

Table des matières

Chapitre 1 — Introduction	1
1.1 Les concepts de base	2
Chapitre 2 — Différenciation des échelles de mesures	7
2.1 Un exemple de questionnaire	8
2.2 Les échelles de mesure	9
2.2.1 Les propriétés des nombres	9
2.2.2 La différenciation des quatre échelles de mesure	10
2.2.3 Quelques commentaires additionnels sur les échelles	14
2.3 La base de données SPSS	16
2.3.1 Avant l'entrée des données	18
2.3.2 Effectuer l'entrée des données	21
2.4 Exercices du chapitre	24
Chapitre 3 — L'analyse univariée	26
3.1 La description d'une variable discrète	26
3.2 La description d'une variable continue	30
3.3 L'inférence liée à une variable discrète	40
3.4 L'inférence liée à une variable continue	48

3.5	Détermination de la taille d'échantillon	52
3.5.1	Variable discrète : proportion	53
3.5.2	Variable continue : moyenne	55
3.6	Tests d'hypothèses sur une moyenne	57
3.6.1	Exemple de calcul d'une <i>p</i> -value	64
3.7	Tests d'hypothèses sur une proportion	66
3.8	Utilisation de l'écart-réduit	70
3.8.1	Tests sur une moyenne	70
3.8.2	Tests sur une proportion	73
3.9	Exercices du chapitre	75
Chapitre 4 — Relation entre deux variables discrètes		78
4.1	La statistique du chi-deux	79
4.1.1	Représentativité de l'échantillon	81
4.2	Les tableaux de contingence	83
4.2.1	Le Chi-deux χ^2 : y a-t-il un lien significatif ?	83
4.2.2	Quelle est la force du lien ?	89
4.2.3	Interprétation du tableau croisé	92
4.3	Le croisement de deux variables ordinaires	95
4.4	Pré-requis : fréquences théoriques	99
4.5	Un autre exemple	101
4.6	L'analyse des correspondances	103
4.7	Exercices du chapitre	110
Chapitre 5 — Relation entre une variable discrète et une continue		112
5.1	Variable dichotomique	113
5.1.1	Pré-requis	116
5.1.2	Comparaison des deux moyennes : Independent Samples T Test .	118
5.2	Variable polychotomique	121

5.2.1	Pré-requis	122
5.2.2	Analyse de la variance : ANOVA	124
5.2.3	Force du lien	127
5.2.4	Comparaisons multiples : analyse Post Hoc	128
5.3	Utilisation de l'écart-réduit	131
5.3.1	Comparaison de deux moyennes	131
5.3.2	Comparaison de deux proportions	137
5.3.3	ANOVA : test de Fisher	140
5.4	Exercices du chapitre	142
Chapitre 6 — Relation entre deux variables continues		144
6.1	L'analyse en corrélation linéaire simple	144
6.1.1	Examen graphique de la relation	148
6.1.2	Limites du coefficient de corrélation	150
6.1.3	Analyse et test d'hypothèses	151
6.2	La régression linéaire simple	154
6.2.1	Le principe des moindres carrés	158
6.2.2	Le modèle statistique de la régression	160
6.2.3	Les paramètres β_0 et β_1	162
6.2.4	Analyse de la variance	167
6.2.5	Utilisation de la droite de régression	170
6.2.6	Introduction aux séries temporelles	184
6.2.7	Association et lien de cause à effet	191
6.3	Régression linéaire avec E-Views	192
6.4	Exercices du chapitre	202
Chapitre 7 — Relation entre trois variables discrètes		204
7.1	Validation de la relation initiale	205
7.1.1	Les relations illusoires	205

7.1.2	Les relations « étouffées »	213
7.1.3	Les relations déformées	217
7.2	Détailler la relation initiale	221
7.3	Exercice du chapitre	225
Chapitre 8 — Relation entre deux variables discrètes et une variable continue		231
8.1	Analyse de la variance à deux facteurs	231
8.2	L'étude des graphiques	241
8.3	Quelques exemples	247
8.4	Exercices du chapitre	257
Chapitre 9 — Relation entre plusieurs variables continues		259
9.1	Le modèle	260
9.1.1	Hypothèses de validité	262
9.2	Une analyse complète	263
9.2.1	Le modèle est-il bon dans son ensemble ?	264
9.2.2	Inférence sur les paramètres du modèle	267
9.2.3	Prédictions et intervalles de confiance	269
9.3	Un autre exemple	271
9.4	Exercices du chapitre	274
Chapitre 10 — Modèles de régression linéaire		275
10.1	Les variables discrètes	275
10.1.1	Les variables discrètes dichotomiques	275
10.1.2	Les variables discrètes polychotomiques	285
10.2	Les variables déphasées	292
10.3	La sélection de variables explicatives	309
10.4	Exercices du chapitre	324

Chapitre 11 — Validité d'un modèle en régression linéaire 326

11.1 Rappel des hypothèses	327
11.2 La multicolinéarité	327
11.2.1 Déetecter la multicolinéarité.	328
11.2.2 Corriger la multicolinéarité	329
11.3 Les résidus	330
11.3.1 Les propriétés des résidus	331
11.4 Vérification de la linéarité	331
11.4.1 Détection de la violation de la linéarité	332
11.4.2 Correction de la violation à la linéarité	338
11.5 Vérification de la variance constante	354
11.5.1 Détection de la variance non constante	354
11.5.2 Correction de la violation à la variance constante	357
11.6 Vérification de la normalité	360
11.6.1 Détection de la violation à la normalité	361
11.6.2 Test d'hypothèses sur la normalité des résidus	366
11.6.3 Correction à la normalité	366
11.7 Les données qui ont beaucoup d'influence	367
11.7.1 Méthodes d'identification des points de types <i>outlier</i> et/ou <i>leverage</i>	371
11.7.2 Que faire avec ces données ?	374
11.8 Vérification de l'indépendance	376
11.8.1 Autocorrélation	376
11.8.2 Détection de l'autocorrélation de premier ordre	377
11.8.3 Correction de l'autocorrélation de premier ordre	382

Chapitre 12 — Les séries temporelles 384

12.1 Quelques généralités	385
12.2 Le choix d'un modèle et taille d'échantillon	386
12.3 Les techniques de moyennes	388

12.3.1	Les moyennes simples	388
12.3.2	Les moyennes mobiles (<i>moving averages</i>)	390
12.3.3	Moyennes mobiles avec E-Views	396
12.4	La décomposition	398
12.4.1	Visualisation de la série	401
12.4.2	Désaisonnalisation des données	402
12.4.3	Détermination de la tendance	411
12.4.4	Détermination de l'effet cyclique	414
12.4.5	Détermination de l'effet irrégulier	420
12.4.6	Établissement des prédictions	425
12.4.7	Décomposition avec E-Views	429
12.5	Le lissage exponentiel	431
12.5.1	Lissage exponentiel simple	432
12.5.2	La méthode de Holt	439
12.5.3	La méthode de Winters	446
12.5.4	Lissage exponentiel avec E-Views	454
12.5.5	La tendance	456
12.6	Les tendances non linéaires	457
12.7	Les données manquantes	467
12.8	Efficacité et comparaison de modèles	472
12.9	Exercices du chapitre	475
Chapitre 13 — Les modèles ARIMA		476
13.1	Introduction	476
13.2	La stationnarité	477
13.3	Le modèle ARIMA de Box-Jenkins	482
13.4	Les fonctions SAC et SPAC	485
13.5	Les comportements théoriques des SAC et SPAC	493
13.6	La méthodologie Box-Jenkins (sans saison)	499

13.6.1	Vérification et obtention de la stationnarité	500
13.6.2	Identification des ordre p et q du modèle ARMA	502
13.6.3	Estimation des paramètres du modèle	503
13.6.4	Validité du modèle	508
13.6.5	Calcul des prédictions	515
13.7	ARIMA sans saisons avec E-Views	544
13.8	Le modèle ARIMA et l'effet des saisons	554
13.9	ARIMA saisonnier avec E-Views	583
13.10	Quelques compléments	587
Chapitre 14 — Les modèles ARCH et GARCH		611
14.1	Formulation du modèle ARCH	611
14.2	Formulation du modèle GARCH	612
14.3	Procédure de création d'un modèle	613
14.4	Un exemple	613
14.5	Autres modèles ARCH	628
14.5.1	Modèle GARCH-M	628
14.5.2	Variables indépendantes dans l'équation de la variance	629
14.5.3	Modèle TARCH	629
14.5.4	Modèle EGARCH	630
14.6	Le test de White	645
Chapitre 15 — Analyse factorielle (analyse en composantes principales)		648
15.1	Généralités	649
15.2	Les étapes d'une ACP	650
15.2.1	La mesure de Kaiser-Meyer-Olkin (KMO)	651
15.2.2	L'extraction de facteurs	652
15.2.3	Interprétation des facteurs	654
15.2.4	Les scores factoriels	656

15.2.5 Un autre exemple et cartes perceptuelles	656
15.2.6 Les cartes perceptuelles	661
15.3 Exercices du chapitre	666
Chapitre 16 — Analyse de la fidélité	667
16.1 Fidélité et validité	667
16.2 Analyse de la fidélité	669
16.2.1 Description des résultats	670
Chapitre 17 — Méthode de classification (<i>clusters analysis</i>)	676
17.1 Généralités	677
17.1.1 La distance et la similarité	677
17.2 La méthode hiérarchique	679
17.2.1 Critères pour combiner deux <i>clusters</i>	681
17.3 La méthode des nuées dynamiques	688
17.4 Exercices du chapitre	704
Chapitre 18 — Analyse discriminante	705
18.1 Introduction	705
18.2 Illustration du concept	707
18.3 Un exemple à deux groupes	713
18.3.1 Survol des données	717
18.3.2 Analyse des différences entre les groupes	718
18.3.3 Estimation des coefficients de la fonction discriminante	721
18.3.4 La classification	724
18.3.5 Le sommaire de la classification	727
18.3.6 Une bonne analyse discriminante	731
18.3.7 Les pré-requis	734
18.4 Un autre exemple	737

18.5 Un exemple à trois groupes	748
18.6 Exercices du chapitre	762
Chapitre 19 — Régression logistique	767
19.1 Le modèle linéaire	768
19.2 Principe du maximum de vraisemblance	770
19.3 La fonction logistique	773
19.4 Un exemple complet	776
19.4.1 Le modèle constant	778
19.4.2 Le modèle est-il bon ?	781
19.4.3 Les détails du modèle	785
19.4.4 Validité du modèle	788
19.4.5 Les courbes ROC	792
19.5 Un autre exemple	796
Chapitre 20 — Régression logistique multinomiale	808
20.1 Les principes de base	808
20.2 L'exemple du chapitre	810
20.3 Exercice du chapitre	823
Chapitre 21 — Modèles d'équations structurelles (MES)	824
21.1 Introduction	824
21.2 Les origines : <i>Path analysis</i>	828
21.3 Le maximum de vraisemblance (MV)	832
21.4 Les mesures d'adéquation	834
21.4.1 Les indices de mesure absolus	837
21.4.2 Les indices incrémentaux	837
21.4.3 Les indices de parcimonie	838
21.5 Concept réflexif ou formatif?	838

21.6 L'analyse factorielle confirmatoire	841
21.6.1 Un exemple	841
21.7 Le modèle structurel	858
21.8 Effets médiateurs et modérateurs	871
21.8.1 Effet médiateur	871
21.8.2 Effet modérateur	875
21.9 Introduction au logiciel AMOS	875
Annexe A — Tables de lois	884
Annexe B — Solutionnaires	888
B.1 Solutions de certains exercices du chapitre 3	888
B.2 Solutions de certains exercices du chapitre 4	893
B.3 Solutions de certains exercices du chapitre 5	899
B.4 Solutions de certains exercices du chapitre 6	907
B.5 Solution de l'exercice du chapitre 7	912
B.6 Solution de l'exercice 3 du chapitre 8	920
B.7 Solution de l'exercice 2 du chapitre 9	925
B.8 Solutions de certains exercices du chapitre 10	931
B.9 Solutions des exercices du chapitre 15	943
B.10 Solutions de certains exercices du chapitre 17	948

Chapitre 1

Introduction

Faire de l'analyse de données s'apparente à délacer une boucle sur un soulier ; tout est simple si on tire sur l'embout du lacet, sinon un double nœud apparaît. La science de la statistique est en fait une façon de penser, une façon de voir les choses, de puiser et d'extraire, de comprendre et de modéliser, de visualiser et de transposer l'information ; voilà le travail d'un statisticien.

Contrairement aux croyances, l'analyse de données prend racine dès l'étape de la planification de l'étude et pendant la préparation du questionnaire qui est l'outil de mesure. À cette étape, il faut préparer et planifier les analyses. Lors de la planification de l'étude, le praticien doit être en mesure de définir la population cible, de dégager avec son client les variables importantes à étudier ainsi que celles susceptibles d'avoir de l'influence sur ces dernières. Lors de la rédaction du questionnaire, il devra tenter de mesurer, par le biais des questions, l'intensité de chacune de ces variables chez les individus. Pour ce faire, il utilisera différentes échelles de mesures et ce sont ces échelles de mesures et non la formulation des questions qui orienteront le praticien sur le choix de la technique statistique à utiliser pour obtenir les réponses recherchées.

Dans le cadre de ces notes, il est proposé au lecteur une approche et une méthode

de travail pour entreprendre et effectuer une analyse de données. Quelques points à ne pas oublier concernant la planification d'une étude seront abordés, notamment la détermination des tailles d'échantillon nécessaires pour que les analyses puissent se dérouler proprement. L'important aspect de la validation des analyses sera aussi à l'ordre du jour.

Ce livre s'adresse plus particulièrement à une clientèle orientée sur la pratique. Pour ce faire, nous utiliserons systématiquement le logiciel SPSS version 14. Les commandes qui amènent le logiciel à réaliser les analyses sont incluses à chacune des techniques présentées.

Ce chapitre présente quelques concepts et notions de base entourant le déroulement d'une enquête précédant l'entrée de données. Bien que nous aborderons quelques sujets inévitables sur le déroulement des enquêtes, mentionnons que cet ouvrage n'est pas un cours de méthodologie à travers lequel les techniques de formulation de question et les techniques d'échantillonnage sont abordées explicitement.

1.1 Les concepts de base

L'objectif d'une étude consiste à recueillir de l'information sur une population cible. Parfois il faut puiser directement l'information sur le terrain (la figure 1.1 illustre toutes les étapes d'une étude classique), parfois l'information est déjà accessible par le biais d'une base de données existante. Bien qu'il soit vrai que les modèles statistiques changent d'expression d'une situation à une autre, l'approche statistique sous-jacente reste immuable.

C'est pourquoi il est courant dans la science de la statistique d'utiliser le terme « individu » au sens large du terme, et ce, peu importe le contexte de l'étude. En effet, le terme individu fait plutôt référence à l'unité de la population qui est observé, que ce soit des humains, des veaux de grains, des bactéries ou encore des boulons en acier. Et ce sont à ces « individus » que s'appliquent les mêmes modèles statistiques.

Les individus étudiés proviennent toujours d'une population d'appartenance. **La po-**

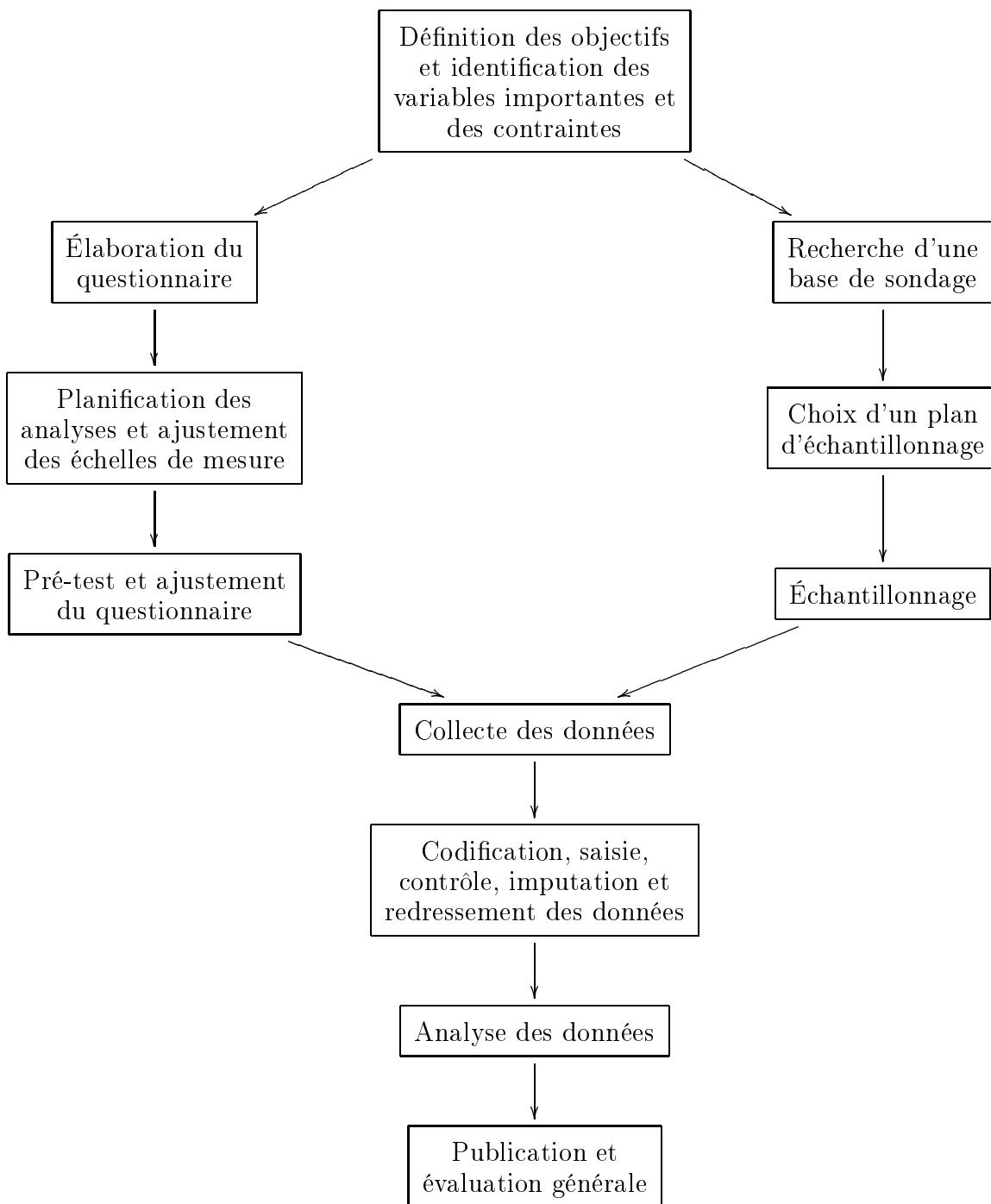


FIG. 1.1 – Les étapes de réalisation d'une étude

**pulation se définit simplement comme l'ensemble de tous les individus aux-
quels l'étude s'intéresse et à qui les résultats se transposeront.** Supposons une étude où un groupe s'intéresse à la rémunération des policiers du Québec. L'ensemble de tous les policiers du Québec forme donc par définition la population cible, tandis que la variable à étudier est le salaire.

Il faut comprendre que les informations recherchées, comme le salaire ou l'âge des individus, varient d'un individu à un autre et c'est pourquoi le praticien utilise le terme « variable ». Par exemple la variable du salaire ou la variable de l'âge des individus. Comme l'humain est complexe, il est facile de comprendre qu'il existe une infinité de variables pouvant être étudiées.

Par chance, parmi toutes les variables susceptibles d'être étudiées, la connaissance des objectifs d'une étude permet au praticien de ne considérer qu'un sous-ensemble de variables. Celui-ci porte le nom de **variables d'intérêts**. Pour formaliser l'information obtenue sur les variables, la science de la statistique quantifie, par le biais de paramètres, le comportement de la variable dans la population cible. Les paramètres de la population sont simplement des valeurs numériques qui permettent de jauger l'amplitude d'une variable dans la population afin de se faire une meilleure idée sur celle-ci.

Par exemple, pour étudier le salaire (la variable) de l'ensemble du corps policier de la province du Québec (la population), le praticien peut s'intéresser au paramètre du salaire moyen (μ_{salaire}) de cette population. Il peut aussi étudier le paramètre de l'écart-type de la population (σ_{salaire}) qui mesure à quel point le salaire de ces individus varie par rapport au salaire moyen. Il peut aussi s'intéresser au paramètre de l'augmentation marginale moyenne de salaire (β_{salaire}) qu'un individu obtient à chaque année d'ancienneté supplémentaire. Il faut bien comprendre que pour une même variable, plusieurs paramètres peuvent être intéressants à connaître. Certains paramètres caractérisent la variable en solitaire (ex. : μ_{salaire} , σ_{salaire}), tandis que d'autres caractérisent la relation qu'elle entretient avec les autres variables (ex. : β_{salaire}). Voici une brève définition d'un paramètre :

Paramètre : Toute valeur numérique pouvant être calculée à partir d'un recensement d'une population (avoir questionné tous les individus d'une population).

Puisque la valeur numérique est issue du recensement d'une population, le paramètre est une valeur numérique fixe. Par exemple le salaire moyen des canadiens $\mu_{\text{salaire canadiens}} = 28\,793 \$$. En statistiques, les paramètres sont représentés algébriquement par des lettres grecques (μ, σ, β, ρ , etc.).

En somme, pour utiliser un paramètre, il faut avoir sous la main la base de données d'un recensement ou les résultats d'un recensement. En général, les données des recensements sont distribuées par les gouvernements et l'information n'est pas toujours gratuite. Parmi les données disponibles, on y retrouve des données générales sur le salaire moyen des individus de la population, sur l'âge moyen, etc.

Cependant, un portefeuille supporte généralement bien mal les coûts d'un recensement. De plus, les organismes gouvernementaux ne disposent pas nécessairement de l'information dont le praticien a besoin pour mener son étude. C'est pourquoi il est courant d'utiliser des échantillons représentatifs de la population. Un échantillon est un sous-ensemble représentatif d'individus de la population. Les individus sont choisis au hasard à partir d'une base de données de la population. L'échantillon agit fondamentalement à titre de photo-réduction de la population, et ce, à une date donnée.

À l'aide de l'échantillon, le praticien propose d'estimer les paramètres de la population par l'entremise des statistiques similaires issues de l'échantillon. Cette transposition de l'échantillon à la population est le principe de l'inférence statistique. Par exemple, pour estimer le paramètre du revenu annuel moyen (μ_{revenu}) de l'éventuelle population d'acheteurs d'un tout nouveau produit, il utilise la statistique de la moyenne échantillonnale (\bar{x}_{revenu}). Voici la définition d'une statistique :

Statistique : Toute valeur numérique pouvant être calculée à partir d'un échantillon.

Il faut bien comprendre que la valeur d'une statistique varie d'un échantillon à un

autre, mais elle variera suivant une loi de probabilité. Par exemple, l'âge moyen des canadiens $\bar{x}_{\text{âge canadiens}}$ estime le paramètre $\mu_{\text{âge canadiens}}$. En somme, avec les statistiques de l'échantillon, on estime les paramètres de la population.

Bien entendu, on utilise les statistiques seulement si les paramètres sont inconnus. En effet, les statistique ne sont que des estimations (des « clones ») des paramètres. Contrairement aux paramètres qui sont représentés par des lettres grecques, les statistiques sont représentées par des lettres latines (\bar{x} , s , r , b , etc.).

Dans l'analyse de données, à l'aide des statistiques, on étudie l'information concernant une variable à la fois (analyse univariée) pour ensuite étudier les relations entre elles. Il existe donc des statistiques qui permettent l'étude de variables seules ou en groupes. Dans le cadre de ces notes, le lecteur sera en mesure d'apprendre à différencier l'utilité des différentes statistiques.

Chapitre 2

Différenciation des échelles de mesures

Il faut comprendre que pour étudier une variable, plusieurs formulations de questions peuvent être élaborées. Il suffit de donner un même mandat à deux groupes pour constater que les deux questionnaires élaborés abordent un même sujet de manière bien différente.

Au lieu de choisir les analyses en fonction de la formulation des questions, le praticien s'inspirera plutôt de l'échelle de mesure liée à cette variable. Contrairement à l'infinité de formulations possibles des questions, les échelles de mesures ne se limitent qu'à quatre : les échelles nominales, ordinales, d'intervalles et de ratios.

Dans le cadre de ce chapitre, nous présentons dans la première section un exemple de questionnaire afin de faire un premier lien entre le questionnaire et l'analyse de données. Dans la seconde section, nous introduisons et différencions chacune des quatre types d'échelles de mesures. Dans la troisième section nous présentons comment entrer les réponses dans une feuille de données SPSS une fois l'étude effectuée.

2.1 Un exemple de questionnaire

À titre d'illustration, supposons qu'une entreprise de chasseurs de têtes désire faire une étude portant sur le salaire de cadres de différentes entreprises privées de même type et de même envergure. Il est probable de trouver les questions suivantes dans un questionnaire :

Q₁ : Quel est votre sexe ?

Masculin₀

Féminin₁

Q₂ : Quel est votre niveau de scolarité ?

Primaire₁

Secondaire₂

Collégial₃

Universitaire₄

Q₃ : Combien d'années d'expérience avez-vous cumulées à ce poste de cadre dans cette entreprise ? _____(ans, mois)

Q₄ : Combien d'années d'expérience avez-vous cumulées à ce type de poste de cadre dans votre carrière ? _____(ans, mois)

Q₅ : Quelle est votre salaire annuel brut ? _____\$

Q₆ : Dans quelle province habitez-vous présentement ?

Québec₁

Ontario₂

Manitoba₃

Alberta₄

Bien d'autres questions peuvent être posées, notamment la zone d'exploitation de l'entreprise (urbaine ou non), le chiffre d'affaires de l'entreprise, les primes à la performance, etc.

2.2 Les échelles de mesure

Poser une question, c'est mesurer la profondeur ou l'intensité d'une variable chez les individus. Pour y arriver, on associe des nombres à des événements. Cette association est construite logiquement de manière à quantifier l'intensité d'une variable (l'événement étudié) chez un individu.

Dans le cadre de cette section nous présentons dans un premier temps les quatre grandes propriétés des nombres. Ces propriétés sont à la base de chacune des échelles de mesures qui seront présentées ensuite et avec lesquelles nous allons travailler tout au long de ce document.

2.2.1 Les propriétés des nombres

En somme, mesurer c'est attribuer des nombres à des événements. Parmi l'ensemble des propriétés des nombres, quatre grandes propriétés nous intéressent plus particulièrement. Elles sont présentées de la plus faible à la plus puissante :

L'identification : 0, 1 et 2 sont des nombres bien différents.

L'ordonnancement : 3 est plus petit que 4.

L'égalité d'intervalle : La longueur de l'intervalle entre les nombres 1 et 3 est la même qu'entre les nombres 98 et 100. La « largeur » de l'intervalle est de 2.

L'égalité des ratios : Le ratio 24/8 est le même que le ratio 15/5.

Ces quatre propriétés définissent complètement les quatre types d'échelles de mesure. Dans l'optique de l'analyse de données, il est important de signaler qu'une propriété supérieure hérite automatiquement des propriétés inférieures.

2.2.2 La différenciation des quatre échelles de mesure

Cette sous-section présente les quatre échelles de mesure sur lesquelles se base la science de la statistique. Rappelons que les échelles associent des nombres aux événements. Les échelles nominales, ordinaires, d'intervalles et de ratios sont présentées dans l'ordre, de la moins à la plus puissante. La puissance pour un statisticien est décrite comme étant le potentiel d'une échelle à détecter, comprendre et exploiter la variabilité d'un phénomène.

L'échelle nominale :

L'échelle nominale n'utilise les nombres que pour différencier et identifier les modalités de réponses d'une question. Cette échelle est discrète au sens où il n'y a rien « entre » les modalités de réponse. Voir dans l'exemple d'introduction la question portant sur le sexe.

Q₁ : Quel est votre sexe ?

Masculin₀

Féminin₁

Une des codifications possibles est la suivante : 0 = Masculin et 1 = Féminin. Il faut comprendre que les nombres 0 et 1 ne servent qu'à différencier les deux modalités au même titre que 1 est différent de 0. De plus, il n'existe pas de catégorie d'individu « 0,5 ».

Toute autre association entre les réponses et les nombres aurait aussi été correcte, par exemple 1 = Masculin et 0 = Féminin, au même titre que 12 = Masculin et 64,2 = Féminin. Cependant, dans le cadre de ce cours, lorsque la question présente seulement deux modalités de réponses, il est préférable d'utiliser les nombres 0 et 1 dans la codification. En effet, d'un point de vue mathématique, cette codification de type « on/off » (binaire) est très utile dans certaines analyses statistiques, le groupe « 0 » agissant mathématiquement à titre de groupe de référence.

Lorsque la question présente trois modalités de réponses ou plus, la codification de

type « on/off » perd toute son utilité mathématique et le praticien peut utiliser les nombres 0, 1, 2, 3, ... pour codifier les modalités de la question, et ce, sans regard à un groupe de référence.

La question 6 est elle aussi de type nominal :

Q₆ : Dans quelle province habitez-vous présentement ?

- Québec₁
- Ontario₂
- Manitoba₃
- Alberta₄

Le praticien peut associer 1 = Québec, 2 = Ontario, 3 = Manitoba et 4 = Alberta. L'association des codes (1, 2, 3 et 4) est arbitraire et aurait pu être fixée tout autrement par un autre analyste.

L'échelle nominale est la plus faible des quatre et elle ne sert essentiellement qu'à identifier les groupes. Cette responsabilité est cependant très importante puisqu'il est courant de comparer deux ou plus de deux groupes les uns par rapport aux autres.

L'échelle ordinale :

L'échelle ordinale possède la propriété d'identification à laquelle s'additionne la propriété d'ordonnancement. Cette échelle est discrète. Ici, l'ordonnancement des nombres de l'association a une importance logique et sera exploitée. La question posée sur le niveau de scolarité en est un exemple.

Q₂ : Quel est votre niveau de scolarité ?

- Primaire₁
- Secondaire₂
- Collégial₃
- Universitaire₄

En suivant l'association des codes écrits en indice, plus le code est petit, plus le niveau de scolarité est bas. Cependant, les codes n'informent pas exactement sur le nombre d'années de scolarité qui différencient vraiment deux personnes. Par exemple, une personne

ayant un baccalauréat et une autre ayant un doctorat obtiennent toutes deux un code 4. Parallèlement, un autre analyste aurait pu associer logiquement un ordonnancement inverse dans sa codification, par exemple :

Q₂ : Quel est votre niveau de scolarité ?

Primaire₄

Secondaire₃

Collégial₂

Universitaire₁

Afin de mieux comprendre la portée restreinte de cette échelle, voyons un exemple supplémentaire. Supposons que la question suivante ait été posée aux individus de notre exemple d'introduction :

Q₇ : Le poste que vous occupez se doit de comporter beaucoup de pression.

Tout à fait en accord₄

En accord₃

En désaccord₂

Tout à fait en désaccord₁

Ce type d'échelle porte le nom d'échelle de Likert en quatre points illustrant différents « niveaux d'accord ». De façon semblable, il existe aussi des échelles de style Likert en 7 points. En suivant les codifications écrites en indice, il est possible de voir que plus la réponse d'un individu est associée à un code élevé, plus il est en accord avec l'affirmation. Il faut comprendre la limite de cette mesure puisque le code n'informe en rien sur la puissance de la différence qui existe entre l'opinion d'un individu « En accord » par rapport à un autre « Totalement en accord ».

L'échelle d'intervalle :

Cette échelle cumule les propriétés d'identification, d'ordonnancement et d'égalité d'intervalles, et nous informe de façon précise de l'écart entre deux modalités (réponses). La première question traitant de l'expérience des cadres en est un exemple :

Q₃ : Combien d'années d'expérience avez-vous cumulées à ce poste de cadre dans cette entreprise ? _____ (ans, mois)

La réponse des individus est un nombre d'années additionné d'une fraction d'année (nombre de mois/12 additionné *a posteriori* par l'analyste) représentant ainsi le temps passé à titre de cadre dans l'entreprise. Ainsi on sait exactement quelle est la différence d'expérience entre deux individus. Aussi, il faut comprendre que lorsque l'individu répond « 0 ans » à cette question, ceci n'implique pas que l'individu ne possède aucune expérience pertinente. En effet, il peut avoir travaillé à un poste similaire de cadre ailleurs. Le zéro d'une échelle d'intervalle n'est pas absolu, il est relatif. Ceci fait que les ratios n'ont pas de sens avec une telle échelle.

Par exemple, l'intervalle entre 1995 et 2000 a un sens, mais pas le ratio 2000/1995. L'an 0 n'est pas un zéro absolu. Donc la variable temporelle des années est mesurée avec l'échelle d'intervalle.

L'échelle d'intervalle peut mesurer des variables continues. La question Q₃ est une variable continue : toutes les valeurs en terme d'années d'expérience sont possibles. L'échelle d'intervalle est plus nuancée et exprime mieux la variabilité de l'expérience d'un individu à un autre. De plus, la moyenne d'années d'expérience de tous les cadres a un sens bien concret, c'est là une grande particularité de ce type d'échelle.

L'échelle d'intervalle peut aussi mesurer des variables discrètes, en autant que celles-ci cumulent les propriétés de cette échelle (de nombreuses variables discrètes ne donnent pas de façon précise l'écart entre deux modalités et ne peuvent donc pas être associées à cette échelle).

L'échelle de ratio :

L'échelle de ratio cumule toutes les propriétés des échelles précédentes. Elle s'apparente énormément à l'échelle d'intervalles mais possède la propriété d'avoir un zéro absolu, ce qui fait que les ratios sont significatifs. La seconde question sur l'expérience ainsi que la question sur le salaire en sont des exemples :

Q₄ : Combien d'années d'expérience avez-vous cumulées à ce type de poste de cadre dans votre carrière ? _____ (ans, mois)

Q₅ : Quelle est votre salaire annuel brut ? _____ \$

Un individu qui note « 0 an » à la question 4 illustre simplement une absence d'expérience à ce type de poste. Un individu qui note 0 de salaire (ce qui est logiquement improbable) illustre une absence de salaire versé par cette entreprise. En somme, le zéro possède ici un caractère absolu. Tout comme pour l'échelle d'intervalles, la notion de moyenne (ici le salaire moyen des cadres) a un sens concret. Et un cadre peut être deux fois mieux payé qu'un autre (ratios significatifs).

Cette échelle peut mesurer des variables discrètes et continues, de la même façon que l'échelle d'intervalles (c'est-à-dire qu'avant de mesurer une variable discrète avec cette échelle il faut s'assurer qu'elle possède toutes les propriétés mentionnées ci-dessus).

2.2.3 Quelques commentaires additionnels sur les échelles

Dans la pratique et dans ces notes, les échelles nominales et ordinaires sont regroupées dans la famille des variables discrètes tandis que les échelles d'intervalles et de ratios peuvent mesurer les deux types de variables (discrète et continue). On identifie le type de variable au type d'échelle (on dira par exemple une variable nominale).

Bien qu'il existe quelques différences entre les analyses des variables nominales et ordinaires, il n'en est pas de même pour les analyses des variables d'intervalles et de ratio. Dans le cadre de ce cours, nous présenterons les analyses de base de la famille des variables discrètes et, lorsque nécessaire, nous mettrons en évidence les différences entre les analyses des échelles nominale et ordinaire. Nous traiterons sans distinction les analyses statistiques des échelles d'intervalles et de ratio.

Compte tenu que les échelles plus puissantes cumulent les propriétés des échelles de moindre puissance, il sera toujours possible pour le praticien de briser une variable continue en variable discrète, et ce, sans problème. Cependant, certains auteurs soutiennent

que les variables discrètes ordinaires peuvent être interprétées comme étant continues. Il soutiennent que la moyenne des réponses d'une telle variable a un sens bien pratique. Voici un exemple où la moyenne des réponses des individus pourrait être exploitée.

Q₇ : Le poste que vous occupez se doit de comporter beaucoup de pression.

Tout à fait en accord₄

En accord₃

En désaccord₂

Tout à fait en désaccord₁

Supposons que la moyenne des réponses des individus se situe à 3,2 (entre 3 = en accord et 4 = tout à fait en accord, mais plus près de 3). Cela pourrait s'interpréter comme suit : « L'ensemble des répondants a tendance à être plus qu'en accord avec l'affirmation de la question Q₇ ». Mais ce n'est pas si simple, car comme nous le verrons, une moyenne n'est rien sans son écart-type. Mathématiquement, cette position accorde trop de puissance dans la tractation de la variation qu'une variable discrète ordinaire ne possède vraiment. C'est un peu comme prétendre que des poutrelles de carton peuvent remplacer des poutrelles d'acier sous un pont ; l'illusion peut parfois être parfaite.

Dans les faits, ce type d'action génère plus souvent qu'autrement des problèmes de validation dans plusieurs analyses qui ont été conçues pour accueillir de véritables variables continues. Il est possible d'atténuer le problème en augmentant le nombre de modalités de réponses de manière à mieux traquer la variation. La morale : plus on a de modalités, moins on a de problèmes. La question ci-dessous en est un exemple.

Q₈ : Le poste que vous occupez se doit de comporter beaucoup de pression.

Tout à fait en accord	<input type="checkbox"/> ₇ <input type="checkbox"/> ₆ <input type="checkbox"/> ₅ <input type="checkbox"/> ₄ <input type="checkbox"/> ₃ <input type="checkbox"/> ₂ <input type="checkbox"/> ₁	Tout à fait en désaccord
--------------------------	---	-----------------------------

L'optique que nous proposons dans ce cours est la suivante : si la moyenne d'une variable ordinaire a vraiment un sens pratique, nous l'utiliserons dans nos analyses à titre de variable continue. Cependant, jouer avec le feu n'est pas toujours sans conséquences. De fait, afin de se protéger des conclusions négatives, des analyses de validités seront

faites systématiquement et nous allons nous rétracter à la moindre problématique.

Mais la meilleure des solutions consiste à utiliser une échelle de mesure continue. La formulation suivante présente une solution intéressante qui a fait ses preuves.

Q₈ : Le poste que vous occupez se doit de comporter beaucoup de pression (mettre un X sur la partie de la droite qui correspond le mieux à votre opinion, en sachant que le 0 correspond à tout à fait en désaccord, et le 10 à tout à fait en accord).



Comme illustré dans cet exemple, l'individu inscrit un « x » sur la droite représentant ainsi son opinion. Une fois le x inscrit par le répondant, le praticien mesure, à l'aide d'une règle, la distance en cm (ou autre unité de mesure) entre le niveau de « Tout à fait en désaccord » (0 cm) jusqu'à l'endroit où le x coupe la droite. C'est justement cette valeur (en cm) qui sera inscrite dans le fichier de données à titre de réponse de l'individu. En somme, plus la distance en cm est élevée, plus l'individu est en accord avec l'affirmation.

Cette technique produit une mesure continue (toutes les valeurs sont possibles, il suffit d'avoir une règle suffisamment précise) et la moyenne a bien un sens. De plus cette technique ne change pas la validité ni la fiabilité d'un questionnaire dans lequel on remplacerait les échelles de Likert par ce type d'échelle continue.

2.3 La base de données SPSS

Une fois les questionnaires complétés par un nombre de répondants, il est maintenant temps d'entrer les réponses dans la base de données SPSS. La figure 2.1 illustre la fenêtre d'une base de données vierge, appelée le **Data View**. Cette fenêtre apparaît par défaut lors du démarrage de SPSS.

Avant d'entrer les données et pour obtenir la configuration finale de la base de données, il est important de définir dans un premier temps à SPSS la codification afin que celle-ci

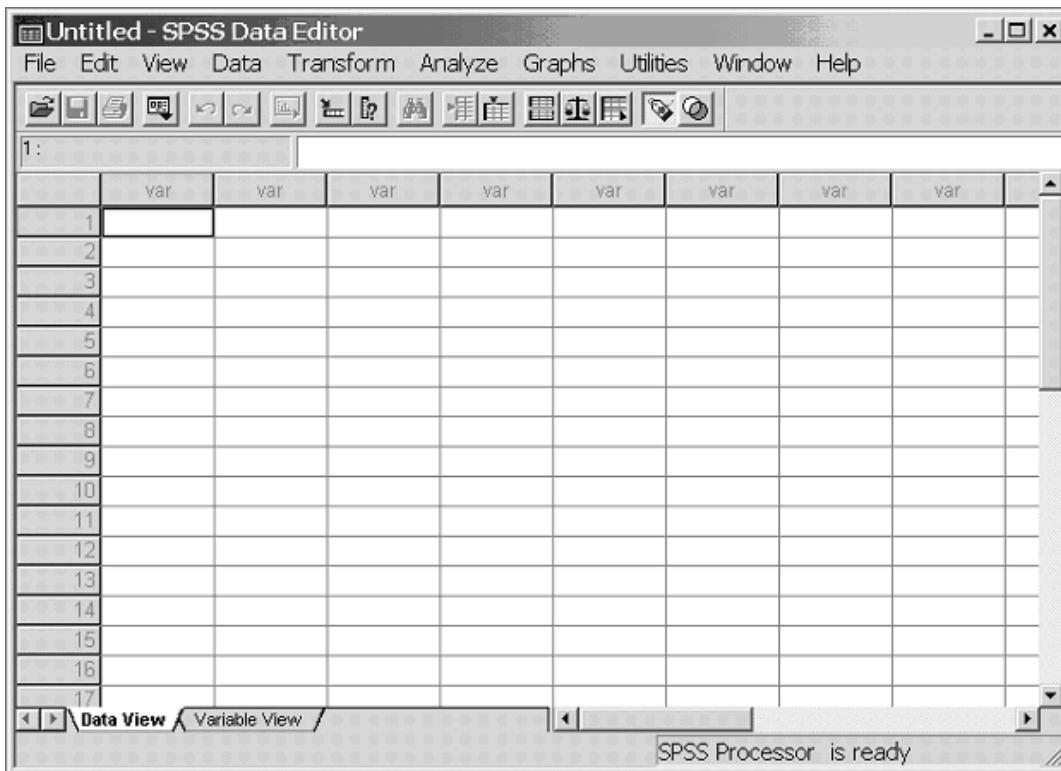


FIG. 2.1 – Fenêtre Data View

facilite la tâche de l'entrée de données. Par exemple, au lieu d'écrire systématiquement les mots « Masculin » et « Féminin », la codification permet de définir des associations simples telles $0 = \text{Masculin}$ et $1 = \text{Féminin}$. Cette association permet ensuite à l'analyste de taper dans la bonne colonne de la base de données les nombres 0 ou 1 pour voir apparaître automatiquement les associations « Masculin » ou « Féminin ».

Mentionnons que la codification peut s'effectuer avant ou après l'entrée de données. Toutefois, il est plus simple de la faire avant. Dans SPSS, l'ordonnancement des colonnes dans le fichier de données n'a aucun impact sur le déroulement, ni sur les résultats des analyses statistiques. Pour codifier les données, il suffit de suivre les instructions incluses dans la sous-section 2.3.1.

2.3.1 Avant l'entrée des données

Dans un premier temps il faut définir à SPSS les variables (les questions) qui seront incluses dans la base de données. En effet, SPSS a besoin de connaître le nom des variables, de savoir à quelle échelle cette variable est associée, etc.

Par l'entremise du menu Démarrer de Windows, ouvrez l'application SPSS et, une fois l'application ouverte, cliquez sur l'onglet de passage **Variable View** (voir figure 2.2), de manière à faire apparaître la fenêtre **Variable View** telle qu'illustrée par la figure 2.3.



FIG. 2.2 – L'onglet de passage de la fenêtre Data View à Variable View

Une fois dans la fenêtre **Variable View**, il faut définir les variables. Il suffit de consulter la figure 2.3 basée sur notre exemple d'introduction. C'est dans cette fenêtre que le praticien spécifie à SPSS la codification à utiliser (ex. : 0 = Masculin et 1 = Féminin).

Dans la fenêtre **Variable View**, chaque ligne représente une nouvelle question (une variable à être étudiée). Afin que SPSS s'y retrouve, il est important de définir chacune des questions du questionnaire. Côté pratique, il faut respecter l'ordre dans lequel se retrouve les questions dans le questionnaire. Cette stratégie ne change rien aux analyses mais facilite grandement l'étape de l'entrée des réponses dans la base de données.

À titre de règle de travail, il est important de définir la première ligne avec le terme **ident** ; il s'agit du numéro d'identification du questionnaire ; c'est l'unique lien qui restera entre la base et les questionnaires.

Voici quelques précisions utiles sur chacun des éléments de la fenêtre **Variable View** :

La colonne **Name** : Écrire le nom de la variable. Ce nom ne doit pas dépasser

8 lettres (sauf à partir de la version 12) et c'est le nom qui apparaîtra en haut de la colonne dans votre base de données.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	ident	Numeric	8	0	Numéro d'identification	None	None	8	Right	Nominal
2	sexe	Numeric	8	0	Sexe de l'individu	{0, Masculin}...	None	8	Right	Nominal
3	scolar	Numeric	8	0	Niveau de scolarité	{1, Primaire}...	None	8	Right	Ordinal
4	expert	Numeric	8	2	Nombre d'années d'expérience dans l'entreprise	None	None	8	Right	Scale
5	salaire	Numeric	8	2	Salaire annuel brut	None	None	8	Right	Scale
6	province	Numeric	8	0	Province de résidence	{1, Québec}...	None	8	Right	Nominal
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										

FIG. 2.3 – Définition des variables

- La colonne **Type** : Précisez la nature des données. Par exemple, vous pouvez spécifier qu'il s'agit de données en \$ ou en %. Par défaut, l'option est fixée à **numeric**. Si vous voulez utiliser le point du pavé numérique pour les décimales, vous devez choisir le type **Comma**. Sinon vous devez utiliser la virgule du clavier.
- La colonne **Width** : Précisez le chiffre significatif à partir duquel SPSS arrondira. Par défaut, cette valeur est fixée à 8 ; l'augmenter au besoin.
- La colonne **Decimal** : Précisez le nombre de décimales qui apparaîtront après la virgule. Par défaut, cette valeur est fixée à 2. Par exemple, dans ce cas, 2,577 deviendra 2,58.
- La colonne **Label** : Écrire au long la signification ou la définition de la variable. Cette colonne est de loin la plus importante. En présence d'un questionnaire, il est recommandé de recopier entièrement les libellés des questions. Vous pouvez utiliser

les commandes « copier/coller » à partir d'un document Word ou autre application Windows.

La colonne Value : Cette colonne permet de définir des associations entre des nombres et des expressions. Par exemple : 0 = Masculin, 1 = Féminin. Le logiciel supporte facilement une centaine d'associations différentes. Une fois la codification définie, elle permettra à l'analyste de simplement taper le « 1 » dans la colonne **sexé** de la fenêtre **Data View** pour faire apparaître automatiquement dans la base de données le mot **Féminin**. Pour définir les codes liés à la variable **sexé**, cliquez une fois dans le champ de la case **Value** : un petit carré gris apparaîtra dans ce champ. Cliquez sur ce carré gris, ce qui vous permettra d'accéder à une nouvelle fenêtre appelée **Value Labels**. Dans les emplacements **Value** et **Value Label** respectivement, tapez :
Value : 0 Value Label : Masculin cliquez : **Add**
Value : 1 Value Label : Féminin cliquez : **Add**

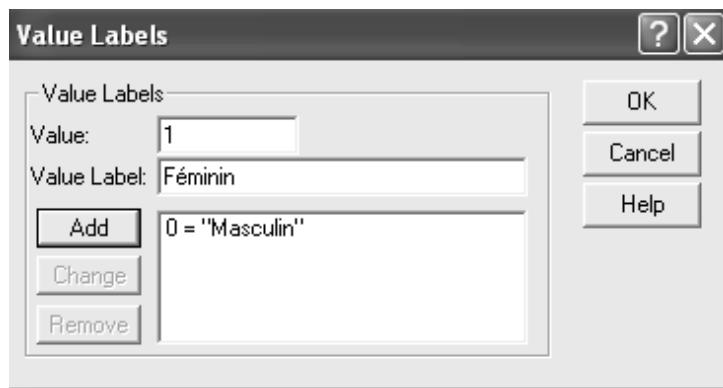


FIG. 2.4 – La fenêtre **Value Labels**

La colonne Missing : Cette colonne permet de définir des codes spéciaux pour identifier les valeurs manquantes. Dans certaines études, il

faut spécifier des codes tels que 99 = « Ne sais pas », 999 = « Ne veut pas répondre », etc. Cette stratégie permet de recenser les raisons liées à la non-réponse. La valeur par défaut est « none ». Dans la plupart des études, en présence d'une valeur manquante, l'analyste laisse simplement la case vide.

La colonne **Column** : Précisez la largeur « physique » (visuelle) de votre colonne. La largeur par défaut est de 8 caractères.

La colonne **Align** : Précisez l'alignement de vos données dans votre fichier. Pour changer l'alignement, cliquez une fois dans le champ de la case **Align**, un petit carré gris apparaîtra dans ce champ. Cliquez sur le carré, et choisissez le type d'alignement.

La colonne **Measure** : Cette colonne permet à l'analyste de spécifier l'échelle de mesure de qu'il compte utiliser. La spécification de l'échelle modifie certaines analyses statistiques. Pour spécifier le type d'échelle lié à la question, cliquez une fois dans le champ de la case **Measure**, un petit carré gris apparaîtra dans ce champ. Cliquez sur le carré, et choisissez le type de mesure s'adaptant à la variable étudiée : **Nominal** pour une variable nominale, **Ordinal** pour une variable ordinale et **Scale** pour une variable d'intervalles ou de ratio.

2.3.2 Effectuer l'entrée des données

Pour accéder à la base de données à partir de la fenêtre **Variable View**, il faut appuyer sur l'onglet de passage **Data View** en base de page. Attention, dans la base de données, chaque colonne représente une question (préalablement définie et dans l'ordre présenté dans la fenêtre **Variable View**). Les réponses d'un individu doivent toutes être

inscrites sur une seule et même ligne. La touche TAB du clavier d'ordinateur permet de changer de colonne. En somme, si 15 individus ont répondu à 5 questions, vous obtiendrez une matrice contenant 15 lignes et 6 colonnes (la colonne d'identification et une colonne pour chacune des questions posées).

À titre d'exemple, basé sur la figure 2.5, l'individu ayant complété le questionnaire numéro 1 est un homme, son niveau de scolarité est le secondaire, il a répondu « 3 ans » à la question sur l'expérience dans l'entreprise actuelle, il gagne présentement un salaire de 60 200 \$ et il demeure au Québec.

	ident	sexe	scolar	expert	salaire	province	var	var
1	1	Masculin	Secondaire	3,00	60200,00	Québec		
2	2	Masculin	Collégial	4,00	63200,00	Québec		
3	3	Féminin	Secondaire	1,50	50000,00	Manitoba		
4	4	Féminin	Collégial	2,50	55000,00	Alberta		
5	5	Féminin	Collégial	1,00	56000,00	Québec		
6	6	Masculin	Collégial	5,00	65300,00	Ontario		
7	7	Masculin	Universitaire	4,00	68000,00	Manitoba		
8	8	Masculin	Universitaire	7,00	87000,00	Ontario		
9	9	Masculin	Collégial	8,00	78500,00	Ontario		
10	10	Féminin	Collégial	3,00	61000,00	Québec		
11	11	Masculin	Universitaire	4,00	67500,00	Alberta		
12	12	Féminin	Universitaire	2,00	53000,00	Manitoba		
13	13	Féminin	Universitaire	2,50	65300,00	Québec		
14	14	Masculin	Universitaire	4,00	68700,00	Alberta		
15	15	Masculin	Universitaire	5,50	89500,00	Ontario		
16								
17								
18								

FIG. 2.5 – Données saisies dans SPSS - bouton Value Label

Prenez note que l'identification ou la numérotation, des questionnaires s'effectue généralement *a posteriori*, à la fois sur les questionnaires et dans la base de données par l'entremise de la colonne **ident**, généralement à l'étape de l'entrée des données. Cette numérotation est importante puisqu'elle suivra intimement le questionnaire tout au long

de l'étude, et ce, malgré les tris de données.

Il faut savoir que les numéros en marge gauche inscrits par SPSS ne sont aucunement liés au questionnaire d'un individu ; au moindre exercice de tri, tout lien est perdu. Il faut aussi savoir que certaines procédures (telle `Merge Files`) effectuent un tri automatiquement. D'où l'importance d'utiliser une colonne d'identification des questionnaires qui permettra de retracer et de corriger, en retracant la bonne réponse dans le bon questionnaire, certaines erreurs d'entrée de données.

Si les codes (ex : Masculin) n'apparaissent toujours pas et que les associations ont pourtant été définies à l'étape précédente dans la fenêtre `Variable View`, appuyez simplement sur le bouton `Value label` de la fenêtre `Data View` (voir la figure 2.5).

2.4 Exercices du chapitre

Exercice 1 Une étude est menée sur les différences de salaires entre les cadres de trois provinces. Pour ce faire, trente cadres, dont les tâches et responsabilités sont similaires, sont choisis au hasard. Vous recevez, par télécopieur, le fichier « Tableau initial des données ». Pour faciliter votre travail d'analyste, vous devez codifier les données et obtenir le fichier « Tableau codifié désiré ».

Tableau initial des données

Tableau codifié désiré

	ident	Province	sexe	salaire		ident	Province	sexe	salaire
1	1	1	0	37500,30	1	1	Alberta	Féminin	37500,30
2	2	1	0	34000,52	2	2	Alberta	Féminin	34000,52
3	3	1	1	32200,00	3	3	Alberta	Masculin	32200,00
4	4	1	0	29650,00	4	4	Alberta	Féminin	29650,00
5	5	1	1	44700,00	5	5	Alberta	Masculin	44700,00
6	6	1	1	43729,56	6	6	Alberta	Masculin	43729,56
7	7	1	0	36540,00	7	7	Alberta	Féminin	36540,00
8	8	1	1	47600,00	8	8	Alberta	Masculin	47600,00
9	9	1	1	53200,00	9	9	Alberta	Masculin	53200,00
10	10	1	1	47000,00	10	10	Alberta	Masculin	47000,00
11	11	2	0	54000,00	11	11	Ontario	Féminin	54000,00
12	12	2	0	61300,00	12	12	Ontario	Féminin	61300,00
13	13	2	1	73400,00	13	13	Ontario	Masculin	73400,00
14	14	2	1	66000,00	14	14	Ontario	Masculin	66000,00
15	15	2	0	49700,00	15	15	Ontario	Féminin	49700,00
16	16	2	0	56500,00	16	16	Ontario	Féminin	56500,00
17	17	2	1	62500,00	17	17	Ontario	Masculin	62500,00
18	18	2	1	61900,00	18	18	Ontario	Masculin	61900,00
19	19	2	0	58750,00	19	19	Ontario	Féminin	58750,00
20	20	2	1	63900,00	20	20	Ontario	Masculin	63900,00
21	21	3	0	45300,00	21	21	Québec	Féminin	45300,00
22	22	3	1	46600,00	22	22	Québec	Masculin	46600,00
23	23	3	0	42090,00	23	23	Québec	Féminin	42090,00
24	24	3	0	41000,00	24	24	Québec	Féminin	41000,00
25	25	3	0	39870,00	25	25	Québec	Féminin	39870,00
26	26	3	1	49450,00	26	26	Québec	Masculin	49450,00
27	27	3	1	47375,00	27	27	Québec	Masculin	47375,00
28	28	3	0	38800,00	28	28	Québec	Féminin	38800,00
29	29	3	1	54980,00	29	29	Québec	Masculin	54980,00
30	30	3	1	46537,00	30	30	Québec	Masculin	46537,00

Exercice 2 Dites si les variables suivantes sont discrètes ou continues, et quelle échelle de mesure leur correspond.

1. Quelle est votre année de naissance? _____

Variable _____

Échelle de mesure _____

2. L'utilisation de Quickplace par les professeurs est une bonne idée.

Tout à fait en accord

En accord

En désaccord

Tout à fait en désaccord

Variable _____

Échelle de mesure _____

3. Quelle est votre nationalité? _____

Variable _____

Échelle de mesure _____

4. Quel est votre revenu brut annuel? _____

Variable _____

Échelle de mesure _____

5. Quel est votre revenu brut annuel?

Moins de 20 000 \$

De 20 000 \$ à moins de 40 000 \$

De 40 000 \$ à moins de 60 000 \$

60 000 \$ et plus

Variable _____

Échelle de mesure _____

Chapitre 3

L'analyse univariée

Les analyses univariées décrivent une variable à la fois, mettant de côté toutes les relations possibles avec les autres variables. Ces techniques d'analyses de données primaires poursuivent deux objectifs : décrire et inférer.

On utilise la description pour synthétiser, résumer et structurer l'information contenue dans les données. Pour cela, on utilise des représentations de données sous forme de tableaux, de graphiques ou d'indicateurs numériques. Entre autres, c'est lors de cette phase qu'on évalue la tendance centrale, la variation et la forme de la distribution des données. D'autre part, on utilise l'inférence pour étendre à la population entière les propriétés constatées sur l'échantillon et pour valider ou infirmer les hypothèses établies a priori ou formulées après une phase de description.

3.1 La description d'une variable discrète

Les variables mesurées à l'aide d'échelles nominales ou ordinaires sont des variables discrètes. Par la suite, nous utiliserons le terme variable discrète pour désigner les va-

riables nominales et ordinaires. Pour décrire et résumer les réponses, le praticien utilise le dénombrement, le mode et les tableaux de distribution de fréquences (diagramme en bâtons ou Bar Charts).

Considérons un exemple d'une course au leadership pour un nouveau parti politique. Une étude est commandée afin de savoir quel candidat, parmi les quatre candidats en liste, a le plus grand potentiel d'impact sur la population. Une firme de consultants mène donc une étude auprès de 400 électeurs répartis un peu partout en province. Une question possible dans ce sondage pourrait bien être la suivante :

- Parmi les candidats suivants, lequel voudriez-vous élire comme chef du parti ?

Candidat 1 (158 votes)

Candidat 2 (175 votes)

Candidat 3 (51 votes)

Candidat 4 (16 votes)

Les nombres entre parenthèses illustrent la répartition finale des réponses des électeurs questionnés. Le fichier SPSS traitant cette étude contient deux colonnes. Plus précisément, une colonne d'identification **ident** (inutile aux traitements statistiques mais nécessaire pour retracer les erreurs de saisie) et une colonne représentant la variable à l'étude pouvant être appelée **candidat**. Le fichier SPSS contient 400 lignes. La figure 3.1 illustre un extrait de la base de donnée SPSS.

	ident	candidat
1	1	candidat 1
2	2	candidat 2
3	3	candidat 2
4	4	candidat 3
5	5	candidat 2
6	6	candidat 4
7	7	candidat 2
8	8	candidat 3
9	9	candidat 1
10	10	candidat 2
11	11	candidat 2
12	12	candidat 2
13	13	candidat 1
14	14	candidat 2
15	15	candidat 4
16	16	candidat 1
17	17	candidat 2

FIG. 3.1 – Extrait de la base de données de l'exemple

Pour être en mesure de connaître la compilation des votes pour chacun des candidats, il faut dégager un tableau de répartition des fréquences. Pour obtenir la sortie 3.2, il faut effectuer les commandes SPSS suivantes :

Menu SPSS :	→ Analyse
	→ Descriptive statistics
	→ Frequencies
Bouton Charts... :	→ Chart type : <input checked="" type="checkbox"/> Bar charts
	→ Chart values : <input checked="" type="checkbox"/> Percentages

Ensuite, sélectionner la variable **candidat** et l'envoyer dans la fenêtre **Variable(s)** (en utilisant la flèche ou simplement en double-cliquant).

Le tableau de la distribution des fréquences (sortie 3.2) met en évidence la popularité de chacun des candidats. On remarque facilement que deux candidats se démarquent des

autres : le candidat 1 avec 158 votes sur 400, représentant 39,5 % des votes, et le candidat 2 avec 175 votes sur 400, représentant 43,8 % des votes.

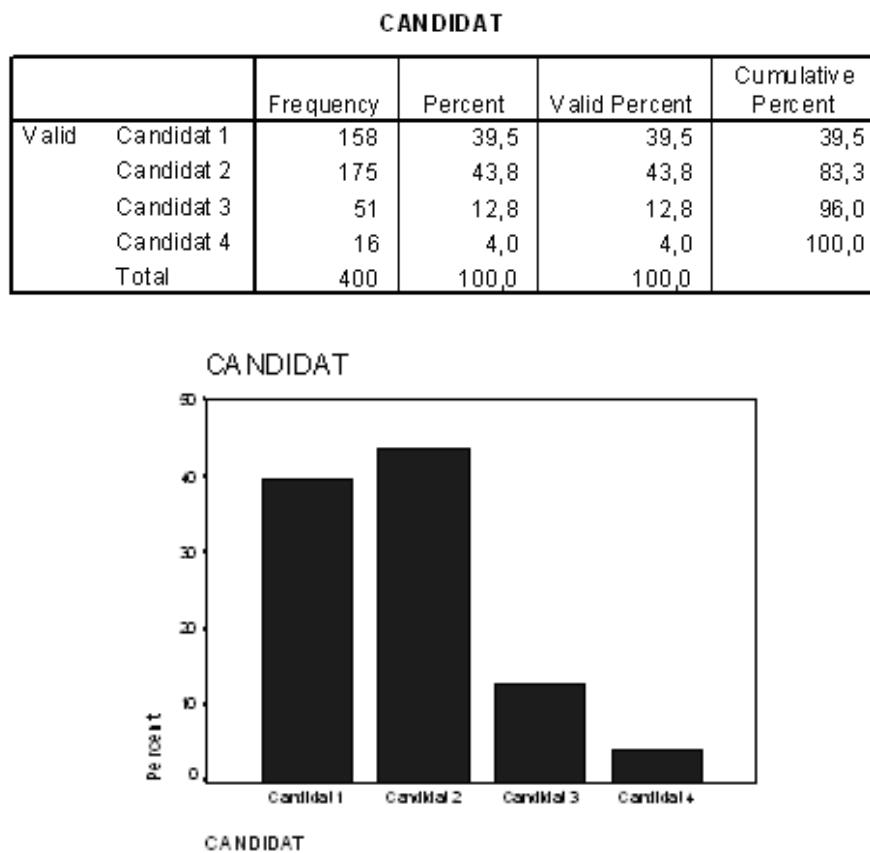


FIG. 3.2 – Distribution des fréquences et diagramme en bâtons

La colonne Valid Percent ajuste les % en excluant les valeurs manquantes (appelées *missing values*). Lorsque les colonnes Percent et Valid Percent sont égales, le fichier de données ne contient pas de valeurs manquantes (une cellule vide dans le fichier). Une valeur manquante peut simplement être une personne rejointe et indécise qui ne fournit aucune réponse.

En se basant sur l'échantillon, l'analyste est en droit de dire que ponctuellement, on peut estimer la proportion de la population en faveur du candidat 2 (le paramètre $\Pi_{\text{candidat } 2}$) à 43,8 %.

Au niveau descriptif d'une variable discrète, il est rare pour un praticien d'effectuer

d'autres analyses.

3.2 La description d'une variable continue

La plupart des variables d'intervalles ou de ratio sont des variables continues ; dans ce qui suit, pour simplifier les choses, on parlera de variables continues pour désigner les variables d'intervalles ou de ratio.

Lorsqu'on veut décrire et résumer une variable continue, on utilise un ensemble de statistiques qui mesurent trois grandes caractéristiques :

- Les statistiques mesurant la tendance centrale ;
- Les statistiques mesurant la dispersion ;
- Les statistiques mesurant la forme.

Le tableau 3.3 contient quelques statistiques servant à la description ainsi que les paramètres de la population qu'elles estiment respectivement. Comme les statistiques varient d'un échantillon à l'autre, il est pertinent que le praticien considère à la fois plusieurs statistiques qui estiment un même paramètre. Le praticien peut alors les comparer et voir si elles semblent être de bons estimateurs ou non (on espère qu'elles soient semblables).

Les paramètres de la population les plus recherchés sont la moyenne μ , la variance σ^2 et l'écart-type σ . Ces paramètres s'expriment de la façon suivante :

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i = \frac{1}{N} (X_1 + X_2 + \cdots + X_N) ,$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2 = \frac{1}{N} [(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_N - \mu)^2]$$

où N est la taille de la population et X_i dénote la i^e observation. L'écart-type est la racine carrée de la variance : $\sigma = \sqrt{\sigma^2}$.

Mesures de tendance centrale, estimateurs de μ	Mesures de dispersion, estimateurs de σ et σ^2	Mesures de la forme de la population	Autres mesures
Moyenne \bar{x} <i>(Mean)</i>	Variance s^2 <i>(Variance)</i>	Asymétrie <i>(Skewness)</i>	Minimum <i>(Minimum)</i>
Médiane <i>(Median)</i>	Écart-type s <i>(Std. Deviation)</i>	Aplatissement <i>(Kurtosis)</i>	Maximum <i>(Maximum)</i>
Moyenne tronquée 5 % <i>(5 % Trimmed Mean)</i>	Intervalle interquartile <i>(Interquartile Range)</i>		Coefficient de variation (CV)

FIG. 3.3 – Les statistiques mesurant certains paramètres

Par exemple, supposons que l'on a une population de quatre personnes qui nous donnent leurs gains de la semaine dernière à la loto : $X_1 = 0 \$$, $X_2 = 2 \$$, $X_3 = 6 \$$ et $X_4 = 8 \$$. Alors, la moyenne de gain pour cette population est de :

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i = \frac{1}{4}(0 \$ + 2 \$ + 6 \$ + 8 \$) = 4 \$.$$

Est-ce qu'il y a beaucoup de variation dans les gains d'un joueur à l'autre de cette population ? La mesure de la variation s'effectue en calculant la différence des gains de chaque individu par rapport au gain moyen.

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2 \\ &= \frac{1}{4} [(0 \$ - 4 \$)^2 + (2 \$ - 4 \$)^2 + (6 \$ - 4 \$)^2 + (8 \$ - 4 \$)^2] \\ &= \frac{1}{4} [16 \$^2 + 4 \$^2 + 4 \$^2 + 16 \$^2] \\ &= 10 \$^2. \end{aligned}$$

et donc $\sigma = \sqrt{10 \$^2} = 3,16 \$$.

La variance est donc une mesure de la variation. Plus les gains des individus sont loin de la moyenne, plus la variance augmente. Cependant, l'unité de mesure de la variance est le carré de celle de la moyenne, ce qui est difficile à interpréter. En effet, qui sait ce que veut dire des \$² ?

C'est pourquoi les praticiens utilisent plutôt l'écart-type σ qui est non seulement une mesure de la variatio, mais qui s'exprime dans les mêmes unités que la moyenne μ . Dans le domaine de la bourse, σ porte le nom de « volatilité des marchés ». Plus la volatilité est grande, plus il y a de risques, plus il y a de variation (à la hausse ou la baisse) dans les titres sur le parquet de la bourse.

Cependant, comme les recensements sont rares, le praticien se contente généralement d'un échantillon représentatif contenant n individus. Il calcule les statistiques qui estimeront les paramètres recherchés à partir de cet échantillon. Pour estimer la moyenne de la population, le praticien utilise la statistique de la moyenne échantillonnale \bar{x} :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

Pour estimer la variance σ^2 ainsi que l'écart-type σ de la population, l'analyste utilise la variance échantillonnale s^2 ainsi que l'écart type échantillonnal s :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

et

$$s = \sqrt{s^2}.$$

La variance échantillonnale s^2 est divisée par $n - 1$ au lieu de n . Il s'agit d'une correction au fait que dans cette formule, μ (le paramètre qui ne varie pas) est remplacé par une statistique \bar{x} qui n'est qu'une estimation de μ , et n'est donc pas parfaite. En d'autres termes, puisque \bar{x} n'est qu'une estimation, elle injecte une variation additionnelle dans la valeur de s^2 . Pour tenir compte de cette variation induite, le praticien prend la « liberté » de diviser la somme par $n - 1$ au lieu de n . Cette stratégie est toujours vraie ; pour chaque paramètre remplacé par sa statistique dans la formule d'une variance, une

unité est enlevée à son dénominateur à titre de correction. C'est là le principe des degrés de liberté.

Exemple 3.2.1 Un administrateur est chargé de planifier les ressources humaines d'un nouvel hôpital en construction. Pour commencer sa planification, il aimeraient connaître le nombre moyen d'employés à temps plein requis pour faire fonctionner son hôpital qui contiendra 30 lits. La figure 3.4 contient les nombres d'employés travaillant à temps plein (`nb_em_tp`) dans 12 hôpitaux tirés au hasard.

	hopital	nb_emp_tp
1	1	69
2	2	95
3	3	102
4	4	118
5	5	126
6	6	125
7	7	138
8	8	178
9	9	156
10	10	184
11	11	176
12	12	225
13		

FIG. 3.4 – Les données

Pour obtenir les sorties 3.5 et 3.6, il faut effectuer les commandes suivantes :

Menu SPSS :

→ Analyse

→ Descriptive Statistics

→ Explore...

Dans la fenêtre Dependent List : → `nb_em_tp`

Display : → Both

Dans le bouton Statistics... : √ Descriptives

Confidence Interval for Mean : 95 % (niveau de confiance)

Dans le bouton Plots : → Boxplots • None
 → Descriptive • Histogram

Cliquez sur Continue, puis sur Ok. On obtient alors les figures 3.5 et 3.6.

Descriptives			
		Statistic	Std. Error
nb_emp_tp	Mean	141,00	12,781
	95% Confidence Interval for Mean	112,87	
		169,13	
	5% Trimmed Mean	140,33	
	Median	132,00	
	Variance	1960,364	
	Std. Deviation	44,276	
	Minimum	69	
	Maximum	225	
	Range	156	
	Interquartile Range	72	
	Skewness	,276	,637
	Kurtosis	-,315	1,232

FIG. 3.5 – Les statistiques descriptives

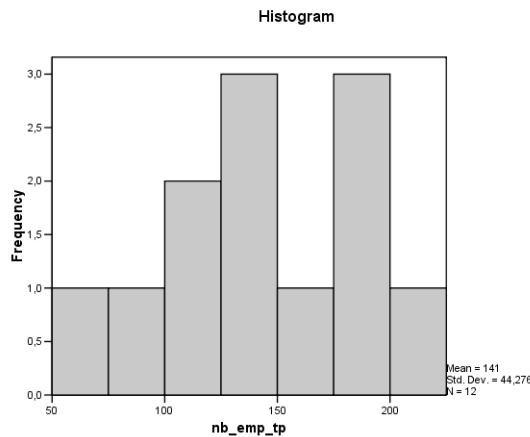


FIG. 3.6 – L'histogramme

Voici l'interprétation de la sortie 3.5. Ponctuellement, il y a en moyenne $\bar{x} = 141$ employés qui travaillent à temps plein dans les hôpitaux. Le plus petit hôpital de l'échan-

tillon contient 69 employés tandis que le plus imposant en contient 225. Il semble y avoir beaucoup de variation d'un hôpital à l'autre...

L'indice de la variabilité (l'écart-type) se quantifie à 44,28 employés.

La moyenne de la population semble-t-elle représentative de tous les hôpitaux ? En d'autres termes, est-il vraisemblable que tous les hôpitaux ont en moyenne 141 employés travaillant à temps plein ? Pour la planification du personnel d'un nouvel hôpital, serait-il prudent de planifier immédiatement l'embauche de 141 employés ?

Une moyenne est **représentative** (ou résume bien ses données) s'il y a peu de variation autour d'elle dans les données. Pour déterminer si une moyenne est représentative, il est utile de considérer la statistique appelée coefficient de variation (CV) :

$$CV = \frac{\sigma}{\mu} \approx \frac{s}{\bar{x}} = \frac{44,28 \text{ employés}}{141,00 \text{ employés}} = 0,314.$$

Si la variation est trop élevée, l'écart-type s sera très grand, démontrant la grande variabilité des données de l'échantillon par rapport à la moyenne. Plus la variation autour de la moyenne est élevée, plus il sera risqué d'utiliser la moyenne dans une planification ou dans un rapport. Si l'écart-type est très grand comparé à la moyenne, le CV le sera lui aussi. Pour savoir si une moyenne est utilisable, il faut consulter la table suivante :

CV en valeur absolue	Interprétation
$0 \leq CV < 0,15$	La moyenne est représentative.
$0,15 \leq CV < 0,30$	Il faut faire attention à l'utilisation de la moyenne.
$0,30 \leq CV$	La moyenne n'est pas représentative.

Cette interprétation est valide en autant que $\bar{x} \geq 5$.

FIG. 3.7 – Le coefficient de variation (CV)

Dans le cadre de notre exemple, à l'aide du CV, il est possible de voir que la moyenne n'est pas représentative des données. Cela signifie simplement qu'il est imprudent pour cet administrateur de planifier l'embauche de 141 employés pour son hôpital. En effet, la moyenne univariée n'est pas en mesure de tenir compte de la taille des hôpitaux.

L'administrateur peut alors tirer un autre échantillon d'hôpitaux de taille similaire à la sienne (environ 30 lits) et recommencer le processus. Il utilisera pour sa planification la statistique de la moyenne, si et seulement si celle-ci est représentative. Nous verrons plus tard qu'il est beaucoup plus intéressant et moins risqué d'effectuer de telles planifications à l'aide d'une régression linéaire. La figure 3.8 présente la signification des autres statistiques descriptives.

Statistique	Interprétation
La médiane (<i>median</i>)	Une fois les données rangées en ordre croissant, la médiane est la valeur exactement au centre de cette série classée. Ainsi 50 % des données lui sont inférieures et 50 % lui sont supérieures. Ex. : Dans la série « 3 6 7 7 9 », la médiane est 7. Dans la série « 3 6 7 9 », la médiane est 6,5. La médiane porte aussi le nom de second quartile : Q_2 .
Q_1 et Q_3	Q_1 et Q_3 sont les premier et troisième quartiles. Ainsi 25 % des données sont inférieures à Q_1 et 75 % lui sont supérieures. De même, 75 % des données sont inférieures à Q_3 et 25 % lui sont supérieures.
5 % <i>Trimmed Mean</i>	Une fois les données rangées en ordre non décroissant, on recalcule une moyenne en enlevant 5 % des données les plus petites et 5 % des données les plus élevées : ceci donne la 5 % <i>Trimmed Mean</i> . Cette technique permet d'enlever des données aberrantes (trop grandes ou trop petites) qui influencent la moyenne normale.
Étendue (<i>Range</i>)	C'est l'écart entre le minimum et le maximum.
<i>Interquartile Range</i>	C'est l'écart entre Q_1 et Q_3 : ceci nous donne donc l'étendue du 50 % des données centrales. Cette statistique est parfois utilisée pour mesurer la variation autour de la médiane, au même titre que l'écart-type est utilisé pour mesurer la variation des données autour de la moyenne.

FIG. 3.8 – Les autres statistiques dans une sortie SPSS de statistiques descriptives

Les statistiques **Mean**, **Median** et **5 % Trimmed Mean** (moyenne tronquée) ont pour objectif d'estimer le même paramètre de la moyenne de la population μ .

En effet, ces trois statistiques mesurent à leur manière la tendance centrale μ . Ainsi,

l'analyste est plus confiant face à l'estimation de μ lorsque les trois statistiques sont semblables. Lorsque les trois estimateurs concordent, le praticien peut être confiant face au fait qu'il a bien estimé la tendance centrale dans la population.

Lorsque les trois estimateurs de la moyennes ne concordent pas, et qu'il faut absolument fournir une estimation, les analystes préfèrent utiliser la médiane à titre d'estimation ponctuelle du paramètre μ . Pourquoi ? Simplement parce que la médiane est dite plus robuste aux valeurs aberrantes. L'analyste sera encore plus confiant de ce choix si la moyenne tronquée se rapproche de celle-ci. Cependant, utiliser cette valeur peut être risqué. Il existe aussi des situations où les trois statistiques sont totalement différentes, c'est alors l'impasse. Sans autres informations, l'analyste peut tenter d'utiliser un histogramme pour essayer de comprendre laquelle des trois statistiques résume le mieux la tendance centrale μ . Comme nous serons en mesure de le constater, ce type d'impasse se contourne facilement avec les analyses bivariées et multivariées ; cependant ce type d'analyse exige la présence de plusieurs variables.

En somme, une statistique, telle la moyenne \bar{x} , peut bien estimer la tendance centrale (la moyenne, la moyenne tronquée et la médiane sont très rapprochées), mais ne pas être représentative (ne pas bien résumer) des données. En effet, il est possible que le paramètre estimé μ ne soit lui-même pas représentatif des données de la population (trop de variation dans la population !). Par exemple, la variable de l'âge de la population est très dispersée autour de l'âge moyen $\mu_{\text{âge}}$; la situation inverse serait catastrophique (!). En somme, une statistique (ou un paramètre) ne peut servir seule à prendre de bonnes décisions que si la dispersion des données est petite autour d'elle.

Les statistiques traitant la forme sont utiles pour vérifier si les données de la population se distribuent suivant la cloche de la loi normale ou non. Ce type de vérification est nécessaire dans plusieurs contextes où il est présupposé que les données de la population se distribuent suivant une loi normale.

La cloche de la loi normale est centrée et parfaitement symétrique par rapport à la moyenne μ . Le coefficient d'asymétrie (Skewness) qui est fourni dans la sortie SPSS

(voir la figure 3.5) vérifie si la distribution des données de l'échantillon n'est pas trop asymétrique par rapport à la moyenne. Cette statistique est centrée à 0 ; ainsi, une grande valeur positive pour la statistique Skewness indique une asymétrie prononcée vers la droite. Inversement, une grande valeur négative indique une asymétrie prononcée vers la gauche.

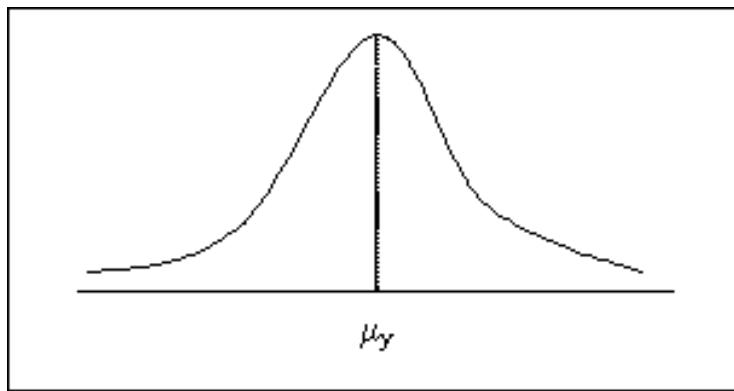


FIG. 3.9 – La forme en cloche de la loi normale

Aussi, la cloche normale n'est ni trop aplatie ni trop pointue. Le coefficient d'aplatissement (Kurtosis) vérifie si l'aplatissement des données ressemble à celui d'une distribution normale. Cette statistique est aussi centrée à 0. Une grande valeur négative pour la statistique Kurtosis indique un aplatissement hors du commun, et une forte valeur positive indique la présence d'un sommet trop pointu.

Plus les statistiques Skewness et Kurtosis sont près de 0, plus la distribution des données ressemble à la distribution d'une loi normale. À titre de règle du pouce, pour être en mesure de juger si ces statistiques sont loin ou près de 0, il est possible d'utiliser le quotient de ces statistiques sur leur écart type (*Std. Error*, dans la colonne de droite). À titre d'ordre de grandeur, on tend à rejeter la normalité des données de la population lorsque

$$\left| \frac{\text{Skewness} - E(\text{Skewness})}{\sqrt{\text{Var}(\text{Skewness})}} \right| = \left| \frac{\text{Skewness} - 0}{\text{Std. Error}_{\text{Skewness}}} \right| > 2$$

ou

$$\left| \frac{\text{Kurtosis} - E(\text{Kurtosis})}{\sqrt{\text{Var}(\text{Kurtosis})}} \right| = \left| \frac{\text{Kurtosis} - 0}{\text{Std. Error}_{\text{Kurtosis}}} \right| > 2.$$

Illustrons ceci à l'aide des sorties de l'exemple 3.2.1.

Descriptives			
		Statistic	Std. Error
nb_emp_tp	Mean	141,00	12,781
	95% Confidence Interval for Mean	Lower Bound 112,87 Upper Bound 169,13	
	5% Trimmed Mean	140,33	
	Median	132,00	
	Variance	1960,364	
	Std. Deviation	44,276	
	Minimum	69	
	Maximum	225	
	Range	156	
	Interquartile Range	72	
	Skewness	,276	,637
	Kurtosis	-,315	1,232

FIG. 3.10 – Les statistiques descriptives de l'exemple 3.2.1

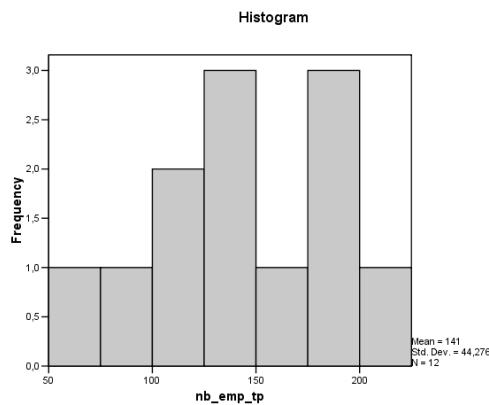


FIG. 3.11 – L'histogramme de l'exemple 3.2.1

La statistique d'asymétrie Skewness a une valeur de 0.276. Pour juger si la courbe illustrée par l'histogramme est asymétrique par rapport à une cloche normale, il suffit de

calculer le ratio $0.276/0.637 = 0.43$. Comme $|0.43| < 2$, nous ne rejetons pas la normalité de la courbe. Donc, il ne semble pas y avoir une asymétrie significative pour conclure à la non normalité.

La statistique d'aplatissement Kurtosis a une valeur de -0.315 . Pour juger si l'aplatissement de la courbe illustrée par l'histogramme est typique d'une loi normale, il suffit de calculer le ratio $-0.315/1.232 = -0.26$. Comme $|-0.26| < 2$, nous ne rejetons pas la normalité de la courbe. Donc, il ne semble pas y avoir un aplatissement significatif pour conclure à la non normalité.

Ainsi, pas de violation à l'asymétrie, pas de violation au niveau de l'aplatissement, tout va bien. Mais cela n'assure pas la normalité pour autant ; elle n'a simplement pas été rejetée. En effet, il faut comprendre que ces statistiques mesurent respectivement l'asymétrie et l'aplatissement de façon indépendante et non conjointe. Elles ne sont que des règles du pouce. Comme nous le verrons plus tard, il existe des tests d'hypothèses qui vérifient la normalité sous toutes ses facettes.

3.3 L'inférence liée à une variable discrète

À la section 3.1, on a estimé de façon ponctuelle une proportion (la proportion des électeurs en faveur du candidat 2). L'estimation ponctuelle, bien qu'utile, ne fournit aucune information sur la précision et la crédibilité des résultats. En effet, un pourcentage peut provenir d'une étude ne contenant que 10 individus. Pour étendre les résultats observés à la population (inférer), l'industrie utilise les intervalles de confiance. Un intervalle de confiance coince la valeur d'un paramètre entre deux valeurs a et b , et ce, avec une forte probabilité d'avoir raison. Plus précisément :

$$P(a \leq \Pi \leq b) = 1 - \alpha.$$

Cette probabilité s'interprète de la manière suivante : la probabilité que le paramètre soit coincé entre les valeurs a et b est de $(1 - \alpha) \times 100\%$. La quantité $1 - \alpha$ est appelée

le niveau de confiance de l'estimation. Cette quantité représente la probabilité d'avoir raison ; l'analyste la veut la plus élevée possible. Mathématiquement, $1 - \alpha$ représente 100 % des chances d'avoir raison moins une possibilité d'erreur α . Cette erreur est liée au fait qu'il est possible que l'échantillon en main soit simplement mauvais. En effet, il faut comprendre qu'il est impossible de ne jamais faire d'erreur en se basant sur les résultats d'un échantillon. L'analyste désire cependant que cette erreur α soit la plus petite possible.

Dans l'entreprise, il est courant pour les analystes de l'industrie d'utiliser l'un des trois niveaux de confiance suivants : $1 - \alpha = 0,90$, $1 - \alpha = 0,95$ ou $1 - \alpha = 0,99$. Le niveau de confiance 95 % est le niveau le plus couramment utilisé. Il est d'ailleurs associé à des phrases standards du type : « Actuellement, la vraie proportion de la population en faveur du candidat 2 devrait être comprise entre a et b , et ce, 19 fois sur 20. », c'est-à-dire avec 95 % des chances d'avoir raison.

À l'aide des données dans l'échantillon, l'analyste est en mesure de calculer les valeurs a et b qui « fourchettent » le paramètre recherché. La structure d'un intervalle de confiance qui estime le paramètre Π entre deux valeurs a et b , et ce, pour un niveau de confiance $1 - \alpha$ fixé à l'avance, est importante à comprendre. Ce squelette se répète pour tous les intervalles de confiance estimant un paramètre (discret ou non). Plus précisément, la structure d'un intervalle de confiance est la suivante : l'estimateur du paramètre plus ou moins un nombre d'écart-types de la statistique, le nombre d'écart-types étant donné par le biais de la loi de probabilité de l'estimateur. Voici donc l'intervalle de confiance pour une proportion :

$$\text{IC} = \left[\underbrace{p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}}_{a}, \underbrace{p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}}_{b} \right]$$

Dans cette structure :

- p représente la proportion échantillonnale ;
- $\frac{p(1-p)}{n}$ représente la variance échantillonnale de la statistique p ;

- $\sqrt{\frac{p(1-p)}{n}}$ représente l'écart-type échantillonnal de la statistique p ;
- $\pm z_{\alpha/2}$ représente le nombre d'écart-types. Ce coefficient est issu de la loi normale qui modélise le comportement de la statistique p lorsqu'elle estime Π . Cette valeur est entièrement déterminée par le niveau de confiance $1 - \alpha$ fixé par l'analyste.
- $\pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$ représente la précision des résultats. Elle détermine la largeur de l'intervalle de confiance. Plus cette quantité est grande, moins l'analyste est précis dans ses conclusions.
- Cet intervalle de confiance est **valide** en autant que $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$.

Les trois premiers points sont entièrement déterminés par l'échantillon, tandis que le coefficient $z_{\alpha/2}$ est indépendant des valeurs de l'échantillon. Cette valeur est déterminée par le niveau de confiance $1 - \alpha$ fixé par l'expérimentateur. Lorsque l'analyste désire changer son niveau de confiance, il lui suffit d'ajuster la valeur du coefficient $z_{\alpha/2}$ à partir de la table de la loi normale.

Pour bien comprendre le rôle de la loi normale et de ses coefficients $z_{\alpha/2}$ dans cet intervalle, il faut s'attarder au comportement de statistique p lorsqu'elle est utilisée pour estimer le paramètre Π .

Il faut se rappeler que p est une statistique et que ses estimations varient d'un échantillon à l'autre. La variation de la statistique p qui estime son paramètre Π se compare à la variation des balles d'un tireur d'élite tirant sur sa cible. La variation de la statistique p se distribue autour du paramètre Π suivant une loi de probabilité en forme de cloche, cette forme illustrant l'idée suivante : la probabilité d'un « bon tir » près de la cible est plus grande que la probabilité d'un « mauvais tir » loin de la cible. La variation d'une statistique est toujours mesurée en terme de nombre d'écart-types.

Comme l'illustre la figure 3.12, par la définition d'une loi de probabilité, l'aire sous la courbe représente toujours une probabilité. L'annexe A présente les valeurs de la loi normale. En se basant sur la loi normale, il est courant d'utiliser les repères suivants :

- 68,26 % des échantillons en main produiront des estimations p à plus ou moins un écart-type de la cible Π , ce qui est excellent ;

- 90 % des échantillons en main produiront des estimations p à plus ou moins 1,645 écart-type de la cible Π ;
- 95 % des échantillons en main produiront des estimations p à plus ou moins 1,96 écart-type de la cible Π ;
- Un peu plus de 99 % des échantillons en main produiront des estimations p à plus ou moins 3 écart-types de la cible Π .

La représentation de gauche de la figure illustre que 95 % des échantillons en main produiront des estimations p qui seront à $\pm z_{\alpha/2} = 1,96$ écart-type de la cible 0, 0 illustrant la vraie valeur Π (Π étant à 0 écart-type de lui-même). La probabilité α d'obtenir une mauvaise estimation de Π est dans cet exemple fixé à 5 %. Cette erreur est divisée également à gauche et à droite ; elle représente la probabilité respective de sous-estimer ou de sur-estimer Π . La représentation de droite de la même figure illustre que 90 % des échantillons produiront des estimations p qui seront à $\pm z_{\alpha/2} = 1,645$ écart-type de la même cible. La probabilité d'erreur de 10 % est répartie également entre la gauche et la droite.

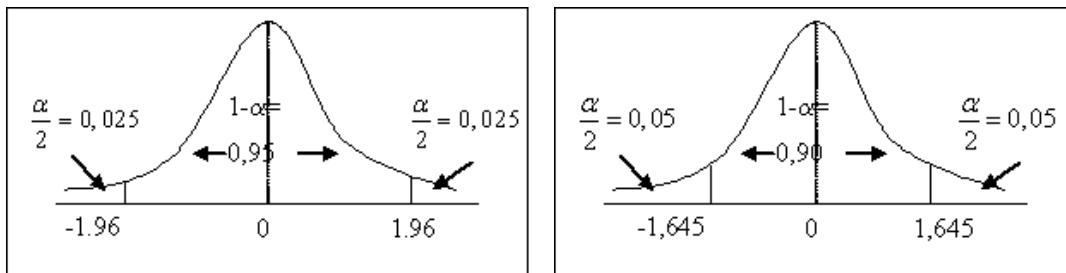


FIG. 3.12 – La loi normale, les niveaux de confiance et les valeurs $z_{\alpha/2}$.

L'analyste espère que l'intervalle contiendra vraiment le paramètre recherché. Si l'échantillon est parmi les « 95 % de bons échantillons », il espère, en additionnant et en retranchant la précision (un nombre d'écart-types) à p , que le paramètre se retrouvera pris dans le filet. À titre d'exemple, voici les détails de la construction d'un intervalle de confiance de niveau de confiance fixé à 95 % pour le paramètre $\Pi_{\text{candidat 2}}$.

$$\begin{aligned}
 \text{IC} &= \left[\underbrace{p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}}_a, \underbrace{p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}}_b \right] \\
 &= \left[0,438 - 1,96 \sqrt{\frac{0,438(1-0,438)}{400}}, 0,438 + 1,96 \sqrt{\frac{0,438(1-0,438)}{400}} \right] \\
 &= [0,389, 0,487].
 \end{aligned}$$

Cet intervalle s'interprète de la manière suivante : « Actuellement, la vraie proportion de la population en faveur du candidat 2 est comprise entre 38,9 % et 48,7 %, et ce, 19 fois sur 20 (avec 95 % des chances d'avoir raison). ». Pour obtenir un intervalle de confiance de niveau de confiance 90 % pour le même paramètre, il suffit d'ajuster les coefficients $z_{\alpha/2}$ de la loi normale dans la structure et d'effectuer les calculs suivants :

$$\begin{aligned}
 \text{IC} &= \left[\underbrace{p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}}_a, \underbrace{p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}}_b \right] \\
 &= \left[0,438 - 1,645 \sqrt{\frac{0,438(1-0,438)}{400}}, 0,438 + 1,645 \sqrt{\frac{0,438(1-0,438)}{400}} \right] \\
 &= [0,397, 0,479].
 \end{aligned}$$

Ainsi, avec le même échantillon, l'analyste est en mesure de dire que la vraie proportion $\Pi_{\text{candidat 2}}$ de la population qui est en faveur du second candidat est comprise entre 39,7 % et 47,9 %, et ce, 18 fois sur 20 (90 % des chances d'avoir raison).

Il faut remarquer que l'intervalle de confiance de niveau de confiance 95 % est plus large (donc moins précis) que l'intervalle de confiance de niveau 90 %. En contrepartie, celui de 90 % est moins sûr. Comme l'illustre la figure 3.12, cette différence entre la confiance et la précision provient de la différence entre les coefficients $z_{\alpha/2}$. Il existe toujours une dualité entre la précision et la confiance.

La morale de l'histoire est bien simple : à vouloir être trop confiant, l'analyste est moins précis (l'intervalle est plus large). C'est pourquoi on voit rarement des intervalles de confiance de niveau de confiance 99 %. Il est important de rappeler que les calculs précédents ne sont valides que pour un échantillon dont la taille est supérieure à 30 unités.

Bien entendu, il est possible de faire effectuer les calculs précédents par le logiciel ; SPSS est en mesure de calculer les intervalles de confiance si et seulement si la variable est codifiée en mode binaire (0 et 1). De plus, le logiciel ne retourne par défaut que la proportion associée à la modalité de réponse codifiée avec des 1 dans la base de données. Si la variable est discrète avec seulement deux modalités de réponse (ex : oui, non ou encore masculin, féminin) et que l'analyste a déjà codifié les modalités en binaire (ex : 0=oui, 1=non ou encore 0=mASCULIN, 1=fÉMININ), SPSS peut calculer directement l'intervalle de confiance lié à la modalité 1 (p_{non} ou encore $p_{\text{fÉMININ}}$). Si, comme dans notre exemple, la variable discrète contient plus de deux modalités de réponses, il faudra simplement procéder à un recodage binaire des données. Pour obtenir de SPSS un intervalle de confiance pour la proportion des gens de la population qui voteront pour le candidat 2, il faut fabriquer de toutes pièces une nouvelle variable binaire suivant cette petite transformation :

Candidat 1 = 0 (pas voté pour le candidat 2)

Candidat 2 = 1 (voté pour le candidat 2)

Candidat 3 = 0 (pas voté pour le candidat 2)

Candidat 4 = 0 (pas voté pour le candidat 2)

Pour créer cette nouvelle variable binaire permettant de dégager l'intervalle de confiance pour $\Pi_{\text{candidat 2}}$ et pour faire exécuter ces changements de façon automatique par le logiciel, il faut effectuer les commandes suivantes :

Menu SPSS : → Transform

→ Recode

→ Into Different Variables...

Dans la fenêtre Input Variable : → candidat (variable à recoder)

Dans la fenêtre Output Variable : → Name : candid2

(nom de la nouvelle variable)

Appuyez sur le bouton **Change** pour officialiser la codification. Ensuite, pour la recodification, appuyez sur le bouton **Old and New Values...** :

Old Value : → **Value** : 2 (la valeur associée au candidat 2)

New Value : → **Value** : 1 (puisque c'est cette proportion que l'on veut)

Cliquez sur **Add**.

Old Value : → **All other values** (tous les autres candidats)

New Value : → **Value** : 0 (ils n'ont pas voté pour le candidat 2)

Cliquez sur **Add**.

Cliquez sur **Continue**.

Cliquez sur **Ok**.

On obtient ainsi la nouvelle variable **candid2**, tel qu'on peut l'entrevoir dans la figure 3.13.

	ident	candidat	candid2	
1	1	candidat 1	,00	
2	2	candidat 2	1,00	
3	3	candidat 2	1,00	
4	4	candidat 3	,00	
5	5	candidat 2	1,00	
6	6	candidat 4	,00	
7	7	candidat 2	1,00	
8	8	candidat 3	,00	
9	9	candidat 1	,00	
10	10	candidat 2	1,00	
11	11	candidat 2	1,00	
12	12	candidat 2	1,00	
13	13	candidat 1	,00	
14	14	candidat 2	1,00	
15	15	candidat 4	,00	
16	16	candidat 1	,00	
17	17	candidat 2	1,00	

FIG. 3.13 – Crédit de la nouvelle variable **candid2**

Cette nouvelle variable ne contient que des 0 et des 1. Il suit de cette recodification que les « 1 » ne sont associés qu'aux gens ayant voté pour le candidat 2. La somme de la

colonne `candid2`, qui contient toujours 400 lignes, est égale au nombre de personnes ayant voté pour le candidat 2. Alors la moyenne de cette colonne sera égale à la proportion des gens qui ont voté pour ce candidat.

Pour obtenir l'intervalle de confiance de niveau 95 % pour $\Pi_{\text{candidat } 2}$, il suffit d'effectuer les commandes SPSS suivantes :

Menu SPSS :	→ Analyse
	→ Descriptive Statistics
	→ Explore...
Dans la fenêtre Dependent List :	→ <code>candid2</code>
Display :	→ Statistics (pas de graphe)
Dans le bouton Statistics... :	✓ Descriptives
Confidence Interval for Mean :	95 % (niveau de confiance)

Cliquez sur **Ok**.

Descriptives		
CANDID_2	Mean	Statistic
	95% Confidence Interval for Mean	Lower Bound
		Upper Bound
		,4863
	5% Trimmed Mean	,4306
	Median	,0000
	Variance	,247
	Std. Deviation	,4967
	Minimum	,00
	Maximum	1,00
	Range	1,00
	Interquartile Range	1,0000
	Skewness	,253
	Kurtosis	-1,946
		Std. Error
		,2483E-02

FIG. 3.14 – Intervalle de confiance pour une proportion

Interprétation de la figure 3.14 : ponctuellement, la proportion (la moyenne) des gens de l'échantillon qui ont voté pour le candidat 2 est de 43,75 %.

Dans la population, la proportion des gens en faveur du candidat 2 est comprise entre 38,87 % et 48,63 %, et ce, 19 fois sur 20 (ou avec 95 % des chances d'avoir raison). Les

autres statistiques ne sont pas utiles à l'analyse d'une variable discrète binaire. Elles ne seront utiles que dans le contexte des variables continues.

3.4 L'inférence liée à une variable continue

À l'aide des statistiques de l'échantillon, le praticien estime les paramètres de la population. Il existe deux façons de faire l'estimation de paramètres : l'estimation ponctuelle et l'estimation par intervalles de confiance.

L'estimation ponctuelle d'un paramètre d'une variable continue consiste à utiliser la valeur d'une statistique correspondante à titre d'estimation. Pour un même paramètre, il peut exister plusieurs estimateurs. Par exemple, pour estimer la moyenne μ de la population, il y a la moyenne, la moyenne tronquée, la médiane et bien d'autres. Pour estimer la variation autour de la moyenne σ dans la population, on utilise la variation échantillonnale s .

Cependant, ces estimations ponctuelles ne transcrivent aucune crédibilité. En effet, pour obtenir une estimation ponctuelle, un analyste peut avoir sondé qu'un seul hôpital ou pire avoir imaginé cet estimation ! C'est pourquoi l'estimation d'un paramètre est généralement effectuée par l'entremise d'un intervalle de confiance. L'intervalle tentera de coincer le paramètre entre deux bornes a et b , et ce, avec la plus grande confiance possible.

Tous les paramètres de la population peuvent être estimés par intervalles de confiance. Par exemple, la section 3.3 a présenté l'estimation du paramètre Π par intervalle de confiance pour le cas de variables discrètes binaires. Bien qu'il soit possible d'obtenir des intervalles de confiance pour tous les paramètres liés à une variable continue, nous nous limiterons dans ce cours à l'estimation du paramètre de la moyenne μ de la population.

Dans la science de la statistique, les intervalles de confiance ont tous un seul et même objectif : coincer un paramètre entre deux valeurs, et ce, avec une probabilité très élevée

(95 % par exemple) de réussir. En d'autres mots, nous voulons que :

$$P(a \leq \mu \leq b) = 1 - \alpha.$$

Tout comme illustré dans la section 3.3, la quantité $1 - \alpha$ représente la probabilité que la vraie moyenne μ de la population soit comprise entre les bornes a et b . La quantité $1 - \alpha$ s'appelle le niveau de confiance. Les niveaux de confiance généralement utilisés sont 90 %, 95 % et 99 %. L'industrie utilise généralement le niveau de confiance 95 %.

Par rapport à l'exemple 3.2.1, pour un niveau de confiance de 95 % ($1 - \alpha = 0,05$), quelles sont les valeurs a et b entre lesquelles devrait se situer le paramètre du nombre moyen d'employés travaillant à temps plein dans les hôpitaux ?

Pour obtenir l'intervalle de confiance de niveau 95 % pour $\mu_{nb_em_tp}$, il suffit d'effectuer les commandes SPSS suivantes :

Menu SPSS :	→ Analyse
	→ Descriptive Statistics
	→ Explore...
Dans la fenêtre Dependent List :	→ nb_em_tp
Display :	→ Statistics (pas de graphe)
Dans le bouton Statistics... :	✓ Descriptives
Confidence Interval for Mean :	95 % (niveau de confiance)

D'après la figure 3.15, on voit que le paramètre du nombre moyen d'employés travaillant à temps plein dans les hôpitaux est compris entre 112.87 et 169.13 employés, et ce, 19 fois sur 20 (avec 95 % des chances d'avoir raison). Voici le détail des calculs. La structure de cet intervalle de confiance ressemble à celle présentée pour coincer le paramètre de la proportion. Un intervalle de confiance est simplement la statistique plus ou moins un nombre d'écart-types.

Descriptives			
nb_emp_tp	Mean	141,00	12,781
	95% Confidence Interval for Mean	Lower Bound 112,87	Upper Bound 169,13
	5% Trimmed Mean	140,33	
	Median	132,00	
	Variance	1960,364	
	Std. Deviation	44,276	
	Minimum	69	
	Maximum	225	
	Range	156	
	Interquartile Range	72	
	Skewness	,276	,637
	Kurtosis	-,315	1,232

FIG. 3.15 – Intervalle de confiance de niveau 95 % pour l'exemple 3.2.1

$$\begin{aligned}
 \text{IC} &= \left(\bar{x} - t_{(n-1); \alpha/2} \sqrt{\frac{s^2}{n}}, \bar{x} + t_{(n-1); \alpha/2} \sqrt{\frac{s^2}{n}} \right) \\
 &= \left(141,0 - t_{(n-1); \alpha/2} \frac{44,28}{\sqrt{12}}, 141,0 + t_{(n-1); \alpha/2} \frac{44,28}{\sqrt{12}} \right) \\
 &= (141,0 - 2,201 \times 12,78, 141,0 + 2,201 \times 12,78) \\
 &= (112,87, 169,13).
 \end{aligned}$$

Contrairement à la statistique de la proportion échantillonale p dont le comportement de « franc-tireur » était modélisé à l'aide de loi normale, le comportement de la statistique \bar{x} est plutôt modélisé par la loi de Student. La loi de Student ressemble à la loi normale, à la différence que plus l'échantillon est petit, plus elle est aplatie (elle adopte une position plus conservatrice). D'ailleurs, lorsque la taille de l'échantillon dépasse 120 unités, la loi de Student est la copie conforme de la loi normale. La loi de Student est donc une loi en cloche. L'utilisation de la loi de Student au lieu de la loi normale provient d'un souci de correction au fait que l'intervalle de confiance pour μ utilise la statistique s^2 (doublement « erronée ») dans son calcul. Les quantiles $t_{(n-1); \alpha/2}$ jouent le même rôle et s'interprètent de la même manière que les coefficients $z_{\alpha/2}$. Ils ne servent qu'à intro-

duire le niveau de confiance dans les estimations. Tout comme pour la loi normale, ce sont les coefficients $t_{(n-1); \alpha/2}$ de la loi de Student qui indiquent le niveau de confiance d'un intervalle de confiance. Ces coefficients dépendent aussi de la taille de l'échantillon, contrairement aux coefficients $z_{\alpha/2}$ de la loi normale.

Ces coefficients se lisent facilement dans la table de Student avec $n-1$ degrés de liberté (n étant la taille de l'échantillon). La quantité est directement en lien avec la pénalité au dénominateur de la statistique de la variation échantillonnale $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

À partir de l'exemple 3.2.1, dans lequel on a 12 hôpitaux, il est possible d'utiliser les repères suivants en se basant sur la loi de Student avec 11 degré de liberté :

- 90 % des échantillon en main produiront des estimations \bar{x} à plus ou moins 1,796 écart-type de la cible μ .
- 95 % des échantillon en main produiront des estimations \bar{x} à plus ou moins 2,201 écart-types de la cible μ .

Contrairement à la loi normale, ces valeurs changent pour chaque taille d'échantillon. À chaque échantillon, il faut retrouver les bonnes valeurs $t_{(n-1); \alpha/2}$. Pour lire ces valeurs dans la table de la loi de Student, il suffit de suivre les instructions à cet effet dans l'annexe A.

Pour obtenir un intervalle de confiance de niveau de confiance 90 % avec SPSS, il faut effectuer les commandes SPSS suivantes :

Menu SPSS :	→ Analyse
	→ Descriptive Statistics
	→ Explore...
Dans la fenêtre Dependent List :	→ nb_em_tp
Display :	→ Statistics (pas de graphe)
Dans le bouton Statistics... :	✓ Descriptives
Confidence Interval for Mean :	90 % (niveau de confiance)

Ainsi, le véritable nombre moyen d'employés travaillant à temps plein dans les hôpitaux est compris entre 118,05 et 163,95 employés avec 90 % des chances d'avoir raison.

Descriptives			
NB_EM_TP	Mean	Statistic	Std. Error
	90% Confidence Interval for Mean	Lower Bound Upper Bound	
		118,05 163,95	
	5% Trimmed Mean	140,33	
	Median	132,00	
	Variance	1960,364	
	Std. Deviation	44,28	
	Minimum	69	
	Maximum	225	
	Range	156	
	Interquartile Range	71,50	
	Skewness	,276	,637
	Kurtosis	-,315	1,232

FIG. 3.16 – Intervalle de confiance de niveau 90 % pour l'exemple 3.2.1

Il faut remarquer que plus l'analyste désire être confiant, moins il est précis. C'est pourquoi l'intervalle de confiance de niveau de confiance 95 % est plus large que celui de niveau 90 %.

3.5 Détermination de la taille d'échantillon

Dans une étude, pour estimer un paramètre (une proportion ou une moyenne, par exemple), il est souvent possible de déterminer à l'avance la taille de l'échantillon qui permettra d'obtenir une précision et un niveau de confiance désirés pour cette estimation. Cette technique est à la base de toute soumission faite par les maisons de sondage. En effet, connaissant la taille de l'échantillon, il est possible d'estimer le temps de collecte de données, les ressources humaines nécessaires, les coûts, etc.

Ici nous présentons la détermination de la taille d'échantillon à priori dans les cas où on veut estimer une proportion et une moyenne.

3.5.1 Variable discrète : proportion

À l'aide de la formule algébrique de la structure d'un intervalle de confiance pour une proportion, il est possible de voir qu'il existe seulement deux façons de jouer sur la précision (la largeur de l'intervalle de confiance obtenu) des résultats.

$$\text{IC} = \left[\underbrace{p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}}_{a}, \underbrace{p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}}_{b} \right]$$

La quantité $z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$ porte le nom de **précision** et sera notée E . C'est cette quantité qui détermine la largeur de l'intervalle. Les deux façons de jouer avec cette quantité sont les suivantes :

- L'analyste peut modifier le niveau de confiance $1 - \alpha$ lié à l'intervalle. Ceci joue directement sur les coefficients $z_{\alpha/2}$; plus le niveau de confiance est élevé, plus la valeur $z_{\alpha/2}$ est grande, et plus l'intervalle est large.
- L'analyste augmente ou diminue la taille n de l'échantillon. Plus cette taille est élevée, plus la précision est petite, ce qui rend l'intervalle moins large et donc plus précis.

Ainsi, si l'on connaît la précision et le niveau de confiance voulus, on solutionne l'équation ci-dessous pour déterminer la taille d'échantillon nécessaire :

$$E = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}.$$

La précision est généralement fixée dans les demandes du client. Par exemple, pour un plan d'affaires, il peut désirer une précision de $\pm 5\%$ sur l'estimation de sa part de marché.

Le niveau de confiance de l'intervalle de confiance est aussi fixé avec le client. Le niveau de confiance standard est de 95 %, mais peu être fixé autrement. Par exemple, si $1 - \alpha = 0,95$, alors $z_{\alpha/2} = 1,96$.

Par contre, la quantité $p(1-p)$ est inconnue avant l'étude, sinon pourquoi faire cette étude ! Mais p est une proportion, et est donc toujours coincée entre 0 et 1. Il est alors

possible de remplacer $p(1 - p)$ par la « pire » des valeurs possibles, c'est-à-dire la plus grande valeur possible, celle qui augmente le plus la quantité E . Cette valeur est $p(1-p) = 0,25$, c'est-à-dire $p = 0,5$. Cette position est très conservatrice. Ce remplacement assure au praticien que, même dans le pire des cas, il réussira à obtenir la précision et la confiance demandées. L'impact de ce choix a souvent pour conséquence de produire des tailles d'échantillon trop grandes.

Une autre alternative serait d'estimer p avec une proportion obtenue d'une autre étude, similaire à l'étude que l'on désire faire, ou d'une étude préalable faite sur un petit échantillon d'une trentaine d'observations. Ceci permettrait sans doute de réduire la taille de l'échantillon. Par contre, il y a le risque qu'il y ait une bonne différence entre cette estimation et la proportion p obtenue dans la présente étude, ce qui dans certains cas pourrait empêcher le praticien d'obtenir la précision voulue avec le niveau de confiance établi.

Exemple 3.5.1 Supposons que le client désire une précision de $\pm 3\%$ sur les estimations de toutes les proportions à l'étude, et ce, 19 fois sur 20. Il suffit alors pour le praticien d'isoler n dans la formule suivante :

$$\begin{aligned} E = 0,03 &= z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \\ &= 1,96 \sqrt{\frac{0,25}{n}} \\ \Rightarrow 0,03^2 &= 1,96^2 \left(\frac{0,25}{n} \right) \\ \Rightarrow n &= 1,96^2 \times \frac{0,25}{0,03^2} = 1067,11 \approx 1067. \end{aligned}$$

Ainsi, une taille d'échantillon de 1067 personnes garantit au client une précision sur sa part de marché de $\pm 3\%$, et ce avec un niveau de confiance de 95 %.

- Pour un niveau de confiance $1 - \alpha = 0,95$, et pour estimer une proportion,
- une taille d'échantillon de 227 unités garantit une précision d'au moins $\pm 6,5\%$ sur les IC ;

- une taille d'échantillon de 384 unités garantit une précision d'au moins $\pm 5\%$ sur les IC ;
- une taille d'échantillon de 600 unités garantit une précision d'au moins $\pm 4\%$ sur les IC ;
- une taille d'échantillon de 1067 unités garantit une précision d'au moins $\pm 3\%$ sur les IC.

3.5.2 Variable continue : moyenne

Contrairement au cas discret, il ne sera pas toujours possible de calculer une taille d'échantillon à priori garantissant un niveau de précision et un niveau de confiance fixés à l'avance. Cependant deux situations le permettent. Pour bien comprendre, observons la structure de la formule de l'intervalle de confiance pour μ :

$$\left(\underbrace{\bar{x} - t_{(n-1);\alpha/2} \sqrt{\frac{s^2}{n}}}_{a}, \underbrace{\bar{x} + t_{(n-1);\alpha/2} \sqrt{\frac{s^2}{n}}}_{b} \right)$$

Dans cet intervalle de confiance, la précision est la quantité $t_{(n-1);\alpha/2} \sqrt{\frac{s^2}{n}}$ est sera elle aussi notée E .

Si l'on connaît la précision et la confiance voulues, on rencontrera une première difficulté en tentant d'isoler n dans l'équation $E = t_{(n-1);\alpha/2} \sqrt{\frac{s^2}{n}}$: le coefficient $t_{(n-1);\alpha/2}$ dépend de la taille de l'échantillon n , ce que l'on cherche justement à déterminer et qui est donc inconnue. Pour contourner ce problème, on peut remplacer le coefficient $t_{(n-1);\alpha/2}$ de la loi de Student par le coefficient $z_{\alpha/2}$ de la loi normale, qui lui ne dépend pas de n , et nous donne une bonne approximation. Ainsi la formule nous permettant d'estimer la taille d'échantillon devient

$$E = z_{\alpha/2} \sqrt{\frac{s^2}{n}}.$$

La précision est fixée par le client. Mais attention : ici on doit exprimer cette précision dans les mêmes unités que celles de la moyenne recherchée. Par exemple, si le client veut

étudier la consommation d'essence en litres d'une nouvelle voiture à mettre sur le marché, il doit formuler la précision en terme de litres. Par exemple, $\pm 0,01$ litre.

Le niveau de confiance de l'intervalle de confiance est aussi fixé avec le client, ce qui fixe $z_{\alpha/2}$. Par exemple, si $1 - \alpha = 0,95$, alors automatiquement $z_{\alpha/2} = 1,96$.

La variance s^2 est ici totalement inconnue et, à moins d'être un devin, il est impossible de borner cette valeur afin de la remplacer par la plus grande valeur qu'elle peut prendre. Si aucune estimation de la grandeur de s^2 n'est disponible, ce qui est souvent le cas, c'est l'impasse. Cependant, deux situations peuvent se présenter où il est possible de contourner le problème d'estimation de s^2 en remplaçant cette valeur par une estimation « intelligente » :

- Si une étude équivalente existe, il est possible de substituer dans cette formule le s^2 inconnu par son équivalent $s_{étude équivalente}^2$. Il suffit ensuite d'isoler le n dans la formule

$$E = z_{\alpha/2} \sqrt{\frac{s_{étude équivalente}^2}{n}}.$$

- Il peut aussi être possible d'effectuer une étude préliminaire avec une quinzaine d'unités. On substitue alors s^2 par $s_{étude prél.}^2$ dans la formule, et on isole n :

$$E = z_{\alpha/2} \sqrt{\frac{s_{étude prél.}^2}{n}}.$$

Exemple 3.5.2 Supposons qu'il y a une étude où un magazine spécialisé tente d'évaluer la consommation moyenne du moteur d'une nouvelle voiture. Le magazine désire une précision de $\pm 0,5$ litre, et ce avec un niveau de confiance de 95 %. Supposons que le magazine finance une étude préliminaire sur un « pré-échantillon » de 20 voitures, et que la moyenne de consommation de ces voitures pour 100 km soit $\bar{x}_{étude prél.} = 8,2$ litres et que la variance échantillonnale soit de $s_{étude prél.}^2 = \frac{1}{20-1} \sum_{i=1}^{20} (x_i - 8,2)^2 = 4$ litres². Maintenant qu'on a une estimation de s^2 , il suffit d'isoler le n dans la formule :

$$\begin{aligned}
 E = 0,5 &= z_{\alpha/2} \sqrt{\frac{s_{\text{étude prél.}}^2}{n}} \\
 &= 1,96 \sqrt{\frac{4}{n}} \\
 \Rightarrow 0,5^2 &= 1,96^2 \left(\frac{4}{n} \right) \\
 \Rightarrow n &= 1,96^2 \times \frac{4}{0,5^2} = 61,5 \text{ voitures} \approx 62 \text{ voitures.}
 \end{aligned}$$

Il suffit pour l'expérimentateur de compléter l'échantillon jusqu'à avoir fait l'essai de 62 voitures. En somme, un échantillon de 62 voitures assure au client que la précision des résultats sera telle que demandée, soit $\pm 0,5$ litre, et ce avec un niveau de confiance de 95 % (en autant que la variance échantillonnale n'augmente pas trop en complétant l'échantillon, auquel cas la précision de $\pm 0,5$ litre pourrait ne pas être respectée au niveau de confiance de 95 %). Si le client trouve l'échantillon nécessaire trop important, il peut réviser à la baisse la précision désirée ou encore il peut diminuer le niveau de confiance $1 - \alpha$ des résultats.

3.6 Tests d'hypothèses sur une moyenne

Trop souvent oubliés, les tests d'hypothèses font partie intégrante de la science de la statistique. Contrairement aux intervalles de confiance, la valeur du paramètre est fixée à une valeur hypothétique par un expert (ou autre). Cette valeur sera considérée comme vraie jusqu'à preuve du contraire. C'est le seul endroit où l'expert se confronte à un échantillon.

Les tests d'hypothèses permettent d'évaluer s'il y a suffisamment d'évidence statistique dans un échantillon pour supporter ou rejeter une hypothèse sur la valeur d'un paramètre de la population. Un test d'hypothèses confronte une hypothèse de base H_0 à une contre-hypothèse H_1 :

H_0 : C'est l'hypothèse nulle ; elle est considérée vraie jusqu'à preuve du contraire (c'est-à-dire que les calculs pour le test se basent sur cette affirmation).

H_1 : C'est la contre-hypothèse ou l'hypothèse alternative. **Lorsque la notion d'égalité s'applique, soulignons que cette hypothèse ne contient jamais le signe de l'égalité.**

Pour savoir si une hypothèse faite sur un paramètre de la population est statistiquement supportée, il faut élaborer un test d'hypothèses et se servir d'une règle de décision pour trancher. Mentionnons que la règle de décision est toujours basée sur le fait que l'on considère que H_0 est vraie, ce qui influence les calculs.

Pour des variables continues, il existe trois types de contre-hypothèses possible, ce qui mène à trois types de tests d'hypothèses. Plus précisément, ce sont les tests bilatéral, unilatéral à gauche et unilatéral à droite.

Test bilatéral :	Intuitivement, il faut rejeter H_0 lorsque la moyenne \bar{x} de l'échantillon, qui estime μ , est soit trop grande, soit trop petite par rapport à la valeur hypothétique centrale μ_0 .
Test unilatéral à gauche :	Intuitivement, il faut rejeter H_0 lorsque la moyenne \bar{x} de l'échantillon, qui estime μ , est trop petite par rapport à la valeur hypothétique μ_0 .
Test unilatéral à droite :	Intuitivement, il faut rejeter H_0 lorsque la moyenne \bar{x} de l'échantillon, qui estime μ , est trop grande par rapport à la valeur hypothétique μ_0 .

Il faut prendre note que la valeur μ_0 est fixée à la valeur de l'hypothèse fournie par l'expert, et est donc toujours connue.

En s'appuyant sur l'échantillon et en se basant sur une règle de décision préalablement établie, un test d'hypothèses fera pencher la balance du côté de H_0 ou de H_1 , et ce, avec un certain risque puisque nous nous basons sur un échantillon et non la population.

Les risques : Il est statistiquement impossible de toujours prendre une bonne décision. Alors nous consentons à l'avance de prendre un risque lors d'un test d'hypothèses. Les risques possibles sont les suivants :

Erreur de type I : Rejeter H_0 alors qu'elle est vraie.

Erreur de type II : Ne pas rejeter H_0 alors qu'elle est fausse.

Il n'est pas toujours possible de calculer ce risque.

Lorsque l'on teste une hypothèse, la probabilité avec laquelle on est disposé à risquer une erreur de type I est appelée **seuil de signification du test**. On désigne cette probabilité par α , qui est en fait la probabilité suivante :

$$\begin{aligned}\alpha &= P(\text{ rejeter } H_0 \mid H_0 \text{ est vraie }) \\ &= P(\text{ choisir } H_1 \mid H_0 \text{ est vraie }).\end{aligned}$$

Ce risque est toujours fixé **avant** de faire le test d'hypothèses. Habituellement, on fixe α à 0,01, 0,05 ou 0,10, tout dépendant des conséquences de rejeter à tort H_0 . Si on choisit par exemple 0,05 comme seuil de signification en construisant un test d'hypothèse, il y a alors 5 chances sur 100 pour que l'on rejette l'hypothèse nulle quand elle doit être acceptée. Cela signifie que l'on est sûr à 95 % d'avoir pris la bonne décision. On dit alors que l'on a **rejeté l'hypothèse à un seuil de signification de 5 %**.

Étant donné qu'il n'est pas toujours possible de calculer l'erreur de type II, il est courant de choisir un seuil de signification de 0,05 car c'est un risque acceptable et qui n'augmente pas trop l'erreur de type II. En effet, si on choisit un seuil de signification de 0,01, on prend très peu de chances de rejeter à tort H_0 , mais on augmente le risque de l'accepter alors qu'elle n'est pas vraie, ce qui est précisément l'erreur de type II que l'on ne peut calculer en général.

Pour les trois types de tests d'hypothèses, nous obtenons les règles de décision suivantes qui sont basées sur la notion statistique de *p*-value. La *p*-value représente la proba-

bilité de rejeter H_0 à tort une fois l'échantillon observé. La figure 3.22 présente les règles de décisions basées sur la p -value, le type de test et le seuil de signification α .

Hypothèses	Règle de décision
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	Lorsque la p -value $< \alpha$, nous rejetons H_0 et admettons H_1 comme vraisemblable. Sinon, nous admettons H_0 comme vraisemblable.
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	Lorsque la p -value $< 2\alpha$ et que $\bar{x} < \mu_0$, nous rejetons H_0 et admettons H_1 comme vraisemblable. Sinon, nous admettons H_0 comme vraisemblable.
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	Lorsque la p -value $< 2\alpha$ et que $\bar{x} > \mu_0$, nous rejetons H_0 et admettons H_1 comme vraisemblable. Sinon, nous admettons H_0 comme vraisemblable.

FIG. 3.17 – Règles de décision avec la p -value

Remarque : Lors de la résolution d'un test d'hypothèses, deux conclusions sont possibles : rejeter H_0 ou ne pas rejeter H_0 . Lorsqu'on rejette H_0 , il faut toujours mentionner le risque que l'on prend, qui est le seuil de signification. On insère ainsi une phrase comme « Au risque de se tromper une fois sur 20... » selon le seuil fixé. Lorsqu'on ne rejette pas H_0 , le risque que l'on prend est celui relié à l'erreur de type II et n'est pas connu en général. **Dans ce cas on ne peut donc pas mentionner le risque de façon précise** (on ne peut pas écrire « Au risque de se tromper une fois sur 20... »). On doit se contenter de mentionner qu'on ne rejette pas H_0 au seuil α .

Exemple 3.6.1 Supposons que deux articles soient parus dernièrement dans deux journaux concurrents. Le premier journaliste « expert » soutient que le nombre moyen d'employés travaillant à temps plein dans les hôpitaux est de 114 employés. Le second stipule plutôt que ce nombre moyen est de 180 employés. Basé sur ces hypothèses et les données de l'exemple 3.2.1, étudier une à une les deux hypothèses, et fixer le seuil de signification α à 0,05.

Pour effectuer un test d'hypothèse avec SPSS, il suffit d'exécuter les commandes présentées ci-dessous. Le calcul du test d'hypothèses compare la moyenne \bar{x} de l'échantillon en main à la moyenne théorique (d'où l'expression **Compare Means** de SPSS). Comme le comportement de \bar{x} est modélisé par la loi de Student (souvent appelée le T de Student), le test s'appelle **One Sample T Test** dans SPSS.

Le premier test comporte les hypothèses suivantes :

$$H_0 : \mu = 114$$

$$H_1 : \mu \neq 114$$

Pour réaliser ce test, les commandes sont les suivantes :

Menu SPSS :	→ Analyse
	→ Compare Means
	→ One Sample T Test
Dans la fenêtre Test Variable(s) :	→ nb_emp_tp
Test Value :	→ 114

On obtient alors les sorties 3.18 et 3.19.

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
nb_emp_tp	12	141,00	44,276	12,781

FIG. 3.18 – Quelques statistiques descriptives

Le premier tableau (figure 3.18) contient quelques statistiques descriptives. C'est sur le deuxième tableau (figure 3.19) que l'on se base pour résoudre le test.

On voit d'abord que la valeur que l'on teste apparaît en haut du tableau (**Test Value = 114**).

One-Sample Test						
	Test Value = 114					95% Confidence Interval of the Difference
	t	df	Sig. (2-tailed)	Mean Difference	Lower	
nb_emp_tp	2,112	11	,058	27,000	-1,13	55,13

FIG. 3.19 – Le tableau contenant la *p*-value (Sig. (2-tailed)) pour résoudre le test

Ici, puisque le test est bilatéral, la règle de décision est de rejeter H_0 si la *p*-value (Sig. (2-tailed)) est plus petite que le seuil de signification $\alpha = 0,05$. Rappelons que α est une probabilité qui représente le plus haut risque (de rejeter H_0 à tort) que l'on tolère. La *p*-value représente la probabilité (basée sur l'échantillon) de se tromper en rejetant H_0 . Il est donc tout à fait normal de passer à l'action et de rejeter H_0 si la *p*-value est inférieure au seuil de tolérance α . Ici, la *p*-value = 0.058 > $\alpha = 0.05$ (basé sur l'échantillon en main, le risque d'erreur observé est supérieur à la tolérance $\alpha = 0.05$). Donc nous ne rejetons pas H_0 . Ainsi, nous considérons vraisemblable que le nombre moyen d'employés dans les hôpitaux soit 114. Rappelons que l'échantillon possède peu d'observations.

Dans notre échantillon, rappelons que la moyenne était de 141,00 avec un écart type de 44,28. Pour être en mesure d'évaluer à quel point 141 est loin de 114, le praticien peut utiliser la cote-*t* = 2,112((141 - 114)/12.78 = 2.112). Cette cote-*t* représente le nombre d'écart-types distançant la valeur \bar{x} de la valeur que l'on teste $\mu_0 = 114$. En général, une valeur $|t| > 2$ tend à faire rejeter H_0 . Plus $|t|$ s'éloigne de 2 (2 écart-types), plus le rejet est puissant. Plus la taille de l'échantillon est grande (30 unités et plus), plus cette règle du pouce est vraie. Ici, l'échantillon est petit ($n = 12$ hôpitaux), ce qui explique la présente discordance.

Mentionnons que l'expression de la cote-*t* est $\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$. Cette statistique obéit aux règles du « 1,2,3 écart-types » suivantes :

- Environ 70 % (en fait 68,26 %) des échantillons en main produiront des estimations

à plus ou moins 1 un écart-type de la cible μ .

- Environ 95 % des échantillons en main produiront des estimations à plus ou moins 2 (en fait 1,96) écart-types de la cible μ .
- Un peu plus de 99 % des échantillons en main produiront des estimations à plus ou moins 3 écart-types de la cible μ .

Ce qui dépasse 2 écart-types devient « douteux ». Plus l'échantillon est grand, plus cette règle du pouce est meilleure.

Il faut maintenant tester les autres hypothèses :

$$H_0 : \mu = 180$$

$$H_1 : \mu \neq 180$$

Pour résoudre ce test, il faut faire les mêmes commandes que précédemment, et maintenant on a **Test Value** : 180. On obtient les tableaux suivants :

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
nb_emp_tp	12	141,00	44,276	12,781

FIG. 3.20 – Les statistiques descriptives demeurent les mêmes

	Test Value = 180					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
nb_emp_tp	-3,051	11	,011	-39,000	-67,13	-10,87

FIG. 3.21 – Le tableau contenant la *p*-value (Sig. (2-tailed)) pour résoudre le test

La règle de décision demeure la même : on rejette H_0 si la *p*-value (Sig. (2-tailed)) est plus petite que le seuil de signification $\alpha = 0,05$.

Ici, la p -value = $0.011 < \alpha = 0.05$; nous rejetons donc H_0 . Ainsi, avec une probabilité d'erreur de 5 %, nous considérons qu'il y a suffisamment d'évidence statistique dans notre échantillon pour rejeter l'hypothèse H_0 qui soutient que le paramètre du nombre moyen d'employés dans les hôpitaux est de 180.

La valeur **Mean Difference** est la différence entre la moyenne échantillonnale et la valeur μ_0 que l'on teste. SPSS offre aussi un intervalle de confiance pour cette différence de moyennes. Lorsque cet intervalle de confiance contient 0, on ne rejette pas H_0 ; dans le cas contraire, on rejettéra H_0 et on admettra H_1 comme vraisemblable.

Les intervalles de confiance et les tests d'hypothèses bilatéraux sont intimement liés. Cependant, certaines choses les différencient. Entre autres, dans les intervalles de confiance, l'accent est mis sur la probabilité d'avoir raison (le niveau de confiance $1 - \alpha$), que nous voulons la plus grande possible, tandis que dans les tests d'hypothèses, l'accent est mis sur la probabilité d'avoir tort (le seuil de signification $\alpha = 0,05$) que nous voulons la plus petite possible.

3.6.1 Exemple de calcul d'une p -value

Nous illustrons ici comment se calcule une p -value. C'est la base de données **satisfactiontravail.sav** et la variable **part_q1** qui sont utilisées pour cet exemple. Prenons les hypothèses suivantes :

$$H_0 : \mu = 2,3$$

$$H_1 : \mu \neq 2,3$$

Résolvons ce test au seuil $\alpha = 0,05$. Voici les sorties SPSS qu'on obtient :

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
part_q1	721	2,0687	2,58760	,09637

One-Sample Test						
	Test Value = 2,3					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
part_q1	-2,401	720	,017	-,23135	-,4205	-,0422

FIG. 3.22 – Les sorties

La cote-*t* est obtenue de la façon suivante :

$$\begin{aligned}
 t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\
 &= \frac{2,0687 - 2,3}{2,5876/\sqrt{721}} \\
 &= -2,401.
 \end{aligned}$$

Ici, puisque la taille d'échantillon est grande, les calculs pour la *p*-value faits avec la loi normale coïncideront avec ceux de SPSS (qui sont faits avec la loi de Student).

Donc tout d'abord, la *p*-value correspond à l'aire sous la courbe de la loi normale avant la cote-*t* qui est ici -2,401. Si cette cote-*t* est plus petite que la valeur critique inférieure (voir la figure 3.23), on se retrouve alors dans la zone de rejet inférieure et on rejette H_0 (il est impossible de se retrouver dans la zone de rejet supérieure puisque la cote-*t* est négative). Si c'est le cas, alors la *p*-value sera plus petite que l'aire avant la valeur critique inférieure, qui dans le cas de l'exemple est de 0,025 (puisque le seuil est de 5 %).

Pour trouver cette aire (la *p*-value), on prend une table de la loi normale :

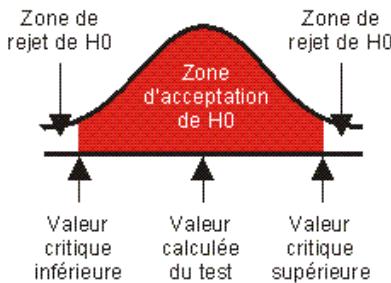


FIG. 3.23 – Zones de rejet

$$\begin{aligned}
 \text{Aire} &= P(Z < -2,401) \\
 &= P(Z > 2,401) \text{(puisque la loi normale est symétrique)} \\
 &= 0,5 - P(0 < Z < 2,401) \\
 &= 0,5 - 0,4918 \\
 &= 0,0082.
 \end{aligned}$$

Puisque $0,0082 < 0,025$, on rejette H_0 . Ou, si on préfère, on peut multiplier cette p -value par deux pour la comparer à notre seuil de 5 % ; c'est ce que fait SPSS. On a $0,0082 \times 2 = 0,0164$, ce qui est effectivement (à peu de chose près) la p -value donnée par SPSS dans la sortie 3.22. Encore une fois, puisque $0,0164 < 0,05$, on arrive à la même conclusion : on rejette H_0 au risque de se tromper une fois sur 20.

3.7 Tests d'hypothèses sur une proportion

Le contenu de la section 3.6 s'applique aussi pour les tests d'hypothèses dans lesquels on compare une proportion à une valeur hypothétique. Voyons un exemple.

Exemple 3.7.1 L'exemple qui suit se base sur une véritable enquête qui a été menée sur le stress au travail d'un échantillon représentatif des employés dans une entreprise

internationale. Le questionnaire a été passé en France et au Canada. La base de données se nomme **stresstravail.sav**.

La figure 3.24 donne un aperçu des variables qui étaient à l'étude.

	Name	Type	Width	Decimals	Label	Values	
1	id	Numeric	8	0	Numéro du questionnaire	None	N
2	pays	Numeric	8	0	Pays dans lequel le questionnaire a été administré	{1, Canada}...	N
3	peur_sup	Numeric	8	2	Peur d'exprimer son désaccord avec un supérieur	None	N
4	accs_sup	Numeric	8	2	Accessibilité des supérieurs (téléphone, rendez-vous)	None	N
5	stress	Numeric	8	2	Stress au travail	None	N
6	sexe	Numeric	8	0	Sexe du répondant	{0, homme}...	N
7	âge	Numeric	8	0	Classe d'âge du répondant	{1, moins de 1	N
8	statut	Numeric	8	0	Statut professionnel du répondant	{1, employé de	N
9	anciennet	Numeric	8	0	Ancienneté dans la société	{1, moins d'1 a	N
10	scolarit	Numeric	8	0	Plus haut niveau du diplôme obtenu	{1, primaireC a	N
11							

FIG. 3.24 – Les variables à l'étude

Il est à noter que les questions **peur_sup**, **accs_sup** et **stress** ont été mesurées à l'aide d'une échelle qui est constituée d'une ligne de 15 cm de long et sur laquelle l'individu est invité à apposer un « X » sur la partie de la droite qui correspond à son opinion. Avec cette droite, toutes les valeurs sont possibles. Voici un exemple lié à la variable **peur_sup** du fichier SPSS :

Q₂ : J'ai peur d'exprimer mon désaccord à mon supérieur immédiat. (Mettre un X sur la partie de la droite qui correspond le mieux à votre opinion, en sachant que le 0 correspond à tout à fait en désaccord, et le 15 à tout à fait en accord.)



On se demande s'il y a beaucoup d'employés qui se disent très stressés. En fait, le directeur des ressources humaines se dit qu'un niveau de stress de 10 ou plus révèle un problème, et il se demande quel est le pourcentage des employés qui sont en haut de cette borne. Il espère que c'est en-dessous de 10 %.

Pour investiguer de ce côté, il commence par recodifier la variable **stress** en une nouvelle variable **stress2** qui ne prend que deux valeurs : 0 lorsque le niveau de stress est inférieur à 10, et 1 lorsqu'il est de 10 ou plus. Puis il veut tester si la proportion des employés qui ont 10 ou plus est supérieure ou égale à 10 %, et ce au seuil $\alpha = 0,05$. Il faut donc résoudre le test suivant :

$$H_0 : \pi_{\text{stress}>10} = 0,10$$

$$H_1 : \pi_{\text{stress}>10} < 0,10$$

Pour obtenir la *p*-value pour ce test il faut effectuer les commandes suivantes (avec la nouvelle variable **stress2**) :

Menu SPSS :	→ Analyse
	→ Compare Means
	→ One Sample T Test
Dans la fenêtre Test Variable(s) :	→ stress2
Test Value :	→ 0,1

On obtient alors les sorties suivantes :

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
stress2	161	,0870	,28265	,02228

FIG. 3.25 – Quelques statistiques descriptives

Dans la première sortie on voit que la proportion des employés qui ont un niveau de stress supérieur ou égal à 10 (la moyenne des « 1 ») est de 0,087, soit 8,7 %.

Puisque le test est unilatéral à gauche, on rejette H_0 au seuil $\alpha = 0,05$ si $p < \pi_0 = 0,10$ et si la *p*-value est plus petite que $2\alpha = 0,1$. La première condition est satisfaite puisque

One-Sample Test						
	Test Value = 0.1					95% Confidence Interval of the Difference
	t	df	Sig. (2-tailed)	Mean Difference	Lower	
stress2	-,586	160	,559	-,01304	-,0570	,0309

FIG. 3.26 – Le tableau contenant la *p*-value (Sig. (2-tailed)) pour résoudre le test

$p = 0,087 < 0,10$. Par contre la *p*-value est égale à 0,559, ce qui n'est pas plus petit que 0,10. Autrement dit, d'après l'échantillon, on a 55,9 % de chances de se tromper en rejetant H_0 . Bien que 8,7 % soit plus petit que 10 %, la différence n'est pas assez grande pour qu'on soit sûr que ce n'est pas dû aux fluctuations échantillonnelles. Donc on admet plutôt qu'au niveau de la population, la proportion des employés qui ont un niveau de stress supérieur ou égal à 10 est de 10 % ou plus.

Que se serait-il passé si le directeur avait plutôt voulu tester si la proportion des employés qui ont 10 ou plus est supérieure à 10 % au seuil $\alpha = 0,05$?

3.8 Utilisation de l'écart-réduit

Dans les sections précédentes, nous avons vu comment résoudre un test d'hypothèses à l'aide d'une p -value. Il est également possible de résoudre un test d'hypothèses en calculant l'écart-réduit approprié, puis en le comparant au seuil de probabilité qui convient selon le contexte. De plus, nous donnons ici les conditions d'application de ces tests.

3.8.1 Tests sur une moyenne

Nous voyons ici comment résoudre un test d'hypothèses avec une moyenne en utilisant l'écart-réduit, et ce à un seuil de signification α . Les tableaux suivants résument les règles de décision selon le contexte et les hypothèses.

Hypothèses statistiques	Écart réduit	Règle de décision
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$	Rejeter H_0 si $Z > z_{\alpha/2}$ ou $Z < -z_{\alpha/2}$
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$	Rejeter H_0 si $Z > z_\alpha$
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$	Rejeter H_0 si $Z < -z_\alpha$

Conditions : Population normale de variance connue,
ou grand échantillon ($n \geq 30$) et variance de la population connue.

Hypothèses statistiques	Écart réduit	Règle de décision
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	Rejeter H_0 si $T > t_{\alpha/2;n-1}$ ou $T < -t_{\alpha/2;n-1}$
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	Rejeter H_0 si $T > t_{\alpha;n-1}$
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	Rejeter H_0 si $T < -t_{\alpha;n-1}$

Conditions : Population normale (ou approximativement normale) de variance inconnue, et échantillon de taille $n \leq 120$.

Hypothèses statistiques	Écart réduit	Règle de décision
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	Rejeter H_0 si $Z > z_{\alpha/2}$ ou $Z < -z_{\alpha/2}$
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	Rejeter H_0 si $Z > z_{\alpha}$
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	Rejeter H_0 si $Z < -z_{\alpha}$

Conditions : Population normale (ou approximativement normale) de variance inconnue, et échantillon de taille $n > 120$.

Rappel

Seuil de signification	Valeurs critiques
$\alpha = 0,01$	$z_{\alpha/2} = z_{0,005} = 2,576$, $z_{\alpha} = z_{0,01} = 2,33$
$\alpha = 0,05$	$z_{\alpha/2} = z_{0,025} = 1,96$, $z_{\alpha} = z_{0,05} = 1,645$
$\alpha = 0,10$	$z_{\alpha/2} = z_{0,05} = 1,645$, $z_{\alpha} = z_{0,1} = 1,28$

Exemple 3.8.1 Un distributeur de bières affirme qu'une nouvelle présentation, affichant en grandeur réelle un chanteur de rock connu, augmentera les ventes dans les supermarchés de 50 caisses par semaine. On fait un test dans 20 supermarchés, et l'augmentation moyenne en une semaine est de 41,3 caisses avec un écart-type de 12,2 caisses. Testez l'affirmation selon laquelle l'augmentation est d'au moins 50 caisses au seuil $\alpha = 0,05$, en précisant sous quelle(s) hypothèse(s) vous pouvez faire ce test.

Solution. On a $n = 20$, $\bar{x} = 41,3$, $s = 12,2$ et $\alpha = 0,05$. On doit supposer que la distribution des données au niveau de la population est approximativement normale puisque σ est inconnue.

Les hypothèses sont les suivantes :

$$H_0 : \mu = 50$$

$$H_1 : \mu < 50$$

On a

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{41,3 - 50}{12,2/\sqrt{20}} = -3,19.$$

On doit comparer T à $-t_{\alpha;n-1} = -t_{0,05;19} = -1,7291$. Puisque $T = -3,19 < -1,7291$, on rejette H_0 au seuil $\alpha = 0,05$. Ainsi au risque de se tromper une fois sur vingt, on peut affirmer que l'augmentation n'est pas d'au moins 50 caisses par semaine.

3.8.2 Tests sur une proportion

Dans cette sous-section, nous voyons comment résoudre un test d'hypothèses avec une proportion en utilisant l'écart-réduit, et ce à un seuil de signification α . Le tableau suivant résume les règles de décision selon les hypothèses.

Hypothèses statistiques	Écart réduit	Règle de décision
$H_0 : \pi = \pi_0$ $H_1 : \pi \neq \pi_0$	$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	Rejeter H_0 si $Z > z_{\alpha/2}$ ou $Z < -z_{\alpha/2}$
$H_0 : \pi = \pi_0$ $H_1 : \pi > \pi_0$	$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	Rejeter H_0 si $Z > z_{\alpha}$
$H_0 : \pi = \pi_0$ $H_1 : \pi < \pi_0$	$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	Rejeter H_0 si $Z < -z_{\alpha}$

Conditions : $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$.

Exemple 3.8.2 Selon un sondage réalisé par le Groupe Everest pour le compte de la Banque Nationale et de *La Presse* (février 2001), deux PME sur cinq affirment avoir adopté Internet dans leur stratégie de marketing. C'est ce que révèle l'enquête auprès de 300 PME québécoises comptant dix à deux cents employés.

- Quelle est, pour un niveau de confiance de 95 %, la marge d'erreur statistique de ce sondage ?
- La présidente de la Chambre de Commerce de la région de Drummondville précise, dans une conférence auprès des membres de la Chambre, que 45 % des PME de la région ont adopté Internet dans leur stratégie de marketing. Est-ce que les données obtenues du sondage mentionné ci-haut sont en contradiction avec l'affirmation de la présidente de la Chambre de Commerce ?

Solution. On a $n = 300 > 30$ et $p = 2/5 = 0,40$. Ainsi

$$np = 300 \cdot 0,4 = 120 > 5,$$

$$n(1 - p) = 300 \cdot 0,6 = 180 > 5$$

et donc on peut faire un test d'hypothèses sur cette proportion.

a) Pour un niveau de confiance $1 - \alpha = 0,95$, on a

$$E = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} = 1,96 \sqrt{\frac{0,4 \cdot 0,6}{300}} = 0,055.$$

Ainsi la marge d'erreur statistique de ce test est de 5,5 %.

b) Répondre à cette question revient à tester les hypothèses suivantes :

$$H_0 : \pi = 0,45$$

$$H_1 : \pi \neq 0,45$$

Fixons le seuil de signification à $\alpha = 0,05$. On a

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0,4 - 0,45}{\sqrt{\frac{0,45 \cdot 0,55}{300}}} = -1,74.$$

On a $z_{\alpha/2} = z_{0,025} = 1,96$. Puisque $Z = -1,74$ est compris entre -1,96 et 1,96, on ne rejette pas H_0 au seuil $\alpha = 0,05$. Donc les données du sondage ne sont pas en contradiction avec l'affirmation de la présidente de la Chambre de Commerce.

Autre solution en utilisant a) : La marge d'erreur trouvée en a) nous indique que l'intervalle de confiance de niveau 95 % pour la proportion des PME qui ont adopté Internet dans leur stratégie de marketing est $[0,4 - 0,055, 0,4 + 0,055] = [0,345, 0,455]$. Puisque le test d'hypothèses à résoudre pour répondre à la question est bilatéral (voir dans la première solution), on dira que le sondage est en contradiction avec l'affirmation si 0,45 n'est pas dans l'intervalle. Or on voit que $0,45 \in [0,345, 0,455]$ (de justesse !), donc les données du sondage ne sont pas en contradiction avec l'affirmation de la présidente de la Chambre de Commerce.

3.9 Exercices du chapitre

Exercice 1 Reprenons le contexte de l'exercice du chapitre 2 (2.4). Une étude est menée sur les différences de salaire entre les cadres de trois provinces. Pour ce faire, trente cadres, dont les tâches et les responsabilités sont similaires, ont été choisis au hasard. À l'aide de SPSS, effectuez les manipulations et analyses suivantes avec la base de données que vous avez créée dans l'exercice du chapitre 2.

1. Combien y a-t-il d'hommes et de femmes (donnez les fréquences) ? Donnez un intervalle de confiance de niveau 95 % sur la proportion des femmes qui occupent ce type de poste. (Attention, le logiciel ne donne les estimations que pour la modalité associée au code « 1 »...)
2. Faites une analyse descriptive de la variable **salaire**.
3. Élaborez un intervalle de confiance pour le véritable salaire moyen des cadres qui occupent ce type de poste.
4. Votre supérieur estime selon sa grande expérience que le véritable salaire moyen pour ce type d'emploi est de 75 000 \$. Appuyez-vous cette affirmation ?

Exercice 2 On veut faire un sondage sur le campus de l'université de Sherbrooke pour savoir si la marque de bière « La Moufette » est connue par les étudiants. Plus précisément, on veut estimer la proportion Π de la population étudiante de l'université de Sherbrooke qui connaît cette marque de bière. On aimerait avoir une précision de 3,5 % sur l'estimation, et ce avec une confiance $1 - \alpha = 0,95$.

1. Quelle est la taille de l'échantillon qu'il vous faut planifier si aucune autre donnée n'est disponible ?
2. Votre temps est restreint, et vous savez que vous ne pourrez sonder plus de 400 étudiants. Vous devrez donc modifier le niveau de confiance ou la précision. Si vous voulez garder le même niveau de confiance, quelle précision sur l'estimation pourrez-vous avoir ?

Exercice 3 On sélectionne au hasard 50 employés d'une entreprise estrienne, et on étudie la variable salaire sur cet échantillon. On obtient une moyenne de 42 323 \$, et une variance de 51 840 000 \$. Il faudra compléter cet échantillon à combien d'employés si l'on désire une précision de 1 000 \$ pour un intervalle de confiance de niveau 95 % ?

Exercice 4 Formuler les hypothèses H_0 et H_1 pertinentes aux affirmations suivantes que l'on voudrait tester sur la base d'un échantillon.

1. Le niveau de stress des employés, évalué sur une échelle de 0 à 10, est d'au plus 5.
2. Le nombre moyen d'employés au Québec pour les entreprises québécoises œuvrant dans la production de logiciels est de 18,5.
3. La période moyenne de recouvrement des achats informatiques par les entreprises est d'au moins 3 ans.
4. Une entreprise envisage le lancement d'un nouveau produit si le taux d'intentions positives d'achat est d'au moins 60 %.
5. Le niveau moyen de satisfaction des employés, évalué sur une échelle de 0 à 10, est de 7,5.
6. La durée de vie d'un pneu automobile est actuellement de 48 000 km. L'introduction d'une nouvelle fibre dans la fabrication du pneu pourrait améliorer la durée de vie.
7. Le temps moyen de réponse d'un ordinateur central est d'au plus 1,4 s.
8. Au moins 60 % des dirigeants de PME affirment que le principal enjeu des relations de travail au sein de leur entreprise est le salaire.

Exercice 5 Une étude est menée à la faculté d'administration sur une expérience de magasinage sur certains sites web. La base de données porte le nom `magasinageweb.sav`.

1. Les étudiants devaient magasiner sur un site qui leur était imposé. Quelle est la répartition des fréquences par rapport à ces sites ?

2. Combien ont acheté un produit, et combien n'en ont pas acheté ? Donnez un intervalle de confiance de niveau 95 % sur la proportion des étudiants qui achètent un produit.
3. Faites une analyse descriptive de la variable `plaisirm`.
4. Élaborez un intervalle de confiance pour le véritable niveau moyen de plaisir des étudiants qui magasinent sur ces sites.
5. Un de vos amis est persuadé que les étudiants ont beaucoup de plaisir à magasiner sur ces sites, et estime donc que la moyenne du plaisir devrait être d'au moins 7 sur 9. Appuyez-vous cette affirmation ?

Exercice 6 Une étude est menée au sein d'une entreprise pour évaluer la satisfaction et la participation des employés. La base de données porte le nom `satisfactiontravail.sav`.

1. Les employés devaient indiquer leur âge ; les réponses sont réparties dans 10 classes d'âges. Quelle est la répartition des fréquences pour ces classes ? Donnez un intervalle de confiance de niveau 95 % sur la proportion des employés qui ont entre 26 et 30 ans.
2. On s'intéresse à savoir à quel point les employés ont l'impression d'influencer la façon de faire leur travail. Pour vous faire une idée, faites une analyse descriptive de la variable `part_q3`.
3. Élaborez un intervalle de confiance pour le véritable niveau moyen d'influence perçue par les employés sur la façon de faire leur travail.
4. Un des dirigeants de l'entreprise estime que le niveau d'influence perçue par les employés sur la façon de faire leur travail est d'au moins 9,5 sur 12,3. Appuyez-vous cette affirmation ?

Chapitre 4

Relation entre deux variables discrètes

L’analyse du lien unissant deux variables discrètes X et Y présentée dans ce chapitre vaut autant pour étudier les relations de dépendance ($X \Rightarrow Y$) que d’interdépendance ($X \Leftrightarrow Y$). Pour y arriver, le praticien utilise les tableaux croisés aussi connus sous le nom de tableaux de contingence.

La loi utilisée pour ces analyses est celle du chi-deux. La première section présente cette statistique, et quelques exemples d’utilisation de cette loi.

La section 4.2 expose une série d’analyses générales pouvant être faites sur tout tableau croisé, et ce, que le croisement contienne des variables nominales ou ordinaires. La section 4.3 s’intéresse au cas particulier où le praticien désire étudier le lien unissant deux variables ordinaires. La section 4.4 présente les pré-requis pour que ces analyses soient valides, et la section 4.5 présente un exemple supplémentaire. Finalement, la section 4.6 présente l’analyse en correspondances. Cette dernière section est optionnelle puisqu’elle introduit l’analyse des correspondances qui utilise un module avancé de SPSS, non disponible avec le module de base.

4.1 La statistique du chi-deux

Le test du chi-deux est très simple. Il mesure la distance entre les valeurs observées d'une expérience et les valeurs théoriques auxquelles on s'attendait.

Supposons donc que nous avons n données, réparties dans k classes différentes. On note $f_{o_1}, f_{o_2}, \dots, f_{o_k}$ le nombre de données qu'il y a dans chaque classe (c'est la fréquence absolue), et on note $f_{t_1}, f_{t_2}, \dots, f_{t_k}$ les fréquences auxquelles on s'attendait (selon le contexte). On remarque que l'on a

$$\sum_{i=1}^k f_{o_i} = n = \sum_{i=1}^k f_{t_i}$$

Pour mesurer à quel point les f_{o_i} sont éloignées ou non des f_{t_i} , on utilise la statistique du chi-deux, qui est la suivante :

$$\begin{aligned}\chi^2 &= \sum_{i=1}^k \frac{(f_{o_i} - f_{t_i})^2}{f_{t_i}} \\ &= \frac{(f_{o_1} - f_{t_1})^2}{f_{t_1}} + \frac{(f_{o_2} - f_{t_2})^2}{f_{t_2}} + \dots + \frac{(f_{o_k} - f_{t_k})^2}{f_{t_k}}\end{aligned}$$

Les seules conditions pour utiliser le test du chi-deux sont les suivantes :

- L'échantillon doit avoir été prélevé au hasard dans la population ;
- Toutes les fréquences théoriques (les f_{t_i}) doivent être plus grandes ou égales à 1, et au plus 20 % des fréquences théoriques sont inférieures à 5.

4.1.1 Hypothèses statistiques et règle de décision

La statistique du chi-deux sert à traiter un test d'hypothèses de la forme suivante :

H_0 : Dans la population, les données suivent la distribution théorique spécifiée.

H_1 : Dans la population, les données ne suivent pas la distribution théorique spécifiée.

Au seuil α , on rejette H_0 si

$$\chi^2 = \sum_{i=1}^k \frac{(f_{o_i} - f_{t_i})^2}{f_{t_i}} > \chi^2_{\alpha; \nu}$$

où $\nu = k - 1 - r$ est le nombre de degrés de liberté du chi-deux, avec r le nombre de paramètres de la loi théorique que l'on doit estimer avec l'échantillon. Dans notre contexte $r = 0$, et donc $\nu = k - 1$.

Exemple 4.1.2 Supposons que nous voulons savoir si un dé à 6 faces est équilibré. Nous effectuons 120 lancés et nous obtenons les résultats suivants :

Face du dé	1	2	3	4	5	6	Total
Fréquences observées	14	25	19	22	27	13	120

Nous voulons tester

H_0 : Le dé est équilibré.

H_1 : Le dé est pipé.

Si le dé est équilibré, alors on devrait obtenir les fréquences suivantes :

Face du dé	1	2	3	4	5	6	Total
Fréquences théoriques	20	20	20	20	20	20	120

Utilisons la statistique du chi-deux pour voir si les valeurs observées sont éloignées ou non des valeurs théoriques au seuil $\alpha = 0,05$. Ici, $\nu = k - 1 = 6 - 1 = 5$.

Il n'y a pas de problème à utiliser le test du chi-deux puisque toutes les fréquences théoriques sont plus grandes que 5. Le chi-deux observé est

$$\begin{aligned}\chi^2 &= \frac{(14 - 20)^2}{20} + \frac{(25 - 20)^2}{20} + \frac{(19 - 20)^2}{20} + \frac{(22 - 20)^2}{20} + \frac{(27 - 20)^2}{20} + \frac{(13 - 20)^2}{20} \\ &= \frac{164}{20} = 8,2.\end{aligned}$$

Or $\chi^2_{\alpha; \nu} = \chi^2_{0,05; 5} = 11,070$. Puisque $\chi^2 = 8,2 < 11,070$, on ne rejette pas H_0 . Donc au seuil $\alpha = 0,05$ on conclut que le dé est équilibré.

4.1.1 Représentativité de l'échantillon

Le test d'ajustement du chi-deux peut également être utilisé pour vérifier si un échantillon est représentatif dans un sondage ou une recherche. Il s'agit de comparer la répartition de l'échantillon (selon une ou des caractéristiques particulières) à celle qui existe dans la population.

Exemple 4.1.3 Dans une enquête sur des préférences liées à la consommation, on a la répartition (en %) suivante de l'échantillon ($n = 2\,265$ répondants) en regard de l'âge des répondants.

	15-24 ans	25-44 ans	45 ans et plus
Échantillon	20,8 %	53,8 %	25,4 %
Population	23,6 %	49 %	27,4 %

Peut-on conclure, au seuil de signification 5 %, que la répartition de l'échantillon en regard de l'âge est représentatif de la population ?

Il faut d'abord calculer les fréquences observées et théoriques pour chaque tranche d'âge (remarquez que pour les fréquences observées on arrondit pour avoir un nombre entier). On a

	15-24 ans	25-44 ans	45 ans et plus	Total
f_o	$20,8 \% \cdot 2\ 265 = 471$	1 219	575	2 265
f_t	$23,6 \% \cdot 2\ 265 = 534,54$	1 109,85	620,61	2 265

On observe tout d'abord que les fréquences théoriques sont plus grandes que 5, on peut donc procéder au test sans problème. Les hypothèses sont

H_0 : La répartition de l'échantillon en fonction de l'âge est représentative de la population.

H_1 : La répartition de l'échantillon en fonction de l'âge n'est pas représentative de la population.

On a

$$\begin{aligned}\chi^2 &= \frac{(471 - 534,54)^2}{534,54} + \frac{(1219 - 1109,85)^2}{1109,85} + \frac{(575 - 620,61)^2}{620,61} \\ &= 21,639.\end{aligned}$$

Or $\chi^2_{\alpha; \nu=k-1} = \chi^2_{0,05; 2} = 5,991$. Puisque $\chi^2 = 21,639 > 5,991$, on rejette H_0 . Donc au risque de se tromper une fois sur vingt, on conclut que la répartition de l'échantillon en regard de l'âge n'est pas représentatif de la population.

4.2 Les tableaux de contingence

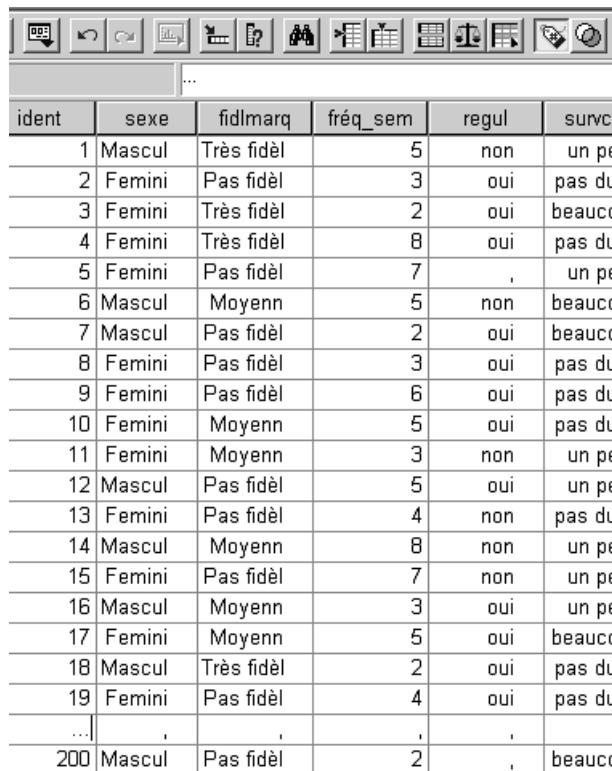
Les tableaux de contingence sont simplement des tableaux permettant de visualiser la répartition des fréquences des observations liées à deux variables discrètes. La statistique du chi-deux nous permettra de tester si cette répartition révèle un lien entre les deux variables. S'il y a un lien, d'autres étapes suivront pour mieux comprendre celui-ci.

4.2.1 Le Chi-deux χ^2 : y a-t-il un lien significatif ?

Voici un exemple concernant une étude menée sur les tablettes de chocolat, toutes marques confondues. Ainsi, chacun des 200 participants devait répondre à quelques questions sur le sujet (sexe, niveau de fidélité à une marque de chocolat, fréquence des consommations par semaine, etc.). La base porte le nom `exchocolat.sav`. La figure 4.1 illustre un extrait simplifié de la base de données du fichier SPSS.

Un des nombreux objectifs de cette étude était d'identifier les variables qui ont une influence sur le niveau de fidélité de la population des consommateurs de chocolat. Une première question se pose : « Est-ce que le sexe a une influence sur le niveau de fidélité des consommateurs à leur marque de chocolat préférée ? ».

Pour mieux visualiser la situation, présentons d'abord le tableau croisé associé à la présente situation. La sortie 4.2 présente un tableau croisé standard illustrant le croisement demandé (`sexe ⇒ niveau de fidélité`). Pour obtenir cette sortie (figure 4.2), il faut effectuer les commandes suivantes :



ident	sexe	fidlmarq	fréq_sem	regul	survca
1	Mascul	Très fidèle	5	non	un peu
2	Femin	Pas fidèle	3	oui	pas du tout
3	Femin	Très fidèle	2	oui	beaucoup
4	Femin	Très fidèle	8	oui	pas du tout
5	Femin	Pas fidèle	7	,	un peu
6	Mascul	Moyenn	5	non	beaucoup
7	Mascul	Pas fidèle	2	oui	beaucoup
8	Femin	Pas fidèle	3	oui	pas du tout
9	Femin	Pas fidèle	6	oui	pas du tout
10	Femin	Moyenn	5	oui	pas du tout
11	Femin	Moyenn	3	non	un peu
12	Mascul	Pas fidèle	5	oui	un peu
13	Femin	Pas fidèle	4	non	pas du tout
14	Mascul	Moyenn	8	non	un peu
15	Femin	Pas fidèle	7	non	un peu
16	Mascul	Moyenn	3	oui	un peu
17	Femin	Moyenn	5	oui	beaucoup
18	Mascul	Très fidèle	2	oui	pas du tout
19	Femin	Pas fidèle	4	oui	pas du tout
...	,	,	,	,	,
200	Mascul	Pas fidèle	2	,	beaucoup

FIG. 4.1 – Extrait de la base de données de l'exemple

Menu SPSS :

→ Analyse

→ Descriptive Statistics

→ Crosstabs...

Dans la fenêtre Row(s) : → fidlmarq (variable dépendante)

Dans la fenêtre Column(s) : → sexe (variable indépendante)

Dans le bouton Cells... : → Percentages √ Column

→ Count √ Observed

√ Expected

→ Residuals √ Standardized

Dans le cas où le modèle étudié est un modèle de dépendance ($X \Rightarrow Y$, ici sexe \Rightarrow niveau de fidélité), il est préférable de **placer la variable indépendante (ici**

			SEXÉ		Total
fidèle à la marque?	Très fidèle		Feminin	Masculin	
		Count	18	58	76
		Expected Count	30,4	45,6	76,0
		% within SEXE	22,5%	48,3%	38,0%
		Std. Residual	-2,2	1,8	
	Moyennement fidèle	Count	18	26	44
		Expected Count	17,6	26,4	44,0
		% within SEXE	22,5%	21,7%	22,0%
		Std. Residual	,1	-,1	
	Pas fidèle	Count	44	36	80
		Expected Count	32,0	48,0	80,0
		% within SEXE	55,0%	30,0%	40,0%
		Std. Residual	2,1	-1,7	
Total		Count	80	120	200
		Expected Count	80,0	120,0	200,0
		% within SEXE	100,0%	100,0%	100,0%

FIG. 4.2 – Le tableau illustrant le croisement sexe \Rightarrow niveau de fidélité

sexe) en position de colonne dans le tableau. Pour étudier le lien (si lien il y a) avec des %, il faut cocher les « % en colonnes ». Cette méthode de travail facilite l’interprétation et la présentation des résultats. En effet, il est alors possible de lire le tableau de gauche à droite et de haut en bas. Dans le cas de l’étude d’une relation d’interdépendance ($X \Leftrightarrow Y$), il est plus simple pour le praticien d’étudier un sens de la relation à la fois.

Pour chacune des cellules du tableau de contingence 4.2, on retrouve d’abord le nombre d’individus observé dans l’échantillon (Count). Ce nombre observé sera comparé à un nombre théorique (Expected Count) qui est calculé selon l’hypothèse que les deux variables X et Y sont indépendantes. Ce tableau contient aussi les % en colonne ainsi que les résidus standardisés, qui serviront à exprimer le lien s’il y a lieu (nous y reviendront à la sous-section 4.2.3).

Dans le cadre de ce tableau de contingence, il est possible de voir que 200 personnes ont participé à l’étude : 80 d’entre elles étaient des femmes et 120 étaient des hommes. Selon le sexe des répondants, ils sont ventilés à travers les lignes du tableau, ce qui permet

de comparer la répartition des femmes et des hommes, et ce, pour chacun des niveaux de « fidélité ».

On remarque que la proportion d'hommes (48,3 %) très fidèles à leur marque de chocolat préféré est supérieure à celle des femmes (22,5 %). On a le phénomène inverse lorsqu'on observe les gens qui passent d'une saveur à l'autre. Plus précisément, 55 % des femmes se qualifient d'infidèles à une marque de chocolat préférée contre seulement 30 % des hommes.

Est-ce que ces fluctuations perçues dans la sortie 4.2 sont palpables au niveau de la population ou ne sont-elles que le fruit de l'imperfection d'un échantillon ? Pour répondre à cette question, on doit résoudre le test d'hypothèses suivant :

H_0 : Dans la population, X est indépendant de Y .

H_1 : Dans la population, X est lié à Y .

lequel, dans le contexte de l'exemple, s'écrit

H_0 : Dans la population, le niveau de fidélité des consommateurs de produits de chocolat est indépendant du sexe des individus.

H_1 : Dans la population, le niveau de fidélité des consommateurs de produits de chocolat est lié au sexe des individus.

La démarche statistique de tout test suppose que la première hypothèse (H_0) est vraie, et ce, jusqu'à preuve du contraire. Ainsi, s'il n'existe pas suffisamment d'évidence dans l'échantillon pour balancer H_0 au profit de H_1 , H_0 sera considérée comme étant vraie, et ainsi, **aucune autre analyse statistique et aucune interprétation ne seront effectuées vu l'absence de lien dans la population.**

Cependant, si dans l'échantillon il existe suffisamment de preuves pour rejeter H_0 , alors nous admettrons l'hypothèse de dépendance dans la population, H_1 , comme étant vraisemblable, et alors, **nous chercherons à trouver comment s'exprime cette dépendance** (sous-section 4.2.3).

C'est la statistique du Chi-deux χ^2 (Chi-square) qui permet de décider s'il y a suffisamment d'évidence statistique dans l'échantillon pour rejeter l'hypothèse d'indépendance H_0 , et pour ainsi admettre l'hypothèse de dépendance H_1 comme étant vraisemblable. Cette statistique est distribuée suivant la loi de probabilité du même nom. Les degrés de cette statistique sont $(r - 1) * (k - 1)$ où r et k représentent le nombre de modalités de réponses de la première et de la seconde variable discrète. Elle se calcule de la façon suivante :

$$\chi^2 = \sum \frac{(\text{Count} - \text{Expected Count})^2}{\text{Expected Count}}$$

S'il y a un lien entre les deux variables X et Y , alors les **Count** devraient s'éloigner des **Expected Count**, et de cette façon le χ^2 tend à devenir grand. C'est lorsque le χ^2 dépasse une borne théorique (le Chi-deux théorique $\chi^2_{(r-1)*(k-1)}$) que l'on rejette l'hypothèse H_0 selon laquelle il n'y a aucun lien entre les deux variables. Dans le cadre de ce cours, on se basera sur une p -value fournie par SPSS pour décider si on doit rejeter H_0 ou non. Pour obtenir cette p -value, il faut ajouter les commandes suivantes à celles que l'on a vues pour générer le tableau croisé :

Menu SPSS :	\rightarrow Analyse
	\rightarrow Descriptive Statistics
	\rightarrow Crosstabs...
Dans le bouton Statistics... : <input checked="" type="checkbox"/>	Chi-square

et alors la sortie 4.3 s'ajoutera à la sortie 4.2.

La sortie 4.3 contient une p -value (Asymp. Sig. (2-sided)) associée au Chi-deux (Pearson Chi-square) qui est calculée à partir de l'échantillon ; cette p -value calcule le risque de se tromper si l'on admet un lien. En effet, puisque nous ne sommes en présence

Chi-Square Tests			
	Value	df	A sym. Sig. (2-sided)
Pearson Chi-Square	15,945 ^a	2	,000
Likelihood Ratio	16,361	2	,000
Linear-by-Linear Association	15,830	1	,000
N of Valid Cases	200		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 17,60.

FIG. 4.3 – Le test d’indépendance avec le Chi-deux χ^2

que d’un échantillon, il est possible que nous détections la présence vraisemblable d’une dépendance alors qu’il n’en n’est rien.

Cette erreur porte le nom d’erreur de première espèce et se note α . Cette erreur est le seuil de signification du test d’hypothèses. Pour tous les tests d’hypothèses, l’industrie fixe cette erreur à l’une des trois quantités suivantes : $\alpha = 0,01$, $\alpha = 0,05$ ou $\alpha = 0,10$. Le standard de l’industrie fixe généralement le seuil de signification à $\alpha = 0,05$, ce qui établit le risque de se tromper, **lors de la découverte d’une dépendance**, à 1 fois sur 20.

Ainsi, si la praticien fixe lui-même son risque d’erreur à $\alpha = 0,05$, il utilisera la règle de décision suivante (basée sur le tableau 4.3) pour trancher son test d’hypothèses :

On rejette H_0 si le Pearson Chi-square admet une p -value (Asymp. Sig. (2-sided)) plus petite que le seuil $\alpha = 0,05$. Si non, on ne rejette pas H_0 .

Dans cet exemple, on remarque que la p -value de la statistique du Chi-deux a une valeur de ,000. Donc ici, peu importe le seuil α fixé, on rejettéra H_0 . Ainsi, au seuil $\alpha = 0,05$, l’analyste peut tirer la conclusion globale suivante :

« Au risque de se tromper 1 fois sur 20, nous sommes en mesure de dire que le sexe influence significativement le niveau de fidélité de la population cible face à la consommation de tablettes de chocolat. »

La *p*-value représente la probabilité empirique *a posteriori* (calcul effectué après avoir observé le comportement de l'échantillon) de commettre une erreur en rejetant H_0 . Dans cet exemple, après observation de l'échantillon, la probabilité de commettre une erreur en rejetant H_0 s'estime à ,000 !

Cependant, la statistique du Chi-deux a une faiblesse. En effet, il a été prouvé qu'une faible relation peut devenir significative si la taille de l'échantillon est élevée. Donc l'analyste a besoin de relativiser la présence significative d'une relation de dépendance. C'est ce que nous verrons dans la prochaine sous-section.

4.2.2 Quelle est la force du lien ?

Si la statistique du Chi-deux nous amène à admettre la présence d'un lien entre X et Y , l'étape suivante sera de quantifier ce lien. Pour ce faire nous utilisons les statistiques de Phi et de Cramer's V :

$$\Phi = \sqrt{\frac{\chi^2}{n}} ,$$

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

où k est le minimum entre les modalités de X et Y , et où n est la taille de l'échantillon. Ces mesures varient toujours entre 0 et 1, et la statistique Φ n'est valide que pour les tableaux 2×2 ; en fait on remarque que Φ n'est qu'un cas particulier de V (c'est-à-dire $\Phi = V$ lorsque $k = 2$).

Pour obtenir ces statistiques, il faut ajouter les commandes suivantes à celles faites pour obtenir le tableau croisé :

Menu SPSS :

→ Analyse

→ Descriptive Statistics

→ Crosstabs...

Dans le bouton **Statistics...** : Nominal : ✓ Phi and Cramer's V

On obtient alors la sortie 4.4.

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	,282	,000
Nominal	Cramer's V	,282	,000
N of Valid Cases		200	

- a. Not assuming the null hypothesis.
- b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 4.4 – Pour mesurer la force de la relation

Dans la sortie 4.4, on interprète directement la valeur du Phi et du Cramer's V (dans la colonne **Value**) et non pas une *p*-value. Contrairement à la statistique du Chi-deux, ces deux statistiques ne se laissent pas influencer par la taille de l'échantillon. Une bonne interprétation de ces statistiques permet d'effectuer une correction dans « l'emballement » du Chi-deux.

La statistique Phi n'est utilisée que pour les tableaux croisés de dimension 2×2 . Cette statistique varie entre 0 et +1 ; 0 représente l'absence de relation de dépendance tandis que 1 représente une dépendance parfaite. Cette statistique ne fournit aucune information quant à la variable qui influence l'autre, pas plus qu'elle ne fournit d'indication sur le « sens » de la relation.

Comme mentionné précédemment, le Phi n'est qu'un cas particulier du Cramer's V ; on peut donc se contenter d'interpréter le Cramer's V qui est valide dans toutes les situations.

La statistique Cramer's V est applicable aux tableaux croisés de toutes les dimensions. Tout comme Phi, cette statistique varie aussi entre 0 et 1, selon que la relation va de nulle (0) à intense (1). Dans cet exemple, les deux statistiques sont égales à 0,282. Mais quelle interprétation faut-il donner à cet indice ? Pour se faire une idée de l'ampleur de la puissance de la relation, il est utile de consulter le schéma 4.5 dit de Davis.

$0,70 \leq \text{Cramer's V ou Phi} \leq 1$	Relation très forte
$0,50 \leq \text{Cramer's V ou Phi} < 0,70$	Relation forte
$0,30 \leq \text{Cramer's V ou Phi} < 0,50$	Relation intéressante
$0,10 \leq \text{Cramer's V ou Phi} < 0,30$	Relation faible
$0,00 \leq \text{Cramer's V ou Phi} < 0,10$	Relation négligeable

FIG. 4.5 – Schéma de Davis : interprétation du Cramer's V et du Phi

Dans le cadre de cet exemple, le Cramer's V a une valeur de 0,282 ; il semble donc que la relation de dépendance soit plus faible que ne le laissait croire la statistique du Chi-deux avec sa p -value associée. En d'autres termes, la relation demeure significative au niveau de la population, mais faible. Ainsi, l'analyste devra soit nuancer son interprétation ou bien rechercher de nouvelles relations dans son étude, en espérant en trouver de meilleures.

Les indices de force de relation sont aussi utiles pour faire des comparaisons entre différentes analyses en tableaux croisés. À titre d'illustration, on pourrait dire ici que la variable revenu (brisée en variable discrète) de l'individu a une influence plus forte que la variable du sexe sur le niveau de fidélité de la clientèle si la relation de dépendance **revenu \Rightarrow niveau de fidélité** admet un Cramer's V de 0,45.

4.2.3 Interprétation du tableau croisé

Lorsque nous admettons qu'il y a un lien entre les variables, on peut procéder à l'analyse du tableau croisé pour voir comment s'exprime cette dépendance. Dans le cadre de l'exemple à propos du niveau de fidélité par rapport à une marque préférée de barre de chocolat, le tableau croisé était le suivant :

			SEXÉ		Total	
			Feminin	Masculin		
fidèle à la marque?	Très fidèle	Count	18	58	76	
		Expected Count	30,4	45,6	76,0	
		% within SEXE	22,5%	48,3%	38,0%	
		Std. Residual	-2,2	1,8		
	Moyennement fidèle	Count	18	26	44	
		Expected Count	17,6	26,4	44,0	
		% within SEXE	22,5%	21,7%	22,0%	
		Std. Residual	,1	-,1		
	Pas fidèle	Count	44	36	80	
		Expected Count	32,0	48,0	80,0	
		% within SEXE	55,0%	30,0%	40,0%	
		Std. Residual	2,1	-1,7		
Total		Count	80	120	200	
		Expected Count	80,0	120,0	200,0	
		% within SEXE	100,0%	100,0%	100,0%	

FIG. 4.6 – Le tableau illustrant le croisement sexe \Rightarrow niveau de fidélité

L'interprétation du tableau se fera en deux étapes : tout d'abord une **analyse descriptive** à l'aide des pourcentages, puis une **analyse pour mettre en évidence la relation entre les deux variables** (à l'aide des résidus standardisés).

4.2.1 Analyse descriptive du tableau

L'analyse descriptive du tableau se fait simplement à l'aide des % en colonne. Cette analyse permet de se faire une première idée de l'expression de la relation unissant les deux variables. Dans le cadre de l'exemple, on obtient ceci :

« 22,5 % des femmes se disent très fidèles à une marque de tablette de chocolat, alors que 48,3 % des hommes se retrouvent dans cette catégorie. Pour les moyennement fidèles les proportions sont semblables : 22,5 % pour les femmes et 21,7 % pour les hommes. Finalement, 55 % des femmes se disent pas fidèle, alors que 30 % des hommes se retrouvent dans cette catégorie. »

4.2.2 Les nuances de la relation

En général, la relation entre les deux variables ne se manifeste pas dans toutes les cellules du tableau croisé. En effet, pour certaines cellules les fréquences observées (**Count**) peuvent être semblables aux fréquences théoriques (**Expected count**) ; ces cellules ne permettent donc pas de comprendre de quelle façon le lien entre les deux variables se manifeste. Dans le cadre de l'exemple, dans la catégorie moyennement fidèle, on retrouve à peu près la même proportion de femmes que d'hommes, et on remarque que les fréquences observées sont près des fréquences théoriques. Analyser ces deux cellules nous informe peu sur la relation ; est-ce les femmes ou les hommes qui sont les plus fidèles à une marque de tablette de chocolat ? Il faudra analyser les autres cellules pour répondre à cette question.

Pour le praticien, il est possible de caricaturer la relation à l'aide des résidus standardisés ; c'est à l'aide de ceux-ci que nous saurons dans quelles cellules se manifeste le plus la relation, et comment l'interpréter. Il est possible que ces résidus apportent une dimension légèrement différente de celle produite par l'utilisation des %. Cependant, il est possible que les résidus n'apportent rien de nouveau.

Les résidus standardisés sont tout simplement des cotes-*t* qui mesurent, en nombre d'écart-types, la différence entre les fréquences observées (**Count**) et les fréquences théoriques (**Expected Count**). Pour chaque cellule, le résidu standardisé se calcule de la façon suivante :

$$\text{Résidu standardisé} = \frac{\text{Count} - \text{Exp. Count}}{\sqrt{\frac{n_i \cdot n_j}{n}}}$$

où n est la taille de l'échantillon, n_i la fréquence associée à la i^e modalité de X , et n_j la fréquence associée à la j^e modalité de Y . L'idée de base justifiant la comparaison entre les fréquences observées (Count) et les fréquences théoriques (Expected Count) provient du fait que si H_0 est vraie, alors les fréquences observées seront près des fréquences théoriques et le résidu s'approchera alors de 0. Lorsque les valeurs observées sont plus grandes que les valeurs théoriques, on obtient des résidus positifs, et vice-versa. Cette graduation en nombre d'écart-types aidera le praticien à interpréter le lien de dépendance, si dépendance il y a bien entendu. Le tableau de la figure 4.7 nous montre comment interpréter les résidus.

Std. Residual < -3	Absence marquée du phénomène
-3 ≤ Std. Residual < -2	Absence significative du phénomène
-2 ≤ Std. Residual < -1,5	Absence visible (tendance) du phénomène
-1,5 ≤ Std. Residual ≤ 1,5	Phénomène normal.
1,5 < Std. Residual ≤ 2	Présence visible (tendance) du phénomène
2 < Std. Residual ≤ 3	Présence significative du phénomène
3 < Std. Residual	Présence marquée du phénomène

FIG. 4.7 – Schéma d'interprétation des résidus standardisés

Voici donc l'interprétation que l'on peut faire du tableau 4.6 à l'aide des résidus et des % en colonne :

« Les femmes évitent significativement d'être fidèles à une marque de tablette de chocolat, plus précisément elles diversifient significativement leur achats. Les hommes ont plus tendance à être fidèles à la tablette de chocolat qu'ils préfèrent. Même qu'ils ont tendance à éviter de changer de sorte. »

4.3 Le croisement de deux variables ordinaires

Que le tableau croisé soit formé à partir de variables nominales ou ordinaires, l'analyse présentée à la section précédente est toujours de mise. Lorsque les deux variables incluses dans un tableau croisé sont mesurées à l'aide d'échelles ordinaires, il est alors possible d'exploiter l'information supplémentaire contenue dans ce type de mesure en calculant la statistique Gamma.

Il existe d'autres statistiques qui poursuivent les mêmes objectifs : Kendall's Tau b, Kendall's Tau c, Somer's d, Spearman rho et le coefficient de corrélation de Spearman. La plupart de ces statistiques prennent les valeurs comprises entre -1 et +1, selon que la relation va de parfaitement négative (-1) à parfaitement positive (+1). Ces statistiques permettent donc de mesurer le sens de la relation.

La statistique Gamma mesure le sens du lien (par son signe), et elle mesure aussi l'apport d'informations supplémentaires fournies par la connaissance d'une première variable, en quantifiant les chances de prédire correctement la seconde. De fait, il est raisonnable de croire que si deux variables sont en grande relation de dépendance, la connaissance de la valeur précise obtenue sur l'une des deux variables chez un individu amènera suffisamment d'informations pour prédire la valeur de la seconde variable. Les statistiques de ce type portent le nom de PRE (Proportional Reduction Error).

Exemple 4.3.1 Considérons maintenant l'exemple d'une étude de dépendance entre le niveau d'alcoolisme (1 = non buveur, 2 = buveur modéré et 3 = buveur invétérée) et le niveau de tabagisme (1 = non fumeur, 2 = fumeur modéré et 3 = fumeur invétérée) des individus d'une population. Fixons le seuil de signification α à 0,05. Pour obtenir les sorties SPSS 4.8, 4.9 et 4.10, il faut procéder aux opérations suivantes :

Menu SPSS :	→ Analyse
	→ Descriptive Statistics
	→ Crosstabs...
Dans la fenêtre Row(s) :	→ fumeur
Dans la fenêtre Column(s) :	→ buveur
Dans le bouton Cells... :	→ Percentages ✓ Column → Count ✓ Observed ✓ Expected → Residuals ✓ Standardized
Dans le bouton Statistics... :	✓ Chi-square Nominal : ✓ Phi and Cramer's V Ordinal : ✓ Gamma

Niveau de fumeur * Niveau de buveur Crosstabulation				
			Niveau de buveur	
			non buveur	buveur modéré
Niveau de fumeur	non fumeur	Count	175	159
		Expected Count	139,4	140,8
		% within Niveau de buveur	85,4%	76,8%
		Std. Residual	3,0	1,5
fumeur modéré		Count	20	8
		Expected Count	36,9	37,3
		% within Niveau de buveur	9,8%	3,9%
		Std. Residual	-2,8	-4,8
fumeur invétérée		Count	10	40
		Expected Count	28,7	29,0
		% within Niveau de buveur	4,9%	19,3%
		Std. Residual	-3,5	2,0
Total		Count	205	207
		Expected Count	205,0	207,0
		% within Niveau de buveur	100,0%	100,0%
		Total	340	340,0
			68,0%	
			18,0%	
			14,0%	
			100,0%	

FIG. 4.8 – Le tableau illustrant le croisement buveur ⇒ fumeur

Vérifions dans un premier temps s'il existe un lien de dépendance (ou d'interdépendance) significatif entre les deux variables au niveau de la population. Pour ce faire, il

Chi-Square Tests			
	Value	df	A symp. Sig. (2-sided)
Pearson Chi-Square	246,294 ^a	4	,000
Likelihood Ratio	234,923	4	,000
Linear-by-Linear Association	95,383	1	,000
N of Valid Cases	500		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12,32.

FIG. 4.9 – Le test d’indépendance avec le Chi-deux χ^2

faut traiter le test d’hypothèses suivant :

H_0 : Dans la population, la consommation d’alcool est indépendante de la consommation de cigarettes.

H_1 : Dans la population, la consommation d’alcool est liée à la consommation de cigarettes.

Le test est résolu à l’aide de la règle de décision basée sur la valeur de la p -value associée à la statistique du Chi-deux (Pearson Chi-Square) de la sortie 4.9. La règle de décision est la suivante :

On rejette H_0 si le Pearson Chi-square admet une p -value (Asymp. Sig. (2-sided)) plus petite que le seuil $\alpha = 0,05$. Si non, on ne rejette pas H_0 .

Compte tenu que la p -value associée au Chi-deux a une valeur de 0,000, ce qui est plus petit que le seuil fixé $\alpha = 0,05$, nous rejetons l’hypothèse d’indépendance H_0 et nous admettons, au risque de nous tromper 1 fois sur 20, que dans la population, l’hypothèse du lien entre les deux variables est vraisemblable.

N’oublions pas qu’étant donné que la statistique du Chi-deux est influencée par la taille de l’échantillon, nous devons vérifier la force du lien. Comme nous sommes en

Symmetric Measures				
	Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Nominal by Nominal Phi	,702			,000
Nominal Cramer's V	,496			,000
Ordinal by Ordinal Gamma	,631	,043	11,617	,000
N of Valid Cases	500			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 4.10 – Mesure de la puissance du lien avec le Cramer's V et Gamma

présence du croisement de deux variables ordinaires, nous utilisons la statistique Cramer's V **et** la statistique Gamma de la sortie 4.10.

Tout d'abord, d'après le schéma d'interprétation 4.5 de Davis, le Cramer's V nous indique que la relation est intéressante (il a une valeur de 0,496).

La statistique Gamma a une valeur de 0,631. En voici la signification : la connaissance de la valeur de la variable **buveur** chez un individu améliore nos prédictions sur la valeur de la variable **fumeur** de 63,1 %, ce qui est beaucoup mieux que la simple chance (prendre au hasard l'une des modalités de la variable dépendante ne donne qu'une probabilité de 33,3 % d'avoir raison).

Cette même statistique nous informe aussi sur la direction de la relation. L'ordre des codes affectés aux variables est important puisque que la direction indiquée par le signe de la statistique Gamma repose sur ce code. Dans cet exemple, en ce qui concerne la variable **buveur**, plus le code est élevé, et plus le répondant boit. Il en va de même pour la variable **fumeur**. Une valeur de Gamma positive indique que plus la première variable est élevée, plus la seconde l'est. Inversement, un Gamma négatif indique que plus la première variable est élevée, moins la seconde l'est. Dans notre exemple, puisque la statistique Gamma est positive (0,631), elle nous indique que plus un individu boit, plus il fume et vice-versa.

Après l'observation de l'ensemble de ces statistiques, l'analyste peut compléter son

interprétation.

« Au risque de se tromper 1 fois sur 20, il existe un lien dans la population entre la façon de boire et la façon de fumer. Ce lien peut être qualifié d'intéressant, et dans cette population, plus un individu boit, plus il fume. On observe que 85,4 % des non-buveurs sont des non-fumeurs, 9,8 % sont des fumeurs modérés et 4,9 % sont des fumeurs invétérés. Pour les buveurs modérés, 76,8 % sont non-fumeurs, 3,9 % sont des fumeurs modérés et 19,3 % sont des fumeurs invétérés. Finalement, 6,8 % des buveurs invétérés sont non-fumeurs, 70,5 % sont des fumeurs modérés et 22,7 % sont des fumeurs invétérés.

Plus précisément, les non-buveurs sont significativement des non-fumeurs. Les buveurs modérés ne sont pas des fumeurs modérés, et ce de façon marquée, et ils ont significativement plus tendance à fumer abondamment. Finalement, les buveurs invétérés fument assurément (ce sont de façon marquée des fumeurs modérés et de façon significative des fumeurs invétérés). »

4.4 Pré-requis : fréquences théoriques

Lors de la conception du questionnaire, l'analyste planifie aussi la taille d'échantillon nécessaire pour que toutes ses analyses se déroulent bien. En effet, il n'est pas toujours nécessaire d'avoir 400 répondants pour que tout fonctionne.

Pour que l'analyse d'un tableau croisé soit valide, une règle du pouce stipule qu'il faut que les **Expected Count** soient tous supérieurs ou égaux à 5. Dans la pratique, l'analyste tolère que jusqu'à 20 % des **Expected Count** soient entre 1 et 5. Lorsque ce 20 % est dépassé, l'analyse ne peut être valide. Une alternative possible (s'il n'est pas possible d'augmenter la taille de l'échantillon) est de recoder au moins l'une des deux variables afin de diminuer le nombre de modalités, et ainsi diminuer le nombre de cases du tableau. Il est aussi utile de remarquer qu'il est toujours indiqué en-dessous du tableau contenant

le Chi-deux (voir par exemple la figure 4.9) quel est le pourcentage de cellules qui ont un **Expected Count** en bas de 5.

			SEXÉ		Total	
			Féminin	Masculin		
fidèle à la marque?	Très fidèle	Count	18	58	76	
		Expected Count	30,4	45,6	76,0	
		% within SEXE	22,5%	48,3%	38,0%	
		Std. Residual	-2,2	1,8		
	Moyennement fidèle	Count	18	26	44	
		Expected Count	17,6	26,4	44,0	
		% within SEXE	22,5%	21,7%	22,0%	
		Std. Residual	,1	-,1		
	Pas fidèle	Count	44	36	80	
		Expected Count	32,0	48,0	80,0	
		% within SEXE	55,0%	30,0%	40,0%	
		Std. Residual	2,1	-1,7		
Total		Count	80	120	200	
		Expected Count	80,0	120,0	200,0	
		% within SEXE	100,0%	100,0%	100,0%	

FIG. 4.11 – Le tableau illustrant le croisement sexe \Rightarrow niveau de fidélité

La sortie 4.11 est le tableau de la relation sexe \Rightarrow niveau de fidélité. Ce tableau contient 6 valeurs **Expected Count** et elles sont toutes supérieures à 5. L'analyse basée sur ces nombres est donc valide.

Au moment d'entreprendre une étude, on peut avoir une idée de la taille d'échantillon minimale requise pour étudier une relation à partir d'un tableau croisé en faisant un simple calcul qui vise à s'assurer que les **Expected Count** soient tous supérieurs à 5.

La formule est la suivante :

$$n_{\min} = [(\# \text{ de modalités de la variable } X) \times (\# \text{ de modalités de la variable } Y)] \times 5 \times 2$$

Le nombre 2 provient d'une règle du pouce, c'est une sorte de minimum. Dans le cadre de l'exemple de cette section, pour planifier une taille d'échantillon idéale assurant la validité d'un tel tableau croisé, l'analyste devait planifier un minimum de 60 unités

(tant mieux s'il y en a plus) :

$$n_{\min} = 2 \times 3 \times 5 \times 2 = 60.$$

4.5 Un autre exemple

On s'intéresse ici à l'influence de la publicité sur les ventes pour une entreprise donnée. La variable `pub` a trois niveau : bas, moyen et élevé, tandis que la variable `ventes` a deux niveaux : bas et élevé. La base de données se nomme `pub.sav`.

On remarque que les variables `pub` et `ventes` sont ordinales. L'exemple présenté ici a pour but d'illustrer que le lien entre deux variables ordinaires n'est pas toujours linéaire, et qu'alors l'apport d'information de la statistique gamma est négligeable.

Mais faisons d'abord l'analyse complète de cette relation. Fixons les seuils de signification à $\alpha = 0,05$. Vérifions dans un premier temps s'il existe un lien de dépendance significatif entre les deux variables au niveau de la population. Pour ce faire, il faut traiter le test d'hypothèses suivant :

H_0 : Dans la population, le niveau de publicité est indépendant du niveau des ventes.

H_1 : Dans la population, le niveau de publicité est lié au niveau des ventes.

On peut utiliser le test du chi-deux sans problème puisqu'il n'y a pas de cellules ayant une fréquence théorique (`Expected Count`) inférieure à 5 (voir la sortie 4.12 ou 4.14). On voit dans la sortie 4.12 que la *p*-value associée au chi-deux est de 0,000 ; puisque $0 < 0,05$, on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on peut affirmer que le niveau de publicité est lié au niveau des ventes.

Quelle est la force de ce lien ? D'après le Cramer's V qui a une valeur de 0,707 (sortie 4.13), on peut dire que le lien est très fort. Par contre la valeur de Gamma étant nulle

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	30,000 ^a	2	,000
Likelihood Ratio	38,191	2	,000
Linear-by-Linear Association	,000	1	1,000
N of Valid Cases	60		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10,00.

FIG. 4.12 – Le test d’indépendance avec le Chi-deux χ^2

Symmetric Measures					
		Value	Asymp. Std. Error ^b	Approx. T ^b	Approx. Sig.
Nominal by Nominal	Phi	,707			,000
Nominal	Cramer's V	,707			,000
Ordinal by Ordinal	Gamma	,000	,193	,000	1,000
N of Valid Cases		60			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 4.13 – Mesure de la puissance du lien

(sortie 4.13), on ne peut parler du sens du lien, car en fait celui-ci n'est pas linéaire. En effet, on observe qu'en passant de niveau bas à élevé pour la publicité, les ventes ont tendance à passer du niveau bas à élevé, mais cette tendance est renversée lorsque l'on passe au niveau élevé de publicité. Comme quoi des tonnes de copies, ça écoûte ;-)

On peut maintenant analyser le tableau croisé. On voit d'abord qu'au niveau bas pour la publicité, 75 % des ventes sont basses, tandis qu'au niveau moyen cette proportion tombe à 0 %, pour ensuite remonter à 75 % au niveau élevé.

Ainsi, un niveau bas de publicité a tendance à amener un niveau bas de ventes. D'autre part, un niveau moyen de publicité amène de façon marquée un niveau élevé de ventes. Finalement, un niveau élevé de publicité semble avoir le même effet qu'un niveau bas de publicité puisqu'il a tendance à amener un niveau bas de ventes.

			Niveau de publicité.			Total	
			Bas	Moyen	Élevé		
Niveau des ventes.	Bas	Count	15	0	15	30	
		Expected Count	10,0	10,0	10,0	30,0	
		% within Niveau de publicité.	75,0%	,0%	75,0%	50,0%	
		Std. Residual	1,6	-3,2	1,6		
	Élevé	Count	5	20	5	30	
		Expected Count	10,0	10,0	10,0	30,0	
		% within Niveau de publicité.	25,0%	100,0%	25,0%	50,0%	
		Std. Residual	-1,6	3,2	-1,6		
Total		Count	20	20	20	60	
		Expected Count	20,0	20,0	20,0	60,0	
		% within Niveau de publicité.	100,0%	100,0%	100,0%	100,0%	

FIG. 4.14 – Le tableau illustrant le croisement pub \Rightarrow ventes

4.6 L'analyse des correspondances

L'analyse des correspondances permet de visualiser les tableaux croisés à deux variables discrètes par l'entremise d'une carte. Cette carte porte le nom de carte perceptive. Elle est très utile pour effectuer les analyses de positionnement.

L'analyse en correspondances est utile lorsque les deux variables ont beaucoup de modalités (habituellement plus de quatre). En effet, lorsqu'il y a beaucoup de modalités, il devient difficile de bien cerner la relation avec un tableau croisé. De plus, la condition sur les fréquences théoriques n'est pas toujours respectée lorsque le tableau est grand ; l'analyse en correspondances possède l'avantage de ne demander aucun pré-requis. Cependant, ce n'est qu'une analyse descriptive.

L'analyse des correspondances positionne sur une même carte les modalités des deux variables à l'étude. Pour obtenir une analyse des correspondances, il faut d'abord créer une nouvelle base de données à partir de celle qui contient les deux variables à l'étude. Les commandes sont les suivantes :

Menu SPSS :	→ Data
	→ Aggregate...
Dans la fenêtre Break Variables(s) :	→ Mettre les deux variables à l'étude
Dans la fenêtre Aggregated Variables :	<input checked="" type="checkbox"/> Number of cases
Dans la fenêtre Save :	<input checked="" type="checkbox"/> Create new data file containing aggregated variables only

(Puis spécifier l'emplacement et le nom du nouveau fichier dans le bouton File)

Une nouvelle base de données est alors créée. Cette nouvelle base contiendra toutes les combinaisons possibles des modalités des deux variables (ceci correspond aux cellules du tableau croisé), et une nouvelle variable qui donne la fréquence observée pour chacune des combinaisons (donc le Count de chacune des cellules).

À partir de cette nouvelle base de données, il faut effectuer les commandes suivantes avant de procéder à l'analyse en correspondances :

Menu SPSS :	→ Data
	→ Weight Cases...
	<input checked="" type="checkbox"/> Weight cases by
	→ Mettre la variable qui indique le Count dans la fenêtre
	Frequency Variable (par défaut cette variable se nomme N_BREAK)

Il est alors possible de faire l'analyse en correspondances avec les commandes suivantes :

Menu SPSS :	→ Analyse
	→ Data Reduction
	→ Correspondence Analysis...

Dans la fenêtre Row : → Nom_var1 (nom d'une des deux variables)

Dans le bouton `Define Range...` : → Indiquer les valeurs minimum et maximum des modalités.

→ Cliquer sur `Update`.

Dans la fenêtre `Column` : → `Nom_var2` (nom de l'autre variable)

Dans le bouton `Define Range...` : → Indiquer les valeurs minimum et maximum des modalités.

→ Cliquer sur `Update`.

Pour interpréter une carte perceptuelle, il suffit de comprendre le principe des aimants : plus les modalités s'éloignent du centre et se regroupent, plus leurs associations entre elles sont présentes. Les éléments qui restent au centre sont soit « repoussés » par tous (ne sont pas clairement associés à quelque chose), soit « attirés » par tous.

Exemple 4.6.1 Afin de bien comprendre le principe de l'analyse en correspondances, nous commencerons avec un exemple simple ; en fait tellement simple qu'habituellement on se serait contenté d'un tableau croisé puisque les deux variables n'ont que trois modalités chacune. Nous reprenons l'exemple 4.3.1 : la relation entre le fait de boire et de fumer. La carte perceptuelle issue de cette relation se retrouve dans la figure 4.15.

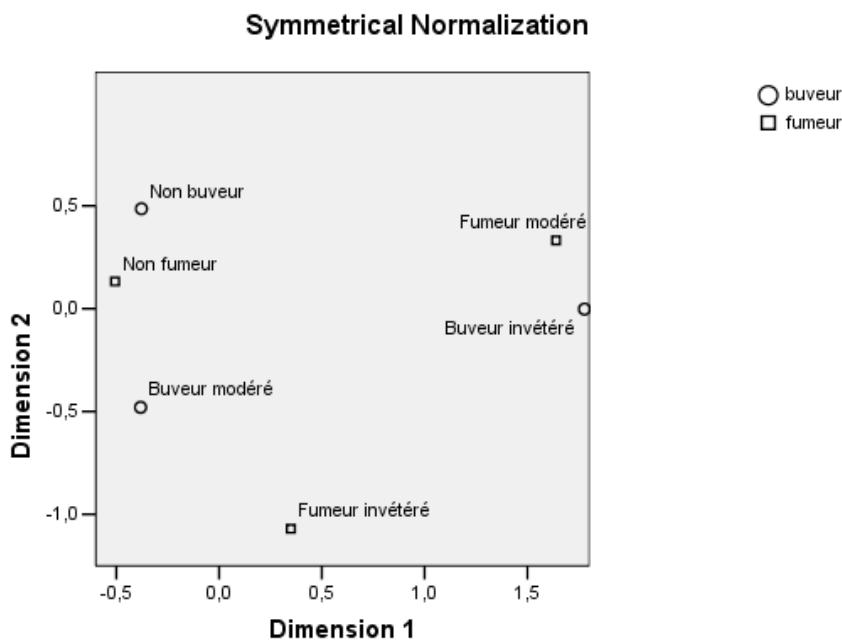


FIG. 4.15 – La carte perceptuelle de la relation entre le fait de boire et de fumer

On constate que non-fumeur est près de non-buveur, et que fumeur modéré est près de buveur invétérée. Cette carte ne fait que confirmer ce qu'on avait déduit du tableau croisé de cette relation.

Exemple 4.6.2 Voici maintenant un exemple où il est vraiment pertinent de considérer une carte perceptuelle. Cette carte perceptuelle (figure 4.16) illustre le lien qui existe entre les activités recherchées par les touristes et la région touristique, au nombre de quatre, à laquelle les gens associent les activités.

En utilisant le principe des aimants, voici une interprétation possible de la carte perceptuelle : « Il est possible de voir que la région 1 est davantage une destination de divertissement. Les activités associées à cette région se déroulent sur des territoires plus compacts. La région 2 est une destination de contemplation. Les activités ne sont pas très intenses et se déroulent sur des territoires moyennement grands. La région 3 est une destination où les activités sont plus intenses. Les activités pratiquées nécessitent un grand territoire pour se dérouler. La région 4 se retrouve au centre et n'est donc pas une

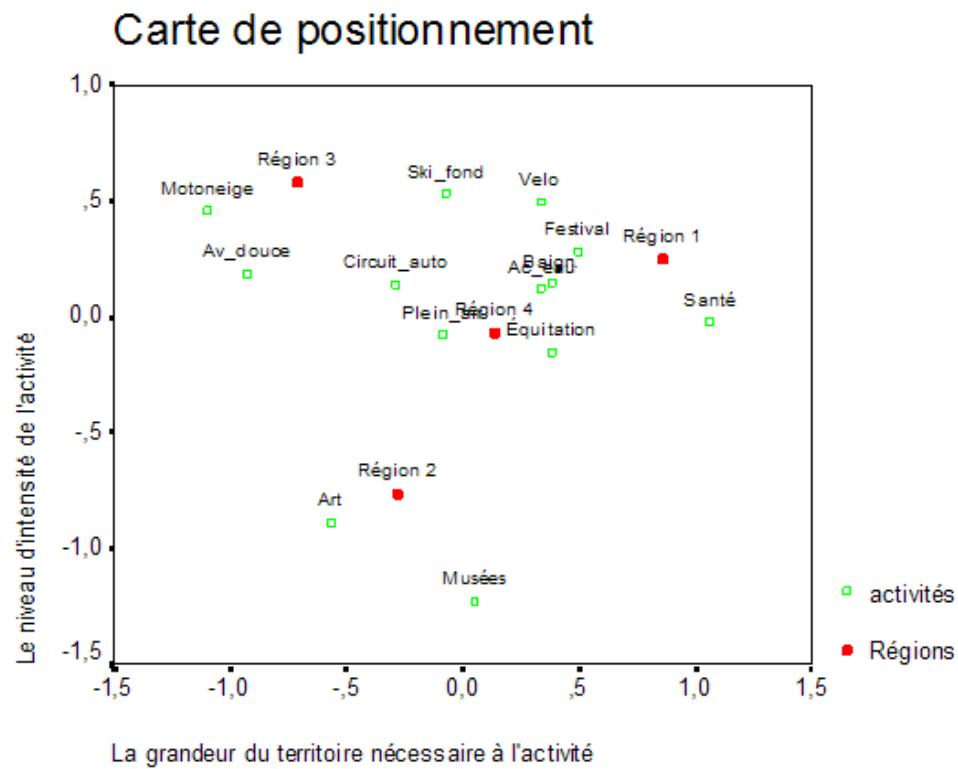


FIG. 4.16 – Liens entre activités touristiques et régions

destination associée à une activité étudiée. »

Dans une telle carte perceptuelle, les deux axes de référence sont nommés par l'analyste. Pour y arriver, il lui faut regarder les modalités qui s'opposent d'un coté à l'autre de l'axe à nommer. Tout est subjectif dans le choix du nom des axes. Deux analystes peuvent arriver à des noms d'axes bien différents. Dans ce type d'analyse spécialisée l'expérience dans le domaine est pratiquement irremplaçable.

Exemple 4.6.3 Voyons maintenant un exemple intéressant et rare d'une carte perceptuelle circulaire. Est-ce que l'âge a un impact sur le choix des activités des touristes ? La carte perceptuelle de la figure 4.17 illustre que oui. En effet, l'évolution de l'âge (brisée en variable discrète ordinaire) se traduit par un changement dans les priorités et donc dans le choix des activités. On voit aussi que certaines activités ou attraits (golf, festival, activités aquatiques, nature, marche pédestre, gastronomie) sont considérés ou pratiqués par toutes les catégories d'âge.

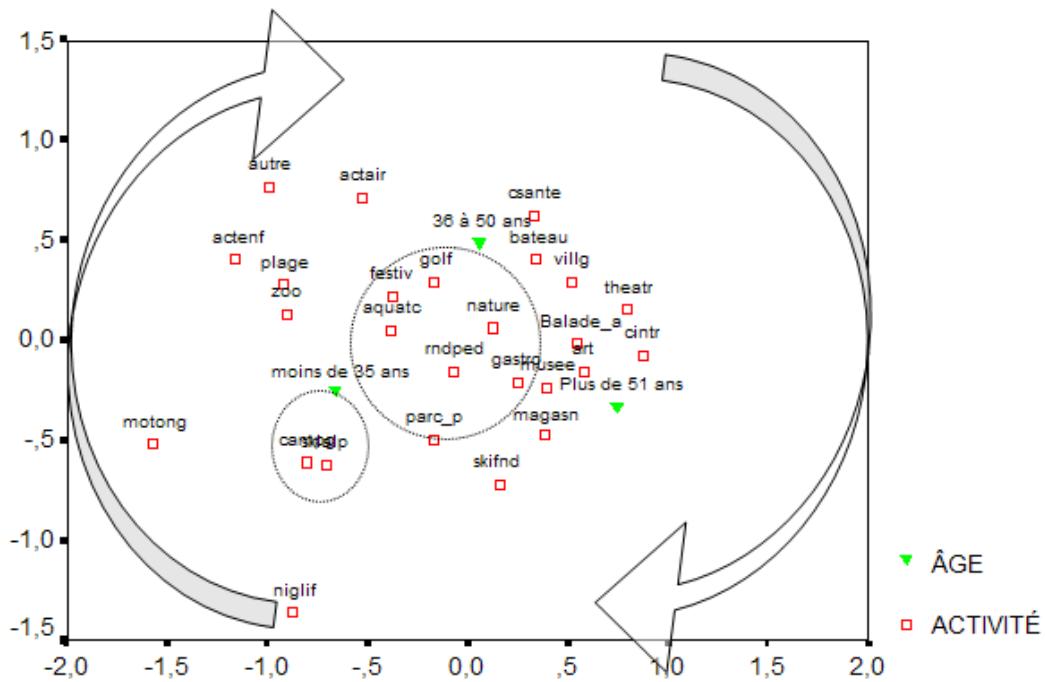


FIG. 4.17 – Liens entre l'âge et les activités touristiques

En interprétant la carte en commençant par les moins de 35 ans et en tournant dans le sens horaire, on voit que les âges se suivent dans l'ordre. Donc en « déroulant » les âges et les activités, on retrouve le schéma de la figure 4.18. Il est intéressant de noter que finalement, le choix des activités semble dépendre des besoins de la famille (jeunes familles, enfants plus vieux puis finalement plus d'enfants à charge).

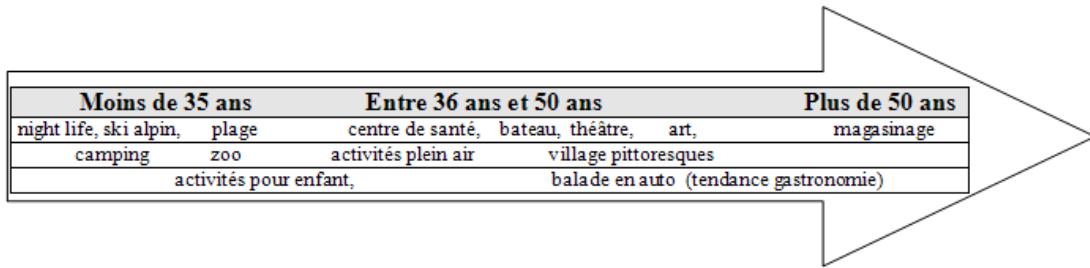


FIG. 4.18 –

4.7 Exercices du chapitre

Exercice 1 Vous êtes responsable des ressources humaines. Vous devez évaluer l'efficacité d'une nouvelle méthode de travail qui a été implantée dans les départements d'administration et de production. Cette nouvelle méthode doit théoriquement améliorer le niveau de satisfaction des employés face à leur travail.

L'implantation s'est effectuée il y a près de dix mois, temps où le niveau de satisfaction était « moyen » dans chacun des deux départements. Depuis l'implantation, un consultant s'occupe personnellement de l'administration globale de la méthode.

Vous devez maintenant prendre une décision sur la pertinence de ces plans, et ce, pour chacun des départements. En effet, selon votre pressentiment, il est possible que la méthode ne fonctionne bien que pour un des départements. Afin de ne pas arrêter la production, une brève étude de la satisfaction a été soumise à un échantillon de 60 employés choisis au hasard.

Faites les analyses appropriées à partir de la base de données `nouvmethode.sav`.

Exercice 2 Utilisez la base de données `satisfactiontravail.sav` pour cet exercice. On se demande si au sein de cet entreprise, il y a un lien entre l'ancienneté à l'emploi et le plus haut niveau d'étude atteint. Pour étudier ce lien, recoder la variable `anciennt` en 4 classes, et la variable `études` en 3 classes (faites attention à la codification : ce serait bien d'utiliser le fait que ces variables peuvent être ordinaires si bien codifiées).

Exercice 3 Une étude est menée sur l'effet de la confiance et du risque perçu sur l'adoption des services bancaires en ligne. Un questionnaire est administré auprès d'un échantillon aléatoire de Québécois. Voici quelques unes des questions qui faisaient partie de ce questionnaire (après chaque question on retrouve le nom de la variable qui lui correspond dans la base de données `servicebancaire.sav`) :

1. Quel est votre âge ? (**âge**)

Moins de 25 ans 1

Entre 25 et 35 ans 2

Entre 35 et 45 ans 3

Plus de 45 ans 4

2. Le service bancaire en ligne est facile à utiliser. (**facile**)

Tout à fait 1 2 3 4 5 6 7 Tout à fait
en désaccord en accord

3. Le service bancaire en ligne permet d'effectuer plus rapidement mes transactions.
(**utile**)

Tout à fait 1 2 3 4 5 6 7 Tout à fait
en désaccord en accord

4. Comment caractérissez-vous la décision d'utiliser le service bancaire en ligne ? (**risque**)

Peu risquée 1

Moyennement risquée 2

Très risquée 3

5. J'ai confiance dans le service bancaire en ligne. (**confiance**)

Tout à fait 1 2 3 4 5 6 7 Tout à fait
en désaccord en accord

6. Combien de fois par mois utilisez-vous le service bancaire en ligne ? _____
(**adoption**)

On se demande si l'âge a une influence sur le risque perçu. Faites l'analyse pertinente pour répondre à cette question ; fixez le seuil à $\alpha = 0,05$.

Chapitre 5

Relation entre une variable discrète et une continue

Peu importe quelle variable est dépendante, l'analyse des liens unissant une variable discrète à une variable continue se ramène généralement à un problème de comparaison de moyennes. Par exemple, supposons que nous voulons savoir si les cadres des provinces du Québec et de l'Ontario (variable discrète) sont rémunérés (variable continue) de la même façon. Le modèle de la relation prendrait la forme suivante :

Province \Rightarrow Salaire des cadres d'un certain type d'entreprise

Nous conclurions à l'influence de la province si le salaire moyen des cadres de la province du Québec était différent du salaire moyen de l'autre province. Dans une autre étude, nous pourrions tout aussi bien étudier la relation entre le salaire (variable continue) des gens et la possession d'une voiture de luxe (oui ou non, variable discrète). Le modèle de la relation à étudier prendrait alors la forme suivante :

Salaire \Rightarrow La possession d'une voiture de luxe

Nous conclurons que le salaire a de l'influence si les gens qui possèdent une automobile de luxe ont un salaire moyen différent (possiblement plus élevé) de ceux qui n'en ont pas. Voilà pourquoi l'étude de la relation entre une variable discrète et une variable continue se ramène généralement à un problème de comparaison des moyennes.

5.1 Variable dichotomique

Une variable dichotomique est une variable discrète qui ne peut prendre que deux états. Par exemple, le sexe (masculin, féminin), une réponse en oui ou non, etc. L'analyse présentée dans cette section est utilisée lorsque la variable discrète est dichotomique, ou lorsque l'on ne considère que deux modalités de réponse d'une variable discrète (par exemple, si la variable province a comme réponses possibles Alberta, Ontario et Québec et que l'on ne s'intéresse qu'aux deux modalités Ontario et Québec).

Considérons l'exemple basé sur l'exercice du chapitre 2 où nous voulons savoir si, pour des entreprises de même type, les cadres du Québec sont aussi bien rémunérés que les cadres de l'Ontario.

Ici, la variable dichotomique est la **province** (Québec ou Ontario), tandis que la variable continue est la variable **salaire**, en dollars. En fait, nous sommes intéressés à étudier le modèle suivant (en utilisant la base de données de l'exercice du chapitre 2) :

$$\text{Province} \Rightarrow \text{Salaire}$$

Si la province n'a aucune influence sur le salaire, il ne devrait pas y avoir de différence entre les salaires moyens des deux provinces. Cette problématique revient à mettre en cause l'égalité des moyennes de salaire entre les deux provinces. Cette question peut être traitée de façon équivalente en étudiant la différence des salaires moyens des deux provinces. Plus précisément, s'il est vrai que la province n'a aucune influence sur les salaires, alors la différence des moyennes devrait être nulle. Ceci revient à dire qu'il est équivalent d'écrire les tests d'hypothèses suivants :

$$\begin{array}{ll} H_0 : \mu_{\text{Québec}} = \mu_{\text{Ontario}} & \text{ou} \\ H_1 : \mu_{\text{Québec}} \neq \mu_{\text{Ontario}} & H_0 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} = 0 \\ & H_0 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} \neq 0 \end{array}$$

Il faudra résoudre ce test pour pouvoir affirmer qu'il y a un lien ou non entre les deux variables. Mais avant de résoudre le test, une analyse descriptive est de mise. Pour obtenir les sorties 5.1 et 5.2, il faut effectuer les commandes suivantes :

Menu SPSS :	→ Analyse
	→ Descriptive Statistics
	→ Explore...
Dans la fenêtre Dependent List :	→ salaire (la variable continue)
Dans la fenêtre Factor List :	→ province (la variable discrète)
Display :	→ Both
Dans le bouton Plots... :	désactivez Stem-and-leaf
Dans le bouton Statistics... :	✓ Descriptives
Confidence Interval for Mean :	95 % (niveau de confiance)

Le tableau de la sortie 5.1 permet de comparer simultanément les statistiques des deux provinces (on oublie simplement ce qui concerne l'Alberta). Premièrement, par leur CV, les moyennes présentées sont représentatives.

La différence entre les moyennes de salaire s'estime ponctuellement à 15 594,80 \$ (60 795 \$ - 45 200,20 \$). Plus précisément, le salaire moyen des cadres québécois est compris entre 41 672,39 \$ et 48 728,01 \$, et ce, 19 fois sur 20, tandis que le salaire moyen des cadres ontariens est compris entre 56 074,47 \$ et 65 515,53 \$, et ce, 19 fois sur 20.

L'analyste remarque aussi que les intervalles de confiance ne se chevauchent pas, ce qui laisse présager une différence significative entre les moyennes. Cette hypothèse reste cependant à valider avec un test d'hypothèses formel.

		Descriptives					
		Province					
		Alberta		Ontario		Québec	
		Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
salaire	Mean	40612,04	2439,103	60795,00	2086,737	45200,20	1559,487
	95% Confidence Interval for Mean	Lower Bound	35094,40		56074,47		41672,39
		Upper Bound	48129,67		65515,53		48728,01
	5% Trimmed Mean		40521,71		60711,11		45012,44
	Median		40614,93		61600,00		45918,50
	Variance		5,9E+07		4,4E+07		2,4E+07
	Std. Deviation		7713,121		6598,840		4931,532
	Minimum		29850,00		49700,00		38800,00
	Maximum		53200,00		73400,00		54980,00
	Range		23550,00		23700,00		16180,00
	Interquartile Range		13599,61		8550,00		7176,25
	Skewness		,120	,687	,200	,687	,590
	Kurtosis		-1,178	1,334	,694	1,334	,234
							1,334

FIG. 5.1 – Statistiques descriptives

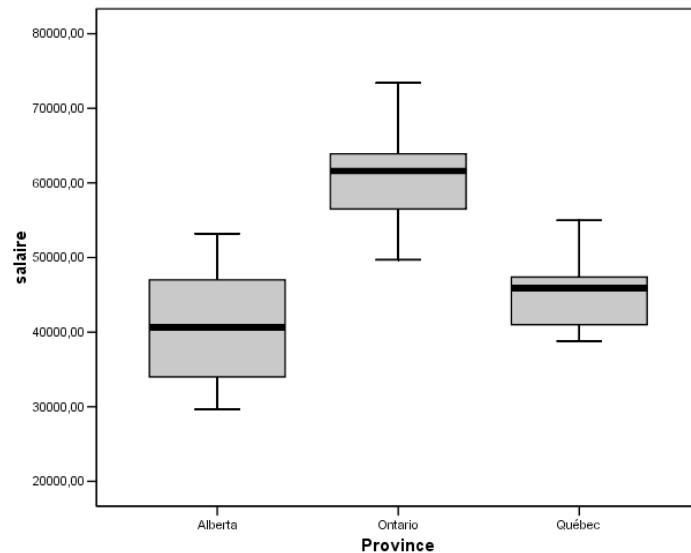


FIG. 5.2 – Visualisation de l'influence des provinces sur les salaires

La figure 5.2 permet de visualiser cette différence. Ainsi les statistiques et le graphique suggèrent l'hypothèse suivant laquelle les salaires des cadres varient en fonction de la province. Avant de faire le test d'hypothèses nous permettant de valider ou non cette affirmation, voyons quel est le pré-requis pour pouvoir faire ce test.

5.1.1 Pré-requis

Pour être en mesure d'utiliser la procédure statistique nous permettant de résoudre le test

$$H_0 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} = 0$$

$$H_1 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} \neq 0$$

il faut que **les deux échantillons proviennent de populations normales**. En fait, lorsque les deux échantillons ont une taille supérieure ou égale à 30, on peut se contenter de populations *approximativement* normale.

Dans tous les cas, on doit vérifier si les **deux** populations suivent une loi normale. Pour vérifier cette condition, on peut se servir des statistiques Skewness et Kurtosis de la sortie 5.1 tel qu'indiqué dans le chapitre 3. Il est encore mieux de tester formellement la normalité en traitant le test d'hypothèses suivant pour chacun des groupes :

H_0 : Les données de la population se répartissent selon une loi normale.

H_1 : Les données de la population ne se répartissent pas selon une loi normale.

Dans le cadre de l'exemple, il faut donc vérifier si les populations des salaires des cadres du Québec et de l'Ontario suivent une loi normale. Pour résoudre le test de normalité, il faut d'abord effectuer les commandes suivantes :

Menu SPSS :	→ Analyse
	→ Descriptive Statistics
	→ Explore...
Dans la fenêtre Dependent List :	→ salaire (la variable continue)
Dans la fenêtre Factor List :	→ province (la variable discrète)
Display :	→ Plots
Dans le bouton Plots... :	✓ Normality plots with tests

On obtient alors (entre autres) le tableau 5.3.

Tests of Normality						
Province	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
salaire						
Alberta	,157	10	,200*	,955	10	,730
Ontario	,131	10	,200*	,982	10	,974
Québec	,136	10	,200*	,948	10	,644

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. 5.3 – Vérification de la normalité des données

Ce sont les *p*-values (**Sig.**) associées aux statistiques de Kolmogorov-Smirnov et Shapiro-Wilk qui nous permettent de résoudre le test d'hypothèses. La règle de décision est la suivante :

Nous rejetons l'hypothèse H_0 si la *p*-value est plus petite que le seuil de signification α fixé (par exemple $\alpha = 0,05$). Sinon, nous ne rejetons pas H_0 et la considérons comme vraisemblable.

La littérature ne s'entend pas sur lequel des deux tests est le plus performant. Il faut donc considérer les deux réponses en même temps. Ainsi, lorsque les deux tests fournissent la même conclusion, l'analyste est confiant. Il peut par contre arriver qu'un seul des deux tests rejette la normalité. Étant donné que la procédure statistique que nous verrons est

assez robuste à la violation de la normalité, nous poursuivrons l'analyse même si l'une des deux statistiques rejette la normalité. Par contre, **dans le cas où la normalité est rejetée par les deux statistiques, nous ne poursuivons pas l'analyse.**

Dans notre exemple, pour les provinces de l'Ontario et du Québec, les deux tests soutiennent de ne pas rejeter H_0 car toutes les p -values sont plus grandes que $\alpha = 0,05$ (elles sont égales à 0,200 et 0,974 pour l'Ontario, et à 0,200 et 0,644 pour le Québec). Ainsi, l'hypothèse de normalité des données n'est pas violée, ce qui nous permet de poursuivre l'analyse.

5.1.2 Comparaison des deux moyennes : Independent Samples T

Test

On peut maintenant résoudre le test

$$H_0 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} = 0$$

$$H_1 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} \neq 0$$

à l'aide du **Independent Samples T Test**. La sortie 5.4 permet le traitement complet de ce test d'hypothèses. Pour obtenir cette sortie, il faut effectuer les commandes suivantes :

Menu SPSS :	→ Analyse
	→ Compare Means
	→ Independant-Samples T Test...
Dans la fenêtre Test Variable(s) :	→ salaire (la variable continue)
Dans la fenêtre Grouping Variable :	→ province (la variable discrète)
Dans le bouton Define Groups... :	inscrire le code des provinces étudiées (2 et 3 dans notre cas).

Independent Samples Test										
	Levene's Test for Equality of Variances		t-test for Equality of Means						95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference			
	,420	,525	5,986	18	,000	15594,800	2605,0855	10121,72	21067,88	
salaire	Equal variances assumed			5,986	16,663	,000	15594,800	2605,0855	10090,07	21099,53
	Equal variances not assumed									

FIG. 5.4 – Independent Samples T Test

Voici, dans l'ordre, comment effectuer l'analyse de la sortie 5.4. On suppose que le seuil de signification α est fixé à 0,05 pour tous les tests.

1. On doit d'abord faire le test de Levene pour vérifier si les variances des deux populations sont égales ou non (deux premières colonnes du tableau). Ceci est nécessaire pour savoir si l'on doit interpréter le tableau selon la première ou la deuxième ligne.

Le test de Levene permet en fait de résoudre le test d'hypothèses suivant :

$$H_0 : \text{Les variances des deux populations sont égales} (\sigma_{\text{Québec}}^2 = \sigma_{\text{Ontario}}^2).$$

$$H_1 : \text{Les variances des deux populations ne sont pas égales} (\sigma_{\text{Québec}}^2 \neq \sigma_{\text{Ontario}}^2).$$

On rejette l'hypothèse H_0 si la p -value (Sig. du test de Levene, voir la deuxième colonne du tableau 5.4) est plus petite que le seuil $\alpha = 0,05$, sinon on ne rejette pas H_0 . Dans notre cas la p -value est égale à 0,525, ce qui est plus grand que 0,05. Ainsi on ne rejette pas H_0 , ce qui signifie que les variances sont égales. Ainsi, dans la suite, on utilisera les statistiques de la première ligne (Equal variances assumed). Si H_0 avait été rejetée, on aurait utilisé les statistiques de la deuxième ligne (Equal variances not assumed).

2. On peut maintenant traiter le test d'hypothèses principal :

$$H_0 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} = 0$$

$$H_1 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} \neq 0$$

Pour ce faire, on utilise la *p*-value (*Sig. (2-tailed)*) de la 5^e colonne du tableau 5.4. On rejette l'hypothèse H_0 si cette *p*-value est plus petite que le seuil $\alpha = 0,05$, sinon on ne rejette pas H_0 . Dans notre cas, la *p*-value est égale à 0,000, et par conséquent on rejette H_0 . Ainsi nous rejetons la nullité de la différence et admettons qu'il y a une différence significative entre les salaires moyens de ces deux provinces, et ce au risque de se tromper 1 fois sur 20.

3. Lorsqu'on admet qu'il y a une différence significative entre les deux moyennes, l'étape suivante est de voir comment s'exprime cette différence. La colonne **Mean Difference** nous donne l'estimation ponctuelle de la différence des moyennes (c'est simplement la différence des moyennes échantillonnelles). Ici on peut donc affirmer que ponctuellement, il y a une différence de 15 594,80 \$ entre la moyenne salariale au Québec et celle en Ontario, en faveur de l'Ontario.

Il est plus intéressant de considérer l'intervalle de confiance pour cette différence de moyennes, pour ainsi étendre à la population l'estimation de la différence des moyennes salariales. L'intervalle est donné dans les deux dernières colonnes : ici on voit que la véritable différence entre les salaires moyens est comprise entre 10 121,72 \$ et 21 067,88 \$ (en faveur de l'Ontario), et ce 19 fois sur 20. Il est possible de changer ce niveau de confiance (en allant dans le bouton **Options...** lorsqu'on effectue les commandes pour obtenir le **Independant Samples T Test**).

On peut ainsi résumer l'analyse que l'on vient de faire :

« Au risque de se tromper 1 fois sur 20, on peut affirmer que pour des emplois similaires, il y a une différence significative entre le salaire moyen des cadres québécois et le salaire moyen des cadres ontariens. On peut affirmer, avec 95 % des chances d'avoir raison, qu'en moyenne les cadres québécois gagnent entre 10 121,72 \$ et 21 067,88 \$ de moins que les cadres ontariens. »

Remarque. Il est également possible de faire des tests unilatéraux :

$$\begin{aligned} H_0 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} &= 0 & \text{ou} & \quad H_0 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} = 0 \\ H_1 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} &< 0 & H_1 : \mu_{\text{Québec}} - \mu_{\text{Ontario}} &> 0 \end{aligned}$$

Il suffit alors de changer la règle de décision :

On rejette H_0 si la différence des moyennes échantillonnelles va dans le même sens que l'hypothèse H_1 , et si la p -value est plus petite que 2α . Sinon on accepte H_0 comme étant vraisemblable.

5.2 Variable polychotomique

Une variable polychotomique est une variable discrète qui admet plus de deux valeurs distinctes. Dans notre exemple introductif, nous pourrions être intéressés à mesurer l'influence de la variable `province` avec toutes ses modalités (`Québec`, `Ontario` et `Alberta`) sur le salaire des cadres. L'hypothèse nulle serait alors

$$H_0 : \mu_{\text{Québec}} = \mu_{\text{Ontario}} = \mu_{\text{Alberta}}.$$

Cependant, il est impossible d'utiliser la stratégie développée dans la section précédente (il est insensé d'écrire $\mu_{\text{Québec}} - \mu_{\text{Ontario}} - \mu_{\text{Alberta}} = 0$). Une stratégie pourrait être de comparer les moyennes deux à deux, mais cette façon de faire augmente la probabilité d'erreurs (de façon multiplicative) à mesure que le nombre de comparaisons augmente.

Il est préférable d'utiliser une autre méthode. Cette méthode porte le nom d'analyse de la variance à un facteur (**One-Way ANOVA**).

La technique de l'analyse de la variance (ANOVA) à un facteur permet de traiter les différences de moyennes d'une variable continue lorsqu'elle est divisée en plusieurs groupes (les groupes étant imposés par les k modalités de la variable discrète). On traite alors le test d'hypothèses suivant :

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_1 : \text{Au moins une des moyennes est différente.}$$

qui dans le cadre de l'exemple s'écrit

$$H_0 : \mu_{\text{Québec}} = \mu_{\text{Ontario}} = \mu_{\text{Alberta}}$$

$$H_1 : \text{Au moins une des moyennes est différente.}$$

La technique de l'analyse de la variance (ANOVA) permet donc de trancher si au moins une des moyennes diffère des autres, illustrant ainsi le lien à étudier. Cette technique sera présentée en détails dans la sous-section 5.6 et permettra de résoudre le test d'hypothèses précédent.

Avant de résoudre le test, une analyse descriptive telle que présentée à la section 5.1 est de mise. Et pour pouvoir appliquer l'analyse de la variance, certaines conditions doivent être respectées ; c'est ce que nous voyons dans la prochaine sous-section.

5.2.1 Pré-requis

Pour qu'une analyse de la variance soit valide, deux hypothèses doivent être vérifiées :

- **Les échantillons proviennent de populations normales ;**
- **Les variances des populations sont égales.**

La vérification de la normalité se fait de la même façon que celle décrite dans la sous-section 5.1.1. Dans le cadre de l'exemple, on sait déjà que la normalité est respectée pour

le Québec et l'Ontario. Si on se réfère au tableau 5.3, on voit que les p -values associées à l'Alberta sont égales à 0,200 et 0,730, ce qui est plus grand que $\alpha = 0,05$. Ainsi on conserve H_0 (qui affirme que les données de la population suivent une loi normale) pour les trois provinces, ce qui nous permet de passer à l'étape suivante.

Pour vérifier si l'hypothèse de l'égalité des variances dans les populations est respectée, il faut utiliser la statistique de Levene. Celle-ci permet de résoudre le test d'hypothèses suivant :

H_0 : Les variances des populations sont égales ($\sigma_{\text{Québec}}^2 = \sigma_{\text{Ontario}}^2 = \sigma_{\text{Alberta}}^2$).

H_1 : Au moins une des variances est différente.

Pour obtenir la sortie 5.5 qui contient la statistique de Levene et la p -value associée, il faut effectuer les commandes suivantes :

Menu SPSS :	→ Analyse
	→ Compare Means
	→ One-Way ANOVA...
Dans la fenêtre Dependent List :	→ salaire (la variable continue)
Dans la fenêtre Factor :	→ province (la variable discrète)
Dans le bouton Options... :	✓ Homogeneity of variance test

On obtient alors, entre autres, la figure 5.5 dans laquelle se retrouve la p -value associée à la statistique de Levene (dernière colonne du tableau).

Test of Homogeneity of Variances			
salaire			
Levene Statistic	df1	df2	Sig.
1,691	2	27	,203

FIG. 5.5 – Vérification de l'égalité des variances

Pour résoudre le test d'hypothèses sur les variances, la règle de décision est la suivante :

Nous rejetons l'hypothèse H_0 si la p -value est plus petite que le seuil de signification α fixé (par exemple $\alpha = 0,05$). Sinon, nous ne rejetons pas H_0 et la considérons comme vraisemblable.

Dans l'exemple qui nous intéresse, la p -value est égale à 0,203, et donc nous ne rejetons pas H_0 . Ainsi l'hypothèse d'égalité des variances dans les populations est vérifiée.

En somme, l'analyse de la variance que nous allons obtenir sera valide puisque les hypothèses de normalité et d'égalité des variances sont vérifiées.

Remarque. Mentionnons que l'ANOVA est assez robuste aux violations. Ainsi, si jamais la normalité ou l'égalité des variances est rejetée au seuil α , mais que la p -value associée est comprise entre $\alpha/2$ et α (par exemple entre 0,025 et 0,05 lorsque $\alpha = 0,05$), on poursuivra l'analyse. On fera alors mention de ce léger accroc, et un ajustement sera nécessaire lors de l'interprétation de l'analyse (sous-section 5.2.4).

5.2.2 Analyse de la variance : ANOVA

Comme vu au début de la présente section, la méthode de l'analyse de la variance nous permettra de résoudre le test d'hypothèses suivant :

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \text{Au moins une des moyennes est différente.}$$

D'abord, pour mieux comprendre comment fonctionne la méthode de l'analyse de la variance, considérons l'expression suivante :

$$\sum_{i=1}^n (x_i - \bar{x}_{\text{totale}})^2 = \underbrace{\sum_{j=1}^k n_j (\bar{x}_j - \bar{x}_{\text{totale}})^2}_{\text{Variation entre les groupes}} + \underbrace{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}_{\text{Variation dans les groupes}}$$

La méthode ANOVA décompose la variation totale de la variable continue (ici la variable **salaire**) en deux sources bien distinctes : la variation attribuable à la **différence entre les groupes** et la variation que l'on retrouve **dans les groupes**.

Si H_0 est vraie, alors la variation entre les groupes devrait être petite. Effectuons les calculs dans le cadre de l'exemple pour illustrer nos propos. Pour simplifier les choses, nous utiliserons la variable `sal_mil` qui est la variable `salaire` exprimée en milliers de dollars (il suffit d'aller dans `Transform` et ensuite `Compute Variable...` et de diviser la variable `salaire` par 1 000). On obtient alors les chiffres suivants (à l'aide de quelques analyses descriptives pour le premier tableau) :

$\bar{x}_{\text{totale}} = 48,87$	$\bar{x}_{\text{Alberta}} = 40,61$	$\bar{x}_{\text{Ontario}} = 60,80$	$\bar{x}_{\text{Québec}} = 45,20$
$s_{\text{totale}}^2 = 116,72$	$s_{\text{Alberta}}^2 = 59,49$	$s_{\text{Ontario}}^2 = 43,55$	$s_{\text{Québec}}^2 = 24,32$

$$\begin{aligned} \text{Variation totale} &= \sum_{i=1}^n (x_i - \bar{x}_{\text{totale}})^2 \\ &= (n-1) \times s_{\text{totale}}^2 = 29 \times 116,72 = 3384,88. \end{aligned}$$

$$\begin{aligned} \text{Variation entre les groupes} &= \sum_{j=1}^k n_j (\bar{x}_j - \bar{x}_{\text{totale}})^2 \\ &= n_{\text{Alberta}} (\bar{x}_{\text{Alberta}} - \bar{x}_{\text{totale}})^2 + \\ &\quad n_{\text{Ontario}} (\bar{x}_{\text{Ontario}} - \bar{x}_{\text{totale}})^2 + \\ &\quad n_{\text{Québec}} (\bar{x}_{\text{Québec}} - \bar{x}_{\text{totale}})^2 \\ &= 10(40,61 - 48,87)^2 + 10(60,80 - 48,87)^2 \\ &\quad + 10(45,20 - 48,87)^2 \\ &\approx 2238,67. \end{aligned}$$

$$\begin{aligned} \text{Variation dans les groupes} &= \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \\ &= \sum_{j=1}^k (n_j - 1) s_j^2 \\ &= (n_{\text{Alberta}} - 1) s_{\text{Alberta}}^2 + \\ &\quad (n_{\text{Ontario}} - 1) s_{\text{Ontario}}^2 + \\ &\quad (n_{\text{Québec}} - 1) s_{\text{Québec}}^2 \\ &= 9 \times 59,49 + 9 \times 43,55 + 9 \times 24,32 \\ &\approx 1146,21. \end{aligned}$$

$$\underbrace{3384,88}_{\text{Variation totale}} = \underbrace{2238,67}_{\text{Variation entre les groupes}} + \underbrace{1146,21}_{\text{Variation dans les groupes}}$$

On retrouve ces variations dans une table ANOVA générée par SPSS. Pour obtenir la sortie 5.6, il faut effectuer les commandes suivantes :

Menu SPSS :	→ Analyse → Compare Means → One-Way ANOVA...
-------------	--

Dans la fenêtre Dependent List :	→ salaire (la variable continue)
----------------------------------	----------------------------------

Dans la fenêtre Factor :	→ province (la variable discrète)
--------------------------	-----------------------------------

ANOVA

sal_mil

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2238,670	2	1119,335	26,367	,000
Within Groups	1146,212	27	42,452		
Total	3384,882	29			

FIG. 5.6 – Table ANOVA

On se sert de la *p*-value (Sig.) de la dernière colonne pour résoudre le test suivant :

$$H_0 : \mu_{\text{Québec}} = \mu_{\text{Ontario}} = \mu_{\text{Alberta}}$$

$$H_1 : \text{Au moins une des moyennes est différente.}$$

La règle de décision est la règle habituelle :

Nous rejetons l'hypothèse H_0 si la *p*-value est plus petite que le seuil de signification α fixé (par exemple $\alpha = 0,05$). Sinon, nous ne rejetons pas H_0 et la considérons comme vraisemblable.

Puisque que la p -value est égale à 0,000, ce qui évidemment plus petit que $\alpha = 0,05$, nous rejetons H_0 . Ainsi, au risque de se tromper 1 fois sur 20, nous pouvons affirmer qu'au moins une des moyennes salariales est significativement différente des autres. Il reste à voir comment s'exprime cette différence, ce qui est fait dans les deux prochaines sous-sections.

5.2.3 Force du lien

C'est la variation entre les groupes (ici entre les provinces) qui nous permet de quantifier le lien entre la variable discrète et la variable continue. Ici, le % de la variation totale expliquée par la différence entre les provinces est de 66,1 % (i.e. $2238,67/3384,88 \times 100$). Plus ce % est élevé et moins l'égalité des moyennes est probable, ce qui amène le rejet de H_0 .

Dans une One-Way Anova, ce pourcentage porte le nom de ETA SQUARE (ETA^2). Cette mesure représente le % de la variation de la variable continue expliquée par les différences entre les classes de la variable discrète. L'analyste ne doit pas confondre le ETA^2 et le r^2 qui est issu d'une régression qui jongle avec une idée de linéarité. L'analyste peut utiliser la table d'interprétation suivante pour interpréter le ETA, qui est la racine carrée du ETA^2 ($\text{ETA} = \sqrt{\text{ETA}^2}$).

$0,70 \leq \text{ETA} \leq 1$	Relation très forte.
$0,50 \leq \text{ETA} < 0,70$	Relation forte.
$0,30 \leq \text{ETA} < 0,50$	Relation correcte.
$0,10 \leq \text{ETA} < 0,30$	Relation faible.
$0,00 \leq \text{ETA} < 0,10$	Relation négligeable.

Ainsi, dans le cadre de notre exemple, on a $\text{ETA} = \sqrt{0,661} = 0,813$. On est donc en mesure de dire que la relation entre les provinces et le salaire des cadres est très forte.

5.2.4 Comparaisons multiples : analyse Post Hoc

L'analyse de la variance nous informe de l'existence d'une différence entre les moyennes. Ce type d'analyse ne nous informe en rien sur l'expression de la différence. En effet, admettre qu'au moins une des moyennes est différente des autres n'informe en rien sur où se situe cette différence.

Les analyses Post Hoc ont été conçues pour palier à ce problème. Ces analyses permettent de dire quelles sont les moyennes qui se distinguent des autres et quelles sont celles qui sont considérées égales. Il y a près d'une quinzaine d'analyses Post Hoc disponibles sur SPSS ; nous n'en présenterons que deux. La première, celle de Scheffe, produit des résultats conservateurs. De plus elle est robuste aux violations des hypothèses de normalité et de l'égalité des variances. La seconde, celle de Bonferroni, offre un portrait plus juste de la réalité. Cependant, elle est moins robuste aux violations que celle de Scheffe. Le choix de la méthode revient à l'analyste. Dans le cadre de ce cours, s'il y a eu une violation à la normalité ou à l'égalité des variances (voir la sous-section 5.2.1), nous utiliserons Scheffe. Dans le cas contraire, c'est Bonferroni qui sera utilisée.

Dans le cadre de l'exemple, les deux analyses soutiennent les mêmes conclusions et s'interprètent de la même façon. Compte tenu que les hypothèses de validité ont été vérifiées, nous ne présenterons que l'interprétation de l'analyse de Bonferroni.

Pour obtenir la sortie 5.7, il faut effectuer les opérations suivantes :

Menu SPSS :	→ Analyse
	→ Compare Means
	→ One-Way ANOVA . . .

Dans la fenêtre Dependent List :	→ salaire (la variable continue)
----------------------------------	----------------------------------

Dans la fenêtre Factor :	→ province (la variable discrète)
--------------------------	-----------------------------------

Dans le bouton Post Hoc . . . :	✓ Bonferroni
---------------------------------	--------------

Multiple Comparisons						
		Dependent Variable: sal_mil				
		Bonferroni				
(I) Province	(J) Province	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
Alberta	Ontario	-20,18296*	2,91384	,000	-27,6204	-12,7455
	Québec	-4,58816	2,91384	,381	-12,0256	2,8493
Ontario	Alberta	20,18296*	2,91384	,000	12,7455	27,6204
	Québec	15,59480*	2,91384	,000	8,1573	23,0323
Québec	Alberta	4,58816	2,91384	,381	-2,8493	12,0256
	Ontario	-15,59480*	2,91384	,000	-23,0323	-8,1573

*. The mean difference is significant at the .05 level.

FIG. 5.7 – Analyse Post Hoc : Bonferroni

Voici comment interpréter la sortie 5.7. Tout d'abord, il faut comparer les provinces deux-à-deux, en parcourant toutes les paires possibles : il y a en fait $\frac{k(k-1)}{2}$ paires à observer, où k est le nombre de modalités de la variable discrète. Ainsi dans notre exemple il y a $\frac{3(3-1)}{2} = 3$ paires à considérer. On remarque que si on fait toutes les comparaisons disponibles dans le tableau 5.7, on fait le double de ce qui est nécessaire car il y a de l'information redondante. Voici donc comment on fait ces comparaisons.

1. μ_{Alberta} et μ_{Ontario} : la p -value associée à la différence des moyennes est de 0,000, donc au seuil 0,05 on conclut que la différence entre ces deux moyennes est significative. La différence des moyennes s'estime ponctuellement à 20 182,96 \$ en faveur de l'Ontario, donc $\mu_{\text{Ontario}} > \mu_{\text{Alberta}}$. Et il y a une probabilité de 95 % de retrouver la différence de ces moyennes entre 12 745,50 \$ et 27 620,40 \$ au niveau de la population.
2. μ_{Alberta} et $\mu_{\text{Québec}}$: la p -value associée à la différence des moyennes est de 0,381, donc au seuil 0,05 on conclut que la différence entre ces deux moyennes n'est pas significative. Donc $\mu_{\text{Alberta}} = \mu_{\text{Québec}}$.
3. μ_{Ontario} et $\mu_{\text{Québec}}$: la p -value associée à la différence des moyennes est de 0,000, donc au seuil 0,05 on conclut que la différence entre ces deux moyennes est significative.

La différence des moyennes s'estime ponctuellement à 15 594,80 \$ en faveur de l'Ontario, donc $\mu_{\text{Ontario}} > \mu_{\text{Québec}}$. Et il y a une probabilité de 95 % de retrouver la différence de ces moyennes entre 8 157,30 \$ et 23 032,30 \$ au niveau de la population.

On peut résumer la situation de la façon suivante : $(\mu_{\text{Québec}} = \mu_{\text{Alberta}}) < \mu_{\text{Ontario}}$.

La visualisation des Box Plots à l'étape descriptive permet de voir rapidement quelle(s) est (sont) les moyennes qui risque(nt) de se détacher des autres.

5.3 Utilisation de l'écart-réduit

5.3.1 Comparaison de deux moyennes

Dans cette sous-section, nous voyons comment effectuer un test d'hypothèses dans lequel on compare deux moyennes issues de deux échantillons indépendants, et ce à un seuil de signification α . Les tableaux suivants résument les règles de décision selon le contexte et les hypothèses.

Cas 1. Le tableau suivant n'est valide que lorsque les variances des deux populations σ_1^2 et σ_2^2 sont connues, ce qui est peu fréquent dans la pratique. Il faut aussi que les populations d'origine soient distribuées normalement, ou que les deux échantillons soient de taille supérieure ou égale à 30.

Hypothèses statistiques	Écart réduit	Règle de décision
$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 \neq d_0$	$Z = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Rejeter H_0 si $Z > z_{\alpha/2}$ ou $Z < -z_{\alpha/2}$
$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 > d_0$	$Z = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Rejeter H_0 si $Z > z_\alpha$
$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 < d_0$	$Z = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Rejeter H_0 si $Z < -z_\alpha$

Conditions : Populations normales de variances connues,

ou grands échantillons ($n_1 \geq 30$ et $n_2 \geq 30$) et variances des populations connues.

Dans ce cas, l'intervalle de confiance de niveau $1 - \alpha$ pour la différence entre les deux moyennes $(\mu_1 - \mu_2)$ est donné par :

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Cas 2. On est ici dans le cas où les variances des populations sont inconnues, et on utilise alors les variances échantillonnelles s_1^2 et s_2^2 pour les estimer. Il faut que les populations d'origine soient distribuées normalement, et alors la statistique utilisée pour faire le test suit une loi de Student, dont le nombre de degrés de liberté est donné par la formule suivante :

$$dl = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

Il est à noter aussi que les estimations sont meilleures lorsque les tailles d'échantillon n_1 et n_2 sont grandes.

Ici l'intervalle l'intervalle de confiance de niveau $1 - \alpha$ pour la différence entre les deux moyennes $(\mu_1 - \mu_2)$ est donné par :

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2; dl} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Le tableau qui suit montre comment effectuer un test d'hypothèses dans ce contexte.

Hypothèses statistiques	Écart réduit	Règle de décision
$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 \neq d_0$	$T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Rejeter H_0 si $T > t_{\alpha/2; \text{dl}}$ ou $T < -t_{\alpha/2; \text{dl}}$
$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 > d_0$	$T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Rejeter H_0 si $T > t_{\alpha; \text{dl}}$
$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 < d_0$	$T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Rejeter H_0 si $T < -t_{\alpha; \text{dl}}$

Conditions : Populations normales de variances inconnues.

Cas 3. On suppose ici que les variances des deux populations sont égales, mais sa valeur est inconnue (si elle est connue, on a alors simplement $\sigma_1^2 = \sigma_2^2$, et on utilise le tableau du cas 1). On soit estimer cette valeur commune à partir des variances échantillonnelles s_1^2 et s_2^2 de la façon suivante :

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

On doit aussi supposer que les populations d'origine se distribuent selon une loi normale.

L'intervalle de confiance de niveau $1 - \alpha$ pour la différence entre les deux moyennes $(\mu_1 - \mu_2)$ est ici donné par :

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(\alpha/2; n_1 + n_2 - 2)} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

Le tableau qui suit montre comment effectuer un test d'hypothèses dans ce contexte.

Hypothèses statistiques	Écart réduit	Règle de décision
$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 \neq d_0$	$T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	Rejeter H_0 si $T > t_{\alpha/2; n_1+n_2-2}$ ou $T < -t_{\alpha/2; n_1+n_2-2}$
$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 > d_0$	$T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	Rejeter H_0 si $T > t_{\alpha; n_1+n_2-2}$
$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 < d_0$	$T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	Rejeter H_0 si $T < -t_{\alpha; n_1+n_2-2}$

Conditions : Populations normales de variances inconnues,
mais supposées égales.

Exemple 5.3.1 Une firme étudie les délais de livraison de deux fournisseurs de matières premières. Elle est globalement satisfaite de ses rapports avec le fournisseur A et est d'accord de garder ce fournisseur, si le délai de livraison est le même ou est inférieur à celui du fournisseur B . Cependant, si la firme découvre que le délai de livraison du fournisseur B est inférieur à celui du fournisseur A , elle passera ses commandes de matières premières au fournisseur B .

- Formuler les hypothèses appropriées pour cette situation.
- Supposez que deux échantillons aléatoires indépendants révèlent les caractéristiques suivantes concernant les délais de livraison des deux fournisseurs.

Fournisseur A	Fournisseur B
$n_1 = 50$	$n_2 = 30$
$\bar{x}_1 = 14$ jours	$\bar{x}_2 = 12,5$ jours
$s_1 = 3$ jours	$s_2 = 2$ jours

Avec $\alpha = 0,05$, quelle est votre conclusion au test d'hypothèses (avec les hypothèses de a)) ? Quel fournisseur recommanderiez-vous ?

Solution.

Exemple 5.3.2 Un nouveau procédé technique pour les moteurs d'automobiles est supposé réduire le taux de monoxyde de carbone qui est émis. Des essais ont été effectués avec le procédé actuel et le nouveau procédé, et on obtient les résultats suivants :

Procédé actuel	Nouveau procédé
$n_1 = 20$	$n_2 = 20$
$\bar{x}_1 = 7,68 \text{ ppm}$	$\bar{x}_2 = 5,58 \text{ ppm}$
$s_1 = 1,33 \text{ ppm}$	$s_2 = 1,23 \text{ ppm}$

Formuler les bonnes hypothèses et les tester avec un seuil de signification $\alpha = 0,05$ (on suppose que les populations sont normales et que leurs variances sont égales). Le nouveau système réduit-il significativement le taux de monoxyde de carbone ?

Solution.

5.3.2 Comparaison de deux proportions

Nous voyons dans cette sous-section comment comparer deux proportions à l'aide d'un test d'hypothèses, et comment construire un intervalle de confiance pour la différence entre deux proportions.

Cas 1. Le tableau qui suit présente les règles de décision à suivre lorsque l'hypothèse nulle du test est de la forme $H_0 : \pi_1 = \pi_2$, ce qui est équivalent à $H_0 : \pi_1 - \pi_2 = 0$. Dans ce cas, on note p_1 et p_2 les proportions issues des deux échantillons qui sont de taille n_1 et n_2 respectivement. Puisque l'hypothèse nulle stipule que ces deux proportions sont supposées égales, on doit estimer la proportion commune p qui servira dans le calcul de l'écart-réduit :

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}.$$

L'échantillon doit être suffisamment grand pour que le test soit valide. On s'en assure en vérifiant que $n_1 p_1 \geq 5$, $n_1(1 - p_1) \geq 5$, $n_2 p_2 \geq 5$ et $n_2(1 - p_2) \geq 5$.

Hypothèses statistiques	Écart réduit	Règle de décision
$H_0 : \pi_1 = \pi_2$ $H_1 : \pi_1 \neq \pi_2$	$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Rejeter H_0 si $Z > z_{\alpha/2}$ ou $Z < -z_{\alpha/2}$
$H_0 : \pi_1 = \pi_2$ $H_1 : \pi_1 > \pi_2$	$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Rejeter H_0 si $Z > z_\alpha$
$H_0 : \pi_1 = \pi_2$ $H_1 : \pi_1 < \pi_2$	$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Rejeter H_0 si $Z < -z_\alpha$

Conditions : $n_1 p_1 \geq 5$, $n_1(1 - p_1) \geq 5$, $n_2 p_2 \geq 5$ et $n_2(1 - p_2) \geq 5$.

Cas 2. Le tableau qui suit présente les règles de décision à suivre lorsque l'hypothèse nulle du test est de la forme $H_0 : \pi_1 - \pi_2 = \pi_d$, où $\pi_d \neq 0$. Tout comme dans le cas 1, on note p_1 et p_2 les proportions issues des deux échantillons qui sont de taille n_1 et n_2 respectivement.

L'échantillon doit être suffisamment grand pour que le test soit valide. On s'en assure en vérifiant que $n_1 p_1 \geq 5$, $n_1(1 - p_1) \geq 5$, $n_2 p_2 \geq 5$ et $n_2(1 - p_2) \geq 5$.

Hypothèses statistiques	Écart réduit	Règle de décision
$H_0 : \pi_1 - \pi_2 = \pi_d$ $H_1 : \pi_1 - \pi_2 \neq \pi_d$	$Z = \frac{(p_1 - p_2) - \pi_d}{\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}}$	Rejeter H_0 si $Z > z_{\alpha/2}$ ou $Z < -z_{\alpha/2}$
$H_0 : \pi_1 - \pi_2 = \pi_d$ $H_1 : \pi_1 - \pi_2 > \pi_d$	$Z = \frac{(p_1 - p_2) - \pi_d}{\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}}$	Rejeter H_0 si $Z > z_\alpha$
$H_0 : \pi_1 - \pi_2 = \pi_d$ $H_1 : \pi_1 - \pi_2 < \pi_d$	$Z = \frac{(p_1 - p_2) - \pi_d}{\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}}$	Rejeter H_0 si $Z < -z_\alpha$

Conditions : $n_1 p_1 \geq 5$, $n_1(1 - p_1) \geq 5$, $n_2 p_2 \geq 5$ et $n_2(1 - p_2) \geq 5$.

Finalement, l'intervalle de confiance de niveau $1 - \alpha$ pour la différence entre deux proportions est donné par :

$$(p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}.$$

Cet intervalle est valide si $n_1 p_1 \geq 5$, $n_1(1 - p_1) \geq 5$, $n_2 p_2 \geq 5$ et $n_2(1 - p_2) \geq 5$.

Exemple 5.3.3 On a utilisé un échantillon de 1545 hommes et un échantillon indépendant de 1691 femmes pour comparer la quantité de tâches ménagères effectuée par les femmes et les hommes dans les couples où les deux personnes travaillent. L'étude a révélé que 67,5 % des hommes et 60,8 % des femmes considéraient la répartition des tâches ménagères équitable. Au seuil $\alpha = 0,05$, peut-on conclure que la proportion d'hommes qui considèrent la répartition des tâches ménagères équitable est supérieure à celle des femmes ?

Solution. On a $p_H = 0,675$ et $p_F = 0,608$. Donc

$$p = \frac{0,675 \cdot 1545 + 0,608 \cdot 1691}{1545 + 1691} = 0,64.$$

Aussi, on a

$$\begin{aligned} n_H p &= 1545 \cdot 0,64 = 988,8, & n_H(1-p) &= 1545 \cdot 0,36 = 556,2, \\ n_F p &= 1691 \cdot 0,64 = 1082,24 & \text{et} & n_F(1-p) = 1691 \cdot 0,36 = 608,76. \end{aligned}$$

Ainsi les hypothèses d'application du test pour comparer deux proportions sont vérifiées puisque toutes ces valeurs sont plus grandes que 5.

Les hypothèses que nous voulons traiter sont les suivantes :

$$H_0 : \pi_H = \pi_F$$

$$H_1 : \pi_H > \pi_F$$

L'écart-réduit est

$$Z = \frac{p_H - p_F}{\sqrt{p(1-p) \left(\frac{1}{n_H} + \frac{1}{n_F} \right)}} = \frac{0,675 - 0,608}{\sqrt{0,64 \cdot 0,36 \left(\frac{1}{1545} + \frac{1}{1691} \right)}} = 3,97.$$

Au seuil $\alpha = 0,05$, on rejette H_0 si $Z > z_{0,05} = 1,645$. Puisque $Z = 3,97 > 1,645$, on rejette H_0 . Donc au risque de se tromper une fois sur 20, on peut conclure que la proportion d'hommes qui considèrent la répartition des tâches ménagères équitable est supérieure à celle des femmes... faut pas juste sortir les poubelles, les gars! ;-)

5.3.3 ANOVA : test de Fisher

Nous avons vu que pour comparer plusieurs moyennes, on doit décomposer la variation totale des données en deux sources de variation :

$$\begin{array}{rcl} \text{Variation totale} & = & \text{variation entre les groupes} + \text{variation dans les groupes} \\ \\ \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 & = & \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \\ \\ \text{SCT} & = & \text{SC}_A + \text{SC}_{\text{Rés}} \end{array}$$

Les sommes de carrés divisées par leur nombre de degrés de liberté sont des variances appelées **carrés moyens**. Les notations sont les suivantes : $\text{CM}_A = \frac{\text{SC}_A}{k-1}$, $\text{CM}_{\text{Rés}} = \frac{\text{SC}_{\text{Rés}}}{n-k}$.

Pour résoudre le test d'hypothèses, les conditions sont les mêmes que celles vues précédemment :

Conditions d'application du test :

- a) Échantillons prélevés au hasard et indépendamment ;
- b) Les populations se distribuent normalement ;
- c) Les variances des populations sont identiques.

Lorsque les conditions sont vérifiées, on peut résoudre le test :

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \text{Au moins une des moyennes est différente.}$$

au seuil α : on calcule d'abord le rapport des carrés moyens :

$$F = \frac{\text{CM}_A}{\text{CM}_{\text{Rés}}}$$

qui est une statistique qui se distribue selon la loi de Fisher avec $(k - 1)$ et $(n - k)$ degrés de libertés.

Règle de décision : au seuil α , on rejette H_0 si $F > F_{\alpha;(k-1),(n-k)}$.

5.4 Exercices du chapitre

Vous décidez, un bon jour, de vous impliquer dans l'organisation de partys pour des levées de fonds pour votre faculté. Ayant suivi un merveilleux cours de statistiques, vous décidez de faire une petite étude pour savoir quels sont les facteurs déterminants d'un bon party. Suite à un party ayant eu lieu sur le campus, vous faites remplir le questionnaire suivant à 75 personnes qui étaient présentes à cette soirée (en échange d'un 2 pour 1 sur la bière pour un prochain party...) :

1. Êtes-vous satisfait de votre soirée ? Inscrivez un X sur la partie de la droite qui correspond le mieux à votre niveau de satisfaction, en sachant que le 0 correspond un à niveau très bas de satisfaction, et le 10 à un niveau de satisfaction très élevé :



2. Combien de bières avez-vous consommées ?
3. Avez-vous aimé la musique ? Inscrivez un X sur la partie de la droite qui correspond le mieux à votre niveau de satisfaction, en sachant que le 0 correspond un à niveau très bas de satisfaction, et le 10 à un niveau de satisfaction très élevé :



4. Êtes-vous célibataire ? Oui Non
5. Êtiez-vous conducteur désigné ? Oui Non
6. Votre sexe ? Masculin Féminin

Note : Les échelles des questions 1 et 3 mesurent 10 cm de long. Une fois les questionnaires remplis, une règle a été utilisée pour mesurer la distance qui sépare le rebord gauche de la droite et le X sur la droite. Ce sont ces distances qui se retrouvent dans la base de données. Plus la distance est grande, plus le niveau de satisfaction est élevé.

La base de données porte le nom `Party.sav`.

Faites les analyses nécessaires pour voir lesquelles des variables discrètes (`bieres`, `celibat`, `auto`, `sexe`) ont une influence sur le niveau de satisfaction. Pour la variable `bieres`, recodez-la pour obtenir une variable discrète à 3 modalités (0 à 2 bières, 3 à 5 bières, 6 bières et plus).

Chapitre 6

Relation entre deux variables continues

L'étude de la relation entre deux variables continues (mesurées à l'aide d'une échelle d'intervalles ou de ratio) peut être menée de deux façons différentes selon qu'il s'agit d'étudier une relation d'interdépendance ou de dépendance.

La section 6.1 présente l'analyse en corrélation. Elle est utile pour étudier les relations d'interdépendance. Elle est aussi pratique pour détecter rapidement les relations de dépendance. La section 6.2 introduit l'analyse en régression linéaire simple qui modélise la relation de dépendance entre deux variables continues.

6.1 L'analyse en corrélation linéaire simple

L'objectif de l'analyse en corrélation est de vérifier si une relation linéaire entre deux variables continues X et Y existe. Le terme linéaire doit être compris dans une optique de description et d'orientation de la relation. Par exemple, « plus la variable X augmente, plus la variable Y augmente » ; il s'agit alors d'une relation positive. Ou encore « plus la variable X augmente, plus la variable Y diminue » ; il s'agit alors d'une relation négative.

Le **coefficient de corrélation de Pearson** r est une statistique qui quantifie la puissance du lien linéaire entre les deux variables tout en indiquant la direction de la relation. Ce coefficient estime le paramètre de la corrélation linéaire ρ dans la population.

Le coefficient de corrélation de Pearson varie de -1 à +1 selon que la relation va de parfaitement négative à parfaitement positive. Lorsque le coefficient de corrélation de Pearson est égal à 0, la relation linéaire entre les deux variables est considérée nulle. Proposé par Cohen (1983), l'ampleur du coefficient de corrélation de Pearson, r , peut être faite via le schéma d'interprétation suivant :

$0,7 \leq r \leq 1$	Interrelation linéaire très forte
$0,5 \leq r < 0,7$	Interrelation linéaire forte
$0,3 \leq r < 0,5$	Interrelation linéaire modérée
$0,1 \leq r < 0,3$	Interrelation linéaire faible
$0 \leq r < 0,1$	Interrelation linéaire négligeable

FIG. 6.1 – Schéma d'interprétation de r

La figure 6.2 illustre comment le coefficient de corrélation de Pearson mesure la force et la direction des liens **linéaires** entre deux variables continues.

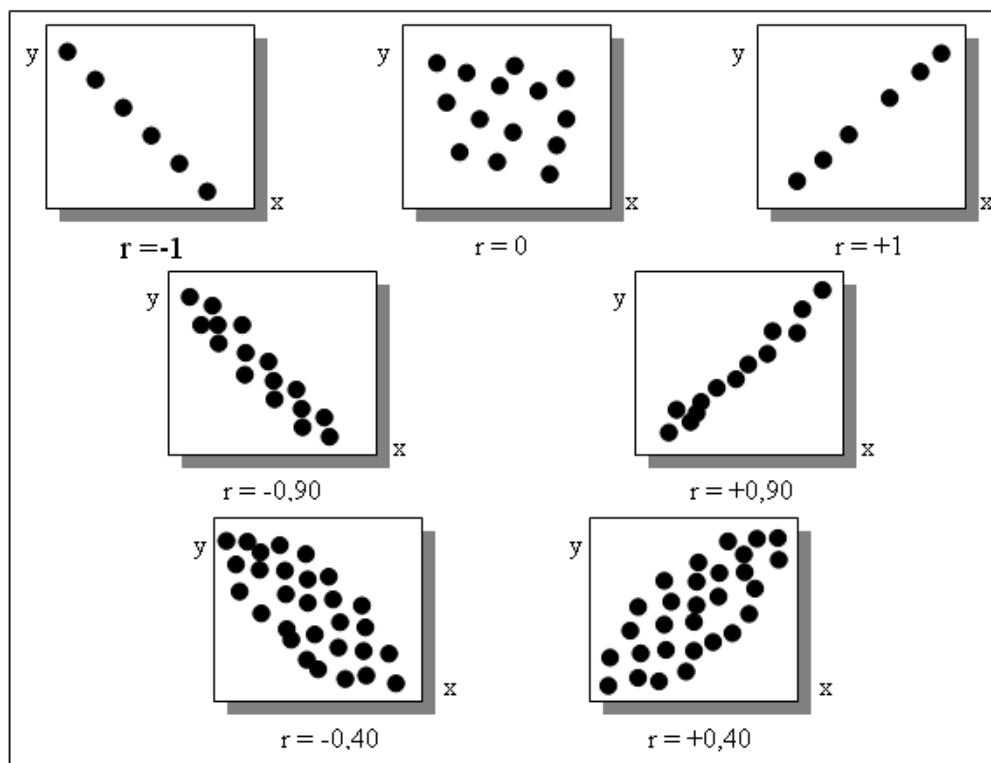


FIG. 6.2 – Quelques corrélations possibles

Il faut bien comprendre que le coefficient de corrélation de Pearson mesure la puissance et la direction des liens **linéaires** entre deux variables continues. Ainsi, il faut interpréter un coefficient $r = 0$ de la façon suivante : « Il y a absence de relation linéaire entre les deux variables ». Cependant, il est possible qu'il existe une relation autre que linéaire entre les deux variables.

Le coefficient de corrélation de Pearson r est une mesure d'intensité de relation linéaire entre deux variables continues qui ne dépend pas de la taille de l'échantillon utilisé. En effet :

$$\begin{aligned}
r &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\left(\frac{1}{n-1}\right) \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\frac{1}{n-1}\right) \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\left(\frac{1}{n-1}\right) \sum_{i=1}^n (y_i - \bar{y})^2}} \\
&= \frac{\text{Cov}(X, Y)}{s_X s_Y}.
\end{aligned}$$

De telles statistiques portent le nom de « *Effect size* » ou « *ES* » puisque seule la taille de la relation entre X et Y est mise en cause. La covariance de deux variables continues, $\text{Cov}(X, Y)$ provient de la définition de la variance du cumul de X et Y :

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y).$$

La covariance mesure à quel point les valeurs élevées (ou faibles) de la variable X correspondent aux valeurs élevées (ou faibles) de la variable Y . Cette co-variation peut être positive et accroître la variation de la somme comme elle peut être négative et réduire la variation du cumul. Cependant, la covariance n'est pas bornée ($-\infty \leq \text{Cov}(X, Y) \leq +\infty$), ce qui la rend moins utilisable. Le fait de la diviser par les écart-types de X et de Y ramène cette mesure entre -1 et $+1$.

Ce coefficient porte aussi le nom de « *Product Moment* ». Pour mieux comprendre, il faut s'arrêter à la constitution même du paramètre ρ . Il est simplement le produit des variables X et Y une fois normalisées (cotes- z) :

$$\rho = E \left[\left(\frac{X - E(X)}{\sigma_X} \right) \left(\frac{Y - E(Y)}{\sigma_Y} \right) \right] = \frac{E(X \cdot Y) - E(X) \cdot E(Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

6.1.1 Examen graphique de la relation

Que la statistique r détecte la forte ou la faible présence d'un lien linéaire, l'analyste doit toujours effectuer une analyse graphique des relations étudiées. Le tableau 6.3 contient le dilemme d'Anscombe, qui illustre à quel point une analyse en corrélation n'est pas complète sans une analyse graphique. Le dilemme compare la conclusion de quatre analyses en corrélation. La force de chacune des quatre relations est quantifiée à $r = 0,816$. À l'aide de la figure 6.4, dites pour laquelle des quatre relations le coefficient de corrélation est une mesure efficace.

Relation 1		Relation 2		Relation 3		Relation 4	
Y1	X1	Y2	X1	Y3	X1	Y4	X2
4,26	4	3,10	4	5,39	4	6,58	8
5,68	5	4,74	5	5,73	5	5,76	8
7,24	6	6,13	6	6,08	6	7,71	8
4,82	7	7,26	7	6,42	7	8,84	8
6,95	8	8,14	8	6,77	8	8,47	8
8,81	9	8,77	9	7,11	9	7,04	8
8,04	10	9,14	10	7,46	10	5,25	8
8,33	11	9,26	11	7,81	11	5,56	8
10,84	12	9,13	12	8,15	12	7,91	8
7,58	13	8,74	13	12,74	13	6,89	8
9,96	14	8,10	14	8,84	14	12,50	19

FIG. 6.3 – Le quartelet de Frank Anscombe

Pour obtenir les graphiques de la figure 6.4, il faut effectuer les opérations suivantes pour chacun des graphiques (en supposant que vous avez saisi les données de la figure 6.3) :

Menu SPSS :
 → Graphs
 → Legacy Dialogs
 → Scatter/Dot...

Sélectionnez Simple Scatter, puis appuyez sur Define

Dans la fenêtre Y Axis : → Y1 (la variable en Y)

Dans la fenêtre X Axis : → X1 (la variable en X)

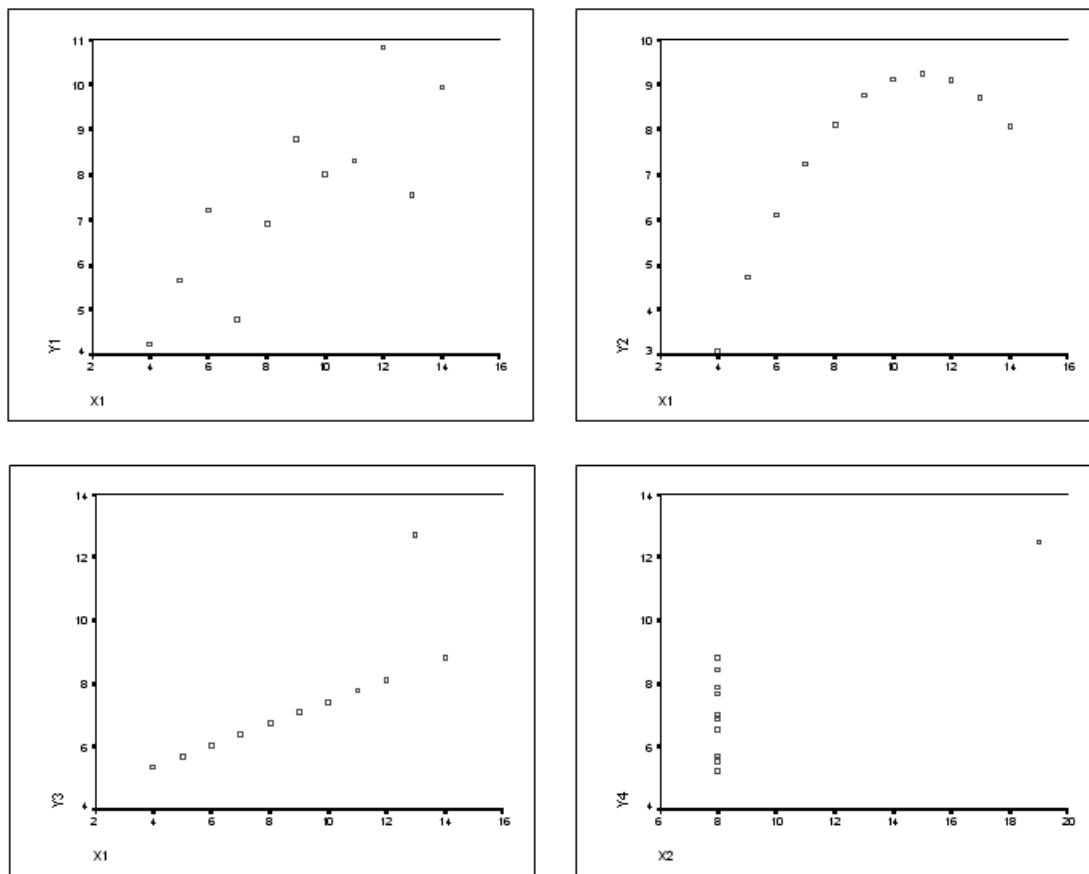


FIG. 6.4 – Le quartelet de Frank Anscombe

Le dilemme provient du fait que les quatre relations obtiennent un même coefficient de corrélation $r = 0,816$. Sans analyse graphique, il est impossible de savoir à quel point

ce coefficient illustre bien la relation, et ce, même si une relation est significative (comme nous le verrons bientôt). Dans le cas qui nous intéresse, seule la relation linéaire (Y_1 , X_1) est bien mesurée par son coefficient de corrélation.

6.1.2 Limites du coefficient de corrélation

L'examen du graphe nous permet donc de détecter certains problèmes qui affectent le calcul du coefficient de corrélation, comme les valeurs aberrantes ou un lien non linéaire. D'autres problèmes peuvent se poser, par exemple lorsqu'on tire une conclusion sur la force de l'association entre des entités, et ce, en fonction du coefficient de corrélation entre les valeurs générales ou les moyennes de ces entités. Par exemple, une conclusion tirée sur la force de la relation qui existe entre les actions et les obligations en fonction de la corrélation entre l'indice global de l'obligation (une moyenne) et l'indice global d'une action (une moyenne) peut être très trompeuse. Les actions et les obligations tendent à connaître une variation beaucoup plus grande au regard des variations globales ou moyennes. On ne doit pas utiliser une corrélation entre l'indice S&P/TSX et l'indice d'une obligation pour en déduire que tous les prix des actions et des obligations ont la même corrélation.

Aussi, il ne faut pas oublier que le coefficient de corrélation est basée sur une formule mathématique. On peut ainsi trouver de fortes corrélations entre des phénomènes qui ne sont pas reliés, comme l'apparition de taches solaires et les cycles économiques, le taux de naissance et le taux de criminalité dans un pays, etc. De telles relations sont dites **illussoires**.

Par ailleurs, constater une forte corrélation entre deux variables n'est pas suffisant pour affirmer qu'il y a une relation de dépendance entre ces deux variables. Une relation de cause à effet entre des variables doit être supportée par la théorie.

6.1.3 Analyse et test d'hypothèses

L'analyse en corrélation possède l'avantage d'être rapide à effectuer et d'être fort simple à interpréter. Par exemple, dans le cadre de l'exemple du quartelet d'Anscombe, on obtient la sortie 6.5 en effectuant les opérations suivantes :

Menu SPSS :	→ Analyse
	→ Correlate
	→ Bivariate...
Dans la fenêtre Variables :	→ X1, X2, Y1, Y2, Y3, Y4

Correlations							
	X1	X2	Y1	Y2	Y3	Y4	
Pearson	X1	1,000	,500	,816**	,816**	,816**	,370
Correlation	X2	,500	1,000	,401	,098	,219	,817**
	Y1	,816**	,401	1,000	,750**	,469	,295
	Y2	,816**	,098	,750**	1,000	,588	,099
	Y3	,816**	,219	,469	,588	1,000	,146
	Y4	,370	,817**	,295	,099	,146	1,000
Sig.	X1	,	,117	,002	,002	,002	,262
(2-tailed)	X2	,117	,	,221	,775	,518	,002
	Y1	,002	,221	,	,008	,146	,379
	Y2	,002	,775	,008	,	,057	,772
	Y3	,002	,518	,146	,057	,	,669
	Y4	,262	,002	,379	,772	,669	,
N	X1	11	11	11	11	11	11
	X2	11	11	11	11	11	11
	Y1	11	11	11	11	11	11
	Y2	11	11	11	11	11	11
	Y3	11	11	11	11	11	11
	Y4	11	11	11	11	11	11

**. Correlation is significant at the 0.01 level (2-tailed).

FIG. 6.5 – L'analyse en corrélation des variables d'Anscombe

La sortie est divisée en 3 tableaux (ou matrices). Le premier est le tableau des corrélations qui contient les coefficients de corrélation pour chacune des variables prises deux-à-deux. Par exemple, pour $X2$ et $Y4$, on a $r = 0,817$.

Cette matrice possède sur sa diagonale des 1,000 qui signifient une relation parfaite entre une même variable (ce qui est normal). Cette matrice est symétrique par rapport à sa diagonale et les termes en dehors de la diagonale correspondent aux coefficients de corrélation entre les variables en tête et en marge gauche du tableau.

Le second tableau de la sortie contient les *p*-values pour la résolution d'un test d'hypothèses effectué pour chacune des relations prises deux-à-deux. Le test confronte les hypothèses suivantes :

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

On rappelle que ρ est le paramètre de corrélation dans la population que r estime. La règle de décision est la suivante :

On rejette H_0 , et nous admettons H_1 comme vraisemblable, si la *p*-value (Sig. (2-tailed)) est plus petite que le seuil α . Sinon on considère H_0 comme vraisemblable.

Par exemple, au risque de se tromper une fois sur 20, on peut affirmer que le coefficient de corrélation entre les variables $Y4$ et $X2$ est significativement différent de 0 (où 0 exprime l'absence de lien linéaire entre les variables étudiées) puisque $0,002 < 0,05$.

Par contre, au seuil $\alpha = 0,05$, on peut affirmer que le coefficient de corrélation entre les variables $Y4$ et $X1$ n'est pas significativement différent de 0 puisque $0,262 \not< 0,05$.

Remarque : Lorsque la corrélation est significative au seuil $\alpha = 0,01$, SPSS imprime ** à côté du coefficient de corrélation, simplement pour faciliter le repérage. Malgré cela, rappelons qu'il faut toujours effectuer une analyse graphique.

Exemple 6.1.1 (Tiré de *Méthodes statistiques pour les sciences de la gestion* de Lind, Marchal, et Mason, Chenelière McGraw-Hill, 2006.) Le débat sur la relation qui existe entre le taux de chômage (TC) et le taux d'inflation (TI) est très contesté au sein de la communauté économique depuis longtemps. Un représentant de la Banque du Canada

vous demande d'estimer la mesure quantitative de cette relation. La base de données TCTI contient les observations trimestrielles du taux de chômage et du taux d'inflation de 1996 à 2000.

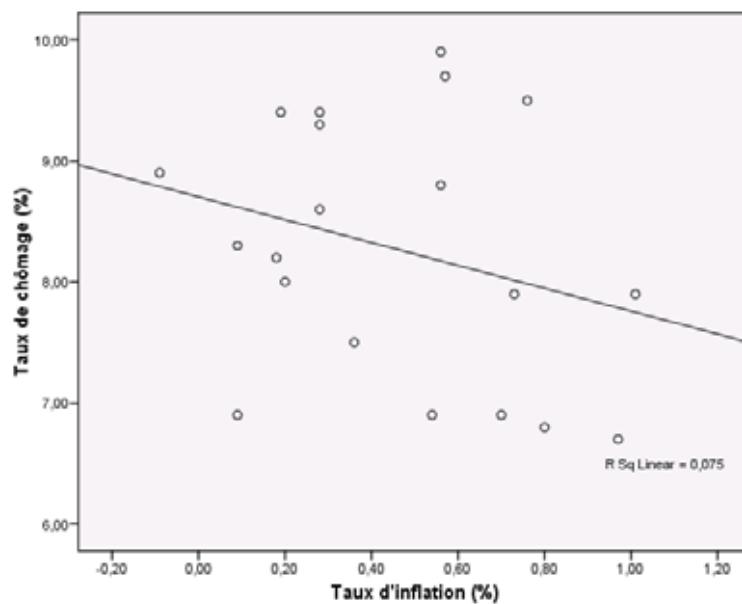


FIG. 6.6 – Le graphe de la relation entre le TC et le TI

Le graphe 6.6 nous montre que la relation entre le TC et le TI semble linéaire négative mais faible car la dispersion des points est très grande.

Correlations			
		TC	TI
TC	Pearson Correlation	1	-.274
	Sig. (2-tailed)		.242
	N	20	20
TI	Pearson Correlation	-.274	1
	Sig. (2-tailed)	.242	
	N	20	20

FIG. 6.7 – La corrélation entre le TC et le TI

La figure 6.7 nous montre que le coefficient de corrélation entre le TC et le TI est estimé à -0,274, donc la relation peut effectivement être qualifiée de faible. De plus, si on

teste

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

au seuil $\alpha = 0,05$, on conclut que cette corrélation n'est pas significativement différente de 0 puisque la p -value est de $0,242 > 0,05$. Donc à partir de cet échantillon nous ne trouvons pas de relation linéaire significative entre le TC et le TI.

Le calcul d'un coefficient de corrélation à la main est fastidieux ; cependant, une fois la valeur de celui-ci obtenue, il est simple de tester à la main s'il est significativement différent de zéro. Le tableau suivant résume les règles de décision selon les hypothèses.

Hypothèses statistiques	Écart réduit	Règle de décision
$H_0 : \rho = 0$ $H_1 : \rho \neq 0$	$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$	Rejeter H_0 si $T > t_{\alpha/2;n-2}$ ou $T < -t_{\alpha/2;n-2}$
$H_0 : \rho = 0$ $H_1 : \rho > 0$	$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$	Rejeter H_0 si $T > t_{\alpha;n-2}$
$H_0 : \rho = 0$ $H_1 : \rho < 0$	$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$	Rejeter H_0 si $T < -t_{\alpha;n-2}$

6.2 La régression linéaire simple

L'analyse en régression linéaire simple permet de faire l'étude des relations de dépendance entre deux variables continues. Dans cette technique, l'analyste doit être en mesure de spécifier quelle variable est dépendante (influençable) et quelle variable est indépendante (celle qui influence). La variable dépendante est, par convention, notée Y tandis que la variable indépendante est notée X. La relation étudiée est de la forme suivante :

$$X \Rightarrow Y$$

La régression linéaire simple est une technique statistique des plus performantes. Contrairement à d'autres techniques, l'analyse en régression aboutit à une formule mathématique qui modélise le lien entre les variables Y et X . En fait, l'ensemble des intervenants en statistiques mentionnent que l'on peut se servir de la technique de régression pour :

- étudier les liens entre les variables ;
- établir des prédictions ;
- prendre des décisions.

La formule obtenue ou l'équation de régression linéaire simple est de la forme suivante :

$$\hat{y}_i = b_0 + b_1 x_i$$

où i est l'indice associé aux réponses du i -ème individu. Cette équation sert à estimer celle que l'on retrouve au niveau de la population :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

L'objectif de l'analyse en régression consiste à utiliser l'information contenue dans la variable X afin d'améliorer nos prédictions sur les valeurs de Y .

Nous verrons dans les sous-sections qui suivent tous les détails concernant l'analyse en régression linéaire simple, et ceci en illustrant chaque étape à l'aide des données de l'exemple suivant :

Exemple 6.2.1 Reprenons l'exemple où un administrateur est chargé de planifier les ressources humaines d'un hôpital nouvellement construit. Il aimeraient connaître le nombre moyen d'employés à temps plein requis pour faire fonctionner un hôpital qui contiendra 30 lits. Par rapport à l'exemple 3.2.1, on a une information supplémentaire : le nombre de lits pour chacun des hôpitaux. On a donc les données suivantes :

	hopital	nb_emp_tp	nb_lit
1	1	69	23
2	2	95	29
3	3	102	29
4	4	118	35
5	5	126	42
6	6	125	46
7	7	138	50
8	8	178	54
9	9	156	64
10	10	184	66
11	11	176	76
12	12	225	78

FIG. 6.8 – Les données de l'exemple

L'analyse en régression permettra de modéliser le lien entre les variables X (`nb_lit`) et Y (`nb_emp_tp`). Plus précisément, nous voulons obtenir une équation du type

$$\hat{y}_{\text{nb_emp_tp}} = b_0 + b_1 x_{\text{nb_lit}}$$

Jetons un coup d'œil au graphe de la relation. Pour obtenir le graphe 6.10, il faut effectuer les commandes suivantes :

Menu SPSS :	→ Graphs
	→ Scatter...
Sélectionnez Simple ,	puis appuyez sur Define
Dans la fenêtre Y Axis :	→ <code>nb_emp_tp</code> (la variable dépendante)
Dans la fenêtre X Axis :	→ <code>nb_lit</code> (la variable indépendante)

Une fois le graphique obtenu dans la fenêtre **Output** de SPSS, on peut ajouter la droite de régression. Pour ce faire, il faut d'abord double-cliquer sur le graphe. La fenêtre **Chart Editor** s'ouvrira. Ensuite, il faut cliquer (une seule fois) en se positionnant sur l'un des points du graphe, n'importe lequel. Cette action aura pour effet de sélectionner tous

les points du graphe. Il suffit ensuite de cliquer sur l'icône `Add fit line` (c'est l'icône encerclé dans la sortie 6.9). Finalement, fermer les fenêtres `Properties` et `Chart Editor`. (Note : ces instructions concernent la version 12 de SPSS. Dans les autres versions, l'icône `Add fit line` est accessible sans avoir besoin de sélectionner les points.)

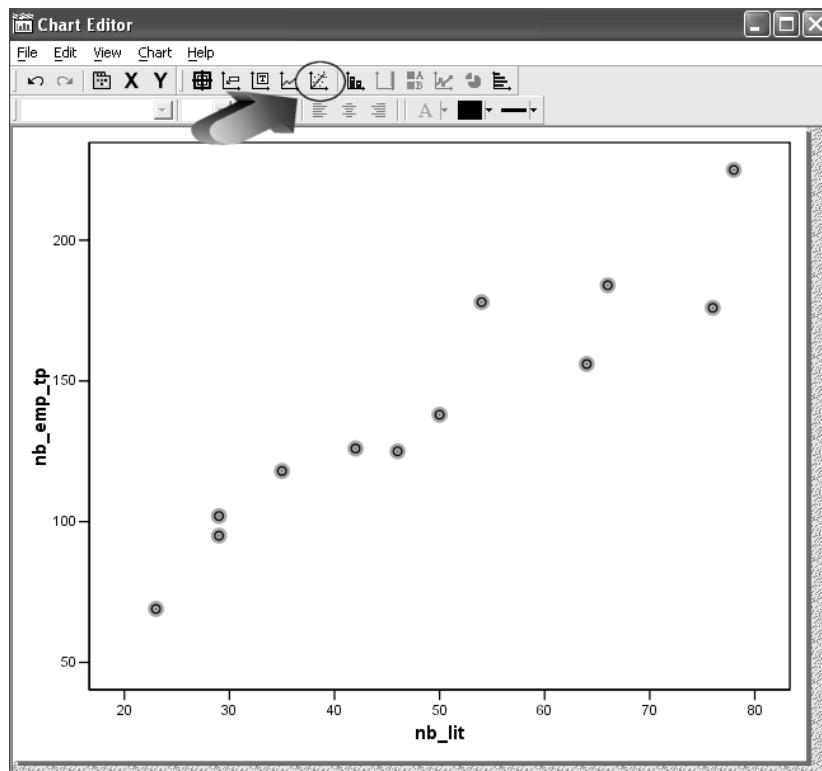


FIG. 6.9 – Pour ajouter la droite

On obtient alors la sortie 6.10, qui est le graphe de la relation.

Puisque les points semblent être répartis de façon uniforme autour de la droite, il est plausible d'affirmer que la relation entre X_{nb_lit} et $Y_{nb_emp_tp}$ est linéaire. On voit de plus que la relation est positive : plus le nombre de lits augmente, plus le nombre d'employés augmente.

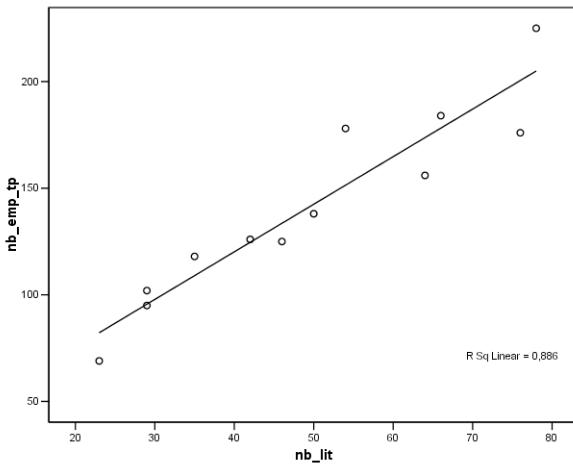


FIG. 6.10 – Le graphe de la relation

6.2.1 Le principe des moindres carrés

La droite que nous avons fait « apparaître » dans la figure 6.10 n'est bien sûr pas choisie au hasard. Elle doit être la meilleure possible pour estimer notre modèle. Mathématiquement, on trouve cette droite en utilisant la **méthode des moindres carrés** : elle permet de minimiser la somme des carrés des écarts entre les valeurs observées y_i et les valeurs de la droite $\hat{y} = b_0 + b_1 x$. Cette façon de faire permet de minimiser l'espace vertical (donc associé aux valeurs que prend Y) entre les points et la droite. Intuitivement, ceci permet de positionner la droite « au milieu » du nuage de points.

Autrement dit, on veut que chaque $e_i = y_i - \hat{y}_i$ soit le plus petit possible. Il faut donc minimiser l'expression suivante :

$$\begin{aligned} SSE &= e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^n e_i^2 \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2 \end{aligned}$$

où SSE signifie *sum of squared errors*.

La droite qui minimise SSE est appelée **droite de régression**. La méthode des

moindres carrés nous donne les valeurs de b_0 et b_1 :

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad b_0 = \bar{y} - b_1 \bar{x}.$$

Les coefficients b_0 et b_1 nous permettent d'écrire la droite $\hat{y}_i = b_0 + b_1 x_i$, qui sert à estimer le modèle que l'on retrouverait au niveau de la population : $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Pour notre exemple, c'est la sortie 6.11 qui nous donne les coefficients de la droite. Pour obtenir cette sortie (ainsi que les autres sorties associées à une régression), il faut effectuer les commandes suivantes :

Menu SPSS :	→ Analyse
	→ Regression
	→ Linear...
Dans la fenêtre Dependant :	→ nb_emp_tp (la variable dépendante)
Dans la fenêtre Independant(s) :	→ nb_lit (la variable indépendante)

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	30,912	13,254		2,332	,042
nb_lit	2,232	,253	,942	8,835	,000

a. Dependent Variable: nb_emp_tp

FIG. 6.11 – La sortie qui contient les coefficients de la droite

Le coefficient b_0 est la constante de la droite, et se retrouve dans la première colonne du tableau vis-à-vis (**Constant**). Ici on a $b_0 = 30,912$. Le coefficient b_1 est lui aussi

dans la première colonne, vis-à-vis la variable indépendante qui ici est `nb_lit`. On a $b_1 = 2,232$. L'équation de la droite est donc

$$\hat{y}_{\text{nb_emp_tp}} = b_0 + b_1 x_{\text{nb_lit}} = 30,912 + 2,232 x_{\text{nb_lit}}.$$

À quel point cette régression arrive-t-elle à modéliser le lien entre Y et X ? Comment interpréter cette droite et le modèle qu'elle représente? C'est ce que nous verrons dans les sous-sections qui suivent.

6.2.2 Le modèle statistique de la régression

Au niveau de la population, le modèle de la régression linéaire simple s'écrit :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Ce modèle signifie que pour chaque couple d'observations (X_i, Y_i) , on peut exprimer Y_i en fonction de X_i (donc la connaissance de la valeur de X_i nous permet de connaître la valeur de Y_i). La première partie de l'équation $(\beta_0 + \beta_1 X_i)$ est l'équation d'une droite. Comme il est impensable que tous les points se retrouvent sur la droite (c'est-à-dire qu'il est à peu près impossible de tomber sur un modèle linéaire parfait, on se retrouve plutôt avec un nuage de points), le terme ϵ_i est ajouté : il représente « l'erreur », c'est-à-dire la distance verticale entre le point et la droite.

Dans l'expression **modèle linéaire simple**, le terme **simple** indique la présence d'une seule variable explicative X (lorsqu'il y a plusieurs variables explicatives X_1, X_2, \dots , on parle alors de régression linéaire **multiple**).

La signification du terme **linéaire** ne se réfère qu'aux paramètres du modèle, soient β_0 et β_1 . Ainsi, les modèles

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \text{ et } Y_i = \beta_0 + \beta_1 X_i^2 + \epsilon_i$$

sont des modèles linéaires. Cependant, le modèle

$$Y_i = \beta_0 \cdot \beta_1^{X_i} + \epsilon_i$$

est non linéaire puisque nous sommes en présence de produits de paramètres. (Ce modèle peut être linéarisé en utilisant une transformation logarithmique :

$$\log(a \cdot b) = \log(a) + \log(b) \text{ et } \log(ax) = x \log(a).$$

Il est aussi fréquent de caractériser la variable X en identifiant l'ordre de l'expression mathématique du modèle utilisé. Ainsi, le modèle

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

est linéaire d'ordre 1 tandis que le modèle suivant est linéaire d'ordre 2 :

$$Y_i = \beta_0 + \beta_1 X_i^2 + \epsilon_i.$$

Étant donnée les N couples d'observations de la population $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$, et en supposant que la relation plausible entre Y et X est linéaire d'ordre 1, le modèle de régression linéaire simple s'écrit

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

pour $i = 1, 2, \dots, N$ où N est la taille de la population. Ce modèle est tel que :

- Y est la variable dépendante (ou expliquée) ayant un caractère aléatoire et dont les valeurs sont conditionnées par celles de la variable explicative X et une composante aléatoire ϵ .
- β_0 et β_1 sont appelés les paramètres du modèle de régression. Nous ne les connaissons pas, nous les estimons donc à l'aide de l'échantillon par b_0 et b_1 respectivement.
- X est la variable explicative (ou indépendante), mesurée **sans erreur** ou dont les valeurs sont fixées avant l'expérience à des valeurs arbitraires.
- ϵ dénote la fluctuation aléatoire, non observable (ou inexplicable) attribuable à un ensemble de facteurs qui ne sont pas pris en considération dans le modèle. Cette fluctuation aléatoire n'est pas expliquée par le modèle et se reflète sur la variable dépendante Y .

Pour pouvoir estimer ce modèle à l'aide d'un échantillon (et s'en servir), il faut que les hypothèses suivantes soient respectées :

On suppose que les erreurs ϵ_i sont indépendantes et suivent une loi normale de moyenne $E(\epsilon_i) = 0$ et de variances identiques $\text{Var}(\epsilon_i) = \sigma_{\text{résiduelle}}^2$ pour **toutes** les valeurs de X .

Ceci revient à dire qu'on a

$$\epsilon_i \sim N(0, \sigma_{\text{résiduelle}}^2)$$

pour toute valeur X_i de la population.

FIG. 6.12 – Les hypothèses de validité du modèle de régression linéaire d'ordre 1

On peut déduire de ceci que, pour toute valeur particulière X_i , la variable dépendante Y_i est une variable aléatoire distribuée selon une loi normale de moyenne $E(Y_i) = \beta_0 + \beta_1 X_i$ et de variance $\text{Var}(Y_i) = \sigma_{\text{résiduelle}}^2$. C'est-à-dire qu'on a

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma_{\text{résiduelle}}^2).$$

Pour qu'une droite de régression soit considérée valide (et donc utilisable pour faire des estimations), les hypothèses de la figure 6.12 doivent être vérifiées. Nous verrons plus tard comment les vérifier dans la pratique.

6.2.3 Les paramètres β_0 et β_1

On a vu qu'au niveau de la population, la droite de régression modélisant le lien entre les variables Y et X s'écrit

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Pour estimer ce modèle à partir d'un échantillon, on utilise la méthode des moindres carrés (sous-section 6.2.1) qui nous donne l'équation suivante :

$$\hat{y}_i = b_0 + b_1 x_i.$$

Ainsi, les statistiques b_1 et b_0 sont les estimateurs ponctuels des paramètres β_1 et β_0 . Les calculs de b_1 et b_0 sont toujours effectués par le logiciel. Mais quelles en sont les interprétations ?

- Dans la régression, b_1 (la pente) représente le taux d'accroissement marginal moyen de la variable dépendante Y pour une augmentation unitaire de X :

$$\hat{y}_i = b_0 + b_1(x_i + 1) = b_0 + b_1x_i + b_1.$$

- b_0 est l'ordonnée à l'origine (valeur de \hat{y} lorsque $x = 0$). L'interprétation n'en est possible que lorsque $x = 0$ est une observation possible ou possédant un certain sens. Lorsqu'elle est interprétable, elle représente souvent des frais ou dépenses fixes.

Exemple 6.2.2 Revenons à notre exemple. On a trouvé que la droite s'écrit

$$\hat{y}_{\text{nb_emp_tp}} = b_0 + b_1x_{\text{nb_lit}} = 30,912 + 2,232x_{\text{nb_lit}}.$$

Ici on pourrait dire que le coefficient $b_0 = 30,912$ représente le nombre d'employés minimum requis pour faire fonctionner un hôpital lorsqu'il y a 0 lit (même si 0 lit est un concept insensé). On peut imaginer que ceci correspond au personnel administratif et de soutien.

Le coefficient b_1 représente l'augmentation marginale moyenne du nombre d'employés à temps plein lorsque le nombre de lits augmente d'une unité. Ainsi environ 2,232 employés s'ajoutent lorsqu'un lit s'ajoute.

Propriétés de b_0 et b_1

Les estimateurs b_0 et b_1 obtenus de la méthode des moindres carrés possèdent trois propriétés qui font de ces derniers les meilleurs estimateurs linéaires existants pour estimer les paramètres d'une droite de régression :

1. Ce sont des estimateurs **linéaires** : en effet, en considérant que les valeurs de X sont des constantes, on peut écrire b_0 et b_1 de façon à montrer que ce sont des combinaisons linéaires des Y_i .
2. Ce sont des estimateurs **sans biais**, c'est-à-dire que $E(b_0) = \beta_0$ et $E(b_1) = \beta_1$.
3. Ces estimateurs sont de **variance minimale** : il est impossible que d'autres estimateurs linéaires et sans biais aient des variances plus petites que celles de b_0 et b_1 (pour estimer β_0 et β_1 respectivement). Autrement dit, les intervalles de confiance pour estimer β_0 et β_1 seront les plus précis possibles.

On peut résumer ceci en disant qu'aucune autre droite ne minimisera mieux l'incertitude. Ces estimateurs sont dits de type BLUE.

Les distributions d'échantillonnage de b_0 et b_1

Afin de pouvoir établir des intervalles de confiance ou des tests d'hypothèses sur un paramètre (β_0 et β_1 dans notre cas), nous devons préciser la distribution d'échantillonnage de l'estimateur ponctuel correspondant (b_0 et b_1 dans notre cas). Il faut donc connaître la forme, la moyenne (l'espérance) et la variance de chacun d'eux.

La distribution d'échantillonnage de b_1

L'estimateur b_1 a une distribution d'échantillonnage qui suit une loi normale d'espérance β_1 et de variance $\frac{\sigma_{\text{résiduelle}}^2}{(n - 1)s_x^2}$ où $\sigma_{\text{résiduelle}}^2$ est la variance des résidus ϵ_i .

Au niveau de l'échantillon, la variance résiduelle $s_{\text{résiduelle}}^2$ est une estimation de la dispersion des valeurs de y autour de la droite de régression $\hat{y} = b_0 + b_1x$, et sert à

estimer $\sigma_{\text{résiduelle}}^2$. Elle se calcule de la façon suivante :

$$s_{\text{résiduelle}}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{(1-r^2)s_y^2(n-1)}{n-2}.$$

L'écart-type résiduel est la racine carrée de la variance résiduelle : $s_{\text{résiduelle}} = \sqrt{s_{\text{résiduelle}}^2}$.

Ainsi, pour la distribution d'échantillonnage de b_1 , nous utilisons $s_{\text{résiduelle}}^2$ pour estimer $\sigma_{\text{résiduelle}}^2$. Le fait d'utiliser un estimateur nous amène à modéliser la distribution d'échantillonnage de b_1 avec la loi de Student qui est plus conservatrice.

Ainsi l'intervalle de confiance de niveau $1 - \alpha$ pour le paramètre β_1 est le suivant :

$$\left[b_1 - t_{(n-2);\alpha/2} \sqrt{\frac{\sigma_{\text{résiduelle}}^2}{(n-1)s_x^2}}, \quad b_1 + t_{(n-2);\alpha/2} \sqrt{\frac{\sigma_{\text{résiduelle}}^2}{(n-1)s_x^2}} \right].$$

(On retrouve $n - 2$ degrés de liberté pour la loi de Student car deux paramètres (β_0 et β_1) doivent être estimés dans le modèle de régression.)

Test statistique sur β_1

Si la droite $Y = \beta_0 + \beta_1 X + \epsilon$ est significative (c'est-à-dire si elle établit un lien linéaire explicatif entre X et Y), alors nécessairement $\beta_1 \neq 0$. On veut donc tester les hypothèses

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Pour résoudre ce test, on calcule l'écart-réduit

$$T = \frac{b_1}{\left(\sqrt{\frac{\sigma_{\text{résiduelle}}^2}{(n-1)s_x^2}} \right)}$$

et au seuil α on rejette H_0 si $T > t_{\alpha/2;n-2}$ ou $T < -t_{\alpha/2;n-2}$.

La distribution d'échantillonnage de b_0

Tout comme l'estimateur b_1 , la distribution d'échantillonnage de b_0 suit une loi normale. Son espérance est β_0 , et sa variance est

$$\sigma^2(b_0) = \sigma_{\text{résiduelle}}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2} \right].$$

Ici encore on utilisera $s_{\text{résiduelle}}^2$ pour estimer $\sigma_{\text{résiduelle}}^2$. Et on utilisera la loi de Student pour modéliser la distribution d'échantillonnage de b_0 .

Ainsi l'intervalle de confiance de niveau $1 - \alpha$ pour le paramètre β_0 est le suivant :

$$[b_0 - t_{(n-2);\alpha/2}s(b_0), b_0 + t_{(n-2);\alpha/2}s(b_0)]$$

où

$$s(b_0) = \sqrt{s_{\text{résiduelle}}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2} \right]}.$$

Test statistique sur β_0

Si la droite $Y = \beta_0 + \beta_1 X + \epsilon$ passe par l'origine, alors $\beta_0 = 0$. On peut donc vouloir tester les hypothèses

$$H_0 : \beta_0 = 0$$

$$H_1 : \beta_0 \neq 0$$

Pour résoudre ce test, on calcule l'écart-réduit

$$T = \frac{b_0}{s(b_0)}$$

et au seuil α on rejette H_0 si $T > t_{\alpha/2;n-2}$ ou $T < -t_{\alpha/2;n-2}$.

6.2.4 Analyse de la variance

Le modèle de régression linéaire simple nous permet d'identifier les variables X qui peuvent contribuer de façon importante à expliquer les fluctuations dans les observations de la variable dépendante Y .

L'analyse de la variance va nous permettre de :

- Quantifier la variation totale dans les observations et la décomposer en deux sources de variation : expliquée et résiduelle.
- Vérifier à l'aide d'un tableau d'analyse de la variance si la source de fluctuation attribuable à la régression est significative.
- Définir un indice qui donne une mesure descriptive de la qualité de l'ajustement des points expérimentaux par la droite de régression (r^2).

La droite de régression décompose la variation de la variable Y en deux sources :

$$\begin{array}{lcl} \text{Variation de } Y & = & \text{Variation expliquée} + \text{Variation} \\ & & \text{par la régression} \qquad \qquad \qquad \text{résiduelle} \end{array}$$

$$\begin{array}{lcl} \sum_{i=1}^n (y_i - \bar{y})^2 & = & \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \text{SCT} & = & \text{SCR} + \text{SC}_{\text{rés}} \end{array}$$

La **proportion de la variation totale expliquée par la droite de régression** est le **coefficient de détermination r^2** :

$$r^2 = \frac{\text{variation expliquée (SCR)}}{\text{variation totale (SCT)}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Ainsi r^2 est le pourcentage de la variation totale de y expliquée par la droite de régression. Par conséquent r^2 varie entre 0 et 1, et plus r^2 est près de 1, meilleure est la droite de régression.

Le coefficient de corrélation r est la racine carrée de r^2 , et son signe est le même que celui de b_1 .

Une table d'analyse de la variance se présente toujours sous la forme suivante :

Source de variation	Somme de carrés	Degrés de liberté	Carrés moyens	Quotient F
Expliquée par la régression	$SCR = \sum(\hat{y}_i - \bar{y})^2$	1	$CMR = SCR/1$	$F = \frac{CMR}{CM_{rés}}$
Résiduelle	$SC_{rés} = \sum(y_i - \hat{y}_i)^2$	$n - 2$	$CM_{rés} = SC_{rés}/n - 2$	
Totale	$SCT = \sum(y_i - \bar{y})^2$	$n - 1$		

où $F = \frac{CMR}{CM_{rés}}$ suit une loi de Fisher $F_{1; n-2}$. Cette statistique nous permet de tester les hypothèses suivantes :

H_0 : Au niveau de la population, la régression n'est pas significative dans son ensemble ($\beta_1 = 0$).

H_1 : Au niveau de la population, la régression est significative dans son ensemble ($\beta_1 \neq 0$).

Au seuil α , on rejette H_0 si $F > F_{\alpha;1,n-2}$.

Remarques :

- Une somme de carrés divisée par son nombre de degrés de liberté donne toujours une variance. Voilà pourquoi $CM_{rés}$ est une variance. En fait $CM_{rés}$ est la variance échantillonnale du résidu ($s_{résiduelle}^2$).
- Par définition, un quotient de variances suit toujours une loi de Fisher de paramètres (DL numérateur, DL dénominateur).

Exemple 6.2.3 Revenons à notre exemple. À l'aide de la sortie 6.13 dans laquelle on retrouve le r et le r^2 , on décrit la force du lien linéaire entre les deux variables. Tout d'abord, dans la première colonne, on retrouve le coefficient de corrélation r . Ici on a

$r = 0,942$, ce qui nous indique que nous sommes en présence d'une relation linéaire très forte (voir le schéma de Davis, figure 6.1).

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,942 ^a	,886	,875	15,649
a. Predictors: (Constant), nb_lit				

FIG. 6.13 – La sortie qui contient le r et le r^2

Ensuite, afin de mesurer l'apport d'information qu'amène X sur la prédiction Y , nous utilisons le coefficient de détermination r^2 , qui représente le % de la variation totale de Y expliquée par la présence de la variable X . Celui-ci se retrouve dans la deuxième colonne de la sortie 6.13. Ici on a $r^2 = 0,886$, ce qui nous indique que 88,6 % de la variation du nombre d'employés à temps plein (Y) est expliquée lorsque le nombre de lits (X) est pris en considération.

La prochaine étape consiste à traiter le test d'hypothèses suivant :

H_0 : La régression est non significative dans la population ($\beta_1 = 0$).

H_1 : La régression est significative dans la population ($\beta_1 \neq 0$).

Pour ce faire, on utilise la sortie 6.14 (table ANOVA). Tel que vu précédemment, une analyse en régression linéaire décompose la variation totale de Y en deux sources :

$$\begin{aligned} \text{Variation totale} &= \text{Variation expliquée par la droite} + \text{Variation résiduelle} \\ 21\ 564,000 &= 19\ 115,063 + 2\ 448,937 \end{aligned}$$

On peut calculer le r^2 à partir de ces variations :

$$r^2 = \frac{\text{Variation expliquée par la droite}}{\text{Variation totale}} = \frac{19115,063}{21564,000} = 0,886.$$

Plus la variation expliquée par la droite sera grande, plus la régression risque d'être significative. Pour résoudre le test d'hypothèses, on utilise la *p*-value (Sig.) de la dernière colonne sur laquelle repose la règle de décision habituelle :

Nous rejetons l'hypothèse H_0 si la *p*-value est plus petite que le seuil de signification α fixé (par exemple $\alpha = 0,05$). Sinon, nous ne rejetons pas H_0 et la considérons comme vraisemblable.

Ici, puisque la *p*-value est égale à 0,000, ce qui est plus petit que $\alpha = 0,05$, on rejette H_0 . Ainsi, au risque de se tromper une fois sur 20, on peut affirmer que la régression est significative.

ANOVA^b					
Model		Sum of Squares	df	Mean Square	F
1	Regression	19115,063	1	19115,063	78,055
	Residual	2448,937	10	244,894	
	Total	21564,000	11		

a. Predictors: (Constant), nb_lit
b. Dependent Variable: nb_emp_tp

FIG. 6.14 – La table ANOVA de la régression

6.2.5 Utilisation de la droite de régression

Après avoir statué qu'une régression est significative, nous disposons alors d'une équation qui modélise le lien qui existe entre les variables Y et X . Il faut bien comprendre qu'il faut connaître les valeurs de X pour utiliser ce modèle. Ainsi, avec ce modèle, nous sommes en mesure de faire les choses suivantes :

- Faire des estimations sur des valeurs moyennes de Y .
- Faire des intervalles de confiance autour des valeurs moyennes de Y .

- Faire de la prévision sur des valeurs précises de Y .
- Faire des intervalles de prévision autour des valeurs précises de Y .

Estimation d'une valeur moyenne de Y

Pour estimer une valeur moyenne de Y selon une valeur fixée de X , il suffit d'utiliser l'équation obtenue. Dans le cadre de notre exemple, l'analyste est en mesure d'utiliser la droite pour faire des estimations sur des valeurs moyennes de Y , compte tenu de la connaissance de X . Par exemple, pour obtenir une estimation du nombre d'employés requis pour un hôpital de 30 lits, il suffit de remplacer $x_{\text{nb_lit}}$ par 30 :

$$\hat{y}_{\text{nb_emp_tp}} = b_0 + b_1 x_{\text{nb_lit}} = 30,912 + 2,232 \times 30 = 97,87.$$

Il faudrait donc prévoir environ 98 employés à temps plein pour un hôpital de 30 lits. Cette estimation ne tient cependant pas compte de l'erreur induite par l'échantillon. Il serait donc plus approprié de construire un intervalle de confiance pour cette prédiction.

Estimation d'une valeur moyenne de Y par intervalle de confiance

Rappelons qu'initialement Y était une variable aléatoire de loi souvent inconnue, qui possédait une variation que nous estimions à l'aide de s_y^2 (estimateur de σ_y^2). Nous avons vu que, lorsqu'une régression est significative entre Y et une autre variable X , l'une des hypothèses du modèle de régression linéaire simple nous permet de dire que

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma_{\text{résiduelle}}^2).$$

L'estimation de $\sigma_{\text{résiduelle}}^2$ se fait par l'entremise de $s_{\text{résiduelle}}^2 = \text{CM}_{\text{rés}}$ qui se trouve dans la table ANOVA.

Ici on s'intéresse à la valeur **moyenne** de Y_i lorsque X_i est fixé. Cette valeur moyenne est donnée par $\hat{y}_i = b_0 + b_1 x_i$.

Pour une valeur fixée x_i , l'intervalle de confiance de niveau $1 - \alpha$ pour la valeur **moyenne** de y est

$$[\hat{y}_i - t_{\alpha/2;n-2} s(\hat{y}_i), \hat{y}_i + t_{\alpha/2;n-2} s(\hat{y}_i)]$$

où $s(\hat{y}_i) = s_{\text{résiduelle}} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_j - \bar{x})^2}}$.

Estimation d'une valeur réelle de Y

L'objectif d'une étude de régression est non seulement d'obtenir des estimations de la moyenne des Y_i , c'est-à-dire de $E(Y_i)$ pour diverses valeurs de X_i , mais également de fournir des prévisions concernant les valeurs éventuelles de la variable dépendante Y .

De façon ponctuelle, l'estimation d'une valeur réelle s'obtient de la même façon que pour une valeur moyenne, soit directement de la droite.

Pour l'intervalle de confiance, cependant, il y a une différence. Il y a plus de fluctuations entre des valeurs réelles qu'entre des valeurs moyennes, et ainsi l'intervalle de confiance pour une valeur réelle sera moins précis (plus large) que l'intervalle de confiance pour une valeur moyenne.

Ainsi, pour une valeur fixée X_i , l'intervalle de confiance de niveau $1 - \alpha$ pour une **unique valeur réelle** de Y est

$$[\hat{y}_i - t_{\alpha/2;n-2} s(d_i), \hat{y}_i + t_{\alpha/2;n-2} s(d_i)]$$

avec $d_i = y_i - \hat{y}_i$ et

$$s(d_i) = \sqrt{s_{\text{résiduelle}}^2 + s^2(\hat{y}_i)} = s_{\text{résiduelle}} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_j - \bar{x})^2}}.$$

Exemple 6.2.4 Reprenons l'exemple de ce chapitre. Pour obtenir les estimations à partir de SPSS, il faut d'abord entrer la valeur de la variable indépendante pour laquelle on veut une prédiction **directement dans la base de données**, dans la colonne de cette variable, et dans une ligne inutilisée (voir la figure 6.15).

	hopital	nb_emp_tp	nb_lit
1	1	69	23
2	2	95	29
3	3	102	29
4	4	118	35
5	5	126	42
6	6	125	46
7	7	138	50
8	8	178	54
9	9	156	64
10	10	184	66
11	11	176	76
12	12	225	78
13	.	.	30
1.1			

FIG. 6.15 – Saisie de la valeur pour laquelle on veut une prédiction

On effectue ensuite les commandes pour une régression linéaire, mais avec quelques ajouts :

Menu SPSS :	→ Analyse
	→ Regression
	→ Linear...
Dans la fenêtre Dependant :	→ nb_emp_tp (la variable dépendante)
Dans la fenêtre Independant(s) :	→ nb_lit (la variable indépendante)
Dans le bouton Save... :	→ Predicted Values
	✓ Unstandardized
	→ Prediction Intervals
	✓ Mean ✓ Individual
	(et préciser le niveau de confiance voulu)

On retrouve ensuite les résultats dans la base de données et non dans une fenêtre Output comme à l'habitude (voir la figure 6.16).

	hopital	nb_emp_tp	nb_lit	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	1	69	23	82,23706	64,32205	100,15207	43,03569	121,43843
2	2	95	29	95,62609	80,38585	110,86632	57,57264	133,67953
3	3	102	29	95,62609	80,38585	110,86632	57,57264	133,67953
4	4	118	35	109,01511	96,11604	121,91418	71,83735	146,19287
5	5	126	42	124,63564	113,75679	135,51449	88,10963	161,16164
6	6	125	46	133,56165	123,32272	143,80059	97,22110	169,90220
7	7	138	50	142,48767	132,41506	152,56028	106,19363	178,78171
8	8	178	54	151,41369	141,01108	161,81629	115,02668	187,80069
9	9	156	64	173,72873	160,71152	186,74593	136,50981	210,94764
10	10	184	66	178,19173	164,43326	191,95021	140,70713	215,67534
11	11	176	76	200,50677	182,43627	218,57727	161,23410	239,77945
12	12	225	78	204,96978	185,95416	223,98540	165,25337	244,68620
13	.	.	30	97,85759	83,03527	112,67991	59,96958	135,74560
14								

FIG. 6.16 – Les prédictions

Les 3 premières colonnes de la figure 6.16 sont celles des variables. La colonne suivante (PRE_1) contient les prédictions ponctuelles. Ainsi vis-à-vis la dernière ligne dans laquelle on a entré la valeur de 30 lits, on retrouve la prédition calculée plus tôt de 97,9 employés.

Les deux colonnes suivantes (LMCI_1 et UMCI_1) sont les bornes de l'intervalle de confiance de niveau 95 % qui veut estimer quelle serait la **moyenne** d'employés à temps plein pour **des** hôpitaux de 30 lits. Ainsi le nombre moyen d'employés à temps plein pour des hôpitaux de 30 lits devrait être compris entre 83,03 et 112,68, et ce 19 fois sur 20.

Les deux dernières colonnes (LICI_1 et UICI_1) sont les bornes de l'intervalle de confiance de niveau 95 % qui veut estimer quelle serait le **nombre réel** d'employés à temps plein pour **un** hôpital de 30 lits. Ainsi le nombre réel d'employés à temps plein pour un hôpital de 30 lits devrait être compris entre 59,97 et 135,75, et ce 19 fois sur 20.

Quelques éléments doivent être pris en considération lors de l'utilisation de la droite pour effectuer des estimations ou des prédictions :

- Pour l'estimation de Y , les valeurs de b_0 et de b_1 ne sont valides que pour les variations observées (sans erreur) de la variable X (c'est-à-dire pour des valeurs de X comprises entre le minimum et le maximum de cette variable dans l'échantillon).

Ce phénomène s'appelle **l'intrapolation**. En particulier, si vous n'avez pas de

valeurs de X près de 0 dans votre échantillon, l'interprétation de b_0 aura bien peu de sens.

- Un phénomène à utiliser avec grande précaution consiste à faire des estimations de Y à partir de valeurs de X se situant en dehors du domaine de validité observé dans l'échantillon. Cela s'appelle faire de l'**extrapolation**.

Exemple 6.2.5 Une société de transport veut établir une politique d'entretien des camions de sa flotte. Tous ses camions sont de même modèle et utilisés à des transports semblables. La direction de la société est d'avis qu'une liaison statistique entre le coût direct de déplacement (cents par km) et l'espace de temps écoulé depuis la dernière inspection de ce camion serait utile. La base de données se nomme `camions.sav`.

	ident	cout	mois
1	1	10	3
2	2	18	7
3	3	24	10
4	4	22	9
5	5	27	11
6	6	13	6
7	7	10	5
8	8	24	8
9	9	25	7
10	10	8	4
11	11	16	6
12	12	20	9
13	13	28	12
14	14	22	8
15	15	19	10
16	16	18	9
17	17	26	11
18	18	14	6
19	19	20	8
20	20	26	10
21	21	30	12
22	22	12	5

FIG. 6.17 – Les données de l'exemple

On se demande si ce lien est linéaire. Regardons donc tout d'abord le graphe de la

relation (figure 6.18). Puisque les points semblent être répartis de façon assez uniforme autour de la droite, il est plausible d'affirmer que la relation entre X_{mois} et Y_{cout} est linéaire. On voit de plus que la relation est positive : plus le nombre de mois depuis la dernière inspection augmente, plus le coût augmente.

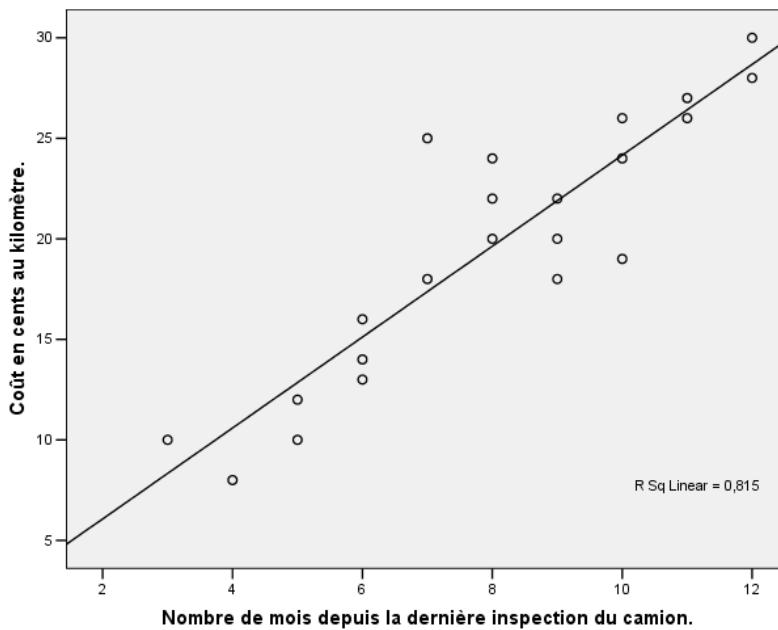


FIG. 6.18 – Graphe de la relation $\text{mois} \Rightarrow \text{cout}$

Ensuite, on voit que $r = 0,903$ (figure 6.19), ce qui indique que la relation linéaire est très forte. Aussi, on a $r^2 = 0,815$. Par conséquent, 81,5 % de la variation des coûts est expliquée par le nombre de mois depuis la dernière inspection.

La table ANOVA nous permet de résoudre le test d'hypothèses suivant :

H_0 : La régression est non significative dans la population ($\beta_1 = 0$).

H_1 : La régression est significative dans la population ($\beta_1 \neq 0$).

Ici, puisque la p -value est égale à 0,000, ce qui est plus petit que $\alpha = 0,05$, on rejette H_0 . Ainsi, au risque de se tromper une fois sur 20, on peut affirmer que la régression est

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,903 ^a	,815	,806	2,826

a. Predictors: (Constant), Nombre de mois depuis la dernière inspection du camion.

FIG. 6.19 – Le r et le r^2 ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	705,391	1	705,391	88,340	,000 ^a
	Residual	159,700	20	7,985		
	Total	865,091	21			

a. Predictors: (Constant), Nombre de mois depuis la dernière inspection du camion.

b. Dependent Variable: Coût en cents au kilomètre.

FIG. 6.20 – La table ANOVA

significative.

Finalement, on peut écrire la droite de régression. On obtient

$$\hat{y}_{\text{cout}} = b_0 + b_1 x_{\text{mois}} = 1,549 + 2,261 x_{\text{mois}}.$$

Ainsi l'augmentation moyenne du coût de transport est de 2,261 cents du km pour chaque mois supplémentaire depuis la dernière inspection.

La valeur de b_0 s'interprète bien ici : le coût de déplacement d'un camion qui vient tout juste d'être inspecté est de 1,549 cents du km.

On voit que dans la table des coefficients, on a les intervalles de confiance de niveau 95 % pour les paramètres de la droite. Voici les commandes pour faire apparaître ces intervalles :

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	1,549	2,016		,768	,451	-2,657	5,756
Nombre de mois depuis la dernière inspection du camion.	2,261	,241	,903	9,399	,000	1,759	2,763

a. Dependent Variable: Coût en cents au kilomètre.

FIG. 6.21 – La table des coefficients

-
- | | |
|----------------------------------|---|
| Menu SPSS : | → Analyse
→ Regression
→ Linear... |
| Dans la fenêtre Dependant : | → cout (la variable dépendante) |
| Dans la fenêtre Independant(s) : | → mois (la variable indépendante) |
| Dans le bouton Statistics... : | → Regression Coefficients
<input checked="" type="checkbox"/> Confidence Intervals |
-

On voit donc que l'intervalle de confiance de niveau 95 % pour β_1 est [1,759, 2,763]. Ceci signifie qu'au niveau de la population, il y a une probabilité de 95 % que l'augmentation du coût de transport soit entre 1,759 et 2,763 cents du km pour chaque mois supplémentaire depuis la dernière inspection.

Supposons que l'on veuille estimer le coût de transport d'un camion qui a eu sa dernière inspection voilà 9 mois. Tout d'abord, la droite nous donne l'estimation ponctuelle de ce coût :

$$\hat{y}_{\text{cout}} = 1,549 + 2,261x_{\text{mois}} = 1,549 + 2,261 \cdot 9 = 21,9.$$

Donc l'estimation du coût moyen est de 21,9 cents du km pour un camion qui a eu sa dernière inspection voilà 9 mois. Quels sont les intervalles de confiance de niveau

95 % pour cette estimation au niveau de la population ? La figure 6.22 nous en donne les bornes.

	ident	cout	mois	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
19	19	20	8	19,63636	18,37966	20,89307	13,60943	25,66329
20	20	26	10	24,15810	22,54988	25,76633	18,04819	30,26801
21	21	30	12	28,67984	26,31179	31,04789	22,32750	35,03218
22	22	12	5	12,85375	10,89282	14,81469	6,64168	19,06583
23	.	.	9	21,89723	20,54406	23,25040	15,84945	27,94502

FIG. 6.22 – Les estimations

Tout d'abord, dans la colonne PRE_1, on retrouve notre estimation ponctuelle calculée ci-haut (21,9). Les colonnes LMCI_1 et UMCI_1 nous donnent l'intervalle de confiance pour le coût **moyen** : ainsi, au niveau de la population, il y a une probabilité de 95 % que le coût moyen de déplacement de plusieurs camions dont la dernière inspection remonte à 9 mois soit entre 20,54 et 23,25 cents du kilomètre. Par contre, si je m'intéresse au coût réel de déplacement d'**un** camion dont la dernière inspection remonte à 9 mois, je dois prendre les bornes LICI_1 et UICI_1 : ainsi au niveau de la population, il y a une probabilité de 95 % que ce coût réel soit entre 15,85 et 27,95 cents du kilomètre.

Exemple 6.2.6 On veut calculer le risque bêta des actions de Bell Canada (BCE) à partir des données sur le cours de clôture des actions de BCE et du S&P/TSX le dernier vendredi de chaque mois pendant 20 mois (novembre 1999 à mai 2001). Les données figurent dans le tableau 6.23.

Date	S&P/TSX (X)	BCE (Y)	Date	S&P/TSX (X)	BCE (Y)
25 mai	8 292,84	39,75	29 juill.	10 342,98	34,60
27 avr.	7 967,34	39,38	30 juin	10 195,45	35,10
30 mars	7 608,00	35,44	26 mai	9 020,88	32,65
23 févr.	8 028,81	39,45	28 avr.	9 347,61	41,66
26 janv.	9 158,19	42,50	31 mars	9 462,39	43,97
29 déc.	8 933,68	43,30	25 fév.	9 141,17	39,17
24 nov.	9 024,43	42,25	28 janv.	8 390,40	34,56
27 oct.	9 321,89	39,90	31 déc.	8 413,75	31,86
29 sept.	10 377,92	35,05	26 nov.	7 889,94	26,48
25 août	11 246,04	33,30			

FIG. 6.23 –

Après avoir calculé le taux de rendement à partir des données du tableau, on obtient le diagramme de dispersion (figure 6.24). La pente semble assez faible, on s'attend donc à un bêta de moins de 1. Cependant la dispersion est forte, ce qui diminue la fiabilité de nos estimations.

On voit dans la figure 6.25 que $r = 0,341$, donc la relation entre les deux taux est qualifiée de modérée. Le r^2 de 0,116 nous indique que d'après cet échantillon, 11,6 % du taux de variation de BCE est expliqué par le taux de variation du S&P/TSX, ce qui n'est pas très élevé.

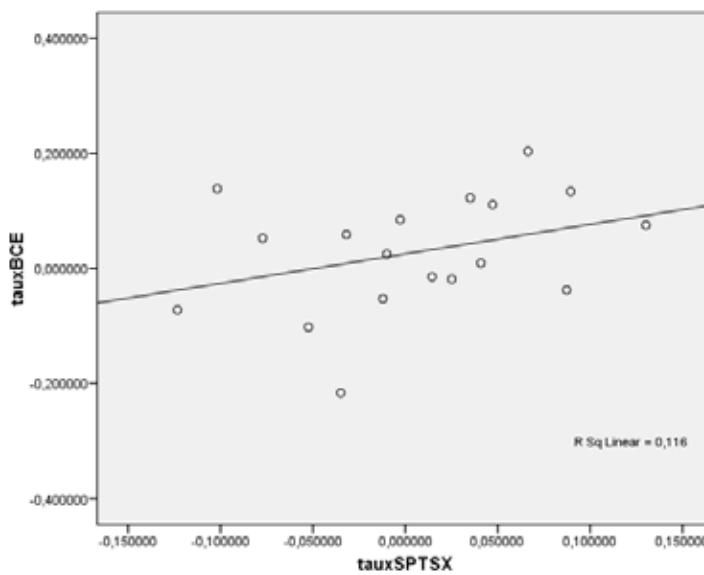


FIG. 6.24 – Le graphe de la relation

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.341 ^a	.116	.061	.099244625

a. Predictors: (Constant), tauxSPTSX

FIG. 6.25 – Le r et le r^2

En fait, étant donné que la p -value de la table ANOVA (figure 6.26) est de $0,166 > 0,05$, la relation entre les deux taux n'est pas considérée significative au seuil $\alpha = 0,05$ puisque nous ne rejetons pas H_0 dans le test suivant :

H_0 : La régression est non significative dans la population ($\beta_1 = 0$).

H_1 : La régression est significative dans la population ($\beta_1 \neq 0$).

La table des coefficients (figure 6.27) nous permet d'écrire l'équation de la droite de régression :

$$\hat{y}_{tauxBCE} = 0,025 + 0,512x_{tauxSPTSX}.$$

ANOVA ^b					
Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	.021	1	.021	2.106	.166 ^a
Residual	.158	16	.010		
Total	.178	17			

a. Predictors: (Constant), tauxSPTSX

b. Dependent Variable: tauxBCE

FIG. 6.26 – La table ANOVA

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	.025	.023		1.079	.296	-.024	.075
tauxSPTSX	.512	.352	.341	1.451	.166	-.236	1.259

a. Dependent Variable: tauxBCE

FIG. 6.27 – La table des coefficients

On voit donc que le bêta est estimé à 0,512, et il a une probabilité de 95 % de se retrouver entre -0,236 et 1,259. On ne peut donc affirmer qu'il est significativement différent de zéro. Ceci est dû à la grande dispersion et à la petite taille d'échantillon ($n = 18$). Pour illustrer ceci, rapportons-nous à la figure 6.28 qui donne les sorties de la régression effectuée avec les mêmes données que l'on a doublées (donc maintenant $n = 36$). On voit que le r et le bêta demeurent inchangés, mais maintenant la régression est considérée comme étant significative au seuil $\alpha = 0,05$ ($p\text{-value} = 0,042 < 0,05$), et donc le bêta est considéré comme étant significativement différent de zéro.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.341 ^a	.116	.090	.096281428

a. Predictors: (Constant), tauxSPTSX

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.041	1	.041	4.476	.042 ^a
	Residual	.315	34	.009		
	Total	.357	35			

a. Predictors: (Constant), tauxSPTSX

b. Dependent Variable: taux BCE

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	.025	.016			-.007	.058
	tauxSPTSX	.512	.242	.341	2.116	.042	.020
							1.003

a. Dependent Variable: taux BCE

FIG. 6.28 – Les sorties après avoir « doublé » l'échantillon

6.2.6 Introduction aux séries temporelles

La majorité des études statistiques sont élaborées sur une courte période de temps. La date de réalisation de l'étude est importante puisque l'information n'est pas valide éternellement. C'est dans le but de rafraîchir l'information qu'il y a des recensements à tous les quatre ans.

Mais il arrive qu'on amasse de l'information sur une certaine période de temps. Lorsque les mêmes données sont prises sur un même « individu » à intervalle **régulier**, nous obtenons une série de mesures. On dit alors qu'elles forment une **série chronologique ou temporelle**. Les prochains chapitres nous permettront d'étudier l'influence du temps et d'autres facteurs sur les variables que nous tentons d'expliquer.

Une analyse en régression nous permet justement d'étudier l'histoire d'une variable. Ceci nous permettra d'améliorer nos estimations et/ou nos prédictions.

Afin d'effectuer des prédictions sur le futur, deux différentes approches sont utilisées :

- Le modèle des causalités ;
- Le modèle en extrapolation.

Le premier modèle utilise les variables qui causent des effets directs sur la variable à prédire. Par exemple, imaginez un modèle de prédition des ventes d'une entreprise qui est basé sur les campagnes publicitaires de ses concurrents. Ici le problème est qu'il faut connaître à l'avance les actions des concurrents, ce qui n'est pas évident.

Le modèle en extrapolation ne se base que sur l'histoire des données déjà collectées pour effectuer les prévisions. Il faut cependant assumer que les mouvements perçus dans le passé se répéteront dans l'avenir. Dans le cas contraire, le modèle donnera de mauvaises estimations. Dans le cadre de ce cours, nous allons étudier le modèle en extrapolation.

La forme la plus simple d'une série temporelle est la forme linéaire où la variable explicative X est simplement le temps t . Voici son équation :

$$Y_i = \beta_0 + \beta_1 t_i + \epsilon_i.$$

Pour des fins de lecture des résultats, il est préférable pour les valeurs de t d'utiliser

les nombres 1, 2, 3, ... plutôt que des valeurs plus complexes telles, par exemple, des années (1999, 2000, 2001, ...). Une translation ne change rien aux résultats.

Voyons maintenant un exemple d'une série temporelle.

Exemple 6.2.7 La compagnie ABX vend des articles de sport. On comptabilise le total des ventes par trimestre (en milliers de \$) du premier trimestre de 1985 jusqu'au dernier trimestre de 1994. Faites une prédiction pour chacun des trimestres de 1995. Les données sont disponibles dans le fichier **ABX.sav**.

On s'intéresse à l'évolution des ventes au cours du temps. On s'intéresse donc au modèle

$$Y_{\text{ventes}} = \beta_0 + \beta_1 X_{\text{index}} + \epsilon.$$

Les données de l'exemple sont les suivantes :

	index	ventes	annee	saison
1	1	221,00	1985	Hiver
2	2	203,50	1985	Printemps
3	3	190,00	1985	Été
4	4	225,50	1985	Automne
5	5	223,00	1986	Hiver
6	6	190,00	1986	Printemps
7	7	206,00	1986	Été
8	8	226,50	1986	Automne
9	9	236,00	1987	Hiver
10	10	214,00	1987	Printemps
11	11	210,50	1987	Été
12	12	237,00	1987	Automne
13	13	245,50	1988	Hiver
14	14	201,00	1988	Printemps
15	15	230,00	1988	Été
16	16	254,50	1988	Automne
17	17	257,00	1989	Hiver
18	18	238,00	1989	Printemps
19	19	228,00	1989	Été
20	20	255,00	1989	Automne
21	21	260,50	1990	Hiver
22	22	244,00	1990	Printemps
23	23	256,00	1990	Été
24	24	276,50	1990	Automne

FIG. 6.29 – Les données de l'exemple

On regarde d'abord le graphe de la relation et le graphe séquentiel (figure 6.30). Ce dernier nous permet de voir s'il y a des fluctuations particulières. Pour obtenir un graphe séquentiel, les commandes sont les suivantes :

Menu SPSS :	→ Analyse
	→ Time Series
	→ Sequence Charts...
Dans la fenêtre Variables :	→ ventes
Dans la fenêtre Time Axis Labels :	→ index (la variable du temps)

Le graphe de la relation (figure 6.30) nous montre que le lien semble linéaire ; en fait on voit que les ventes augmentent de façon linéaire avec le temps.

Le graphe séquentiel, pour sa part, nous laisse voir qu'il y a des fluctuations qui semblent cycliques. Ce sont en fait des fluctuations saisonnières. Nous verrons plus tard comment tenir compte des saisons dans le modèle (il est important d'en tenir compte car ici l'hypothèse d'indépendance des résidus n'est pas vérifiée).

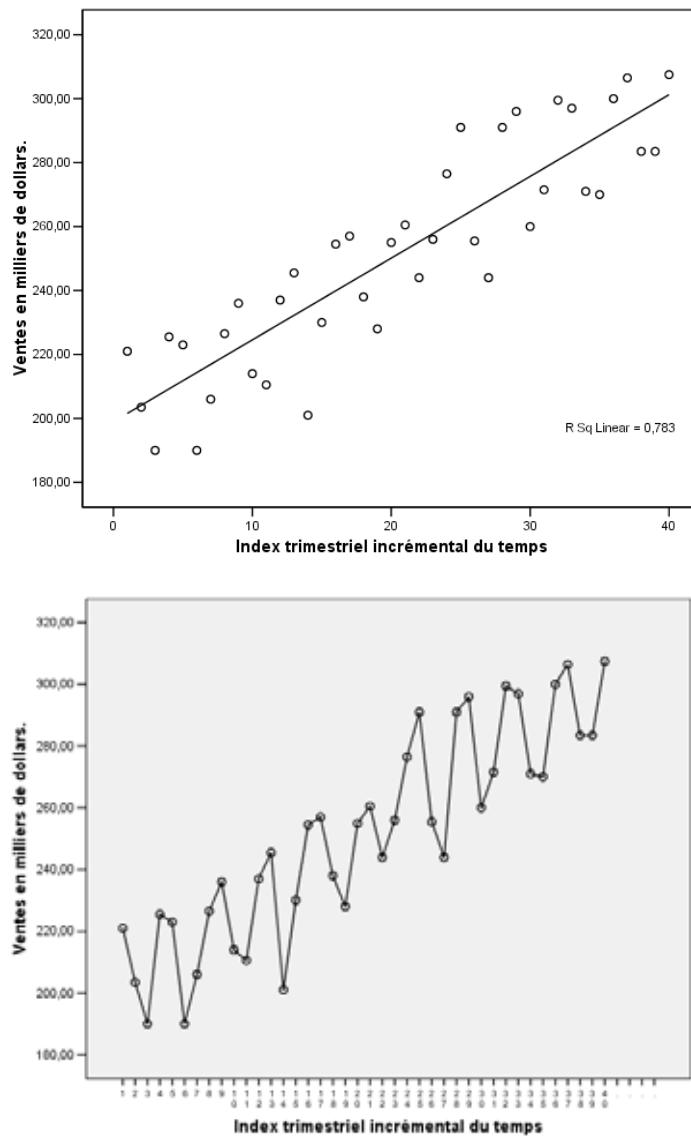


FIG. 6.30 – Le graphe de la relation et le graphe séquentiel

Pour l'instant, faisons l'analyse de ce modèle-ci à l'aide des sorties qui suivent.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,885 ^a	,783	,778	15,91264

a. Predictors: (Constant), index

b. Dependent Variable: ventes

ANOVA^b

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	34817,869	1	34817,869	137,505	,000 ^a
Residual	9622,058	38	253,212		
Total	44439,927	39			

a. Predictors: (Constant), index

b. Dependent Variable: ventes

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1 (Constant)	199,017	5,128		38,811	,000	188,636	209,398
Index trimestriel incrémental du temps	2,556	,218	,885	11,726	,000	2,115	2,997

a. Dependent Variable: Ventes en milliers de dollars.

FIG. 6.31 – Les sorties de la régression

On voit que $r = 0,885$, ce qui indique que la relation linéaire est très forte. Aussi, on a $r^2 = 0,783$. Par conséquent, 78,3 % de la variation des ventes est expliquée par le temps.

La table ANOVA nous permet de résoudre le test d'hypothèses suivant :

H_0 : La régression est non significative dans la population ($\beta_1 = 0$).

H_1 : La régression est significative dans la population ($\beta_1 \neq 0$).

Ici, puisque la p -value est égale à 0,000, ce qui est plus petit que $\alpha = 0,05$, on rejette H_0 . Ainsi, au risque de se tromper une fois sur 20, on peut affirmer que la régression est

significative au niveau de la population.

La table des coefficients nous permet d'écrire la droite de régression. On obtient

$$\hat{y}_{\text{ventes}} = b_0 + b_1 x_{\text{index}} = 199,017 + 2,556x_{\text{index}}.$$

Ainsi l'augmentation moyenne du montant des ventes est de 2,556 milliers de dollars à chaque trimestre. Et la valeur de b_0 nous indique que le montant des ventes du dernier trimestre de 1984 a dû être d'environ 199 017 \$.

On voit aussi que l'intervalle de confiance de niveau 95 % pour β_1 est [2,115, 2,997]. Ainsi l'augmentation moyenne du montant des ventes d'un trimestre au suivant devrait être comprise entre 2 115 \$ et 2 996 \$, et ce 19 fois sur 20.

De même, le montant des ventes du dernier trimestre de 1984 devrait être entre 188 636 \$ et 209 398 \$, et ce avec une probabilité de 95 %.

Passons maintenant aux prévisions demandées. On voit que les index correspondant aux trimestres de 1995 sont 41, 42, 43 et 44. Donc pour les prévisions ponctuelles il suffit de faire les calculs suivants (en utilisant l'équation de la droite de régression) :

$$\hat{y}_{\text{ventes}, 41} = 199,017 + 2,556 \cdot 41 = 303,81 \text{ milliers de \$}$$

$$\hat{y}_{\text{ventes}, 42} = 199,017 + 2,556 \cdot 42 = 306,36 \text{ milliers de \$}$$

$$\hat{y}_{\text{ventes}, 43} = 199,017 + 2,556 \cdot 43 = 308,92 \text{ milliers de \$}$$

$$\hat{y}_{\text{ventes}, 44} = 199,017 + 2,556 \cdot 44 = 311,48 \text{ milliers de \$}$$

	index	ventes	annee	saison	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
39	39	283,50	1994	été	298,6959	289,0743	308,3176	265,0763	332,3156
40	40	307,50	1994	automne	301,2518	291,2531	311,2505	267,5223	334,9813
41	41	.	1995	hiver	303,8077	293,4268	314,1885	269,9629	337,6524
42	42	.	1995	printemps	306,3635	295,5960	317,1310	272,3982	340,3289
43	43	.	1995	été	308,9194	297,7612	320,0776	274,8282	343,0106
44	44	.	1995	automne	311,4752	299,9227	323,0278	277,2529	345,6976

FIG. 6.32 – Les prévisions

Les intervalles de confiance correspondant à ces prévisions ponctuelles sont disponibles dans la figure 6.32. Ainsi, si on s'intéresse aux ventes **réelles** que l'on devrait obtenir pour les trimestres de 1995, les intervalles de confiance de niveau 95 % sont les suivants :

Trimestres de 1995	Ventes réelles en milliers de \$ (IC de niveau 95 %)
Premier	[269,9629, 337,6524]
Deuxième	[272,3982, 340,3289]
Troisième	[274,8282, 343,0106]
Quatrième	[277,2529, 345,6976]

6.2.7 Association et lien de cause à effet

Une association entre deux variables Y et X n'est pas toujours un lien de causalité. D'autres explications peuvent être plausibles :

- L'inverse est vrai ($Y \Rightarrow X$). La personne responsable des analyses a mal interprété le lien entre les variables. Par exemple, le lien entre la productivité dans une usine et le nombre de décibels.
- Il existe une troisième variable, inconnue, qui produit souvent des changements sur Y et X à la fois, et pourtant, X et Y n'ont aucun lien direct entre eux. Par exemple, avant que le vaccin de Salk ne soit développé contre la polio, une étude a démontré qu'il y avait une augmentation de nouveaux cas de polio à chaque fois que la vente de liqueurs douces augmentait. Est-ce que les liqueurs douces transmettaient la polio ? Non, il y avait une troisième variable, la saison, qui avait un effet sur les deux variables en même temps.

6.3 Régression linéaire avec E-Views

Le logiciel E-Views a été développé par des économistes et est utilisé surtout pour la modélisation des séries temporelles et les prévisions. C'est un outil puissant qui s'utilise facilement. Reprenons les données de l'exemple 6.2.7 (la série ABX) pour voir comment faire un modèle de régression avec E-Views.

Tout d'abord, la feuille de travail de cet exemple est le fichier abx.wf1. La figure 6.33 montre ce qu'on obtient en ouvrant ce fichier avec E-Views. Le premier item de la liste est la série des ventes de ABX. L'item **c** est toujours là par défaut et sert pour l'évaluation de la constante dans un modèle ; l'item **resid** est la série dans laquelle sont stockés les résidus des modèles, et finalement l'item **t** est une série créée pour représenter le passage d'un trimestre à l'autre (index incrémental).

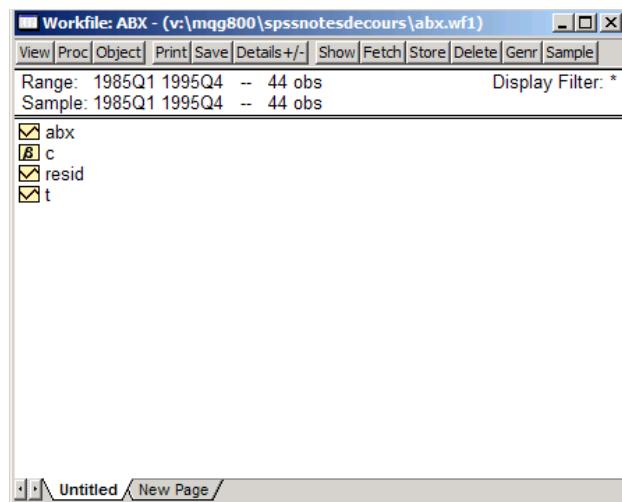


FIG. 6.33 – Feuille de travail abx.wf1

Aussi, dans le haut de la feuille de travail on voit que l'échantillon est défini du premier trimestre de 1985 au dernier trimestre de 1995. En fait la série des ventes se termine au dernier trimestre de 1994, et 1995 est là dans le but de faire des prévisions.

Pour visualiser une série, il suffit de la sélectionner d'un clic droit puis de sélectionner **Open** (figure 6.34). On voit alors la série comme dans la figure 6.35.

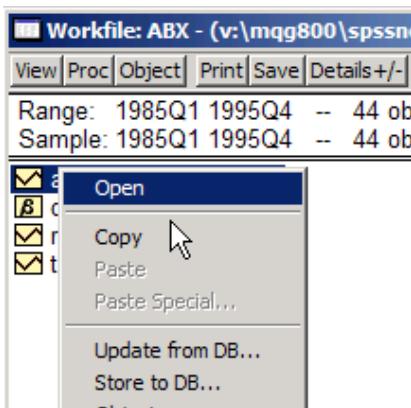


FIG. 6.34 –

A screenshot of the EViews software interface showing a data grid titled 'abx'. The grid contains a table of data with two columns: 'Year' and 'Value'. The data spans from 1985Q1 to 1989Q2. The values are as follows:

Year	Value
1985Q1	221.0000
1985Q2	203.5000
1985Q3	190.0000
1985Q4	225.5000
1986Q1	223.0000
1986Q2	190.0000
1986Q3	206.0000
1986Q4	226.5000
1987Q1	236.0000
1987Q2	214.0000
1987Q3	210.5000
1987Q4	237.0000
1988Q1	245.5000
1988Q2	201.0000
1988Q3	230.0000
1988Q4	254.5000
1989Q1	257.0000
1989Q2	

FIG. 6.35 – Visualisation des données

À partir du menu **View**, plusieurs options s'offrent à nous. Par exemple, si on sélectionne **View → Graph → Line**, on obtient le graphe séquentiel de la série (figure 6.36).

La figure 6.36 présente aussi un graphe dans lequel on relie les ventes associées à un même trimestre (trimestre 1, 2, 3 ou 4). On l'obtient en sélectionnant **View → Graph → Seasonal Split Line**. Ce graphe nous permet de voir qu'il y a de bonnes différences selon les trimestres ; les trimestres 1 et 4 ont des ventes assez semblables, qui sont différentes des ventes des trimestres 2 et 3.

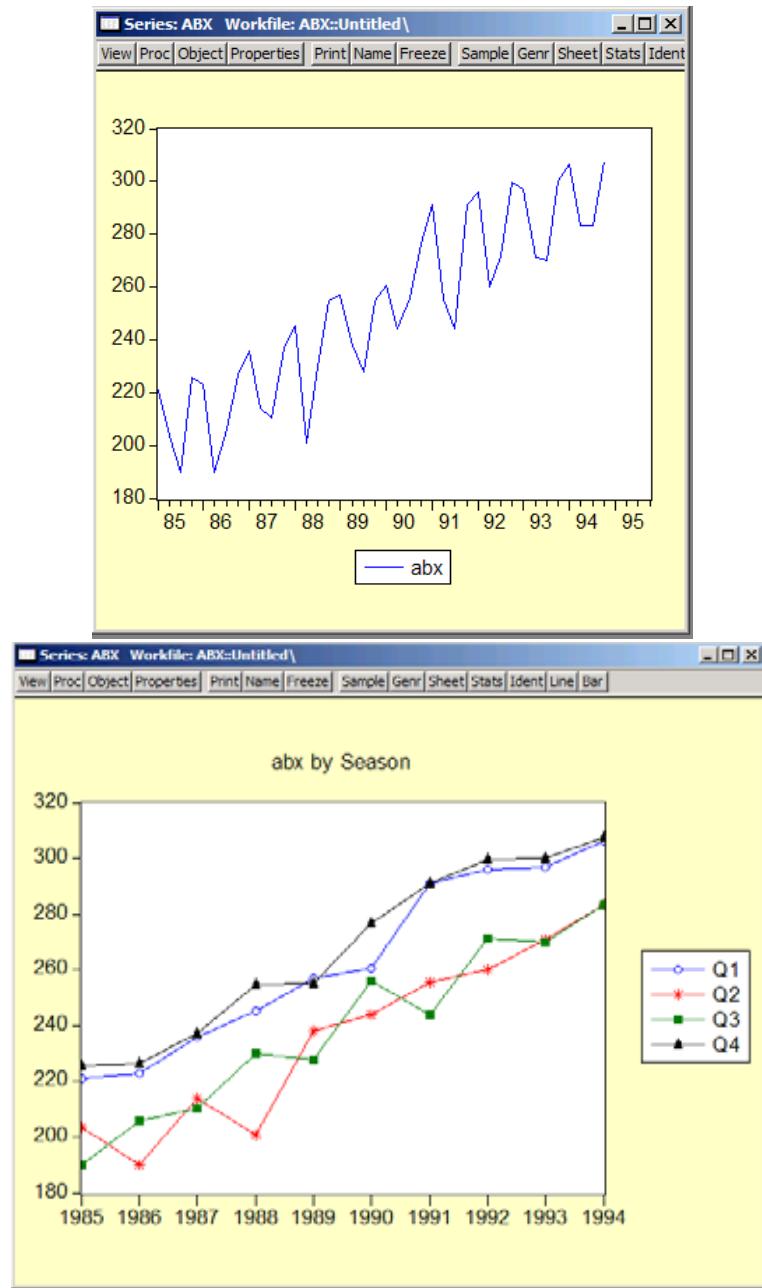


FIG. 6.36 – Graphe séquentiel de la série, et le graphe de la série selon les trimestres

On peut aussi facilement obtenir les statistiques de la série (figure 6.37) en faisant View → Descriptive Statistics → Histogram and Stats.

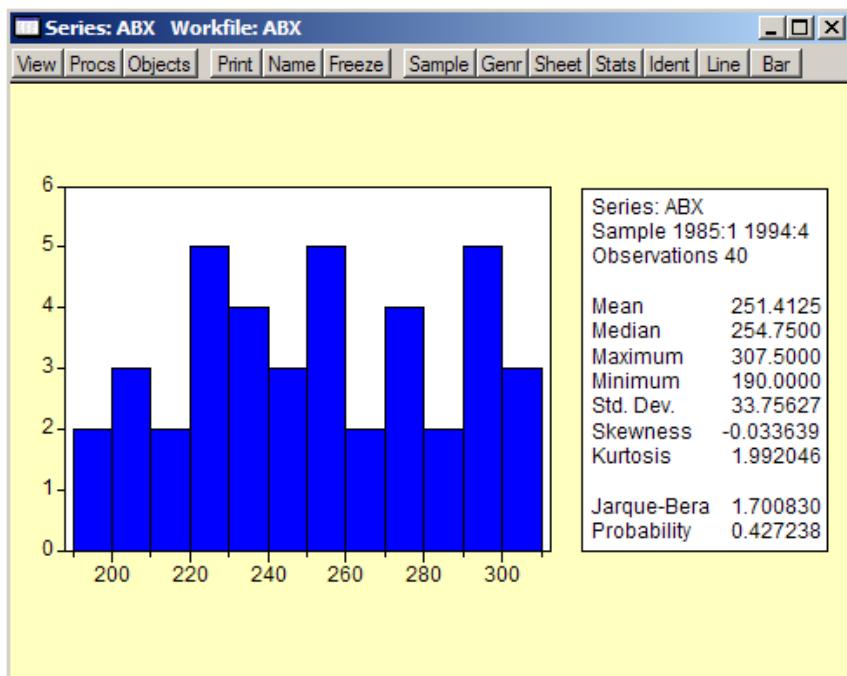


FIG. 6.37 – Les statistiques de la série

Pour estimer un modèle, on doit sélectionner Quick → Estimate Equation... (dans le haut de la fenêtre). Ensuite, il faut écrire l'équation qui nous intéresse en donnant d'abord le nom de la variable dépendante, puis on indique c si on veut une constante, puis le nom des variables indépendantes. Ici, on veut reproduire le modèle vu dans l'exemple 6.2.7, et donc l'équation est `abx c t` (rappelons que `t` est la série de l'index incrémental indiquant le passage d'un trimestre à l'autre). La méthode d'estimation est celle des moindres carrés, qui est l'option par défaut de E-Views. On peut visualiser la saisie de l'équation dans la figure 6.38.

En cliquant sur OK on obtient alors la sortie de la figure 6.39. D'après la colonne Coefficient, on peut voir que l'équation de la droite de régression est

$$\hat{y} = 199,0173 + 2,5559t$$

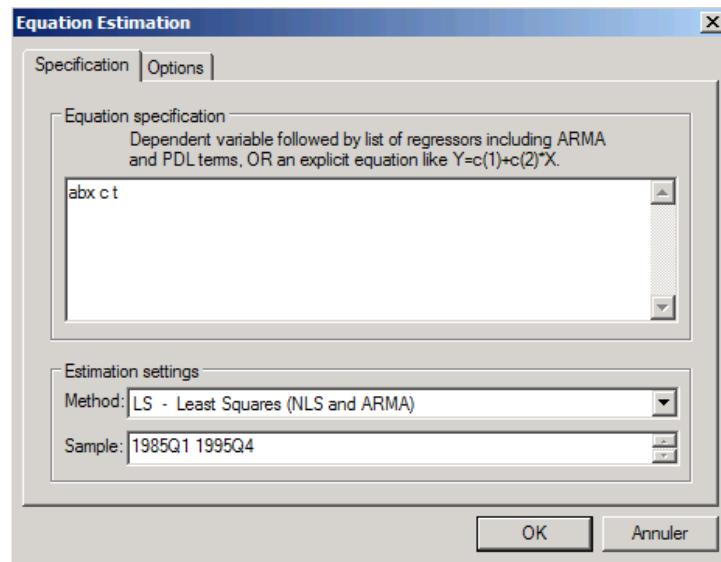


FIG. 6.38 – Équation du modèle de régression

View Proc Object Print Name Freeze Estimate Forecast Stats Resids				
Dependent Variable: ABX				
Method: Least Squares				
Date: 09/18/07 Time: 16:38				
Sample (adjusted): 1985Q1 1994Q4				
Included observations: 40 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	199.0173	5.127875	38.81087	0.0000
T	2.555863	0.217961	11.72624	0.0000
R-squared	0.783482	Mean dependent var	251.4125	
Adjusted R-squared	0.777784	S.D. dependent var	33.75627	
S.E. of regression	15.91264	Akaike info criterion	8.420811	
Sum squared resid	9622.061	Schwarz criterion	8.505255	
Log likelihood	-166.4162	F-statistic	137.5048	
Durbin-Watson stat	1.970249	Prob(F-statistic)	0.000000	

FIG. 6.39 – Estimations des paramètres du modèle et statistiques

ce qui est bien la même équation que celle obtenue dans l'exemple 6.2.7.

La sortie contient bien d'autres informations. Par exemple, on voit que la *p*-value de la statistique *F* est nulle (**Prob(F-statistic)**), ce qui signifie que le modèle est significatif (comme dans une table ANOVA). On voit aussi que $r^2 = 0,7835$.

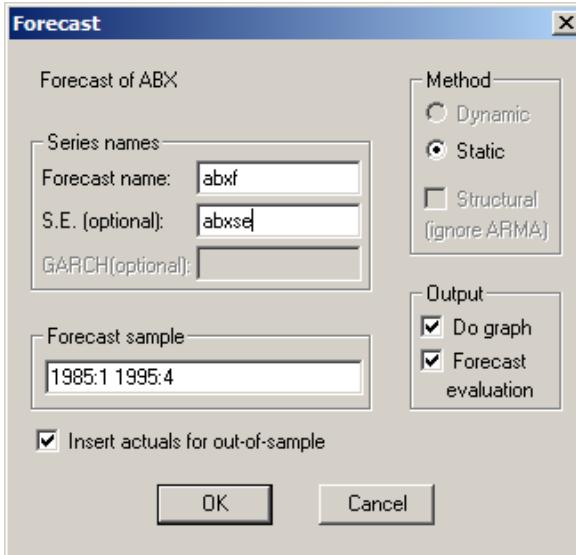


FIG. 6.40 – Pour faire des prévisions

Maintenant, si on veut faire les prévisions des ventes pour 1995, il faut aller dans le menu **Forecast** (on voit le bouton **Forecast** dans le haut de la fenêtre de la figure 6.39). On obtient alors la fenêtre de la figure 6.40. Pour l'instant on laisse les options par défaut, et pour sauvegarder la série des prévisions et ses **Standard Errors** il faut simplement indiquer un nom pour chacune dans la fenêtre **Series name**. Ici la série des prévisions se nomme **abxf** et la série des S.E. se nomme **abxse**. On voit aussi qu'on peut indiquer sur quelle période se définit l'échantillon pour les prévisions ; ici on les demande pour tous les trimestres jusqu'à la fin de 1995.

En cliquant sur **OK**, on obtient le graphe de la figure 6.41. Ceci montre en fait la droite de la régression jusqu'en 1995, ce qui représente effectivement les prévisions. Les deux autres lignes représentent un intervalle d'un niveau de confiance d'environ 95 % pour les prévisions. Les valeurs présentées à la droite du graphe mesurent l'ajustement du modèle

à la série. Nous en reparlerons plus longuement dans un chapitre subséquent.

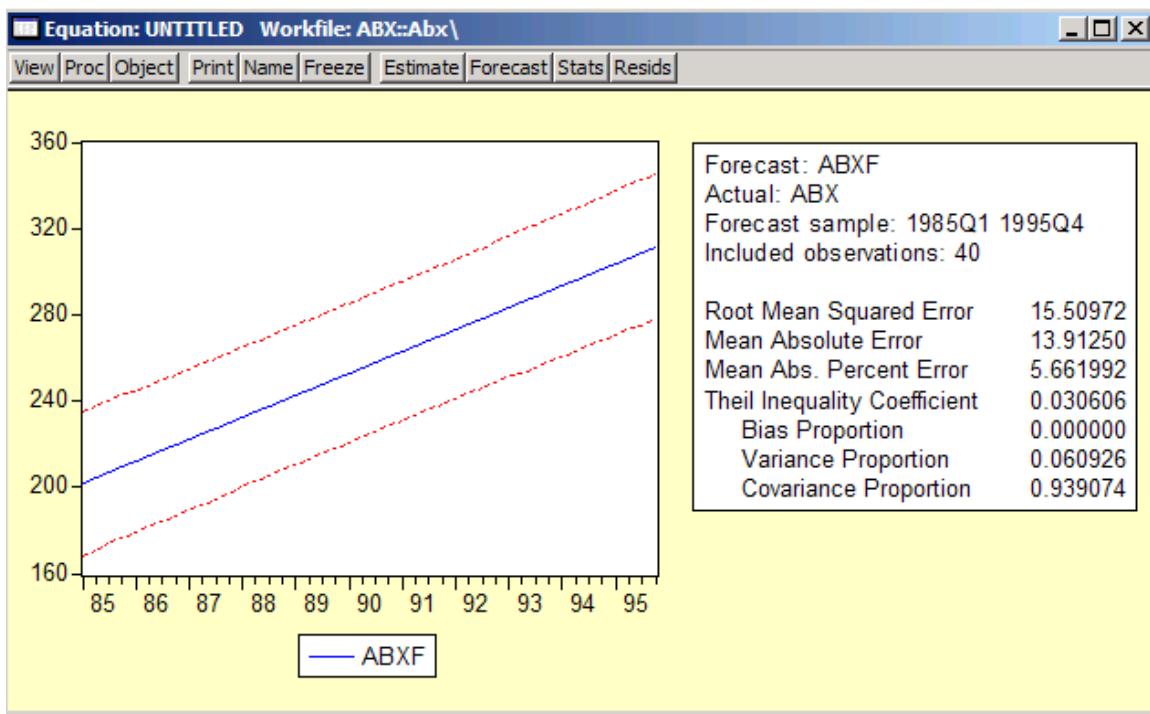
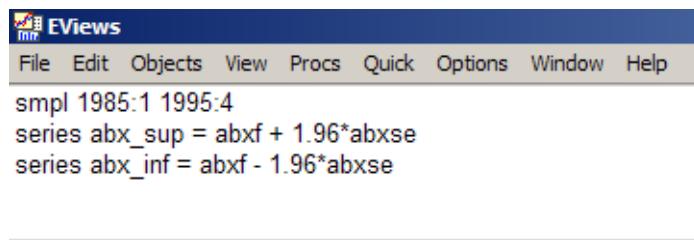


FIG. 6.41 – Le graphe des prévisions

Maintenant que la série `abxf` des prévisions a été créée, il est possible d'aller la visualiser. La figure 6.42 en présente un extrait. On y voit en particulier les prévisions pour 1995, qui coïncident avec celles qu'on avait obtenues avec SPSS. Il serait intéressant d'avoir les intervalles de confiance pour ces prévisions. Il faut aller les créer en inscrivant les commandes que l'on voit dans la figure 6.43 (dans le haut de la fenêtre de E-Views). Celles-ci nous permettent de créer les bornes inférieures et supérieures des intervalles de confiance de niveau 95 % (d'où le 1,96) pour les prévisions.

ABXF			
1992Q1	1992Q1	273.1373	
1992Q2	1992Q2	275.6932	
1992Q3	1992Q3	278.2491	
1992Q4	1992Q4	280.8049	
1993Q1	1993Q1	283.3608	
1993Q2	1993Q2	285.9167	
1993Q3	1993Q3	288.4725	
1993Q4	1993Q4	291.0284	
1994Q1	1994Q1	293.5842	
1994Q2	1994Q2	296.1401	
1994Q3	1994Q3	298.6960	
1994Q4	1994Q4	301.2518	
1995Q1	1995Q1	303.8077	
1995Q2	1995Q2	306.3636	
1995Q3	1995Q3	308.9194	
1995Q4	1995Q4	311.4753	

FIG. 6.42 – Extrait des prévisions



```

EViews
File Edit Objects View Procs Quick Options Window Help
smpl 1985:1 1995:4
series abx_sup = abxf + 1.96*abxse
series abx_inf = abxf - 1.96*abxse

```

FIG. 6.43 – Créations des bornes des intervalles de confiance

Et en effet, lorsque ces commandes sont effectuées, les séries `abx_inf` et `abx_sup` sont créées et représentent respectivement les bornes inférieures et supérieures des intervalles de confiance. En allant sélectionner les séries `abx`, `abxf`, `abx_inf` et `abx_sup` en même temps (avec la touche Shift ou Ctrl), on peut les ouvrir pour visualiser toutes ces valeurs dans une seule fenêtre (en sélectionnant `Open → as Group` avec le bouton droit, voir figure 6.44). On peut alors visualiser les quatre séries comme dans la figure 6.45. On voit par exemple que pour le quatrième trimestre de 1995, on prévoit des ventes de 311 475,30 \$, et que d'après ce modèle (s'il est valide, ce qu'on n'a pas vérifié encore !) on est sûr à 95 % que les ventes du quatrième trimestre de 1995 seront comprises entre 278 341,50 \$ et 344 609 00 \$.

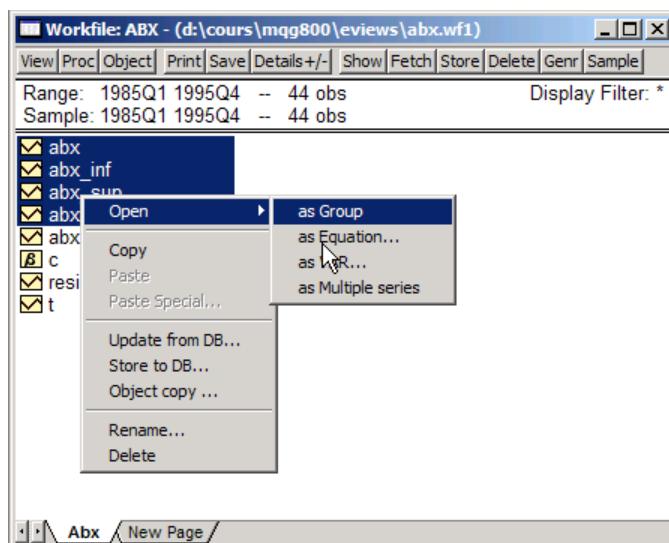


FIG. 6.44 –

Group: UNTITLED Workfile: ABX

obs	ABX	ABX_INF	ABX_SUP	ABXF
1991:3	244.0000	236.3275	299.7237	268.0256
1991:4	291.0000	238.8431	302.3198	270.5815
1992:1	296.0000	241.3530	304.9217	273.1373
1992:2	260.0000	243.8572	307.5292	275.6932
1992:3	271.5000	246.3558	310.1423	278.2491
1992:4	299.5000	248.8488	312.7611	280.8049
1993:1	297.0000	251.3362	315.3854	283.3608
1993:2	271.0000	253.8181	318.0152	285.9167
1993:3	270.0000	256.2944	320.6506	288.4725
1993:4	300.0000	258.7653	323.2914	291.0284
1994:1	306.5000	261.2308	325.9377	293.5842
1994:2	283.5000	263.6909	328.5893	296.1401
1994:3	283.5000	266.1457	331.2462	298.6960
1994:4	307.5000	268.5952	333.9084	301.2518
1995:1	NA	271.0395	336.5759	303.8077
1995:2	NA	273.4786	339.2485	306.3636
1995:3	NA	275.9126	341.9262	308.9194
1995:4	NA	278.3415	344.6090	311.4753

FIG. 6.45 – Extrait des séries des ventes, des prévisions et des bornes des IC à 95 %

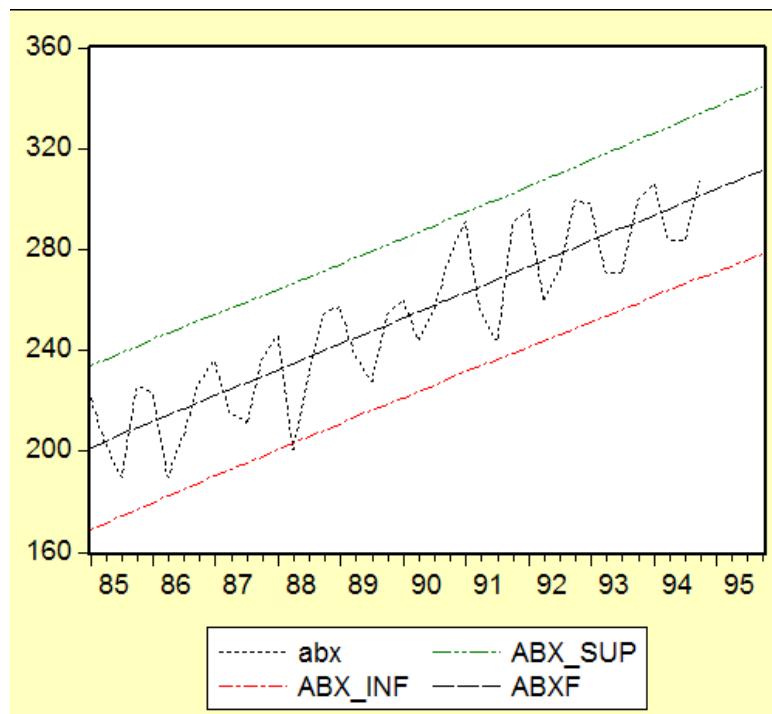


FIG. 6.46 – Graphe séquentiel des quatre séries

Encore plus intéressant, en allant dans `View → Graph → Line`, on obtient le graphe de la figure 6.46, qui nous permet de voir si les bornes des intervalles de confiance contiennent effectivement les véritables ventes de 1985 à 1994. En fait il semble qu'il n'y ait qu'une seule valeur qui ne soit pas comprise dans l'intervalle de confiance ; en vérifiant, on voit que c'est au deuxième trimestre de 1988 ; les ventes ont alors été de 201 000 \$ alors que la borne inférieure de l'intervalle de confiance pour ce trimestre est de 203 101,30 \$.

6.4 Exercices du chapitre

Exercice 1 Une entreprise fabrique des toiles métalliques pour des usines de pâte et papier. Afin de mieux répartir leur personnel, le responsable de la production aimerait utiliser la régression pour estimer le temps requis, en moyenne, pour la finition des toiles. Une variable importante pouvant affecter le temps de finition est la surface de la toile.

Le tableau ci-dessous donne l'information sur 15 toiles qui ont été fabriquées par l'usine. Le responsable de la production postule qu'une relation linéaire est plausible entre les deux variables.

	ident	tempsfin	surface
1	1	5,5	10
2	2	5,9	15
3	3	5,8	12
4	4	6,3	16
5	5	7,0	18
6	6	7,5	25
7	7	5,5	12
8	8	7,2	22
9	9	6,5	17
10	10	6,5	16
11	11	7,1	20
12	12	7,0	17
13	13	6,9	18
14	14	6,8	17
15	15	6,6	18

FIG. 6.47 – Les données de l'exercice

Faites l'analyse en régression pour les variables `tempsfin` et `surface` en répondant aux questions suivantes :

1. Le graphique de la relation est-il bien linéaire ?
2. Quel est le % de la variation constatée des temps de finition d'une toile à l'autre qui est expliqué par la surface de la toile ?
3. Quelle est l'équation de la droite ? Peut-on conclure, au seuil de signification de 5 %, que la droite est significative ?
4. Estimez le temps moyen de finition pour des toiles dont la surface est de 13 m².

5. Calculez un intervalle de confiance, de niveau 95 %, qui contient les temps moyens de finition des toiles de 13 m².

Exercice 2 Le fichier `s&pdata.wf1` contient les données mensuelles du S&P Composite index returns de janvier 1995 à septembre 2001. Les séries de cette feuille de travail sont `inf` (consumer price index (CPI) inflation rate), `dt_bill` (three-month Treasury bill (T-bill) rate) et `ret` pour le S&P Composite index returns. Faites deux modèles de régression linéaire simple avec `ret` comme variable dépendante.

Exercice 3 Reprenons le contexte de la base de données `party.sav`. Étudiez le lien `musique` \Rightarrow `satis`.

Exercice 4 À l'aide des données de `magasinageweb.sav`, étudiez le lien entre la satisfaction et la frustration (`sitstam` \Rightarrow `frustm`).

Exercice 5 À l'aide des données de `satisfactiontravail.sav`, étudiez le lien entre le niveau de satisfaction par rapport à la possibilité d'organiser soi-même son travail et le niveau d'influence perçu sur la façon de faire son travail (`sati_q3` \Rightarrow `part_q3`).

Chapitre 7

Relation entre trois variables discrètes

L'étude de la relation entre trois variables discrètes est utilisée à titre de complément à l'analyse de la relation entre deux variables discrètes. En fait, l'idée est d'introduire une troisième variable discrète afin de mieux comprendre la relation entre les deux premières variables discrètes. Ceci permet de :

- Déetecter si la relation entre deux variables est illusoire ou erronée, ce qui permet d'éviter d'émettre de fausses interprétations.
- Préciser et détailler la relation entre deux variables discrètes qui, a priori, est trop générale.

Les techniques d'analyses sont les mêmes que celles présentées au chapitre 4; l'introduction de la troisième variable discrète a simplement pour effet de produire k tableaux de contingence, avec k le nombre de modalités de cette variable.

Le chapitre est divisé en deux sections : la section 7.1 présente des exemples où l'introduction de la 3^e variable permet de vérifier si la relation entre les deux variables est illusoire ou non. La section 7.2 présente des exemples où l'introduction de la 3^e variable permet de préciser la relation initiale.

7.1 Validation de la relation initiale

L'étude de l'authenticité d'une relation entre deux variables discrètes peut être faite par l'introduction d'une troisième variable, appelée **variable de contrôle**, dans la relation. La variable ajoutée peut intervenir dans la relation initiale de différentes façons. Dans le cadre de l'analyse de la validité de la relation initiale, trois situations peuvent se présenter :

- La relation originale était illusoire ;
- La relation originale était étouffée ;
- La relation originale est déformée par une distorsion.

7.1.1 Les relations illusoires

L'introduction d'une variable de contrôle à titre de troisième variable peut avoir pour effet de faire disparaître la relation originale. Cette situation montre que la relation originale était de nature illusoire. À l'inverse, il est aussi possible que la relation originale tienne toujours malgré l'ajout d'une variable de contrôle. Cette dernière situation montre que la relation originale est stable, ce qui sécurise l'analyste.

Exemple 7.1.1 Prenons l'exemple d'une entreprise au prise avec des problèmes d'absentéisme chronique les lundis. On pense que les employés membres de clubs sont significativement plus absents que les non-membres.

La base de données se nomme `absentlundi.sav`. Cette base de données ne contenant que des variables discrètes, elle a été faite avec le principe du poids : pour chaque combinaison possible des réponses des variables, on indique combien de répondants correspondent à cette combinaison. Par exemple, il y a 4 individus qui ont entre 20 et 35 ans, ne sont pas membre d'un club et sont rarement absents le lundi (voir la figure 7.1).

Une telle base de donnée peut s'obtenir d'une base de données traditionnelle par la méthode vue à la section 4.6. Et tout comme dans la section 4.6, il faut effectuer les

commandes suivantes afin d'attribuer le poids aux variables :

Menu SPSS :	→ Data
	→ Weight Cases...
	<input checked="" type="checkbox"/> Weight cases by
Dans la fenêtre Frequency Variable	→ poids

	club	age	absent	poids
1	non	20-35 ans	rarement	4
2	non	20-35 ans	parfois	12
3	non	20-35 ans	souvent	16
4	non	35-50 ans	rarement	40
5	non	35-50 ans	parfois	40
6	non	35-50 ans	souvent	20
7	non	50 ans et +	rarement	56
8	non	50 ans et +	parfois	22
9	non	50 ans et +	souvent	16
10	oui	20-35 ans	rarement	4
11	oui	20-35 ans	parfois	32
12	oui	20-35 ans	souvent	56
13	oui	35-50 ans	rarement	20
14	oui	35-50 ans	parfois	10
15	oui	35-50 ans	souvent	8
16	oui	50 ans et +	rarement	10
17	oui	50 ans et +	parfois	9
18	oui	50 ans et +	souvent	2

FIG. 7.1 – La base de données avec une variable poids

On peut maintenant vérifier si le fait d'être membre ou pas d'un club influence le fait d'être absent le lundi. Fixons les seuils à $\alpha = 0,05$ pour l'exemple en entier. On obtient les sorties 7.2 et 7.3.

Il faudrait ici faire l'analyse complète de la relation ; la p -value du test du χ^2 étant nulle, on voit que les deux variables sont liées. Le Cramer's V ayant une valeur de 0,255, cette relation est qualifiée de faible. D'après le tableau croisé, on peut penser que le fait d'être membre d'un club a comme effet d'être plus souvent absent le lundi par rapport à ceux qui ne sont pas membres d'un club.

			dub		Total	
			non	oui		
absent	rarement	Count	100	34	134	
		Expected Count	80,3	53,7	134,0	
		% within club	44,2%	22,5%	35,5%	
		Std. Residual	2,2	-2,7		
	parfois	Count	74	51	125	
		Expected Count	74,9	50,1	125,0	
		% within club	32,7%	33,8%	33,2%	
		Std. Residual	-,1	,1		
	souvent	Count	52	66	118	
		Expected Count	70,7	47,3	118,0	
		% within club	23,0%	43,7%	31,3%	
		Std. Residual	-2,2	2,7		
Total		Count	226	151	377	
		Expected Count	226,0	151,0	377,0	
		% within club	100,0%	100,0%	100,0%	

FIG. 7.2 – Tableau croisé de la relation club \Rightarrow absent

Chi-Square Tests				Symmetric Measures		
	Value	df	Asymp. Sig. (2-sided)		Value	Approx. Sig.
Pearson Chi-Square	24,448 ^a	2	,000	Nominal by Nominal	Phi	,255 ,000
Likelihood Ratio	24,870	2	,000	Nominal	Cramer's V	,255 ,000
Linear-by-Linear Association	24,382	1	,000	N of Valid Cases		377
N of Valid Cases	377					

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 47,26.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.3 – χ^2 et Cramer's V de la relation club \Rightarrow absent

Introduisons maintenant la variable indiquant la classe d'âge de l'individu comme variable de contrôle. Pour ce faire il faut effectuer les commandes comme si on voulait obtenir le tableau croisé de la relation `club` \Rightarrow `absent`, mais en plus on ajoute la variable `age` dans la fenêtre Layer 1 of 1. On obtient alors les sorties des figures 7.4 et 7.5.

			absent * club * age Crosstabulation		
			club		Total
			non	oui	
20-35 ans	absent	rarement	Count	4	4
			Expected Count	2,1	5,9
			% within club	12,5%	4,3%
			Std. Residual	1,3	-.8
	parfois		Count	12	32
			Expected Count	11,4	32,6
			% within club	37,5%	34,8%
	souvent		Count	16	56
			Expected Count	18,6	53,4
			% within club	50,0%	60,9%
	Total		Count	32	92
			Expected Count	32,0	92,0
			% within club	100,0%	100,0%
35-50 ans	absent	rarement	Count	40	20
			Expected Count	43,5	16,5
			% within club	40,0%	52,6%
	parfois		Count	40	10
			Expected Count	36,2	13,8
			% within club	40,0%	26,3%
	souvent		Count	20	8
			Expected Count	20,3	7,7
			% within club	20,0%	21,1%
	Total		Count	100	38
			Expected Count	100,0	38,0
			% within club	100,0%	100,0%
50 ans et +	absent	rarement	Count	56	10
			Expected Count	53,9	12,1
			% within club	59,6%	47,6%
	parfois		Count	,3	-,6
			Expected Count	22	9
			% within club	23,4%	42,9%
	souvent		Count	22	9
			Expected Count	25,3	5,7
			% within club	23,4%	31,0
	Total		Count	-,7	1,4
			Expected Count	16	2
			% within club	17,0%	9,5%
			Count	,3	-,7
			Expected Count	14,7	3,3
			% within club	17,0%	18,0
			Count	94	21
			Expected Count	94,0	21,0
			% within club	100,0%	100,0%

FIG. 7.4 – Tableau croisé de la relation `club` \Rightarrow `absent` avec la variable de contrôle `age`

Chi-Square Tests

age		Value	df	Asymp. Sig. (2-sided)
20-35 ans	Pearson Chi-Square	2,978 ^a	2	,226
	Likelihood Ratio	2,682	2	,262
	Linear-by-Linear Association	2,250	1	,134
	N of Valid Cases	124		
35-50 ans	Pearson Chi-Square	2,449 ^b	2	,294
	Likelihood Ratio	2,506	2	,286
	Linear-by-Linear Association	,628	1	,428
	N of Valid Cases	138		
50 ans et +	Pearson Chi-Square	3,454 ^c	2	,178
	Likelihood Ratio	3,272	2	,195
	Linear-by-Linear Association	,061	1	,805
	N of Valid Cases	115		

- a. 1 cells (16,7%) have expected count less than 5. The minimum expected count is 2,06.
 b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,71.
 c. 1 cells (16,7%) have expected count less than 5. The minimum expected count is 3,29.

Symmetric Measures

age		Value	Approx. Sig.
20-35 ans	Nominal by Nominal	Phi	,155
		Cramer's V	,155
	N of Valid Cases		124
35-50 ans	Nominal by Nominal	Phi	,133
		Cramer's V	,133
	N of Valid Cases		138
50 ans et +	Nominal by Nominal	Phi	,173
		Cramer's V	,173
	N of Valid Cases		115

- a. Not assuming the null hypothesis.
 b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.5 – χ^2 et Cramer's V de la relation club \Rightarrow absent avec la variable de contrôle age

On a maintenant trois relations à étudier : une par classe d'âge. Or, pour les trois classes, la p -value du test du χ^2 est supérieure à $\alpha = 0,05$ (elles ont comme valeur 0,226, 0,294 et 0,178 respectivement). Donc la relation initiale ne tient plus lorsqu'on tient compte de l'âge des employés. Pour mieux comprendre ce phénomène, examinons les relations `age` \Rightarrow `absent` et `age` \Rightarrow `club`.

			absent * age Crosstabulation			Total	
			age				
			20-35 ans	35-50 ans	50 ans et +		
absent	rarement	Count	8	60	66	134	
		Expected Count	44,1	49,1	40,9	134,0	
		% within age	6,5%	43,5%	57,4%	35,5%	
		Std. Residual	-5,4	1,6	3,9		
	parfois	Count	44	50	31	125	
		Expected Count	41,1	45,8	38,1	125,0	
		% within age	35,5%	36,2%	27,0%	33,2%	
		Std. Residual	,5	,6	-1,2		
	souvent	Count	72	28	18	118	
		Expected Count	38,8	43,2	36,0	118,0	
		% within age	58,1%	20,3%	15,7%	31,3%	
		Std. Residual	5,3	-2,3	-3,0		
Total			124	138	115	377	
			124,0	138,0	115,0	377,0	
			100,0%	100,0%	100,0%	100,0%	

FIG. 7.6 – Tableau croisé de la relation `age` \Rightarrow `absent`

D'abord les figures 7.6 et 7.7 nous montrent que la relation `age` \Rightarrow `absent` existe et est qualifiée d'intéressante. Ce sont les 20-35 ans qui s'absentent le lundi, et ce de façon marquée.

Chi-Square Tests				Symmetric Measures		
	Value	df	A Sympl. Sig. (2-sided)		Value	Approx. Sig.
Pearson Chi-Square	92,064 ^a	4	,000	Nominal by Nominal	Phi	,494
Likelihood Ratio	101,887	4	,000	Cramer's V		,349
Linear-by-Linear Association	79,131	1	,000	N of Valid Cases		377
N of Valid Cases	377					

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 35,99.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.7 – χ^2 et Cramer's V de la relation age \Rightarrow absent

			age			Total
			20-35 ans	35-50 ans	50 ans et +	
club	non	Count	32	100	94	226
		Expected Count	74,3	82,7	68,9	226,0
		% within age	25,8%	72,5%	81,7%	59,9%
		Std. Residual	-4,9	1,9	3,0	
	oui	Count	92	38	21	151
		Expected Count	49,7	55,3	46,1	151,0
		% within age	74,2%	27,5%	18,3%	40,1%
		Std. Residual	6,0	-2,3	-3,7	
	Total	Count	124	138	115	377
		Expected Count	124,0	138,0	115,0	377,0
		% within age	100,0%	100,0%	100,0%	100,0%

FIG. 7.8 – Tableau croisé de la relation age \Rightarrow club

Ensuite les figures 7.8 et 7.9 nous montrent que la relation age \Rightarrow club existe et est elle aussi qualifiée d'intéressante. Ce sont les 20-35 ans qui sont membres de clubs, et ce de façon marquée.

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	91,945 ^a	2	,000
Likelihood Ratio	94,242	2	,000
Linear-by-Linear Association	79,012	1	,000
N of Valid Cases	377		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 46,06.

Symmetric Measures		
	Value	Approx. Sig.
Nominal by Nominal Phi	,494	,000
Nominal Cramer's V	,494	,000
N of Valid Cases	377	

- a. Not assuming the null hypothesis.
- b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.9 – χ^2 et Cramer's V de la relation age \Rightarrow club

Ainsi, la relation originale était illusoirement significative par la présence d'une plus grande concentration de jeunes employés membres de clubs.

7.1.2 Les relations « étouffées »

Lorsque nous sommes en présence d'une absence de relation entre deux variables discrètes alors qu'il devrait logiquement en exister une, l'analyste doit soupçonner qu'il existe une troisième variable dont l'absence étouffe la relation initiale. Voyons un exemple de ceci.

Exemple 7.1.2 Une compagnie qui conçoit des logiciels veut lancer deux nouveaux jeux. Le jeu 1 vise les 11-14 ans et le jeu 2 vise les 15-18 ans. La base données se nomme `jeux.sav`.

On recueille donc un échantillon de 83 jeunes dans le but de vérifier si la clientèle est bien ciblée. On demande à chaque jeune d'indiquer si c'est le jeu 1 ou 2 qui suscite le plus son intérêt. Les figures 7.10 et 7.11 présentent les sorties pour étudier la relation `age` \Rightarrow `jeux`.

			age		Total	
			11-14 ans	15-18 ans		
jeux	Jeu 1	Count	23	14	37	
		Expected Count	19,2	17,8	37,0	
		% within age	53,5%	35,0%	44,6%	
		Std. Residual	,9	-,9		
	Jeu 2	Count	20	26	46	
		Expected Count	23,8	22,2	46,0	
		% within age	46,5%	65,0%	55,4%	
		Std. Residual	-,8	,8		
Total		Count	43	40	83	
		Expected Count	43,0	40,0	83,0	
		% within age	100,0%	100,0%	100,0%	

FIG. 7.10 – Tableau croisé de la relation `age` \Rightarrow `jeux`

La p -value du test du χ^2 étant de 0,09, la relation est inexisteante au seuil $\alpha = 0,05$. Et au seuil $\alpha = 0,10$, elle est qualifiée de faible. Ainsi la clientèle ne semble pas très bien ciblée. Pour mieux comprendre, on décide d'introduire la variable `sexé` pour voir si celle-ci influence la relation.

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2,867 ^b	1	,090		
Continuity Correction ^a	2,168	1	,141		
Likelihood Ratio	2,888	1	,089		
Fisher's Exact Test				,122	,070
Linear-by-Linear Association	2,833	1	,092		
N of Valid Cases	83				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 17,83.

Symmetric Measures

	Value	Approx. Sig.
Nominal by Nominal	Phi ,186	,090
Nominal	Cramer's V ,186	,090
N of Valid Cases	83	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.11 – χ^2 et Cramer's V de la relation age \Rightarrow jeux

Les figures 7.12 et 7.13 présentent les sorties obtenues lorsqu'on introduit la variable sexe.

jeux * age * sexe Crosstabulation					
sexe	jeux	Jeu 1	age		Total
			11-14 ans	15-18 ans	
Masculin	jeux	Count	17	5	22
		Expected Count	11,3	10,7	22,0
		% within age	81,0%	25,0%	53,7%
		Std. Residual	1,7	-1,7	
	Jeu 2	Count	4	15	19
		Expected Count	9,7	9,3	19,0
		% within age	19,0%	75,0%	46,3%
		Std. Residual	-1,8	1,9	
	Total	Count	21	20	41
		Expected Count	21,0	20,0	41,0
		% within age	100,0%	100,0%	100,0%
Féminin	jeux	Count	6	9	15
		Expected Count	7,9	7,1	15,0
		% within age	27,3%	45,0%	35,7%
		Std. Residual	-,7	,7	
	Jeu 2	Count	16	11	27
		Expected Count	14,1	12,9	27,0
		% within age	72,7%	55,0%	64,3%
		Std. Residual	,5	-,5	
	Total	Count	22	20	42
		Expected Count	22,0	20,0	42,0
		% within age	100,0%	100,0%	100,0%

FIG. 7.12 – Tableau croisé de la relation `age` \Rightarrow `jeux` avec la variable de contrôle `sexe`

On voit qu'au seuil $\alpha = 0,05$, la relation `age` \Rightarrow `jeux` existe, mais seulement pour les garçons (la p -value est nulle pour les garçons, et elle a une valeur de 0,231 pour les filles). La relation est même qualifiée de forte pour le groupe des garçons. La relation initiale existe donc, mais pas pour les filles. Ainsi la compagnie pourra ajuster ses stratégies de marketing en fonction de ces résultats.

Chi-Square Tests

sexé		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Masculin	Pearson Chi-Square	12,897 ^b	1	,000		
	Continuity Correction ^a	10,745	1	,001		
	Likelihood Ratio	13,675	1	,000		
	Fisher's Exact Test				,001	,000
	Linear-by-Linear Association	12,583	1	,000		
	N of Valid Cases	41				
Féminin	Pearson Chi-Square	1,434 ^c	1	,231		
	Continuity Correction ^a	,766	1	,382		
	Likelihood Ratio	1,440	1	,230		
	Fisher's Exact Test				,336	,191
	Linear-by-Linear Association	1,400	1	,237		
	N of Valid Cases	42				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,27.

c. 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,14.

Symmetric Measures

sexé			Value	Approx. Sig.
Masculin	Nominal by Nominal	Phi Cramer's V	,561 ,561	,000 ,000
	N of Valid Cases		41	
Féminin	Nominal by Nominal	Phi Cramer's V	-,185 ,185	,231 ,231
	N of Valid Cases		42	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.13 – χ^2 et Cramer's V de la relation age \Rightarrow jeux avec la variable de contrôle sexe

7.1.3 Les relations déformées

Parfois, l'introduction d'une variable de contrôle peut carrément changer la relation initiale et alors mettre en évidence la véritable nature de la relation.

Exemple 7.1.3 Il est reconnu que les familles disposent souvent d'un compte dans l'institution financière la plus près de leur maison. Supposons que pour une banque vous conduisez une étude auprès de 400 foyers : 200 foyers ont été sélectionnés à moins d'un kilomètre de la banque et 200 ont été sélectionnés à plus d'un kilomètre. La base de données se nomme `banque.sav`.

Supposons que vous obteniez les figures 7.14 et 7.15, qui illustrent que, pour cette banque, les familles qui habitent à plus d'un kilomètre ont plus tendance à avoir un compte à cette banque que celles qui habitent plus près. Définitivement, cette situation est plutôt incompréhensible.

			proximité			
			M oins de 1 km	Plus de 1 km	Total	
banque	Oui	Count	102	122	224	
		Expected Count	112,0	112,0	224,0	
		% within proximité	51,0%	61,0%	56,0%	
		Std. Residual	,9	,9		
	Non	Count	98	78	176	
		Expected Count	88,0	88,0	176,0	
		% within proximité	49,0%	39,0%	44,0%	
		Std. Residual	1,1	-1,1		
Total		Count	200	200	400	
		Expected Count	200,0	200,0	400,0	
		% within proximité	100,0%	100,0%	100,0%	

FIG. 7.14 – Tableau croisé de la relation proximité \Rightarrow banque

Le tableau croisé illustre qu'il y a effectivement un lien (mais faible) entre la possession d'un compte et la proximité du domicile ; cependant, ce ne sont pas les gens qui habitent à proximité qui possèdent un compte. L'analyste espère alors trouver une variable de contrôle qui viendra expliquer ce phénomène. Dans cet exemple, la variable de

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,058 ^b	1	,044		
Continuity Correction ^a	3,663	1	,056		
Likelihood Ratio	4,066	1	,044		
Fisher's Exact Test				,055	,028
Linear-by-Linear Association	4,048	1	,044		
N of Valid Cases	400				

a. Computed only for a 2x2table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 88,00.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-,101	,044
Nominal	Cramer's V	,101	,044
N of Valid Cases		400	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.15 – χ^2 et Cramer's V de la relation proximite \Rightarrow banque

la proximité de la concurrence agit à titre de variable de contrôle. Les figures 7.16 et 7.17 présentent comment cette variable agit sur la relation originale.

banque * proximite * competition Crosstabulation

			proximite		Total	
			Moins de 1 km	Plus de 1 km		
Oui	banque	Oui	Count	29	4	
			Expected Count	24,8	8,3	
			% within proximite	24,2%	10,0%	
			Std. Residual	,9	-1,5	
	Non	Non	Count	91	36	
			Expected Count	95,3	31,8	
			% within proximite	75,8%	90,0%	
			Std. Residual	-,4	,8	
	Total		Count	120	40	
			Expected Count	120,0	40,0	
			% within proximite	100,0%	100,0%	
Non	banque	Oui	Count	73	118	
			Expected Count	63,7	127,3	
			% within proximite	91,3%	73,8%	
			Std. Residual	1,2	-,8	
	Non	Non	Count	7	42	
			Expected Count	16,3	32,7	
			% within proximite	8,8%	26,3%	
			Std. Residual	-2,3	1,6	
	Total		Count	80	160	
			Expected Count	80,0	160,0	
			% within proximite	100,0%	100,0%	

FIG. 7.16 – Tableau croisé de la relation proximite \Rightarrow banque avec la variable de contrôle competition

La variable de contrôle permet ici à la relation initiale de prendre le sens auquel on s'attendait ; remarquez le signe des résidus.

Chi-Square Tests

competition		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Oui	Pearson Chi-Square	3,678 ^b	1	,055		
	Continuity Correction ^a	2,863	1	,091		
	Likelihood Ratio	4,138	1	,042		
	Fisher's Exact Test				,071	,040
	Linear-by-Linear Association	3,655	1	,056		
	N of Valid Cases	160				
Non	Pearson Chi-Square	10,052 ^c	1	,002		
	Continuity Correction ^a	9,004	1	,003		
	Likelihood Ratio	11,256	1	,001		
	Fisher's Exact Test				,001	,001
	Linear-by-Linear Association	10,010	1	,002		
	N of Valid Cases	240				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 8,25.

c. 0 cells (,0%) have expected count less than 5. The minimum expected count is 16,33.

Symmetric Measures

competition		Value	Approx. Sig.
Oui	Nominal by Nominal	Phi	,152
		Cramer's V	,152
	N of Valid Cases		160
Non	Nominal by Nominal	Phi	,205
		Cramer's V	,205
	N of Valid Cases		240

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.17 – χ^2 et Cramer's V de la relation proximite \Rightarrow banque avec la variable de contrôle competition

7.2 Détailler la relation initiale

L'ajout d'une troisième variable discrète peut aussi permettre de préciser la relation entre deux variables discrètes. C'était le cas dans l'exemple 7.1.2 ; non seulement l'introduction de la variable `sexe` a-t-elle permis de constater que la relation initiale existait dans le groupe des garçons, mais elle a aussi permis de préciser la relation entre l'âge et l'intérêt pour les jeux. Voici un autre exemple.

Exemple 7.2.1 On suppose ici que les bons vendeurs ont un certain type de personnalité. Pour vérifier cette hypothèse, on fait passer un test de personnalité à 194 nouveaux vendeurs. Après un certains temps, des informations sont fournies sur le succès de chaque vendeur. La base de données se nomme `personnalite.sav`. Les figures 7.18 et 7.19 contiennent les sorties illustrant la relation entre les résultats au test de personnalité et le succès du vendeur dans son travail.

			personnalité		Total	
			Bas	Élevé		
ventes	Peu	Count	69	37	106	
		Expected Count	49,2	56,8	106,0	
		% within personnalité	76,7%	35,6%	54,6%	
		Std. Residual	2,8	-2,6		
	Beaucoup	Count	21	67	88	
		Expected Count	40,8	47,2	88,0	
		% within personnalité	23,3%	64,4%	45,4%	
		Std. Residual	-3,1	2,9		
Total		Count	90	104	194	
		Expected Count	90,0	104,0	194,0	
		% within personnalité	100,0%	100,0%	100,0%	

FIG. 7.18 – Tableau croisé de la relation `personnalité` \Rightarrow `ventes`

Les résultats viennent confirmer l'hypothèse : un bon résultat au test de personnalité semble entraîner un meilleur succès dans les ventes. La relation est qualifiée d'intéressante, et la connaissance du résultat au test de personnalité donne une probabilité de

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	32,867 ^b	1	,000		
Continuity Correction ^a	31,230	1	,000		
Likelihood Ratio	34,083	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	32,697	1	,000		
N of Valid Cases	194				

a. Computed only for a 2x2table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 40,82.

Symmetric Measures					
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Nominal by Nominal	Phi	,412			,000
Nominal	Cramer's V	,412			,000
Ordinal by Ordinal	Gamma	,712	,079	6,333	,000
N of Valid Cases		194			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.19 – χ^2 , Cramer's V et gamma de la relation personnalité \Rightarrow ventes

71,2 % de prédire le succès ou l'insuccès du vendeur.

Une autre hypothèse a été avancée suivant laquelle un bon vendeur est une personne qui a une grande empathie. On décide donc d'introduire la variable `empathie` qui indique le résultat à la section du test de personnalité qui concerne l'empathie. Les figures 7.20 et 7.21 contiennent les sorties illustrant cette relation.

ventes * personnalité * empathie Crosstabulation

empathie			personnalité		Total	
			Bas	Élevé		
Bas	ventes	Peu	Count	57	23	80
			Expected Count	51,2	28,8	80,0
			% within personnalité	89,1%	63,9%	80,0%
			Std. Residual	,8	-1,1	
	Beaucoup		Count	7	13	20
			Expected Count	12,8	7,2	20,0
			% within personnalité	10,9%	36,1%	20,0%
			Std. Residual	-1,6	2,2	
	Total		Count	64	36	100
			Expected Count	64,0	36,0	100,0
			% within personnalité	100,0%	100,0%	100,0%
Élevé	ventes	Peu	Count	12	14	26
			Expected Count	7,2	18,8	26,0
			% within personnalité	46,2%	20,6%	27,7%
			Std. Residual	1,8	-1,1	
	Beaucoup		Count	14	54	68
			Expected Count	18,8	49,2	68,0
			% within personnalité	53,8%	79,4%	72,3%
			Std. Residual	-1,1	,7	
	Total		Count	26	68	94
			Expected Count	26,0	68,0	94,0
			% within personnalité	100,0%	100,0%	100,0%

FIG. 7.20 – Tableau croisé de la relation personnalité \Rightarrow ventes avec la variable empathie

On voit que la relation reste présente dans les deux groupes (résultat bas et résultat élevé à la section concernant l'empathie). L'introduction de la variable empathie permet de faire certaines nuances dans la relation personnalité \Rightarrow ventes ; lesquelles ?

Chi-Square Tests

empathie		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Bas	Pearson Chi-Square	9,125 ^b	1	,003		
	Continuity Correction ^a	7,620	1	,006		
	Likelihood Ratio	8,802	1	,003		
	Fisher's Exact Test				,004	,003
	Linear-by-Linear Association	9,034	1	,003		
	N of Valid Cases	100				
Élevé	Pearson Chi-Square	6,144 ^b	1	,013		
	Continuity Correction ^a	4,933	1	,026		
	Likelihood Ratio	5,827	1	,016		
	Fisher's Exact Test				,020	,015
	Linear-by-Linear Association	6,078	1	,014		
	N of Valid Cases	94				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,20.

c. 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,19.

Symmetric Measures

empathie		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Bas	Nominal by Nominal	Phi	,302		,003
		Cramer's V	,302		,003
	Ordinal by Ordinal	Gamma	,643	,155	2,789
	N of Valid Cases		100		,005
Élevé	Nominal by Nominal	Phi	,256		,013
		Cramer's V	,256		,013
	Ordinal by Ordinal	Gamma	,536	,176	2,272
	N of Valid Cases		94		,023

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.21 – χ^2 , Cramer's V et gamma de la relation personnalité ⇒ ventes avec la variable empathie

7.3 Exercice du chapitre

Les sorties qui suivent vous permettront d'étudier des relations entre les variables **sexe**, **salairecl** (salaire en classes) et **fonction**. Faites les analyses appropriées et tirez vos conclusions. Attention aux pré-requis.

			sexe		Total	
			féminin	masculin		
salaried	250\$-450\$	Count	10	14	24	
		Expected Count	4,1	19,9	24,0	
		% within sexe	29,4%	8,4%	12,0%	
	450\$-650\$	Std. Residual	2,9	-1,3		
		Count	18	102	120	
		Expected Count	20,4	99,6	120,0	
	650\$ et +	% within sexe	52,9%	61,4%	60,0%	
		Std. Residual	-,5	,2		
		Count	6	50	56	
	Total	Expected Count	9,5	46,5	56,0	
		% within sexe	17,6%	30,1%	28,0%	
		Std. Residual	-1,1	,5		
			Count	34	166	
			Expected Count	34,0	166,0	
			% within sexe	100,0%	100,0%	
					100,0%	

FIG. 7.22 – Tableau croisé de la relation **sexe** \Rightarrow **salairecl**

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	12,257 ^a	2	,002
Likelihood Ratio	10,167	2	,006
Linear-by-Linear Association	8,392	1	,004
N of Valid Cases	200		

a. 1 cells (16,7%) have expected count less than 5. The minimum expected count is 4,08.

Symmetric Measures		
	Value	Approx. Sig.
Nominal by Phi	,248	,002
Nominal Cramer's V	,248	,002
N of Valid Cases	200	

- a. Not assuming the null hypothesis.
- b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.23 – χ^2 et Cramer's V de la relation sexe \Rightarrow salairecl

salairecl * sexe * fonction Crosstabulation				
fonction	salairecl	sexe		Total
		féminin	masculin	
administration	250\$-450\$	Count	7	14
		Expected Count	4,5	14,0
		% within sexe	70,0%	33,3%
		Std. Residual	1,2	,8
	450\$-650\$	Count	3	17
		Expected Count	5,5	17,0
		% within sexe	30,0%	66,7%
		Std. Residual	-1,1	,7
	Total		10	31
	Count	10	31,0	
	Expected Count	10,0	31,0	
	% within sexe	100,0%	100,0%	
production	250\$-450\$	Count	3	10
		Expected Count	1,6	10,0
		% within sexe	14,3%	6,5%
		Std. Residual	1,1	,5
	450\$-650\$	Count	13	91
		Expected Count	14,8	91,0
		% within sexe	61,9%	72,2%
		Std. Residual	,5	,2
	650\$ et +	Count	5	28
		Expected Count	4,6	28,0
		% within sexe	23,8%	21,3%
		Std. Residual	,2	,1
	Total		21	129
	Count	21	129	
	Expected Count	21,0	129,0	
	% within sexe	100,0%	100,0%	
direction	450\$-650\$	Count	2	12
		Expected Count	,9	12,0
		% within sexe	66,7%	27,0%
		Std. Residual	1,2	,3
	650\$ et +	Count	1	28
		Expected Count	2,1	28,0
		% within sexe	33,3%	73,0%
		Std. Residual	,8	,2
	Total		3	40
	Count	3	40,0	
	Expected Count	3,0	40,0	
	% within sexe	100,0%	100,0%	

FIG. 7.24 – Tableau croisé de la relation sexe \Rightarrow salairecl avec la variable de contrôle fonction

Chi-Square Tests						
	fonction	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
administration	Pearson Chi-Square	3,677 ^a	1	,055		
	Continuity Correction ^b	2,346	1	,126		
	Likelihood Ratio	3,733	1	,053	,121	,063
	Fisher's Exact Test					
	Linear-by-Linear Association	3,659	1	,059		
	N of Valid Cases	31				
production	Pearson Chi-Square	1,698 ^c	2	,428		
	Likelihood Ratio	1,487	2	,476		
	Linear-by-Linear Association	,178	1	,874		
	N of Valid Cases	129				
direction	Pearson Chi-Square	2,076 ^d	1	,150		
	Continuity Correction ^b	,618	1	,432		
	Likelihood Ratio	1,869	1	,172		
	Fisher's Exact Test				,209	,209
	Linear-by-Linear Association	2,024	1	,155		
	N of Valid Cases	40				

- a. Computed only for a 2x2 table
 b. 1 cells (25,0%) have expected count less than 5. The minimum expected count is 4,52.
 c. 2 cells (33,3%) have expected count less than 5. The minimum expected count is 1,63.
 d. 2 cells (50,0%) have expected count less than 5. The minimum expected count is ,90.

Symmetric Measures

	fonction	Value	Approx. Sig.
administration	Nominal by Nominal	Phi	,344
	Nominal	Cramer's V	,344
	N of Valid Cases		31
production	Nominal by Nominal	Phi	,115
	Nominal	Cramer's V	,115
	N of Valid Cases		129
direction	Nominal by Nominal	Phi	,228
	Nominal	Cramer's V	,228
	N of Valid Cases		40

- a. Not assuming the null hypothesis.
 b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.25 – χ^2 et Cramer's V de la relation sexe \Rightarrow salairecl avec la variable de contrôle fonction

			fonction			Total	
			administratiion	production	direction		
salairecl	250\$-450\$	Count	14	10	0	24	
		Expected Count	3,7	15,5	4,8	24,0	
		% within fonction	45,2%	7,8%	,0%	12,0%	
		Std. Residual	5,3	-1,4	-2,2		
	450\$-650\$	Count	17	91	12	120	
		Expected Count	18,6	77,4	24,0	120,0	
		% within fonction	54,8%	70,5%	30,0%	60,0%	
		Std. Residual	-,4	1,5	-2,4		
	650\$ et +	Count	0	28	28	56	
		Expected Count	8,7	36,1	11,2	56,0	
		% within fonction	,0%	21,7%	70,0%	28,0%	
		Std. Residual	-2,9	-1,4	5,0		
Total		Count	31	129	40	200	
		Expected Count	31,0	129,0	40,0	200,0	
		% within fonction	100,0%	100,0%	100,0%	100,0%	

FIG. 7.26 – Tableau croisé de la relation fonction \Rightarrow salairecl

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79,381 ^a	4	,000
Likelihood Ratio	75,189	4	,000
Linear-by-Linear Association	61,931	1	,000
N of Valid Cases	200		

a. 2 cells (22,2%) have expected count less than 5. The minimum expected count is 3,72.

Symmetric Measures		
	Value	Approx. Sig.
Nominal by Nominal Phi	,630	,000
Nominal Cramer's V	,445	,000
N of Valid Cases	200	

- a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.27 – χ^2 et Cramer's V de la relation fonction \Rightarrow salairecl

			sexe		Total	
			féminin	masculin		
fonction	administration	Count	10	21	31	
		Expected Count	5,3	25,7	31,0	
		% within sexe	29,4%	12,7%	15,5%	
		Std. Residual	2,1	-,9		
	production	Count	21	108	129	
		Expected Count	21,9	107,1	129,0	
		% within sexe	61,8%	65,1%	64,5%	
		Std. Residual	-,2	,1		
	direction	Count	3	37	40	
		Expected Count	6,8	33,2	40,0	
		% within sexe	8,8%	22,3%	20,0%	
		Std. Residual	-1,5	,7		
Total		Count	34	166	200	
		Expected Count	34,0	166,0	200,0	
		% within sexe	100,0%	100,0%	100,0%	

FIG. 7.28 – Tableau croisé de la relation sexe \Rightarrow fonction

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,721 ^a	2	,021
Likelihood Ratio	7,437	2	,024
Linear-by-Linear Association	7,268	1	,007
N of Valid Cases	200		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 5,27.

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	,196	,021
Nominal	Cramer's V	,196	,021
N of Valid Cases		200	

- a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 7.29 – χ^2 et Cramer's V de la relation sexe \Rightarrow fonction

Chapitre 8

Relation entre deux variables discrètes et une variable continue

Dans le cadre de ce chapitre, nous nous intéressons à analyser l'influence de deux variables discrètes sur une variable continue. Contrairement aux analyses vues précédemment, ici les données correspondant à la variable continue seront obtenues en fixant d'abord les deux variables discrètes à certaines valeurs : c'est ce que nous appellerons des traitements.

La section 8.1 présente ce type d'analyse (ANOVA à deux facteurs), et la section 8.2 présente l'étude des graphiques associés à ces analyses.

8.1 Analyse de la variance à deux facteurs

Nous analyserons l'influence de deux variables discrètes sur une variable continue par l'entremise d'une ANOVA à deux facteurs, les facteurs étant les variables discrètes.

Illustrons nos propos à l'aide d'un exemple. Supposons qu'une entreprise veut évaluer l'effet du prix de vente (première variable discrète) et de la publicité (deuxième variable discrète) sur les ventes totales de leur nouveau photocopieur AMINE-P (variable continue). Supposons aussi que les deux prix proposés pour une telle imprimante sont de 600 \$ et 700 \$. Finalement, on considère trois types de médias publicitaires : la télévision, la radio et le journal.

Les variables discrètes sont appelées facteurs. Il est conventionnel de noter les facteurs par A et B . Chaque facteur contient un certain nombre de valeurs discrètes appelées niveaux. Par exemple, le facteur prix contient deux niveaux tandis que le facteur publicité contient trois niveaux.

Dans la terminologie ANOVA, chaque combinaison des facteurs A et B est appelée un traitement. Ainsi, le nombre de traitements possibles est le produit des niveaux des facteurs. Dans notre problème, il y a $2 \times 3 = 6$ différents traitements :

- prix de vente de 600 \$ et publicité à la télévision ;
- prix de vente de 600 \$ et publicité à la radio ;
- prix de vente de 600 \$ et publicité dans les journaux ;
- prix de vente de 700 \$ et publicité à la télévision ;
- prix de vente de 700 \$ et publicité à la radio ;
- prix de vente de 700 \$ et publicité dans les journaux.

Lorsqu'on recueille une valeur de la variable continue pour un traitement en particulier, on dit alors qu'on a fait un **essai** pour ce traitement. Ainsi, dans notre exemple, pour chaque prix de vente et type de publicité fixés dans une région, faire un essai revient à recueillir le montant des ventes dans cette région (ici on doit choisir une région différente pour chaque traitement).

Cette façon de procéder porte le nom de *factorial design*. Lorsque tous les traitements

sont étudiés, cette expérience porte le nom de *complete factorial experiment*. Lorsque seulement un sous-ensemble de traitements est étudié, nous sommes en présence d'un plan expérimental plus complexe. Ces plans utilisent généralement une approche basée sur les carrés latins (par exemple la méthode *Taguchi*). Dans le cadre de ce cours d'introduction, seuls les *complete factorial experiment* seront étudiés.

Pour chacun des traitements, au moins deux essais doivent être faits (c'est-à-dire que pour chacun des traitements, il faut obtenir au moins deux valeurs pour la variable continue).

Aussi, dans le cadre de ce cours, nous supposons que le nombre d'essais est égal pour chaque expérience, et ce, malgré qu'il est possible d'utiliser des essais en nombres différents. Mentionnons que, dans ce dernier cas, il faut utiliser des statistiques pondérées et qu'alors la décomposition de la variation totale de la variable continue ne satisfait pas l'égalité. Un modèle ANOVA à deux facteurs s'écrit ainsi :

$$x_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijl}$$

où

- x_{ijl} est la valeur de la variable continue obtenue au l^e essai du traitement correspondant au i^e niveau du facteur A et au j^e niveau du facteur B ;
- μ est la valeur moyenne commune à tous les traitements ;
- α_i représente l'effet du facteur A lorsque son niveau est fixé à i ;
- β_j représente l'effet du facteur B lorsque son niveau est fixé à j ;
- $(\alpha\beta)_{ij}$ représente l'effet interactif des facteurs A et B lorsque leurs niveaux respectifs sont fixés à i et j ;
- ϵ_{ijl} représente une distorsion aléatoire.

Les effets des facteurs A et B sont appelés effets principaux. Dans une analyse de la variance à deux facteurs, il est possible de déterminer si l'effet des facteurs A et B sont significatifs ; cependant, la vraie puissance de cette approche réside dans le fait qu'elle

permet de vérifier la puissance de l'interaction entre A et B , ce qui est impossible avec une ANOVA à un facteur. La présence d'une interaction indique que les changements perçus sur la variable continue dépendent à la fois de tous les niveaux des facteurs A et B . Si le modèle ne possède pas d'interaction, les facteurs A et B agissent alors de façon indépendante sur la variable continue.

Pour voir s'il y a un lien entre les facteurs et la variable continue, et pour voir si ce lien provient des facteurs de façon indépendante ou s'il y a une interaction, on décompose la variation de la variable continue en plusieurs sources. Le tableau 8.1 présente cette décomposition. L'analyse décompose la variation totale de la variable continue (appelée variable dépendante) en quatre sources de variation :

- variation expliquée par le facteur A (SSA) ;
- variation expliquée par le facteur B (SSB) ;
- variation expliquée l'interaction des facteurs ($SSINT$) ;
- variation inexpliquée (SSE).

Analyse de la variance à deux facteurs		
Décomposition de la variation		
	somme des carrés (sum of squares)	Degrés de liberté (degrees of freedom)
TOTAL :	$SST = \sum_i \sum_j \sum_l (x_{ijl} - \bar{x})^2$	$KHL - 1$
FACTEUR A :	$SSA = HL \sum_{i=1}^K (\bar{x}_{i\bullet\bullet} - \bar{x})^2$	$K - 1$
FACTEUR B :	$SSB = KL \sum_{j=1}^H (\bar{x}_{\bullet j\bullet} - \bar{x})^2$	$H - 1$
INTERACTION :	$SSINT = L \sum_{i=1}^K \sum_{j=1}^H (\bar{x}_{ij\bullet} - \bar{x}_{i\bullet\bullet} - \bar{x}_{\bullet j\bullet} + \bar{x})^2$	$(K - 1)(H - 1)$
ERREUR :	$SSE = \sum_i \sum_j \sum_l (x_{ijl} - \bar{x}_{ij\bullet})^2$	$KH(L - 1)$
Alors $SST = SSA + SSB + SSINT + SSE$		

FIG. 8.1 – ANOVA à deux facteurs

Cette décomposition de la variation nous permet de résoudre des tests d'hypothèses pour voir s'il y a un lien entre les facteurs et la variable continue, et de quelle façon les facteurs A et B influencent la variable continue. Pour illustrer la démarche à suivre, reprenons l'exemple de la présente section. Ainsi on veut évaluer l'effet du prix de vente et de la publicité sur le montant des ventes d'un nouveau photocopieur. Supposons que nous avons recueilli les données suivantes (les montants des ventes sont en milliers de dollars) :

	Télévision	Radio	Journal
Prix 600 \$	18,0 16,8	12,0 13,2	7,8 8,9
Prix 700 \$	14,0 14,7	10,8 9,6	9,8 8,4

FIG. 8.2 – Les données de l'exemple

Ces données seront saisies de la façon suivante dans SPSS :

	ident	prix	media	vente
1	1	600\$	Télévision	18,0
2	2	600\$	Télévision	16,8
3	3	600\$	Radio	12,0
4	4	600\$	Radio	13,2
5	5	600\$	Journal	7,8
6	6	600\$	Journal	8,9
7	7	700\$	Télévision	14,0
8	8	700\$	Télévision	14,7
9	9	700\$	Radio	10,8
10	10	700\$	Radio	9,6
11	11	700\$	Journal	9,8
12	12	700\$	Journal	8,4
13				

FIG. 8.3 – Saisie dans SPSS

L'analyse se fera à l'aide des sorties de la figure 8.4. Les commandes SPSS à effectuer

pour les obtenir sont les suivantes :

Menu SPSS :	→ Analyse
	→ General Linear Model
	→ Univariate...
Dans la fenêtre Dependant Variable :	→ vente (la variable continue)
Dans la fenêtre Fixed Factors :	→ prix, media (les variables discrètes)
Dans le boutons Plots... :	Horizontal Axis : media (facteur <i>B</i>) Separate Lines : prix (facteur <i>A</i>)
Cliquer sur Add.	

Dans la table ANOVA de la figure 8.4, les nombres de la première colonne représentent la décomposition de la variation. Ainsi le nombre vis-à-vis **Corrected Model** (ici 120,030) correspond à la variation expliquée par les facteurs ($SSA + SSB + SSINT$). Le nombre vis-à-vis le nom du facteur *A* (ici **prix**, et le nombre est 7,363) correspond à la variation expliquée par le facteur *A* (SSA). De même, le nombre vis-à-vis le nom du facteur *B* correspond à la variation expliquée par le facteur *B* (SSB). Le nombre vis-à-vis le nom du facteur *A* multiplié par le nom du facteur *B* (ici **prix*media**) correspond à la variation expliquée par l'interaction entre les deux facteurs ($SSINT$). Finalement, le nombre vis-à-vis **Error** correspond à SSE , et celui vis-à-vis **Corrected Total** correspond à la variation totale (SST). (Les autres valeurs de cette colonne ne sont pas utilisées dans le cadre de ce cours.)

Dans le graphe, on retrouve les moyennes pour chaque traitement (les points). Elles sont placées vis-à-vis leur valeur (axe vertical) et vis-à-vis le niveau du facteur *B* qui leur correspond (axe horizontal). Les moyennes correspondant à un même niveau du facteur *A* sont reliées par une ligne.

Tests of Between-Subjects Effects

Dependent Variable: Montant des ventes en milliers de dollars.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	120,030 ^a	5	24,006	36,099	,000
Intercept	1728,000	1	1728,000	2598,496	,000
prix	7,363	1	7,363	11,073	,016
media	104,405	2	52,203	78,500	,000
prix * media	8,262	2	4,131	6,212	,035
Error	3,990	6	,665		
Total	1852,020	12			
Corrected Total	124,020	11			

a. R Squared = ,968 (Adjusted R Squared = ,941)

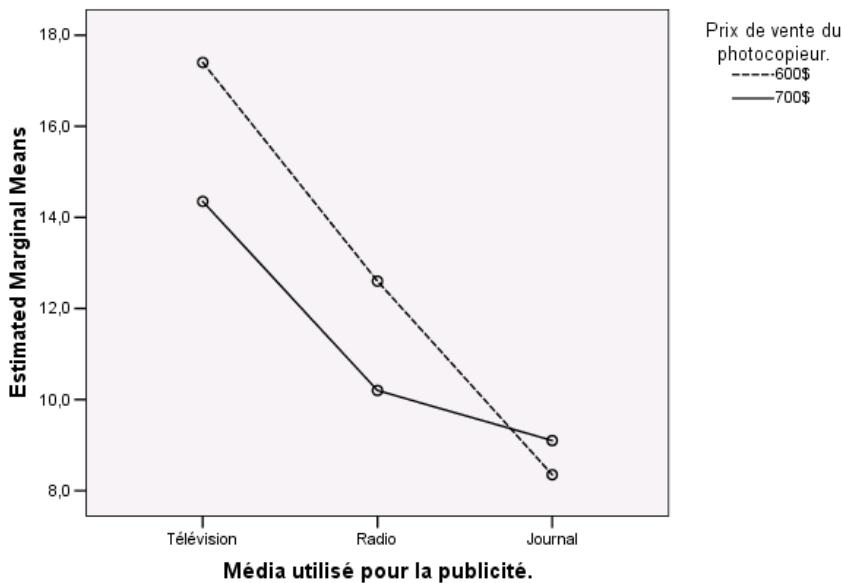
Estimated Marginal Means of Montant des ventes en milliers de dollars.

FIG. 8.4 – Sorties de l'exemple (ANOVA à 2 facteurs)

Revenons à l'exemple. Fixons les seuils de signification à $\alpha = 0,05$. Nous devons d'abord voir s'il y a un lien entre les ventes et les facteurs (le modèle est-il significatif?). Le test à résoudre pour ceci est le suivant :

H_0 : Le modèle n'est pas significatif au niveau de la population.

H_1 : Le modèle est significatif au niveau de la population.

Pour le résoudre, on utilise la première sortie (table ANOVA à 2 facteurs) de la figure 8.4 ; on prend la *p*-value de la première ligne, vis-à-vis **Corrected Model**, dans la colonne **Sig.** La règle de décision est la règle habituelle :

Nous rejetons l'hypothèse H_0 si la *p*-value est plus petite que le seuil de signification α fixé (par exemple $\alpha = 0,05$). Sinon, nous ne rejetons pas H_0 et la considérons comme vraisemblable.

Ici la *p*-value est égale à 0,000, donc au risque de se tromper une fois sur 20 on rejette H_0 . Ainsi on admet que le modèle est significatif dans son ensemble, et on peut donc poursuivre l'investigation.

L'étape suivante consiste à vérifier si l'interaction entre les facteurs A et B est significative. Si l'interaction est significative, il sera alors inutile d'effectuer des tests pour savoir si les facteurs A et B apportent une contribution solitaire significative. De fait, lorsqu'il y a une interaction, les effets des facteurs ne peuvent être considérés séparément.

Pour tester l'interaction on doit résoudre le test suivant :

H_0 : Aucune interaction n'existe entre les facteurs A et B au niveau de la population (tous les $(\alpha\beta)_{ij} = 0$).

H_1 : Les facteurs A et B interagissent au niveau de la population (au moins un des $(\alpha\beta)_{ij} \neq 0$)

qui dans le cadre de cet exemple s'écrit

H_0 : Aucune interaction n'existe entre la publicité et le prix au niveau de la population (tous les $(\alpha\beta)_{ij} = 0$).

H_1 : La publicité et le prix interagissent au niveau de la population (au moins un des $(\alpha\beta)_{ij} \neq 0$)

On rejette H_0 si la p -value qui est vis-à-vis (facteur A) * (facteur B) (ici `prix * media`) est plus petite que le seuil fixé α . Dans le cadre de l'exemple, la p -value est égale à 0,035, ce qui est plus petit que 0,05. On rejette donc H_0 , et au risque de se tromper une fois sur 20, on admet que le prix et le média interagissent. L'interaction étant significative, l'analyse solitaire de l'influence des facteurs A et B n'est pas appropriée. En effet, les ventes dépendent à la fois du niveau du prix et du média utilisé. En d'autres mots, l'effet du prix sur les ventes est différent d'un média à l'autre !

Lorsque l'interaction n'est pas significative, ce qui n'est pas le cas dans l'exemple, il faut alors vérifier l'influence des facteurs principaux l'un après l'autre. Dans le cadre de l'exemple, **si l'interaction n'avait pas été significative**, on aurait eu à résoudre les deux tests suivants à l'aide de la règle habituelle :

H_0 : Au niveau de la population, le niveau du prix n'a aucune influence sur les ventes (tous les $\alpha_i = 0$).

H_1 : Au niveau de la population, le niveau du prix a une influence sur les ventes (au moins un des $\alpha_i \neq 0$).

Pour ce premier test on prendrait la p -value vis-à-vis `prix`. Puisque celle-ci a une valeur de 0,016, ce qui est plus petit que $\alpha = 0,05$, on rejettterait H_0 , et au risque de se tromper une fois sur 20 on admettrait que le prix a une influence sur les ventes.

H_0 : Au niveau de la population, le niveau du média n'a aucune influence sur les ventes (tous les $\beta_j = 0$).

H_1 : Au niveau de la population, le niveau du média a une influence sur les ventes (au moins un des $\beta_j \neq 0$).

Pour ce deuxième test on prendrait la *p*-value vis-à-vis `media`. Puisque celle-ci a une valeur de 0,000, ce qui est plus petit que $\alpha = 0,05$, on rejette H_0 , et au risque de se tromper une fois sur 20 on admettrait que le média a une influence sur les ventes.

Dans cet exemple, nous concluons que les variables `prix` et `media` agissent toutes deux sur les ventes. En examinant le graphe, nous sommes même en mesure de dire que, pour espérer avoir les meilleures ventes possible, il faut utiliser le média de la télévision ET annoncer un prix de 600 \$. Aussi, les ventes sont supérieures lorsqu'un prix de 600 \$ est annoncé à la radio. Cependant, un prix de 700 \$ est plus performant dans les journaux.

Reste maintenant à l'équipe de finance et de marketing à bien utiliser cette information. Par exemple, pour des raisons de budget restreint ou d'image, une entreprise peut choisir de ne faire de la publicité que dans les journaux, et à ce moment le meilleur choix pour le prix serait 700 \$. En se basant sur une meilleure compréhension de la relation et dépendant des circonstances, l'entreprise est maintenant en mesure de prendre les choix finaux.

8.2 L'étude des graphiques

L'étude du graphique de l'interaction des facteurs peut aider à comprendre ce qui se passe dans une relation, et ce, surtout dans le cas où la relation d'interaction est significative. En effet, si l'interaction n'est pas significative, il suffit d'étudier l'influence des facteurs principaux *A* et *B* de façon indépendante.

Une interaction significative

Une interaction significative s'illustre par un entrecouplement des lignes des moyennes (ici les moyennes des ventes, voir le graphe 8.5). Les ventes dépendent à la fois des niveaux de prix et de media. Ainsi, lorsque le produit est annoncé à la télévision ou à la radio, les ventes seront meilleures si le prix annoncé est de 600 \$. Cependant, dans les journaux, un prix de 700 \$ générera de plus grandes ventes.

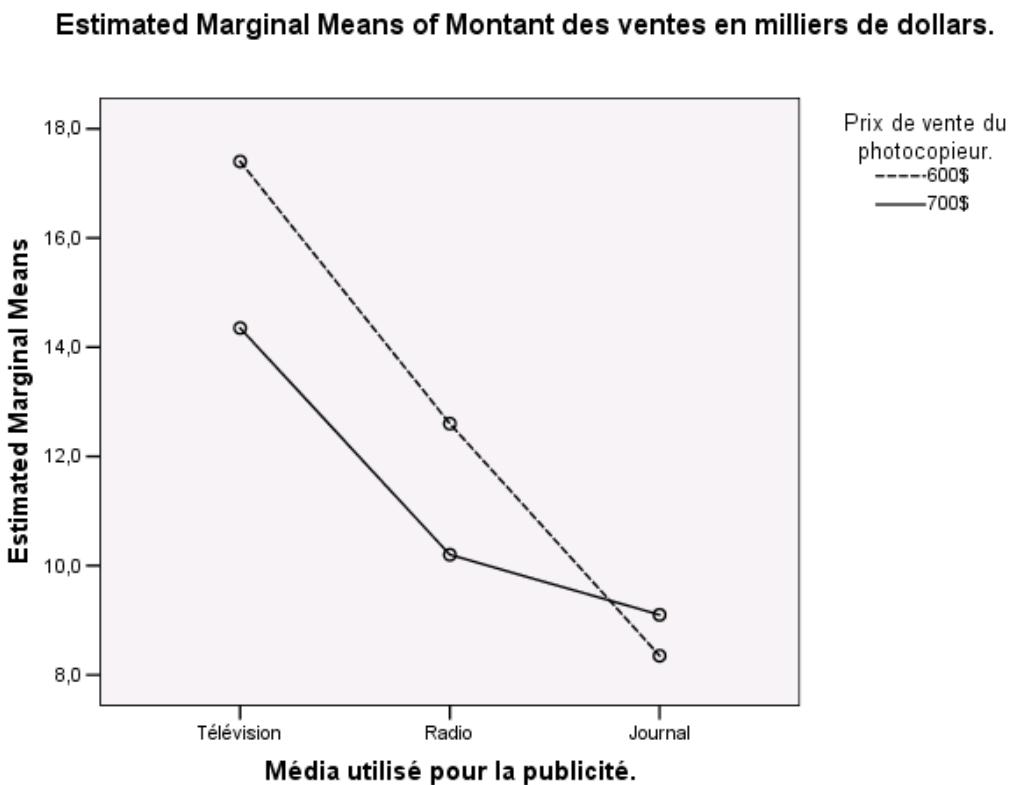


FIG. 8.5 – Interaction significative

Aucune interaction entre les facteurs

Lorsque les lignes sont parallèles, l'interaction est inexisteante. Dans le graphe 8.6, on voit que peu importe le média choisi, les ventes mensuelles sont toujours plus élevées lorsque le prix est égal à 600 \$. Aussi, peu importe le prix, la télévision est le média qui influence le plus les ventes mensuelles, et c'est le journal qui génère le moins de ventes.

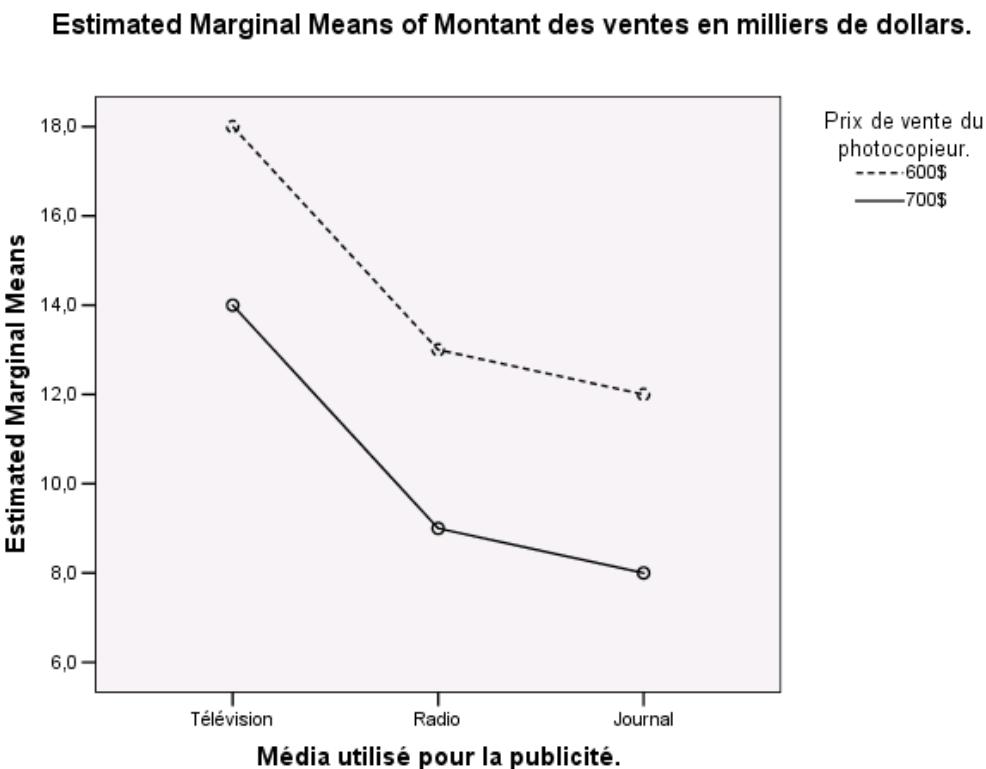


FIG. 8.6 – Aucune interaction

Une interaction faible entre les facteurs

Lorsque les lignes ne sont pas parallèles et qu'elles ne se croisent pas, nous sommes en présence d'une interaction faible. Dans ce type de situation, il est préférable de considérer qu'il n'existe pas d'interaction.

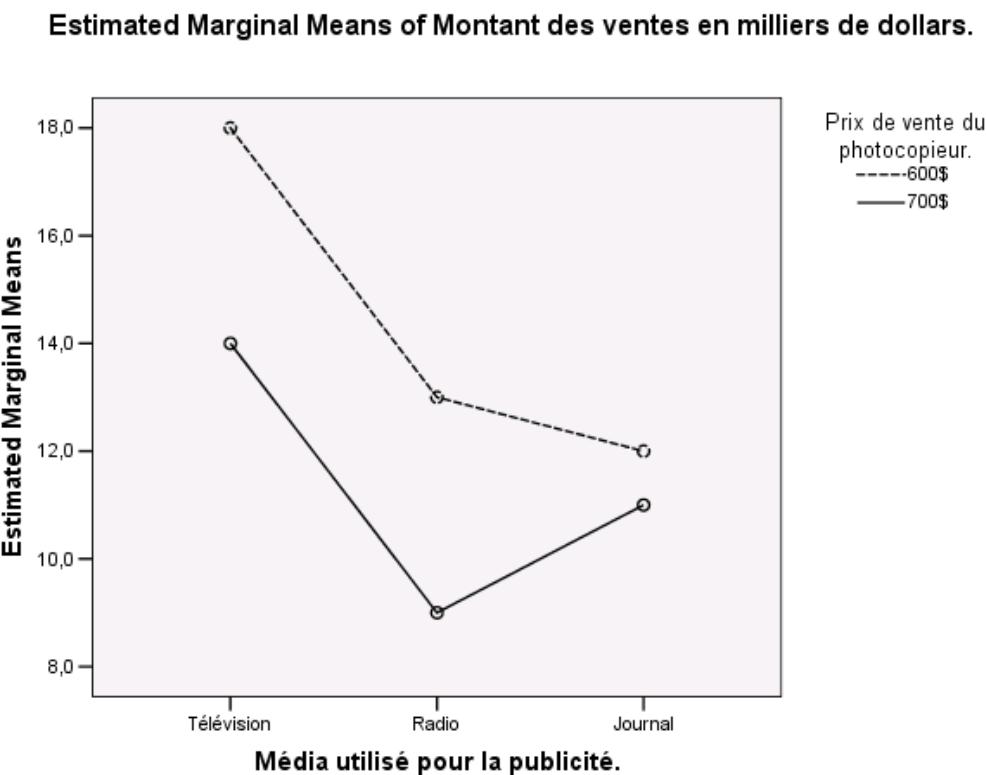


FIG. 8.7 – Interaction faible

Aucune interaction et aucun effet du facteur A

Si le facteur A n'influence pas la variation de la variable continue, alors les deux lignes des ventes moyennes devraient être superposées (en supposant que le facteur A a été utilisé pour identifier les lignes).

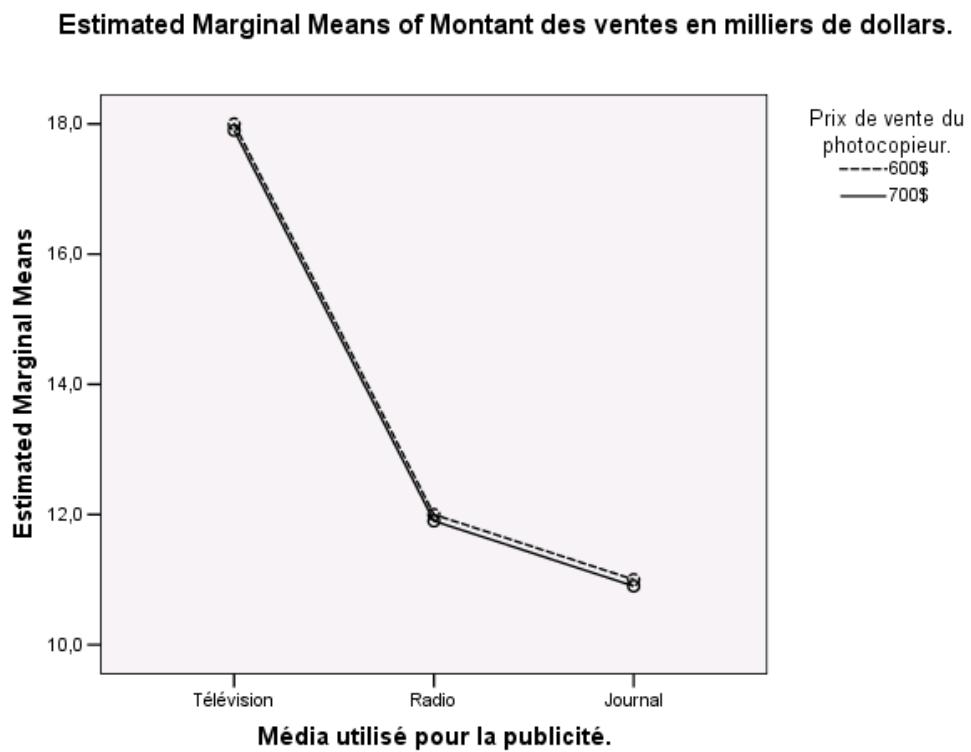


FIG. 8.8 – Aucun effet du facteur A

Aucune interaction et aucun effet du facteur B

Si le facteur B n'influence pas la variation de la variable continue, alors les deux lignes des ventes moyennes devraient être horizontales (en supposant que le facteur B a été utilisé pour l'axe horizontal).

Estimated Marginal Means of Montant des ventes en milliers de dollars.

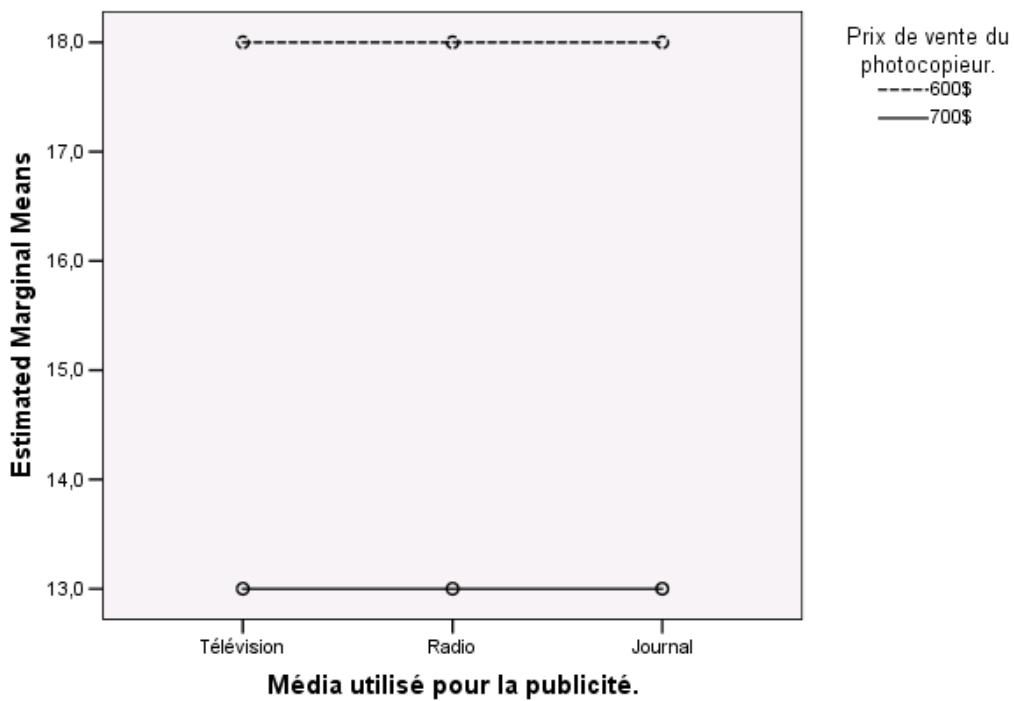


FIG. 8.9 – Aucun effet du facteur B

Ainsi, si les deux facteurs ne sont pas significatifs, alors ces deux lignes seront à la fois confondues (absence d'effet du facteur A) et horizontales (absence d'effet du facteur B).

8.3 Quelques exemples

Exemple 8.3.1 Un supermarché désire étudier l'efficacité de ses stratégies de vente en regardant l'effet de deux facteurs sur ses ventes. Le premier facteur est le prix affiché et a trois niveaux : régulier, réduit et prix du fabricant. Le deuxième facteur est la présentation du produit et a également trois niveaux : espace de présentation normal, espace de présentation normal plus espace en bout d'allée, et deux fois l'espace normal. La variable dépendante est le nombre de ventes hebdomadaires. Faites les analyses nécessaires pour voir comment les facteurs influencent les ventes. Les données sont disponible dans le fichier qui porte le nom `supermarche.sav`.

On s'intéresse ici à la relation (`present`, `prix`) \Rightarrow `ventes`. Cette analyse se fera à l'aide d'une ANOVA à deux facteurs puisque les variables `present` et `prix` sont discrètes tandis que la variable `ventes` est continue.

Tout d'abord, une courte analyse descriptive est de mise pour prendre le pouls des données.

Dans le tableau 8.10 on voit d'abord les statistiques descriptives de la variable `ventes` selon les groupements induits par la variable `prix`. Lorsque le prix affiché est régulier, la moyenne du nombre de ventes est de 1 144,11. Lorsque le prix est réduit, la moyenne est de 1 535,44 et lorsque c'est le prix du fabricant on a une moyenne de 1 972,22. Le CV pour ces groupes sont respectivement de 0,09, 0,2 et 0,21. Donc la moyenne est représentative pour le prix régulier, et il faut faire attention à l'utilisation de la moyenne pour les deux autres groupes.

Ensuite dans le tableau 8.11 on voit les statistiques descriptives de la variable `ventes` selon les groupements induits par la variable `present`. Lorsque l'espace de présentation est normal, la moyenne du nombre de ventes est de 1 265,11. Lorsque l'espace est « normal + », la moyenne est de 1 874,56 et lorsqu'il est deux fois l'espace normal on a une moyenne de 1 512,11. Le CV pour ces groupes sont respectivement de 0,2, 0,299 et 0,18. Donc il faut faire attention à l'utilisation de la moyenne pour les trois groupes.

			Descriptives					
			Prix affiché.					
			régulier		réduit		prix du fabricant	
			Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Nombre de ventes hebdomadaires.	Mean		1144,11	33,046	1535,44	100,968	1972,22	139,363
	95% Confidence Interval for Mean	Lower Bound	1067,91		1302,61		1650,85	
		Upper Bound	1220,31		1768,28		2293,59	
	5% Trimmed Mean		1147,79		1533,38		1964,36	
	Median		1191,00		1501,00		1833,00	
	Variance		9828,111		91750,028		174797,2	
	Std. Deviation		99,137		302,903		418,088	
	Minimum		989		1182		1559	
	Maximum		1233		1926		2527	
	Range		244		744		968	
	Interquartile Range		196		672		914	
	Skewness		-,801	,717	,202	,717	,601	,717
	Kurtosis		-1,494	1,400	-1,667	1,400	-1,708	1,400

FIG. 8.10 – Statistiques descriptives (prix affiché)

			Descriptives					
			Espace de présentation du produit.					
			normal		normal +		2 * normal	
			Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Nombre de ventes hebdomadaires.	Mean		1265,11	82,997	1874,56	187,160	1512,11	90,570
	95% Confidence Interval for Mean	Lower Bound	1073,72		1442,96		1303,26	
		Upper Bound	1456,50		2306,15		1720,97	
	5% Trimmed Mean		1261,96		1876,28		1511,68	
	Median		1211,00		1910,00		1501,00	
	Variance		61996,861		315259,3		73826,861	
	Std. Deviation		248,992		561,480		271,711	
	Minimum		989		1191		1180	
	Maximum		1598		2527		1852	
	Range		609		1336		672	
	Interquartile Range		541		1275		603	
	Skewness		,457	,717	-,083	,717	,052	,717
	Kurtosis		-1,682	1,400	-1,712	1,400	-1,670	1,400

FIG. 8.11 – Statistiques descriptives (espace de présentation du produit)

On perçoit donc certaines fluctuations du nombre moyen de ventes selon le prix affiché et l'espace de présentation alloué. L'analyse avec l'ANOVA à deux facteurs nous permettra de voir si ces fluctuations sont présentes dans la population, et s'il y a une interaction entre le prix affiché et l'espace de présentation. Fixons le seuil α à 0,05 pour tous les tests.

Dans la base de données on voit que pour chaque traitement on a 3 essais, ce qui est plus grand que le minimum requis qui est de 2 essais par traitement.

On doit d'abord voir si au moins un des facteurs influence les ventes. Le test à résoudre est le suivant :

H_0 : Le modèle n'est pas significatif au niveau de la population.

H_1 : Le modèle est significatif au niveau de la population.

On résout ce test à l'aide de la table ANOVA (figure 8.12). Puisque la p -value = $0,000 < 0,05$ (vis-à-vis **Corrected Model**), on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet que le modèle est significatif au niveau de la population.

Tests of Between-Subjects Effects					
Dependent Variable: Nombre de ventes hebdomadaires.					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5291151,185 ^a	8	661393,898	1336,849	,000
Intercept	64917109,5	1	64917109,48	131214,4	,000
present	1691392,519	2	845696,259	1709,373	,000
prix	3089053,852	2	1544526,926	3121,892	,000
present * prix	510704,815	4	127676,204	258,067	,000
Error	8905,333	18	494,741		
Total	70217166,0	27			
Corrected Total	5300056,519	26			

a. R Squared = ,998 (Adjusted R Squared = ,998)

FIG. 8.12 – Table ANOVA à deux facteurs

On doit maintenant vérifier s'il y a une interaction entre les facteurs. Pour ce faire on résout le test suivant :

H_0 : Aucune interaction n'existe entre le prix affiché et l'espace de présentation au niveau de la population (tous les $(\alpha\beta)_{ij} = 0$).

H_1 : Le prix affiché et l'espace de présentation interagissent au niveau de la population (au moins un des $(\alpha\beta)_{ij} \neq 0$)

On résout ce test à l'aide de la table ANOVA (figure 8.12). Puisque la p -value = $0,000 < 0,05$ (vis-à-vis `present*prix`), on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet qu'il y a une interaction significative entre les deux facteurs.

Pour décrire cette interaction on peut s'appuyer sur le graphe de la figure 8.13.

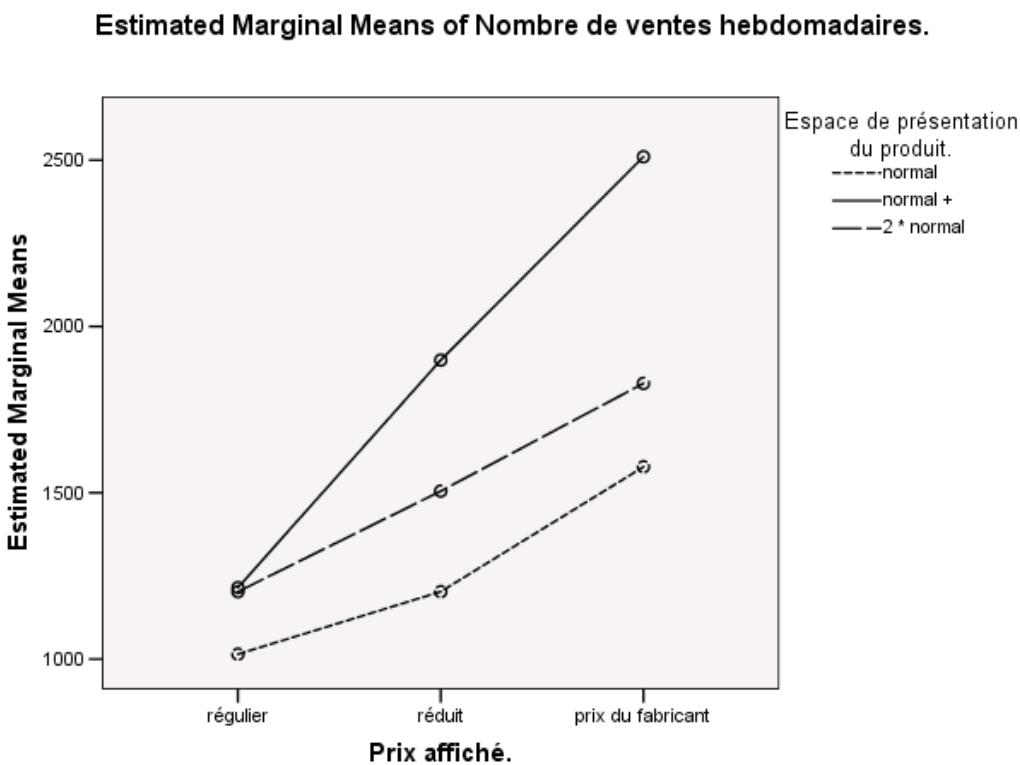


FIG. 8.13 – Graphe de la relation

On voit d'abord que pour un espace de présentation fixé, c'est le prix du fabricant qui génère le plus de ventes, ensuite le prix réduit et finalement le prix normal.

Lorsque le prix est fixé à réduit ou du fabricant, le meilleur nombre de ventes est

obtenu avec un espace de présentation « normal + », suivi par deux fois l'espace normal, tandis qu'avec l'espace normal on a le moins grand nombre de ventes. Par contre, lorsque le prix affiché est régulier, le nombre de ventes est le même lorsqu'on utilise l'espace « normal + » ou deux fois l'espace normal ; c'est là que se manifeste l'interaction des deux facteurs. Ainsi lorsque le produit est au prix régulier, on a le choix entre l'espace « normal + » et deux fois l'espace normal pour générer le plus de ventes possibles, alors qu'avec les deux autres types de prix c'est l'espace « normal + » qu'il faut choisir pour générer le plus de ventes.

Exemple 8.3.2 On désire implanter un nouveau logiciel pour les courriels et la messagerie instantanée à l'université. Trois logiciels ont retenu l'attention du comité qui est chargé de sélectionner le produit qui répondra le plus aux attentes des étudiants, des professeurs et du personnel régulier de l'université. Ce comité décide de mener une étude sur le campus pour voir lequel des logiciels semble être le plus apprécié. Voici un extrait des résultats de cette étude ; ceux-ci ont été obtenus auprès des étudiants de certaines des MSc de la faculté d'administration. Chaque étudiant participant à l'étude devait donner son appréciation du logiciel en donnant une note sur 100.

	Logiciel I			Logiciel II			Logiciel III		
Commerce électronique	60	68	61	80	82	87	75	70	69
Marketing	78	75	77	69	65	70	50	59	49
ICO	78	85	83	59	67	63	45	51	46
Systèmes d'information	57	59	55	79	85	83	72	78	79

Est-ce que l'appréciation est la même selon le logiciel et selon le type de MSc à laquelle l'étudiant est inscrit ?

Tout d'abord, une courte analyse descriptive est de mise pour prendre le pouls des données.

Dans la figure 8.14 on voit d'abord les statistiques descriptives de la variable `appreciation`

selon les groupements induits par la variable **option**. Pour l'option commerce électronique, la moyenne de l'appréciation est de 72,44. Pour l'option marketing, la moyenne est de 65,78, et pour les ressources humaines elle est de 64,11. Finalement, pour l'option systèmes d'information, la moyenne de l'appréciation est de 71,89. Les CV pour ces groupes sont respectivement de 0,13, 0,17, 0,24 et 0,16. Donc la moyenne est représentative pour l'option commerce électronique, et il faut faire attention à l'utilisation de la moyenne pour les trois autres groupes.

Ensuite, dans le deuxième tableau de la figure 8.14, on voit les statistiques descriptives de la variable **appreciation** selon les groupements induits par la variable **logiciel**. Pour le logiciel 1, la moyenne de l'appréciation est de 69,67. Pour le logiciel 2, la moyenne est de 74,08 et pour le logiciel 3 on a une moyenne de 61,92. Les CV pour ces groupes sont respectivement de 0,16, 0,13 et 0,21. Donc il faut faire attention à l'utilisation de la moyenne pour les logiciels 1 et 3, tandis que la moyenne pour le logiciel 2 est représentative.

On perçoit donc certaines fluctuations pour l'appréciation moyenne selon l'option et le logiciel testé. L'analyse avec l'ANOVA à deux facteurs nous permettra de voir si ces fluctuations sont présentes dans la population, et s'il y a une interaction entre l'option et le logiciel. Fixons le seuil α à 0,05 pour tous les tests.

		Descriptives							
		option							
		Commerce électronique		Marketing		ICO		Systèmes d'information	
		Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
appréciation	Mean	72.44	3.096	65.78	3.662	64.11	5.119	71.89	3.921
	95% Confidence Interval for Mean	Lower Bound	65.30	57.33		52.31		62.85	
		Upper Bound	79.58	74.22		75.92		80.93	
	5% Trimmed Mean		72.33	66.03		64.01		72.10	
	Median		70.00	69.00		63.00		78.00	
	Variance		86.278	120.694		235.861		138.361	
	Std. Deviation		9.289	10.986		15.358		11.763	
	Minimum		60	49		45		55	
	Maximum		87	78		85		85	
	Range		27	29		40		30	
	Interquartile Range		17	22		32		23	
	Skewness		.167	.717	-.585	.717	.144	.717	-.561
	Kurtosis		-1.037	1.400	-1.093	1.400	-1.556	1.400	-1.642
									1.400

		Descriptives							
		logiciel							
		Logiciel 1		Logiciel 2		Logiciel 3			
		Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error		
appréciation	Mean	69.67	3.132	74.08	2.770	61.92	3.809		
	95% Confidence Interval for Mean	Lower Bound	62.77	67.99		53.53			
		Upper Bound	76.56	80.18		70.30			
	5% Trimmed Mean		69.63	74.20		61.91			
	Median		71.50	74.50		64.00			
	Variance		117.697	92.083		174.083			
	Std. Deviation		10.849	9.596		13.194			
	Minimum		55	59		45			
	Maximum		85	87		79			
	Range		30	28		34			
	Interquartile Range		19	17		25			
	Skewness		-.024	.637	-.137	.637	-.036	.637	
	Kurtosis		-1.763	1.232	-1.595	1.232	-1.918	1.232	

FIG. 8.14 – Statistiques descriptives

En regardant les données, on voit que pour chaque traitement on a 3 essais, ce qui est plus grand que le minimum requis qui est de 2 essais par traitement.

Tests of Between-Subjects Effects					
Dependent Variable: appréciation					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4834,889 ^a	11	439,535	35,399	,000
Intercept	169195,111	1	169195,111	13626,452	,000
option	483,333	3	161,111	12,975	,000
logiciel	910,389	2	455,194	36,660	,000
option * logiciel	3441,167	6	573,528	46,190	,000
Error	298,000	24	12,417		
Total	174328,000	36			
Corrected Total	5132,889	35			

a. R Squared = ,942 (Adjusted R Squared = ,915)

FIG. 8.15 – Table ANOVA à deux facteurs

On doit d'abord voir si au moins un des facteurs influence les ventes. Le test à résoudre est le suivant :

H_0 : Le modèle n'est pas significatif au niveau de la population.

H_1 : Le modèle est significatif au niveau de la population.

On résout ce test à l'aide de la table ANOVA (figure 8.15). Puisque la p -value = $0,000 < 0,05$ (vis-à-vis **Corrected Model**), on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet que le modèle est significatif au niveau de la population.

On doit maintenant vérifier s'il y a une interaction entre les facteurs. Pour ce faire on résout le test suivant :

H_0 : Aucune interaction n'existe entre l'option et le logiciel au niveau de la population (tous les $(\alpha\beta)_{ij} = 0$).

H_1 : L'option et le logiciel interagissent au niveau de la population (au moins un des $(\alpha\beta)_{ij} \neq 0$)

On résout ce test à l'aide de la table ANOVA (figure 8.15). Puisque la p -value =

$0,000 < 0,05$ (vis-à-vis option*logiciel), on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet qu'il y a une interaction significative entre les deux facteurs.

Pour décrire cette interaction on peut s'appuyer sur le graphe de la figure 8.16. (Il aurait aussi été possible de le faire en prenant le graphe de la figure 8.17.)

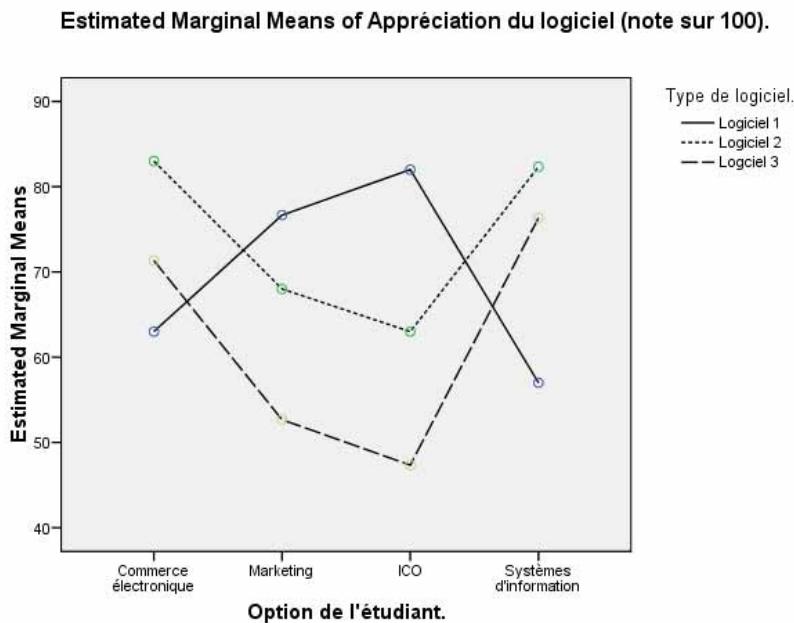


FIG. 8.16 – Graphe de la relation

On remarque que l'appréciation des logiciels est semblable pour les options commerce électronique et systèmes d'information, et semblable pour les options marketing et ressources humaines. En effet, les étudiants de commerce électronique et systèmes d'information préfèrent d'abord le logiciel 2, suivi du logiciel 3 puis finalement du logiciel 1. Pour les étudiants de marketing et ressources humaines, c'est le logiciel 1 qui est d'abord préféré, suivi du logiciel 2 puis du logiciel 3.

Si on veut choisir un logiciel en s'appuyant sur ces données, le logiciel 3 ne semble pas le meilleur choix puisque c'est celui qui atteint les plus basses moyennes (et au niveau descriptif on a vu que c'est le logiciel qui a la plus basse moyenne lorsque les options sont confondues), et le seul qui n'arrive premier pour aucune des options. Resterait à choisir

entre les logiciels 1 et 2. Des informations complémentaires seraient utiles pour faire le meilleur choix possible (n'oublions pas que ces données ne représentent qu'un extrait de l'étude).

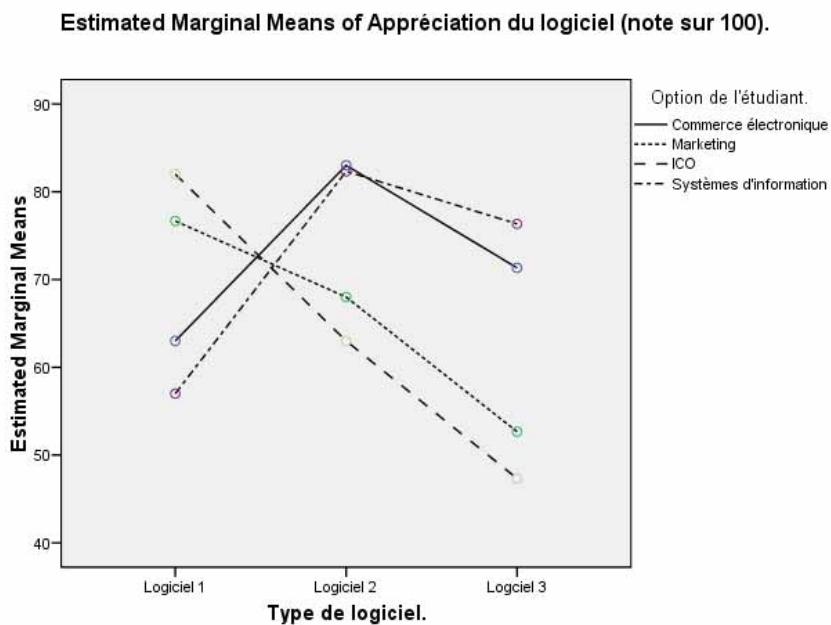


FIG. 8.17 – Autre graphe possible

8.4 Exercices du chapitre

Exercice 1 Une entreprise développe un certain type de logiciel pour un grand nombre d'entreprises. Le logiciel est distribué par trois distributeurs différents. L'entreprise veut connaître l'influence des deux variables discrètes suivantes sur le niveau de satisfaction d'un échantillon de leurs clients :

- Le type d'industrie qui utilise le logiciel ;
- Le distributeur qui a distribué le logiciel.

Les entreprises clientes peuvent être classées en 4 types, trois fournisseurs différents distribuent le logiciel et la satisfaction est un score sur 100. Notez qu'un nombre de deux essais par traitement a été effectué ; c'est le minimum, ce qui n'est pas toujours souhaitable. La base de données s'appelle `logiciel.sav`.

Exercice 2 On s'intéresse à la consommation d'essence de trois types d'automobiles (autos de type X, Y et Z). On décide d'étudier la relation type d'auto \Rightarrow consommation (fichier `autos.sav`). Ensuite, on se demande s'il serait pertinent de considérer aussi l'influence de la classe d'âge du conducteur sur la consommation.

Dites s'il est pertinent de considérer l'influence de la classe d'âge du conducteur en plus de celle du type d'auto, et expliquer pourquoi.

Ensuite, faire l'analyse de la relation `(auto, conducteur) \Rightarrow consommation`.

Exercice 3 On veut mettre en ligne un nouveau site web. On veut tester trois interfaces et trois ensembles de couleurs. On teste toutes les combinaisons possibles auprès d'individus qui doivent donner une note sur 10 après avoir navigué quelques minutes sur le site. On obtient les résultats suivants :

	Interface I		Interface II		Interface III	
Couleurs I	6.5	7.5	7.0	6.0	8.0	10
Couleurs II	5.0	5.5	8.5	9.5	7.0	8.0
Couleurs III	3.5	5.0	5.5	6.0	2.0	4.0

Est-ce que l'appréciation est la même selon l'interface et selon les couleurs utilisées ? Lequel des deux facteurs a le plus grand impact sur l'appréciation ? Faites l'analyse nécessaire pour répondre à ces questions, et fournissez toutes les sorties SPSS appropriées. Fixez le seuil à $\alpha = 0,05$.

Chapitre 9

Relation entre plusieurs variables

continues

Au chapitre 6, nous avons utilisé la régression linéaire simple pour modéliser le lien entre deux variables continues (Y et X). En fait, nous avons utilisé l'information obtenue sur X pour améliorer nos estimations sur Y . Cependant, dans tous les domaines d'application, il est raisonnable de penser qu'il puisse exister plus d'une variable explicative (X) pour modéliser le comportement de la variable dépendante Y . Ainsi dans ce chapitre nous allons tenter d'expliquer les variations d'une variable dépendante Y à l'aide de plusieurs variables explicatives (les variables indépendantes) X_1, X_2, \dots, X_k au moyen de la **régression linéaire multiple**.

9.1 Le modèle

Lors de l'étude de la régression linéaire simple, nous aboutissons à une équation de la forme

$$\hat{y} = b_0 + b_1 x$$

qui est une estimation de la droite de régression au niveau de la population qui s'écrit

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

De façon très semblable, lors d'une analyse en régression linéaire multiple, nous obtenons une équation de la forme

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

qui estime le modèle

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

au niveau de la population, où :

- Y est la variable dépendante (ou expliquée) ayant un caractère aléatoire et dont les valeurs sont conditionnées par celles des variables explicatives X_1, X_2, \dots, X_k et la composante aléatoire ϵ ;
- $\beta_0, \beta_1, \dots, \beta_k$ sont appelés les paramètres du modèle de régression multiple ;
- X_1, X_2, \dots, X_k sont les variables explicatives (ou indépendantes) mesurées sans erreur ou dont les valeurs sont fixées avant l'expérience à des valeurs arbitraires ;
- ϵ dénote la fluctuation aléatoire non observable (ou inexplicable) attribuable à un ensemble de facteurs dont on ne tient pas compte dans le modèle. Cette fluctuation aléatoire n'est pas expliquée par le modèle et se reflète sur la variable dépendante Y .

Le terme linéaire se réfère toujours aux paramètres du modèle ($\beta_0, \beta_1, \beta_2, \dots, \beta_k$). Ainsi les modèles

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

et

$$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

sont linéaires. Aussi, il est intéressant de mentionner que :

- Tout comme pour la régression linéaire simple, les coefficients du modèle $b_0, b_1, b_2, \dots, b_k$ sont calculés à l'aide de la méthode des moindres carrés afin de minimiser la somme de carré suivante :

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- La variation de Y se décompose en deux sources : la variation expliquée par la droite de régression multiple et la variation résiduelle (variation non expliquée par le modèle).

L'interprétation des paramètres de l'équation est semblable à celle de la régression linéaire simple :

- Pour les mêmes raisons que dans le cas de la régression linéaire simple, β_0 est souvent difficile à interpréter. Cependant, si l'étude possède des observations pour des valeurs nulles des variables explicatives, β_0 représente souvent (par exemple) des frais ou des revenus fixes ;
- β_j représente le changement sur Y correspondant à une variation unitaire de la variable explicative X_j lorsque les autres variables explicatives demeurent inchangées.

Le lien étudié est linéaire ; donc tout comme pour la régression linéaire simple, nous pourrons conclure s'il y a un lien **linéaire** entre Y et X_1, X_2, \dots, X_k , mais d'autres types

de lien peuvent exister.

Nous verrons que toutes les propriétés que nous avons étudiées dans le cadre de la régression linéaire simple se transposent à la régression linéaire multiple. En fait, la régression linéaire simple n'est qu'un cas particulier de la régression linéaire multiple. Par contre, avec plusieurs variables explicatives la complexité du modèle augmente, ce qui fait que quelques étapes supplémentaires s'ajouteront lors de l'analyse. En particulier, il faudra vérifier que le modèle est valide.

9.1.1 Hypothèses de validité

Pour qu'un modèle de régression linéaire multiple soit valide, il faut que l'hypothèse suivante soit respectée :

On suppose que les ϵ_i sont des variables aléatoires normales et indépendantes de moyenne $E(\epsilon) = 0$ et de variances identiques $\text{Var}(\epsilon) = \sigma_{\text{résiduelle}}^2$.

On peut donc déduire que, pour toute valeur particulière que prend chacune des variable explicatives X_j ($j = 1, 2, \dots, k$), la variable dépendante Y_i est une variable aléatoire distribuée selon une loi normale de moyenne

$$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

et de variance $\text{Var}(Y_i) = \sigma_{\text{résiduelle}}^2$.

Dans la pratique, comment vérifier si cette hypothèse est vérifiée ? Et comment vérifier à quel point le modèle sied les données ? C'est par l'analyse des résidus, qui estiment les ϵ_i , qu'il est possible d'effectuer ces vérifications. Nous verrons comment faire cette analyse au chapitre 11.

9.2 Une analyse complète

Nous présenterons toutes les étapes d'une analyse en régression linéaire multiple à l'aide d'un exemple (on suppose que le modèle est valide). Les seuils de signification sont tous fixés à $\alpha = 0,05$.

Exemple 9.2.1 Meddicorp est une entreprise qui vend du matériel médical aux hôpitaux et aux cliniques médicales. La direction de l'entreprise veut évaluer l'efficacité de son nouveau programme de bonus selon la performance à la vente. La direction veux savoir si les bonus versés ont une influence sur les ventes. Afin de ne pas être induit en erreur, la direction veut mettre en relief l'effet de la publicité.

Les données sont les suivantes (aussi disponibles dans le fichier `bonus.sav`) :

	ident	ventes	publicite	bonus
1	1	963,50	374,27	230,98
2	2	893,00	408,50	236,28
3	3	1057,25	414,31	271,57
4	4	1183,25	448,42	291,20
5	5	1419,50	517,88	282,17
6	6	1547,75	637,60	321,16
7	7	1580,00	635,72	294,32
8	8	1071,50	446,86	305,69
9	9	1078,25	489,59	238,41
10	10	1122,50	500,56	271,38
11	11	1304,75	484,18	332,64
12	12	1552,25	618,07	261,80
13	13	1040,00	453,39	235,63
14	14	1045,25	440,86	249,68
15	15	1102,25	487,79	232,99
16	16	1225,25	537,67	272,20
17	17	1508,00	612,21	266,64
18	18	1564,25	601,46	277,44
19	19	1634,75	585,10	312,35
20	20	1159,25	524,56	292,87
21	21	1202,75	535,17	268,27
22	22	1294,25	486,03	309,85
23	23	1467,50	540,17	291,03
24	24	1583,75	583,85	289,29
25	25	1124,75	499,15	272,55

FIG. 9.1 – Les données de l'exemple

Pour générer les sorties 9.2, 9.3 et 9.4 qui serviront à l'analyse complète de cet exemple, les commandes à effectuer sont les suivantes :

Menu SPSS :	→ Analyse
	→ Regression
	→ Linear...
Dans la fenêtre Dependant :	→ ventes (la variable dépendante)
Dans la fenêtre Independant(s) :	→ publicite, bonus (les variables indépendantes)

9.2.1 Le modèle est-il bon dans son ensemble ?

Comme dans le cas de la régression linéaire simple, nous nous intéressons à quantifier la performance du modèle, et à voir si le modèle est significatif.

Le modèle est-il significatif ?

Tout comme pour la régression linéaire simple, la variation de Y est décomposée en deux sources pour pouvoir tester si le modèle estime au moins en partie cette variation (table ANOVA de la figure 9.2). Ainsi on a

$$\begin{aligned} \text{Variation totale} &= \text{Variation expliquée par la régression} + \text{Variation résiduelle} \\ 1\ 248\ 973,740 &= 1\ 067\ 797,321 + 181\ 176,419 \end{aligned}$$

ANOVA ^b					
Model		Sum of Squares	df	Mean Square	F
1	Regression	1067797	2	533898,660	64,831
	Residual	181176,4	22	8235,292	
	Total	1248974	24		

a. Predictors: (Constant), bonus, publicite

b. Dependent Variable: ventes

FIG. 9.2 – La table ANOVA

Table ANOVA				
Source de variation	Somme de carrés	Degrés de liberté	Carrés moyens	Quotient F
Expliquée par la régression	SCR	k	$CMR = SCR/k$	$F = \frac{CMR}{CM_{rés}}$
Résiduelle	$SC_{rés}$	$n - k - 1$	$CM_{rés} = SC_{rés}/(n - k - 1)$	
Totale	SCT	$n - 1$		

Plus la variation expliquée par la droite sera grande, plus la régression risque d'être significative. Pour résoudre le test d'hypothèses suivant, on utilise la *p*-value (*Sig.*) de la dernière colonne de la table ANOVA avec la règle de décision habituelle.

H_0 : La régression est non significative dans la population (tous les $\beta_j = 0$).

H_1 : La régression est significative dans la population (au moins un des $\beta_j \neq 0$).

Ici, puisque la *p*-value est égale à 0,000, ce qui est bien sûr plus petit que le seuil $\alpha = 0,05$, on rejette H_0 . Ainsi, au risque de se tromper une fois sur 20, on peut affirmer que la régression est significative. Ainsi il est sensé de poursuivre l'analyse.

Les coefficients r , r^2 et r_{aj}^2 .

Dans le cas de la régression linéaire multiple, le coefficient r peut être interprété comme étant le coefficient de corrélation entre les valeurs réelles de Y et les valeurs estimées par le modèle \hat{y} . Ici, puisque $r = 0,925$ (voir la sortie 9.3), la relation entre les ventes et les valeurs estimées par le modèle peut être qualifiée de très forte (schéma de Davis, sortie 6.1, chapitre 6).

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.925 ^a	.855	.842	90,74851

a. Predictors: (Constant), bonus, publicité
b. Dependent Variable: ventes

FIG. 9.3 – Les coefficients r et r_{aj}^2

Le coefficient

$$r^2 = \frac{\text{SCR}}{\text{SCT}} = 1 - \frac{\text{SC}_{\text{rés}}}{\text{SCT}}$$

représente le % de la variation de Y expliquée par la régression, et peut aussi être interprété comme étant le coefficient de détermination entre les valeurs réelles de Y et les valeurs estimées par le modèle \hat{y} .

Ajouter des variables explicatives X au modèle de régression ne peut que faire augmenter le r^2 car la SCR ne peut qu'augmenter avec plus de variables, et donc la SC_{rés} ne peut que diminuer. Il est alors suggéré d'utiliser une mesure qui s'ajuste selon le nombre de variables explicatives utilisées dans le modèle. Le **coefficient de détermination ajusté**, noté r_{aj}^2 , ajuste le r^2 en divisant chaque somme de carrés par son nombre de degrés de liberté :

$$r_{\text{aj}}^2 = 1 - \frac{\text{SC}_{\text{rés}}}{\frac{n - k - 1}{n - 1}} = 1 - \left(\frac{n - 1}{n - k - 1} \right) \frac{\text{SC}_{\text{rés}}}{\text{SCT}}.$$

Ainsi l'ajout d'une variable explicative X qui ne fait diminuer que très peu la SC_{rés}

peut faire diminuer le r_{aj}^2 si cette diminution n'est pas assez grande pour compenser la perte d'un degré de liberté au dénominateur ($n - k - 1$).

Dans le cadre de l'exemple on a $r_{\text{aj}}^2 = 0,842$, ce qui signifie que 84,2 % de la variation des ventes est expliquée par les bonus et la publicité.

Le coefficient r_{aj}^2 est souvent utilisé pour comparer les performances (ou la qualité de l'ajustement) de deux modèles de régression linéaire multiple ayant un nombre différent de variables explicatives.

9.2.2 Inférence sur les paramètres du modèle

La prochaine étape consiste à voir quels sont les β_j du modèle qui sont significatifs (c'est-à-dire lesquels sont non-nuls). On doit donc tester les hypothèses suivantes pour tous les β_j :

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Pour ce faire, on prend la p -value de la table 9.4 qui est sur la ligne correspondant au β_j en question. Dans le cadre de l'exemple, on a deux tests à faire :

$$H_0 : \beta_1 = 0 \quad \text{et} \quad H_0 : \beta_2 = 0$$

$$H_1 : \beta_1 \neq 0 \quad H_1 : \beta_2 \neq 0$$

La p -value pour le premier test est celle correspondante à la variable `publicite`, elle est égale à 0,000, donc au seuil $\alpha = 0,05$ on rejette H_0 puisque $0,000 < 0,05$ (règle de décision habituelle). Ainsi au risque de se tromper une fois sur 20 on peut affirmer que $\beta_1 \neq 0$.

La p -value pour le deuxième test est celle correspondante à la variable `bonus`, elle est égale à 0,017, donc au seuil $\alpha = 0,05$ on rejette H_0 puisque $0,017 < 0,05$. Ainsi au risque de se tromper une fois sur 20 on peut affirmer que $\beta_2 \neq 0$.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant) -516,444	189,876		-2,720	,013		
	publicite 2,473	,275	,803	8,983	,000	,825	1,213
	bonus 1,856	,716	,232	2,593	,017	,825	1,213

a. Dependent Variable: ventes

FIG. 9.4 – Table des coefficients

Dans un modèle de régression multiple avec plusieurs variables, lorsque nous admettons $H_0 : \beta_j = 0$ comme étant vraisemblable, il ne faut pas conclure que la variable explicative X_j n'apporte aucune contribution significative. Une régression linéaire simple entre Y et X_j peut amener une conclusion contraire. Il faut apporter une conclusion plus nuancée : on devra plutôt conclure que la contribution marginale de X_j , lorsqu'elle est introduite à la suite des autres variables explicatives, est non significative. Son ajout est superflu.

Une fois ces tests effectués, il est possible de classer les variables explicatives selon l'importance de l'influence qu'elles ont sur la variable dépendante Y ; ceci se fait en utilisant leur cote- t (table 9.4). Habituellement on ne fait ce classement que pour les variables dont le coefficient est significatif. Ainsi la variable la plus importante est celle dont la cote- t est la plus grande (en valeur absolue), et ainsi de suite pour les autres.

Dans le cadre de l'exemple, c'est la publicité avec une cote- t de 8,983 qui influence le plus les ventes (vient ensuite la deuxième variable (bonus) qui est significative mais dont la cote- t de 2,593 est moins élevée).

On utilise également la sortie 9.4 pour écrire la droite de régression en utilisant la colonne des coefficients (de la même façon que celle vue pour la régression linéaire simple). Ainsi la droite de l'exemple est la suivante :

$$\hat{y}_{\text{ventes}} = -516,444 + 2,473x_{\text{publicite}} + 1,856x_{\text{bonus}}.$$

L'interprétation du b_0 n'est pas aisée ici. On peut penser que sans publicité et bonus

il y aurait des pertes... mais des ventes négatives, ce n'est pas très sensé !

On a $b_1 = 2,473$. Ainsi, lorsque le bonus est fixe et qu'on ajoute 100 \$ à la publicité, il s'ajoute en moyenne 2 473 \$ aux ventes.

On a $b_2 = 1,856$. Donc lorsque la publicité est fixe et qu'on ajoute 100 \$ aux bonus, il s'ajoute en moyenne 1 856 \$ aux ventes.

Remarque. En ce qui concerne les tests faits pour chacun des β_j et le classement des variables explicatives, des problèmes peuvent survenir s'il y a présence de **multicolinéarité**. Nous verrons comment détecter et remédier à la multicolinéarité au chapitre 11.

9.2.3 Prédictions et intervalles de confiance

Après avoir statué qu'une régression multiple est significative, nous disposons alors d'une équation qui modélise le lien entre la variable Y et les variables X_j . Comme dans le cas de la régression linéaire simple, il est alors possible de

- faire des estimations pour des valeurs moyennes de Y ;
- faire des intervalles de confiance autour des valeurs moyennes de Y ;
- faire des intervalles de confiance autour d'une valeur réelle de Y .

La procédure est la même que celle vue au chapitre 6, à ceci près qu'au lieu d'entrer une seule valeur dans la base de données on doit maintenant entrer une pour chacune des variables explicatives.

Reprendons l'exemple de Meddicorp. Supposons que nous nous intéressons à la valeur moyenne et réelle des ventes lorsque la publicité est fixée à 400 unités et les bonus sont fixés à 300 unités (on remarque que c'est de l'interpolation, il n'y a donc pas de problème à utiliser le modèle). Fixons le niveau de confiance à 95 %.

Tout d'abord, on peut trouver l'estimation ponctuelle en utilisant directement l'équation :

$$\begin{aligned}\hat{y}_{\text{ventes}} &= -516,444 + 2,473x_{\text{publicite}} + 1,856x_{\text{bonus}} \\ &= -516,444 + 2,473 \times 400 + 1,856 \times 300 \\ &= 1029,556.\end{aligned}$$

Ainsi lorsque la publicité est fixée à 40 000 \$ et les bonus à 30 000 \$, les ventes sont estimées à 1 029 556 \$.

Pour les intervalles de confiance, on entre d'abord les valeurs pour la publicité et les bonus dans la base de données, puis on effectue les mêmes commandes que celles vues au chapitre 6 :

Menu SPSS :	→ Analyse
	→ Regression
	→ Linear...
Dans la fenêtre Dependant :	→ ventes
Dans la fenêtre Independant(s) :	→ publicite, bonus
Dans le bouton Save... :	→ Predicted Values
	✓ Unstandardized
	→ Prediction Intervals
	✓ Mean ✓ Individual
	(et préciser le niveau de confiance voulu)

On obtient alors les données suivantes :

	ident	ventes	publicite	bonus	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
24	24	1583,75	583,85	289,29	1464,49448	1412,38336	1516,60560	1269,21226	1659,77669
25	25	1124,75	499,15	272,55	1223,94391	1185,41877	1262,46905	1031,84039	1416,04742
26	.	.	400,00	300,00	1029,68049	935,60325	1123,75774	819,27588	1240,08511
27									

FIG. 9.5 – Les prédictions

Ainsi, si la publicité et les bonus sont fixés à 40 000 \$ et 30 000 \$ pour **plusieurs** magasins, alors les ventes moyennes des ces magasins devraient se retrouver entre 935 603,25 \$ et 1 123 757,74 \$ et ce 19 fois sur 20. D'autre part, si la publicité et les bonus sont fixés à ces mêmes valeurs mais pour un **seul** magasin, alors les ventes réelles de ce magasin devraient se retrouver entre 819 275,88 \$ et 1 240 085,11 \$, et ce 19 fois sur 20.

9.3 Un autre exemple

Vous êtes nouvellement responsable des ressources humaines dans une grande entreprise. Le premier dossier dont vous avez à vous occuper est celui de la satisfaction des employés. Selon les ouï-dire de certains cadres, le niveau de satisfaction des employés est *extrêmement* décourageant.

Ainsi, afin de faire le point sur la situation, vous décidez d'effectuer une étude sur la satisfaction. Aussi, vous aimeriez vous donner les outils nécessaires pour intervenir adéquatement sur la satisfaction des employés. En fait, pour mieux intervenir, vous aimeriez quantifier l'influence de certains facteurs sur le niveau de satisfaction des employés.

Parmi l'ensemble de toutes les variables susceptibles d'influencer la satisfaction d'un employé, vous décidez de ne vérifier que les variables du tableau 9.6, qui sont mesurées de façon continue entre 0 et 12 : 0 représente une **insatisfaction absolue**, tandis que 12 représente un niveau de **satisfaction extrêmement élevé**. Somme toute, 6 représente un niveau de satisfaction moyen théorique.

Après avoir rédigé votre questionnaire, vous effectuez la passation à 100 employés choisis au hasard. Vous êtes maintenant prêt(e) à effectuer l'analyse et l'interprétation de vos données. Faites le point sur la situation. Commentez-la et énoncez aux cadres vos objectifs d'interventions.

Les variables	Nom de la variable
Niveau de satisfaction de l'employé face à son travail.	satisfac
Niveau de satisfaction de l'employé face à la distribution du travail.	distrtrv
Niveau de satisfaction de l'employé face à ses chances d'avancement.	chan_av
Niveau de satisfaction de l'employé face à l'estime témoignée par son supérieur pour un travail bien fait.	estime
Niveau de satisfaction de l'employé face à la paye qu'il reçoit pour le type de travail effectué.	paye

FIG. 9.6 – Les variables de l'exemple

La base de données se nomme **satisftrav.sav**. Les sorties SPSS pour ce problème sont à la page suivant. Faites-en l'analyse.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,842 ^a	,709	,697	,85252

a. Predictors: (Constant), paye, estime, chan_av, distrtrv

b. Dependent Variable: satisfac

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	168,079	4	42,020	57,815	,000 ^a
	Residual	69,046	95	,727		
	Total	237,124	99			

a. Predictors: (Constant), paye, estime, chan_av, distrtrv

b. Dependent Variable: satisfac

FIG. 9.7 – Le modèle dans son ensemble

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
1	(Constant)	1,990	,341	5,829	,000		
	distrtrv	,189	,047	,250	,000	,806	1,240
	chan_av	,278	,040	,423	,000	,830	1,204
	estime	,242	,034	,430	,000	,848	1,179
	paye	,070	,040	,098	,089	,949	1,053

a. Dependent Variable: satisfac

FIG. 9.8 – Les coefficients

9.4 Exercices du chapitre

Exercice 1 Une nouvelle bannière d'agents d'immeubles désire pénétrer de nouveaux quartiers. Il est connu que la valeur d'une maison dépend du quartier dans laquelle elle a été construite, de l'âge de la maison, du nombre de pieds carrés de la maison, du terrain, etc...

Pour chacun des quartiers, l'entreprise désire obtenir une évaluation du prix moyen des maisons. Pour un de ces quartier, on obtient des données pour un échantillon de maisons : le prix de vente, la surface en pieds carrés de la maison et l'âge de la maison. La base de données se nomme `maison.sav`. Faites une analyse complète en régression linéaire multiple pour étudier le modèle

$$Y_{\text{prix}} = \beta_0 + \beta_1 X_{\text{surface}} + \beta_2 X_{\text{age}} + \epsilon.$$

Exercice 2 Dans une municipalité du centre de la Mauricie, on a effectué une étude sur l'endettement des familles. L'endettement a été mesuré à l'aide du total des soldes qui comprend les soldes d'une société prêteuse, banque, caisse populaire, cartes de crédit, ..., à l'exception des prêts hypothécaires. On aimerait, à l'aide de la régression multiple, déterminer une équation de régression qui pourrait établir un lien entre le solde de l'endettement et le revenu mensuel, le nombre de personnes dans le ménage et le niveau de scolarité du chef de famille (en années). Faites les étapes nécessaires pour l'étude de ce modèle avec la base de données `endettement.sav`.

Chapitre 10

Modèles de régression linéaire

10.1 Les variables discrètes

Le modèle de régression linéaire multiple vu au chapitre 9 sera repris ici, et nous verrons comment y incorporer des variables discrètes.

Il en effet parfois très intéressant de considérer une variable discrète dans un modèle de régression linéaire multiple. Par exemple, on peut souhaiter expliquer le salaire des employés d'une entreprise en fonction de leur ancienneté (variable continue) et en fonction de leur sexe (variable discrète) si l'on soupçonne que le salaire est différent selon que l'on est un homme ou une femme.

10.1.1 Les variables discrètes dichotomiques

Prenons l'exemple suivant suggéré dans l'introduction : voyons si l'on peut élaborer un modèle de régression pour expliquer le salaire (hebdomadaire) des employés d'une entreprise (`salaire`) en fonction de leur ancienneté (`ancien`) et de leur sexe (`sexé`). La

base de données est disponible dans le dossier Chapitre 8, elle se nomme `employés.sav`. Les commandes SPSS à effectuer sont les mêmes que celles vues au chapitre 9, il suffit de mettre la variable `sexé` avec les variables indépendantes. Nous verrons également que les étapes pour l'analyse sont les mêmes que celles vues en régression linéaire multiple, il n'y a que l'interprétation des coefficients et de l'équation qui changent.

Ici on veut donc arriver à estimer une équation du type

$$Y_{\text{salaire}} = \beta_0 + \beta_1 X_{\text{ancien}} + \beta_2 X_{\text{sexé}} + \epsilon$$

qui est un modèle dit **additif**. En effet, on ne fait qu'ajouter la variable `sexé`, on ne considère pas sa possible interaction avec la variable `ancien`. Pour ce faire il faudrait plutôt considérer un modèle **multiplicatif**; nous y reviendrons plus tard.

Lorsque l'on veut introduire une variable discrète dans les variables explicatives d'une régression, celle-ci **doit** être codée de façon binaire (0 et 1).

Ici la variable discrète est la variable `sexé`, et on a 0 = féminin, 1 = masculin. Il faut tenir compte de cette codification pour bien interpréter les résultats, comme nous le verrons bientôt.

Comme toute régression linéaire multiple, il faut d'abord tester la normalité des résidus et regarder si la répartition de ceux-ci est uniforme.

On obtient les sorties suivantes :

	Tests of Normality					
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ZRE_1	,046	200	,200*	,977	200	,002

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. 10.1 – Normalité des résidus

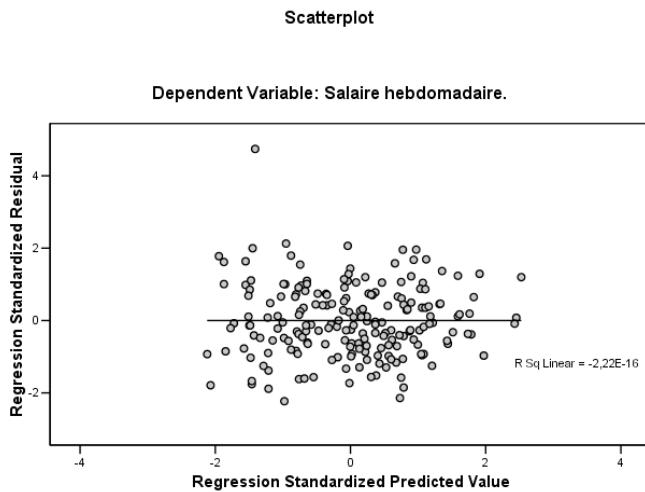


FIG. 10.2 – Répartition des résidus

Rapidement, on voit que l'un des deux tests rejette la normalité, mais surtout, plus important, l'on constate dans le graphe 10.2 qu'il y a un résidu qui se détache nettement des autres (plus de 4 écarts-type). Si on jette un coup d'œil dans la base de données, ce résidu correspond à un individu qui n'a que 10 mois d'ancienneté et qui gagne 927 \$ par semaine, le plus grand salaire de cette base. Il serait possible de continuer l'analyse avec cette donnée puisque seul l'un des deux tests rejette la normalité, et le restant des résidus est distribué de façon uniforme. Par contre, ici, la décision est de retirer cette donnée jugée aberrante et qui rend nécessairement le modèle moins juste. (Il n'est pas nécessaire de supprimer la ligne correspondant à cet individu : il suffit d'aller dans *Data → Select Cases...*, d'inscrire une condition dans *If condition is satisfied* (ici la condition *ZRE_1 < 3* donne le résultat escompté), puis de cocher l'option *Filtered*. Par la suite toutes les analyses se feront sans les données de l'individu en question. Il faut revenir à *Select Cases...* et sélectionner *All cases* pour revenir à la base complète.)

Poursuivons donc l'exemple avec la « nouvelle » base. On effectue les commandes à nouveau, et on a maintenant les sorties suivantes pour les résidus :

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ZRE_2	,038	199	,200*	,992	199	,354

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. 10.3 – Normalité des résidus

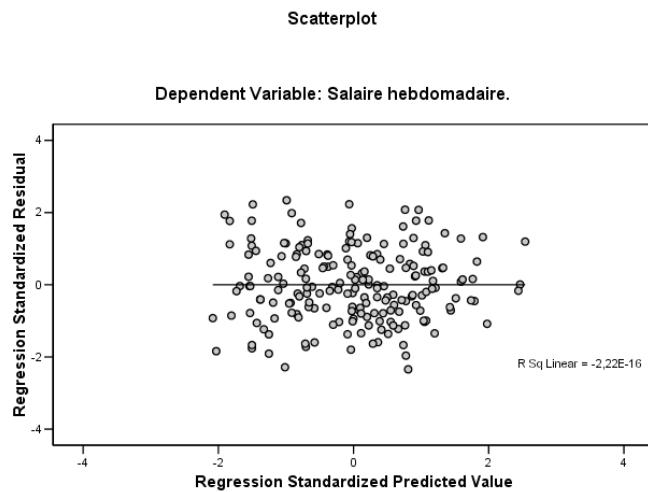


FIG. 10.4 – Répartition des résidus

Fixons les seuils de signification à $\alpha = 0,05$. Testons la normalité :

H_0 : Au niveau de la population, les résidus se distribuent selon une loi normale.

H_1 : Au niveau de la population, les résidus ne se distribuent pas selon une loi normale.

Puisque les deux p -values du test sont plus grandes que $\alpha = 0,05$ ($0,200$ et $0,354$), on conserve H_0 et ainsi on admet la normalité des résidus.

Le graphe 10.4 nous montre bien que les résidus sont répartis de façon uniforme, et il n'y a pas de valeur aberrante. On poursuit donc l'analyse.

La sortie 10.5 nous permet de constater que notre modèle explique à 39,7 % les variations de la variable salaire (c'est le $r^2_{\text{ajusté}}$). Ce n'est pas très fort, ce qui nous indique que le salaire n'est pas seulement fonction de l'ancienneté et du sexe.

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,635 ^a	,403	,397	88,05486

a. Predictors: (Constant), sexe, ancien
b. Dependent Variable: salaire

FIG. 10.5 – Le r^2 ajusté

La sortie 10.6, quant à elle, nous permet de résoudre le test d'hypothèses suivant :

H_0 : La régression est non significative dans la population (tous les $\beta_j = 0$).

H_1 : La régression est significative dans la population (au moins un des $\beta_j \neq 0$).

ANOVA ^b					
Model		Sum of Squares	df	Mean Square	F
1	Regression	1024465	2	512232,308	66,063
	Residual	1519717	196	7753,658	
	Total	2544182	198		

a. Predictors: (Constant), sexe, ancien
b. Dependent Variable: salaire

FIG. 10.6 – Table ANOVA de la régression

Puisque la p -value de la table ANOVA est de 0,000, ce qui est plus petit que $\alpha = 0,05$, on rejette H_0 . Ainsi, au risque de se tromper une fois sur 20, on peut admettre que la régression est significative.

Maintenant, la sortie 10.7 nous permet de voir lesquels des paramètres du modèle sont significatifs, et d'écrire l'équation. *Ici on voit que la valeur commune des deux VIF est de 1,014, ce qui nous indique qu'il n'y a pas de problème de multicolinéarité puisque c'est plus petit que 10.*

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	425,807	17,801		23,920	,000		
ancien	,900	,083	,605	10,886	,000	,986	1,014
sexe	39,873	16,700	,133	2,388	,018	,986	1,014

a. Dependent Variable: salaire

FIG. 10.7 – La table des coefficients

La *p*-value associée à la variable `ancien` étant de 0,000, on peut conclure que le paramètre β_1 est significatif. De même, la *p*-value associée à la variable `sexe` étant de 0,018, on conclut que le paramètre β_2 est significatif. Ceci nous indique qu'il y a une différence entre le salaire des hommes et celui des femmes, et qu'il était pertinent d'introduire cette variable discrète.

Il est maintenant temps d'écrire **les équations**. En effet, il est fondamental de comprendre que ce modèle correspond en réalité à deux droites de régression. À première vue, on peut penser qu'il n'y a qu'une équation, qui dans l'exemple serait

$$\hat{y}_{\text{salaire}} = 425,807 + 0,9x_{\text{ancien}} + 39,873x_{\text{sexe}}.$$

Mais il faut observer que lorsque $X_{\text{sexe}} = 0$, c'est-à-dire lorsqu'on considère les femmes, le modèle s'écrit alors

$$\hat{y}_{\text{salaire}} = 425,807 + 0,9x_{\text{ancien}} + 39,873 \times 0 = 425,807 + 0,9x_{\text{ancien}}$$

alors que si $X_{\text{sexe}} = 1$, c'est-à-dire lorsqu'on considère les hommes, le modèle s'écrit

$$\hat{y}_{\text{salaire}} = 425,807 + 0,9x_{\text{ancien}} + 39,873 \times 1 = 465,68 + 0,9x_{\text{ancien}}.$$

En fait, le paramètre β_2 correspondant à la variable binaire et qui mesure l'effet du sexe, correspond mathématiquement à la différence entre les ordonnées à l'origine (les constantes des deux équations) des deux droites de régression.

On peut donc conclure que β_2 mesure la différence de salaire selon le sexe, et ce peu importe l'ancienneté (en effet on voit que la pente (b_1) des deux droites est la même).

Cependant, il est facile de concevoir que parfois le modèle approprié pourrait être constitué de deux droites non parallèles. Par exemple, on pourrait se demander si l'écart de salaire entre les femmes et les hommes diminue ou augmente selon l'ancienneté. Si c'est le cas, il y a alors **interaction** entre les variables `sexe` et `ancien`, et à ce moment il faut considérer le modèle **multiplicatif** qui s'écrit

$$Y_{\text{salaire}} = \beta_0 + \beta_1 X_{\text{ancien}} + \beta_2 X_{\text{sexe}} + \beta_3 X_{\text{ancien}} X_{\text{sexe}} + \epsilon.$$

On note que s'il y a effet d'interaction ($\beta_3 \neq 0$), il n'est plus possible de parler de l'effet d'une variable sans nécessairement faire référence à la valeur de l'autre variable.

Dans le cas où l'on soupçonne l'effet d'interaction, on conseille généralement d'utiliser la procédure suivante :

- On construit la variable binaire et le produit de la variable explicative continue par cette variable binaire ;
- On estime le modèle interactif et on teste la signification du paramètre associé à l'interaction ;
- Si le paramètre de l'interaction est significatif, on interprète les résultats dans le modèle interactif en prenant garde au fait que l'effet de chaque variable ne dépende pas de la valeur de l'autre ;
- Si le paramètre de l'interaction n'est pas significatif, on peut conclure que les deux droites sont parallèles et que le modèle à utiliser est additif.

Dans le cadre de l'exemple, lorsqu'on considère le modèle multiplicatif, la table des coefficients est donnée par la figure 10.8.

La variable `ancien*sexe` est le produit des variables `ancien` et `sexe`. La *p*-value qui lui est associée étant de 0,922, on peut conclure qu'ici le paramètre associé à l'effet

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	427,783	26,943	15,877	,000
	Ancienneté dans l'entreprise, en mois.	,883	,196	,593	,000
	Sexe.	37,373	30,529	,124	,222
	anciensex	,021	,216	,016	,098
					,922

a. Dependent Variable: Salaire hebdomadaire.

FIG. 10.8 – La table des coefficients du modèle multiplicatif

d'interaction est nul, et donc que c'est le modèle additif qu'il convient de considérer dans cet exemple.

Exemple 10.1.1 Reprenons l'exemple 9.2.1 : Meddicorp est une entreprise qui vend du matériel médical aux hôpitaux et aux cliniques médicales. La direction de l'entreprise veut évaluer l'efficacité de son nouveau programme de bonus selon la performance à la vente. La direction veux savoir si les bonus versés ont une influence sur les ventes. Afin de ne pas être induit en erreur, la direction veut mettre en relief l'effet de la publicité et de la surface des magasins.

Compte tenu que le programme de bonus et de publicité sont indépendants de la surface des magasins, la direction suppose qu'il n'existe pas d'interaction entre les variables explicatives. Meddicorp veut donc étudier le modèle suivant :

$$Y_{\text{ventes}} = \beta_0 + \beta_1 X_{\text{publicite}} + \beta_2 X_{\text{bonus}} + \beta_3 X_{\text{surface}} + \epsilon.$$

Les données sont les suivantes (aussi disponibles dans la base de données `bonus.sav`) :

	ident	ventes	publicite	bonus	surface
1	2	893,00	408,50	236,28	petite
2	1	963,50	374,27	230,98	petite
3	13	1040,00	453,39	235,63	petite
4	14	1045,25	440,86	249,68	petite
5	3	1057,25	414,31	271,57	petite
6	8	1071,50	446,86	305,69	petite
7	9	1078,25	489,59	238,41	petite
8	15	1102,25	487,79	232,99	petite
9	10	1122,50	500,56	271,38	petite
10	25	1124,75	499,15	272,55	petite
11	20	1159,25	524,56	292,87	grande
12	4	1183,25	448,42	291,20	petite
13	21	1202,75	535,17	268,27	grande
14	16	1225,25	537,67	272,20	petite
15	22	1294,25	486,03	309,85	grande
16	11	1304,75	484,18	332,64	grande
17	5	1419,50	517,88	282,17	grande
18	23	1467,50	540,17	291,03	grande
19	17	1508,00	612,21	266,64	grande
20	6	1547,75	637,60	321,16	grande
21	12	1552,25	618,07	261,80	grande
22	18	1564,25	601,46	277,44	grande
23	7	1580,00	635,72	294,32	grande
24	24	1583,75	583,85	289,29	grande
25	19	1634,75	585,10	312,35	grande

FIG. 10.9 – Les données de l'exemple

On obtient les sorties suivantes (étudiez le modèle à partir de celles-ci) :

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,939 ^a	,882	,865	83,79467

a. Predictors: (Constant), surface, bonus, publicite

b. Dependent Variable: ventes

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1101521	3	367173,750	52,292	,000 ^a
	Residual	147452,5	21	7021,547		
	Total	1248974	24			

a. Predictors: (Constant), surface, bonus, publicite

b. Dependent Variable: ventes

FIG. 10.10 – Le r^2 ajusté et la table ANOVA**Coefficients^a**

Model	Unstandardized Coefficients		Beta	t	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
1	(Constant)	-106,922	256,237		,417	,681	
	publicite	1,963	,345	,637	5,691	,000	,448 2,231
	bonus	1,092	,747	,136	1,461	,159	,645 1,550
	surface	124,064	56,610	,277	2,192	,040	,351 2,848

a. Dependent Variable: ventes

FIG. 10.11 – La table des coefficients

10.1.2 Les variables discrètes polychotomiques

Cette section est simplement la généralisation de la précédente. Si la variable discrète possède c modalités (par exemple la variable des saisons possède 4 modalités), il faudra utiliser $c - 1$ variables binaires pour représenter la variable discrète.

S'il y a $k > 1$ variables continues dans le modèle, il faudra veiller à s'interroger sur l'interaction de chacune avec la variable discrète. C'est-à-dire qu'il y aura $k * (c - 1)$ paramètres d'interaction...

S'il y a plusieurs variables discrètes à inclure dans le modèle, chacune doit faire l'objet d'un codage selon la technique des variables binaires.

On peut remarquer que s'il y a plusieurs variables continues et plusieurs variables discrètes, le modèle s'alourdit assez rapidement. Dans ce cas, il est courant d'éliminer des termes multiplicatifs par simple présupposé d'absence d'interaction.

Exemple 10.1.2 Reprenons l'exemple 6.2.7 : la compagnie ABX vend des articles de sport. On comptabilise le total des ventes par trimestre (en millier de \$) du premier trimestre de 1985 jusqu'au dernier trimestre de 1994. La base de données se nomme **ABX.sav**.

Dans l'exemple 6.2.7, on a vu que le lien semble linéaire entre le temps et les ventes, mais aussi qu'il semblait il y avoir des fluctuations saisonnières.

On veut maintenant tenir compte de cet effet. Pour ce faire on doit codifier correctement la variable de saisons : on utilise $4 - 1 = 3$ variables binaires (voir la figure 10.12).

On s'intéresse alors au modèle suivant :

$$Y_{\text{ventes}} = \beta_0 + \beta_1 X_{\text{index}} + (\beta_2 X_{\text{hiver}} + \beta_3 X_{\text{printemps}} + \beta_4 X_{\text{été}}) + \epsilon.$$

	index	ventes	annee	saison	hiver	printemps	été	
1	1	221,00	1985	Hiver	1	0	0	
2	2	203,50	1985	Printemps	0	1	0	
3	3	190,00	1985	Été	0	0	1	
4	4	225,50	1985	Automne	0	0	0	
5	5	223,00	1986	Hiver	1	0	0	
6	6	190,00	1986	Printemps	0	1	0	
7	7	206,00	1986	Été	0	0	1	
8	8	226,50	1986	Automne	0	0	0	
9	9	236,00	1987	Hiver	1	0	0	
10	10	214,00	1987	Printemps	0	1	0	
11	11	210,50	1987	Été	0	0	1	
12	12	237,00	1987	Automne	0	0	0	
13	13	245,50	1988	Hiver	1	0	0	
14	14	201,00	1988	Printemps	0	1	0	
15	15	230,00	1988	Été	0	0	1	
16	16	254,50	1988	Automne	0	0	0	
17	17	257,00	1989	Hiver	1	0	0	
18	18	238,00	1989	Printemps	0	1	0	
19	19	228,00	1989	Été	0	0	1	

FIG. 10.12 – La codification de la variable polychotomique

Puisqu'on a 4 variables explicatives, l'échantillon devrait contenir au minimum 40 données, ce qui est le cas (il y en a exactement 40). Fixons les seuils à $\alpha = 0,05$.

On doit maintenant vérifier les hypothèses de validité relatives aux résidus. Fixons tous les seuils à $\alpha = 0,05$. On doit résoudre le test suivant :

H_0 : *Au niveau de la population, les résidus se distribuent selon une loi normale.*

H_1 : *Au niveau de la population, les résidus ne se distribuent pas selon une loi normale.*

Les p-values de Kolmogorov-Smirnov et Shapiro-Wilk étant respectivement de 0,200 et 0,485 (figure 10.13, on ne rejette pas H_0 au seuil $\alpha = 0,05$. Ainsi on admet que les résidus suivent une loi normale.

En jetant un coup d'œil au graphe de la figure 10.13, on voit que la répartition des

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ZRE_2	,107	40	,200*	,974	40	,485

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

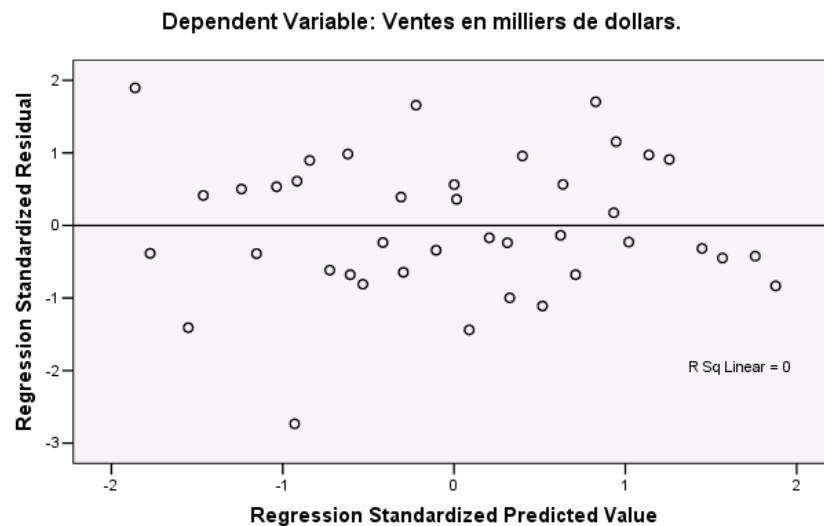


FIG. 10.13 – Les résidus

résidus est assez uniforme, il ne semble pas il y avoir de problème de ce côté. Et tous les points se situent dans les ± 3 écarts-types, nous pouvons donc poursuivre l'analyse sans problème.

Qualifions d'abord le modèle dans son ensemble. On voit que $r^2_{\text{ajusté}} = 0,955$ (première sortie de la figure 10.14). Ainsi 95,5 % de la variation de la variable ventes est expliquée par le modèle, ce qui est excellent, et mieux que le 78,3 % qui avait été obtenu avec le modèle ne comportant que la variable explicative index.

Pour voir si le modèle est significatif dans son ensemble, on résout le test suivant :

H_0 : La régression est non significative dans la population (tous les $\beta_j = 0$).

H_1 : La régression est significative dans la population (au moins un des $\beta_j \neq 0$).

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,979 ^a	,959	,955	7,19028

a. Predictors: (Constant), index, été, printemps, hiver

b. Dependent Variable: ventes

ANOVA^b

Model	Sum of Squares	df	Mean Square	F	Sig.
					,000 ^a
1	Regression	42630,421	4	10657,605	206,143
	Residual	1809,506	35	51,700	
	Total	44439,927	39		

a. Predictors: (Constant), index, été, printemps, hiver

b. Dependent Variable: ventes

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
1	(Constant)	210,846	3,148	66,980	,000		
	Index trimestriel incrémental du temps	2,566	,099	,889	25,932	,000	,991 1,009
	Variable binaire associée à l'hiver.	3,748	3,229	,049	1,161	,254	,661 1,513
	Variable binaire associée au printemps.	-26,118	3,222	-,339	-8,107	,000	,664 1,506
	Variable binaire associée à l'été.	-25,784	3,217	-,335	-8,015	,000	,666 1,501

a. Dependent Variable: Ventes en milliers de dollars.

FIG. 10.14 – Les sorties de la régression avec la variable polychotomique

Puisque la p -value de la table ANOVA (deuxième sortie de la figure 10.14) est de $0,000 < 0,05$, on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet que la régression est significative.

On peut maintenant examiner chacun des paramètres du modèle à l'aide de la table des coefficients de la figure 10.14. *Les VIF étant inférieurs à 10 (leurs valeurs étant de 1,009, 1,513, 1,506, 1,501), on n'a pas de problème de multicolinéarité, on peut donc tester si les paramètres sont significatifs ou non.*

La p -value associée à la variable `index` étant nulle, on peut conclure que l'apport d'information de la variable `index` dans ce modèle est significative. Les variables `printemps` et `été` sont elles aussi jugées significatives puisque leurs p -values sont nulles. Finalement, on voit que la p -value de la variable `hiver` est de 0,254, ce qui est plus grand que notre seuil. Mais attention : ici les variables `hiver`, `printemps`, `été` représentent la variable des saisons (le groupe de référence est l'automne), et donc dès que l'une de ces variables est significative, on peut en dire autant de la variable des saisons. Le fait que la p -value de la variable `hiver` soit plus grande que notre seuil signifie simplement que les ventes de cette saison ne se distinguent pas significativement des ventes de la saison de référence qui est l'automne.

Bref l'évolution dans le temps explique de façon significative une partie de la variation dans les ventes (par le biais de la variable `index`), et il y aussi une partie de la variation des ventes qui est expliquée par les saisons.

L'équation de la régression s'écrit

$$\hat{y}_{\text{ventes}} = 210,846 + 2,566x_{\text{index}} + 3,748x_{\text{hiver}} - 26,118x_{\text{printemps}} - 25,784x_{\text{été}}.$$

On peut tirer de cette équation les équations suivantes :

automne	$\hat{y}_{\text{ventes}} = 210,846 + 2,566x_{\text{index}}$
hiver	$\hat{y}_{\text{ventes}} = 214,594 + 2,566x_{\text{index}}$
printemps	$\hat{y}_{\text{ventes}} = 184,728 + 2,566x_{\text{index}}$
été	$\hat{y}_{\text{ventes}} = 185,062 + 2,566x_{\text{index}}$

Ici, le b_0 de la droite générale peut s'interpréter comme étant les ventes lorsque tout est à 0, c'est-à-dire que les ventes de l'automne 1984 ont été d'à peu près 210 846 \$ (si cette interprétation a du sens dans ce contexte).

Le coefficient $b_1 = 2,566$ signifie que le fait d'augmenter l'index d'une unité augmente en moyenne les ventes de 2 566 \$. De façon plus générale, on peut retenir de ce coefficient que les ventes augmentent avec le temps.

L'effet des saisons est représenté par les autres coefficients. Par exemple, par rapport à l'automne, les ventes augmentent en moyenne de 3 748 \$ l'hiver, tandis qu'au printemps (toujours par rapport à l'automne) elles diminuent en moyenne de 26 118 \$. On a une interprétation semblable pour l'été.

Voici comment faire ce même exemple avec E-Views. On reprend la feuille de travail `abx.wf1` utilisée à la section 6.3, et cette fois-ci pour estimer le modèle on ajoute les variables `@seas(1)`, `@seas(2)` et `@seas(3)` qui représentent les 3 premiers trimestres, et ainsi le quatrième trimestre sert de groupe de référence. On peut voir l'équation dans la figure 10.15.

Puis une fois estimé, on voit que le modèle correspond parfaitement à ce qu'on vient de voir avec SPSS (voir figure 10.16). L'interprétation se fait bien sûr de la même façon.

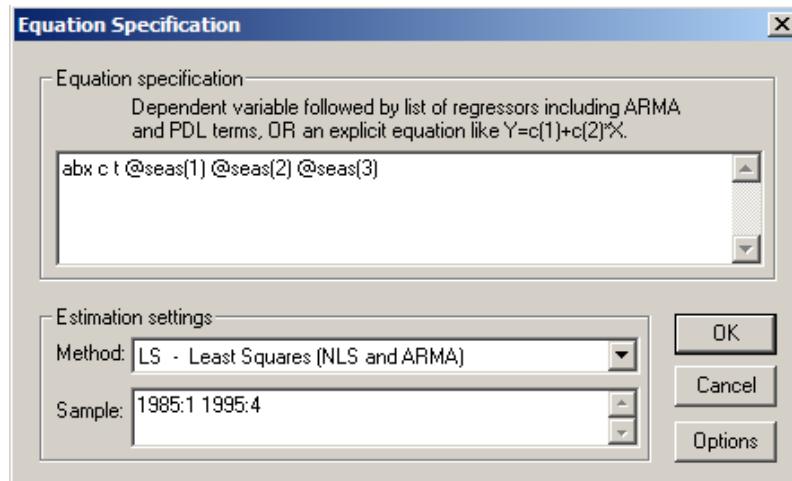


FIG. 10.15 – Pour ajouter l'effet des trimestres

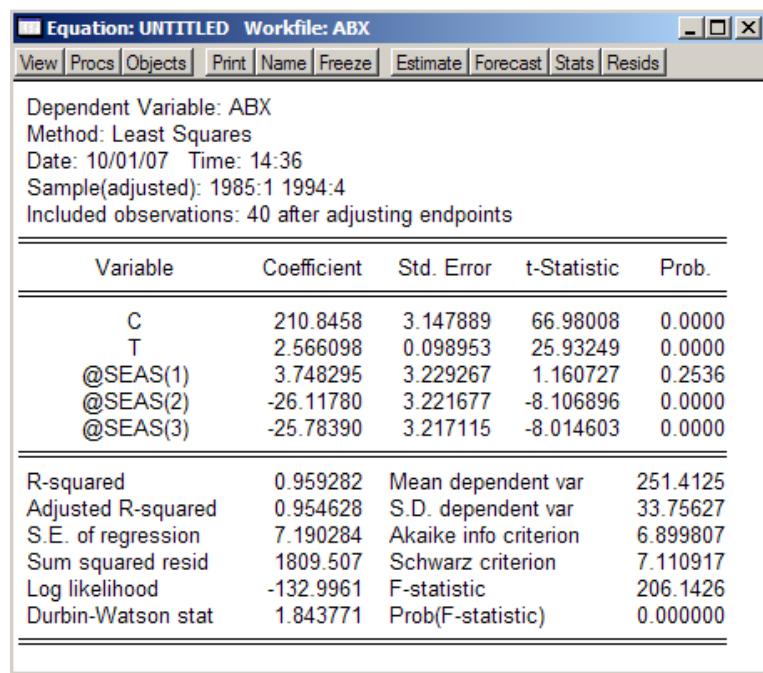


FIG. 10.16 – Les estimations

10.2 Les variables déphasées

Lorsqu'on utilise des données prises à intervalles réguliers, il est possible que les valeurs de la variable dépendante soient fonction des valeurs de la même variable explicative prise à des temps différents. Par exemple, le montant des ventes peut dépendre de la publicité du mois courant, du mois précédent ou d'il y a deux mois.

Cet effet peut facilement être inclus dans un modèle de régression. Par exemple, supposons que nous voulons expliquer les ventes mensuelles Y par la publicité X . Alors :

- Y_i représente le montant des ventes au mois i ;
- X_i représente le montant alloué à la publicité au mois i ;
- X_{i-1} représente le montant alloué à la publicité au mois $i - 1$;
- X_{i-2} représente le montant alloué à la publicité au mois $i - 2$.

Alors, pour modéliser les ventes du mois courant en fonction de la publicité du mois courant et des deux mois précédents, nous obtenons l'équation suivante :

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{i-1} + \beta_3 X_{i-2} + \epsilon_i.$$

Il faut créer les variables déphasées dans la base de données. Voici un exemple de ce qu'on pourrait obtenir :

	mois	ventes	publicite	publicite_1	publicite_2
1	1	963,50	374,27		
2	2	893,00	408,50	374,27	
3	3	1057,25	414,31	408,50	374,27
4	4	1183,25	448,42	414,31	408,50
5	5	1419,50	517,88	448,42	414,31
6	6	1547,75	637,60	517,88	448,42
7	7	1580,00	635,72	637,60	517,88
8	8	1071,50	446,86	635,72	637,60
9	9	1078,25	489,59	446,86	635,72
10	10	1122,50	500,56	489,59	446,86
11	11	1304,75	484,18	500,56	489,59
12	12	1552,25	618,07	484,18	500,56
13	13	1040,00	453,39	618,07	484,18
14	14	1045,25	440,86	453,39	618,07
...

FIG. 10.17 – La base de données avec deux variables déphasées

Les commandes à effectuer dans SPSS pour créer une variable déphasée sont les suivantes :

Menu SPSS :	→ Transform
	→ Create Time Series...
Dans la fenêtre New Variable(s) :	→ publicite (la variable qu'on veut déphaser)
Dans la fenêtre Name :	→ publicite_1 (le nom de la nouvelle variable)
Dans la fenêtre Function :	→ choisir Lag
Dans la fenêtre Order :	→ 1 (pour un déphasage d'un mois)
Cliquer sur Change, puis sur Ok.	

Ensuite, pour créer la variable déphasée de deux mois `publicite_2`, refaire ces commandes en changeant le nom de la variable et l'ordre du Lag (2 au lieu de 1).

Les variables `publicite_1` et `publicite_2` sont donc les variables déphasées, qu'on appelle aussi (en bon français) des *lagged variables*.

Cependant, utiliser trop de variables déphasées peut amener deux types de problème :

- L'utilisation d'un nombre élevé de variables déphasées amène un problème de colinéarité entre les variables ;
- L'utilisation des variables déphasées cause la perte d'une donnée pour chaque variable créée. Cependant, si vous avez beaucoup de données dans votre fichier, cet aspect pose peu de problème.

De façon semblable, il est possible d'utiliser les valeurs de la variable dépendante Y à titre de variable déphasée. En effet, pourquoi le montant de ventes du mois courant ne serait pas dépendant des ventes du mois précédent ? Le modèle à considérer est alors le suivant :

$$Y_i = \beta_0 + \beta_1 Y_{i-1} + \epsilon_i.$$

On peut aussi combiner les deux modèles précédents :

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{i-1} + \beta_3 X_{i-2} + \beta_4 Y_{i-1} + \epsilon_i.$$

Il est clair que la régression linéaire multiple est un outil de modélisation et de prévision extrêmement adaptable et puissant lorsque bien utilisé.

Exemple 10.2.1 Prenons la base de données `chomage.sav` dans laquelle on retrouve les taux de chômage chez les jeunes de 15 à 24 ans de 1976 à 2000. La figure 10.18 présente le graphe de ce taux de chômage en fonction du temps.

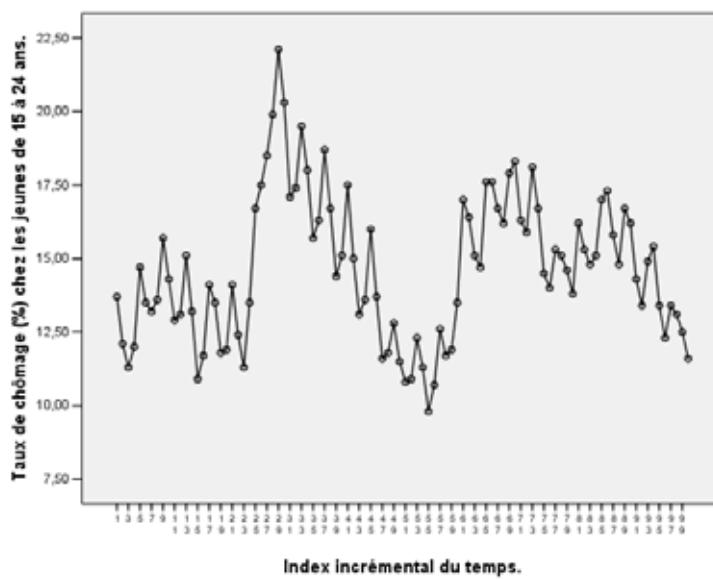


FIG. 10.18 – L'évolution du taux de chômage chez les jeunes de 15 à 24 ans

Est-il possible d'étudier ce taux de chômage à l'aide d'une régression ? On pourrait par exemple étudier le lien entre le taux de chômage du mois courant et celui du mois précédent, ce qui donnerait le modèle suivant :

$$Y_{\text{tauxchoj}_i} = \beta_0 + \beta_1 Y_{\text{tauxchoj}_{i-1}} + \epsilon_i.$$

Pour ce faire il faut d'abord créer la variable déphasée `tauxchoj_1`. La figure 10.19 donne un aperçu de ce qu'on obtient dans la base de données.

	index	annee tri	trimestre	tauxchoj	tauxchoj_1
1	1	1976/1	1	13,70	.
2	2	1976/2	2	12,10	13,70
3	3	1976/3	3	11,30	12,10
4	4	1976/4	4	12,00	11,30
5	5	1977/1	1	14,70	12,00
6	6	1977/2	2	13,50	14,70
7	7	1977/3	3	13,20	13,50
8	8	1977/4	4	13,60	13,20
9	9	1978/1	1	15,70	13,60
10	10	1978/2	2	14,30	15,70
11	11	1978/3	3	12,90	14,30
12	12	1978/4	4	13,10	12,90
13	13	1979/1	1	15,10	13,10

FIG. 10.19 – La base de données avec la variable tauxchoj_1

On peut maintenant faire l'analyse en régression linéaire simple. Le graphe de la relation $\text{tauxchoj_1} \Rightarrow \text{tauxchoj}$ est le suivant :

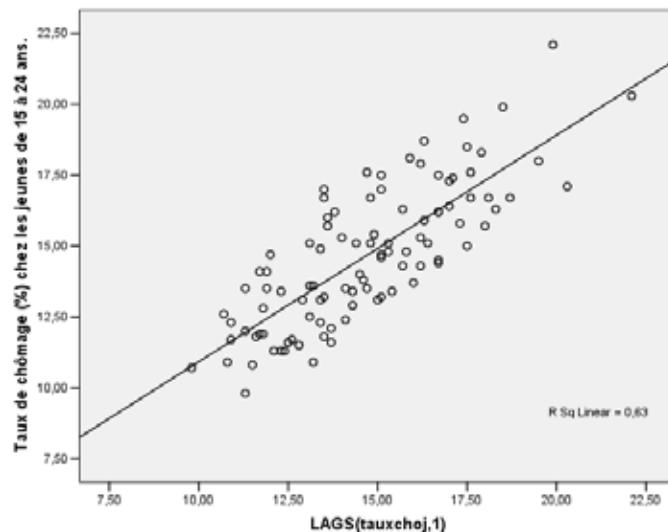


FIG. 10.20 – Le graphe de la relation

On voit que la relation semble linéaire, les points sont distribués de façon uniforme autour de la droite.

La figure 10.21 nous montre que le r de la relation est de 0,794, ce qui démontre une

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,794 ^a	,630	,626	1,50421

a. Predictors: (Constant), LAGS(tauxchoj,1)

FIG. 10.21 – Le r et le r^2

relation linéaire très forte entre le taux de chômage (chez les jeunes) du mois courant et celui du mois précédent. Le r^2 de 0,63 nous indique que 63 % de la variation du taux de chômage est expliquée par le taux de chômage du mois précédent.

ANOVA ^b					
Model		Sum of Squares	df	Mean Square	F
1	Regression	373,404	1	373,404	165,030
	Residual	219,476	97	2,263	
	Total	592,880	98		

a. Predictors: (Constant), LAGS(tauxchoj,1)
b. Dependent Variable: Taux de chômage (%) chez les jeunes de 15 à 24 ans.

FIG. 10.22 – La table ANOVA de la régression

La table ANOVA nous permet de confirmer que ce modèle est significatif. En effet, on peut résoudre le test suivant avec la p -value de la table :

H_0 : La régression est non significative dans la population ($\beta_1 = 0$).

H_1 : La régression est significative dans la population ($\beta_1 \neq 0$).

Fixons le seuil de signification à $\alpha = 0,05$. Puisque la p -value est égale à 0 ce qui est plus petit que $\alpha = 0,05$, on rejette H_0 . Ainsi au risque de se tromper une fois sur 20, on peut affirmer que la régression est significative dans la population.

La table des coefficients nous permet d'écrire la droite :

$$\hat{y}_{tauxchoj_i} = 2,924 + 0,799y_{tauxchoj_{i-1}}.$$

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	2,924	,926		3,159	,002	1,087	4,762
LAGS(tauxchoj,1)	,799	,062	,794	12,846	,000	,676	,923

a. Dependent Variable: Taux de chômage (%) chez les jeunes de 15 à 24 ans.

FIG. 10.23 – La table des coefficients

L'interprétation du b_0 est théoriquement la suivante : si le taux de chômage du mois précédent est de 0 %, alors celui du mois suivant devrait être de 2,924 %. Cette interprétation ne peut être que théorique car aucune donnée n'approche le 0 % dans cet échantillon.

Ensuite, le b_1 de 0,799 nous indique que d'après ce modèle, lorsque le taux du mois précédent augmente de 1 %, celui du mois courant devrait augmenter de 0,799 %.

D'après vous, quelle est la principale faiblesse de ce modèle ?

Exemple 10.2.2 La base de données `QuarterlyEarnings.sav` contient les bénéfices par action trimestriels d'une société sur une période de sept ans. On décide de modéliser cette série temporelle avec le modèle linéaire suivant :

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-4}$$

où y représente la variable `earnings`. Fixons le seuil à $\alpha = 0,05$.

Tout d'abord, la première sortie de la figure 10.24 nous montre qu'on a $r^2_{\text{ajusté}} = 0,657$. Ainsi 65,7 % de la variation des bénéfices par action est expliquée par le modèle.

Pour voir si le modèle est significatif dans son ensemble, on résout le test suivant :

H_0 : La régression est non significative dans la population (tous les $\beta_j = 0$).

H_1 : La régression est significative dans la population (au moins un des $\beta_j \neq 0$).

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,829 ^a	,687	,657	,13642

a. Predictors: (Constant), earnings_4, earnings_1

b. Dependent Variable: earnings

ANOVA^b

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	,857	2	,429	23,038	,000 ^a
Residual	,391	21	,019		
Total	1,248	23			

a. Predictors: (Constant), earnings_4, earnings_1

b. Dependent Variable: earnings

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
1 (Constant)	,129	,150		,857	,401		
earnings_1	-,046	,124	-,046	-,374	,712	,970	1,031
earnings_4	,874	,132	,820	6,611	,000	,970	1,031

a. Dependent Variable: earnings

FIG. 10.24 – Sorties de la régression

Puisque la p -value de la table ANOVA (deuxième sortie de la figure 10.24) est de $0,000 < 0,05$, on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet que la régression est significative.

On peut maintenant examiner chacun des paramètres du modèle à l'aide la table des coefficients de la figure 10.24. Les VIF étant inférieurs à 10 (leur valeur étant de 1,031), on n'a pas de problème de multicolinéarité, on peut donc tester si les paramètres sont significatifs ou non. La p -value associée à la variable earnings_1 étant égale à 0,712 ($> 0,05$), on peut conclure que cette variable n'est pas significative dans ce modèle. Par contre, la p -value associée à la variable earnings_4 étant égale à 0,000 ($< 0,05$), on peut conclure que cette variable est significative dans ce modèle.

L'équation de la régression s'écrit

$$\hat{y}_t = 0,129 - 0,046y_{t-1} + 0,874y_{t-4}.$$

L'interprétation du b_0 n'a pas vraiment de sens : en effet, pour que $y_t = 0,129$, il faudrait que y_{t-1} et y_{t-4} soient nuls, autrement dit que les bénéfices par action du trimestre précédent et du même trimestre l'année d'avant soient nuls.

Le b_1 s'interprète comme suit : pour chaque unité de plus pour les bénéfices par action du trimestre $t - 1$, on enlève en moyenne 0,046 aux bénéfices par action du trimestre t .

Le b_2 s'interprète comme suit : pour chaque unité de plus pour les bénéfices par action du trimestre $t - 4$, on ajoute en moyenne 0,874 aux bénéfices par action du trimestre t .

La figure 10.25 montre les prédictions pour la huitième année. Elles sont, pour les 4 trimestres dans l'ordre, de 0,69964, 0,62054, 1,08736 et 0,68122.

	earnings	index	earnings 4	earnings 1	PRE 4
26	,60	26,00	,44	,69	,48117
27	1,13	27,00	1,02	,60	,99218
28	,69	28,00	,63	1,13	,62681
29	-	29,00	,69	,69	,69964
30	-	30,00	,60	,70	,62054
31	-	31,00	1,13	,62	1,08736
32	-	32,00	,69	1,09	,68122

FIG. 10.25 – Prédictions

Finalement, le graphe de la figure 10.26 nous montre les bénéfices par action et les prédictions établies par le modèle. Ceci nous montre que le modèle arrive à modéliser les fluctuations trimestrielles des bénéfices par action (par exemple, c'est le troisième trimestre qui a toujours les plus grandes valeurs), mais pas toujours de très près. En fait les estimations de chaque année ressemblent beaucoup aux bénéfices de l'année d'avant, ce qui n'est pas étonnant puisque seule la variable y_{t-4} est significative dans ce modèle.

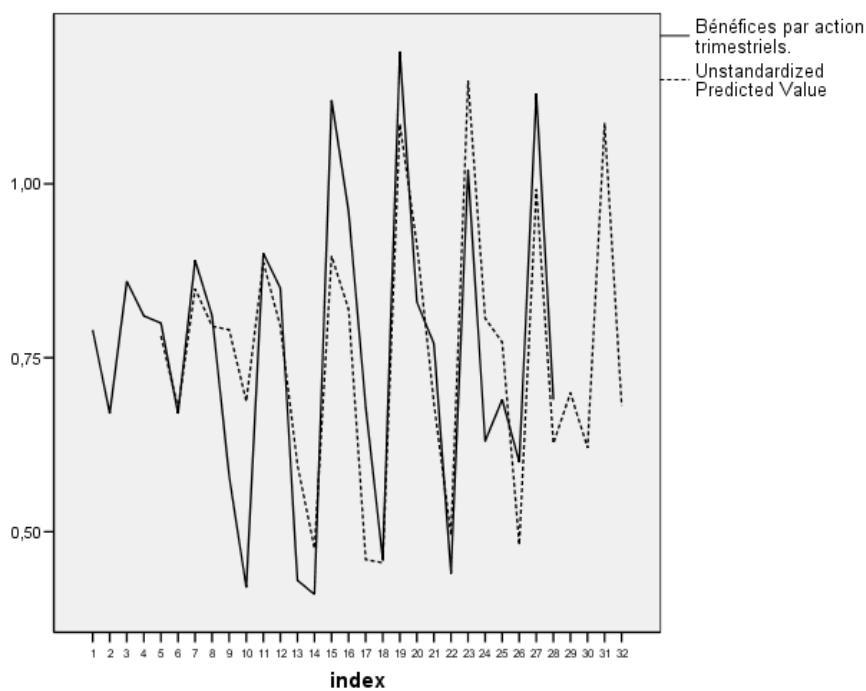


FIG. 10.26 – Graphe séquentiel

Voyons maintenant comment faire cet exemple avec E-Views. La feuille de travail se nomme `QuaterlyEarnings.wf1`, et il est décidé de définir l'intervalle de temps des sept années de 1995 à 2001 ; ce choix est arbitraire. La série des bénéfices par action est nommée `earnings` comme dans la base de données SPSS.

On veut donc développer un modèle pour estimer les bénéfices par action trimestriels en se servant de la valeur de ceux-ci au trimestre précédent et l'année précédente au même trimestre ; ces variables déphasées se notent respectivement `earnings(-1)` et `earnings(-4)`, comme on peut le voir dans l'équation du modèle (figure 10.27).

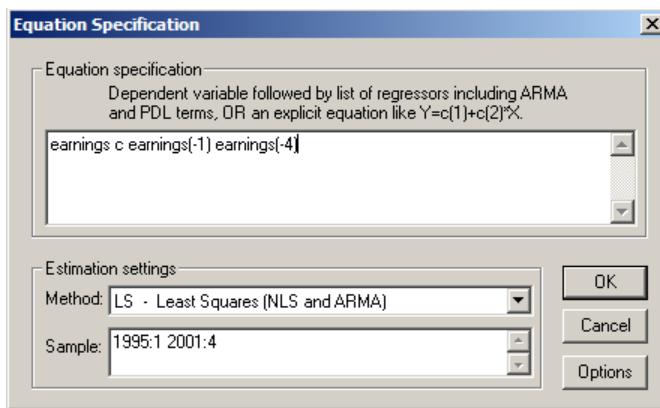


FIG. 10.27 – L'équation du modèle

Une fois le modèle estimé, on obtient la figure 10.28 qui nous montre que c'est bien le même modèle que celui obtenu dans SPSS.

Equation: UNTITLED Workfile: UNTITLED				
View Procs Objects Print Name Freeze Estimate Forecast Stats Resids				
Dependent Variable: EARNINGS				
Method: Least Squares				
Date: 10/01/07 Time: 14:53				
Sample(adjusted): 1996:1 2001:4				
Included observations: 24 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.128661	0.150091	0.857219	0.4010
EARNINGS(-1)	-0.046356	0.123981	-0.373899	0.7122
EARNINGS(-4)	0.873859	0.132183	6.610982	0.0000
R-squared	0.686926	Mean dependent var	0.748750	
Adjusted R-squared	0.657109	S.D. dependent var	0.232964	
S.E. of regression	0.136417	Akaike info criterion	-1.029739	
Sum squared resid	0.390799	Schwarz criterion	-0.882482	
Log likelihood	15.35686	F-statistic	23.03835	
Durbin-Watson stat	1.269486	Prob(F-statistic)	0.000005	

FIG. 10.28 – Les estimations

Il faut maintenant établir les prévisions pour la 8e année, c'est-à-dire pour 2002. Il faut aller dans le menu **Forecast** comme montré dans la section 6.3. Mais dans l'exemple présent nous avons le choix entre faire des prédictions avec la méthode statique ou dynamique (cette dernière est sélectionnée par défaut). La méthode statique prend seulement les véritables valeurs des bénéfices par action pour établir les prévisions, alors que la méthode dynamique utilise les estimations du modèle pour établir les prévisions. Ici nous utilisons la méthode statique pour tout d'abord établir les estimations de 1995 jusqu'au premier trimestre de 2002 puisque les véritables bénéfices par action sont connus jusqu'au dernier trimestre de 2001. On voit donc dans la figure 10.29 que la méthode statique est sélectionnée, et que le **Forecast sample** s'arrête au premier trimestre de 2001.

On obtient alors la deuxième sortie de la figure 10.29. On vient aussi de créer la série **earningsf1**. La figure 10.30 nous montre la visualisation des séries **earnings** et **earningsf1**. Maintenant, pour générer les prévisions pour les trois derniers trimestres de 2002, on n'a pas le choix d'y aller avec la méthode dynamique car notre dernier bénéfice par action observé est celui du dernier trimestre de 2001. Cependant, étant donné que nous devons utiliser au moins une première observation réelle, nous indiquons comme échantillon de prévision l'intervalle du premier trimestre de 2002 au dernier trimestre de 2002. La prévision pour le premier trimestre de 2002 devrait coïncider avec la prévision obtenue avec la méthode statique. Dans la figure 10.31 on voit le menu **Forecast** avec ces options, et on voit que la nouvelle série se nomme **earningsf2**.

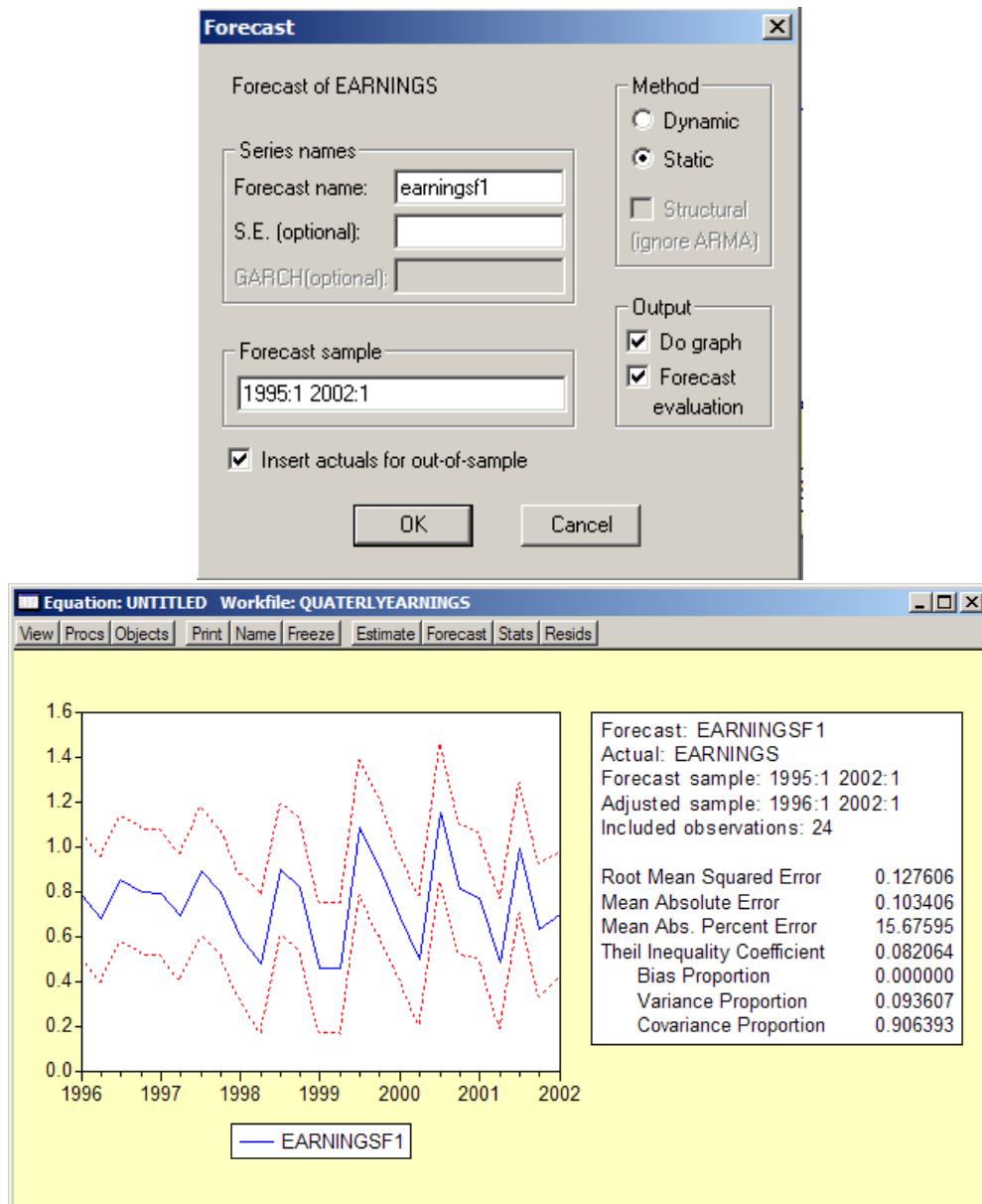


FIG. 10.29 – Pour les prévisions statiques

obs	EARNINGS	EARNINGSF1
1998:3	1.120000	0.896128
1998:4	0.960000	0.819522
1999:1	0.680000	0.459918
1999:2	0.460000	0.455421
1999:3	1.190000	1.086059
1999:4	0.830000	0.912402
2000:1	0.770000	0.684409
2000:2	0.440000	0.494942
2000:3	1.020000	1.148157
2000:4	0.630000	0.806680
2001:1	0.690000	0.772328
2001:2	0.600000	0.481173
2001:3	1.130000	0.992183
2001:4	0.690000	0.626809
2002:1	NA	0.699638
2002:2	NA	NA
2002:3	NA	NA
2002:4	NA	NA

FIG. 10.30 – Les séries earnings et earningsf1

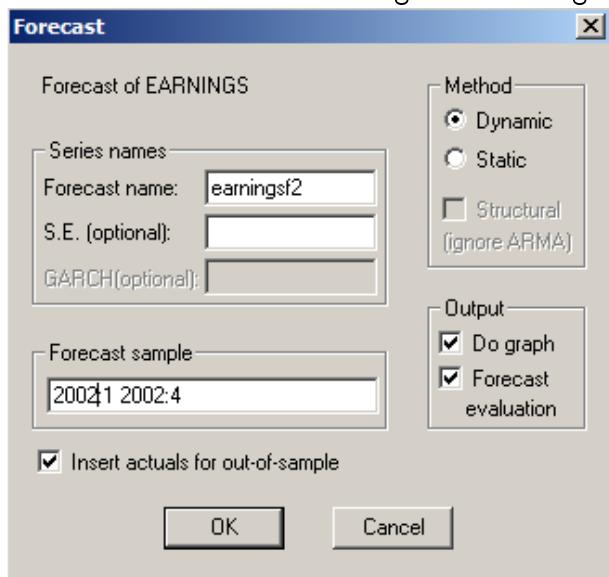


FIG. 10.31 – Pour les prévisions dynamiques

obs	EARNINGS	EARNINGSF1	EARNINGSF2
1998:3	1.120000	0.896128	1.120000
1998:4	0.960000	0.819522	0.960000
1999:1	0.680000	0.459918	0.680000
1999:2	0.460000	0.455421	0.460000
1999:3	1.190000	1.086059	1.190000
1999:4	0.830000	0.912402	0.830000
2000:1	0.770000	0.684409	0.770000
2000:2	0.440000	0.494942	0.440000
2000:3	1.020000	1.148157	1.020000
2000:4	0.630000	0.806680	0.630000
2001:1	0.690000	0.772328	0.690000
2001:2	0.600000	0.481173	0.600000
2001:3	1.130000	0.992183	1.130000
2001:4	0.690000	0.626809	0.690000
2002:1	NA	0.699638	0.699638
2002:2	NA	NA	0.620544
2002:3	NA	NA	1.087356
2002:4	NA	NA	0.681218

FIG. 10.32 – Visualisation de la série originale et des deux séries de prévision

La figure 10.32 nous présente les trois séries, c'est-à-dire les bénéfices par action observés jusqu'en 2001, les prévisions statiques puis les prévisions dynamiques pour 2002. On voit que les prévisions statique et dynamique pour le premier trimestre de 2002 coïncident, et que la série `earningsf2` des prévisions dynamiques présente les valeurs observées jusqu'en 2001 puisque nous n'avons demandé des prévisions que pour 2002.

Finalement, si on visualise ces trois séries (`View → Graph → Line`), on obtient le graphe de la figure 10.33. On voit que `earnings` et `earningsf2` se confondent jusqu'à la fin de 2001.

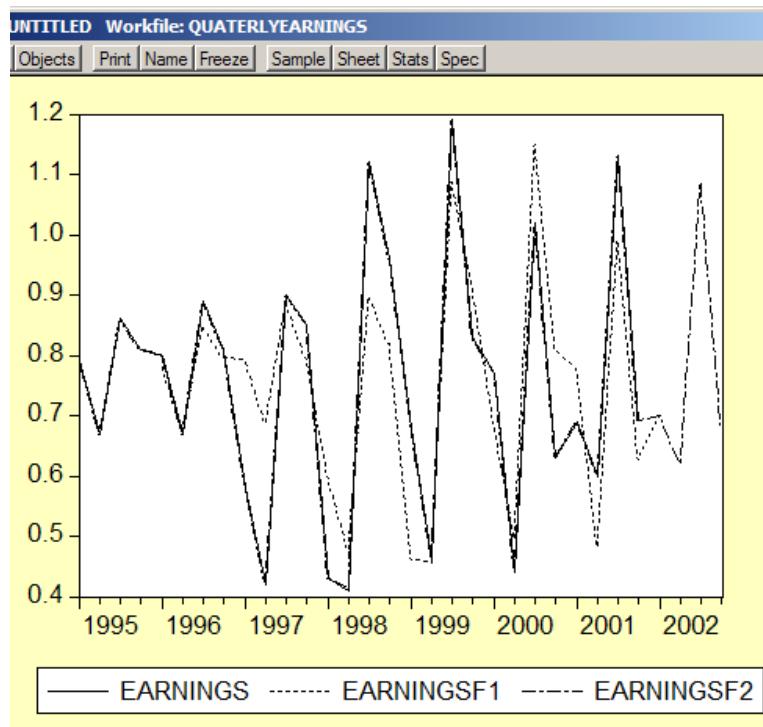


FIG. 10.33 – Graphe des trois séries

Pour visualiser les résidus standardisés, il suffit d'aller dans `View → Actual,Fitted, Residual → Standardized Residual Graph`. Pour ne pas avoir simplement la ligne reliant les résidus mais aussi un point pour chaque résidu, il faut ensuite aller dans `Objects → View Options → Options..., puis dans l'onglet Line & Symbols` il faut sélectionner `Line & Symbols` dans la fenêtre `Line attributes` (figure 10.34). On obtient alors le graphe tel que dans la figure 10.35.

Ici tous les résidus sont entre ± 3 écarts type, il ne semble donc pas il y avoir de outliers. Cependant la répartition des résidus ne semble pas aléatoire (plusieurs résidus positifs suivis de plusieurs négatifs). On verra comment traiter ceci au chapitre 11.

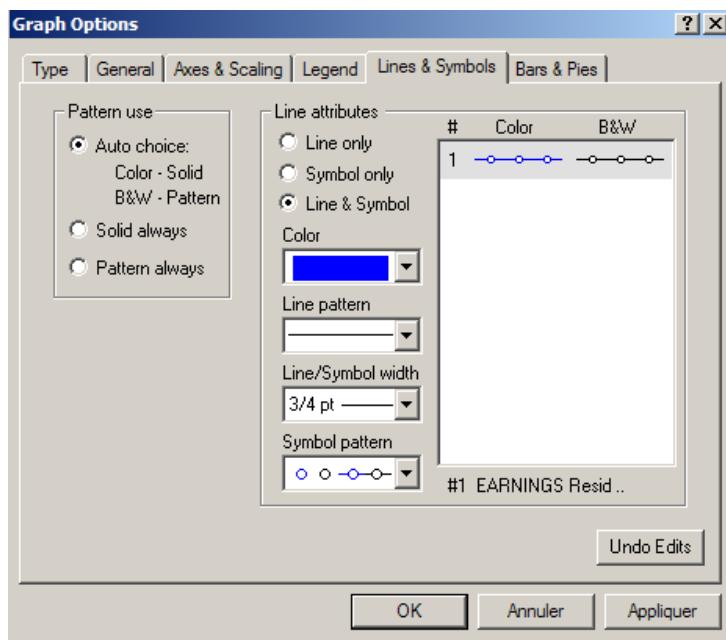


FIG. 10.34 – Pour faire apparaître les points

Pour tester la normalité des résidus, il faut aller dans View → Residual Tests → Histogram - Normality Test. E-Views performe le test de Jarque-Bera pour la normalité. Il permet de résoudre le test d'hypothèses suivant :

H_0 : Les résidus se distribuent selon une loi normale au niveau de la population.

H_1 : Les résidus ne se distribuent pas selon une loi normale au niveau de la population.

Puisque la p-value = 0,752236 > $\alpha = 0,05$, on ne rejette pas H_0 . Ainsi on ne rejette pas la normalité des résidus.

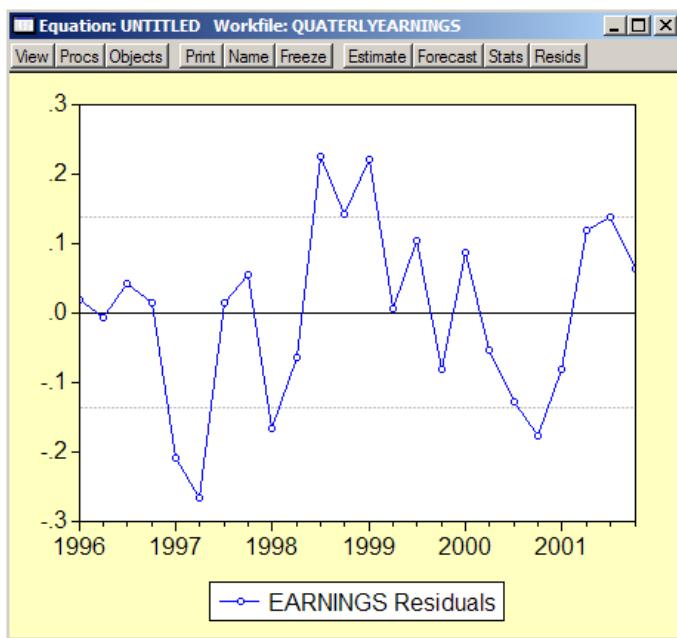


FIG. 10.35 – Graphe des résidus standardisés

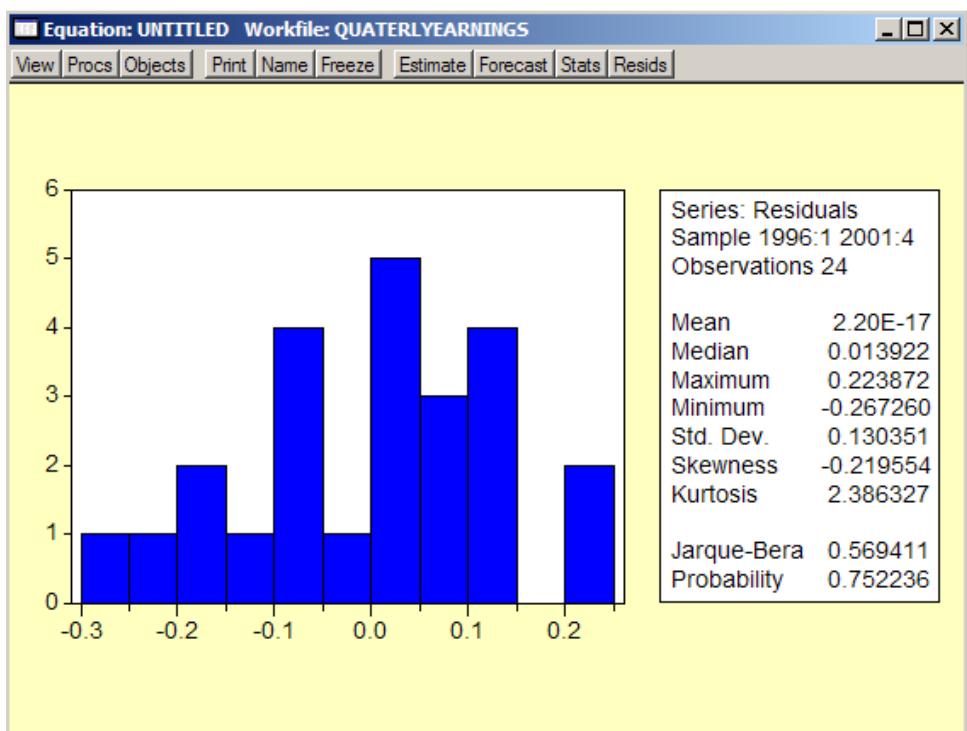


FIG. 10.36 – Histogramme et test de normalité des résidus

10.3 La sélection de variables explicatives

Lorsque l'analyste est en présence d'un nombre restreint de variables, le choix des variables à inclure ou à exclure du modèle est relativement simple. La technique de l'essai et de l'erreur est assez simple. Il suffit de créer un ensemble de modèles, de tester la signification des paramètres β_i et d'utiliser le coefficient r_{aj}^2 afin de trouver le modèle qui s'ajuste le mieux à notre intuition, et le tour est joué.

Cependant, lorsque le nombre de variables explicatives devient important, cette technique atteint rapidement ses limites. Il est alors plus probable de commettre les erreurs suivantes :

- Ne pas inclure une variable explicative dominante dans le modèle ;
- Inclure une variable inutile dans le modèle.

Des méthodes automatiques existent pour aider l'analyste dans son choix. Ces méthodes permettent de faire une pré-sélection qui permettra d'orienter l'analyste. Le choix du meilleur modèle est souvent plus une question d'interprétation que de performance absolue, ce que la machine est incapable de faire.

Quatre méthodes automatisées existent, mais elles ne sont pas toutes offertes par les logiciels :

All regressions. Cette méthode calcule toutes les régressions possibles, commençant par les régressions linéaires à une variable, puis à deux variables, et ainsi de suite. Les logiciels présentent un résumé de toutes les régressions.

Backward. Cette méthode est très intéressante. À la première étape, elle inclut toutes les variables disponibles dans le modèle. Elle compare les coefficients t^2 (cotes- t au carré, elles suivent approximativement une loi de Fisher), enlève la variable la moins significative et recommence la régression avec une variable en moins. Elle recommence de cette façon jusqu'à ce que toutes les variables restantes aient un t^2 significatif (selon le seuil voulu). Le critère s'appelle *F-to-remove*. Une variable qui a été enlevée ne revient plus.

Forward. Cette méthode procède à l'inverse de la méthode *backward*. En effet, cette

méthode examine toutes les régressions linéaires simples et choisit la variable la plus significative (celle qui a le coefficient t^2 le plus grand). Cette première variable étant choisie, la procédure recalcule toutes les régressions possibles lorsque nous ajoutons une deuxième variable à la première. La méthode choisira la variable qui aura le t^2 le plus grand. La méthode se termine lorsqu'aucune variable ne possède un t^2 assez grand. Le critère s'appelle *F-to-enter*. Une variable entrée ne quitte plus.

Stepwise. La méthode stepwise commence comme la *forward*, mais elle applique un critère *backward* à chaque tour. Entre autres, une variable qui est entrée est susceptible de sortir. Une variable qui est entrée et sortie ne reviendra plus. Ici, il y a à la fois un critère d'entrée *F-to-enter* et un critère de sortie *F-to-remove*.

Exemple 10.3.1 Pour illustrer l'utilisation des méthodes *Backward*, *Forward* et *Stepwise*, reprenons la base de données *chomage*. On cherche encore à expliquer le taux de chômage chez les jeunes (`tauxchoj`), et cette fois-ci on aimerait tenir compte des variables déphasées d'ordre 1 à 4 pour les variables `tauxcho` et `tauxchoj`. Ceci fait 8 variables explicatives (voir la figure 10.37). Utilisons les méthodes *Backward*, *Forward* et *Stepwise* pour voir quelles sont les variables parmi ces huit que nous devrions utiliser.

Voyons d'abord ce qui ce passe avec la méthode ***Backward***. Tout d'abord, les commandes à effectuer sont les suivantes :

Menu SPSS :	→ Analyse
	→ Regression
	→ Linear...
Dans la fenêtre Dependant :	→ <code>tauxchoj</code> (la variable dépendante)
Dans la fenêtre Independant(s) :	→ mettre les 8 variables déphasées
Dans la fenêtre Method :	→ sélectionner Backward

trimestre	tauxchoj	tauxcho	tauxchoj_1	tauxchoj_2	tauxchoj_3	tauxchoj_4	tauxcho_1	tauxcho_2	tauxcho_3	tauxcho_4
1	13,70	7,90	7,90	.	.
2	12,10	6,90	13,70
3	11,30	6,60	12,10	13,70	.	.	6,90	7,90	.	.
4	12,00	6,80	11,30	12,10	13,70	.	6,60	6,90	7,90	.
1	14,70	8,70	12,00	11,30	12,10	13,70	6,80	6,60	6,90	7,90
2	13,50	7,90	14,70	12,00	11,30	12,10	8,70	6,80	6,60	6,90
3	13,20	7,60	13,50	14,70	12,00	11,30	7,90	8,70	6,80	6,60
4	13,60	7,80	13,20	13,50	14,70	12,00	7,60	7,90	8,70	6,80
1	15,70	9,40	13,60	13,20	13,50	14,70	7,80	7,60	7,90	8,70
2	14,30	8,60	15,70	13,60	13,20	13,50	9,40	7,80	7,60	7,90
3	12,90	7,80	14,30	15,70	13,60	13,20	8,60	9,40	7,80	7,60
4	13,10	7,60	12,90	14,30	15,70	13,60	7,80	8,60	9,40	7,80
1	15,10	9,00	13,10	12,90	14,30	15,70	7,60	7,80	8,60	9,40
2	13,20	7,70	15,10	13,10	12,90	14,30	9,00	7,60	7,80	8,60
3	10,90	6,50	13,20	15,10	13,10	12,90	7,70	9,00	7,60	7,80
4	11,70	6,80	10,90	13,20	15,10	13,10	6,50	7,70	9,00	7,60
1	14,10	8,50	11,70	10,90	13,20	15,10	6,80	6,50	7,70	9,00
2	13,50	7,80	14,10	11,70	10,90	13,20	8,50	6,80	6,50	7,70

FIG. 10.37 – Un aperçu des 8 variables déphasées.

La méthode *Backward* commence par inclure toutes les variables dans le modèle, puis enlève celles qui ne sont pas significative de façon itérative. Dans la figure 10.38, on voit que seulement deux itérations ont eu lieu, et que seule la variable `tauxcho_1` a été enlevée du modèle. Il semble donc que les sept autres variables soient significatives.

La figure 10.39 nous montre que lorsqu'on enlève la variable `tauxcho_1` du modèle, le r^2_{aj} passe de 0,81 à 0,811. On obtient donc um modèle moins complexe et plus performant.

Variables Entered/Removed ^b			
Model	Variables Entered	Variables Removed	Method
1	tauxcho_4, tauxchoj_2, tauxcho_1, tauxcho_3, tauxchoj_4, tauxchoj_1, tauxchoj_3, tauxcho_2 ^a	.	Enter
2	.	tauxcho_1	Backward (criterion: Probability of F-to-remove >= .100).

a. All requested variables entered.

b. Dependent Variable: tauxchoj

FIG. 10.38 – Les itérations (Backward)

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,909 ^a	,826	,810	1,06622
2	,908 ^b	,825	,811	1,06338

a. Predictors: (Constant), tauxcho_4, tauxchoj_2, tauxcho_1, tauxcho_3, tauxchoj_4, tauxchoj_1, tauxchoj_3, tauxcho_2

b. Predictors: (Constant), tauxcho_4, tauxchoj_2, tauxcho_3, tauxchoj_4, tauxchoj_1, tauxchoj_3, tauxcho_2

FIG. 10.39 – Le r et r_{aj}^2 des modèles (Backward)

La figure 10.40 présente simplement la table ANOVA des deux modèles.

ANOVA ^a					
Model		Sum of Squares	df	Mean Square	F
1	Regression	468,317	8	58,540	51,494
	Residual	98,903	87	1,137	
	Total	567,220	95		
2	Regression	467,712	7	66,816	59,089
	Residual	99,508	88	1,131	
	Total	567,220	95		

a. Predictors: (Constant), tauxcho_4, tauxchoj_2, tauxcho_1, tauxcho_3, tauxchoj_4, tauxchoj_1, tauxchoj_3, tauxcho_2

b. Predictors: (Constant), tauxcho_4, tauxchoj_2, tauxcho_3, tauxchoj_4, tauxchoj_1, tauxchoj_3, tauxcho_2

c. Dependent Variable: tauxchoj

FIG. 10.40 – Les tables ANOVA des modèles (Backward)

La figure 10.41 nous présente les tables des coefficients des deux modèles. Ceci nous permet de constater qu'en effet, la variable `tauxcho_1` n'était pas significative dans le premier modèle.

Coefficients ^a					
Model	Unstandardized Coefficients			t	Sig.
	B	Std. Error	Beta		
1	(Constant)	1,511	,773		,054
	tauxchoj_1	1,418	,276	5,132	,000
	tauxchoj_2	-1,681	,305	-5,511	,000
	tauxchoj_3	2,060	,287	7,177	,000
	tauxchoj_4	-,920	,268	-3,432	,001
	tauxcho_1	-,291	,399	-,729	,468
	tauxcho_2	1,208	,432	2,794	,006
	tauxcho_3	-2,024	,419	-4,828	,000
	tauxcho_4	1,140	,401	2,839	,006
2	(Constant)	1,544	,770		,048
	tauxchoj_1	1,230	,100	12,288	,000
	tauxchoj_2	-1,560	,255	-6,124	,000
	tauxchoj_3	2,079	,285	7,293	,000
	tauxchoj_4	-,884	,263	-3,364	,001
	tauxcho_2	1,056	,378	2,796	,006
	tauxcho_3	-2,127	,393	-5,406	,000
	tauxcho_4	1,118	,399	2,800	,006

a. Dependent Variable: tauxchoj

FIG. 10.41 – Les tables des coefficients des modèles (Backward)

Excluded Variables ^b					
Model	Beta In	t	Sig.	Partial	Collinearity Statistics
				Correlation	Tolerance
2 tauxcho_1	,200 ^a	,729	,468	,078	,027

a. Predictors in the Model: (Constant), tauxchoj_4, tauxchoj_2, tauxchoj_3, tauxchoj_4, tauxchoj_1, tauxchoj_3, tauxchoj_2
b. Dependent Variable: tauxchoj

FIG. 10.42 – Les statistiques de la variable exclue (Backward)

Finalement, la dernière figure (10.42) nous montre les statistiques relatives à la variable exclue.

Regardons maintenant ce qui se passe si nous appliquons la méthode **Forward**. Celle-ci fonctionne à l'inverse de la méthode *Backward*; elle commence avec un modèle à une seule variable, celle qui est la plus significative, puis ajoute des variables une à une de façon itérative.

La figure 10.43 nous montre que la première variable à avoir été incluse est **tauxchoj_1**, suivie de **tauxchoj_4**, ensuite de **tauxchoj_2** et finalement de **tauxchoj_3**. Il semble donc que suite à ce choix de variables, aucune des quatre restantes ne pouvait être incluse dans le modèle. Ceci signifie que tous les modèles à 5 variables construits à partir de celui-ci ont leur 5e variable non-significative.

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	taux choj_1		Forward (Criterion: Probability-of- F-to-enter ≤ .050)
2	taux choj_4		Forward (Criterion: Probability-of- F-to-enter ≤ .050)
3	taux choj_2		Forward (Criterion: Probability-of- F-to-enter ≤ .050)
4	taux choj_3		Forward (Criterion: Probability-of- F-to-enter ≤ .050)

a. Dependent Variable: tauxchoj

FIG. 10.43 – Les itérations (Forward)

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,789 ^a	,622	,618	1,51060
2	,825 ^b	,681	,674	1,39465
3	,837 ^c	,701	,691	1,35773
4	,874 ^d	,764	,754	1,21201

- a. Predictors: (Constant), tauxchoj_1
- b. Predictors: (Constant), tauxchoj_1, tauxchoj_4
- c. Predictors: (Constant), tauxchoj_1, tauxchoj_4, tauxchoj_2
- d. Predictors: (Constant), tauxchoj_1, tauxchoj_4, tauxchoj_2, tauxchoj_3

FIG. 10.44 – Le r et r_{aj}^2 pour les modèles (Forward)

La figure 10.44 nous montre les r_{aj}^2 de chacun des modèles. Ainsi on voit que c'est le 4e modèle (avec 4 variables explicatives) qui est le plus performant avec un $r_{aj}^2 = 0,754$.

La figure 10.45 nous présente les tables ANOVA des quatre modèles.

ANOVA ^e					
Model		Sum of Squares	df	Mean Square	F
1	Regression	352,721	1	352,721	154,573
	Residual	214,499	94	2,282	
	Total	567,220	95		
2	Regression	386,331	2	193,166	99,312
	Residual	180,888	93	1,945	
	Total	567,220	95		
3	Regression	397,623	3	132,541	71,899
	Residual	169,596	92	1,843	
	Total	567,220	95		
4	Regression	433,544	4	108,386	73,784
	Residual	133,676	91	1,469	
	Total	567,220	95		

- a. Predictors: (Constant), tauxchoj_1
- b. Predictors: (Constant), tauxchoj_1, tauxchoj_4
- c. Predictors: (Constant), tauxchoj_1, tauxchoj_4, tauxchoj_2
- d. Predictors: (Constant), tauxchoj_1, tauxchoj_4, tauxchoj_2, tauxchoj_3
- e. Dependent Variable: tauxchoj

FIG. 10.45 – Les tables ANOVA des modèles (Forward)

La figure 10.46 nous montre la table des coefficients des quatre modèles. Une petite surprise nous attend dans le 4e modèle... la variable `tauxchoj_4` n'est plus significative suite à l'introduction de la variable `tauxchoj_3` dans le modèle.

Model	Coefficients ^a				
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	3,091	,950	3,253	,002
	tauxchoj_1	,790	,064		,000
2	(Constant)	1,364	,971	1,406	,163
	tauxchoj_1	,603	,074		,000
	tauxchoj_4	,305	,073		,000
3	(Constant)	1,894	,969	1,955	,054
	tauxchoj_1	,769	,098		,000
	tauxchoj_4	,335	,072		,000
	tauxchoj_2	-,232	,094		,015
4	(Constant)	1,563	,867	1,801	,075
	tauxchoj_1	1,054	,105		,000
	tauxchoj_4	-,069	,104		,507
	tauxchoj_2	-,758	,135		,000
	tauxchoj_3	,668	,135		,000

a. Dependent Variable: `tauxchoj`

FIG. 10.46 – Les tables des coefficients pour les modèles (Forward)

La figure 10.47 nous présente les statistiques relatives aux variables exclues. Ainsi on voit qu'effectivement, aucune des variables exclues du dernier modèle ne sont significatives. Que pensez-vous de ce résultat par rapport à celui obtenu avec la méthode *Backward* ?

Excluded Variables ^e						
Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics	
						Tolerance
1	tauxchoj_2	-,160 ^a	-1,555	,123	-,159	,376
	tauxchoj_3	,241 ^a	3,295	,001	,323	,680
	tauxchoj_4	,307 ^a	4,157	,000	,396	,629
	tauxcho_1	,142 ^a	,880	,381	,091	,156
	tauxcho_2	-,063 ^a	-,665	,508	-,069	,444
	tauxcho_3	,177 ^a	2,378	,019	,239	,689
	tauxcho_4	,258 ^a	3,584	,001	,348	,687
2	tauxchoj_2	-,233 ^b	-2,475	,015	-,250	,366
	tauxchoj_3	,075 ^b	,769	,444	,080	,365
	tauxcho_1	,060 ^b	,400	,690	,042	,153
	tauxcho_2	-,156 ^b	-1,749	,084	-,179	,420
	tauxcho_3	-,003 ^b	-,032	,975	-,003	,432
	tauxcho_4	-,011 ^b	-,072	,943	-,007	,145
3	tauxchoj_3	,673 ^c	4,945	,000	,460	,140
	tauxcho_1	,073 ^c	,497	,620	,052	,153
	tauxcho_2	,054 ^c	,364	,716	,038	,148
	tauxcho_3	,230 ^c	2,065	,042	,212	,254
	tauxcho_4	,053 ^c	,350	,727	,037	,141
4	tauxcho_1	-,028 ^d	-,210	,834	-,022	,149
	tauxcho_2	,115 ^d	,865	,389	,091	,146
	tauxcho_3	-,171 ^d	-1,284	,202	-,134	,145
	tauxcho_4	,109 ^d	,800	,426	,084	,140

a. Predictors in the Model: (Constant), tauxchoj_1

b. Predictors in the Model: (Constant), tauxchoj_1, tauxchoj_4

c. Predictors in the Model: (Constant), tauxchoj_1, tauxchoj_4, tauxchoj_2

d. Predictors in the Model: (Constant), tauxchoj_1, tauxchoj_4, tauxchoj_2, tauxchoj_3

e. Dependent Variable: tauxchoj

FIG. 10.47 – Les statistiques des variables exclues (Forward)

Il nous reste à voir quels sont les résultats suite à l'utilisation de la méthode *Stepwise*. Cette méthode fonctionne d'abord comme le *Forward*, mais vérifie à chaque itération si les variables sont demeurées significatives. On voit dans la figure 10.48 que les modèles considérés sont les mêmes que lors de l'utilisation du *Forward*, sauf qu'une itération a été ajoutée pour enlever la variable `tauxchoj_4` qui n'est plus significative à l'avant-dernière itération. Ceci constitue une amélioration par rapport au *Forward*.

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	tauxchoj_1	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	tauxchoj_4	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
3	tauxchoj_2	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
4	tauxchoj_3	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
5	.	tauxchoj_4	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: tauxchoj

FIG. 10.48 – Les itérations (Stepwise)

La figure 10.49 nous présente les r_{aj}^2 de chacun des modèles. On voit que c'est le dernier modèle qui est le plus performant avec un r_{aj}^2 de 0,755.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,789 ^a	,622	,618	1,51060
2	,825 ^b	,681	,674	1,39465
3	,837 ^c	,701	,691	1,35773
4	,874 ^d	,764	,754	1,21201
5	,874 ^e	,763	,755	1,20834

a. Predictors: (Constant), tauxchoj_1
 b. Predictors: (Constant), tauxchoj_1, tauxchoj_4
 c. Predictors: (Constant), tauxchoj_1, tauxchoj_4,
 tauxchoj_2
 d. Predictors: (Constant), tauxchoj_1, tauxchoj_4,
 tauxchoj_2, tauxchoj_3
 e. Predictors: (Constant), tauxchoj_1, tauxchoj_2,
 tauxchoj_3

FIG. 10.49 – Le r et r_{aj}^2 des modèles (Stepwise)

La figure 10.50 nous présente les tables ANOVA de chacun des modèles.

ANOVA ^f					
Model		Sum of Squares	df	Mean Square	F
1	Regression	352,721	1	352,721	154,573
	Residual	214,499	94	2,282	
	Total	567,220	95		
2	Regression	386,331	2	193,166	99,312
	Residual	180,888	93	1,945	
	Total	567,220	95		
3	Regression	397,623	3	132,541	71,899
	Residual	169,596	92	1,843	
	Total	567,220	95		
4	Regression	433,544	4	108,386	73,784
	Residual	133,676	91	1,469	
	Total	567,220	95		
5	Regression	432,892	3	144,297	98,828
	Residual	134,328	92	1,460	
	Total	567,220	95		

a. Predictors: (Constant), tauxchoj_1

b. Predictors: (Constant), tauxchoj_1, tauxchoj_4

c. Predictors: (Constant), tauxchoj_1, tauxchoj_4, tauxchoj_2

d. Predictors: (Constant), tauxchoj_1, tauxchoj_4, tauxchoj_2, tauxchoj_3

e. Predictors: (Constant), tauxchoj_1, tauxchoj_2, tauxchoj_3

f. Dependent Variable: tauxchoj

FIG. 10.50 – Les tables ANOVA des modèles (Stepwise)

La figure 10.51 nous présente les tables des coefficients des 5 modèles qui ont été considérés.

Model	Coefficients ^a				
	B	Std. Error	Standardized Coefficients	t	Sig.
1	(Constant)	3,091	,950		,002
	tauxchoj_1	,790	,064	,789	,000
2	(Constant)	1,364	,971		,163
	tauxchoj_1	,603	,074	,602	,000
3	tauxchoj_4	,305	,073	,307	,000
	tauxchoj_1	,769	,098	,767	,000
4	tauxchoj_4	,335	,072	,337	,000
	tauxchoj_2	-,232	,094	-,233	,015
5	(Constant)	1,563	,867		,075
	tauxchoj_1	1,054	,105	1,052	,000
4	tauxchoj_4	-,069	,104	-,070	,507
	tauxchoj_2	-,758	,135	-,761	,000
5	tauxchoj_3	,668	,135	,673	,000
	tauxchoj_1	1,012	,084	1,010	,000
5	tauxchoj_2	-,708	,113	-,711	,000
	tauxchoj_3	,597	,083	,602	,000

a. Dependent Variable: taux choj

FIG. 10.51 – Les tables des coefficients des modèles (Stepwise)

La figure 10.52 nous présente les statistiques relatives aux variables exclues.

Model	Excluded Variables					
	Beta In	t	Sig.	Partial Correlation	Colinearity Statistics	
					Tolerance	
1	tauxchoj_2	-.160 ^a	-1,555	,123	-,159	,376
	tauxchoj_3	,241 ^a	3,295	,001	,323	,680
	tauxchoj_4	,307 ^a	4,157	,000	,396	,629
	tauxcho_1	,142 ^a	,880	,381	,091	,156
	tauxcho_2	-,063 ^a	-,665	,508	-,069	,444
	tauxcho_3	,177 ^a	2,378	,019	,239	,689
	tauxcho_4	,258 ^a	3,584	,001	,348	,687
2	tauxchoj_2	-,233 ^b	-2,475	,015	-,250	,366
	tauxchoj_3	,075 ^b	,769	,444	,080	,365
	tauxcho_1	,060 ^b	,400	,690	,042	,153
	tauxcho_2	-,156 ^b	-1,749	,084	-,179	,420
	tauxcho_3	-,003 ^b	-,032	,975	-,003	,432
	tauxcho_4	-,011 ^b	-,072	,943	-,007	,145
3	tauxchoj_3	,673 ^c	4,945	,000	,460	,140
	tauxcho_1	,073 ^c	,497	,620	,052	,153
	tauxcho_2	,054 ^c	,364	,716	,038	,148
	tauxcho_3	,230 ^c	2,065	,042	,212	,254
	tauxcho_4	,053 ^c	,350	,727	,037	,141
4	tauxcho_1	-,028 ^d	-,210	,834	-,022	,149
	tauxcho_2	,115 ^d	,865	,389	,091	,146
	tauxcho_3	-,171 ^d	-1,284	,202	-,134	,145
	tauxcho_4	,109 ^d	,800	,426	,084	,140
5	tauxchoj_4	-,070 ^e	-,666	,507	-,070	,235
	tauxcho_1	-,024 ^e	-,182	,856	-,019	,149
	tauxcho_2	,097 ^e	,739	,462	,077	,151
	tauxcho_3	-,173 ^e	-1,300	,197	-,135	,145
	tauxcho_4	-,002 ^e	-,019	,985	-,002	,358

a. Predictors in the Model: (Constant), tauxchoj_1

b. Predictors in the Model: (Constant), tauxchoj_1, tauxchoj_4

c. Predictors in the Model: (Constant), tauxchoj_1, tauxchoj_4, tauxchoj_2

d. Predictors in the Model: (Constant), tauxchoj_1, tauxchoj_4, tauxchoj_2, tauxchoj_3

e. Predictors in the Model: (Constant), tauxchoj_1, tauxchoj_2, tauxchoj_3

f. Dependent Variable: tauxchoj

FIG. 10.52 – Les statistiques des variables exclues (Stepwise)

Suite aux différents résultats observés, quel modèle prendriez-vous, et pourquoi ?

10.4 Exercices du chapitre

Exercice 1 Reprenez l'exercice 2 du chapitre 9 (avec la base de données `endettement.sav`) et incorporez la variable `proprio` au modèle de régression.

Exercice 2 Le responsable du département d'assemblage de l'entreprise Sigmatex est préoccupé par les différences assez marquées entre les individus de son département concernant l'assemblage de montages transistorisés. Ces montages sont effectués pour le compte d'une entreprise de la région de Montréal, laquelle vient d'obtenir un important contrat d'une firme américaine.

L'entreprise Sigmatex se voit dans l'obligation d'accroître son personnel pour répondre à la demande de ce produit. Pour assurer éventuellement une meilleure sélection à l'embauche, on envisage d'abord d'étudier, à partir des individus œuvrant déjà dans le département d'assemblage, s'il existe des facteurs pouvant expliquer la bonne ou mauvaise performance des individus en ce qui a trait au nombre moyen de montages assemblés par semaine.

On a donc décidé de soumettre à un certain nombre d'individus une batterie de tests d'aptitudes. Les aptitudes mesurées sont les suivantes :

x_1 : **Spatialisation.** (Aptitude à concevoir visuellement des formes géométriques et à comprendre la représentation d'objets en deux dimensions.)

x_2 : **Perception des formes.** (Aptitude à percevoir les détails pertinents des objets, reproduction ou documents écrits.)

x_3 : **Coordination visuo-motrice.** (Aptitude à coordonner les mouvements des yeux, des mains ou des doigts rapidement et avec précision.)

La variable dépendante y représente le nombre total de montages assemblés au cours des deux dernières semaines. On a également noté le sexe de l'individu qui est identifié ici par la variable auxiliaire x_4 avec $x_4 = 1$ si le sexe est féminin et $x_4 = 0$ si le sexe est masculin.

Les données sont disponibles dans le fichier `tests.sav`.

Développez d'abord un modèle de régression qui servira de modèle prévisionnel pour estimer le nombre de montages en fonction des diverses variables explicatives que l'on vient de présenter. Répondez ensuite à la question suivante :

- Trois individus se sont présentés au service d'embauche de l'entreprise et ont subi les différents tests. Les résultats obtenus sont les suivants :

	Spatialisation	Perception des formes	Coordination visuo-motrice	Sexe
Individu A	110	85	78	Féminin
Individu B	104	102	98	Masculin
Individu C	112	96	101	Féminin

Un individu se voit offrir un emploi en autant que le nombre de montages prévu selon les résultats aux tests se situe au-dessus de la moyenne globale obtenue des 45 individus qui ont participé à l'étude. Est-ce qu'il y a une candidature intéressante ? Justifiez votre choix.

Exercice 3 Reprenons l'exemple de la sous-section 10.1.1 : on a utilisé la base de données `employés.sav` pour élaborer le modèle suivant :

$$Y_{\text{salaire}} = \beta_0 + \beta_1 X_{\text{ancien}} + \beta_2 X_{\text{sexé}} + \epsilon.$$

On s'intéresse encore à élaborer un modèle pour expliquer la variable salaire, mais cette fois-ci en tenant compte de la fonction de l'employé (variable `fonction`). On veut donc expliquer la variable `salaire` à l'aide des variables `ancien`, `sexé` et `fonction`. Faites l'analyse complète. Est-il préférable de prendre un modèle additif ou multiplicatif ?

Exercice 4 Reprenons la base de données `chomage.sav` (exemple 10.2.1). Étudiez à nouveau la relation `tauxchoj_1` \Rightarrow `tauxchoj` mais en tenant compte cette fois-ci de l'effet du trimestre.

Chapitre 11

Validité d'un modèle en régression linéaire

Au chapitre précédent, nous avons étudié le modèle de la régression multiple qui s'exprime ainsi :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon.$$

Le modèle a été utilisé à plusieurs reprises en supposant que les hypothèses de bases liées à la régression étaient toutes vérifiées. Il faut bien comprendre que si les hypothèses ne sont pas vérifiées, alors le modèle obtenu ne sera pas aussi efficace que souhaité et peut mener à des estimations erronées.

Aucune analyse en régression n'est considérée comme complète sans la vérification des hypothèses. Ce chapitre est entièrement consacré à la vérification de chacune des hypothèses du modèle de régression linéaire.

11.1 Rappel des hypothèses

Pour qu'un modèle de régression linéaire multiple soit valide, il faut que l'hypothèse suivante soit respectée :

On suppose que les ϵ_i sont des variables aléatoires normales et indépendantes de moyenne $E(\epsilon) = 0$ et de variances identiques $\text{Var}(\epsilon) = \sigma_{\text{résiduelle}}^2$.

Il faut aussi s'assurer que les variables explicatives ne soient pas trop fortement corrélées entre elles (problème de multicolinéarité).

On peut décomposer ces hypothèses de la façon suivante :

- a) La relation à modéliser est linéaire ;
- b) Les résidus (ϵ_i) ont tous la même variance $\sigma_{\text{résiduelle}}^2$;
- c) Les résidus sont distribués normalement ;
- d) Les résidus sont indépendants ;
- e) Les variables explicatives ne sont pas (trop) interdépendantes.

Nous verrons dans les sections qui suivent comment vérifier dans la pratique si ces conditions sont respectées.

11.2 La multicolinéarité

Souvent, deux ou plusieurs des variables indépendantes d'une régression linéaire multiple apportent de l'information semblable. C'est-à-dire qu'il y a parfois existence d'un lien linéaire entre deux des variables explicatives, et elles sont alors corrélées. Par exemple, si on tente d'expliquer la satisfaction générale des employés au sein d'une entreprise, on

peut prendre parmi les variables explicatives la satisfaction par rapport à l'entente entre le supérieur immédiat et ses employés (X_1) et la satisfaction par rapport à l'estime témoignée pour un travail bien fait (X_2). Ces deux variables apportent de l'information par rapport au lien avec le supérieur immédiat, et ont de bonnes chances d'être corrélées.

Lorsque des variables sont ainsi corrélées, on dit qu'il y a de la **multicolinéarité**. En pratique, il n'est pas rare qu'il y ait de la multicolinéarité, mais lorsque celle-ci est trop importante quelques problèmes peuvent survenir. En effet, la présence d'une forte colinéarité rend instables les valeurs numériques des coefficients de régression b_0, b_1, \dots, b_k . Ainsi, l'ajout ou le retrait de quelques observations produit des changements très marqués.

La colinéarité présente des inconvénients concernant l'étude de l'apport individuel des variables explicatives dans le modèle. Par exemple, on peut conclure que la régression est significative, mais que l'apport marginal de chacune des variables est non significatif.

Toutefois, si l'objectif de l'étude est d'obtenir simplement une équation de régression pour des fins d'estimations et de prévisions au lieu d'évaluer l'effet marginal de chaque variable explicative, alors une forte colinéarité entre les variables explicatives présente moins d'inconvénients.

11.2.1 Déetecter la multicolinéarité.

Il existe une statistique appelée VIF (*variance inflation factor*) qui est très utile pour évaluer la présence de multicolinéarité. Plus précisément, supposons que nous avons le modèle suivant :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

Pour chacune des variables explicatives, une régression multiple avec cette variable comme variable dépendante et avec les autres variables explicatives est calculée. On obtient ainsi

un r_i^2 pour chaque variable X_i :

$$\begin{aligned} r_1^2 &\text{ provient de } X_1 = \beta'_0 + \beta'_1 X_2 + \beta'_2 X_3 + \epsilon; \\ r_2^2 &\text{ provient de } X_2 = \beta''_0 + \beta''_1 X_1 + \beta''_2 X_3 + \epsilon; \\ r_3^2 &\text{ provient de } X_3 = \beta'''_0 + \beta'''_1 X_1 + \beta'''_2 X_2 + \epsilon. \end{aligned}$$

Ensuite le coefficient r_i^2 est utilisé dans les calculs de la statistique VIF de la façon suivante :

$$\text{VIF}_i = \frac{1}{1 - r_i^2}.$$

Si une variable explicative est indépendante des autres variables explicatives, alors le coefficient r_i^2 de la régression obtenue entre les variables explicatives sera petit, et ainsi le VIF sera proche de 1 ($\text{VIF} = 1/(1-0)=1$). D'autre part, si le r_i^2 prend une grande valeur, le VIF sera grand. Une règle du pouce tend à montrer qu'un $\text{VIF} > 10$ démontre une grande multicolinéarité. Par conséquent, **dès que l'un des VIF est plus grand que 10, on considère que les tests sur les β_j et le classement des variables explicatives ne sont pas valides.**

Pour obtenir les VIF lors d'une analyse en régression, il suffit de cocher l'option **Collinearity diagnostics** dans le bouton **Statistics**.

Dans les deux exemples du chapitre 9, on retrouve les VIF dans les tables des coefficients, dans la dernière colonne. Ainsi on voit que pour l'exemple 9.2.1, les VIF associés à la publicité et au bonus ont tous deux une valeur de 1,213, ce qui est plus petit que 10. Les tests et le classement faits dans cet exemple sont donc valides. De même, on peut voir que les VIF de l'exemple 9.3 sont tous inférieurs à 2.

11.2.2 Corriger la multicolinéarité

Une solution rapide consiste à enlever du modèle la ou les variables explicatives qui ont une trop grande multicolinéarité avec les autres. Par la perte d'information, cette

solution revêt vraisemblablement un caractère impopulaire. Ajouter des données (un plus grand échantillon) à l'étude brise parfois le lien de multicolinéarité entre les variables explicatives d'un modèle. Cependant, cette solution est souvent impossible. L'analyste peut aussi opter pour une analyse en composante principale et créer des indices.

Toutefois, même s'il y a présence de forte multicolinéarité, il est possible d'utiliser l'équation de la régression afin de faire des estimations et des prévisions.

11.3 Les résidus

Le résidu e_i est la différence (ou la distance) entre la vraie valeur de y_i et son estimé $\hat{y}_i = b_0 + \sum_{j=1}^k b_j x_{ji}$ (appelé *fit*) issu du modèle de régression linéaire :

$$e_i = y_i - \hat{y}_i.$$

Les résidus sont simplement la mesure de la variation inexplicable par le modèle de régression. Ainsi les résidus sont un indicateur de la performance du modèle. Il est intuitif et logique de croire qu'une bonne régression obtient de faibles erreurs dans ses estimations (donc de faibles résidus), tandis qu'une régression de faible qualité (qui modélise moins bien la relation entre Y et X) laisse plus d'erreurs dans ses estimations (donc de plus grands résidus).

Afin de faciliter l'analyse des résidus, il est courant d'utiliser les résidus standardisés (*standardized residuals*). Ainsi, s'il est vrai que les résidus sont distribués normalement, les résidus standardisés suivront une loi $N(0, 1)$.

Par conséquent, les résidus standardisés devraient respecter les balises suivantes :

- 68,26 % des résidus se situeront entre ± 1 écarts-type ;
- 95,44 % des résidus se situeront entre ± 2 écarts-type ;
- 99,72 % des résidus se situeront entre ± 3 écarts-type.

11.3.1 Les propriétés des résidus

Propriété 1 La moyenne des résidus est toujours égale à zéro. Ceci est dû à la façon dont b_0 et b_1 sont calculés (la méthode des moindres carrés).

Propriété 2 Si les hypothèses a) à c) (vues à la section 11.1) sont vraies, alors les résidus devraient être distribués un peu partout, autour de 0 (leur moyenne) sans qu'aucune forme particulière ne se dégage du graphique des résidus.

Propriété 3 Si les hypothèses a) à d) (vues à la section 11.1) sont vraies, alors les résidus devraient être distribués normalement (selon une cloche) autour de la moyenne.

Si certaines hypothèses sont violées, elles laisseront des traces dans les résidus. Par exemple, considérons les graphes des figures 11.1 et 11.2 (dans lesquels on retrouve les résidus standardisés pour l'axe vertical et les estimés (*fitted values*) pour l'axe horizontal). Le premier graphe contient des résidus qui respectent les hypothèses a) à c) car leur répartition est assez uniforme autour de 0. Le deuxième graphe illustre qu'au moins une des conditions a) à c) n'a pas été respectée car il est évident que leur répartition n'est pas uniforme.

Les sections qui suivent montrent la façon de procéder pour vérifier les hypothèses de validité d'un modèle de régression à partir des résidus.

11.4 Vérification de la linéarité

La première hypothèse vue à la section 11.1 stipule que la relation entre la variable dépendante Y et les variables explicatives doit être linéaire.

Après avoir effectué une régression, cette hypothèse peut être vérifiée par l'entremise de l'étude des résidus. Cette section est entièrement consacrée à la détection et à la

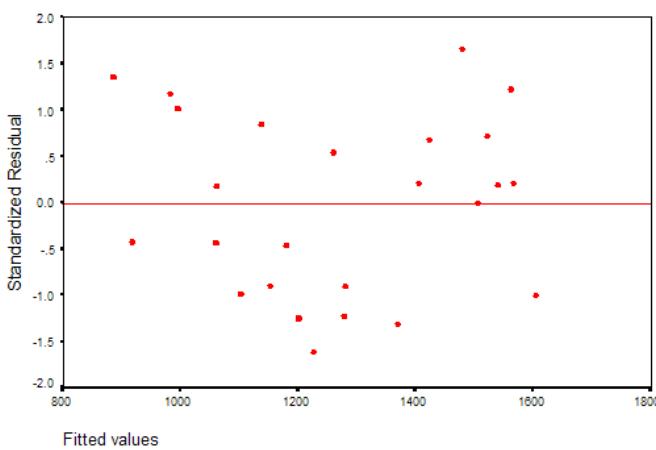


FIG. 11.1 – Répartition assez uniforme des résidus

correction de la violation de l'hypothèse de linéarité.

11.4.1 Détection de la violation de la linéarité

Observer les graphes de la variable dépendante y en fonction de chacune des variables explicatives peut donner une bonne indication à savoir si la relation est linéaire ou pas. Cependant, une faible déviation à la linéarité peut être difficile à voir à l'aide des graphiques. Une telle déviation sera visible dans les résidus. Voyons ceci dans un exemple.

Exemple 11.4.1 Pour aider à la vente de ses produits, une entreprise utilise la technique du télémarketing. Le chef des ressources humaines est intéressé à faire l'étude du nombre d'appels réussis par employé. Les données sur le nombre d'appels moyen réussi par jour et le temps moyen en secondes pour chaque appel (basé sur une moyenne de 20 jours de travail) ainsi que sur l'ancienneté des employés (en mois) sont disponibles pour l'étude. La base de données se nomme `appels.sav`.

La figure 11.3 donne un aperçu des données.

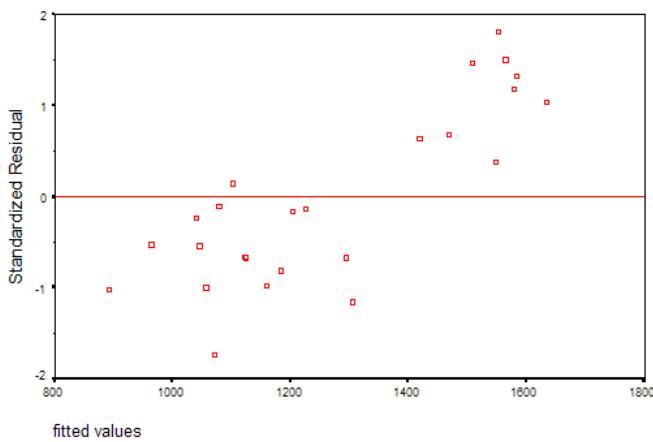


FIG. 11.2 – Répartition non-uniforme des résidus

Le modèle qui nous intéresse est le suivant :

$$Y_{\text{nbappel}} = \beta_0 + \beta_1 X_{\text{moisexp}} + \beta_2 X_{\text{temps}} + \epsilon.$$

Nous ferons donc une analyse en régression linéaire multiple, et nous examinerons le graphe des résidus en fonction des prédictions pour voir si le modèle semble bien linéaire. Pour générer les sorties de cet exemple, les commandes sont les suivantes :

Menu SPSS :	→ Analyse → Regression → Linear...
Dans la fenêtre Dependant :	→ nbappels (la variable dépendante)
Dans la fenêtre Independant(s) :	→ moisexp, temps (les variables explicatives)
Dans le bouton Statistics... :	✓ Collinearity diagnostics (et laisser les autres crochets)
Dans le bouton Plots... :	→ Y : ZRESID (les résidus standardisés) → X : ZPRED (les prédictions standardisés)
Dans le bouton Save... :	→ Residuals ✓ Standardized

	ident	moisexp	nbappel	temps
1	1	10,00	18,00	29,00
2	2	10,00	19,00	34,00
3	3	11,00	22,00	40,00
4	4	14,00	23,00	43,00
5	5	15,00	25,00	39,00
6	6	17,00	28,00	48,00
7	7	18,00	29,00	50,00
8	8	20,00	29,00	51,00
9	9	20,00	31,00	55,00
10	10	21,00	31,00	54,00
11	13	24,00	31,00	60,00
12	17	25,00	31,00	54,00
13	12	22,00	32,00	55,00
14	14	25,00	32,00	62,00
15	15	25,00	32,00	61,00
16	11	22,00	33,00	65,00
17	16	25,00	33,00	60,00
18	18	28,00	33,00	63,00
19	19	29,00	33,00	59,00
20	20	30,00	34,00	55,00

FIG. 11.3 – Les données de l'exemple

Fixons les seuils à $\alpha = 0,05$ pour cette analyse.

La figure 11.4 nous montre que le modèle a un r_{aj}^2 de 0,936 ; il semble donc que 93,6 % de la variation du nombre d'appels réussis soit expliqué par les mois d'expérience et la durée moyenne de chaque appel, ce qui est excellent.

On peut maintenant résoudre le test suivant :

H_0 : La régression est non significative dans la population (tous les $\beta_j = 0$).

H_1 : La régression est significative dans la population (au moins un des $\beta_j \neq 0$).

Puisque la p -value de la table ANOVA est de 0,000, ce qui est plus petit que $\alpha = 0,05$, on rejette H_0 . Ainsi, au risque de se tromper une fois sur 20, on peut admettre que la régression est significative.

Maintenant, la sortie 11.6 nous permet de vérifier les VIF, de voir lesquels des paramètres du modèle sont significatifs et d'écrire l'équation.

Les VIF ont une valeur de 4,57. Puisque c'est plus petit que 10, on peut considérer

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,971 ^a	,942	,936	1,24070

a. Predictors: (Constant), temps, moisexp
b. Dependent Variable: nbappel

FIG. 11.4 – Le r et le r_{aj}^2

ANOVA ^b					
Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	428,781	2	214,391	139,274
	Residual	26,169	17	1,539	
	Total	454,950	19		,000 ^a

a. Predictors: (Constant), temps, moisexp
b. Dependent Variable: nbappel

FIG. 11.5 – La table ANOVA

que les tests sur les β_j sont valides.

La p -value associée à la variable `moisexp` étant de 0,003, on peut conclure que le paramètre β_1 est significatif au seuil $\alpha = 0,05$. De même, la p -value associée à la variable `temps` étant de 0,000, on conclut que le paramètre β_2 est significatif. La cote- t de `temps` étant supérieure à celle de `moisexp` (4,512 comparativement à 3,529), il semble que ce soit la variable `temps` qui ait le plus d'impact sur la variable `nbappel`. Que pensez-vous de ce classement ?

L'équation du modèle est la suivante :

$$\hat{y}_{\text{nbappel}} = 7,728 + 0,349x_{\text{moisexp}} + 0,271x_{\text{temps}}.$$

Le b_0 de 7,728 n'a pas vraiment de sens ici car il faudrait considérer un employé qui n'a aucun mois d'expérience et dont le temps moyen en secondes pour chaque appel est nul...

Le b_1 de 0,349 nous indique que lorsque l'expérience augmente d'un mois et que le

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	7,728	1,648		4,689	,000		
moisexp	,349	,099	,439	3,529	,003	,219	4,570
temps	,271	,060	,561	4,512	,000	,219	4,570

a. Dependent Variable: nbappel

FIG. 11.6 – La table des coefficients

temps moyen pour chaque appel demeure constant, le nombre d'appels réussis augmente en moyenne de 0,349.

Le b_2 de 0,271 nous indique que lorsque le temps moyen pour chaque appel augmente d'une seconde et que l'expérience demeure constante, le nombre d'appels réussis augmente en moyenne de 0,271. On remarque que ces interprétations sont assez limitées...

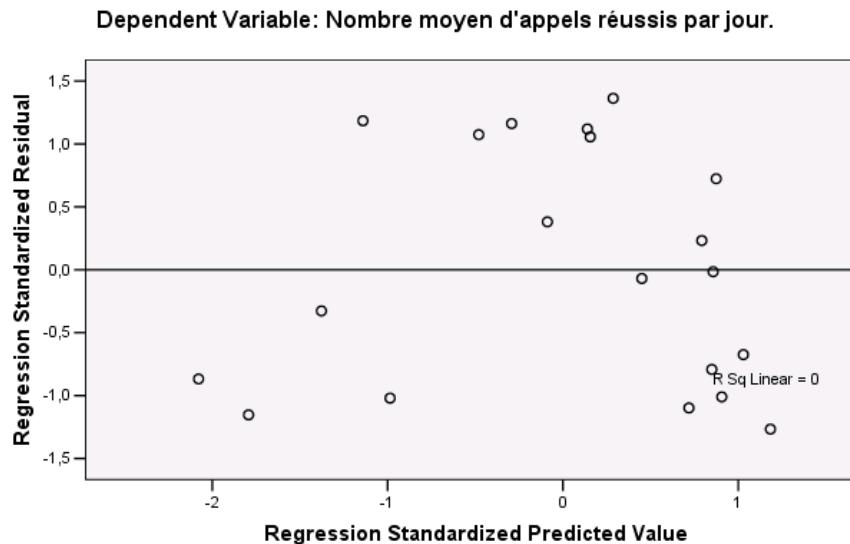


FIG. 11.7 – Le graphe des résidus et des prédictions standardisés

Finalement, la figure 11.7 nous permet d'étudier les résidus. Ceux-ci ne sont manifestement pas répartis uniformément. On remarque la forme d'une courbe, il semble qu'une

relation curvilinéaire soit présente. Il faudra maintenant examiner le graphe de chaque variable explicative versus les résidus pour voir laquelle des variables génère cette courbe (ça pourrait être les deux).

Après l'examen des deux graphes, il semble que ce soit la variable `moisexp` qui est responsable de la violation à la linéarité du modèle (premier graphe).

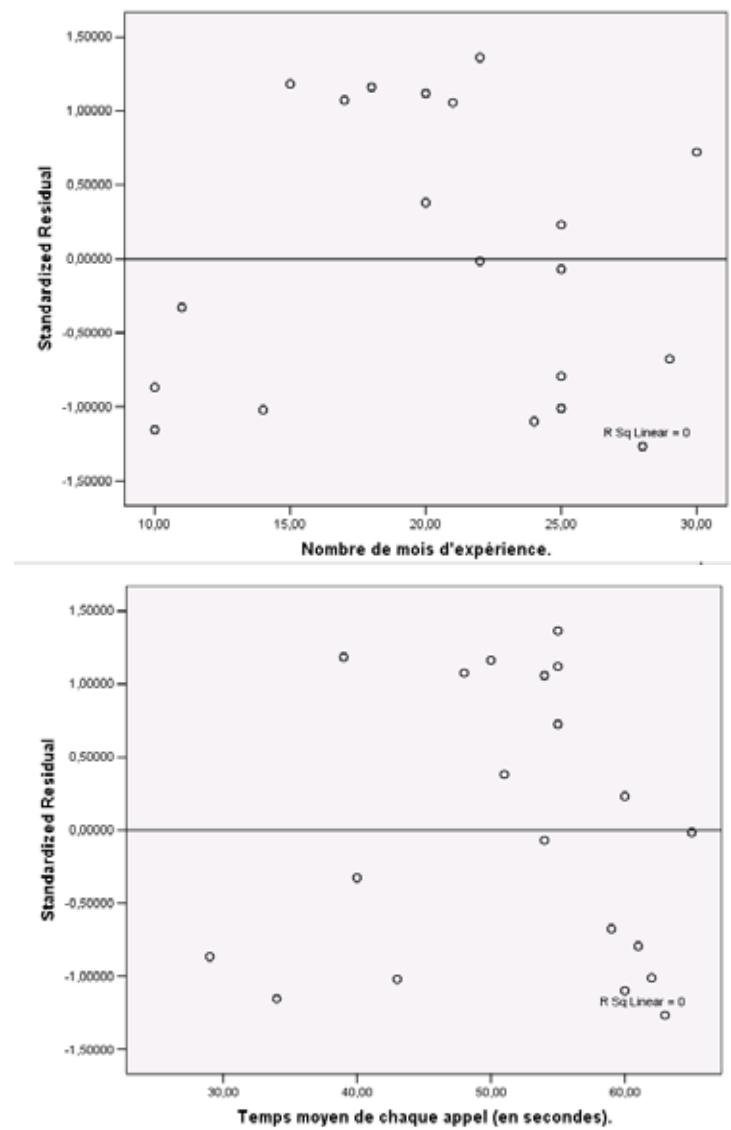


FIG. 11.8 – Les graphes avec les variables explicatives

Maintenant que nous avons vu comment détecter la violation à la linéarité, nous pouvons voir comment la corriger.

11.4.2 Correction de la violation à la linéarité

Lorsque l'hypothèse de linéarité est violée, il n'est pas toujours clair de savoir comment corriger la situation. Le problème vient du fait qu'il existe une infinité de modèles qui peuvent faire l'affaire.

Pour corriger la violation à la linéarité, plusieurs méthodes sont proposées :

- La régression polynomiale ;
- La transformation réciproque ;
- La transformation logarithmique de x ;
- La transformation logarithmique de x et de y.

Après avoir effectué une correction, il faut ensuite regarder les graphes des résidus afin de voir si les motifs ont disparu. Si aucun motif n'est présent, la méthode a corrigé la violation.

La régression polynomiale

Une correction couramment utilisée consiste à ajouter un terme avec une puissance différente de 1 à l'équation de régression multiple, selon la forme de la courbe de la relation. Par exemple, on pourrait avoir les modèles suivants :

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

ou encore

$$Y = \beta_0 + \beta_1 X + \beta_2 \sqrt{X} + \epsilon.$$

En théorie, nous pouvons ajouter autant de termes et de puissances différentes que voulu, mais en pratique, il est fréquent que l'ajout d'une seule puissance soit suffisant.

Aussi, l'utilisation de plusieurs puissances d'une même variable introduit de la colinéarité, ce qui n'est pas souhaitable. Cependant l'utilisation de variables centrées réduit l'effet de la colinéarité. En d'autres mots, au lieu d'utiliser les variables X , X^2 , X^3 et X^4 , on peut utiliser les variables centrées suivantes dans le modèle : $(X - \bar{X})$, $(X - \bar{X})^2$, $(X - \bar{X})^3$ et $(X - \bar{X})^4$ où \bar{X} est la moyenne de la variable X .

Exemple 11.4.2 L'analyse des résidus de l'exemple de l'entreprise qui utilise le télémarketing montrait que l'hypothèse de linéarité n'est pas respectée, et que c'est la variable `moisexp` qui semble être en cause.

Nous allons ici essayer deux types de correction polynomiale, et voir lequel effectue la meilleure correction. Les deux modèles seront les suivants :

$$\text{Modèle 1 : } Y_{\text{nbappel}} = \beta_0 + \beta_1 X_{\text{moisexp}} + \beta_2 X_{\text{moisexp}}^2 + \beta_3 X_{\text{temps}} + \epsilon.$$

$$\text{Modèle 2 : } Y_{\text{nbappel}} = \beta_0 + \beta_1 X_{\text{moisexp}} + \beta_2 \sqrt{X_{\text{moisexp}}} + \beta_3 X_{\text{temps}} + \epsilon.$$

Il faut d'abord créer les variables X_{moisexp}^2 et $\sqrt{X_{\text{moisexp}}}$ en allant dans `Transform` puis `Compute` :

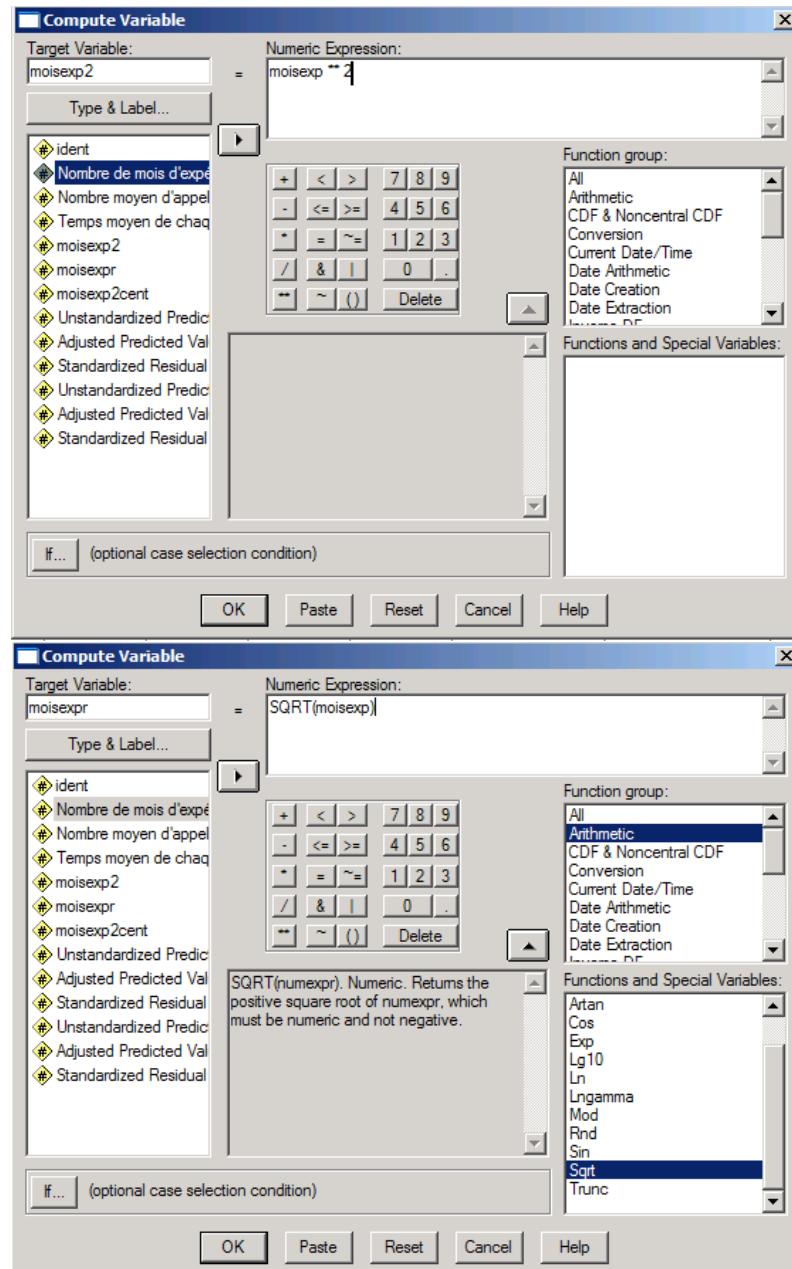


FIG. 11.9 – La création des variables X_{moisexp}^2 et $\sqrt{X_{\text{moisexp}}}$.

Étude du modèle 1 : $Y_{\text{nbappel}} = \beta_0 + \beta_1 X_{\text{moisexp}} + \beta_2 X_{\text{moisexp}}^2 + \beta_3 X_{\text{temps}} + \epsilon$.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,985 ^a	,970	,964	,92600

a. Predictors: (Constant), moisexp2, temps, moisexp

b. Dependent Variable: nbappel

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	441,230	3	147,077	171,524	,000 ^a
	Residual	13,720	16	,857		
	Total	454,950	19			

a. Predictors: (Constant), moisexp2, temps, moisexp

b. Dependent Variable: nbappel

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,800	2,195	,364	,720		
	moisexp	1,733	,371	2,179	4,676	,000	,009 115,2
	temps	,119	,060	,247	1,989	,064	,122 8,176
	moisexp2	-,030	,008	-1,479	-3,810	,002	,013 79,96

a. Dependent Variable: nbappel

FIG. 11.10 – Les sorties de la régression.

On a d'abord les sorties habituelles de la régression. On peut remarquer que le r^2_{aj} est passé de 0,936 à 0,964. Aussi, on voit que le fait d'introduire la variable X_{moisexp}^2 a fait « exploser » les VIF... Le modèle s'écrit maintenant

$$\hat{y}_{\text{nbappel}} = 0,800 + 1,733x_{\text{moisexp}} - 0,030x_{\text{moisexp}}^2 + 0,119x_{\text{temps}}.$$

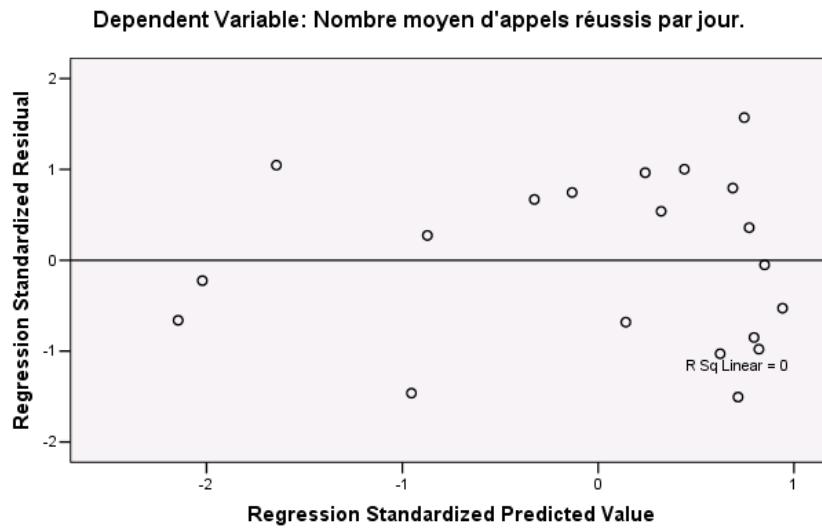


FIG. 11.11 – Les résidus.

L'examen des résidus en fonction des valeurs prédictes (figure 11.11) nous montre que la violation à la linéarité semble avoir été corrigée, et ceci est confirmé par l'examen des graphes des résidus en fonction de chacune des variables explicatives (figure 11.12).

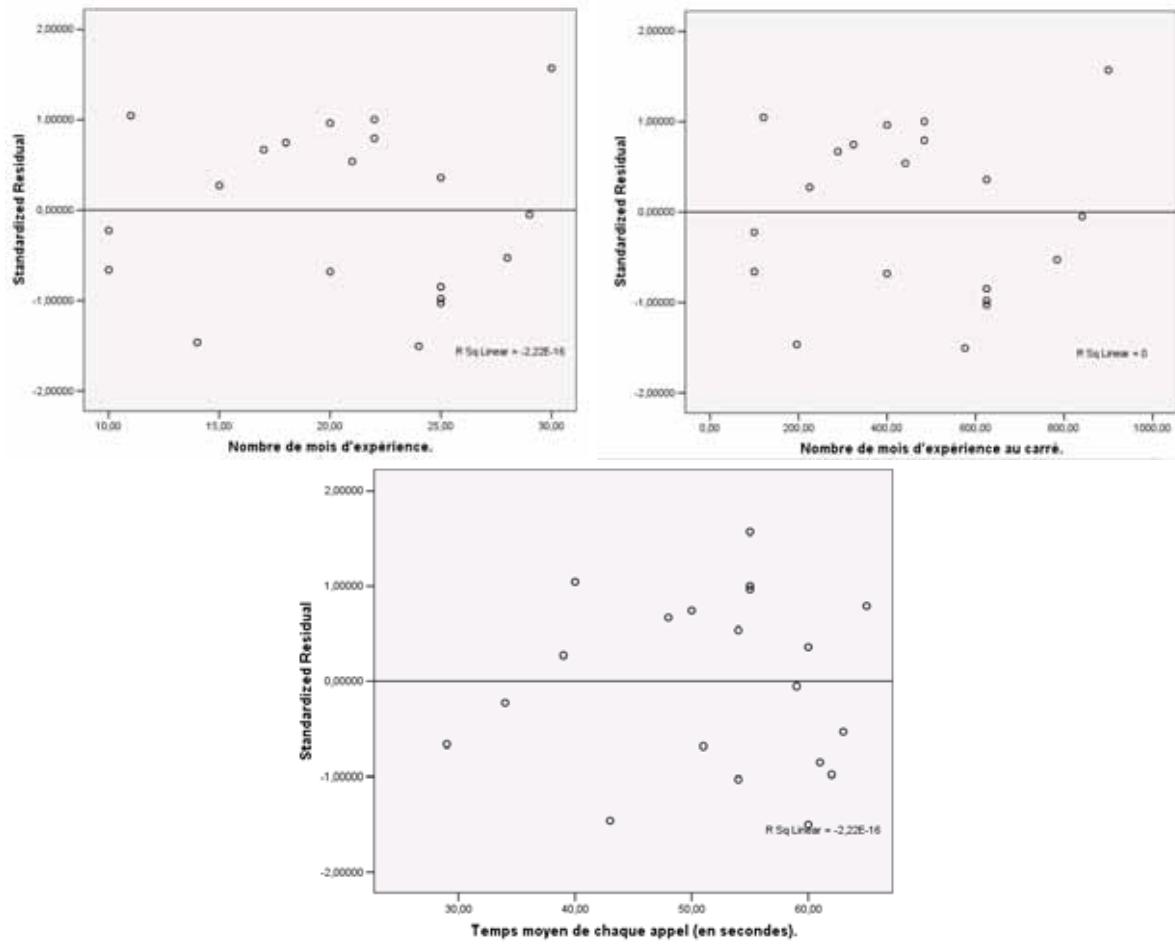


FIG. 11.12 – Les résidus en fonction de chacune des variables.

Étude du modèle 2 : $Y_{\text{nbappel}} = \beta_0 + \beta_1 X_{\text{moisexp}} + \beta_2 \sqrt{X_{\text{moisexp}}} + \beta_3 X_{\text{temps}} + \epsilon$.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,986 ^a	,972	,967	,88800

a. Predictors: (Constant), moisexpr, temps, moisexp

b. Dependent Variable: nbappel

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	442,333	3	147,444	186,985	,000 ^a
	Residual	12,617	16	,789		
	Total	454,950	19			

a. Predictors: (Constant), moisexpr, temps, moisexp

b. Dependent Variable: nbappel

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
1	(Constant) -29,333	9,017		-3,253	,005		
	temps ,130	,055	,269	2,367	,031	,134	7,438
	moisexp -1,673	,493	-2,103	-3,394	,004	,005	221,5
	moisexpr 19,186	4,628	2,809	4,146	,001	,004	264,9

a. Dependent Variable: nbappel

FIG. 11.13 – Les sorties de la régression.

Regardons d'abord les sorties habituelles de la régression. On peut remarquer que le r^2_{aj} est maintenant de 0,967, ce qui est légèrement supérieur au $r^2_{\text{aj}} = 0,964$ du modèle 1.

Aussi, on voit qu'encore une fois, les VIF sont très grands. Le modèle 2 s'écrit

$$\hat{y}_{\text{nbappel}} = -29,333 - 1,673x_{\text{moisexp}} + 19,186\sqrt{x_{\text{moisexp}}} + 0,130x_{\text{temps}}.$$

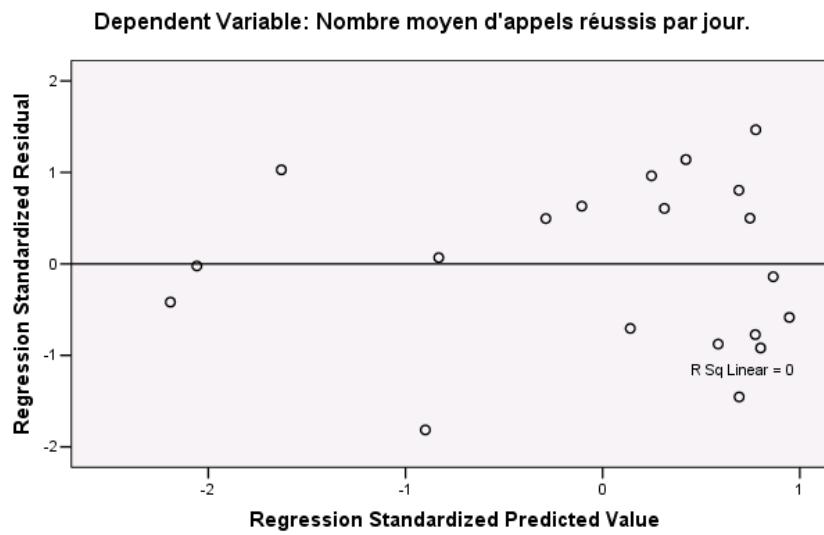


FIG. 11.14 – Les résidus.

L'examen des résidus en fonction des valeurs prédictes (figure 11.14) nous montre que la violation à la linéarité semble avoir été corrigée, et ceci est confirmé par l'examen des graphes des résidus en fonction de chacune des variables explicatives (figure 11.15).

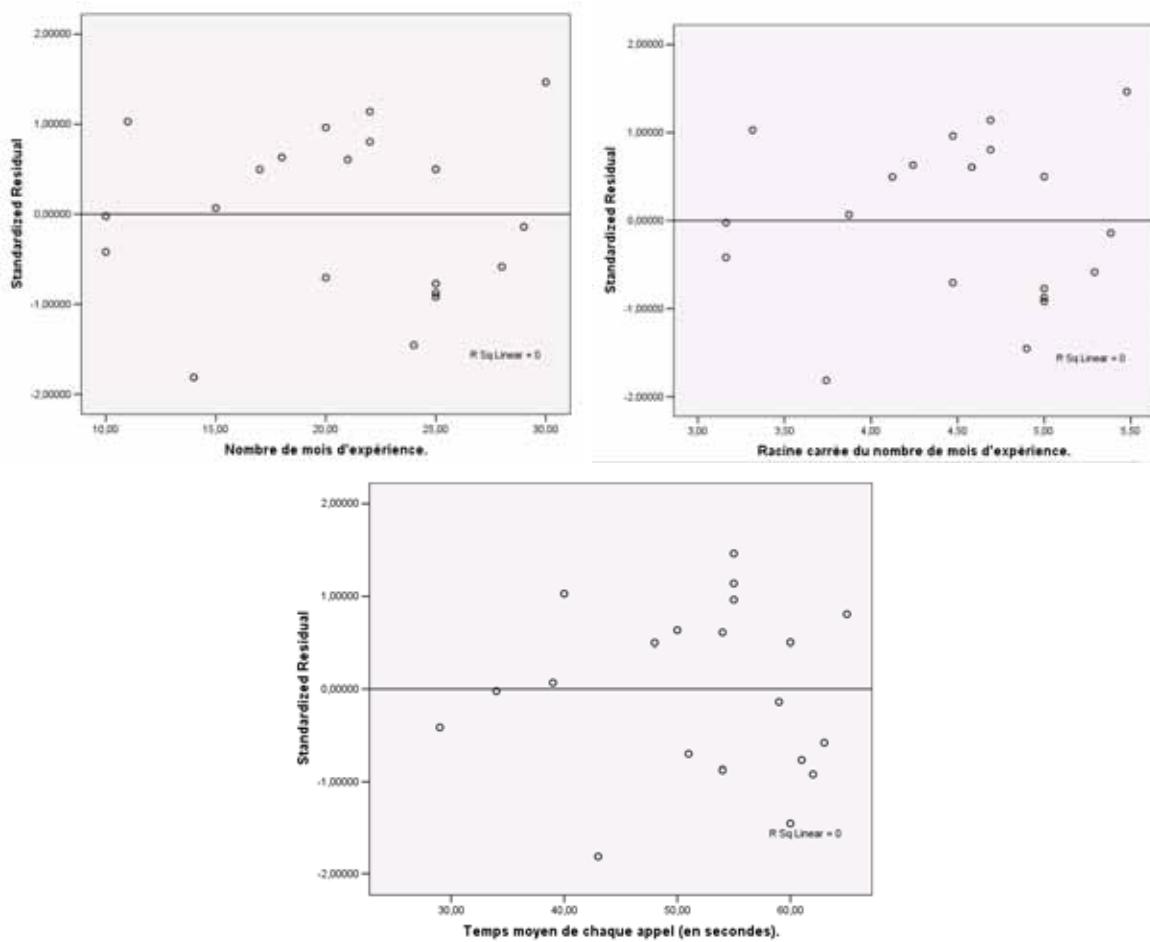


FIG. 11.15 – Les résidus en fonction de chacune des variables.

Quel modèle choisir ?

L'étude des résidus a montré que les deux modèles ont corrigé la violation à la linéarité. Les deux modèles ont un r^2_{aj} très semblable. Nous verrons plus tard que le deuxième modèle respecte mieux l'hypothèse de normalité des résidus (la vérification de la normalité des résidus se fera dans une section subséquente).

La transformation réciproque

La transformation réciproque est une autre transformation qui est souvent utilisée pour corriger la violation à la linéarité. Celle-ci consiste à utiliser la variable $\frac{1}{x}$ dans le modèle :

$$Y = \beta_0 + \beta_1 \left(\frac{1}{x} \right) + \epsilon.$$

Exemple 11.4.3 Une étude a été menée sur la distance franchie par gallon d'essence (Y) en fonction du poids de la voiture (X). La base de données se nomme `consommation.sav`. Le graphe de la relation illustre la relation entre la variable dépendante et la variable explicative :

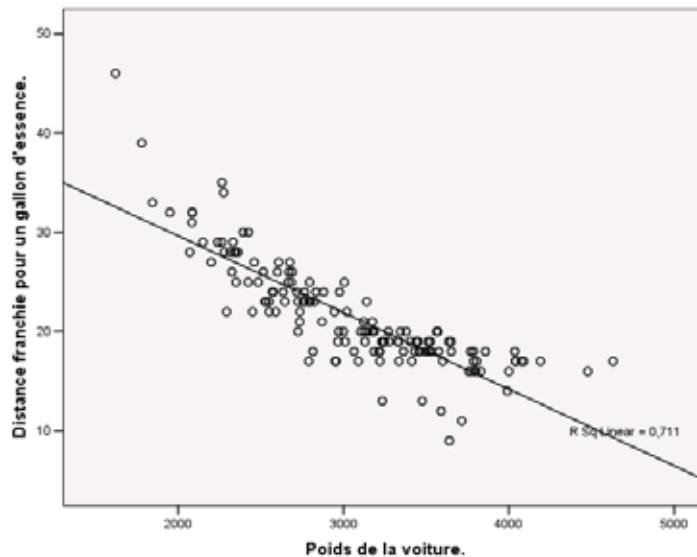


FIG. 11.16 – Le graphe de la relation.

On voit que plus le poids augmente, plus la distance franchie pour un gallon diminue. Cependant cette diminution semble plutôt curvilinéaire. Voici les sorties associées au modèle

$$Y_{\text{consom}} = \beta_0 + \beta_1 X_{\text{poids}} + \epsilon.$$

Model Summary ^b					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	,843 ^a	,711	,709	2,943	

a. Predictors: (Constant), poids
b. Dependent Variable: consom

ANOVA ^b					
Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression 3022,733	1	3022,733	348,908	,000 ^a
	Residual 1230,205	142	8,663		
	Total 4252,938	143			

a. Predictors: (Constant), poids
b. Dependent Variable: consom

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant) 45,078	1,278		35,266	,000
	poids -,008	,000	-,843	-18,679	,000

a. Dependent Variable: consom

FIG. 11.17 – Les sorties de la régression.

On voit que $r = 0,843$ et $r^2 = 0,711$. Selon ces mesures et les autres sorties le modèle semble bon, mais d'après le graphe nous savons que le modèle semble ne pas respecter l'hypothèse de linéarité.

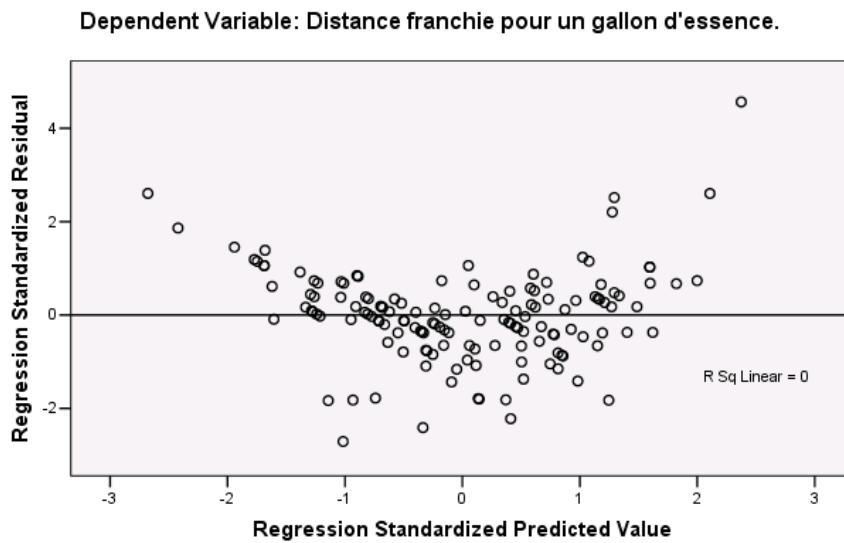
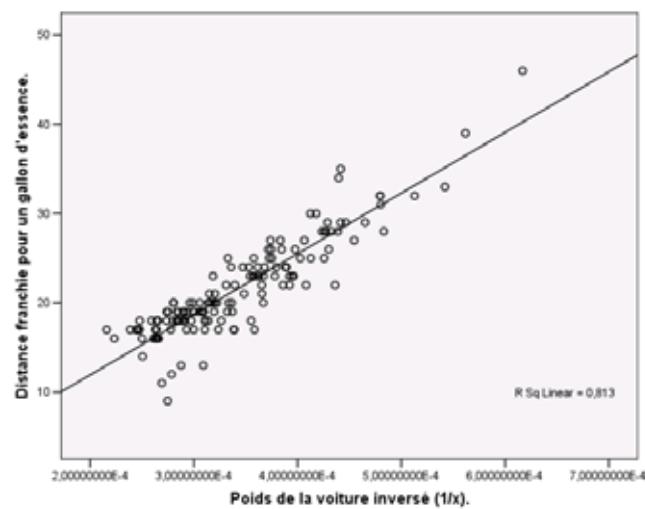


FIG. 11.18 – Les résidus.

Le graphe des résidus (figure 11.18) ne fait que confirmer ce qu'on a vu dans le graphe de la relation. Essayons donc maintenant le modèle

$$Y_{\text{consom}} = \beta_0 + \beta_1 \frac{1}{X_{\text{poids}}} + \epsilon.$$

Le graphe de la nouvelle relation (figure 11.19) nous montre que la violation à la linéarité semble avoir été corrigée.

FIG. 11.19 – Le graphe de la nouvelle relation (avec $1/x$).

Les sorties de la régression (figures 11.20 et 11.21) nous montrent que ce modèle est meilleur que le précédent. Entre autres, le r^2 est passé de 0,711 de à 0,813.

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,902 ^a	,813	,812	2,367

a. Predictors: (Constant), poidsinv
b. Dependent Variable: consom

ANOVA ^b					
Model		Sum of Squares	df	Mean Square	F
1	Regression	3457,522	1	3457,522	617,247
	Residual	795,416	142	5,602	
	Total	4252,938	143		

a. Predictors: (Constant), poidsinv
b. Dependent Variable: consom

FIG. 11.20 – Le r et le r^2 et la table ANOVA (avec $1/x$).

Model	Coefficients ^a				
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-1,664	,959		-1,735	,085
poidsinv	67931,047	2734,253	,902	24,844	,000

a. Dependent Variable: consom

FIG. 11.21 – La table des coefficients de la nouvelle régression (avec $1/x$).

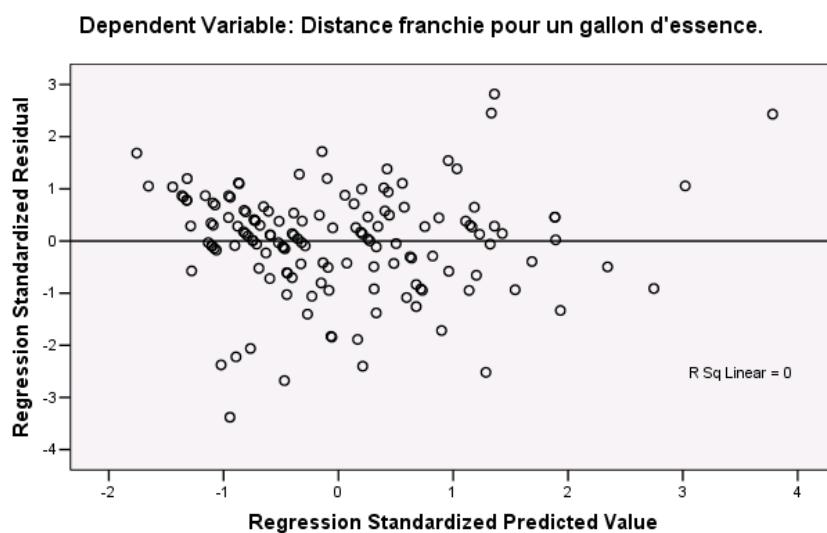


FIG. 11.22 – Les résidus de la nouvelle relation (avec $1/x$).

Les résidus confirment que la violation à la linéarité a été corrigée puisqu'il n'y a plus de motif observable, ils sont distribués de façon assez aléatoire.

La transformation logarithmique

Une autre transformation intéressante peut-être réalisée, celle de la transformation logarithmique de la variable explicative :

$$Y = \beta_0 + \beta_1 \ln(X) + \epsilon$$

où $\ln(X)$ est le logarithme naturel de X . Il faut cependant noter que cette fonction n'est définie que pour des valeurs positives. Autrement dit, $\ln(X)$ n'existe pas lorsque X est négatif. Donc si la variable X admet des valeurs négatives, il ne sera pas possible d'appliquer directement le logarithme à X . Il faudra créer une nouvelle variable X' qui sera une translation de la variable X qui permettra de ne plus avoir de valeurs négatives :

$$X' = X + |\min(X)|$$

où $\min(X)$ est la plus petite valeur (et donc la plus négative) de X .

La transformation double logarithmique

Il est aussi possible d'utiliser le logarithme de Y pour essayer d'obtenir un lien linéaire tout en utilisant le logarithme d'une variable explicative :

$$\ln(Y) = \beta_0 + \beta_1 \ln(X) + \epsilon.$$

Quelques précautions doivent être prises avant d'effectuer des comparaisons. Premièrement, toutes les valeurs de Y et de X doivent être positives (si ce n'est pas le cas, utilisez une translation). Aussi, puisqu'un changement d'unité est appliqué à la variable dépendante, il sera plus difficile de comparer l'efficacité du modèle utilisant $\ln(Y)$ avec l'efficacité d'un modèle utilisant Y . Prenez note que ce problème ne survient pas lorsqu'on ne transforme pas Y .

Le cas des séries chronologiques

Compte tenu qu'une série chronologique n'est qu'une régression où les variables explicatives font intervenir le temps, il est tout aussi de mise d'effectuer l'analyse des résidus.

Toutes les techniques utilisées pour corriger la violation à la linéarité sont applicables aux séries chronologiques.

Par exemple, une série chronologique peut prendre la forme suivante :

$$\hat{y} = b_0 + b_1 t.$$

Une tendance curvilinéaire peut être modélisée ainsi :

$$\hat{y} = b_0 + b_1 t + b_2 t^2.$$

Cependant, certaines tendances prennent la forme d'un S :

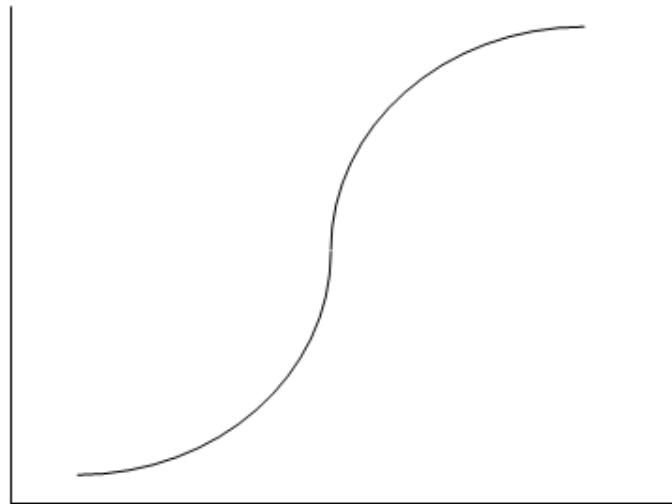


FIG. 11.23 –

Ceci peut représenter une faible demande au départ, puis une forte croissance suivie d'une saturation. Dans ce cas, l'estimation de la droite ne peut être faite directement. Il faut d'abord estimer la droite suivante :

$$Y' = \ln(Y) = \beta_0 + \beta_1 \left(\frac{1}{t} \right).$$

Une fois b_0 et b_1 trouvés, il est possible d'écrire le modèle dans sa forme originale :

$$Y = \exp \left(\beta_0 + \beta_1 \left(\frac{1}{t} \right) \right) = e^{(\beta_0 + \beta_1 (\frac{1}{t}))}.$$

Ensuite, pour faire des prédictions sur des valeurs de Y sachant t à l'aide de SPSS, il suffit de les faire via le modèle calculé dans SPSS :

$$\ln(Y) = \beta_0 + \beta_1 \left(\frac{1}{t} \right)$$

puis d'appliquer la fonction exponentielle en base e (qui est l'inverse de la fonction \ln) aux prédictions obtenues de manière à retrouver les unités originales.

11.5 Vérification de la variance constante

L'hypothèse b) de la section 11.1 stipule que les résidus ont tous la même variance $\sigma_{\text{résiduelle}}^2$. En d'autres mots, il faut que les résidus se dispersent de façon constante autour du modèle de régression.

Dans un graphique résiduel, si la variance est constante, les points devraient se répartir uniformément autour de la droite représentant leur moyenne nulle, et ce, sans motif apparent. Par contre, on peut voir que la constance de la variance n'est pas respectée lorsque par exemple les résidus se dispersent autour de la droite suivant la forme d'un cône. Dans la littérature, le mot hétéroscédasticité est employé, par opposition à homoscédasticité, pour représenter la non constance de la variance.

11.5.1 Détection de la variance non constante

Illustrons à l'aide d'un exemple comment il est possible de détecter le non respect de l'hypothèse de variance constante des résidus.

Exemple 11.5.1 Une étude sur le prix des voitures en fonction de leur puissance (HP) de moteur a été effectuée. Un échantillon de 152 voitures a été étudié. La base de données se nomme `puissance.sav`. Le graphique du prix en fonction de la puissance est le suivant :

Non seulement il semble il y avoir un problème de linéarité pour le modèle $Y_{\text{prix}} = \beta_0 + \beta_1 X_{\text{puissance}} + \epsilon$, il semble aussi que les données sont de plus en plus dispersées à

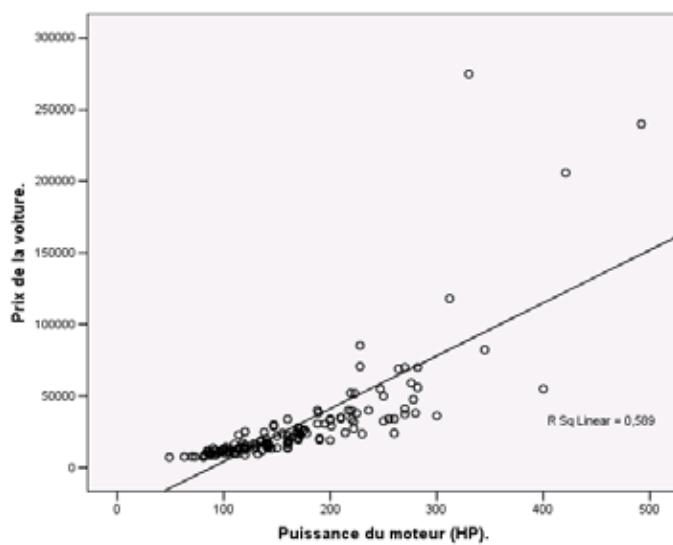


FIG. 11.24 – Le graphe de la relation

mesure que la puissance augmente. Voici les sorties de ce modèle.

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,768 ^a	,589	,587	22564,288

^a. Predictors: (Constant), puissance
^b. Dependent Variable: prix

FIG. 11.25 – Le r et le r^2 .

ANOVA ^b					
Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	1E+011	1	1,096E+011	215,224	,000 ^a
Residual	8E+010	150	509147091,6		
Total	2E+011	151			

a. Predictors: (Constant), puissance
b. Dependent Variable: prix

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-32776,3	4543,999		-7,213	,000
puissance	369,473	25,185	,768	14,671	,000

a. Dependent Variable: prix

FIG. 11.26 – La table ANOVA et des coefficients.

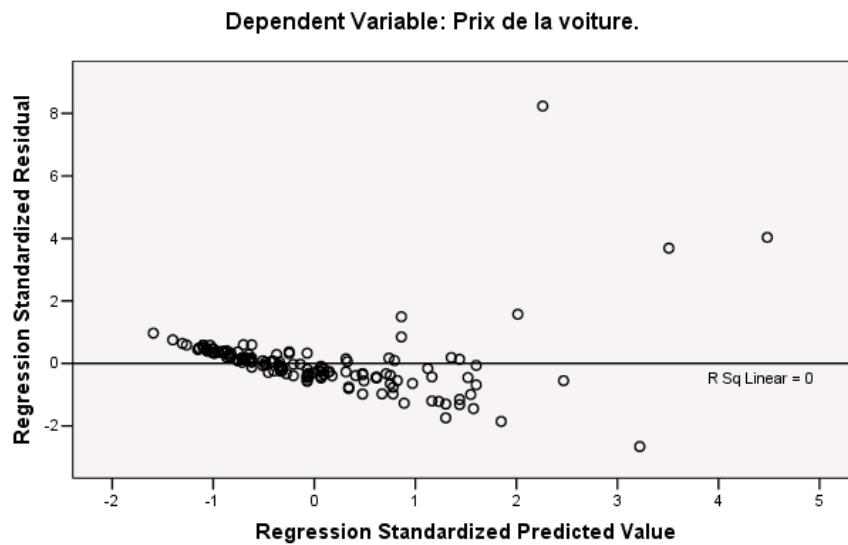


FIG. 11.27 – Les résidus.

Il est évident que ce modèle a avantage à être amélioré... En fait, lorsque les résidus ne sont pas constants tout autour de la droite de régression, il se trouve que les statistiques b_0, b_1, \dots, b_k ne sont plus des estimateurs à variance minimum ; pire encore, l'estimation de leur variance est biaisée.

Deux conséquences sont immédiates :

- Il existe de meilleurs estimateurs que b_0, b_1, \dots, b_k pour estimer $\beta_0, \beta_1, \dots, \beta_k$. Donc, il existe une meilleure droite de régression que celle connue.
- Compte tenu que l'estimation de la variance des β_i est biaisée, alors le calcul suivant donne de mauvais résultats (et donc de mauvaises conclusions) :

$$t = \frac{b_i}{s(b_i)}.$$

11.5.2 Correction de la violation à la variance constante

Il existe plusieurs façons de corriger la non constance de la variance. Cependant, toutes les procédures nécessitent la transformation de la variable dépendante Y . Cette transformation rendra la comparaison et l'interprétation des modèles légèrement plus difficile.

Voici deux transformations possibles :

- La transformation logarithmique ($\ln(Y)$) réduit la variabilité initiale des Y , ce qui stabilise la variation résiduelle. Cette transformation est à utiliser lorsque $CV_Y \sim 1$.
- On peut aussi transformer Y en prenant sa racine carré : \sqrt{Y} .

Ces deux transformations ne sont définies que si les valeurs de la variable dépendante sont positives. Mais ce problème est facilement résolu en utilisant une translation (translation qu'il faut enlever lors de l'interprétation finale des données).

L'emploi d'une de ces transformations peut aussi aider à rendre plus linéaire une relation qui semblait curvilinéaire. D'un autre côté, si la relation entre X et Y est déjà linéaire mais ne respecte pas l'hypothèse de variance constante des résidus, la transformation de Y peut briser cette linéarité, ce qui entraînera la nécessité de transformer X

aussi.

Ces deux transformations ne règlent pas tous les problèmes. Dans les faits, elles minimisent simplement l'ampleur du problème ; mince consolation.

Exemple 11.5.2 Poursuivons l'exemple 11.5.1. Essayons la transformation logarithmique pour modifier le modèle :

$$\ln(Y_{\text{prix}}) = \beta_0 + \beta_1 X_{\text{puissance}} + \epsilon.$$

Voici les sorties obtenues :

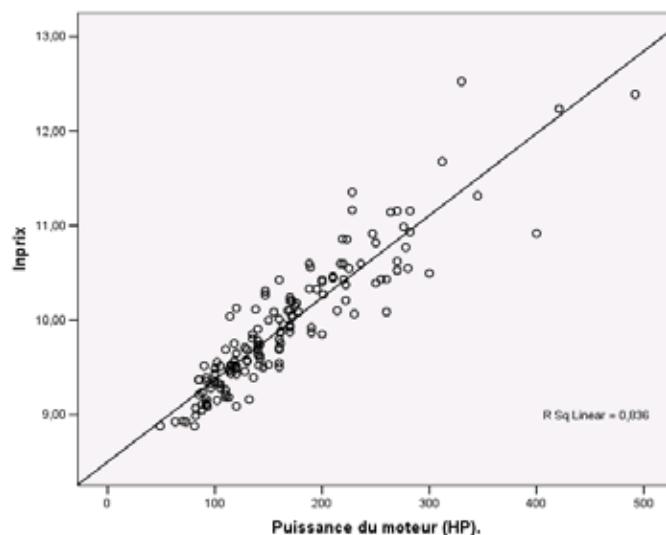


FIG. 11.28 – Le graphe de la relation

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,914 ^a	,836	,835	,28162

a. Predictors: (Constant), puissance

b. Dependent Variable: Inprix

FIG. 11.29 – Le r et le r^2 .

ANOVA ^b					
Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	60,675	1	60,675	765,008	,000 ^a
Residual	11,897	150	,079		
Total	72,572	151			

a. Predictors: (Constant), puissance

b. Dependent Variable: Inprix

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	8,499	,057		149,860	,000
puissance	,009	,000	,914	27,659	,000

a. Dependent Variable: Inprix

FIG. 11.30 – La table ANOVA et des coefficients.

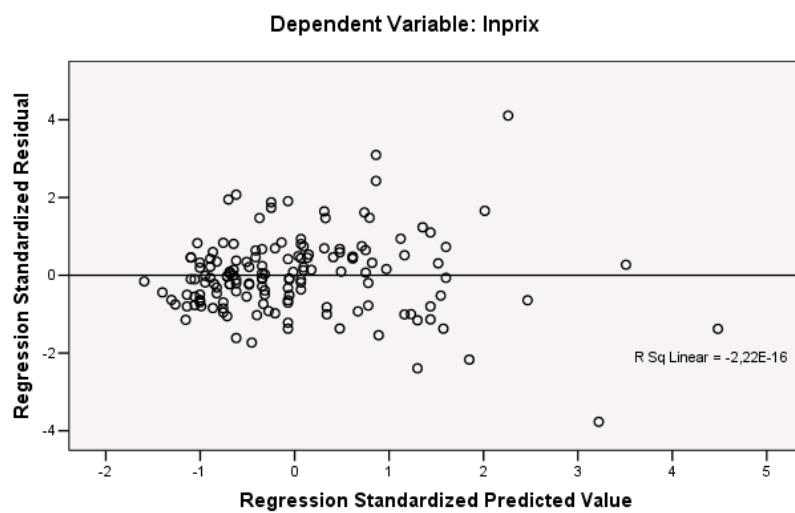


FIG. 11.31 – Les résidus.

On remarque que la forme de cône est toujours présente, mais de moindre ampleur. Notez que la transformation à l'aide de la racine carrée n'a pas donné de meilleurs résultats (essayez-la!). C'est donc le mieux que nous puissions faire avec ces transformations. Dans ce cas, nous disons que la variance est « relativement constante » autour de la droite.

Aussi, 83,6 % de la variabilité de $\ln(Y)$ a été expliquée par la droite de régression. Cette statistique n'est hélas pas comparable au modèle précédent (qui avait un r^2 de 0,589) puisque la nature même de Y a été changée. Ainsi, aucune conclusion ne peut être élaborée à savoir lequel des deux modèles est le plus performant. Gardons à l'idée que le deuxième modèle est « préférable » étant donné qu'il respecte mieux l'hypothèse de la variance constante.

Il faut aussi se rappeler que la droite de régression $\ln(Y_{\text{prix}}) = \beta_0 + \beta_1 X_{\text{puissance}} + \epsilon$ donnera des estimations pour $\ln(Y)$ et non pour Y .

Justement, quelle est l'estimation du prix d'une voiture de 200 HP ? Il faut d'abord utiliser l'équation trouvée :

$$\ln(y_{\text{prix}}) = 8,499 + 0,009x_{\text{puissance}} = 8,499 + 0,00869 \cdot 200 = 10,237.$$

Donc pour une voiture de 200 HP, l'estimation du logarithme du prix est de 10,237. Il faut donc appliquer la fonction inverse à $\ln(y_{\text{prix}})$ pour trouver \hat{y}_{prix} . La fonction inverse de \ln est la fonction exponentielle. Donc $\hat{y}_{\text{prix}} = e^{\ln(y_{\text{prix}})} = e^{10,237} = 27\,917,25 \$$.

11.6 Vérification de la normalité

Essentiellement, nous avons besoin de l'hypothèse de la normalité des résidus pour élaborer suivant les règles de l'art les intervalles de confiance et les intervalles de prédiction. Le graphique des résidus standardisés en fonction des valeurs prédictes peut être utilisé pour établir graphiquement que l'hypothèse de la normalité des résidus est respectée ou non.

Si les résidus standardisés suivent effectivement une loi normale, ils devraient sensiblement respecter les balises suivantes (comme vu précédemment) :

- 68,26 % des résidus se situeront entre ± 1 écarts-type ;
- 95,44 % des résidus se situeront entre ± 2 écarts-type ;
- 99,72 % des résidus se situeront entre ± 3 écarts-type.

11.6.1 Détection de la violation à la normalité

Nous illustrons comment détecter la violation à la normalité à l'aide de l'exemple qui suit.

Exemple 11.6.1 La base de données `salaireent.sav` contient des données à propos des employés d'une entreprise. Voici un aperçu de ces données.

id	sexé	datenaïs	educ	jobcat	salaire	saldebut	ancien	ancienprec
1	homme	02/03/52	15	Gestion	\$57,000	\$27,000	98	144
2	homme	05/23/58	16	Administrat	\$40,200	\$18,750	98	36
3	femme	*****	12	Administrat	\$21,450	\$12,000	98	381
4	femme	04/15/47	8	Administrat	\$21,900	\$13,200	98	190
5	homme	02/09/55	15	Administrat	\$45,000	\$21,000	98	138
6	homme	08/22/58	15	Administrat	\$32,100	\$13,500	98	67
7	homme	04/26/56	15	Administrat	\$36,000	\$18,750	98	114
8	femme	05/06/66	12	Administrat	\$21,900	\$9,750	98	0
9	femme	01/23/46	15	Administrat	\$27,900	\$12,750	98	115
10	femme	02/13/46	12	Administrat	\$24,000	\$13,500	98	244
11	femme	02/07/50	16	Administrat	\$30,300	\$16,500	98	143
12	homme	01/11/66	8	Administrat	\$28,350	\$12,000	98	26
13	homme	07/17/60	15	Administrat	\$27,750	\$14,250	98	34
14	femme	02/26/49	15	Administrat	\$35,100	\$16,800	98	137
15	homme	08/29/62	12	Administrat	\$27,300	\$13,500	97	66
16	homme	11/17/64	12	Administrat	\$40,800	\$15,000	97	24
17	homme	07/18/62	15	Administrat	\$46,000	\$14,250	97	48

FIG. 11.32 – Un aperçu des données

	Name	Type	Width	Decimals	Label	Values
1	id	Numeric	4	0	Code de l'employé.	None
2	sexe	Numeric	8	0	Sexe.	{0, femme}...
3	datenais	Date	8	0	Date de naissance.	None
4	educ	Numeric	2	0	Niveau d'éducation en années.	None
5	jobcat	Numeric	1	0	Type de poste.	{1, Administrat...
6	salaire	Dollar	8	0	Salaire.	None
7	saldebut	Dollar	8	0	Salaire à l'embauche.	None
8	ancien	Numeric	2	0	Nombre de mois depuis l'embauche.	None
9	ancienprec	Numeric	6	0	Expérience précédente en mois.	None

FIG. 11.33 – Le variable view

On s'intéresse à modéliser le salaire des employés à l'aide de l'équation suivante :

$$Y_{\text{salaire}} = \beta_0 + \beta_1 X_{\text{ancien}} + \beta_2 X_{\text{ancienprec}} + \beta_3 X_{\text{saldebut}} + \beta_4 X_{\text{educ}} + \beta_5 X_{\text{sexe}} + \epsilon.$$

Pour générer les sorties de cet exemple, les commandes sont les suivantes :

Menu SPSS :

→ Analyse

→ Regression

→ Linear...

Dans la fenêtre Dependant : → salaire (la variable dépendante)

Dans la fenêtre Independant(s) : → ancien, ancienprec, saldebut, educ, sexe

Dans le bouton Statistics... : ✓ Collinearity diagnostics
✓ Casewise diagnostics
✓ Outliers outside : 2 standard deviations

Dans le bouton Plots... : → Y : ZRESID (les résidus standardisés)
→ X : ZPRED (les prédictions standardisées)
✓ Histogram
✓ Normal probability plot

Dans le bouton Save... : → Residuals
✓ Standardized

Voici donc les sorties :

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,902 ^a	,814	,812	\$7,410.457

a. Predictors: (Constant), sexe, ancien, ancienprec, saldebut, educ

b. Dependent Variable: salaire

ANOVA^b

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	112216333930	5	2,244E+010	408,692	,000 ^a
Residual	25700161507	468	54914875,01		
Total	137916495436	473			

a. Predictors: (Constant), sexe, ancien, ancienprec, saldebut, educ

b. Dependent Variable: salaire

Coefficients^b

Model	Unstandardized Coefficients			t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	-14782,9	3267,788		-4,524	,000		
ancien	154,536	34,085	,091	4,534	,000	,987	1,013
ancienprec	-19,436	3,583	-,119	-5,424	,000	,827	1,210
saldebut	1,723	,061	,794	28,472	,000	,512	1,953
educ	593,031	166,630	,100	3,559	,000	,502	1,990
sexe	2232,917	792,078	,065	2,819	,005	,744	1,343

a. Dependent Variable: salaire

FIG. 11.34 – Le r_{aj}^2 , la table ANOVA et des coefficients

À l'aide du « Casewise Diagnostic » (figure 11.35), il est possible de voir que 21 employés sur un total de 474 (un peu moins de 5 %) ont des résidus au-delà de 2 écarts-types, ce qui est semblable à ce qu'on pourrait s'attendre de données issues d'une loi normale. Cependant, l'histogramme des résidus (figure 11.36) n'illustre pas la cloche de la normalité.

De plus, si les résidus se distribuaient réellement suivant une loi normale, les points s'aligneraient exactement sur la droite du graphique *Normal P-P Plot* (figure 11.36). La droite représente les valeurs hypothétiques des résidus si ceux-ci se distribuaient selon une loi normale. Mentionnons que le fait que les résidus s'alignent suivant une droite dans ce graphique n'a rien à voir avec l'étude de la linéarité de la relation entre Y et

Casewise Diagnostics ^a				
Case Number	Std. Residual	salaire	Predicted Value	Residual
18	6,179	\$103,750	\$57,961.66	\$45,788.344
32	2,952	\$110,625	\$88,745.85	\$21,879.154
35	2,368	\$81,250	\$63,700.56	\$17,549.444
53	2,289	\$73,750	\$56,785.51	\$16,964.488
100	2,650	\$78,250	\$58,615.85	\$19,634.150
103	3,404	\$97,000	\$71,773.26	\$25,226.738
106	3,794	\$91,250	\$63,138.14	\$28,111.862
160	-3,121	\$66,000	\$89,128.03	-\$23,128.030
205	-3,839	\$66,750	\$95,196.45	-\$28,446.450
217	-2,686	\$34,620	\$54,521.62	-\$19,901.625
218	6,006	\$80,000	\$35,490.38	\$44,509.621
240	2,216	\$54,375	\$37,950.80	\$16,424.199
274	5,042	\$83,750	\$46,384.08	\$37,365.918
290	-2,465	\$51,450	\$69,715.95	-\$18,265.953
371	2,922	\$58,125	\$36,473.88	\$21,651.117
383	2,961	\$78,500	\$56,559.54	\$21,940.462
446	2,615	\$100,000	\$80,624.67	\$19,375.327
449	3,497	\$70,000	\$44,084.52	\$25,915.478
454	3,829	\$90,625	\$62,249.47	\$28,375.529
464	-2,346	\$47,550	\$64,934.88	-\$17,384.876
468	2,352	\$55,750	\$38,317.33	\$17,432.668

a. Dependent Variable: salaire

FIG. 11.35 – Les données qui se distinguent des autres

les variables explicatives. Ainsi, si les données s’alignent visuellement suivant une ligne droite, l’hypothèse de la normalité semble plausible. Sinon, le graphique montrera des écarts.

Les données inscrites dans le tableau *Casewise Diagnostics* ne sont pas des données aberrantes et il ne faut surtout pas les effacer de l’étude instantanément. Cela veut plutôt dire que, pour certaines raisons inconnues, ces données sont légèrement différentes des autres et qu’elles méritent une attention plus particulière.

Il est possible d’effectuer un test d’hypothèses afin de vérifier si effectivement les données se distribuent suivant une loi normale. Pour ce faire, il faut avoir sauvégardé les résidus standardisés.

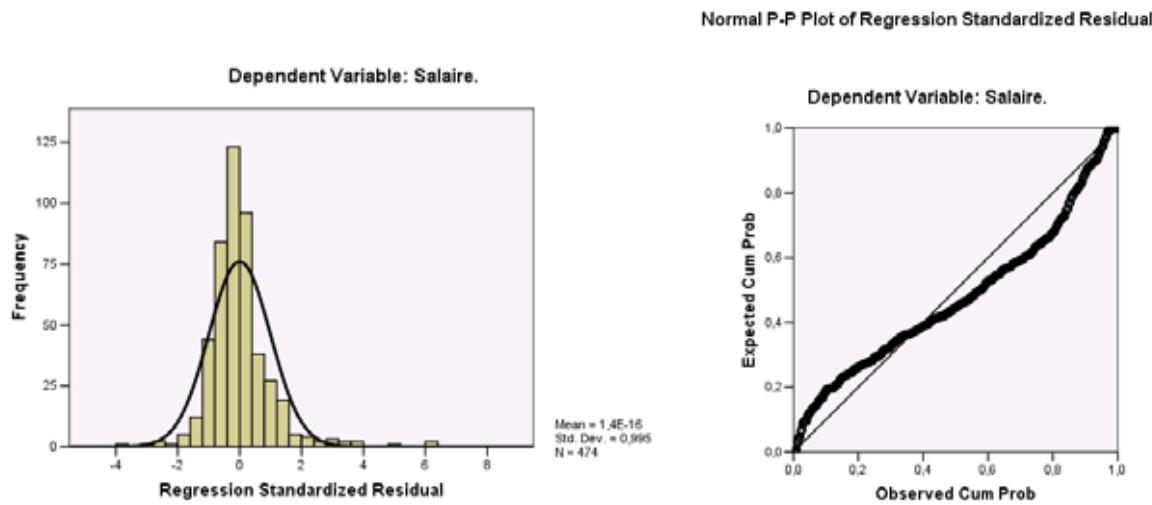


FIG. 11.36 – L'histogramme des résidus et le « Normal P-P Plot » des résidus

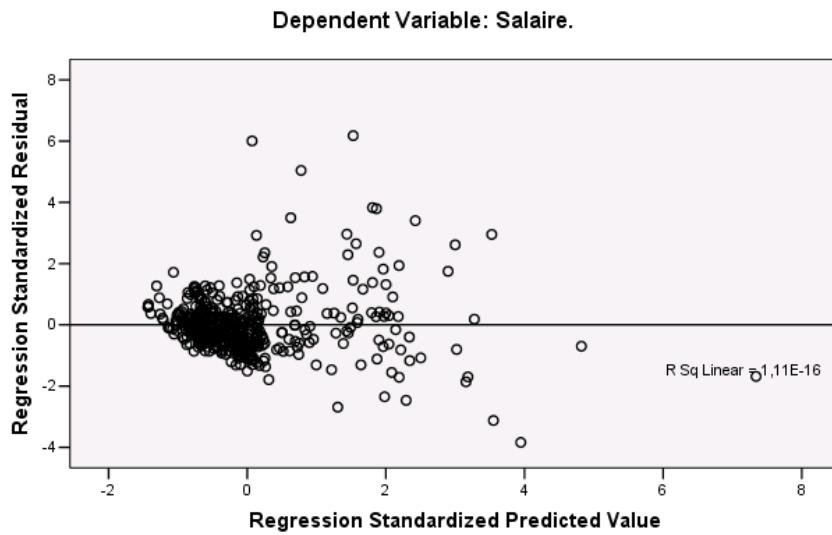


FIG. 11.37 – Le graphe des résidus

11.6.2 Test d'hypothèses sur la normalité des résidus

Si les résidus se distribuent suivant une loi normale, alors ils devraient s'aligner suivant une droite dans le graphique *Normal P-P Plot*. La qualité de l'alignement peut être mesurée à l'aide d'un test d'hypothèses :

H_0 : Les résidus se distribuent selon une loi normale au niveau de la population.

H_1 : Les résidus ne se distribuent pas selon une loi normale au niveau de la population.

Pour résoudre ce test, on prend les *p*-values des tests de Kolmogorov-Smirnov et Shapiro-Wilk :

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ZRE_1	,127	474	,000	,866	474	,000

a. Lilliefors Significance Correction

FIG. 11.38 – Le test de normalité

On rejette H_0 lorsque les *p*-values sont plus petites que le seuil de signification. Fixons ici le seuil à $\alpha = 0,05$. Puisque les *p*-values sont plus petites que 0,05 (elles sont nulles), on rejette H_0 avec une chance sur 20 de se tromper. Ainsi les résidus ne suivent pas une loi normale, et nous ne pourrons pas avoir confiance dans les intervalles de confiance et de prédiction produites par le modèle.

11.6.3 Correction à la normalité

La non normalité des résidus provient bien souvent du manque de données dans l'échantillon. Une trentaine de données avec une douzaine de données supplémentaires pour chaque variable explicative X donne une bonne idée de la taille minimale de l'échantillon requise.

Avant de corriger la normalité, il faut avoir fait les étapes suivantes :

- Être sûr d'avoir choisi le bon modèle de régression.
- Avoir corrigé les violations à la linéarité.
- Avoir corrigé les violations à la constance de la variance des résidus.

Assez souvent la correction des problèmes précédents corrige la normalité du même coup. Si le problème de normalité persiste, trois possibilités s'offrent :

- Augmenter la taille de l'échantillon. Cette tâche est souvent impossible.
- Effectuer l'étude approfondie des données de la table « Casewise diagnostics ». Cette étude fera l'objet de la prochaine section.
- Effectuer une transformation de la variable dépendante à l'aide d'un logarithme : $\ln(Y)$. Il existe d'autres types de transformation qui corrige la normalité. Elles sont connues sous le nom de *Box-Cox transformations*, mais elle ne seront pas vues dans ce cours.

11.7 Les données qui ont beaucoup d'influence

Considérons les deux définitions suivantes :

Outlier Un *outlier* est une donnée dont la valeur en « Y » est très différente des autres observations de l'échantillon.

Leverage Un *leverage* point est une donnée dont sa valeur en « X » est placée de telle façon qu'elle a un potentiel d'influencer la droite de régression.

On remarque qu'un *outlier* est étiqueté de par son écart marqué aux autres valeurs de Y , tandis qu'un *leverage* point est étiqueté de par son écart marqué aux autres valeurs de X . Ainsi, il est possible qu'une valeur soit à la fois un *outlier* et un *leverage* point. Surtout, il ne faut pas supprimer ces valeurs sans les avoir bien étudiées.

Pour illustrer ces deux concepts, voici un exemple théorique.

Exemple 11.7.1 Considérons les données suivantes et leur graphe :

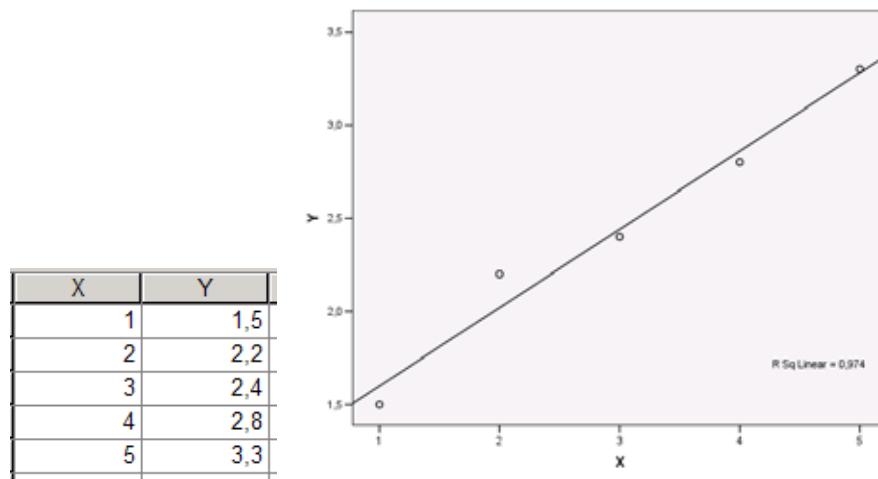


FIG. 11.39 – Une relation sans *outlier* ni *leverage*

Afin de comprendre l'influence sur la droite d'un point *outlier*, changeons la valeur en Y du point qui a 3 pour valeur en X ; faisons-la passer de 2,4 à 5 :

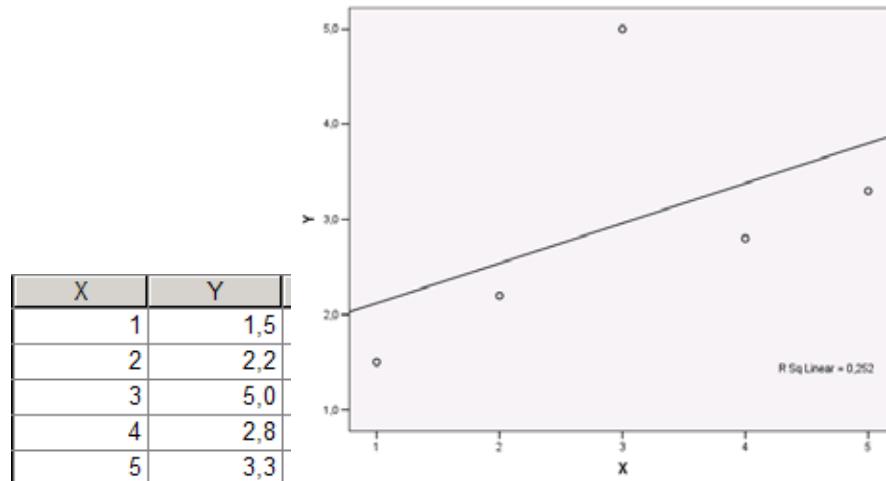


FIG. 11.40 – Une relation avec un *outlier* mais sans *leverage*

Par rapport au premier graphique, on remarque que :

- La droite de régression s'est déplacée dans la même direction que le point en question. Ceci est dû à la méthode des moindres carrés qui minimise les écarts à la verticale entre les points et la droite.
- Le coefficient r^2 est passé de 0.974 à 0.252, ce qui démontre une dégradation de la performance.

Les points *outlier* peuvent être vus de deux façons :

- Ils peuvent amener de l'information inattendue sur une sous-population de l'étude. Par exemple, si tous les *outlier* appartiennent à des hommes de plus de 50 ans, on prêtera une attention plus particulière à ces gens.
- Une telle valeur peut aussi semer la confusion et masquer de l'information importante qui aurait pu être obtenue à l'aide d'une régression sans ce point. Par exemple, une valeur a été mal entrée et se distingue beaucoup des autres, ce qui amène l'analyste à douter de son modèle linéaire.

Voyons maintenant quelle est l'influence d'un point de type *leverage*. Ajoutons un tel point et observons son effet sur la droite de régression :

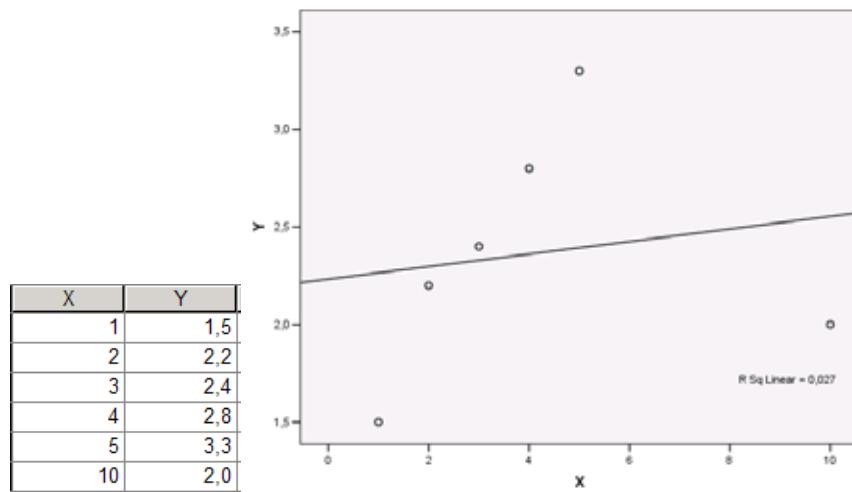


FIG. 11.41 – Une relation sans *outlier* mais avec un *leverage*

Par rapport au premier graphique, on remarque que :

- La droite de régression s'est déplacée dans la même direction que le point en question. Ceci est dû encore une fois à la méthode des moindres carrés qui minimise les écarts à la verticale entre les points et la droite.
- La pente a changée de façon drastique (ceci est typique des points de type *leverage*).
- Le coefficient r^2 est passé de 0.974 à 0.027, ce qui démontre une très grande dégradation de la performance (il n'y a pas beaucoup de points, l'influence se fait sentir plus facilement).

Les points *leverage* peuvent avoir une forte influence sur la pente de la droite de régression. Les points de type *leverage* peuvent eux aussi être vus de deux façons : ils peuvent amener de l'information inattendue sur une sous-population de l'étude, ou semer la confusion et masquer de l'information importante qui aurait pu être obtenue à l'aide d'une régression sans ce point.

Voyons maintenant l'influence d'un point à la fois de type *leverage* et de type *outlier*. Ajoutons un tel point et observons son effet sur la droite de régression :

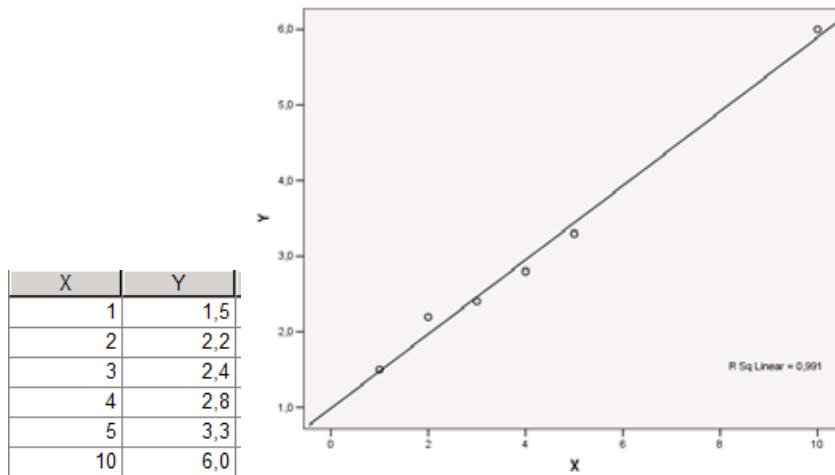


FIG. 11.42 – Une relation avec un point à la fois de type *leverage* et de type *outlier*

Ici, nous sommes en présence d'un point étant à la fois *outlier* et *leverage*. En effet,

autant la valeur en Y que la valeur en X de ce point se distinguent des autres valeurs. Pourtant, tout semble parfait. Mais il peut il y avoir anguille sous roche... Ainsi, ce point doit être étudié avec autant de soin que les précédents.

11.7.1 Méthodes d'identification des points de types *outlier* et/ou

leverage

Les points *outlier* sont reliés aux résidus puisque les résidus mesure l'erreur faite par rapport à la variable Y . Si les résidus proviennent effectivement d'une loi normale, alors les résidus standardisés ayant une valeur plus grande que 2 ne devraient constituer qu'au plus 5 % des résidus.

Les points de type *leverage* sont des points pouvant avoir beaucoup d'influence sur la position de la droite. Mais un point peut avoir un caractère *leverage* très grand sans pourtant affecter la position de la droite de régression. La statistique h_i mesure le *leverage* de la donnée i . Elle est offerte dans la plupart des logiciels statistiques.

Ces méthodes identifient de façon indépendante si un point est de type *outlier* ou de type *leverage*. Pourtant l'influence sur la droite de régression est déterminée par la combinaison de ces deux effets.

Ainsi quelques statistiques ont été développées pour combiner les indications données par les résidus (la mesure associée aux *outlier*) et les h_i (la mesure associée aux *leverage*). Il s'agit de la statistique $DFFit_i$ et de la statistique Cook's D. Ces deux mesures se basent sur l'omission d'une valeur pour mesurer son influence (le modèle sans cette valeur versus le modèle avec cette valeur, c'est un peu ce qu'on a fait en comparant les modèles avec et sans valeurs influentes). Il est possible de faire calculer ces mesures par SPSS lors d'une régression.

L'objectif de ces deux statistiques est de permettre l'identification des données qui ont de l'influence sur la droite de régression. Tout comme les autres statistiques présentées

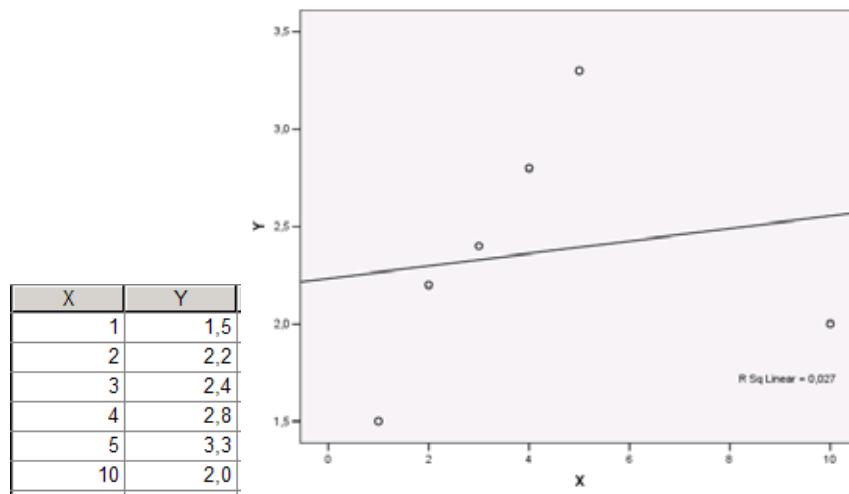
auparavant, l'identification d'une valeur ne signifie pas qu'elle soit mauvaise ; cela signifie qu'il faut étudier avec soin cette valeur.

La $DFFIT_i$ mesure l'influence des données de niveau i sur l'estimation \hat{y}_i . Pour la calculer, on considère deux modèles : celui calculé à partir de toutes les données, et celui calculé à partir de toutes les données sauf celles du niveau i . Ensuite $DFFIT_i$ mesure le nombre d'écart-types qui séparent la valeur estimée \hat{y}_i par le modèle conçu à partir de toutes les données et la valeur estimée $\hat{y}_{i(i)}$ par le modèle conçu à partir de toutes les données sauf celle du niveau i . Ainsi si les données de niveau i influencent beaucoup le modèle, il devrait il y avoir une bonne différence entre ces deux estimations.

La distance de Cook's, elle, mesure l'influence des valeurs du niveau i sur **toutes** les valeurs estimées. Ainsi, pour tout niveau j , Cook's mesure la distance entre la valeur estimée \hat{y}_j par le modèle conçu à partir de toutes les données et la valeur estimée $\hat{y}_{j(i)}$ par le modèle conçu à partir de toutes les données sauf celle du niveau i . Ensuite ces distances sont élevées au carré et additionnées, puis divisées par l'écart-type leur correspondant pour ainsi avoir une mesure standardisée (comme pour la $DFFIT_i$). On mesure ainsi l'effet global des valeurs en i sur le modèle.

Pour utiliser et visualiser les statistiques $DFFIT_i$ et Cook's D, il est recommandé d'utiliser un graphe séquentiel avec l'index en abscisse (l'axe des X). Remarquez que cette stratégie sert simplement à visualiser les statistiques précédentes qui n'ont rien à voir avec des séries chronologiques. L'index représente alors le numéro des observations, ce qui facilite le repérage des valeurs à haute influence.

Exemple 11.7.2 Reprenons l'une des relations précédentes. La donnée (10, 2) avait une grande influence sur la droite de régression.

FIG. 11.43 – Une relation sans *outlier* mais avec un *leverage*

Pour obtenir les mesures DfFit et Cook's, il faut ajouter les commandes suivantes aux commandes de la régression :

Dans le bouton Save... : → Distances
 Cook's
→ Influence Statistics
 DfFit

On obtient alors les mesures Cook's et DfFit dans la base de données (figure 11.44).

	X	Y	COO_1	DFF_1
1	1	1,5	.54801	-.43766
2	2	2,2	.00463	-.03392
3	3	2,4	.00155	.01699
4	4	2,8	.04831	.08808
5	5	3,3	.22967	.19944
6	10	2,0	9,94511	-2,82590

FIG. 11.44 – Les mesures Cook's et DfFit

Comme suggéré précédemment, considérons les graphes séquentiels pour ces deux mesures (le but est de facilement repérer les valeurs à haute influence, ce qui serait vraiment utile dans un contexte où il y aurait beaucoup d'observations).

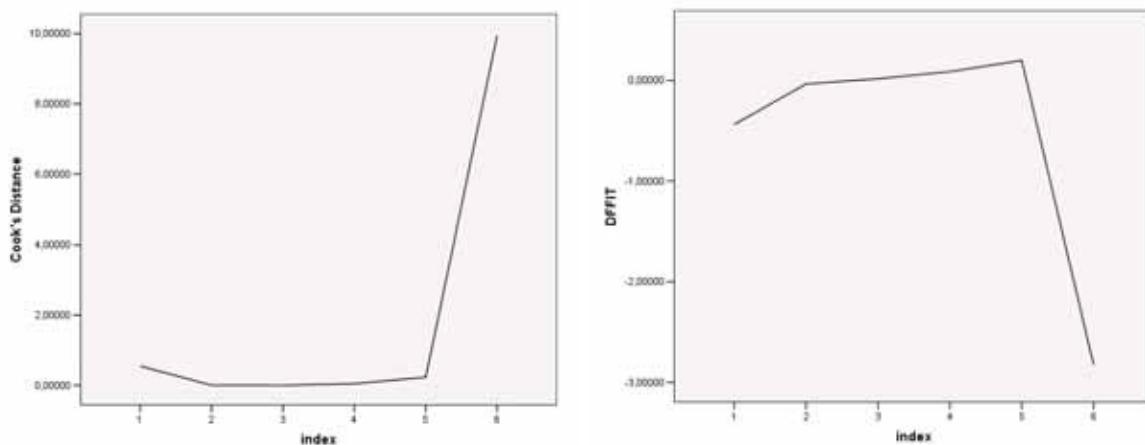


FIG. 11.45 – Les graphes séquentiels pour Cook's et DFFIT

Les deux graphiques illustrent que la sixième donnée possède une grande influence. N'oubliez pas que pour utiliser ces statistiques, il faut bien entendu les avoir créées lors de l'élaboration de la régression.

11.7.2 Que faire avec ces données ?

Comme mentionné auparavant, le fait qu'une donnée soit différente des autres ne signifie pas qu'elle doive être supprimée de la base de données. Ceci est vrai peu importe la méthode utilisée pour mettre en évidence les données aberrantes.

Il peut y avoir plusieurs raisons qui font qu'un point soit différent :

- Les violations à la linéarité et/ou à la constance de la variance peuvent être à l'origine du problème. Un bonne transformation des données pourrait corriger le problème.
- La donnée peut avoir été mal entrée. Il suffit alors de trouver la bonne donnée et de la remplacer. Si la donnée exacte n'est pas disponible, supprimer cette donnée qui est en fait une erreur.
- Une sous-population existe et ne fait pas tout à fait partie de l'étude.

Si aucune violation persiste et que la donnée est bien colligée, le problème est plus corsé. L'option de supprimer des données peut parfois être appropriée.

Par exemple, supposons qu'on s'intéresse à l'élaboration d'un modèle de régression pour prédire le prix des maisons ayant un terrain entre 1 500 et 2 500 pieds carrés. Si dans l'échantillon il y a une maison avec 4 500 pieds carrés de terrain, elle peut être rayée de l'étude.

Aussi, en ingénierie, il est courant que les phénomènes de chauffage ou de climatisation prennent un certain temps avant d'atteindre le régime permanent (par exemple une voiture ne chauffe pas instantanément l'hiver, il faut lui laisser un peu de temps). Si une mesure de la performance est désirée, les données prises avant d'atteindre le régime permanent peuvent être enlevées de l'étude.

11.8 Vérification de l'indépendance

La vérification de l'hypothèse de l'indépendance des résidus ne possède de l'intérêt que dans le cadre des séries temporelles. Dans ce cas, il est courant qu'un événement dans une période donnée ait une influence sur l'événement d'une période subséquente. Il est possible de vérifier si un tel phénomène est présent en étudiant le lien entre les résidus successifs.

On parle alors d'**autocorrélation**.

11.8.1 Autocorrélation

L'influence d'une période sur la suivante se modélise ainsi :

$$\epsilon_t = \rho\epsilon_{t-1} + \nu_t$$

où

- ϵ_t = l'erreur (le résidu) au temps t ;
- ρ = le coefficient d'autocorrélation (ce coefficient mesure la corrélation entre les erreurs qui ne sont séparées que d'une période) ;
- ν_t = le résidu au temps t (qui respecte les hypothèses de validité, dont celle d'indépendance).

Ainsi, le modèle linéaire simple s'écrit

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \text{ où } \epsilon_t = \rho\epsilon_{t-1} + \nu_t.$$

Lors de l'étude d'une série temporelle, lorsque l'hypothèse d'indépendance est violée, les estimations des β_i seront sans biais, mais les estimations de leurs écart-types seront biaisées. Ainsi, les intervalles de confiance, les intervalles de prédictions et les tests d'hypothèses ne seront pas aussi fiables que voulu.

Le coefficient d'autocorrélation ρ représente la force du lien entre les résidus successifs. Comme tout coefficient de corrélation, il varie entre -1 et 1. Une valeur nulle pour ρ est

espérée puisqu'elle démontre une indépendance linéaire entre les résidus (et donc que l'hypothèse d'indépendance des résidus est respectée).

11.8.2 Détection de l'autocorrélation de premier ordre

Reprendons le modèle

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \text{ où } \epsilon_t = \rho \epsilon_{t-1} + \nu_t.$$

Le test de Durbin-Watson est un test bien connu qui confronte les hypothèses suivantes :

$$H_0 : \rho = 0 \text{ (absence d'autocorrélation)}$$

$$H_1 : \rho > 0 \text{ (autocorrélation positive)}$$

Le test de Durbin-Watson est non-paramétrique et unilatéral. Aucun test d'hypothèses n'existe pour le cas $\rho < 0$.

La statistique de Durbin-Watson s'écrit de la façon suivante :

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

où

- $e_t = Y_t - \hat{y}_t$ est le résidu au temps t ;
- $e_{t-1} = Y_{t-1} - \hat{y}_{t-1}$ est le résidu au temps $t - 1$.

Lorsque les résidus sont indépendants, la valeur de DW se situe autour de 2. Lorsque les résidus sont en autocorrélation positive, la valeur de DW aura tendance à être plus petite que 2.

La procédure pour résoudre le test d'hypothèses est différente de celles vues jusqu'à maintenant, et il ne sera pas toujours possible de conclure s'il y a autocorrélation ou non.

Les étapes pour résoudre le test sont les suivantes :

- Identifier les bornes d_L et d_U du test de Durbin-Watson (table statistique C-6 du livre) selon la taille de l'échantillon, le seuil de signification et le nombre de variables indépendantes dans le modèle ;
- Comparer la valeur de DW (qui peut être calculée par SPSS) à ces bornes :
 1. Si $DW > d_U$, on conclut que $H_0 : \rho = 0$ est vraie ;
 2. Si $DW < d_L$, on conclut que $H_1 : \rho > 0$ est vraie ;
 3. Si DW est entre les deux bornes ($d_L \leq DW \leq d_U$), le test n'est pas concluant.

Exemple 11.8.1 Les données suivantes représentent les profits corporatifs et le produit intérieur brut (en billions de dollars) d'un pays pour les années 1960 à 1991. La base de données se nomme **pib.sav**.

	année	corprof	pib
1	1960	28,40	516,60
2	1961	28,20	535,40
3	1962	32,40	575,80
4	1963	34,90	607,70
5	1964	40,00	653,00
6	1965	47,90	708,10
7	1966	51,40	774,90
8	1967	49,20	819,80
9	1968	51,20	895,50
10	1969	49,40	965,60
11	1970	44,00	1017,10
12	1971	52,40	1104,90
13	1972	62,60	1215,70
14	1973	81,60	1362,30
15	1974	91,00	1474,30
16	1975	89,50	1599,10
17	1976	109,50	1785,50
18	1977	130,30	1994,60
19	1978	154,40	2254,50
20	1979	173,40	2520,80
21	1980	156,10	2742,10
22	1981	147,00	2662,80

FIG. 11.46 – Extrait de la base de données.

On s'intéresse au modèle suivant :

$$Y_{\text{corprof}} = \beta_0 + \beta_1 X_{\text{pib}} + \epsilon.$$

Les commandes pour obtenir les sorties sont les mêmes qu'à l'habitude (pour une régression linéaire), il suffit de cocher **Durbin-Watson** dans le bouton **Statistics** :

- Menu SPSS :
- Analyse
 - Regression
 - Linear...
- Dans la fenêtre Dependant : → salaire (la variable dépendante)
- Dans la fenêtre Independant(s) : → ancien, ancienprec, saldebut, educ, sexe
- Dans le bouton Statistics... : ✓ Collinearity diagnostics
✓ Durbin-Watson
✓ Casewise diagnostics
✓ Outliers outside : 2 standard deviations
- Dans le bouton Plots... : → Y : ZRESID (les résidus standardisés)
→ X : ZPRED (les prédictions standardisées)
✓ Histogram
✓ Normal probability plot
- Dans le bouton Save... : → Residuals
✓ Standardized

Model Summary ^b					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,907 ^a	,823	,817	24,99091	,478

a. Predictors: (Constant), pib

b. Dependent Variable: corprof

FIG. 11.47 – Le r , r^2 et le Durbin-Watson

La sortie 11.47 contient la statistique de Durbin-Watson. On voit que $DW = 0,478$. On doit utiliser cette valeur pour traiter le test

$$H_0 : \rho = 0 (\text{ absence d'autocorrélation})$$

$$H_1 : \rho > 0 (\text{ autocorrélation positive})$$

ANOVA ^b					
Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	87178,160	1	87178,160	139,587	,000 ^a
Residual	18736,364	30	624,545		
Total	105914,5	31			

a. Predictors: (Constant), pib
b. Dependent Variable: corprof

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	30,051	7,604		3,952	,000
pib	,032	,003	,907	11,815	,000

a. Dependent Variable: corprof

FIG. 11.48 – La table ANOVA et des coefficients

Fixons le seuil à $\alpha = 0,05$. On a $n = 32$, et $k = 1$ (le nombre de variables indépendantes). On trouve (table C-6) $d_L = 1,37$ et $d_U = 1,50$. La règle de décision est la suivante : on rejette H_0 si $DW < 1,37$, on accepte H_0 si $DW > 1,50$ et on ne peut rien dire si $1,37 < DW < 1,50$.

Puisque $0,478 < 1,37$, on rejette H_0 . Au seuil considéré, il est significatif de dire qu'il y a un problème d'autocorrélation de premier ordre. Ce problème doit donc être corrigé avant toute inférence.

Casewise Diagnostics ^a				
Case Number	Std. Residual	corprof	Predicted Value	Residual
19	2,117	154,40	101,4873	52,91273
20	2,540	173,40	109,9252	63,47477
27	-2,173	111,30	165,5942	-54,29420

a. Dependent Variable: corprof

FIG. 11.49 – Casewise Diagnostics

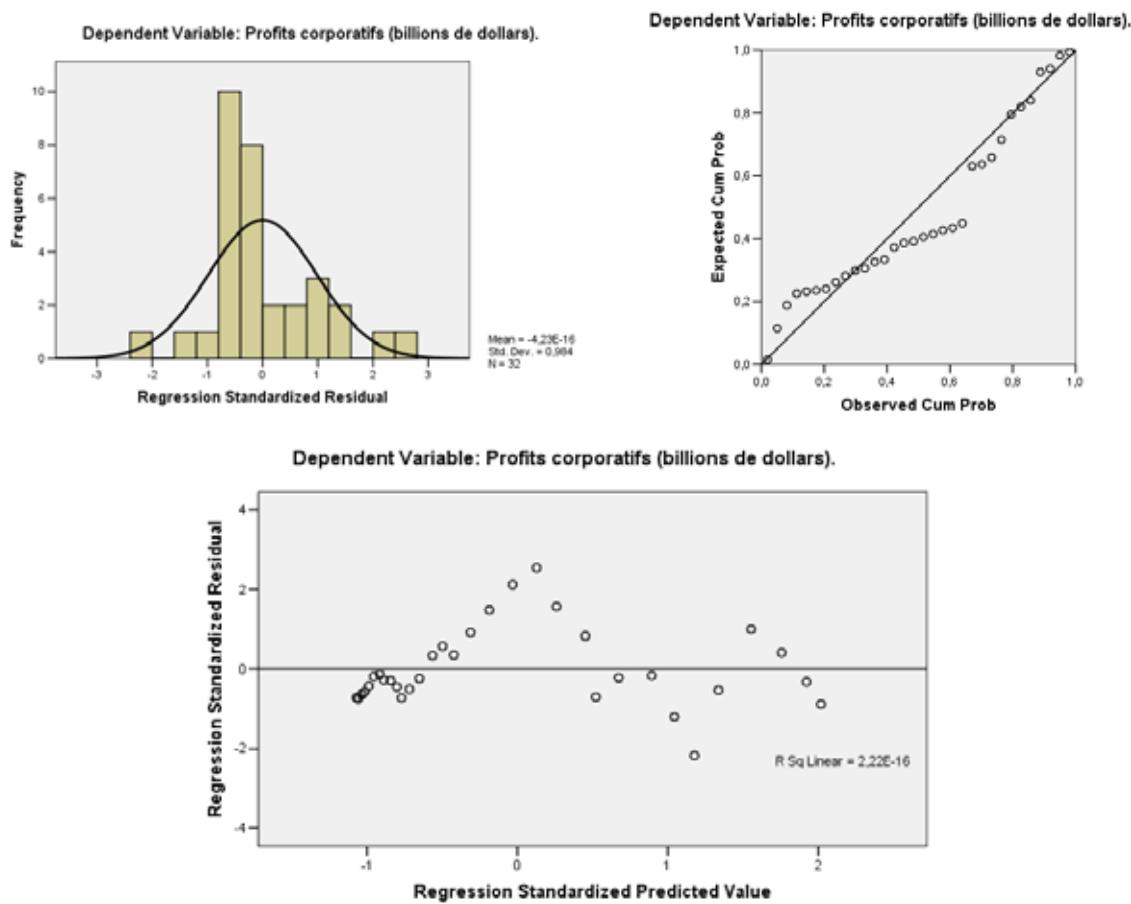


FIG. 11.50 – Graphes illustrant la distribution des résidus

11.8.3 Correction de l'autocorrélation de premier ordre

Plusieurs phénomènes peuvent expliquer l'autocorrélation. Entre autres, le phénomène fait surface lorsqu'une variable importante, positivement corrélée avec la variable dépendante, est manquante dans le modèle. Pour remédier à ce problème, il suffit de trouver la variable en question et de l'ajouter au modèle. Cette dernière solution est évidemment plus facile à dire qu'à faire.

Pour corriger l'autocorrélation, il est courant d'utiliser une transformation qui enlève les quantités autocorrélées. Considérons un modèle

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \text{ où } \epsilon_t = \rho \epsilon_{t-1} + \nu_t$$

dans lequel il y a de l'autocorrélation ($\rho > 0$). Pour enlever l'autocorrélation, il faut créer de nouvelles variables Y^* et X^* en deux étapes.

Tout d'abord, pour les périodes $t = 2$ jusqu'à n , on transforme X et Y de la façon suivante pour obtenir Y^* et X^* :

$$Y_t^* = Y_t - \rho Y_{t-1} \text{ et } X_t^* = X_t - \rho X_{t-1}.$$

Il reste ensuite à transformer les données pour le temps $t = 1$ de la façon suivante :

$$Y_1^* = \sqrt{1 - \rho^2} Y_1 \text{ et } X_1^* = \sqrt{1 - \rho^2} X_1.$$

Le nouveau modèle à étudier sera alors

$$Y_t^* = \beta_0 + \beta_1 X_t^* + \nu_t.$$

On obtient une estimation de β_0 et de β_1 en faisant une analyse en régression linéaire comme à l'habitude. Mais il ne faut pas oublier d'effectuer la transformation inverse avant l'interprétation des données.

Il existe aussi plusieurs types de procédures automatisées qui effectuent les transformations précédentes ou similaires. Toute transformation tenant compte de la première période dans ses calculs est recommandable. La littérature mentionne qu'il faut éviter

toute transformation qui omet d'utiliser la première période dans ses calculs. Il semble que cette stratégie entraîne une perte importante d'information, ce qui est toujours à éviter.

Il est aussi possible d'ajouter la variable dépendante déphasée aux variables indépendantes pour corriger l'autocorrélation.

Chapitre 12

Les séries temporelles

Plusieurs méthodes sont utilisées pour tenter de maîtriser l'influence du temps sur le déroulement des choses. Ce chapitre présente quelques outils de base permettant à l'analyste de s'ajuster avec souplesse aux différentes situations en modélisant bon nombre de « courbes ».

Suivant l'expression anglaise « *Nuts and Bolts* », les techniques présentées dans le cadre de cette section représentent un coffre à outil en tout genre pour le modéliste averti. Plus précisément, l'analyste peut utiliser ces techniques seules, ensemble ou de manière concomitante. La compréhension, l'intuition ainsi que la créativité de l'analyste prend ici une place des plus importantes. Tout au long de ce chapitre, l'analyste doit se rappeler qu'un bon modèle est surtout celui qui fait le travail que l'analyste désire. Plusieurs approches et modèles peuvent être utilisés et créés et des comparaisons peuvent ensuite être effectuées.

Dans le cadre de ce chapitre, plusieurs sujets seront abordés. L'approche de Box-Jenkins et les méthodes ARCH et GARCH seront présentées dans des chapitres subséquents. Les sujets sont simplement présentés suivant un ordre pédagogique.

12.1 Quelques généralités

Une série temporelle (*time serie*) est une suite de données observées et étudiées à intervalles temps égaux. Quatre types d'effets peuvent survenir dans une série : un effet de **tendance**, un effet **cyclique**, un effet **saisonnier** et un effet **irrégulier**.

- Un **effet de tendance** (*trend*) est une composante qui représente la croissance ou la décroissance à long terme d'une série. La croissance du Produit National Brut (PNB) est un exemple macroéconomique de tendance croissante à long terme.
- Un **effet cyclique** (*cyclical component*) représente un effet de vagues autour de la tendance. Cet effet représente l'impact de la santé de l'économie sur la tendance et revient à tous les deux, trois années ou plus. Les changements dans la mode ou encore la vie utile d'un produit en sont des exemples.
- Un **effet saisonnier** (*seasonal component*) est un tracé qui se répète d'année en année en fonction des saisons (température, vacances, fêtes, etc.). Par exemple, l'industrie de la construction est très influencée par l'impact des saisons.
- Un **effet irrégulier** (*irregular effect*) représente les fluctuations résiduelles du tracé de la série une fois que les effets de tendance, cyclique et saisonnier ont été enlevés. Dans des conditions idéales, l'irrégularité du tracé est de nature aléatoire. Cependant, des attaques terroristes, des grèves, des nouvelles lois peuvent causer et expliquer les effets aléatoires dans une série.

Finalement, une **série stationnaire** (*stationary serie*) est une suite de données temporelles n'ayant pas d'effet de tendance à long terme, ni à la hausse, ni à la baisse, et la série est de type homoscédastique. À long terme, une telle série est, en moyenne, égale à elle-même. Ce type de situation se produit par exemple lorsque la demande pour un produit est stable.

12.2 Le choix d'un modèle et taille d'échantillon

Comme il sera possible de le voir dans le cadre de ce chapitre, plusieurs techniques existent pour modéliser le comportement des données issues de séries chronologiques. Le choix de la technique dépend des besoins de l'analyste. Un bon modèle est un modèle qui fait le travail que l'analyste désire et le choix repose souvent sur le jugement du modéliste.

L'analyste qui développe le modèle n'est souvent pas celui qui l'utilisera. Ainsi, il est impératif de se poser la question à savoir quels sont les qualifications de l'utilisateur ; parfois la simplicité doit être de mise. On doit se demander quelles sont les caractéristiques des données en fonction du temps (graphique : effet de saison, de cycle, etc.) ? Aussi, existe-t-il beaucoup de données manquantes ? En effet, cette question est importante car la plupart des techniques présentées dans le cadre de ce chapitre ne supportent pas les données manquantes. Une section de ce chapitre sera dédiée à cet effet.

On doit également se demander quelle est la longueur des prévisions (*time frame* ou *time horizon*) à établir : immédiate (moins de 1 mois), court terme (de 1 à 3 mois), moyen terme (entre trois mois et deux ans) ou long terme (deux ans et plus) ? Les techniques ne se comportant pas toutes de la même façon, la longueur des prévisions voulues influence directement le choix de la technique. On doit aussi se demander quelle est la taille d'observations historiques minimale et quel est le coût d'acquisition des données. Désire-t-on des prédictions ponctuelles ou par intervalle de confiance ? Cette dernière question est importante puisque certaines techniques produisent des intervalles de confiance, d'autres pas.

Finalement, quelle est la précision désirée sur les prédictions ? Dans certains cas, 20 % de probabilité d'erreur peut être acceptable, dans d'autres, 2 % peut avoir un effet catastrophique. Tels sont les interrogations à résoudre avant de procéder à une modélisation.

	Type d'effet*	Horizon de prédiction**	Type de modèle***	Minimum de périodes sans effet de saison	Minimum de périodes avec effet de saison
Modèle naïf M_t^1	St, T, S	Imm	SC	1	
Moyenne simple M_t^t	St	Imm, C-T	SC	30	
Moyenne mobile M_t^n	St	Imm, C-T	SC	4-20	
Lissage exponentiel simple	St	Imm, C-T	SC	3	
Lissage exponentiel Holt	T	Imm, C-T	SC	3	
Lissage exponentiel quadratique	T	Imm, C-T	SC	4	
Lissage exponentiel Winter	T,S	Imm, C-T	SC		(2 à 4)*L (SPSS 4*L)
Régression linéaire simple	T	C-T, M-T	C	10	
Régression linéaire multiple	T,C,S	C-T, M-T, L-T	C	10*V	6*L
Décomposition	T,C,S	Imm, C-T	SC		(3 à 5)*L (SPSS 4*L)
Courbe en « S »	T	M-T, L-T	SC	10	
Courbe de croissance	T	M-T, L-T	SC	10	
Box-Jenkins	St,T,C,S	Imm, C-T	SC	24	(3 ou 4) *L (SPSS 4*L)
Delphi		L-T	QU		

* St = Stationnaire, T = Tendance, S = Saison, C = Cycle

** Imm = Immédiat, C-T = Court Terme, M-T = Moyen Terme, L-T = Long Terme

*** SC = Série Chronologique, C = Causal

12.3 Les techniques de moyennes

12.3.1 Les moyennes simples

Le modèle des moyennes simples a comme prémissse que les ventes éventuelles seront, en moyenne, les mêmes que par le passé. Plus précisément, ce type de modèle s'applique lorsque l'analyste est en présence de données stationnaires.

L'équation représentant ce modèle est

$$\hat{Y}_{t+1} = \frac{1}{t} \sum_{i=1}^t Y_i$$

où \hat{Y}_{t+1} représente la valeur de Y estimée au temps $t + 1$ en fonction de la moyenne de toutes les valeurs passées.

Exemple 12.3.1 Reprenons l'exemple de la compagnie ABX (la base de données se nomme **ABX.sav**). Au chapitre 8 nous avons développé un modèle très performant pour estimer les ventes de cette compagnie ; celui-ci tenait compte du temps (index) et des saisons.

Voyons maintenant comment un modèle basé sur la moyenne simple estime les ventes de la compagnie ABX. Prenons par exemple les 4 dernières données de la base de données :

index	ventes	annee
37	306,50	1994
38	283,50	1994
39	283,50	1994
40	307,50	1994

FIG. 12.1 – Les quatre dernières données

Avec le modèle de la moyenne simple, les estimations pour les trimestres de 1994 seront :

$$\hat{y}_{37} = \frac{1}{36} \sum_{t=1}^{36} Y_t = 246,54, \quad \hat{y}_{38} = \frac{1}{37} \sum_{t=1}^{37} Y_t = 248,16,$$

$$\hat{y}_{39} = \frac{1}{38} \sum_{t=1}^{38} Y_t = 249,09, \quad \hat{y}_{40} = \frac{1}{39} \sum_{t=1}^{39} Y_t = 249,97.$$

En comparant ces valeurs aux ventes réelles, on voit bien que le modèle des moyennes simples n'est pas bien adapté à ce contexte. Ceci vient du fait que la série n'est pas stationnaire. Le graphe suivant nous montre bien que le modèle de la moyenne simple sous-estime beaucoup les ventes réelles.

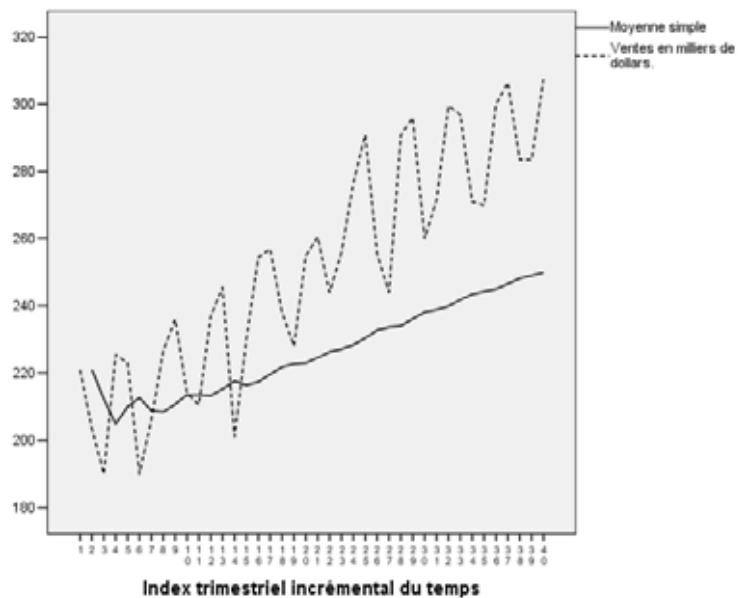


FIG. 12.2 – Les moyennes simples comparées aux ventes réelles

12.3.2 Les moyennes mobiles (*moving averages*)

Le modèle de la moyenne simple utilise l'ensemble des données historiques pour calculer la prédiction. Cependant, lorsque le marché est en mouvement, les données moins récentes fournissent rarement une bonne image du futur. Dans ces conditions, il peut être pertinent de n'utiliser que les valeurs des périodes les plus récentes. Ce modèle porte le nom de moyennes mobiles (*moving averages*). Plus précisément, de période en période, la valeur la plus récente remplace la valeur la plus ancienne et une nouvelle moyenne est recalculée à partir de ces dernières périodes.

Le modèle des moyennes mobiles d'ordre k a comme équation :

$$M_t^k = \hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + \cdots + Y_{t-k+1}}{k}$$

où

- k est le nombre de périodes considérées dans les calculs ;
- M_t^k est la valeur de la moyenne mobile au temps t issue des k valeurs les plus récentes ;
- \hat{Y}_{t+1} est la prédiction pour la période $t + 1$;
- Y_i est la valeur de la variable au temps i (pour i entre $(t - k + 1)$ et t).

La moyenne mobile d'ordre k est simplement la moyenne arithmétique des k données les plus récentes. Par construction, cette moyenne associe le même poids aux valeurs prises en considération. Plus l'ordre k est grand, moins la moyenne réagira rapidement aux mouvements de la série. Par exemple, pour des données mensuelles ou trimestrielles, une moyenne mobile d'ordre 12 ou 4 retournera, dans les deux cas, une moyenne annuelle qui annulera l'effet des saisons. Cependant, une moyenne mobile d'ordre k avec k petit réagira rapidement aux sursauts de la série. Ce type d'opération porte le nom de lissage (*smoothing*).

Exemple 12.3.2 Reprenons l'exemple de la compagnie ABX. Étudions les moyennes mobiles d'ordre 4 et d'ordre 8 pour les ventes trimestrielles. Pour créer ces moyennes

(elles formeront deux nouvelles variables dans la base de données), les commandes sont les suivantes :

Menu SPSS : → Transform
→ Create Time Series...

Dans la fenêtre New Variable(s) : → ventes

Dans la fenêtre Name : → m4_t

Dans la fenêtre Function : → sélectionnez Prior moving average

Dans la fenêtre Span : → 4 (l'ordre de la moyenne mobile)

Appuyez sur Change.

On répète pour la moyenne mobile d'ordre 8 :

Dans la fenêtre New Variable(s) : → ventes

Dans la fenêtre Name : → m8_t

Dans la fenêtre Function : → sélectionnez Prior moving average

Dans la fenêtre Span : → 8 (l'ordre de la moyenne mobile)

Appuyez sur Change.

Voici un aperçu du résultat dans la base de données :

	index	ventes	annee	saison	m4_t	m8_t
1	1	221,00	1985	hiver	-	-
2	2	203,50	1985	printemps	-	-
3	3	190,00	1985	été	-	-
4	4	225,50	1985	automne	-	-
5	5	223,00	1986	hiver	210,00	-
6	6	190,00	1986	printemps	210,50	-
7	7	206,00	1986	été	207,12	-
8	8	226,50	1986	automne	211,12	-
9	9	236,00	1987	hiver	211,37	210,69
10	10	214,00	1987	printemps	214,62	212,56

FIG. 12.3 – Les données résultantes

Lorsque la série de données présente une tendance positive, les moyennes mobiles au-

ront tendance à sous-estimer les valeurs réelles, et de façon croissante avec l'ordre (plus l'ordre de la moyenne est grand, plus elle aura tendance à sous-estimer les ventes réelles). Donc ce type de modèle est davantage utilisé lorsque les données sont de nature stationnaire. Par contre ce modèle réagit mieux aux tendances que le modèle de la moyenne simple.

La figure 12.4 illustre bien ceci.

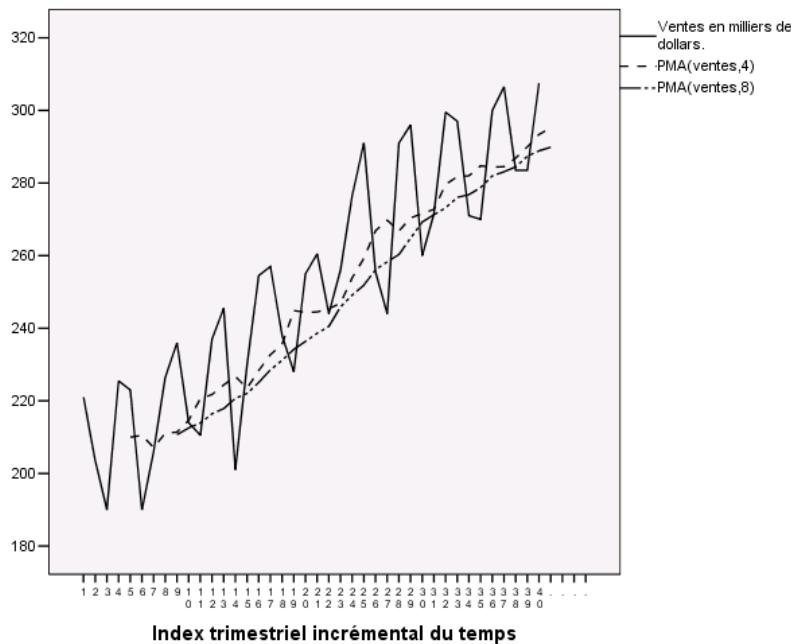


FIG. 12.4 – Le graphe avec les variables ventes, m4_t et m8_t

Le jugement de l'analyste ainsi que les objectifs à atteindre sont les seuls critères aidant à fixer l'ordre de la moyenne mobile. Plus l'ordre est élevé, plus le lissage est important. Le modèle $M_t^1 = \hat{Y}_{t+1} = Y_t$ existe et porte le nom de modèle naïf. Ce modèle utilise seulement la dernière valeur à titre d'estimation de la subséquente. En données trimestrielles, ce modèle peut être très mauvais, notamment en présence d'un effet de saison important.

Dans certaines circonstances, il peut être intéressant pour l'analyste de calculer une moyenne mobile double. Certaines techniques, notamment la décomposition, utilisent

cette stratégie. La moyenne double calcule d'abord une moyenne mobile d'ordre k et, de ces valeurs obtenues, une seconde moyenne mobile d'ordre l est alors calculée.

Le modèle de la double moyenne mobile d'ordre $k * l$ s'écrit

$$M_t^{k*l} = \hat{Y}_{t+1} = \frac{(M_t^k + M_{t-1}^k + \dots + M_{t-l+1}^k)}{l}$$

où

- k est le nombre de périodes considérées lors du calcul de la première moyenne mobile ;
- l est le nombre de périodes considérées lors du calcul de la seconde moyenne mobile ;
- M_t^k est la valeur de la moyenne mobile au temps t issue des k valeurs les plus récentes (relativement au temps t) ;
- M_t^{k*l} est la valeur de la double moyenne mobile au temps t ;
- \hat{Y}_{t+1} est la prédition pour la période $t + 1$.

En général, il est courant d'utiliser des doubles moyennes mobiles d'ordre $k * k$.

Exemple 12.3.3 Reprenons l'exemple de la compagnie ABX. Pour obtenir une double moyenne mobile d'ordre $4 * 4$ (notée `m4_4_t`) des ventes avec SPSS, il faut d'abord créer la première moyenne mobile `m4_t` comme précédemment et effectuer les commandes suivantes :

Menu SPSS :	→ Transform
	→ Create Time Series...
Dans la fenêtre New Variable(s) :	→ <code>m4_t</code>
Dans la fenêtre Name :	→ <code>m4_4_t</code>
Dans la fenêtre Function :	→ sélectionnez Prior moving average
Dans la fenêtre Span :	→ 4
Appuyez sur Change .	

index	ventes	annee	m4_t	m4_4_t
1	221,00	1985		
2	203,50	1985		
3	190,00	1985		
4	225,50	1985		
5	223,00	1986	210,00	
6	190,00	1986	210,50	
7	206,00	1986	207,12	
8	226,50	1986	211,12	
9	236,00	1987	211,37	209,69
10	214,00	1987	214,62	210,03
11	210,50	1987	220,62	211,06
12	237,00	1987	221,75	214,44
13	245,50	1988	224,37	217,09
14	201,00	1988	226,75	220,34
15	230,00	1988	223,50	223,37
16	254,50	1988	228,37	224,09
17	257,00	1989	232,75	225,75
18	238,00	1989	235,62	227,84
19	228,00	1989	244,87	230,06
20	255,00	1989	244,37	235,41
21	260,50	1990	244,50	239,41

FIG. 12.5 – Aperçu des données résultantes

La figure 12.5 nous donne un aperçu de la nouvelle variable `m4_4_t` qui est alors créée.

Examinons le graphe séquentiel pour comparer la moyenne mobile et la double moyenne mobile (figure 12.6).

On remarque que la double moyenne mobile est moins « brusque » dans ses changements. Le modèle de la double moyenne mobile simule un effet de type « linéaire », ce qui, dans certaines situations pourrait être utile pour remplacer l'équation d'une droite de régression. Cette stratégie de remplacement, contrairement à une droite de régression, n'a aucun présupposé à vérifier.

Comme souligné précédemment, ces techniques sont généralement efficaces lorsque les données sont stationnaires.

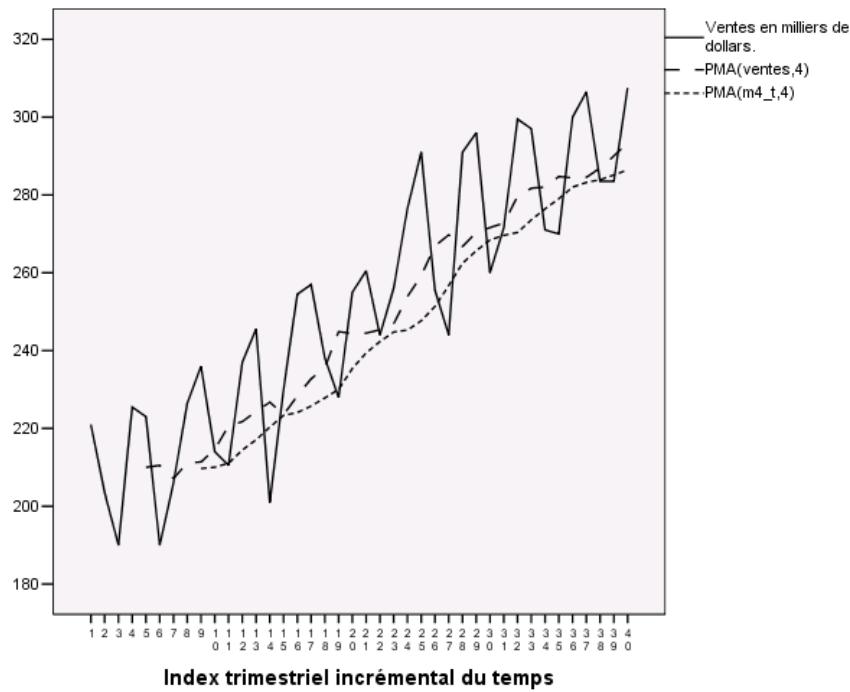


FIG. 12.6 – La moyenne mobile d'ordre 4 et la double moyenne mobile d'ordre 4^*4

Exemple 12.3.4 Voici un ensemble de données stationnaires (des ventes mensuelles) ainsi que les estimations obtenues à l'aide d'une moyenne mobile d'ordre 4 (figure 12.7). La base de données se nomme `exstationnaire.sav`.

Le graphe séquentiel nous permet de comparer les estimations avec les ventes réelles.

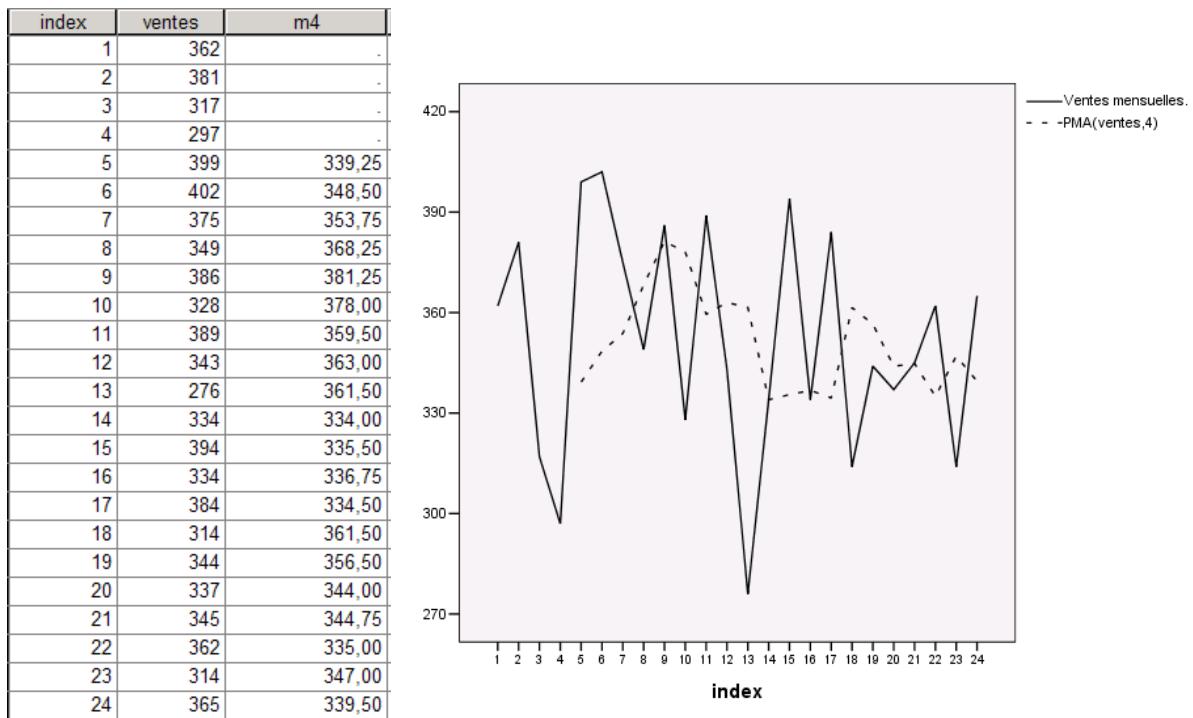


FIG. 12.7 – Estimations à l'aide d'une moyenne mobile d'ordre 4 (série stationnaire) et graphe séquentiel

Nous avons développé plusieurs types de modèles, parfois pour estimer une même variable dépendante. Il sera utile de pouvoir comparer l'efficacité de ces modèles. Bien entendu, l'efficacité est bien relative en ce sens qu'un modèle est efficace s'il effectue le travail que l'analyste désire. Une analyse de la performance du modèle est toujours de mise.

L'analyse d'efficacité sera présentée dans une section subséquente. Il est toujours conseillé de construire plus d'un modèle et de comparer les performances respectives.

12.3.3 Moyennes mobiles avec E-Views

Pour créer une série avec des moyennes mobiles, il suffit d'utiliser les expressions suivantes dans le haut de la fenêtre de E-Views :

- Pour des moyennes mobiles antérieures (*prior moving average*) :

```
series nommoymob = @movav(nomserie, k)
```

où **nommoymob** est le nom que l'on donne à la nouvelle série, où **nomserie** est le nom de la série à partir de laquelle on calcule les moyennes, et où *k* est l'ordre des moyennes mobiles.

- Pour des moyennes mobiles centrées :

```
series nommoymob = @movavc(nomserie, k)
```

où **nommoymob** est le nom que l'on donne à la nouvelle série, où **nomserie** est le nom de la série à partir de laquelle on calcule les moyennes, et où *k* est l'ordre des moyennes mobiles.

La figure 12.8 montre les commandes pour produire des moyennes mobiles antérieures d'ordre 3, et des moyennes mobiles centrées d'ordre 3, toutes deux à partir de la série ABX. On voit aussi les séries résultantes.

The screenshot shows the EViews interface with the following content:

```
File Edit Object View Proc Quick Options Windows
smpl 1985:1 1994:4
series abxmovav3 = @movav(abx,3)
series abxcentmovav3 = @movavc(abx,3)
```

Below the commands is a data table titled "Group: UNTITLED Workfile: ABX::Abx". The table has two columns: "obs" and "ABXMOVAV3" (for prior moving average) and "ABXCENTMOAVAV3" (for centered moving average). The data spans from 1985Q1 to 1990Q2.

obs	ABXMOVAV3	ABXCENTMOAVAV3
1985Q1	NA	NA
1985Q2	NA	204.8333
1985Q3	204.8333	206.3333
1985Q4	206.3333	212.8333
1986Q1	212.8333	212.8333
1986Q2	212.8333	206.3333
1986Q3	206.3333	207.5000
1986Q4	207.5000	222.8333
1987Q1	222.8333	225.5000
1987Q2	225.5000	220.1667
1987Q3	220.1667	220.5000
1987Q4	220.5000	231.0000
1988Q1	231.0000	227.8333
1988Q2	227.8333	225.5000
1988Q3	225.5000	228.5000
1988Q4	228.5000	247.1667
1989Q1	247.1667	249.8333
1989Q2	249.8333	241.0000
1989Q3	241.0000	240.3333
1989Q4	240.3333	247.8333
1990Q1	247.8333	253.1667
1990Q2	253.1667	253.5000

FIG. 12.8 – Moyennes mobiles (antérieures (*prior*) et centrées)

12.4 La décomposition

La méthode de la décomposition est une méthode intuitive qui décompose une série chronologique en isolant les effets de tendance, cyclique, saisonnier et irrégulier. Cette méthode est efficace lorsque le tracé général de la série se répète à travers le temps.

Lorsque les données sont de nature annuelle, le modèle multiplicatif de la décomposition s'écrit de la manière suivante :

$$Y_t = T_t \times C_t.$$

Lorsque les données sont mensuelles ou trimestrielles, le modèle prend alors l'expression suivante :

$$Y_t = T_t \times C_t \times S_t \times I_t.$$

Dans ces expressions on a :

- Y_t est la valeur de la série observée au temps t ;
- T_t est la composante de la tendance au temps t ;
- C_t est la composante cyclique au temps t ;
- S_t est la composante saisonnière au temps t ;
- I_t est la composante d'irrégularité au temps t .

Lorsque les données mensuelles ou trimestrielles ne sont pas disponibles, il est alors impossible de coincer les effets des saisons et d'irrégularité à court terme. Et même si les données sont disponibles sur une base mensuelle, il n'y a pas toujours un effet saisonnier. Par exemple, il est possible que les ventes d'un produit ne soient pas influencées par les saisons. La consommation du lait est justement un exemple où l'effet des saisons sur les ventes est nul. Dans ce type de situation, le terme S_t est égal à 1 dans le modèle de décomposition. Aussi, il faut voir le terme effet saisonnier au sens large ; les mois ou les journées peuvent faire office de « saison » dans un modèle. Par exemple, il peut il y avoir plus d'achalandage le jeudi dans un centre d'achat.

Pour mettre en œuvre la décomposition, il est impératif d'isoler les effets les uns après les autres. Pour illustrer la méthode, nous utiliserons à nouveau la base de données **ABX.sav**.

Voici les étapes de la décomposition. Il est préférable d'effectuer ces étapes dans l'ordre.

- Visualisation de la série temporelle.
- Désaisonnalisation des données (si les données le permettent).
- Détermination de la tendance (linéaire ou non linéaire).
- Détermination de l'effet cyclique.
- Détermination de l'effet irrégulier.
- Établissement des prédictions.

On doit d'abord préciser la nature des données à SPSS. Il faut spécifier à SPSS que les données font parties d'une série chronologique. Il faut lui préciser le type de dates et il calculera lui-même ces dernières, ce qui minimise l'entrée de données. Les commandes sont les suivantes :

Menu SPSS :	→ Data
	→ Define Dates...
Dans la fenêtre Cases Are :	→ Years, quarters
Dans la fenêtre First Case Is :	Year : 1985 Quarter : 1

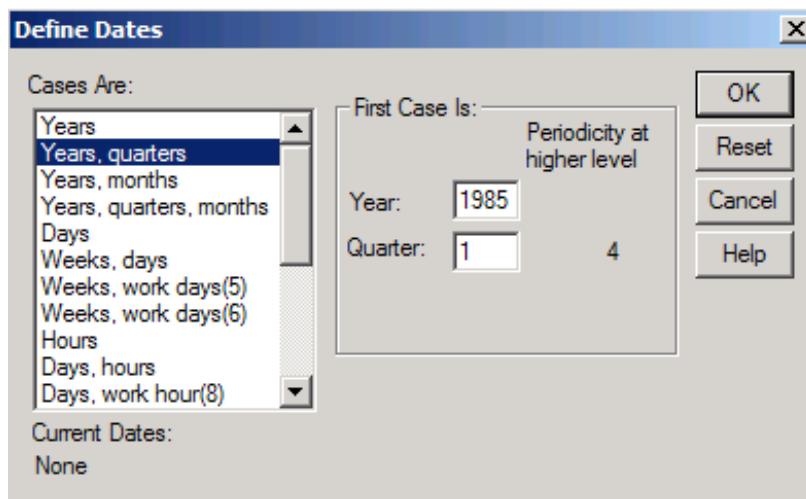


FIG. 12.9 – Définition des dates

Il faut savoir que seules les dates générées par SPSS peuvent servir de référentiel à SPSS. Sinon, bon nombre d'analyses ne seront pas possibles. Les commandes SPSS précédentes ont pour effet de créer les variables `YEAR_`, `QUARTER_` et `DATE_`.

	index	ventes	annee	YEAR_	QUARTER	DATE_
1	1	221,00	1985	1985	1	Q1 1985
2	2	203,50	1985	1985	2	Q2 1985
3	3	190,00	1985	1985	3	Q3 1985
4	4	225,50	1985	1985	4	Q4 1985
5	5	223,00	1986	1986	1	Q1 1986
6	6	190,00	1986	1986	2	Q2 1986
7	7	206,00	1986	1986	3	Q3 1986
8	8	226,50	1986	1986	4	Q4 1986
9	9	236,00	1987	1987	1	Q1 1987
10	10	214,00	1987	1987	2	Q2 1987
11	11	210,50	1987	1987	3	Q3 1987
12	12	237,00	1987	1987	4	Q4 1987
13	13	245,50	1988	1988	1	Q1 1988
14	14	201,00	1988	1988	2	Q2 1988
15	15	230,00	1988	1988	3	Q3 1988

FIG. 12.10 – Les nouvelles variables qui sont créées

Rappelons qu'une série temporelle contient des données à intervalles de longueur fixe. C'est ce que suppose SPSS. Si des données sont manquantes à des dates connues, il faut

alors laisser des espaces vides à ces endroits dans la base de données. Il est important de noter que plusieurs procédures informatisées telles les méthodes de lissage et la décomposition ne supportent pas les données manquantes. Pour ce faire, l'analyste doit imputer les données manquantes à la moyenne mobile. Ces techniques de remplacement des données manquantes sont présentées plus loin dans ce chapitre.

12.4.1 Visualisation de la série

Revoici le graphe séquentiel des ventes trimestrielles de ABX, mais cette fois-ci en fonction de la variable DATE_ au lieu de la variable index.

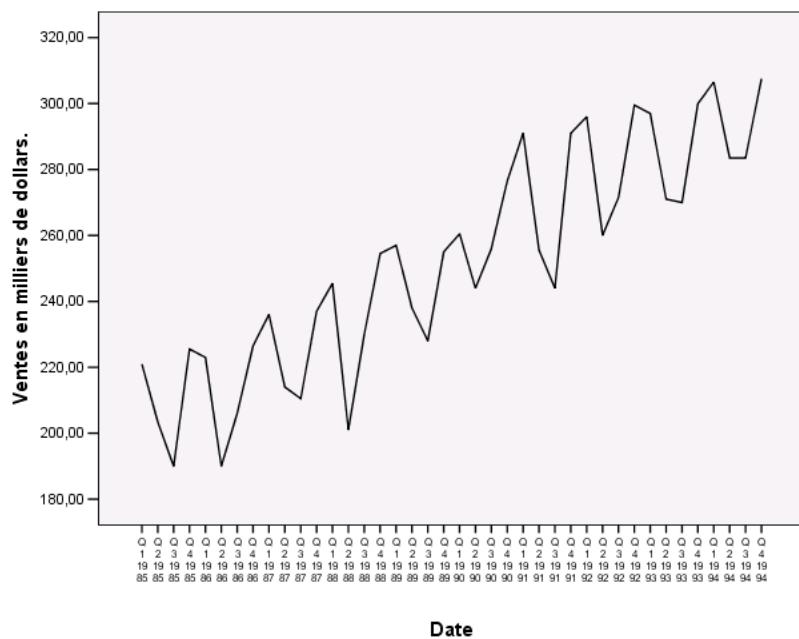


FIG. 12.11 – Visualisation de la série

L'examen de la figure 12.11 permet de comprendre la progression de la série. Ici on peut détecter un effet de saison et une tendance globalement linéaire dans les ventes, phénomène causé par la notoriété. On tentera donc une régression linéaire à titre de ligne de base au terme T_t . Une double moyenne mobile ferait probablement l'affaire aussi.

Mentionnons que la technique de la décomposition fonctionne même si la tendance n'est pas linéaire. Il suffit d'ajuster le bon modèle à la ligne de base de la tendance T_t dans le modèle. Quelques modèles non linéaires sont présentés plus loin dans ce chapitre.

12.4.2 Désaisonnalisation des données

La désaisonnalisation des données s'effectue suivant quelques algorithmes standards. L'algorithme le plus courant porte le nom de *Ratio-to-moving-average method*. L'algorithme isole les variations liées aux saisons.

Pour ce faire, l'algorithme présente un index représentant le poids, en terme de pourcentage, de chacun des trimestres (ou chacun des mois de l'année, chaque mois représentant une « saison »). À la base, chaque trimestre représente $\frac{1}{4}$ des ventes totales de l'année. Ainsi, un trimestre ayant index de 100 % représente un trimestre qui, en moyenne, a atteint $\frac{1}{4}$ des ventes annuelles centrées en ce trimestre. Un index de 150 % associé à un trimestre illustre que ce trimestre obtient, en moyenne, des ventes 50 % supérieures à $\frac{1}{4}$ des ventes annuelles centrées en ce trimestre. À l'opposé, un index de 80 % associé à un trimestre illustre que ce trimestre obtient, en moyenne, des ventes de 20 % inférieures au $\frac{1}{4}$ des ventes annuelles totales centrées en ce trimestre.

L'algorithme se décompose en trois étapes A, B et C :

Étape A : Calcul itératif des moyennes mobiles de comparaison.

– Si les données sont trimestrielles :

$$\bar{y}_{\text{trimestre_central}_t} = \frac{(0,5 \cdot y_{t-2}) + y_{t-1} + y_t + y_{t+1} + (0,5 \cdot y_{t+2})}{4}.$$

– Si les données sont mensuelles :

$$\bar{y}_{\text{mois_central}_t} = \frac{(0,5 \cdot y_{t-6}) + y_{t-5} + \cdots + y_{t-1} + y_t + y_{t+1} + \cdots + y_{t+5} + (0,5 \cdot y_{t+6})}{12}.$$

Dans le cadre de cet exemple, on a

$$\bar{y}_{\text{trimestre_central}_3} = \frac{(0, 5 \cdot 221) + 203, 5 + 190 + 225, 5 + (0, 5 \cdot 223)}{4} = 210, 25.$$

$$\bar{y}_{\text{trimestre_central}_4} = \frac{(0, 5 \cdot 203, 5) + 190 + 225, 5 + 223 + (0, 5 \cdot 190)}{4} = 208, 81.$$

Il est possible d'obtenir ces moyennes avec SPSS en effectuant les commandes suivantes :

Menu SPSS :	→ Transform
	→ Create Time Series...
Dans la fenêtre New Variable(s) :	→ ventes
Dans la fenêtre Name :	→ moycent (mettre un nom représentatif)
Dans la fenêtre Function :	→ sélectionnez Centered moving average
Dans la fenêtre Span :	→ 4 (la largeur d'un cycle)
Appuyez sur Change.	

La figure 12.12 donne un aperçu de la variable `moycent`.

index	ventes	YEAR	QUARTER	DATE	moycent
1	221,00	1985	1	Q1 1985	.
2	203,50	1985	2	Q2 1985	.
3	190,00	1985	3	Q3 1985	210,25
4	225,50	1985	4	Q4 1985	208,81
5	223,00	1986	1	Q1 1986	209,12
6	190,00	1986	2	Q2 1986	211,25
7	206,00	1986	3	Q3 1986	213,00
8	226,50	1986	4	Q4 1986	217,62
9	236,00	1987	1	Q1 1987	221,19
10	214,00	1987	2	Q2 1987	223,06
11	210,50	1987	3	Q3 1987	225,56
12	237,00	1987	4	Q4 1987	225,12
13	245,50	1988	1	Q1 1988	225,94
14	201,00	1988	2	Q2 1988	230,56
15	230,00	1988	3	Q3 1988	234,19
16	254,50	1988	4	Q4 1988	240,25
17	257,00	1989	1	Q1 1989	244,62
18	238,00	1989	2	Q2 1989	244,44

FIG. 12.12 – Aperçu de la moyenne centrée

Étape B : Calcul de l'index saisonnier.

- Si les données sont trimestrielles :

$$\text{index}_{\text{trimestre_central}_t} = \frac{y_t}{\bar{y}_{\text{trimestre_central}_t}}.$$

- Si les données sont mensuelles :

$$\text{index}_{\text{mois_central}_t} = \frac{y_t}{\bar{y}_{\text{mois_central}_t}}.$$

Dans le cadre de cet exemple, on a :

$$\text{index}_{\text{trimestre_central}_3} = \frac{190}{210,25} = 0,903686.$$

$$\text{index}_{\text{trimestre_central}_4} = \frac{225,5}{208,81} = 1,079929.$$

Il est possible de faire effectuer ces calculs par SPSS de la manière suivante :

Menu SPSS :	→ Transform
	→ Compute...
Dans la fenêtre Target Variable :	→ index_s
Dans la fenêtre Numeric Expression :	→ ventes/moycent

YEAR	QUARTER	DATE	moycent	index_s
1985	1	Q1 1985	.	.
1985	2	Q2 1985	.	.
1985	3	Q3 1985	210,25	,9037
1985	4	Q4 1985	208,81	1,0799
1986	1	Q1 1986	209,12	1,0663
1986	2	Q2 1986	211,25	,8994
1986	3	Q3 1986	213,00	,9671
1986	4	Q4 1986	217,62	1,0408
1987	1	Q1 1987	221,19	1,0670
1987	2	Q2 1987	223,06	,9594
1987	3	Q3 1987	225,56	,9332

FIG. 12.13 – Aperçu de l'index saisonnier

Étape C : Calcul de l'index saisonnier ajusté.

Afin de trouver l'index saisonnier ajusté, la méthode calcule dans un premier temps une moyenne tronquée pour chaque trimestre (en prenant les valeurs obtenues pour l'index de ce trimestre). Cette technique permet simplement plus de robustesse mais peut être optionnelle si le nombre d'années ne permet pas au moins 4 données par moyenne. Ainsi, isolément pour chacun des trimestres, une moyenne est calculée en enlevant les valeurs minimale et maximale, d'où la troncation.

Par exemple, dans le cadre de l'exemple ABX, l'index saisonnier prend les valeurs suivantes :

trimestre	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
1		1,0663	1,067	1,0866	1,0506	1,0404	1,0848	1,0717	1,0437	1,0508
2		0,8994	0,9594	0,8718	0,9737	0,9510	0,9514	0,9263	0,9528	0,9633
3	0,9037	0,9671	0,9332	0,9821	0,9308	0,9732	0,9004	0,9632	0,9451	
4	1,0799	1,0408	1,0527	1,0593	1,0361	1,0305	1,0691	1,0569	1,0401	

FIG. 12.14 – L'index saisonnier pour chaque trimestre et chaque année

Ainsi, pour le premier et le deuxième trimestre, les moyennes tronquées sont

$$\bar{y}_{\text{tronquée_trimestre}_1} = \frac{1,0663 + 1,067 + 1,0506 + 1,0848 + 1,0717 + 1,0437 + 1,0508}{7} \\ = 1,0621.$$

$$\bar{y}_{\text{tronquée_trimestre}_2} = \frac{0,8994 + 0,9594 + 0,9510 + 0,9514 + 0,9263 + 0,9528 + 0,9633}{7} \\ = 0,9434.$$

La moyenne tronquée ne peut être calculée qu'avec au moins quatre données. SPSS n'effectue pas les calculs si moins de quatre cycles complets sont présents dans la base. Lorsque moins de quatre cycles sont disponibles, l'analyste doit effectuer lui-même les calculs en remplaçant la moyenne tronquée par une moyenne ordinaire. Il n'est pas recommandé d'utiliser le modèle de décomposition avec moins de deux cycles complets.

En théorie, suivant la construction effectuée, la somme des quatre moyennes tronquées devrait donner 4 (ou 12 dans le cas de données mensuelles). La somme ici donne 4,0014 suite aux erreurs d'arrondissement. Une règle d'équivalence ramène cette somme à 4 de

la manière suivante :

$$\text{Ajustement} = \frac{\text{Somme théorique}}{\text{Somme observée}} = \frac{4}{4,0014} = 0,9997.$$

Finalement, les moyennes tronquées sont révisées en fonction de cet ajustement pour obtenir l'index saisonnier ajusté :

$$\text{Index saisonnier ajusté}_{\text{trimestre}_i} = \bar{y}_{\text{tronquée}_i} \times \text{Ajustement}$$

pour $i = 1, 2, 3, 4$.

trimestre	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	Moyenne tronquée	Index saisonnier ajusté.
1		1,0663	1,067	1,0866	1,0506	1,0404	1,0848	1,0717	1,0437	1,0508	1,0621	1,0618
2		0,8994	0,9594	0,8718	0,9737	0,9510	0,9514	0,9263	0,9528	0,9633	0,9434	0,9430
3	0,9037	0,9671	0,9332	0,9821	0,9308	0,9732	0,9004	0,9632	0,9451		0,9452	0,9449
4	1,0799	1,0408	1,0527	1,0593	1,0361	1,0305	1,0691	1,0569	1,0401		1,0507	1,0503
											Somme	4,0014
											Ajustement	0,9997

FIG. 12.15 – L'index saisonnier ajusté

Il est possible de faire effectuer directement ces calculs par SPSS, et ce, sans passer par les étapes A et B. Il faut utiliser les commandes suivantes :

Menu SPSS :	→ Analyse
	→ Time Series
	→ Seasonnal Decomposition...
Dans la fenêtre Variable(s) :	ventes
Model :	✓ Multiplicative
Moving Average Weight :	✓ Endpoints weighted by .5
Dans le bouton Save... :	✓ Add to file
	✓ Display casewise listing

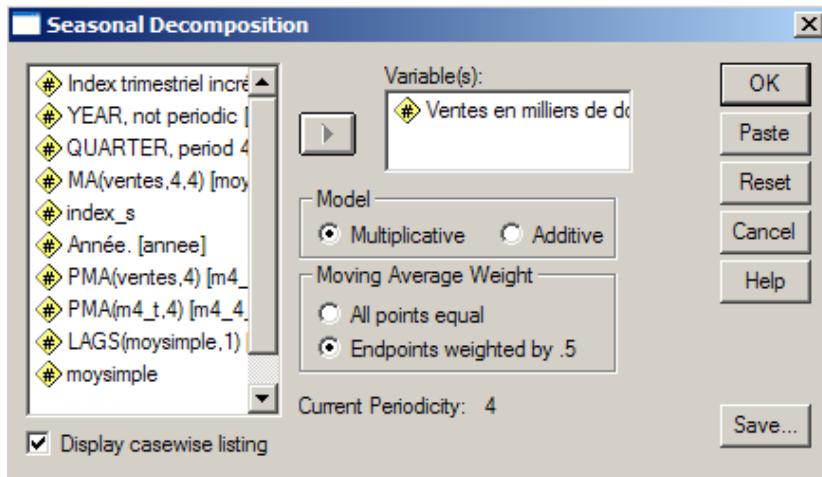


FIG. 12.16 – Pour créer l'index saisonnier ajusté

Ces commandes produiront une sortie ainsi que de nouvelles variables dans le fichier de données. Les nouvelles variables se nomment `ERR_1`, `SAS_1`, `SAF_1` et `STC_1`. Plus précisément :

SAS_1 : Cette variable est formée des données désaisonnalisées qui sont le quotient des données brutes par l'index saisonnier ajusté (chaque donnée est divisée par l'index ajusté correspondant à son trimestre).

index	ventes	YEAR	QUARTER	DATE	moycent	ERR_1	SAS_1	SAF_1	STC_1
1	221,00	1985	1	Q1 1985	.	,98384	208,14426	1,06176	211,56231
2	203,50	1985	2	Q2 1985	.	1,03577	215,79438	,94303	208,34219
3	190,00	1985	3	Q3 1985	210,25	,96142	201,08794	,94486	209,15620
4	225,50	1985	4	Q4 1985	208,81	1,02582	214,69054	1,05035	209,28627
5	223,00	1986	1	Q1 1986	209,12	1,00463	210,02792	1,06176	209,05911
6	190,00	1986	2	Q2 1986	211,25	,95898	201,47878	,94303	210,09651
7	206,00	1986	3	Q3 1986	213,00	1,02165	218,02166	,94486	213,40082
8	226,50	1986	4	Q4 1986	217,62	,99226	215,64260	1,05035	217,32467
9	236,00	1987	1	Q1 1987	221,19	1,00385	222,27170	1,06176	221,41819

FIG. 12.17 – Un aperçu des données

SAF_1 : Cette variable contient les index saisonniers ajustés (dans le cadre de l'exemple, puisqu'il y a quatre trimestres, cette variable n'a que quatre valeurs différentes qui se répètent).

STC_1 : Cette variable est formée des double moyennes mobiles M_t^{3*3} calculées à partir des données désaisonnalisées (variable **SAS_1**). Cette double moyenne mobile simule une certaine linéarisation représentant l'effet de la tendance. Plus de précisions seront apportées sur cet effet dans une prochaine section. Cette double moyenne, ayant peu de mémoire, poursuit lentement les cycles. Il peut être intéressant d'utiliser **STC_1** à titre de composante T_t*C_t dans le modèle $Y_t = T_t*C_t*S_t*I_t$. Malgré que ce ne soit pas une procédure standard, elle possède l'avantage de ne rien supposer sur la nature linéaire ou non de la tendance. De plus cette courbe lente s'ajuste aux cycles économiques, c'est pourquoi SPSS propose ce modèle.

ERR_1 : Cette variable est le quotient $\frac{\text{SAS}_1}{\text{STC}_1}$. Ce terme représente un terme d'irrégularité I_t (pas toujours aléatoire) lorsque le terme **STC_1** est utilisé à titre de terme $T_t * C_t$ dans le modèle de décomposition.

Au départ, pour des données trimestrielles, chaque trimestre représente $\frac{1}{4}$ des ventes totales de l'année. Dans le cadre de l'exemple, le premier trimestre obtient un indice saisonnier ajusté de 1,06176, ce qui s'interprète ainsi : en moyenne, le premier trimestre obtient un niveau de vente de 6,176 % supérieur à $\frac{1}{4}$ des ventes annuelles centrées en ce trimestre. Les deux autres trimestres obtiennent des index similaires d'environ 0,94,

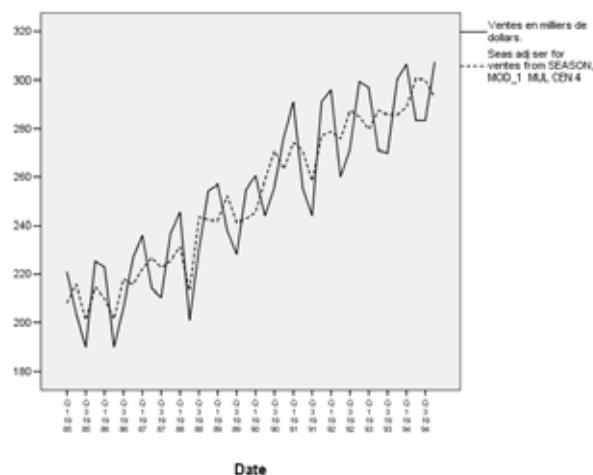
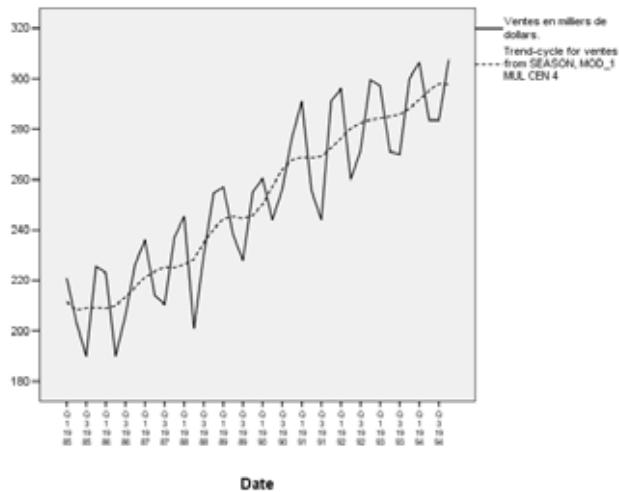


FIG. 12.18 – Les données désaisonnalisées

montrant que ces trimestres admettent en moyenne un niveau de vente moyen d'environ 6 % inférieur au $\frac{1}{4}$ des ventes annuelles centrées en ces trimestres. L'interprétation de l'index saisonnier ajusté du dernier trimestre s'effectue de la même manière.

FIG. 12.19 – Les double moyennes mobiles M_t^{3*3} des données désaisonnalisées

12.4.3 Détermination de la tendance

Il est important de déterminer la tendance à long terme T_t d'une série temporelle. L'objectif consiste à établir une équation modélisant la tendance (qui n'est pas toujours linéaire). L'équation remplacera le terme T_t dans le modèle de décomposition. Cette section utilise la droite de régression pour estimer la tendance linéaire. Mais des modèles logarithmiques, de Weibull, de Gompertz (courbe de croissance d'une population), logistiques ou autres peuvent être utilisés. Il faut simplement que les paramètres du modèle soient déterminés de manière à obtenir une équation. Une section subséquente présente ces modèles ainsi que l'estimation de leurs paramètres.

Pour obtenir les paramètres de la régression, l'analyste utilise l'algorithme des moindres carrés standards. Si aucun effet saisonnier n'est visible, une dizaine de données sont nécessaires ; en présence d'un effet saisonnier, il est recommandé d'utiliser 6 cycles complets pour estimer la tendance. L'analyste peut effectuer les calculs avec moins d'observations mais il devra alors être plus sévère lors de l'analyse de ses résidus.

Pour visualiser la tendance, on peut regarder les graphes des données brutes et des données désaisonnalisées en fonction du temps. Dans le cadre de l'exemple ABX, on obtient les graphes de la figure 12.20.

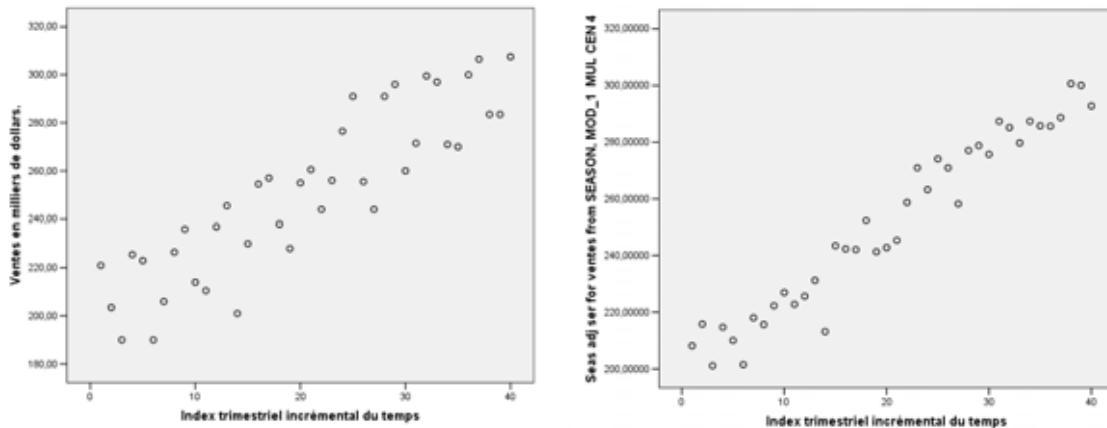


FIG. 12.20 – Les ventes réelles et les ventes désaisonnalisées

Dans les deux cas la nature linéaire de la tendance est bien visible. On remarque aussi

que les données désaisonnalisées sont beaucoup moins dispersées autour de la tendance.

Il est donc tout à fait pertinent d'utiliser la régression linéaire pour estimer cette tendance. Certains experts préfèrent modéliser la tendance à partir des données brutes, d'autres à partir des données désaisonnalisées. Dans le cadre de ce document, la série désaisonnalisée sera utilisée. Cette stratégie permet simplement d'obtenir une vision plus nette de la droite de régression qui, en enlevant la variation due à l'effet de saison, est plus stable. Ceci est davantage vérifique s'il existe des différences marquées aux ventes moyennes trimestrielles (ou mensuelles) issues d'un petit nombre de périodes durant l'année ; ces écarts influencent beaucoup la régression.

Dans le cadre de l'exemple on fait donc la régression linéaire avec les données désaisonnalisées et l'index, et on obtient les sorties des figures 12.21 et 12.22.

Model Summary ^b					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,973 ^a	,947	,946	7,19639635	1,900

a. Predictors: (Constant), index
b. Dependent Variable: SAS_1

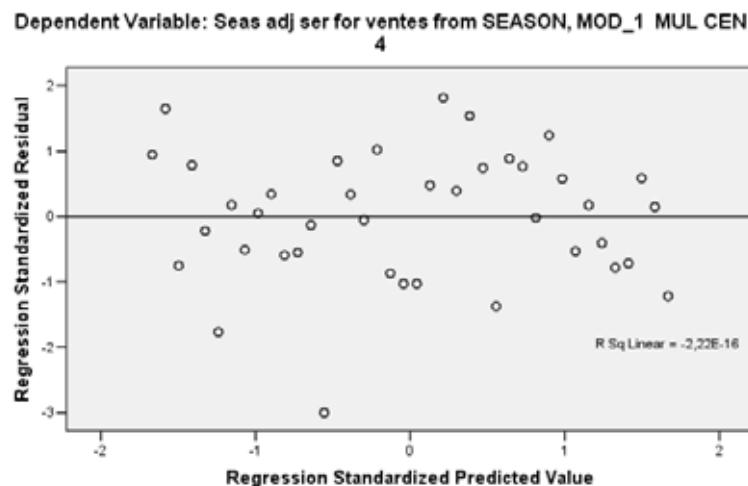
ANOVA ^b					
Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression 35179,098	1	35179,098	679,289	,000 ^a
	Residual 1967,949	38	51,788		
	Total 37147,046	39			

a. Predictors: (Constant), index
b. Dependent Variable: SAS_1

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant) 198,764	2,319		85,709	,000
	index 2,569	,099	,973	26,063	,000

a. Dependent Variable: SAS_1

FIG. 12.21 – Sorties de la régression



Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ZRE_1	,061	40	,200*	,975	40	,504

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. 12.22 – Les résidus

Il faudrait faire ici une analyse complète de cette régression. On voit que 94,7 % de la variation des données désaisonnalisées est expliquée par la variable `index`. Le test de Durbin-Watson a comme conclusion qu'il n'y a pas d'autocorrélation. Les résidus sont répartis de façon assez aléatoire et sont distribués selon une loi normale. Aussi, indépendamment de l'effet des saisons, plus le temps passe, plus les ventes augmentent ; la notoriété de cette entreprise est donc grandissante.

L'équation de la droite de régression est la suivante :

$$\hat{y}_{SAS_1} = 198,764 + 2,569x_{\text{index}} = T_t.$$

Cette équation s'interprète comme une droite de régression standard. Ainsi à chaque trimestre additionnel les ventes désaisonnalisées augmentent en moyenne de 2,569 unités, donc de 2 569 \$.

12.4.4 Détermination de l'effet cyclique

Les effets cycliques ont un effet de vagues sur la série et les changements de direction sont lents, soit sur des périodes de plus de huit mois et pouvant atteindre des périodes de quatre ans et plus. Ces variations cycliques représentent l'impact de l'économie sur la série. De l'équation du modèle de décomposition il est possible d'isoler les termes $C_t * I_t$:

$$Y_t = T_t * C_t * S_t * I_t \Leftrightarrow \frac{Y_t}{T_t * S_t} = C_t * I_t.$$

Afin d'isoler la composante C_t qui varie lentement par hypothèse, il est courant d'utiliser une moyenne mobile centrée. Lorsque les données sont mensuelles, des moyennes mobiles d'ordre 3, 5, 7 ou 9 sont utilisées. Le choix de l'ordre dépend de la lenteur du mouvement du phénomène, ce choix revient à l'analyste. Plus l'ordre est petit, plus l'ajustement est rapide. Lorsque les données sont trimestrielles, une moyenne mobile d'ordre 3 est utilisée. La valeur impaire de l'ordre permet simplement d'obtenir une valeur centrale. Plus précisément, on a

- si les données sont trimestrielles :

$$\bar{c}_{\text{trimestre_central}_t} = \frac{y_{t-1} + y_t + y_{t+1}}{3};$$

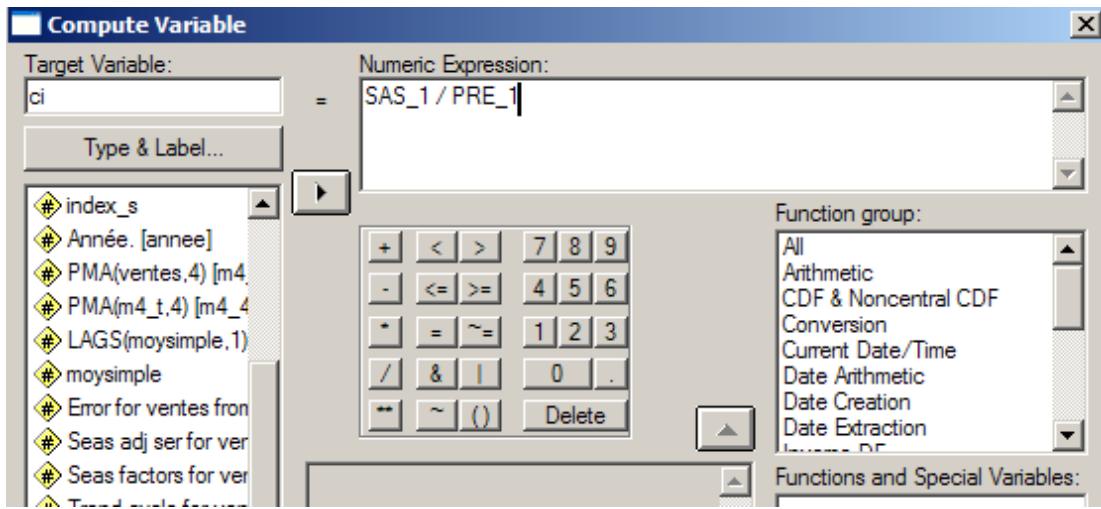
- si les données sont mensuelles :

$$\bar{c}_{\text{mois_central}_t} = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}}{5}$$

(pour le choix de l'ordre 5).

Il est possible de calculer la composante $C_t * I_t$ avec SPSS. Il suffit d'observer que $C_t * I_t = \frac{Y_t}{T_t * S_t} = \frac{Y_t}{S_t} * \frac{1}{T_t}$, ce qui en terme de variables correspond à **SAS_1 / PRE_1** où **PRE_1** est la variable des prédictions obtenues du modèle estimant la composante de la tendance T_t .

On obtient donc la composante $C_t * I_t$ en calculant **SAS_1 / PRE_1** à l'aide des menus **Transform** puis **Compute...** (voir la figure 12.23).

FIG. 12.23 – Calcul de la composante $C_t * I_t$

Une fois le terme $C_t * I_t$ isolé, il est possible de calculer une moyenne mobile centrée d'ordre 3 pour isoler l'effet cyclique. Voici les commandes pour une moyenne mobile centrée :

Menu SPSS :	→ Transform
	→ Create Time Series...
Dans la fenêtre New Variable(s) :	→ ci
Dans la fenêtre Name :	→ cycle
Dans la fenêtre Function :	→ sélectionnez Centered moving average
Dans la fenêtre Span :	→ 3 (l'ordre de la moyenne mobile)
Appuyez sur Change.	

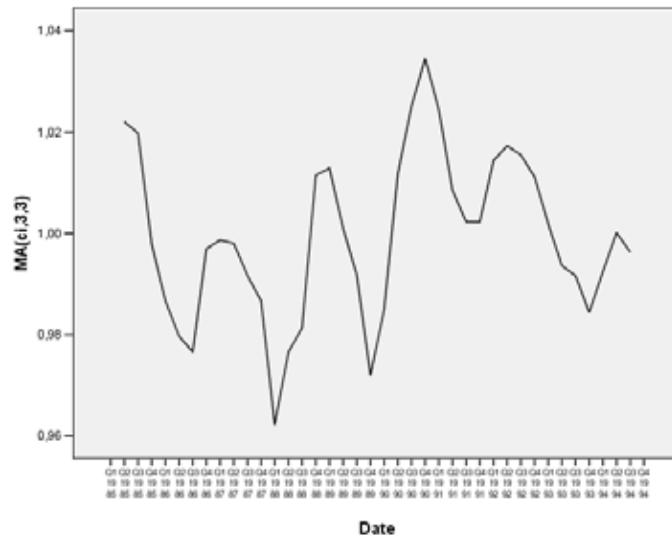
Comme l'ordre de la moyenne mobile est impair, SPSS n'attribue pas de poids de 0,5 aux extrémités. Ce poids est attribué aux valeurs de chaque extrémité seulement lorsque l'ordre est pair.

Lorsqu'il n'y a pas d'effet cyclique dans une série, les valeurs que prend C_t tendent vers 1 (puisque l'effet multiplicatif de 1 est neutre). Un graphe séquentiel nous permet

DATE	moycent	index_s	ERR_1	SAS_1	SAF_1	STC_1	PRE_1	ci	cycle
Q1 1985	.	.	,98384	208,1443	1,06176	211,5623	201,33355	1,03	.
Q2 1985	.	.	1,03577	215,7944	,94303	208,3422	203,90264	1,06	1,02
Q3 1985	210,25	,9037	,96142	201,0879	,94486	209,1562	206,47172	,97	1,02
Q4 1985	208,81	1,0799	1,02582	214,6905	1,05035	209,2863	209,04081	1,03	1,00
Q1 1986	209,12	1,0663	1,00463	210,0279	1,06176	209,0591	211,60990	,99	,99
Q2 1986	211,25	,8994	,95898	201,4788	,94303	210,0965	214,17898	,94	,98
Q3 1986	213,00	,9671	1,02165	218,0217	,94486	213,4008	216,74807	1,01	,98
Q4 1986	217,62	1,0408	,99226	215,6426	1,05035	217,3247	219,31716	,98	1,00
Q1 1987	221,19	1,0670	1,00385	222,2717	1,06176	221,4182	221,88624	1,00	1,00

FIG. 12.24 – Un aperçu des données

de visualiser quel semble être l'effet cyclique pour la série ABX (figure 12.25).

FIG. 12.25 – L'effet cyclique (composante C_t)

Le graphe illustre que les coefficients C_t tournent autour de 1. Étudions ceci plus attentivement à l'aide des statistiques descriptives.

L'analyse de la sortie 12.26 montre que la valeur du cycle moyen est coincée entre 0,9940 et 1,0048, et ce avec une probabilité de 95 %. Comme la valeur 1 est dans cet intervalle, un t -test au seuil de 5 % ne rejette pas que la moyenne est de 1. De plus,

Descriptives		
cycle	Mean	,9994
	95% Confidence Interval for Mean	,9940 1,0048
	Lower Bound	
	Upper Bound	
	5% Trimmed Mean	,9995
	Median	,9983
	Variance	,000
	Std. Deviation	,01635
	Minimum	,96
	Maximum	1,03
	Range	,07
	Interquartile Range	,03
	Skewness	-,020 ,383
	Kurtosis	-,355 ,750

FIG. 12.26 – Les statistiques descriptives de la composante C_t

par son coefficient de variation inférieur à 0,15 ($CV = 0,01635/0,9994 = 0,01636$), cette moyenne est (très !) représentative.

Cependant, les statistiques précédentes ne tiennent pas compte du fait que la série de données est temporelle. Afin de tenir compte de l'ordonnancement de la série, on utilisera un test de *Runs*. Pour ce faire, on considère la moyenne 0,9994 et une fonction $f(C_t)$; lorsqu'une donnée de la série est supérieure ou égale à la moyenne on pose $f(C_t) = +$, et lorsqu'une donnée est strictement inférieure à cette moyenne on pose $f(C_t) = -$. Dans le cadre de l'exemple, on obtient le tableau de la figure 12.27.

S'il n'y a pas d'effet cyclique, on s'attend à ce que la suite créée par la fonction $f(C_t)$ ait une distribution aléatoire, un peu comme si on avait tiré à pile ou face : $+---++-++-+-+--$. S'il y a vraiment un effet de cycle, alors la suite ne sera pas aléatoire et ressemblera davantage à ceci : $----++++--++++$.

Pour tester ceci formellement, posons d'abord m comme étant le nombre de $+$ obtenus dans la série, et n le nombre de $-$. On résout alors le test d'hypothèses suivant :

Index	C_t	$f(C_t)$									
2	1, 02	+	11	, 99	-	21	, 98	-	30	1, 02	+
3	1, 02	+	12	, 99	-	22	1, 01	+	31	1, 02	+
4	1, 00	+	13	, 96	-	23	1, 03	+	32	1, 01	+
5	, 99	-	14	, 98	-	24	1, 03	+	33	1, 00	+
6	, 98	-	15	, 98	-	25	1, 02	+	34	, 99	-
7	, 98	-	16	1, 01	+	26	1, 01	+	35	, 99	-
8	1, 00	+	17	1, 01	+	27	1, 00	+	36	, 98	-
9	1, 00	+	18	1, 00	+	28	1, 00	+	37	, 99	-
10	1, 00	+	19	, 99	-	29	1, 01	+	38	1, 00	+
			20	, 97	-				39	1, 00	+

FIG. 12.27 – La composante C_t et la fonction $f(C_t)$ (test de *Runs*)

H_0 : La séquence de + et de - de la fonction $f(C_t)$ a été générée par $m + n$ expériences indépendantes de Bernoulli de probabilité $\pi_+ = \pi_- = 0,5$.

H_1 : La suite n'est pas indépendante.

En résumé, le test de *Runs* s'intéresse au nombre de changements de signe dans la suite. Par exemple, la suite ordonnée —————+———— contient seulement 4 *runs* en 16 essais, ce qui n'apparaît pas aléatoire. Pour traiter le test d'hypothèses, l'algorithme calcule la probabilité d'obtenir seulement 4 *runs* parmi toutes les permutations possibles d'une suite de 16 essais.

En somme, la suite C_t se comporte aléatoirement autour de 1 (aucun effet de cycle) si l'hypothèse H_0 n'est pas rejetée. Pour obtenir le test de *Runs*, il suffit d'effectuer les commandes suivantes :

Menu SPSS :

- Analyse
- Nonparametric tests
- Runs...

Dans la fenêtre Test Variable List : → cycle

Dans la fenêtre Cut Point : ✓ Median

 ✓ Mean

Runs Test		Runs Test 2	
	cycle		cycle
Test Value ^a	1,00		,9994
Cases < Test Value	19		20
Cases >= Test Value	19		18
Total Cases	38		38
Number of Runs	10		8
Z	-3,125		-3,776
Asymp. Sig. (2-tailed)	,002		,000

FIG. 12.28 – Les sorties du test de *Runs*

Pour résoudre le test, il suffit de comparer les p -values des deux tests avec notre seuil α fixé à 0,05 (on fait aussi le test autour de la médiane pour plus de certitude). Puisque les p -values (**Asymp. Sig (2-tailed)**) des deux tests sont inférieures à 0,05 (0,002 et 0,000), on rejette H_0 , et donc au risque de se tromper une fois sur 20 on peut dire qu'un effet non aléatoire (l'effet cyclique) semble intervenir dans les hauts et les bas de la série. Cependant, il faut avoir en tête que le test de *Runs* ne tient pas compte de l'ampleur de l'effet cyclique. Rassemblant les résultats du graphique, de l'analyse univariée et du résultat des tests d'hypothèses, on peut conclure qu'il existe un effet cyclique périodique mais que celui-ci est marginal (puisque la moyenne est très représentative, l'effet de cycle ne s'éloigne jamais beaucoup de la valeur 1). Ainsi, dans le cadre de cet exemple, le terme C_t sera fixé à 1 pour effectuer les prédictions.

Lorsque le terme C_t est non aléatoire et que l'ampleur est importante, l'analyste peut tenter d'étudier et de modéliser, avec un lissage ou une régression par exemple, le comportement de la variable observée C_t . L'objectif de cet exercice consiste à trouver les variables économiques qui ont un impact sur la série, ce qui n'est pas une mince tâche dans la plupart des cas. Aussi, le lissage possède ses limitations lorsque des prédictions à long terme sont nécessaires. Parallèlement, lorsque le lissage est impossible ou peu prudent, il est possible de réunir des experts via des groupes de discussion (*focus group*) dans l'objectif d'avoir une meilleure vision de la progression des cycles à venir. D'autres techniques de nature qualitative existent, tel la méthode Delphi, mais sont au delà de l'objectif de ce chapitre.

12.4.5 Détermination de l'effet irrégulier

Le terme I_t des irrégularités représente les effets non prévisibles autres que les effets saisonnier, cyclique et de tendance. Des événements tels des grèves ou des attaques comme celle du 11 septembre sont représentés par le terme I_t . Idéalement, ce terme est entièrement aléatoire (c'est-à-dire $I_t \sim 1$ aléatoirement). Les analystes l'appellent le bruit blanc ou encore le bruit de fonds créé par le marché. Un terme aléatoire est incontrôlable. Cependant, lorsque le bruit n'est pas aléatoire, l'analyste peut étudier et même modéliser l'impact de ce terme sur la série. C'est probablement ainsi que les analystes financiers sont en mesure de dire que les effets du 11 septembre tendent à diminuer sur les marchés (c'est-à-dire que $I_t \rightarrow 1$ aléatoirement depuis le 11 septembre).

Pour isoler le terme I_t dans le modèle de décomposition, il suffit de l'isoler dans l'équation du modèle :

$$Y_t = T_t \times C_t \times S_t \times I_t \Leftrightarrow I_t = \frac{Y_t}{T_t \times S_t \times C_t}.$$

Rappelons que dans le cadre de l'exemple ABX, la composante $C_t * I_t$ a été obtenue en calculant `SAS_1 / PRE_1` et a été nommée `ci`. De plus, en 12.4.4, l'effet cyclique C_t

a été déterminé (variable `cycle`). Donc $I_t = \frac{C_t \times I_t}{C_t}$, ce qu'on peut obtenir en divisant `ci` par `cycle` à l'aide du menu **Compute**. Puisqu'on avait conclut que $C_t = 1$, on pourrait se contenter d'identifier la variable `ci` comme représentant la composante irrégulière. Mais pour plus de précision, on prend la peine de vraiment isoler la composante irrégulière I_t , et le fait que C_t ait été fixé à 1 n'interviendra que pour les prédictions. On crée donc la variable `irr` pour représenter l'effet irrégulier (figure 12.29).

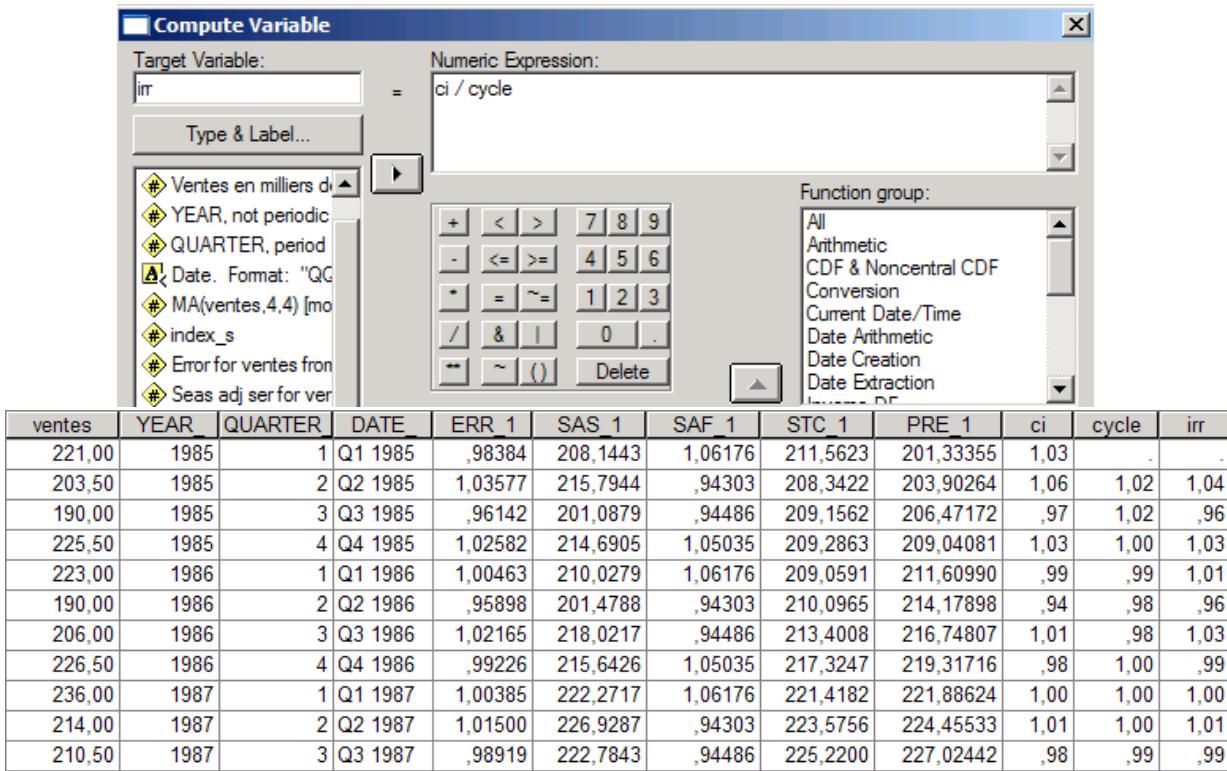
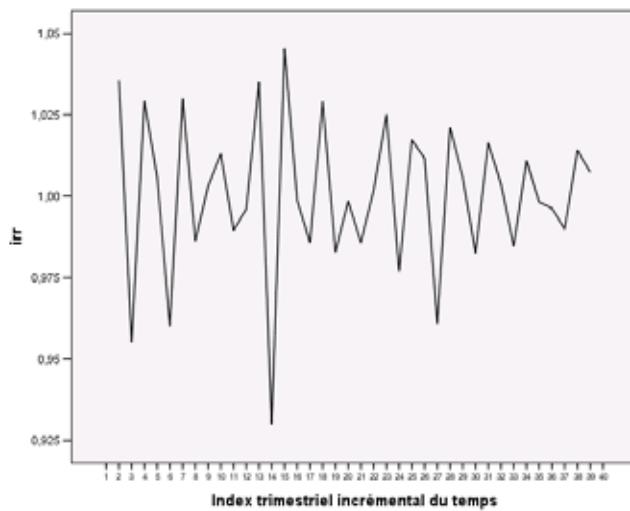


FIG. 12.29 – Création de la variable `irr` pour représenter la composante I_t

Pour visualiser et étudier la progression des irrégularités dans la série, il suffit de générer un graphe séquentiel et les statistiques descriptives pour la variable `irr` (figures 12.30 et 12.31).

L'analyse du graphique et des statistiques descriptives montrent que l'effet moyen des irrégularités tourne autour de 1. L'oscillation est aussi très rapide, ce qui suggère un bruit blanc. Par son coefficient de variation de 0,0243, cette moyenne est représentative.

FIG. 12.30 – L'effet irrégulier (composante I_t)

Descriptives			
		Statistic	Std. Error
irr	Mean	1,0005	,00394
	95% Confidence Lower Bound	,9925	
	Interval for Mean Upper Bound	1,0085	
	5% Trimmed Mean	1,0015	
	Median	1,0024	
	Variance	,001	
	Std. Deviation	,02428	
	Minimum	,93	
	Maximum	1,05	
	Range	,12	
	Interquartile Range	,03	
	Skewness	-,654	,383
	Kurtosis	,848	,750

FIG. 12.31 – Les statistiques descriptives de la composante I_t

Cependant, les statistiques précédentes ne tiennent pas compte du fait que les données forment une série temporelle. Comme pour l'effet cyclique, on considère la moyenne 1,0005 et la fonction $f(I_t)$; lorsqu'une donnée de la série est supérieure ou égale à la moyenne on pose $f(I_t) = +$, et lorsqu'une donnée est strictement inférieure à cette moyenne on pose $f(I_t) = -$. On obtient alors le tableau de la figure 12.32.

Index	I_t	$f(I_t)$	Index	I_t	$f(I_t)$	Index	I_t	$f(I_t)$	Index	C_t	$f(C_t)$
2	1,04	+	11	,99	-	21	,99	-	30	,98	-
3	,96	-	12	1,00	-	22	1,00	-	31	1,02	+
4	1,03	+	13	1,04	+	23	1,03	+	32	1,00	-
5	1,01	+	14	,93	-	24	,98	-	33	,98	-
6	,96	-	15	1,05	+	25	1,02	+	34	1,01	+
7	1,03	+	16	1,00	-	26	1,01	+	35	1,00	-
8	,99	-	17	,99	-	27	,96	-	36	1,00	-
9	1,00	-	18	1,03	+	28	1,02	+	37	,99	-
10	1,01	+	19	,98	-	29	1,01	+	38	1,01	+
			20	1,00	-				39	1,01	+

FIG. 12.32 – La composante I_t et la fonction $f(I_t)$ (test de *Runs*)

Rappelons que le test de *Runs* confronte les hypothèses suivantes :

H_0 : La séquence de + et de - de la fonction $f(I_t)$ a été générée par $m + n$ expériences indépendantes de Bernoulli de probabilité $\pi_+ = \pi_- = 0,5$.

H_1 : La suite n'est pas indépendante.

avec m le nombre de + obtenus dans la série, et n le nombre de -.

En somme, la suite I_t réagit comme un bruit blanc si l'hypothèse H_0 n'est pas rejetée.

Rappelons les commandes pour obtenir le test de *Runs* :

Menu SPSS :

- Analyse
- Nonparametric tests
- Runs . . .

Dans la fenêtre Test Variable List : → irr

Dans la fenêtre Cut Point : ✓ Median

 ✓ Mean

Runs Test		Runs Test 2	
	irr		irr
Test Value ^a	1,00	Test Value ^a	1,0005
Cases < Test Value	19	Cases < Test Value	18
Cases >= Test Value	19	Cases >= Test Value	20
Total Cases	38	Total Cases	38
Number of Runs	25	Number of Runs	25
Z	1,480	Z	1,502
Asymp. Sig. (2-tailed)	,139	Asymp. Sig. (2-tailed)	,133

FIG. 12.33 – Les sorties du test de *Rungs*.

Puisque les p -values sont de 0,139 et 0,133, ce qui est plus grand que $\alpha = 0,05$, on ne rejette pas H_0 . Donc au seuil $\alpha = 0,05$ on conclut que le terme I_t se comporte aléatoirement autour de 1. Et, compte tenu que la moyenne des irrégularités est fortement représentative autour de 1, le terme sera alors fixé à 1 dans l'équation du modèle de décomposition, ce qui facilitera les prédictions futures.

Tout comme le terme cyclique, lorsque le terme I_t est non aléatoirement centré à 1, l'analyste peut tenter d'étudier et de modéliser la variable observée **irr** afin d'effectuer de meilleures prédictions. Par exemple, il peut utiliser une variable binaire dans une régression pour isoler les zones de fortes turbulences des autres. L'analyste peut aussi avoir recours à des groupes de discussion avec des experts pour parvenir à estimer le terme I_t à l'avenir.

12.4.6 Établissement des prédictions

Une fois les composantes du modèle de décomposition estimées, il suffit d'inverser le processus pour obtenir des estimations. La figure 12.34 présente la synthèse des résultats des sous-sections précédentes à propos des composantes du modèle de décomposition dans le cadre de l'exemple ABX.

Les colonnes **LICI_1** et **UICI_1** sont les bornes des intervalles de prédictions issus de la régression linéaire élaborée à la sous-section 12.4.7 pour estimer la tendance.

Supposons que nous désirons effectuer des prédictions pour l'année 1995 en utilisant le modèle $Y_t = T_t \times C_t \times S_t \times I_t$. On sait que $C_t = I_t = 1$. Il suffit donc de calculer le terme $T_t \times S_t$ pour les trimestres de 1995 :

$$\begin{aligned}\hat{y}_{1995.1} &= T_{41} \times S_{41} = 304,097 \times 1,06176 = 322,878. \\ \hat{y}_{1995.2} &= T_{42} \times S_{42} = 306,6661 \times 0,94303 = 289,195. \\ \hat{y}_{1995.3} &= T_{43} \times S_{43} = 309,2352 \times 0,94486 = 292,184. \\ \hat{y}_{1995.4} &= T_{44} \times S_{44} = 311,8043 \times 1,05035 = 327,504.\end{aligned}$$

Ainsi, les prédictions pour les ventes de 1995 sont, dans l'ordre pour chacun des trimestres, 322 878 \$, 289 195 \$, 292 184 \$ et 327 504 \$.

En calculant $Y_t = T_t \times C_t \times S_t \times I_t = T_t \times S_t$ pour chacun des points de la base de données, il est possible d'obtenir un aperçu de l'efficacité du présent modèle en comparant les deux courbes (avec un graphe séquentiel).

	index	ventes	YEAR	QUARTER	DATE_	ERR_1	SAS_1	SAF_1	STC_1	PRE_1	LICI_1	UICI_1	ci	cycle	irr
1	1	221,0	1985	1	Q1 1985	,98384	208,1443	1,06176	211,5623	201,3335	186,0796	216,5875	1,03	.	.
2	2	203,5	1985	2	Q2 1985	1,0358	215,7944	,94303	208,3422	203,9026	188,6983	219,1069	1,06	1,02	1,04
3	3	190,0	1985	3	Q3 1985	,96142	201,0879	,94486	209,1562	206,4717	191,3146	221,6288	,97	1,02	,96
4	4	225,5	1985	4	Q4 1985	1,0258	214,6905	1,05035	209,2863	209,0408	193,9285	224,1532	1,03	1,00	1,03
5	5	223,0	1986	1	Q1 1986	1,0046	210,0279	1,06176	209,0591	211,6099	196,5398	226,6800	,99	,99	1,01
6	6	190,0	1986	2	Q2 1986	,95898	201,4788	,94303	210,0965	214,1790	199,1485	229,2094	,94	,98	,96
7	7	206,0	1986	3	Q3 1986	1,0217	218,0217	,94486	213,4008	216,7481	201,7548	231,7414	1,01	,98	1,03
8	8	226,5	1986	4	Q4 1986	,99226	215,6426	1,05035	217,3247	219,3172	204,3584	234,2759	,98	1,00	,99
9	9	236,0	1987	1	Q1 1987	1,0039	222,2717	1,06176	221,4182	221,8862	206,9595	236,8130	1,00	1,00	1,00
10	10	214,0	1987	2	Q2 1987	1,0150	226,9287	,94303	223,5756	224,4553	209,5579	239,3527	1,01	1,00	1,01
11	11	210,5	1987	3	Q3 1987	,98919	222,7843	,94486	225,2200	227,0244	212,1538	241,8951	,98	,99	,99
12	12	237,0	1987	4	Q4 1987	1,0028	225,6393	1,05035	224,9996	229,5935	214,7470	244,4400	,98	,99	1,00
13	13	245,5	1988	1	Q1 1988	1,0214	231,2191	1,06176	226,3810	232,1626	217,3375	246,9876	1,00	,96	1,04
14	14	201,0	1988	2	Q2 1988	,93272	213,1433	,94303	228,5169	234,7317	219,9254	249,5379	,91	,98	,93
15	15	230,0	1988	3	Q3 1988	1,0361	243,4222	,94486	234,9359	237,3008	222,5107	252,0909	1,03	,98	1,05
16	16	254,5	1988	4	Q4 1988	1,0080	242,3004	1,05035	240,3742	239,8698	225,0932	254,6465	1,01	1,01	1,00
17	17	257,0	1989	1	Q1 1989	,99010	242,0501	1,06176	244,4707	242,4389	227,6731	257,2048	1,00	1,01	,99
18	18	238,0	1989	2	Q2 1989	1,0283	252,3787	,94303	245,4360	245,0080	230,2503	259,7658	1,03	1,00	1,03
19	19	228,0	1989	3	Q3 1989	,98643	241,3055	,94486	244,6248	247,5771	232,8247	262,3295	,97	,99	,98
20	20	255,0	1989	4	Q4 1989	,98745	242,7764	1,05035	245,8615	250,1462	235,3965	264,8959	,97	,97	1,00
21	21	260,5	1990	1	Q1 1990	,98081	245,3465	1,06176	250,1466	252,7153	237,9656	267,4649	,97	,98	,99
22	22	244,0	1990	2	Q2 1990	1,0060	258,7412	,94303	257,2020	255,2844	240,5320	270,0367	1,01	1,01	1,00
23	23	256,0	1990	3	Q3 1990	1,0262	270,9395	,94486	264,0235	257,8535	243,0957	272,6112	1,05	1,03	1,03
24	24	276,5	1990	4	Q4 1990	,98330	263,2458	1,05035	267,7154	260,4225	245,6567	275,1884	1,01	1,03	,98
25	25	291,0	1991	1	Q1 1991	1,0194	274,0723	1,06176	268,8621	262,9916	248,2150	277,7683	1,04	1,02	1,02
26	26	255,5	1991	2	Q2 1991	1,0086	270,9359	,94303	268,6364	265,5607	250,7706	280,3508	1,02	1,01	1,01
27	27	244,0	1991	3	Q3 1991	,95899	258,2392	,94486	269,2828	268,1298	253,3236	282,9360	,96	1,00	,96
28	28	291,0	1991	4	Q4 1991	1,0170	277,0507	1,05035	272,4264	270,6989	255,8738	285,5239	1,02	1,00	1,02
29	29	296,0	1992	1	Q1 1992	1,0087	278,7814	1,06176	276,3827	273,2680	258,4214	288,1145	1,02	1,01	1,01
30	30	260,0	1992	2	Q2 1992	,98406	275,7078	,94303	280,1743	275,8371	260,9664	290,7077	1,00	1,02	,98
31	31	271,5	1992	3	Q3 1992	1,0173	287,3441	,94486	282,4710	278,4061	263,5087	293,3035	1,03	1,02	1,02
32	32	299,5	1992	4	Q4 1992	1,0053	285,1433	1,05035	283,6272	280,9752	266,0485	295,9020	1,01	1,01	1,00
33	33	297,0	1993	1	Q1 1993	,98444	279,7233	1,06176	284,1446	283,5443	268,5856	298,5031	,99	1,00	,98
34	34	271,0	1993	2	Q2 1993	1,0088	287,3724	,94303	284,8710	286,1134	271,1201	301,1067	1,00	,99	1,01
35	35	270,0	1993	3	Q3 1993	1,0001	285,7565	,94486	285,7385	288,6825	273,6520	303,7129	,99	,99	1,00
36	36	300,0	1993	4	Q4 1993	,99108	285,6193	1,05035	288,1902	291,2516	276,1814	306,3217	,98	,98	1,00
37	37	306,5	1994	1	Q1 1994	,98999	288,6706	1,06176	291,5896	293,8207	278,7083	308,9330	,98	,99	,99
38	38	283,5	1994	2	Q2 1994	1,0180	300,6275	,94303	295,2991	296,3898	281,2327	311,5468	1,01	1,00	1,01
39	39	283,5	1994	3	Q3 1994	1,0075	300,0444	,94486	297,8106	298,9588	283,7545	314,1631	1,00	1,00	1,01
40	40	307,5	1994	4	Q4 1994	,98354	292,7598	1,05035	297,6578	301,5279	286,2740	316,7819	,97	.	.
41	41	1,06176	.	304,0970	288,7909	319,4031	.	.	.
42	42	,94303	.	306,6661	291,3055	322,0267	.	.	.
43	43	,94486	.	309,2352	293,8176	324,6527	.	.	.
44	44	1,05035	.	311,8043	296,3274	327,2811	.	.	.

FIG. 12.34 – Synthèse des résultats

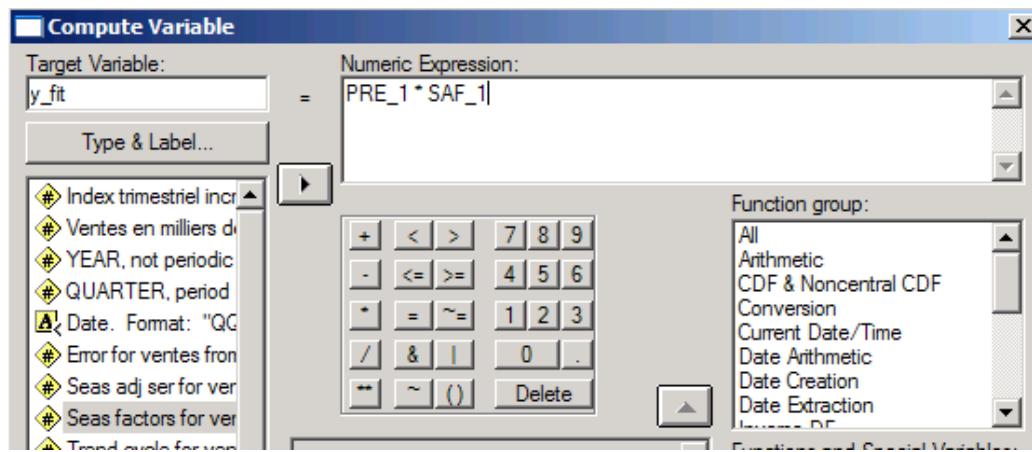


FIG. 12.35 – Création de la variable *y_fit* qui représente les prédictions du modèle de décomposition

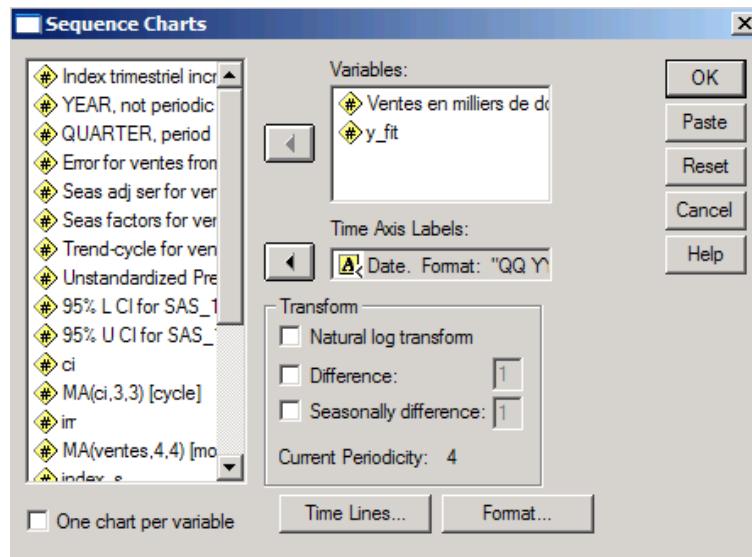


FIG. 12.36 – Création du graphe séquentiel pour comparer les ventes réelles aux prédictions du modèle de décomposition

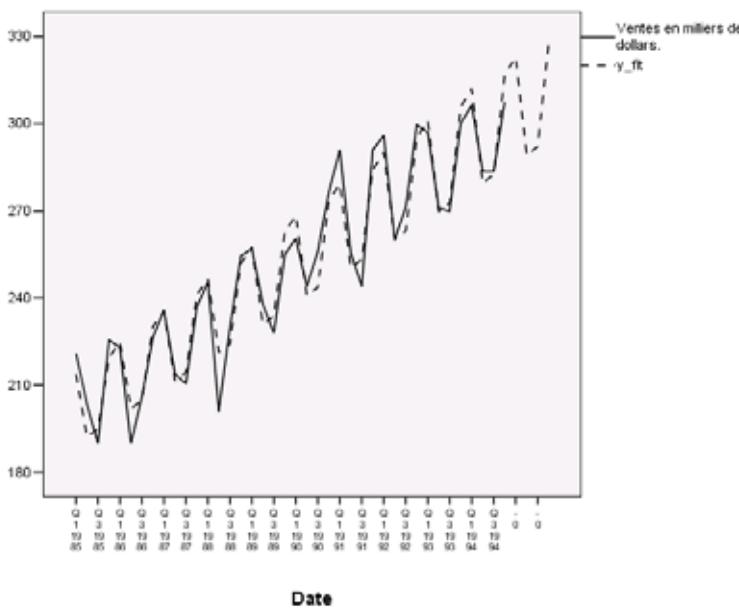


FIG. 12.37 – Comparaison des ventes réelles et des prédictions du modèle de décomposition

Le graphe de la figure 12.37 montre que le modèle semble bien performer.

Il n'existe pas vraiment de théorie sur laquelle s'appuyer pour bâtir des intervalles de prédiction à partir de ce modèle de décomposition. Il est quand même possible d'obtenir une estimation intéressante des intervalles de prédiction en utilisant les intervalles de prédiction issus de la régression qui estime la tendance (colonnes `LICI_1` et `UICI_1`). Pour ce faire, on utilise la moitié de la largeur de cet intervalle (c'est-à-dire sa précision) puis on ajoute et retranche cette valeur à la prédiction ponctuelle $Y_t = T_t \times S_t$.

Par exemple, pour la période 41, la prédiction ponctuelle du modèle de décomposition est $\hat{y}_{1995.1} = T_{41} \times S_{41} = 304,097 \times 1,06176 = 322,878$ \$. D'autre part, la régression représentant la tendance estime que les ventes moyennes devraient être comprises entre 288 791 \$ et 319 403 \$ et ce avec une probabilité de 95 %.

La moitié de la largeur de cet intervalle est $\frac{319\,403\, \$ - 288\,791\, \$}{2} = 15\,306\, \$$. Ainsi

il est possible d'estimer que le niveau de vente moyen pour le premier trimestre de 1995 devrait être compris entre $322\ 878\ $ - 15\ 306\ $ = 307\ 572\ $$ et $322\ 878\ $ + 15\ 306\ $ = 338\ 184\ $$ et ce avec une probabilité d'environ 95 %.

12.4.7 Décomposition avec E-Views

Désaisonnalisation

Il est possible de calculer les indices saisonniers ajustés à l'aide de E-Views. Ils sont calculés de façon semblable à celle décrite dans les notes de cours, excepté qu'à l'étape C la moyenne est faite sur tous les indices disponibles (au lieu de calculer une moyenne tronquée), et ensuite l'ajustement est fait à l'aide de la moyenne géométrique (pour que la multiplication des indices donne un). Pour plus de détails vous pouvez consulter l'aide de E-Views (si vous consultez le guide de l'utilisateur de la version 6, habituellement disponible en format pdf à partir du menu de l'aide, voir la page 354 du volume I).

Pour obtenir ces indices, il faut d'abord ouvrir la série à partir de laquelle on veut les calculer, puis aller dans `Proc → Seasonal Adjustment → Moving Average Methods...`. Il faut ensuite sélectionner `Ratio to moving average - Multiplicative`, et écrire le nom que l'on veut donner à la série désaisonnalisée dans la fenêtre `Adjusted series`, et si l'on veut sauvegarder la série des indices ajustés il faut la nommer dans la fenêtre `Factors`. On voit ceci dans la figure 12.38 où l'on veut calculer les ventes désaisonnalisées de la série `ABX`. On voit dans la partie droite de la figure 12.38 les quatre indices obtenus, qui sont légèrement différents de ceux que l'on avait obtenus avec SPSS (étant donné les différences mentionnées auparavant). Et la figure 12.39 montre les séries créées (ventes désaisonnalisées et indices ajustés).

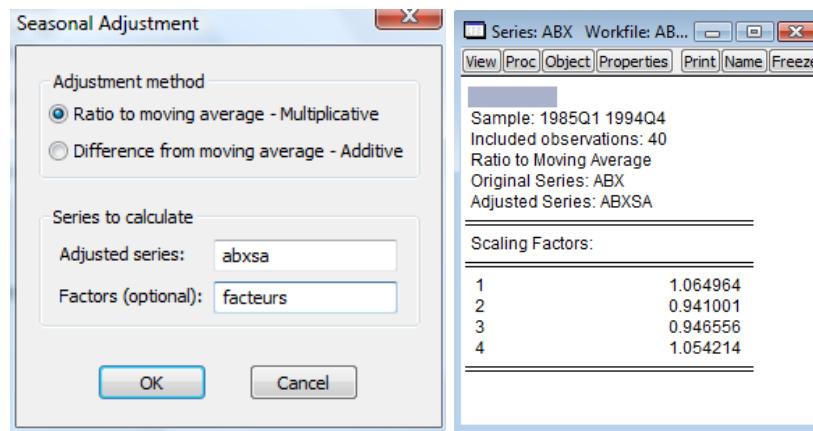


FIG. 12.38 – Calcul des indices saisonniers ajustés de la série ABX

obs	ABXSA	FACTEURS
1985Q1	207.5188	1.064964
1985Q2	216.2591	0.941001
1985Q3	200.7277	0.946556
1985Q4	213.9035	1.054214
1986Q1	209.3968	1.064964
1986Q2	201.9126	0.941001
1986Q3	217.6310	0.946556
1986Q4	214.8521	1.054214
1987Q1	221.6038	1.064964
1987Q2	227.4174	0.941001
1987Q3	222.3851	0.946556
1987Q4	224.8121	1.054214
1988Q1	230.5243	1.064964
1988Q2	213.6023	0.941001
1988Q3	242.9861	0.946556
1988Q4	241.4121	1.054214
1989Q1	241.3227	1.064964
1989Q2	252.9222	0.941001
1989Q3	240.8732	0.946556
1989Q4	241.8864	1.054214
1990Q1	244.6092	1.064964
1990Q2	259.2983	0.941001
1990Q3	270.4541	0.946556
1990Q4	262.2808	1.054214
1991Q1	273.2487	1.064964
1991Q2	271.5194	0.941001

FIG. 12.39 – Ventes désaisonnalisées et indices ajustés (ABX)

Calcul de la tendance

Suffit de prendre la série désaisonnalisée puis de générer une série qui représente la tendance de la série (avec par exemple les prévisions issues d'une régression, ou une série créée par le calcul de moyennes mobiles, ou à l'aide de tout autre modèle jugé pertinent). Si par exemple cette série s'appelle `abxtendance` et que la série des ventes désaisonnalisées s'appelle `abxsa`, alors il suffit de taper `series ci = abxsa/abxtendance` dans le haut de la fenêtre de E-Views pour obtenir la série `ci` qui représente l'effet de cycle et l'effet irrégulier.

Calcul de l'effet de cycle et de l'effet irrégulier

On procède de façon semblable à ce qui est décrit pour le calcul de la tendance, en utilisant les moyennes mobiles pour isoler l'effet de cycle, puis une fois celui-ci isolé, on peut aussi isoler l'effet irrégulier. Il faut ensuite décider si on modélise l'effet de cycle ou pas. Malheureusement, E-Views n'a pas de test équivalent au test de Runs. Il y a le `Binomial sign test`, mais celui-ci se contente de tester que la proportion des données en-dessous de la médiane et au-dessus de la médiane est de 50 %. Ce test est trop limité car ne tient pas compte du nombre d'observations d'affilée qui sont au-dessus ou en-dessous de la médiane. Autrement dit, ce n'est pas parce que 50 % des valeurs de l'effet de cycle sont en-dessous de la médiane que celui-ci est aléatoire. Il faut donc se contenter d'un examen descriptif de la série pour prendre une décision. De même pour l'effet irrégulier.

12.5 Le lissage exponentiel

Le concept du lissage exponentiel repose sur une idée simple : on suppose que les observations influencent d'autant moins la prévision qu'elles sont éloignées de la date t à laquelle on fait la prévision ; en outre, on suppose que cette influence décroît exponentiellement. Cette technique est plus efficace lorsque les composantes de la série (tendance

et effet saisonnier) changent avec le temps, et lentement.

En effet, lorsque les fluctuations sont importantes, les paramètres permettant de modéliser la série temporelle doivent être mis à jour à la fin de chaque période de temps de manière à tenir compte du mouvement des observations les plus récentes. C'est ce que permet le lissage exponentiel.

Historiquement, les méthodes de lissage exponentiel sont des méthodes intuitives ne reposant sur aucun modèle statistique formel. Cependant, des travaux ont par la suite permis la construction d'intervalles de confiance.

Plusieurs types de lissage exponentiel existent selon les caractéristiques de la série temporelle à modéliser. C'est ce que nous verrons dans les sous-sections qui suivent.

12.5.1 Lissage exponentiel simple

La méthode du lissage exponentiel simple s'applique pour les séries temporelles qui n'ont pas de tendance, mais dont la valeur moyenne peut changer avec le temps.

Supposons donc que nous avons une série temporelle dont le comportement sans tendance est décrit par le modèle suivant :

$$Y_t = \beta_0 + \epsilon_t.$$

Ce modèle exprime que la méthode des moindres carrés appliquée à une série sans tendance pour estimer les paramètres de la droite $Y_t = \beta_0 + \beta_1 t + \epsilon_t$ retourne comme résultat $\beta_1 = 0$. Dans ce cas on a $\beta_0 = \bar{y}$. Ainsi un poids égal est donné aux observations y_1, y_2, \dots, y_n dans ce modèle.

Lorsque la valeur du paramètre β_0 change avec le temps, la méthode de la pondération égale n'est pas adaptée ; elle est équivalente à la méthode de la moyenne simple.

La méthode de lissage exponentiel alloue une pondération plus importante aux valeurs les plus récentes de manière à permettre une mise à jour temporelle du paramètre β_0 . De cette manière, les lents changements de valeur de ce paramètre peuvent être saisis et incorporés dans les prédictions.

Modèle du lissage exponentiel simple

Considérons une série temporelle y_1, y_2, \dots, y_n n'ayant pas de tendance ou d'effet saisonnier mais dont le niveau moyen β_0 semble changer lentement avec le temps. Alors l'estimation $a_0(t)$ de β_0 au temps t est donnée par **l'équation de lissage**

$$\hat{y}_{t+1} = a_0(t) = \alpha y_t + (1 - \alpha)a_0(t - 1)$$

où $0 \leq \alpha \leq 1$ est la **constante de lissage** et où $a_0(t - 1)$ représente l'estimation de β_0 au temps $t - 1$. Au temps $t = 0$ on a $a_0(0) = \bar{y}$; c'est la constante d'initialisation du processus.

Lorsque $\alpha = 1$, seule la dernière observation est utilisée pour établir la prédiction ; on se retrouve alors avec le modèle naïf. Lorsque $\alpha = 0$, on se retrouve avec le modèle de la moyenne simple (la même pondération sera donnée à toutes les valeurs de la série).

Pour mieux comprendre l'équation de lissage, développons-la :

$$\begin{aligned} a_0(t) &= \alpha y_t + (1 - \alpha)a_0(t - 1) \\ &= \alpha y_t + (1 - \alpha)(\alpha y_{t-1} + (1 - \alpha)a_0(t - 2)) \\ &= \alpha y_t + \alpha(1 - \alpha)y_{t-1} + (1 - \alpha)^2 a_0(t - 2) \\ &= \cdots \\ &= \alpha y_t + \alpha(1 - \alpha)y_{t-1} + \alpha(1 - \alpha)^2 y_{t-2} + \cdots + \alpha(1 - \alpha)^{t-1} y_1 + (1 - \alpha)^t a_0(0). \end{aligned}$$

Ainsi, on voit que l'estimation $a_0(t)$ de β_0 au temps t peut être exprimée en fonction des observations y_1, y_2, \dots, y_n et de l'estimé initial $a_0(0) = \bar{y}$. Les poids mesurant la contribution des valeurs y_1, y_2, \dots, y_n sont respectivement $\alpha(1 - \alpha)^{t-1}, \dots, \alpha(1 - \alpha)^2, \alpha(1 - \alpha)$ et α . Puisque $0 \leq \alpha \leq 1$, cette pondération diminue avec « l'âge » de l'observation, c'est-à-dire que sa plus grande valeur est pour la valeur la plus récente, et diminue ensuite (de façon exponentielle, d'où le nom de la méthode).

L'estimation de la constante α s'effectue à l'aide d'une méthode de type itérative suivant le principe de minimisation de la somme des carrés des erreurs (*sum of squared error*) SSE ou son homologue, la somme des carrés moyens (*mean squared error*) MSE.

Plus précisément,

$$\text{SSE} = \sum_{t=1}^n (y_t - \hat{y}_t)^2 \quad \text{et} \quad \text{MSE} = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2.$$

Au départ, on sait que $0 \leq \alpha \leq 1$. Par conséquent, l'algorithme effectue les calculs pour comparer les erreurs des modèles $a_0(t) = \alpha y_t + (1 - \alpha)a_0(t - 1)$ en faisant varier α entre 0 et 1 par incrémentation de 0,1. Ainsi 10 modèles sont calculées, et pour chaque modèle la SSE (ou la MSE) sont calculées. Le meilleur modèle parmi les 10 est celui qui obtient la plus petite SSE (ou la plus petite MSE).

Exemple 12.5.1 La base de données `inventaire.sav` contient les inventaires journaliers de 149 jours d'un fournisseur. La variable `inv_jour` représente l'inventaire journalier.

Observons d'abord le graphe séquentiel de la variable `inv_jour` en fonction de l'index (figure 12.40). La série ne semble pas contenir d'effet de tendance (aucune pente), ni d'effet saisonnier (effet répétitif). Pour des données journalières, l'analyste peut soupçonner de tels effets aux 7, 30, 90 ou 365 jours, ce qui n'est pas apparent ici. Donc l'équation de la régression pour approximer cette série serait vraisemblablement de la forme $Y_t = \beta_0 + \epsilon_t$.

La série semble aussi varier lentement de jour en jour, c'est-à-dire que l'inventaire de deux journées qui se suivent sont semblables. Ceci semble indiquer la présence d'autocorrélation positive.

Il semble donc approprié d'appliquer le lissage exponentiel simple à cette série. Voici les commandes SPSS requises :

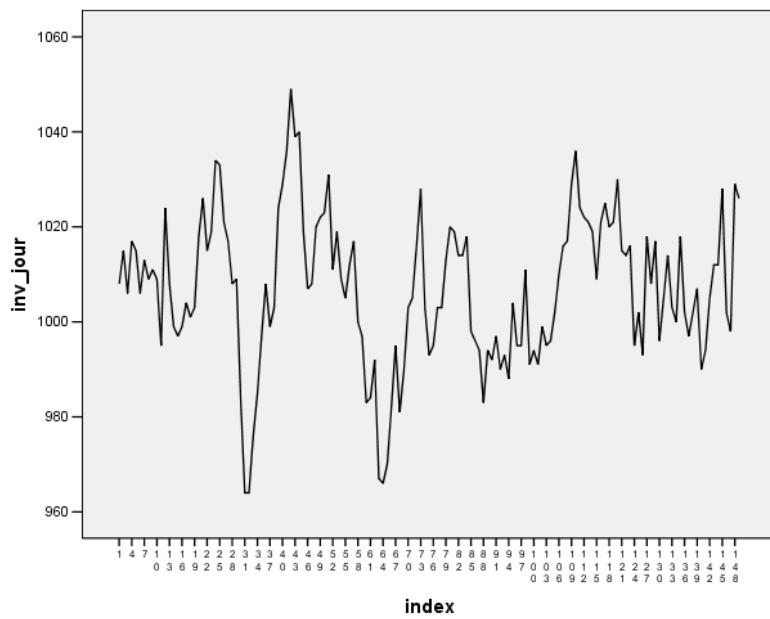


FIG. 12.40 – L'inventaire journalier

-
- Menu SPSS :
- Analyse
 - Time Series
 - Exponential Smoothing...
- Dans la fenêtre Variable(s) : inv_jour
- Dans la fenêtre Model : Simple
- Dans le bouton Parameters... : → General (Alpha)
 Grid Search : Start : 0 Stop : 1 By : ,1
- Dans le bouton Save... : → Create Variables
 Add to file
→ Predict Cases
 Predict through : Observation : 152
(pour avoir des prédictions pour les journées 150 à 152)
-

On voit tout d'abord dans la figure 12.41 que c'est un modèle simple (pas de tendance)

Model Description		
Model Name	1	MOD_2
Series	Trend	inv_jour
Simple Model	Seasonality	None
		None

Applying the model specifications from MOD_2

Initial Smoothing State	
	inv_jour
Level	1006,906

FIG. 12.41 – Le modèle et la valeur initiale

ni d'effet saisonnier), et que la valeur initiale est 1006,906 (la moyenne de la variable `inv_jour`).

Smallest Sums of Squared Errors			
Series	Model rank	Alpha (Level)	Sums of Squared Errors
inv_jour	1	,80000	17291,25
	2	,90000	17435,96
	3	,70000	17470,24
	4	1,00000	17879,20
	5	,60000	18033,12
	6	,50000	19089,29
	7	,40000	20820,59
	8	,30000	23510,67
	9	,20000	27541,48
	10	,10000	32919,01

FIG. 12.42 – Le classement des modèles

La figure 12.42 nous donne le classement des modèles en ordre décroissant d'efficacité au sens de la SSE. Ainsi le meilleur modèle est celui avec $\alpha = 0,8$ et le pire est celui avec $\alpha = 0,1$.

La figure 12.43 ne fait que présenter le meilleur modèle. La colonne `df_error` donne le nombre de degrés de liberté associés à la SSE : c'est $n - 1$ car il y a ici un paramètre

Smoothing Parameters			
Series	Alpha (Level)	Sums of Squared Errors	df error
inv_jour	,80000	17291,25	148

Shown here are the parameters with the smallest Sums of Squared Errors. These parameters are used to forecast.

FIG. 12.43 – Le meilleur modèle

$(a_0(0))$ estimé pour calculer \hat{y}_t .

Deux nouvelles variables sont créées : la variable **FIT_1** contient les prédictions \hat{y}_t issues du meilleur modèle (ici c'est avec $\alpha = 0,8$), et la variable **ERR_1** contient les résidus $e_t = y_t - \hat{y}_t$. Plus précisément, le modèle s'écrit

$$\hat{y}_{t+1} = a_0(t) = \alpha y_t + (1 - \alpha)a_0(t - 1) = 0,8y_t + (1 - 0,8)a_0(t - 1).$$

La valeur $\alpha = 0,8$ illustre à quel point la dernière valeur observée est importante dans le processus de la prédiction, ce qui sera toujours le cas lorsqu'un phénomène d'autocorrélation sera présent.

Issu du tableau de la figure 12.44, il est possible de voir par exemple que

$$\hat{y}_3 = 0,8 \cdot 1015 + 0,2 \cdot 1007,78121 = 1013,55624$$

et que

$$e_3 = 1006 - 1013,55624 = -7,55624.$$

Dans la base de données apparaissent trois prédictions pour les journées 150 à 152. Ces prédictions sont les mêmes. En effet, pour la journée 151, la prédiction s'écrit $\hat{y}_{151} = 0,8 \cdot y_{150} + (1 - 0,8)a_0(150)$; or y_{150} n'est pas connue. Elle donc estimée par l'estimation la plus récente de β_0 , qui est justement $a_0(150)$. Donc $\hat{y}_{151} = 0,8 \cdot y_{150} + (1 - 0,8)a_0(150) = 0,8 \cdot a_0(150) + (1 - 0,8)a_0(150) = 1 \cdot a_0(150)$, c'est pourquoi les prédictions sont les mêmes. Ce n'est pas une mauvaise chose, puisque, par hypothèse, β_0 change lentement dans le temps.

index	inv_jour	FIT_1	ERR_1
1	1008	1006,90604	1,09396
2	1015	1007,78121	7,21879
3	1006	1013,55624	-7,55624
4	1017	1007,51125	9,48875
5	1015	1015,10225	-10225
6	1006	1015,02045	-9,02045
7	1013	1007,80409	5,19591
8	1009	1011,96082	-2,96082
9	1011	1009,59216	1,40784
10	1009	1010,71843	-1,71843
11	995	1009,34369	-14,34369
12	1024	997,86874	26,13126
13	1008	1018,77375	-10,77375
14	999	1010,15475	-11,15475
15	997	1001,23095	-4,23095

FIG. 12.44 – Aperçu des données (nouvelles variables FIT_1 et ERR_1)

index	inv_jour	FIT_1	ERR_1
145	1028	1011,6307	16,36932
146	1002	1024,7261	-22,72614
147	998	1006,5452	-8,54523
148	1029	999,70905	29,29095
149	1026	1023,1418	2,85819
.	.	1025,4284	.
.	.	1025,4284	.
.	.	1025,4284	.

FIG. 12.45 – Les prédictions

On peut visualiser avec un graphe séquentiel les courbes de l'inventaire réel et des estimations du modèle (figure 12.46). Le résultat semble très satisfaisant.

Il est aussi possible de calculer un intervalle de confiance autour de la prédiction $\hat{y}_{t+\tau} = a_0(t)$. L'intervalle de niveau $1 - \alpha$ pour la période $t + \tau$ est donné par

$$\left[a_0(t) \pm z_{\alpha/2} s_t \sqrt{1 + (\tau - 1)\alpha^2} \right]$$

où

$$s_t = \sqrt{\frac{\text{SSE}}{t - 1}}.$$

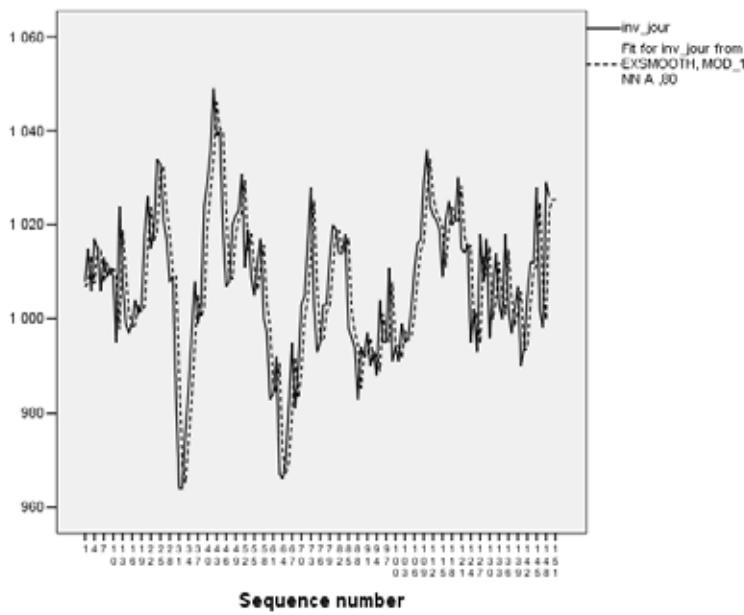


FIG. 12.46 – L’inventaire réel versus les estimés du modèle de lissage exponentiel simple

12.5.2 La méthode de Holt

La méthode de Holt est une méthode de lissage flexible qui utilise deux paramètres permettant d’ajuster les prévisions en fonction d’une tendance changeante.

Supposons donc que nous avons une série temporelle y_1, y_2, \dots, y_n dont le comportement, sans effet de saison, est décrit par le modèle suivant :

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$

où les paramètres β_0 et β_1 changent lentement avec le temps.

Le modèle de la méthode de Holt

Soient α et γ des constantes de lissage (on a donc $0 \leq \alpha \leq 1$ et $0 \leq \gamma \leq 1$). L'estimation de β_0 au temps t est donnée par

$$a_0(t) = \alpha y_t + (1 - \alpha) [a_0(t - 1) + b_1(t - 1)]$$

et l'estimation de β_1 au temps t est donnée par

$$b_1(t) = \gamma [a_0(t) - a_0(t-1)] + (1-\gamma)b_1(t-1).$$

La prédiction pour une valeur future $\hat{y}_{t+\tau}$ à partir du temps t est donnée par

$$\hat{y}_{t+\tau} = a_0(t) + b_1(t)\tau$$

et donc en particulier on a

$$\hat{y}_{t+1} = a_0(t) + b_1(t).$$

Finalement, un intervalle de confiance de niveau $1 - \alpha$ pour $\hat{y}_{t+\tau}$ peut être calculé de la manière suivante :

$$\left[\hat{y}_{t+\tau} \pm z_{\alpha/2} s_t \sqrt{1 + \sum_{j=1}^{\tau-1} \alpha^2 (1+j\gamma)^2} \right]$$

où

$$s_t = \sqrt{\frac{\text{SSE}}{t-2}}.$$

En particulier, l'intervalle pour \hat{y}_{t+1} est

$$[\hat{y}_{t+1} \pm z_{\alpha/2} s_t].$$

Le paramètre α possède la même interprétation que pour le modèle exponentiel simple. Le paramètre γ est utilisé seulement si la série démontre une tendance. Plus γ est grand, plus l'importance est donnée aux données récentes pour calculer la pente qui servira dans le calcul pour les estimations.

Dépendant des algorithmes, les valeurs initiales $a_0(0)$ et $b_1(0)$ sont calculées à partir de la moitié de la base de données ou encore à partir de la base de données entière. Ensuite, tout comme le modèle exponentiel simple, le meilleur modèle est sélectionné suivant l'idée de minimiser les quantités SSE ou MSE.

Exemple 12.5.2 La base de données `ventesholt.sav` contient une série temporelle constituée de 48 ventes mensuelles. On se servira des 36 premiers mois pour construire le modèle, et les 12 derniers serviront à valider celui-ci.

Pour sélectionner les 36 premiers mois, les commandes à effectuer sont les suivantes :

Menu SPSS :	→ Data
	→ Select Cases...
Dans la fenêtre Select :	<input checked="" type="checkbox"/> Based on time or case range
Dans le bouton Range... :	Observation : First Case 1 Last Case 36

Visualisons maintenant la série constituée des 36 premiers mois (figure 12.47). Cette série n'est pas stationnaire, le graphe montre clairement une tendance. De plus il ne semble pas il y avoir d'effet saisonnier. Il est donc approprié d'appliquer le modèle de Holt.

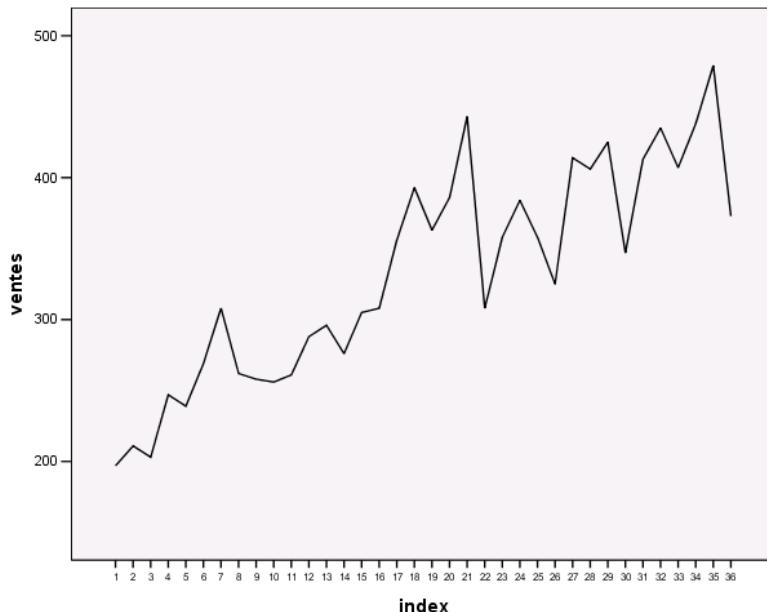


FIG. 12.47 – Les ventes des 36 premiers mois

Voici les commandes pour obtenir un modèle de Holt :

Menu SPSS :

- Analyse
- Time Series
- Exponential Smoothing...

Dans la fenêtre Variable(s) : ventes

Dans la fenêtre Model : ✓ Holt

Dans le bouton Parameters... : → General (Alpha)

- ✓ Grid Search : Start : 0 Stop : 1 By : ,1
- Trend (Gamma)
- ✓ Grid Search : Start : 0 Stop : 1 By : ,1

Dans le bouton Save... : → Create Variables

- ✓ Add to file
- Predict Cases
- ✓ Predict through : Observation : 48

Model Description		
Model Name		MOD_2
Series	1	ventes
Holt's Model	Trend	Linear
	Seasonality	None

Applying the model specifications from MOD_2

Initial Smoothing State		
	ventes	
Level	194,48571	
Trend	5,02857	

FIG. 12.48 – Le modèle et les valeurs initiales

On voit tout d'abord dans la figure 12.48 que c'est un modèle Holt, et que les valeurs initiales sont $a_0(0) = 194,48571$ et $b_1(0) = 5,02857$.

Smallest Sums of Squared Errors				
Series	Model rank	Alpha (Level)	Gamma (Trend)	Sums of Squared Errors
ventes	1	,20000	,00000	48470,66
	2	,30000	,00000	48972,08
	3	,40000	,00000	50571,90
	4	,10000	,20000	51595,41
	5	,10000	,40000	52114,69
	6	,20000	,20000	52195,28
	7	,10000	,00000	52589,05
	8	,50000	,00000	52850,82
	9	,30000	,20000	54992,51
	10	,60000	,00000	55756,83

FIG. 12.49 – Le classement des modèles

La figure 12.49 nous donne le classement des modèles en ordre décroissant d'efficacité au sens de la SSE. Ainsi le meilleur modèle est celui avec $\alpha = 0,2$ et $\gamma = 0$ (aucun ajustement temporel à la pente initiale).

Smoothing Parameters				
Series	Alpha (Level)	Gamma (Trend)	Sums of Squared Errors	df error
ventes	,20000	,00000	48470,66	34

Shown here are the parameters with the smallest Sums of Squared Errors. These parameters are used to forecast.

FIG. 12.50 – Le meilleur modèle

La figure 12.50 ne fait que présenter le meilleur modèle. La colonne `df error` donne le nombre de degrés de liberté associés à la SSE : c'est $n - 2$ car il y a ici deux paramètres ($a_0(0)$ et $b_1(0)$) estimés pour calculer \hat{y}_t .

Tout comme dans le cadre du lissage exponentiel simple, deux nouvelles variables sont créées : la variable `FIT_1` contient les prédictions \hat{y}_t issues du meilleur modèle, et la

variable `ERR_1` contient les résidus $e_t = y_t - \hat{y}_t$. Plus précisément, le modèle s'écrit

$$\hat{y}_{t+1} = a_0(t) + b_1(t)$$

avec

$$a_0(t) = 0, 2y_t + (1 - 0, 2)[a_0(t-1) + b_1(t-1)]$$

et

$$b_1(t) = 0[a_0(t) - a_0(t-1)] + (1 - 0)b_1(t-1) = b_1(t-1)$$

et donc $b_1(t)$ est constant et égal à $b_1(0) = 5,02857$.

index	ventes	FIT_1	ERR_1
1	197	199,51429	-2,51429
2	211	204,04000	6,96000
3	203	210,46057	-7,46057
4	247	213,99703	33,00297
5	239	225,62619	13,37381
6	269	233,32953	35,67047
7	308	245,49219	62,50781
8	262	263,02233	-1,02233
9	258	267,84643	-9,84643
10	256	270,90572	-14,90572
11	261	272,95315	-11,95315
12	288	275,59109	12,40891
13	296	283,10144	12,89856
14	276	290,70972	-14,70972
15	305	292,79635	12,20365

FIG. 12.51 – Aperçu des données (nouvelles variables FIT_1 et ERR_1)

On peut par exemple faire le calcul pour \hat{y}_2 . On a

$$\begin{aligned} a_0(1) &= 0, 2 \cdot y_1 + (1 - 0, 2)[a_0(0) + b_1(0)] \\ &= 0, 2 \cdot 197 + 0, 8[194,48571 + 5,02857] \\ &= 199,011424. \\ b_1(1) &= 5,02857. \end{aligned}$$

Et donc

$$\hat{y}_2 = a_0(1) + b_1(1) = 199,011424 + 5,02857 = 204,04.$$

	index	ventes	FIT_1	ERR_1
37	37	467	435,24788	31,75212
38	38	500	440,27645	59,72355
39	39	535	445,30502	89,69498
40	40	525	450,33359	74,66641
41	41	449	455,36216	-6,36216
42	42	557	460,39073	96,60927
43	43	543	465,41931	77,58069
44	44	433	470,44788	-37,44788
45	45	475	475,47645	-47645
46	46	592	480,50502	111,49498
47	47	548	485,53359	62,46641
48	48	520	490,56216	29,43784

FIG. 12.52 – Les prédictions

Afin d'évaluer la qualité du modèle, il est important d'observer attentivement les prédictions pour les périodes 37 à 48 (figure 12.52), périodes qui n'ont pas servi à l'élaboration du modèle. Les ratures associées à ces périodes dans le fichier SPSS sont apparues suite aux commandes effectuées dans **Select Cases...** (ceci indique que ces lignes ne sont pas considérées pour les analyses).

Il serait bien de visualiser la différence entre les ventes réelles et les estimations pour les données 37 à 48. On peut faire ceci avec un graphe séquentiel, mais il faut tout d'abord indiquer à SPSS que l'on veut que toutes les données soient sélectionnées :

Menu SPSS : → Data
→ Select Cases...
Dans la fenêtre Select : All Cases

On peut alors visualiser avec un graphe séquentiel les courbes des ventes réelles et des estimations du modèle (figure 12.53).

À l'aide du graphe, il est possible de voir que le modèle n'est pas en mesure de faire de bonnes prédictions. Compte tenu que le paramètre $\gamma = 0$ signifie qu'aucun ajustement temporel n'est effectué sur la pente initiale (calculée à partir des 36 premiers mois), le modèle n'a pas été en mesure de s'adapter correctement à la situation de changement de

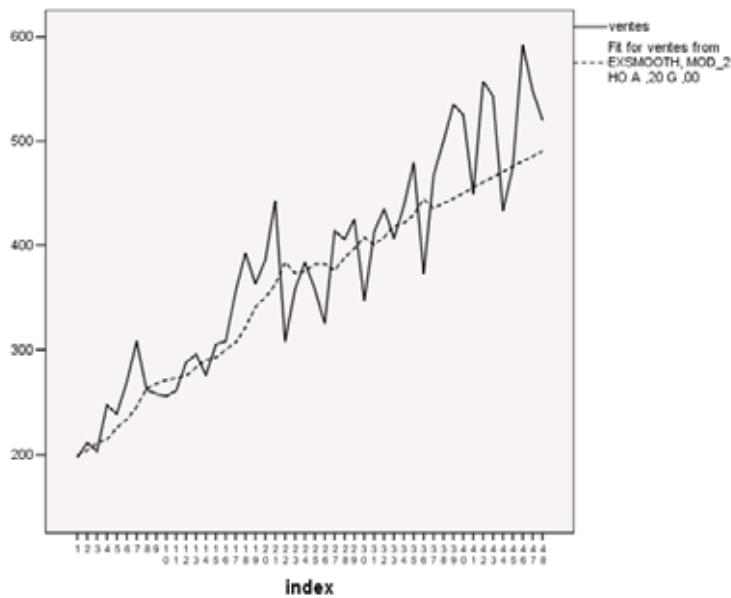


FIG. 12.53 – Les ventes réelles versus les estimés du modèle de Holt

la pente, d'où l'importance de se laisser une plage d'essai. Il faut bien comprendre que le modèle s'adapte si les paramètres lui en laissent la chance. Le modèle doit être revu avec toutes les données ou un autre modèle issu d'une autre technique doit être créé.

12.5.3 La méthode de Winters

La méthode de Winters est une approche de lissage exponentiel en mesure de tenir compte de la tendance et de l'effet saisonnier. Tout comme les autres méthodes de lissage exponentiel, aucun fondement statistique n'est à la base de cette méthode. Cependant, cette approche est généralement considérée comme l'une des meilleures méthodes pour traiter des séries temporelles obéissant à un modèle du type

$$Y_t = (\beta_0 + \beta_1 t)S_t + \epsilon_t$$

dans lequel les paramètres changent lentement avec le temps. Aussi, puisque l'effet de saison est modélisé de façon multiplicative, il est supposé que cet effet croît ou décroît

lentement avec le temps (si ce n'est pas le cas, il serait plus approprié de considérer le modèle additif de Winters, mais seul le modèle multiplicatif sera présenté dans le cadre de ce cours).

Le modèle de la méthode de Winters

Supposons que nous avons une série temporelle y_1, y_2, \dots, y_n dont le comportement est décrit par le modèle suivant :

$$Y_t = (\beta_0 + \beta_1 t)S_t + \epsilon_t$$

où les paramètres β_0 , β_1 et S_t changent lentement avec le temps.

Soient α , γ et δ des constantes de lissage (on a donc $0 \leq \alpha \leq 1$, $0 \leq \gamma \leq 1$ et $0 \leq \delta \leq 1$).

L'estimation de β_0 au temps t est donnée par

$$a_0(t) = \alpha \frac{y_t}{S_{(t-L)}} + (1 - \alpha)[a_0(t-1) + b_1(t-1)]$$

où

- $S_{(t-L)}$ est l'estimé le plus récent de l'effet de saison pour la saison correspondant à cette période ; L représente le nombre de « saisons » dans une année (donc $L = 12$ pour des données mensuelles, et $L = 4$ pour des données trimestrielles).
- $\frac{y_t}{S_{(t-L)}}$ est la donnée désaisonnalisée au temps t .

L'estimation de β_1 au temps t est donnée par

$$b_1(t) = \gamma [a_0(t) - a_0(t-1)] + (1 - \gamma)b_1(t-1).$$

La nouvelle estimation de S_t au temps t est donnée par

$$S_{(t)} = \delta \frac{y_t}{a_0(t)} + (1 - \delta)S_{(t-L)}.$$

La prédition pour une valeur future $\hat{y}_{t+\tau}$ à partir du temps t est donnée par

$$\hat{y}_{t+\tau} = [a_0(t) + b_1(t)\tau] S_{(t+\tau-L)}$$

et donc en particulier on a

$$\hat{y}_{t+1} = [a_0(t) + b_1(t)] S_{(t+1-L)}.$$

Finalement, une approximation d'un intervalle de confiance de niveau $1 - \alpha$ pour $\hat{y}_{t+\tau}$ peut être calculé de la manière suivante :

$$[\hat{y}_{t+\tau} \pm z_{\alpha/2} s_t(\sqrt{c_\tau}) S_{(t+\tau-L)}]$$

où

$$c_1 = (a_0(t) + b_1(t))^2,$$

$$c_\tau = \left[\sum_{j=1}^{\tau-1} \alpha^2 (1 + [\tau - j]\gamma)^2 (a_0(t) + jb_1(t))^2 \right] + (a_0(t) + \tau b_1(t))^2 \quad (2 \leq \tau \leq L)$$

et

$$s_t = \sqrt{\frac{\sum_{i=1}^t \left[\frac{y_t - \hat{y}_t}{\hat{y}_t} \right]^2}{t-3}}.$$

Les valeurs des paramètres α et γ ont la même interprétation que pour le modèle de Holt. Dans le modèle de Winters, le paramètre δ s'interprète ainsi : plus δ est grand, plus l'importance est donnée aux saisons récentes pour actualiser l'estimation de l'effet saisonnier.

Comme les autres méthodes présentées dans le cadre de cette section, des valeurs initiales sont calculées pour initialiser le processus. Ces méthodes utilisent la moitié de la base ou la base entière. Pour de plus amples détails sur la détermination de ces valeurs, voir Bowerman et O'Connell (1993).

Exemple 12.5.3 Reprenons la série des ventes trimestrielles de la compagnie ABX. Nous savons déjà que cette série a un effet saisonnier et une tendance ; voyons comment nous pouvons modéliser ces ventes avec le modèle de Winters.

Si ce n'est déjà fait, il faut définir les dates à SPSS (menu `Define dates...`).

Les commandes pour obtenir le modèle multiplicatif de Winters ainsi que des prédictions pour l'année 1995 sont les suivantes :

Menu SPSS :	→ Analyse
	→ Time Series
	→ Exponential Smoothing...
Dans la fenêtre Variable(s) :	ventes
Dans la fenêtre Seasonal Factors :	ne rien spécifier
Dans la fenêtre Model :	✓ Winters
Dans le bouton Parameters... :	→ General (Alpha) ✓ Grid Search : Start : 0 Stop : 1 By : ,1 → Trend (Gamma) ✓ Grid Search : Start : 0 Stop : 1 By : ,1 → Seasonal (Delta) ✓ Grid Search : Start : 0 Stop : 1 By : ,1
Dans le bouton Save... :	→ Create Variables ✓ Add to file → Predict Cases ✓ Predict through : Year : 1995 Quarter : 4

On voit tout d'abord dans la figure 12.54 que c'est un modèle de Winters multiplicatif, et que les valeurs initiales sont $a_0(0) = 205,26387$, $b_1(0) = 2,36806$, $S_1 = 106,15760$, $S_2 = 94,23881$, $S_3 = 94,60814$ et $S_4 = 104,99546$.

Comme pour le modèle multiplicatif de décomposition, pour des données trimestrielles, chaque trimestre représente $\frac{1}{4}$ des ventes totales de l'année. On voit que le premier trimestre obtient un indice initial saisonnier de 106,1576 %, ce qui s'interprète ainsi :

Model Description			
Model Name		MOD_1	
Series	1	ventes	
Winters's Multiplicative Model	Trend Seasonality	Linear Multiplicative	
Length of Seasonal Period			4

Applying the model specifications from MOD_1

Initial Smoothing State			
			ventes
Seasonal Indices	1 2 3 4	106,15760 94,23881 94,60814 104,99546	
Level		205,26387	
Trend		2,36806	

FIG. 12.54 – Le modèle et les valeurs initiales

initialement, le premier trimestre obtient un niveau de vente de 6,1576 % supérieur au $\frac{1}{4}$ des ventes annuelles centrées en ce trimestre. L'interprétation des autres indices initiaux s'effectue de la même manière.

La figure 12.55 nous donne le classement des modèles en ordre décroissant d'efficacité au sens de la SSE. Ainsi le meilleur modèle est celui avec $\alpha = 0,2$, $\gamma = 0$ et $\delta = 0$ (aucun ajustement temporel à la pente et aux indices saisonniers).

La valeur $\alpha = 0,2$ indique qu'un peu plus de poids est donné aux données les plus récentes afin d'ajuster β_0 . Ceci se répercute dans une légère élévation de la droite, ce qui ne change pas la pente qui varie indépendamment. Les valeurs $\gamma = 0$ et $\delta = 0$ illustrent que la pente et l'effet saisonnier calculés initialement ne varieront pas. Cela ne signifie pas, par exemple, que la pente est nulle ou que l'effet saisonnier est marginal. Cela signifie que la pente initiale (2,36806) ainsi que l'effet saisonnier (quatre indices initiaux) sont présents et demeurent simplement constants tout au long de la série.

Si on avait eu $\delta > 0$, les indices saisonnier auraient évolué avec le temps. Cependant les nouveaux indices calculés ne sont pas sauvegardés dans SPSS, et il faut donc les calculer

Smallest Sums of Squared Errors					
Series	Model rank	Alpha (Level)	Gamma (Trend)	Delta (Season)	Sums of Squared Errors
ventes	1	,20000	,00000	,00000	2092,479
	2	,10000	,00000	,00000	2132,455
	3	,30000	,00000	,00000	2140,055
	4	,40000	,00000	,00000	2237,546
	5	,20000	,10000	,00000	2260,964
	6	,30000	,10000	,00000	2264,519
	7	,10000	1,00000	,00000	2283,984
	8	,20000	,30000	,00000	2289,399
	9	,20000	,40000	,00000	2292,281
	10	,10000	,90000	,00000	2292,365

FIG. 12.55 – Le classement des modèles

à la main lorsque nécessaire. On remarque que les indices trouvés dans cet exemple sont très semblables à ceux trouvés avec la méthode de décomposition.

Smoothing Parameters					
Series	Alpha (Level)	Gamma (Trend)	Delta (Season)	Sums of Squared Errors	df error
ventes	,20000	,00000	,00000	2092,479	35

Shown here are the parameters with the smallest Sums of Squared Errors.
These parameters are used to forecast.

FIG. 12.56 – Le meilleur modèle

La figure 12.56 présente simplement le meilleur modèle.

Tout comme dans le cadre du lissage exponentiel simple et du modèle de Holt, deux nouvelles variables sont créées : la variable **FIT_1** contient les prédictions \hat{y}_t issues du meilleur modèle, et la variable **ERR_1** contient les résidus $e_t = y_t - \hat{y}_t$. Plus précisément, le modèle s'écrit

$$\hat{y}_{t+1} = [a_0(t) + b_1(t)] S_{(t+1-L)}$$

avec

$$a_0(t) = 0,2 \frac{y_t}{S_{(t-L)}} + (1 - 0,2) [a_0(t-1) + b_1(t-1)],$$

$$b_1(t) = 0 [a_0(t) - a_0(t-1)] + (1 - 0)b_1(t-1) = b_1(t-1)$$

(donc $b_1(t)$ est constant et égal à $b_1(0) = 2,36806$) et finalement

$$S_{(t)} = 0 \frac{y_t}{a_0(t)} + (1 - 0)S_{(t-L)}.$$

Donc tel que mentionné précédemment les indices saisonniers sont constants.

index	ventes	YEAR	QUARTER	DATE	FIT_1	ERR_1
1	221,00	1985	1	Q1 1985	220,41707	,58291
2	203,50	1985	2	Q2 1985	198,00497	5,49501
3	190,00	1985	3	Q3 1985	202,12466	-12,1247
4	225,50	1985	4	Q4 1985	224,11172	1,38826
5	223,00	1986	1	Q1 1986	229,38690	-6,38691
6	190,00	1986	2	Q2 1986	204,73026	-14,7303
7	206,00	1986	3	Q3 1986	204,81539	1,18460
8	226,50	1986	4	Q4 1986	230,05198	-3,55200
9	236,00	1987	1	Q1 1987	234,39392	1,60606
10	214,00	1987	2	Q2 1987	210,59423	3,40575
11	210,50	1987	3	Q3 1987	214,34376	-3,84378
12	237,00	1987	4	Q4 1987	239,51042	-2,51044
13	245,50	1988	1	Q1 1988	244,16767	1,33232
14	201,00	1988	2	Q2 1988	219,22203	-18,2220
15	230,00	1988	3	Q3 1988	218,66287	11,33712
16	254,50	1988	4	Q4 1988	247,67326	6,82672
17	257,00	1989	1	Q1 1989	254,30896	2,69101

FIG. 12.57 – Aperçu des données (nouvelles variables FIT_1 et ERR_1)

index	ventes	YEAR	QUARTER	DATE	FIT_1	ERR_1
39	283,50	1994	3	Q3 1994	281,75298	1,74699
40	307,50	1994	4	Q4 1994	315,56162	-8,06165
.	.	1995	1	Q1 1995	319,93810	.
.	.	1995	2	Q2 1995	286,24884	.
.	.	1995	3	Q3 1995	289,61105	.
.	.	1995	4	Q4 1995	323,89469	.

FIG. 12.58 – Les prédictions

Pour donner un exemple de calcul, établissons la prédition pour le premier trimestre

de 1995. On a

$$\begin{aligned}
 \hat{y}_{41} &= [a_0(40) + b_1(40)] S_{(1)} \\
 &= \left[\left(0, 2 \frac{\hat{y}_{40}}{S_{(4)}} + 0,8 [a_0(39) + b_1(39)] \right) + 2,36806 \right] \cdot 1,0615760 \\
 &= \left[\left(0, 2 \frac{307,50}{1,0499546} + 0,8 [a_0(39) + b_1(39)] \right) + 2,36806 \right] \cdot 1,0615760 \\
 &= \left[\left(0, 2 \frac{307,50}{1,0499546} + 0,8 [300, 54787^*] \right) + 2,36806 \right] \cdot 1,0615760 \\
 &= 319,9381.
 \end{aligned}$$

*En effet, $\hat{y}_{40} = [a_0(39) + b_1(39)] S_{(4)}$

Donc $[a_0(39) + b_1(39)] = \frac{\hat{y}_{40}}{S_{(4)}} = \frac{315,56162}{1,0499546} = 300,54787.$

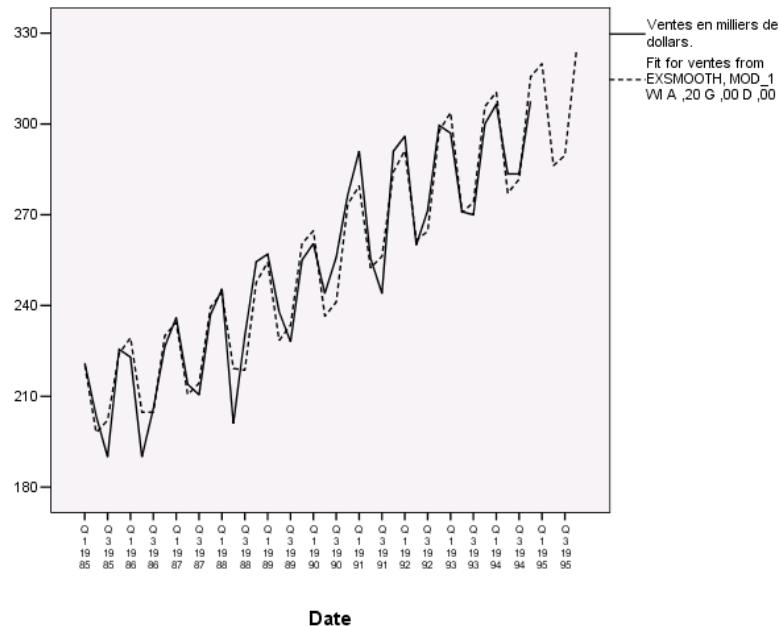


FIG. 12.59 – Les ventes réelles versus les estimés du modèle de Winters

La figure 12.59 nous montre que le modèle s'ajuste assez bien aux données.

12.5.4 Lissage exponentiel avec E-Views

Lissage exponentiel simple

Pour estimer un modèle de lissage exponentiel simple avec E-Views, il faut d'abord ouvrir la série à modéliser, puis aller dans **Proc → Exponential Smoothing...**. Il faut ensuite sélectionner **Single**, puis si on veut changer le nom qui sera donné à la série, il faut le taper dans la fenêtre **Smoothed series** (E-Views en donne un par défaut). Dans la fenêtre **Alpha** on laisse le E qui signifie que la constante de lissage sera estimée par E-Views (si on veut imposer une valeur à cette constante il suffit de l'inscrire dans cette fenêtre). On peut définir l'échantillon sur lequel on veut que le modèle soit estimé dans la fenêtre **Estimation sample**.

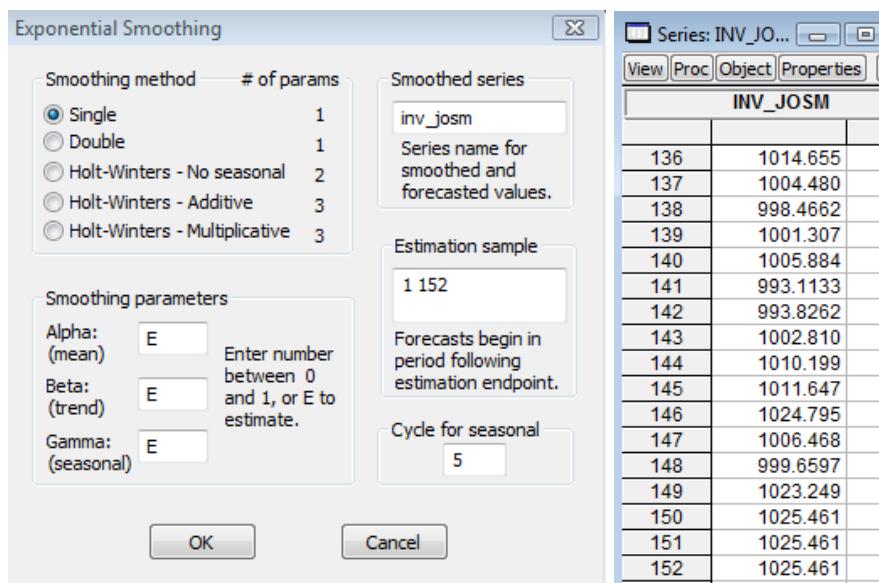


FIG. 12.60 – Lissage exponentiel simple (exemple 12.5.1)

La figure 12.60 présente la fenêtre d'options pour faire un modèle de lissage exponentiel (dans ce cas-ci on voulait un lissage exponentiel simple), ainsi que la série obtenue (données de l'exemple 12.5.1). L'échantillon avait été défini pour 152 observations, ce qui fait que le lissage a automatiquement produit des prévisions pour les périodes 150 à 152 (dans la série originale les observations s'arrêtent à la ligne 149). On voit que celles-

ci coïncident (à quelques décimales près) aux estimations que l'on avait obtenues avec SPSS. Il est à noter que la valeur initiale ($a_0(0)$) qui est la moyenne de toutes les données dans SPSS n'est pas la même dans E-Views (il prend la moyenne des $(n + 1)/2$ données initiales de l'échantillon, où n est la taille de l'échantillon).

La figure 12.61 nous montre les détails du modèle. Ainsi on voit que la constante de lissage α a été estimée à 0,8040 (avec SPSS on avait obtenu 0,8 lorsqu'on demandait une précision de 0,1).

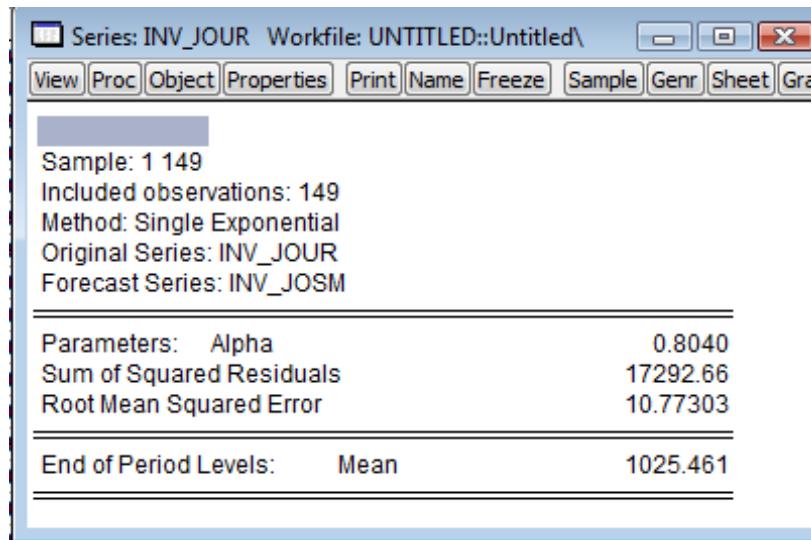


FIG. 12.61 – Résultats des estimations pour le lissage

Modèle de Holt

Il suffit de suivre la même procédure que celle décrite pour le lissage simple, mais en sélectionnant le modèle **Holt-Winters - No seasonal**. Il est à noter que les résultats obtenus seront différents de ceux de SPSS car les valeurs initiales ne sont pas calculées de la même façon.

Modèle de Winters

Il suffit de suivre la même procédure que celle décrite pour le lissage simple, mais en sélectionnant le modèle **Holt-Winters - Multiplicative**. Il est à noter que les résultats obtenus seront différents de ceux de SPSS car les valeurs initiales ne sont pas calculées de la même façon.

12.5.5 La tendance

Dans le cadre de cette section, la tendance était supposée linéaire. Cependant, lorsque la croissance de la variable étudiée est moins rapide que la progression linéaire, il est possible d'ajuster un modèle de lissage pour une tendance dite « *damped* ». Inversement, lorsque la progression est définitivement plus rapide que la progression linéaire, il est possible d'ajuster le modèle pour une tendance dite exponentielle. La figure 12.62 illustre les différences entre les trois types de tendances.

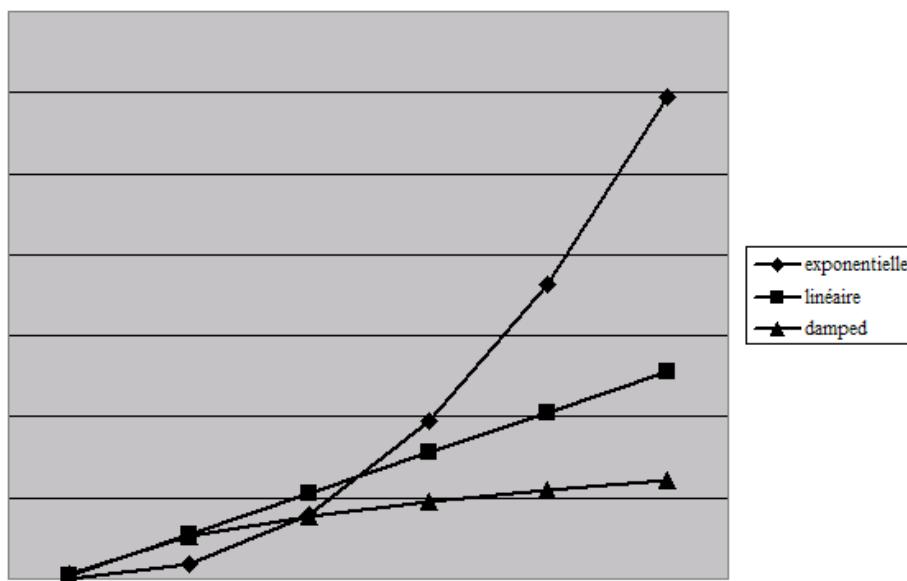


FIG. 12.62 – Les types de tendance

Pour une tendance *damped*, SPSS remplace la constante de lissage γ par une autre

constante ϕ . Cette constante provient d'un modèle où l'analyste estime que le phénomène à modéliser contient une pente ET que la croissance est plus lente que la croissance linéaire. Comme pour les autres constantes, lorsque ϕ est élevée, le modèle réagira rapidement à toute indication que la courbe s'estompe. Pour l'approche exponentielle, ce sont les mêmes constantes α , γ et δ qui sont utilisées. Ces deux modèles sont disponibles via le bouton **Custom**....

12.6 Les tendances non linéaires

Rappelons que dans le modèle de décomposition $Y_t = T_t \times S_t \times C_t \times I_t$, le terme T_t peut ne pas être linéaire. Par exemple, la croissance d'un produit évolue davantage suivant une courbe en forme de « S ». D'autres courbes sont utilisées comme les courbes exponentielles, logarithmiques, logistiques, de Weibull, de Gompertz, etc.

Il est possible d'estimer ces courbes à l'aide de SPSS. Pour ce faire, l'analyste doit d'abord reconnaître la forme de la courbe à modéliser. Ensuite, il doit trouver l'équation algébrique associée ainsi que les paramètres du modèle théorique qu'il désire utiliser. Finalement, il doit estimer les paramètres du modèle de manière à ajuster de manière optimale la courbe aux points expérimentaux.

Exemple 12.6.1 Voici les données concernant une étude de la tolérance des consommateurs d'un produit donné par rapport à l'augmentation du prix. L'indice de tolérance varie entre 0 et 100 points.

id	augmentation	tolerance
1	,5	86,30
2	1,5	85,79
3	3,5	83,33
4	6,0	79,11
5	9,0	67,58
6	13,0	47,45
7	18,0	33,25
8	23,0	24,14
9	28,0	16,99
10	35,5	11,98
11	45,5	9,27
12	55,5	7,22
13	68,0	5,68
14	88,0	3,89

FIG. 12.63 – Les données

La première étape consiste à observer le graphe de ces données (tolérance en fonction de l'augmentation du prix).

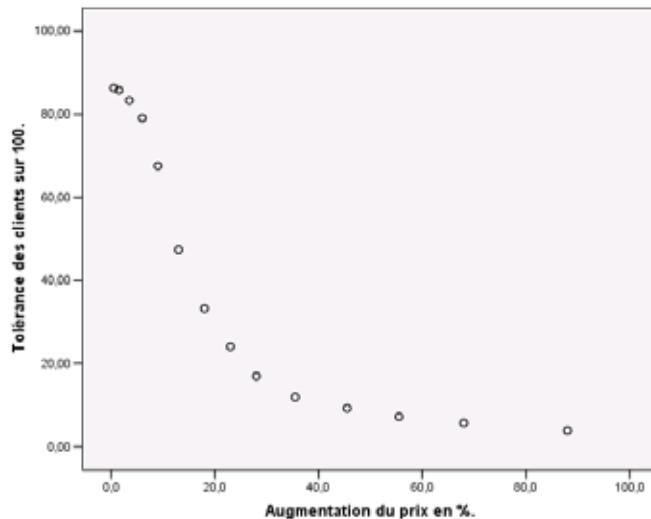


FIG. 12.64 – Le graphe

L'analyste observe (sans surprise) que plus les prix augmentent, plus la tolérance diminue, mais cette relation n'est pas linéaire. Elle présente la forme d'un « S » et est strictement décroissante. Ce type de décroissance (ou de croissance lorsque le « S » se présente dans l'autre sens) porte le nom de courbe de croissance et est connu depuis plus d'un siècle. Plusieurs modèles existent pour modéliser différents types de courbes et ce chapitre en présente quelques uns ainsi que la méthode permettant l'estimation de leurs paramètres.

Dans le cadre de cet exemple, afin de présenter la méthodologie, le modèle de survie de Weibull sera ajusté aux données.

Le modèle de survie de Weibull(α, λ) est un modèle paramétrique. En d'autres termes, il faut estimer les paramètres α et λ pour pouvoir ajuster la courbe aux données. La formule algébrique de ce modèle est la suivante :

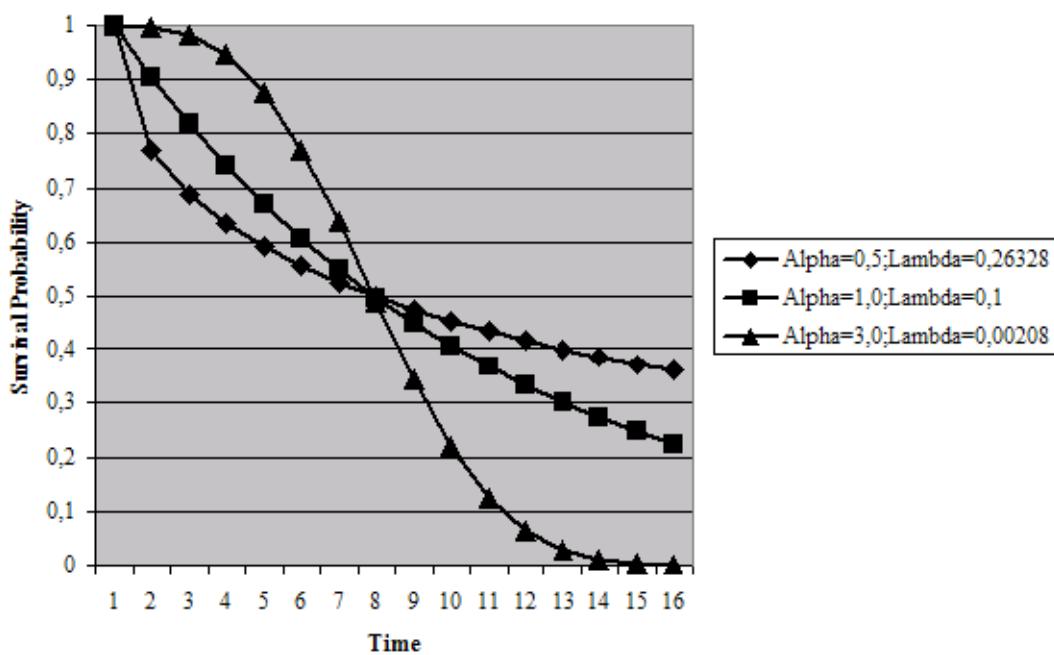
$$S(t) = e^{-\lambda t^\alpha} \text{ pour } \lambda > 0, \alpha > 0.$$

Les paramètres α et λ ajustent respectivement la forme de la courbe et l'échelle du modèle. $S(t) = P(T > t)$ est une probabilité et représente dans les études de survie la probabilité qu'un patient survive plus de t mois à sa maladie. Cette courbe est clairement décroissante comme l'illustre la figure ci-dessous présentant trois courbes ajustées avec différents valeurs pour les paramètres α et λ .

Dans le cadre de cet exemple, pour appliquer le modèle de Weibull, on doit ramener les données entre 0 et 1. Pour ce faire un facteur homothétique γ est ajouté au modèle initial de la manière suivante :

$$\frac{\text{tolerance}}{\gamma} = e^{-\lambda(\text{augmentation})^\alpha} \Leftrightarrow \text{tolerance} = \gamma e^{-\lambda(\text{augmentation})^\alpha}.$$

On doit donc estimer α , λ et γ afin d'ajuster la courbe aux données. Les commandes sont les suivantes :

FIG. 12.65 – Le modèle de Weibull avec différentes valeurs pour α et λ

Menu SPSS :	→ Analyse
	→ Regression
	→ Nonlinear...
Dans la fenêtre Dependent :	tolerance
Dans la fenêtre Model Expression :	b1 * EXP(-b2 * augmentation ** b3)
Dans le bouton Parameters... :	Name : b1 Starting Value : 1 Add Name : b2 Starting Value : 1 Add Name : b3 Starting Value : 1 Add
Dans le bouton Constraints... :	<input checked="" type="checkbox"/> Define parameter constraint : b2 \geq 0 Add b3 \geq 0 Add
Dans le bouton Save... :	<input checked="" type="checkbox"/> Predicted values <input checked="" type="checkbox"/> Residuals

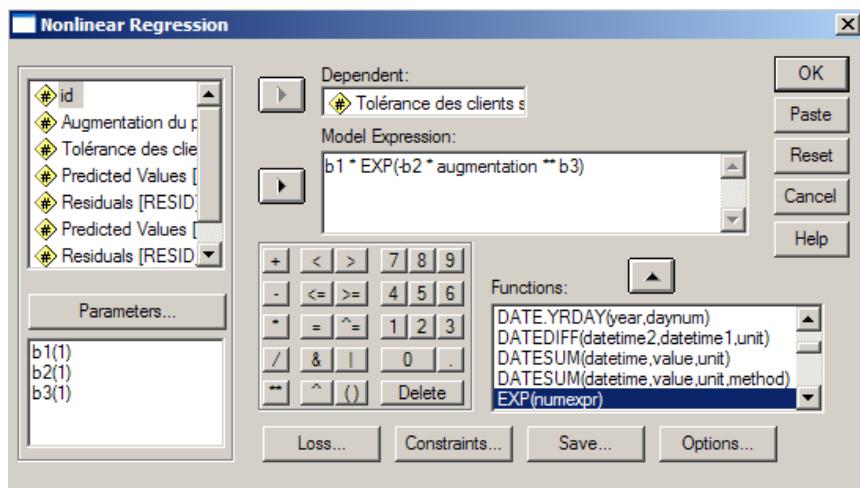


FIG. 12.66 – Le modèle de Weibull

L'estimation s'effectue suivant un algorithme itératif de type non linéaire visant à minimiser $\text{SSE} = \sum_i (y_i - \hat{y}_i)^2$.

La figure 12.67 présente les itérations. Le processus a convergé après 28 itérations. La dernière ligne présente la solution optimale. La somme des carrés des erreurs est alors égale à 198,349, et on a $b1 = \gamma = 90,186$, $b2 = \lambda = 0,018$ et $b3 = \alpha = 1,344$. Donc le modèle est

$$\hat{y}_{\text{tolerance}} = 90,186 e^{-0,018 x_{\text{aug}}^{1,344}}.$$

Iteration Number ^a	Residual Sum of Squares	Parameter		
		b1	b2	b3
		1,000	,353	,353
0.3	36986,397	1,000	1,000	1,000
1.2	4232,868	96,430	,353	,353
2.1	1700,263	127,735	,353	,485
3.1	1444,179	135,216	,333	,480
4.1	1198,368	121,469	,223	,566
5.1	820,356	102,821	,141	,688
6.1	815,596	102,646	,139	,690
7.2	695,562	111,716	,104	,820
8.1	600,723	106,155	,088	,839
9.1	514,318	102,217	,094	,819
10.1	462,918	100,537	,084	,854
11.1	397,461	98,036	,063	,929
12.1	324,965	97,473	,055	,984
13.1	301,108	95,738	,045	1,035
14.1	246,672	93,230	,037	1,107
15.1	234,874	92,834	,032	1,152
16.1	218,107	93,616	,027	1,229
17.1	210,587	92,636	,024	1,258
18.1	204,456	90,720	,023	1,260
19.1	201,981	90,893	,022	1,279
20.1	200,399	91,115	,021	1,294
21.1	199,444	90,662	,019	1,322
22.1	198,500	90,424	,019	1,330
23.1	198,363	90,211	,018	1,341
24.1	198,350	90,189	,018	1,343
25.1	198,349	90,185	,018	1,344
26.1	198,349	90,186	,018	1,344
27.1	198,349	90,186	,018	1,344
28.1	198,349	90,186	,018	1,344

Derivatives are calculated numerically.

a. Major iteration number is displayed to the left of the decimal, and minor iteration number is to the right of the decimal.

b. Run stopped after 28 iterations. Optimal solution is found.

FIG. 12.67 – Les itérations

Parameter Estimates				
Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
b1	90,186	3,281	82,964	97,408
b2	,018	,009	-,001	,038
b3	1,344	,151	1,011	1,677

FIG. 12.68 – Les paramètres

La figure 12.68 présente la solution optimale avec plus de détails, dont les intervalles de confiance pour chacun des paramètres.

ANOVA ^a			
Source	Sum of Squares	df	Mean Squares
Regression	36936,043	3	12312,014
Residual	198,349	11	18,032
Uncorrected Total	37134,392	14	
Corrected Total	14575,712	13	

Dependent variable: tolerance

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = ,986.

FIG. 12.69 – Table ANOVA

Dans le bas de la sortie 12.69 est donné un pseudo r^2 de 0,986, ce qui montre un peu comme dans une régression que 98,6 % de la variation totale de la tolérance est expliquée par le modèle.

À partir du modèle

$$\hat{y}_{\text{tolerance}} = 90,186e^{-0,018x_{\text{aug}}^{1,344}}.$$

il est maintenant possible d'estimer la tolérance pour tout % d'augmentation du prix et vice versa. Par exemple, si le prix augmente de 6 %, la tolérance s'estime ainsi :

id	augmentation	tolerance	PRED_	RESID
1	,5	86,30	89,54	-3,24
2	1,5	85,79	87,40	-1,61
3	3,5	83,33	81,77	1,56
4	6,0	79,11	73,67	5,44
5	9,0	67,58	63,63	3,95
6	13,0	47,45	50,91	-3,46
7	18,0	33,25	37,20	-3,95
8	23,0	24,14	26,33	-2,19
9	28,0	16,99	18,14	-1,15
10	35,5	11,98	9,93	2,05
11	45,5	9,27	4,15	5,12
12	55,5	7,22	1,61	5,61
13	68,0	5,68	,46	5,22
14	88,0	3,89	,05	3,84

FIG. 12.70 – Les prédictions

$$\hat{y}_{\text{tolerance}} = 90,186e^{-0,018 \cdot 6^{1,344}} = 73,835.$$

Les estimations se retrouvent dans la colonne PRED_ (figure 12.70). La différence entre l'estimation calculée pour une augmentation de 6 % et celle que l'on retrouve dans la base de donnée provient du fait que les paramètres ont été arrondis pour le calcul à la main.

Il est possible de comparer les estimations du modèle avec les valeurs observées en produisant un graphe séquentiel (figures 12.71 et 12.72). On voit que l'ajustement est très bon.

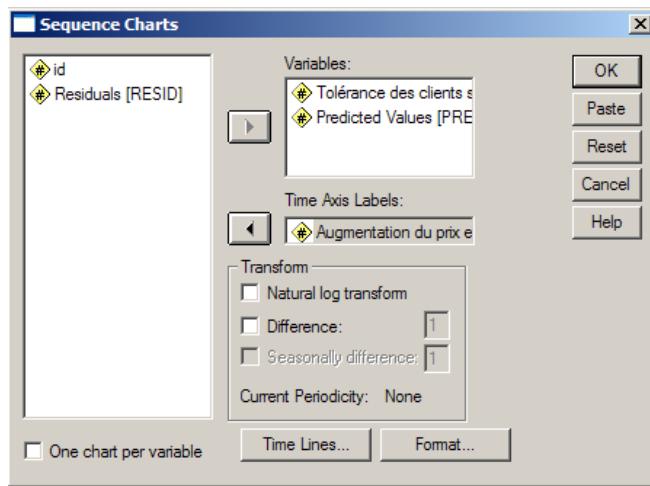


FIG. 12.71 – Pour produire le graphe séquentiel

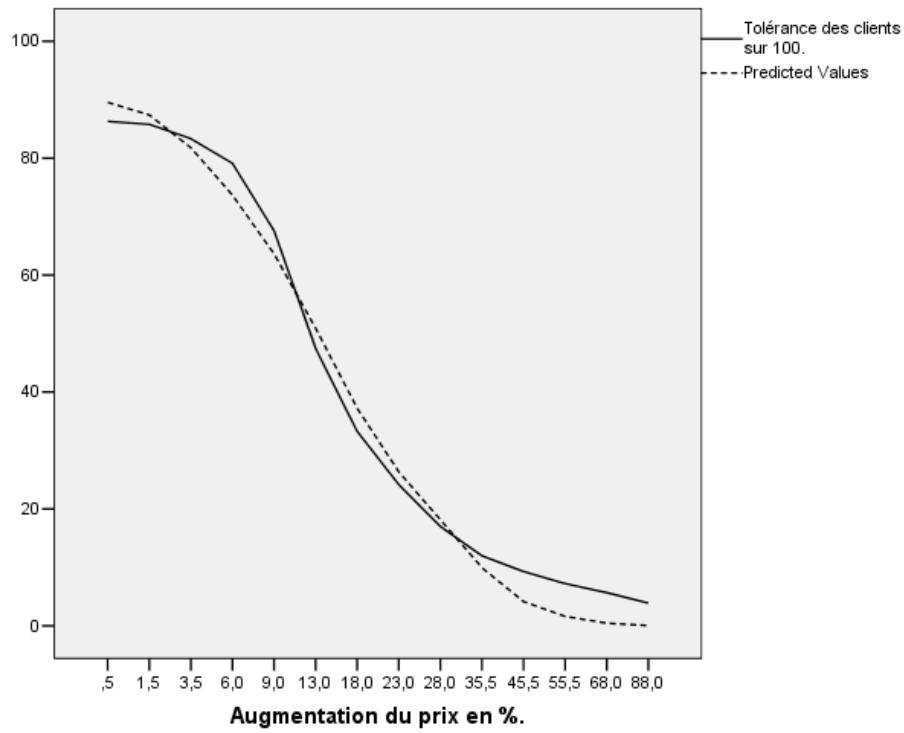


FIG. 12.72 – Comparaison des estimés et des tolérances observées

D'autres modèles simulant la courbe « S » existent. Il est fortement conseillé de comparer la performance de plus d'un modèle afin de choisir celui qui sied le mieux à ses données.

Les fonctions $S(t)$ et $F(t)$ présentées dans le tableau 12.73 sont issues de loi de probabilités. Ainsi leurs valeurs sont coincées entre 0 et 1. Pour des données non coincées entre 0 et 1, il suffit d'inclure un paramètre homothétique γ dans le modèle comme celui utilisé dans le cadre de cette section. Voici quelques modèles en forme de S décroissant $S(t)$ et croissants $F(t)$ les plus utilisés :

Nom	Paramètres	Condition	$S(t)$	$F(t)$
Exponentiel	$\lambda > 0$	$t \geq 0$	$e^{-\lambda t}$	$1 - e^{-\lambda t}$
Weibull	$\alpha, \lambda > 0$	$t \geq 0$	$e^{-\lambda t^\alpha}$	$1 - e^{-\lambda t^\alpha}$
Gompertz	$\alpha, \lambda > 0$	$t \geq 0$	$e^{(\frac{\lambda}{\alpha}(1-e^{-\alpha t}))}$	$1 - e^{(\frac{\lambda}{\alpha}(1-e^{-\alpha t}))}$
Log Logistique	$\alpha, \lambda > 0$	$t \geq 0$	$\frac{1}{1 + \lambda t^\alpha}$	$1 - \frac{1}{1 + \lambda t^\alpha}$

FIG. 12.73 – Quelques modèles en forme de S

La figure 12.74 présente quelques autres modèles utiles pour le lecteur averti. Contrairement aux modèles présentés dans le tableau précédent, ce ne sont pas toutes des courbes en formes de S. Les modèles présentés ici sont des versions généralisées de ceux présentés dans la littérature. Il est possible d'estimer les paramètres de ces modèles avec SPSS. Par exemple, pour le modèle de Weibull, le modèle proposé contient déjà le paramètre homothétique $\gamma = b_2$ ainsi qu'un paramètre de centralisation b_1 ramenant les données à l'origine.

Il faut prendre note qu'avec SPSS les valeurs initiales des paramètres sont impor-

Name	Model expression
Asymptotic Regression	$b_1 + b_2 * \exp(b_3 * x)$
Asymptotic Regression	$b_1 - (b_2 * (b_3 ** x))$
Density	$(b_1 + b_2 * x) ** (-1/b_3)$
Gauss	$b_1 * (1 - b_3 * \exp(-b_2 * x ** 2))$
Gompertz	$b_1 * \exp(-b_2 * \exp(-b_3 * x))$
Johnson-Schumacher	$b_1 * \exp(-b_2 / (x + b_3))$
Log-Modified	$(b_1 + b_3 * x) ** b_2$
Log-Logistic	$b_1 - \ln(1 + b_2 * \exp(-b_3 * x))$
Metcherlich Law of Diminishing Returns	$b_1 + b_2 * \exp(-b_3 * x)$
Michaelis-Menten	$b_1 * x / (x + b_2)$
Morgan-Mercer-Florin	$(b_1 * b_2 + b_3 * x ** b_4) / (b_2 + x ** b_4)$
Peal-Reed	$b_1 / (1 + b_2 * \exp(-(b_3 * x + b_4 * x ** 2 + b_5 * x ** 3)))$
Ratio of Cubics	$(b_1 + b_2 * x + b_3 * x ** 2 + b_4 * x ** 3) / (b_5 * x ** 3)$
Ratio of Quadratics	$(b_1 + b_2 * x + b_3 * x ** 2) / (b_4 * x ** 2)$
Richards	$b_1 / ((1 + b_3 * \exp(-b_2 * x)) ** (1/b_4))$
Verhulst	$b_1 / (1 + b_3 * \exp(-b_2 * x))$
Von Bertalanffy	$(b_1 ** (1 - b_4) - b_2 * \exp(-b_3 * x)) ** (1/(1 - b_4))$
Weibull	$b_1 - b_2 * \exp(-b_3 * x ** b_4)$
Yield Density	$(b_1 + b_2 * x + b_3 * x ** 2) ** (-1)$

FIG. 12.74 – Quelques modèles de courbes

tantes pour pouvoir faire converger l'algorithme non linéaire. Lorsque l'analyste n'a pas besoin d'un paramètre il est préférable de ne pas l'utiliser, ce qui facilite la convergence. Par exemple, dans notre exemple, le paramètre de centralisation b_1 de la loi de Weibull généralisée n'a pas été utilisé.

12.7 Les données manquantes

Les techniques présentées telle que la moyenne mobile, le modèle de décomposition ou le lissage exponentiel ne supportent pas les valeurs manquantes dans la série temporelle. Il est alors impossible de produire un modèle lorsqu'une seule donnée manquante se glisse dans le fichier. Il faut donc remplacer (imputer) les données manquantes avant de procéder à une analyse.

Lorsque peu de données sont manquantes, il suffit de remplacer celles-ci par la moyenne

des données les entourant. Les commandes suivantes remplacent les données manquantes par une moyenne avec un certain rayon d'action :

Menu SPSS :	→ Transform
	→ Replace Missing Values...
Dans la fenêtre New Variable(s) :	choisir la variable qui a des valeurs manquantes
Dans la fenêtre Name and Method :	Name : choisir un nom pour la nouvelle variable sans valeurs manquantes
	Method : Mean of nearby points
	Span of nearby points : <input checked="" type="checkbox"/> Number : 2

Ceci aura pour effet de remplacer la valeur manquante y_t par la moyenne suivante :

$$\frac{y_{t-2} + y_{t-1} + y_{t+1} + y_{t+2}}{4}.$$

Cependant, lorsque le fichier est « trouvé » de manière importante et que l'analyste entrevoit des patrons à différents intervalles de temps, il peut tenter de faire passer une courbe à travers les données. Le choix du modèle dépend de l'allure des données expérimentales. Aussi, rien n'empêche l'analyste d'isoler une séquence de la base de données. Pour ce faire, on peut utiliser un des modèles de la section précédente ou encore faire passer une courbe de régression polynomiale du type $\hat{y} = b_0 + b_1 t + b_2 t^2 + b_3 t^3$. Une fois la courbe déterminée, comme pour une régression, il suffit de choisir les temps t pour lesquels y_t est manquant et calculer l'estimé \hat{y} . La courbe peut être modélisée de plusieurs façons (avec l'aide d'un des nombreux modèles déjà vus), et une nouvelle approche est proposée dans l'exemple qui suit.

Exemple 12.7.1 Reprenons les données de l'exemple 12.6.1, mais cette fois-ci avec une donnée manquante.

id	augmentation	tolerance
1	,5	86,30
2	1,5	85,79
3	3,5	83,33
4	6,0	79,11
5	9,0	67,58
6	13,0	47,45
7	18,0	33,25
8	23,0	24,14
9	28,0	-
10	35,5	11,98
11	45,5	9,27
12	55,5	7,22
13	68,0	5,68
14	88,0	3,89

FIG. 12.75 – Les données

Puisque nous savons que les autres données semblent suivre un « S » décroissant, nous allons essayer de modéliser les données avec un modèle cubique et en S avec les commandes suivantes :

-
- | | |
|--------------------------------|---------------------------|
| Menu SPSS : | → Analyse |
| | → Regression |
| | → Curve Estimation... |
| Dans la fenêtre Dependent(s) : | tolerance |
| Dans la fenêtre Independent : | ✓ Variable : augmentation |
| Dans la fenêtre Models : | ✓ Cubic |
| | ✓ S |
| Dans le bouton Save : | ✓ Predicted values |
-

Les sorties de la figure 12.76 ne font que nous décrire les données.

Model Description	
Model Name	MOD_2
Dependent Variable	1
Equation	tolerance
	Cubic
	S ^a
Independent Variable	augmentation
Constant	Included
Variable Whose Values Label Observations in Plots	Unspecified
Tolerance for Entering Terms in Equations	,0001

a. The model requires all non-missing values to be positive.

Case Processing Summary

	N
Total Cases	14
Excluded Cases ^a	1
Forecasted Cases	0
Newly Created Cases	0

a. Cases with a missing value in any variable are excluded from the analysis.

Variable Processing Summary

	Variables	
	Dependent	Independent
	tolerance	augmentation
Number of Positive Values	13	14
Number of Zeros	0	0
Number of Negative Values	0	0
Number of Missing Values	User-Missing System-Missing	0 1

FIG. 12.76 – Les modèles, les données manquantes, etc...

Model Summary and Parameter Estimates								
Dependent Variable: tolerance								
Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Cubic	,986	210,625	3	9	,000	94,755	-4,305	,069
S	,247	3,608	1	11	,084	2,969	1,043	,000

The independent variable is augmentation.

FIG. 12.77 – Description des modèles

La figure 12.77 nous donne plusieurs informations, dont les r^2 pour chacun des modèles. Ainsi le modèle cubique semble de loin le meilleur puisque son $r^2 = 0,986$ versus un r^2 de 0,247 pour le modèle en S. On peut tirer de cette même sortie l'équation du modèle cubique :

$$\hat{y}_{\text{tolerance}} = 94,755 - 4,305x_{\text{aug}} + 0,069x_{\text{aug}}^2 - 0,000356x_{\text{aug}}^3.$$

Tolérance des clients sur 100.

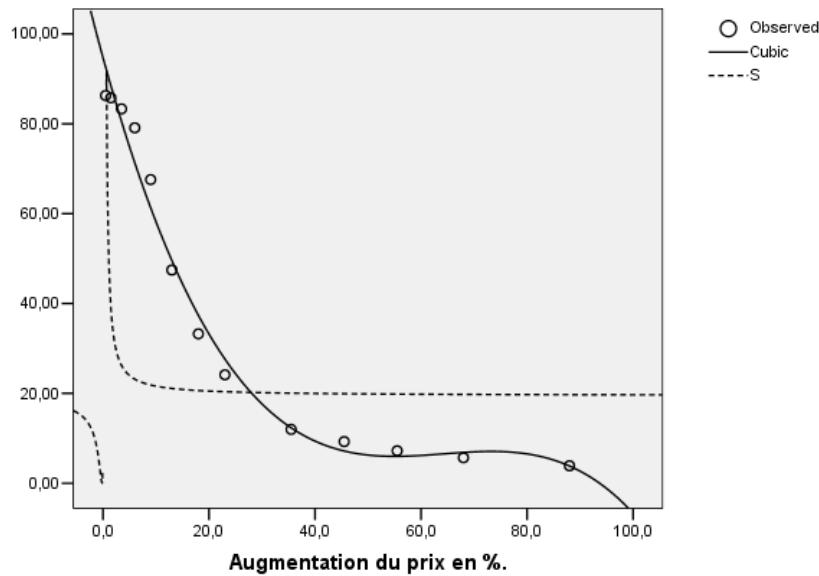


FIG. 12.78 – Les estimations des deux modèles versus les données observées

Le fait que le modèle cubique semble le meilleur est confirmé par le graphe de la figure

12.78. En fait on voit que le modèle en S est carrément « dans le champ ».

On choisit donc d'estimer la valeur manquante avec le modèle cubique :

$$\hat{y}_{28} = 94,755 - 4,305 \cdot 28 + 0,069 \cdot 28^2 - 0,000356 \cdot 28^3 = 20,496.$$

On peut aussi obtenir cette estimation dans la base de données (colonne FIT_1 de la figure 12.79). Encore une fois le fait d'arrondir les paramètres fait qu'il y a une différence entre la prédiction calculée à la main et celle calculée par SPSS.

id	augmentation	tolerance	FIT_1	FIT_2
1	,5	86,30	92,61982	156,88093
2	1,5	85,79	88,45059	39,03139
3	3,5	83,33	80,51145	26,22998
4	6,0	79,11	71,31421	23,16623
5	9,0	67,58	61,29925	21,86161
6	13,0	47,45	49,58513	21,09556
7	18,0	33,25	37,38459	20,63047
8	23,0	24,14	27,64828	20,37214
9	28,0	-	20,10918	20,20778
10	35,5	11,98	12,33622	20,04933
11	45,5	9,27	7,17038	19,92024
12	55,5	7,22	5,98709	19,83810
13	68,0	5,68	6,85420	19,76967
14	88,0	3,89	3,82879	19,70085

FIG. 12.79 – Les prédictions

12.8 Efficacité et comparaison de modèles

Cette dernière section s'intéresse aux méthodes courantes permettant l'évaluation de la qualité de prédiction des modèles en séries temporelles. Elles permettent de comparer des modèles entre eux afin de trouver, par exemple, la solution optimale. Bien entendu, ces méthodes sont entièrement basées sur les résidus $e_t = y_t - \hat{y}_t$ aussi appelés erreurs de prédiction. Un bon modèle obtient de faibles résidus. Les indices présentés dans le

cadre de cette section sont valides pour tous les modèles qui peuvent servir à établir des prédictions. Entre autre, ces techniques fonctionnent pour tous les modèles vus dans les sections précédant ce chapitre.

Lorsque possible, il est conseillé de diviser son fichier de données en deux parties ; une partie sert à la construction du modèle et une autre sert à tester l'efficacité du modèle. L'indice de performance MAD (*Mean Absolute Deviation*) ainsi que son homologue SAD (*Sum of Absolute Deviation*) sont les premiers indices de performance étudiés. Ces indices sont utilisés lorsque l'analyste désire mesurer l'ampleur de l'erreur de son modèle dans la même unité que la série étudiée. Ces indices s'écrivent

$$\text{SAD} = \sum_{t=1}^n |y_t - \hat{y}_t|$$

et

$$\text{MAD} = \frac{\sum_{t=1}^n |y_t - \hat{y}_t|}{n}.$$

L'indice MSE (*Mean Squared Error*) ainsi que son homologue SSE (*Sum of Squared Error*) sont deux indicateurs fortement utilisés par les logiciels statistiques. Par l'élévation de l'erreur au carré, ces indices fournissent à l'analyste une fonction de pénalité intéressante. Généralement, un meilleur modèle possède un plus petit SSE. Ces indices s'écrivent

$$\text{SSE} = \sum_{t=1}^n (y_t - \hat{y}_t)^2$$

et

$$\text{MSE} = \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{n}.$$

L'analyste peut aussi calculer les erreurs en terme de pourcentage. L'indice MAPE (*Mean Absolute Percentage error*) est utilisé à cet effet. Cet indice permet de comparer l'ampleur des erreurs aux valeurs de la série. Cet indice s'écrit

$$\text{MAPE} = \frac{1}{n} \cdot \frac{\sum_{t=1}^n |y_t - \hat{y}_t|}{y_t}.$$

Un bon modèle produit des estimations près des valeurs réelles certes, mais réparties tout autour des valeurs observées. Un modèle est dit biaisé si les estimations produites

surestiment ou sous-estiment toujours les valeurs observées. L’indice MPE (*Mean Percentage Error*) permet d’évaluer ceci. En effet, les écarts positifs produiront des pourcentages positifs tandis que les écarts négatifs produiront des pourcentages négatifs, et s’annuleront lorsqu’on les additionne. Un bon modèle produira donc un MPE près de zéro, et les mauvais modèles produiront des valeurs s’éloignant de zéro selon qu’ils surestiment ou sous-estiment les valeurs réelles de la série. Cet indice s’écrit

$$\text{MPE} = \frac{1}{n} \cdot \frac{\sum_{t=1}^n (y_t - \hat{y}_t)}{y_t}.$$

Finalement, un bon modèle produit des estimations aléatoirement distribuées autour des valeurs de la série. Le graphique des autocorrélations est utile pour détecter si les données sont réparties aléatoirement, pour déceler si les données ont une tendance (non stationnaire) et pour détecter la présence d’un effet saisonnier. Cet outil sera présenté dans le chapitre suivant.

Finalement, à partir de quand un modèle est-il acceptable ? Il faut que l’analyste place les indicateurs relativement au problème étudié. Par exemple, un indice MAD de 2 500 unités peut être énorme si les unités mensuelles moyennes sont de 10 000 unités. Ce même indice peut être considéré satisfaisant lorsque les unités mensuelles sont par exemple de 70 000 unités. Le tout dépend du problème et de la qualité des estimations à obtenir. Le jugement de l’analyste prend ici toute son importance.

Finalement, pour les méthodes de type lissage exponentiel, mentionnons que des systèmes de traçage envoient un signal à l’analyste lui permettant de revoir le modèle dans le plus bref délai. Cependant, ces méthodes de traçage vont au delà des objectifs de cette section.

12.9 Exercices du chapitre

Exercice 1 La base de données `immatriculation.sav` contient une série de données représentant le nombre de véhicules immatriculés par mois de janvier 1986 à décembre 1992. Il faut dans un premier temps imputer les données manquantes pour ensuite tenter un ou plusieurs modèles.

Exercice 2 La base de données `breuvage.sav` contient les ventes trimestrielles d'un breuvage sportif du premier trimestre de 1996 au dernier trimestre de 2003. Essayez de modéliser cette série.

Chapitre 13

Les modèles ARIMA

13.1 Introduction

Les modèles ARIMA (*AutoRegressive Integrated Moving Average models*) forment une classe de modèles linéaires qui utilisent les valeurs passées et présentes de la variable dépendante pour produire des prédictions à court-terme. Ces modèles fournissent des méthodes sophistiquées pour modéliser les tendances et les effets saisonniers, produisant habituellement de meilleurs résultats que les lissages exponentiels. Il est de plus possible d'inclure des variables indépendantes (un peu comme dans un modèle de régression), améliorant l'horizon de prédition et la stabilité du modèle. De plus, les modèles ARIMA supportent les valeurs manquantes.

La méthodologie pour l'identification d'un modèle ARIMA a été développée au début des années 1970 par George Box et Gwilym Jenkins. L'approche proposée par Box-Jenkins est une méthode itérative qui tente d'identifier un modèle permettant de faire de bonnes prévisions pour la variable dépendante sélectionnée. Le modèle sera efficient si les résidus sont petits, de moyenne nulle, distribués aléatoirement suivant une loi normale,

de variance constante en toutes périodes et indépendants entre eux. Si le modèle n'est pas satisfaisant, l'approche de Box-Jenkins est répétée jusqu'à la satisfaction de l'analyste.

Il y a trois composantes de base dans un modèle ARIMA : l'autorégression (AR), la différentiation ou intégration (I), et le processus de moyenne mobile (MA). Nous verrons en détail chacune de ces composantes, et les étapes de la méthode Box-Jenkins pour bien identifier le modèle ARIMA.

13.2 La stationnarité

La méthodologie de base de Box-Jenkins tente de décrire le comportement d'une série stationnaire. Ainsi, dans un premier temps, il est impératif de vérifier si la série est stationnaire. Sinon, l'analyste doit chercher à obtenir cette stationnarité.

Une série **stationnaire** possède trois caractéristiques :

- À long terme, la série est en moyenne égale à elle même sans effet de tendance.
- La variance de la série est homoscédastique à travers le temps. Ainsi, si la variance change avec le temps, la série sera considérée non stationnaire.
- L'autocorrélation entre les données est la même avec le temps. Par exemple, si une série y_t est telle que y_t est fortement relié à y_{t-4} , ceci est supposé demeurer vrai pour tout t .

En présence d'une série non stationnaire, l'analyste doit alors effectuer une ou des transformations pour obtenir la stationnarité. La transformation proposée par Box-Jenkins pour résoudre les problèmes de tendance est la différentiation. Cette différentiation est associée au terme *Integrated* dans l'acronyme ARIMA, et est en fait l'approximation de la dérivée de la série.

D'autre part, le problème lié à l'hétéroscédasticité peut être réglé, comme en régression, en transformant la variable dépendante avec un logarithme ou une racine carrée.

Définition de la différentiation

Soient y_1, y_2, \dots, y_n les valeurs observées d'une série chronologique. La **définition d'ordre 1** prend la forme suivante :

$$z_t = y_t - y_{t-1} \text{ pour } t = 2, 3, \dots, n.$$

Souvent, la différentiation d'ordre 1 solutionne le problème lié à la tendance. S'il n'est pas réglé, l'analyste fait de nouveau une différentiation à partir de la série différentiée, ce qui donne une **définition d'ordre 2** :

$$\begin{aligned} z'_t &= z_t - z_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \quad \text{pour } t = 3, 4, \dots, n. \end{aligned}$$

Exemple 13.2.1 Le tableau qui suit montre des différentiations d'ordre 1 et 2.

Série originale	Différentiation de premier ordre $z_t = y_t - y_{t-1}$	Différentiation de deuxième ordre $z'_t = z'_t - z'_{t-1}$
y_t	$z_t = y_t - y_{t-1}$	$z'_t = z'_t - z'_{t-1}$
4		
7	3	
8	1	-2
2	-6	-7
5	3	9

Exemple 13.2.2 La base de données `ventesmensuelles.sav` contient les ventes mensuelles d'une petite entreprise. Les ventes sont en centaines de dollars. La figure 13.1 présente le graphe séquentiel de ces ventes. On voit que la série n'est pas stationnaire car elle présente une tendance à la hausse.

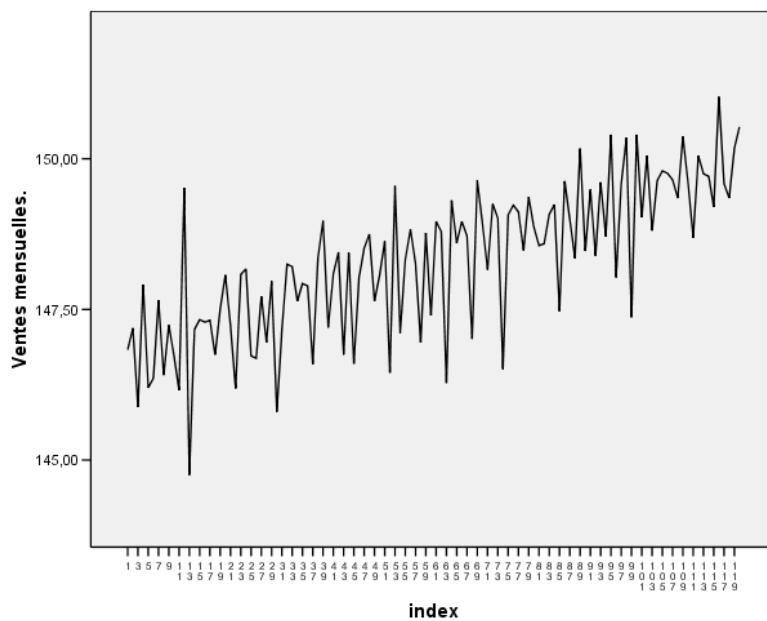


FIG. 13.1 – La série (non stationnaire)

Pour rendre cette série stationnaire il faut effectuer une différentiation. Les commandes sont les suivantes :

Menu SPSS :

→ Transform

→ Create Time Series...

Dans la fenêtre Function : → sélectionnez Difference

Dans la fenêtre Name : → ventes_1

Dans la fenêtre Order : → 1 (pour une différentiation d'ordre 1)

Appuyez sur Change (si nécessaire) puis sur OK.

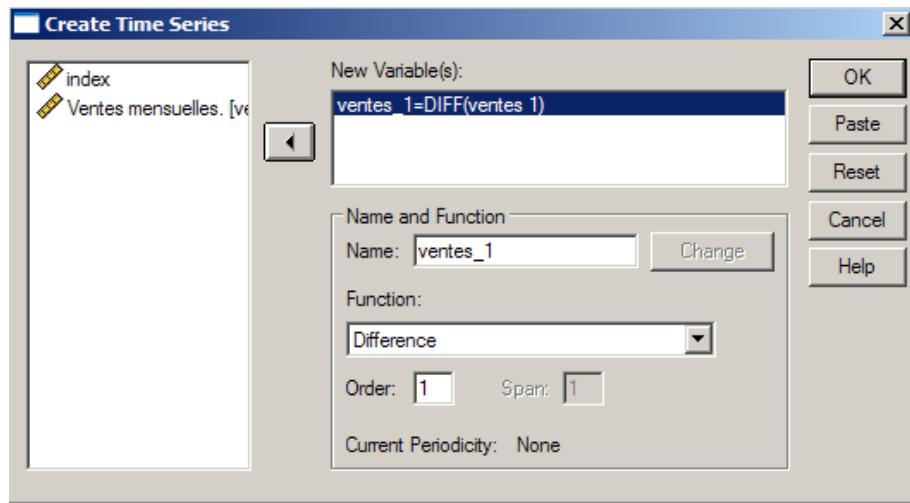


FIG. 13.2 – Pour créer la série différentiée

La figure 13.3 présente un aperçu des valeurs de la série différentiée, et son graphe séquentiel : on voit que cette nouvelle série ne présente pas de tendance. Elle semble aussi homoscédastique (la variance est constante), et l'autocorrélation entre les valeurs semble demeurer la même (les fluctuations demeurent semblables en terme de « largeur »). Ainsi la stationnarité semble atteinte, il n'est donc pas nécessaire d'essayer une différentiation de second ordre. Nous verrons plus tard un outil pour mieux juger de la stationnarité d'une série.

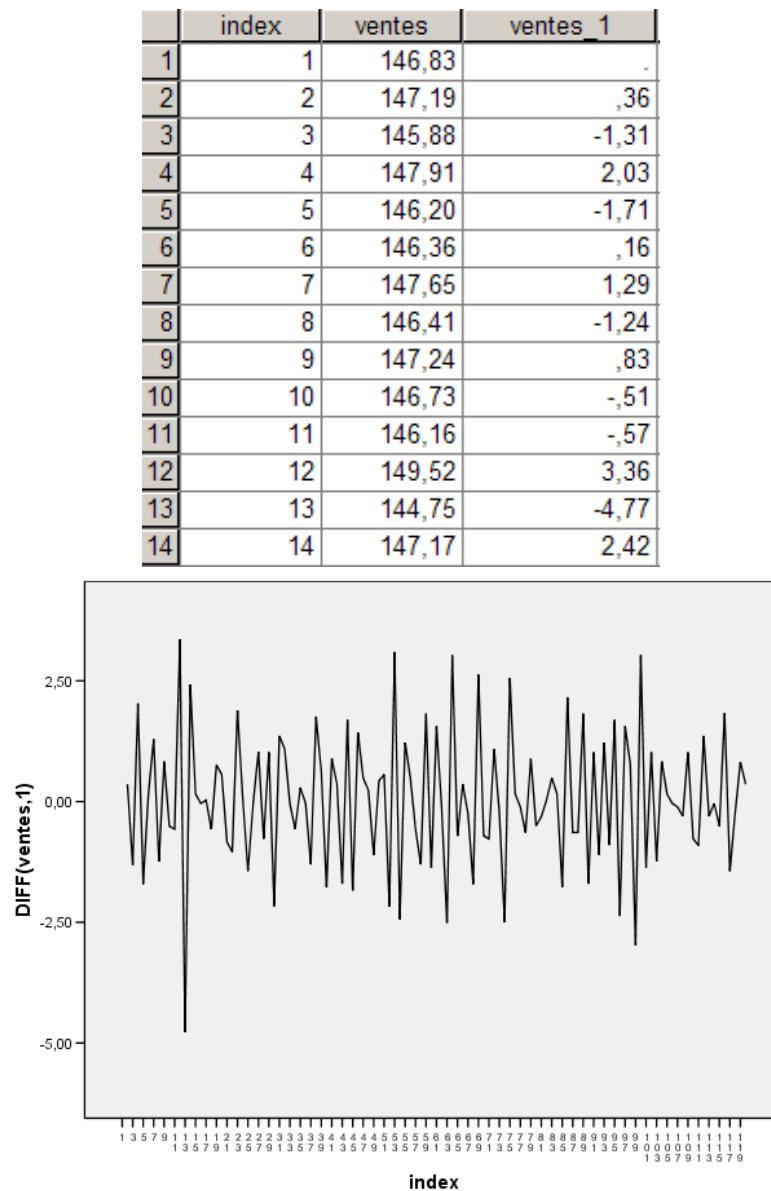


FIG. 13.3 – Aperçu et graphe de la série différentiée

13.3 Le modèle ARIMA de Box-Jenkins

Tel que mentionné dans l'introduction, il y a trois composantes de base dans un modèle ARIMA :

- AR** Le modèle autorégressif (*AutoRegressive*);
- I** La différentiation (*Integrated*);
- MA** Le modèle de moyenne mobile (*Moving Average*).

Tous les modèles ARIMA font partie de la famille générale ARIMA(p, d, q) où p, d et q sont des entiers plus grand ou égal à zéro tels que :

- p est l'ordre du modèle autorégressif ;
- d est l'ordre de la différentiation nécessaire pour obtenir une série stationnaire ;
- q est l'ordre du modèle de moyenne mobile.

Nous définissons maintenant ce que sont les modèles autorégressifs et de moyenne mobile, puis nous présentons les étapes de la méthodologie Box-Jenkins qui permettront justement de déterminer l'ordre de ces modèles.

Les modèles autorégressifs

Un **modèle autorégressif d'ordre p** , noté AR(p), est de la forme

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t$$

où

- Y_t est la variable dépendante ;
- Y_{t-i} est la variable dépendante déphasée de i périodes ($i = 1, \dots, p$) ;
- $\phi_0, \phi_1, \phi_2, \dots, \phi_p$ sont les coefficients à estimer ;
- ϵ_t est l'erreur d'estimation au temps t ; les hypothèses à propos du terme d'erreur sont les mêmes que celles du modèle de régression linéaire. Rappelons qu'au niveau de l'échantillon l'erreur au temps t est notée e_t .

On voit que l'équation d'un modèle autorégressif est celle d'un modèle linéaire dont les variables explicatives sont des variables déphasées formées à partir de la variable dépendante. Cependant, ce modèle n'est pas traité comme le serait un modèle de régression linéaire. En effet, les paramètres d'une régression linéaire sont estimés avec la méthode standard des moindres carrés ; or en présence de variables déphasées, on rencontre souvent des problèmes de multicolinéarité qui rendent l'estimation des paramètres instable lorsqu'elle est faite avec les moindres carrés standards.

Dans l'approche proposée par Box-Jenkins, les paramètres ϕ_i du modèle autorégressif AR(p) sont évalués à l'aide de la méthode itérative non linéaire du moindre carré. Cette méthode tient compte du fait que les variables explicatives Y_{t-i} sont corrélées entre elles, ce qui évite les problèmes d'instabilité mentionnés précédemment.

Les modèles de moyenne mobile

Un **modèle de moyenne mobile d'ordre q** , noté MA(q), est de la forme

$$Y_t = \omega_0 + \epsilon_t - \omega_1\epsilon_{t-1} - \omega_2\epsilon_{t-2} + \cdots + \omega_q\epsilon_{t-q}$$

où

- Y_t est la variable dépendante ;
- $\omega_0, \omega_1, \omega_2, \dots, \omega_p$ sont les coefficients à estimer ;
- ϵ_t est l'erreur d'estimation au temps t ; les hypothèses à propos du terme d'erreur sont les mêmes que celles du modèle de régression linéaire ;
- ϵ_{t-i} est le terme d'erreur déphasé de i périodes ($i = 1, \dots, q$), c'est-à-dire les erreurs d'estimation faites en des périodes de temps précédentes.

Ainsi, un modèle de moyenne exprime le fait que la variable dépendante Y_t dépend des erreurs d'estimation passées. Le fait que les coefficients sont négatifs est simplement une convention, il est donc possible d'obtenir des valeurs positives.

Il est important de ne pas confondre ce modèle avec les moyennes mobiles vues au chapitre 12, ce sont deux techniques complètement différentes.

Il est important de remarquer que par hypothèse, $E(\epsilon_t) = 0$ pour tout t , et ainsi on a

$$\begin{aligned} E(Y_t) &= E(\omega_0) + E(\epsilon_t) - E(\omega_1\epsilon_{t-1}) - E(\omega_2\epsilon_{t-2}) + \cdots + E(\omega_q\epsilon_{t-q}) \\ &= \omega_0 + E(\epsilon_t) - \omega_1 E(\epsilon_{t-1}) - \omega_2 E(\epsilon_{t-2}) + \cdots + \omega_q E(\epsilon_{t-q}) \\ &= \omega_0 + 0 - \omega_1 \cdot 0 - \omega_2 \cdot 0 + \cdots + \omega_q \cdot 0 \\ &= \omega_0. \end{aligned}$$

Les modèles ARMA(p, q)

Un modèle ARMA(p, q) est une combinaison d'un modèle autorégressif et de moyenne mobile, qui s'applique sur une série stationnaire. Le modèle a alors la forme suivante :

$$Y_t = \underbrace{\phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p}}_{\omega_0} + \epsilon_t - \omega_1 \epsilon_{t-1} - \omega_2 \epsilon_{t-2} + \cdots + \omega_q \epsilon_{t-q}.$$

Ainsi un modèle ARMA(p, q) est un modèle mixte qui tient compte des valeurs passées et des erreurs passées pour faire les prédictions. Le modèle ARIMA(p, d, q) est simplement un modèle ARMA(p, q) appliqué à une série différentiée d'ordre d ; rappelons qu'en effet une série doit être stationnaire pour la modéliser avec un modèle ARMA, et lorsqu'elle ne l'est pas on applique d'abord une ou des différentiations afin d'obtenir une série stationnaire. En somme, la différentiation enlève la tendance de la série, ce qui permet d'évaluer le modèle, puis ensuite on applique l'opération inverse (l'intégration) afin d'établir les estimations sur la série originale. De là l'origine du terme *Integrated* dans l'acronyme ARIMA.

Finalement, le problème auquel l'analyste est confronté est de savoir quels sont les ordres p et q qui donneront le modèle ARMA adéquat pour la série étudiée. Ce choix se basera sur les autocorrélations et les autocorrélations partielles de la série, ce dont nous discutons dans la prochaine section.

13.4 Les fonctions SAC et SPAC

L'identification d'un modèle ARIMA se fait à l'aide de la fonction SAC (*Sample Auto-Correlation function*) et de la fonction SPAC (*Sample Partial AutoCorrelation function*). Commençons donc par définir ce que sont ces fonctions.

Soit Y_t une série temporelle, et Z_t la série temporelle stationnaire issue de la série Y_t ; supposons sans perte de généralité que z_1, z_2, \dots, z_n sont les valeurs de la variable Z_t . Rappelons que pour obtenir Z_t il peut être nécessaire d'appliquer une ou des transformations mathématiques selon les problèmes identifiés. Par exemple, si Y_t présente une tendance, on applique une ou des différentiations, et si elle semble hétéroscléastique, on peut appliquer un logarithme.

Définition de l'autocorrélation

L'autocorrélation r_k est la corrélation entre une variable Z_t et cette même variable déphasée de k périodes, c'est-à-dire Z_{t-k} . Elle s'exprime sous la forme suivante :

$$r_k = \frac{\sum_{t=k+1}^n (z_t - \bar{z})(z_{t-k} - \bar{z})}{\sum_{t=1}^n (z_t - \bar{z})^2} \quad \text{avec } k = 0, 1, 2, \dots$$

L'écart-type de la distribution d'échantillonnage de r_k est le suivant (estimation de Bartlett) :

$$s_{r_k} = \sqrt{\frac{1 + 2 \sum_{j=1}^{k-1} r_j^2}{n}}$$

La mesure r_k varie entre -1 et 1 et s'interprète comme une corrélation. Elle mesure le lien linéaire entre des observations et les mêmes observations déphasées de k périodes. Par exemple, si on constate qu'un taux de rendement observé deux mois auparavant influence positivement le taux du mois présent (c'est-à-dire s'ils varient dans le même

sens, que ce soit à la hausse ou à la baisse), l'autocorrélation entre le taux au temps t et au temps $t - 2$ sera positive.

Pour juger si une autocorrélation est significative ou pas, l'écart-réduit $t_k = \frac{r_k}{s_{r_k}}$ est calculé. Au seuil de 5 %, une autocorrélation sera jugée significativement différente de 0 si $|t_k| > 1,96$.

La **fonction SAC** est simplement formée des valeurs que prennent les r_k , et est la plupart du temps accompagnée d'un graphe représentant ces valeurs. Les commandes pour obtenir le graphe ainsi que les valeurs de la fonction SAC dans SPSS sont les suivantes (dans le contexte de l'exemple 13.2.2, base de données `ventesmensuelles.sav`) :

Menu SPSS :	\rightarrow Graphs
	\rightarrow Time Series
	\rightarrow Autocorrelations...
Dans la fenêtre Variables :	\rightarrow ventes
Dans la fenêtre Display :	\checkmark Autocorrelations (seulement)
Dans le bouton Options... :	Maximum Number of Lags : 16
	Standard Error Method
	\checkmark Bartlett's approximation

On obtient alors les sorties de la figure 13.4. On voit par exemple que les ventes et les ventes déphasées d'un mois ont une autocorrélation positive de 0,342. Cette autocorrélation est significativement différente de 0 au seuil de 5 % puisque

$$t_1 = \frac{r_1}{s_{r_1}} = \frac{0,342}{0,091} = 3,76 > 1,96.$$

Autocorrelations						
Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic			Sig. ^b
			Value	df	Sig.	
1	,342	,091	14,420	1	,000	
2	,516	,101	47,399	2	,000	
3	,579	,121	89,301	3	,000	
4	,472	,142	117,370	4	,000	
5	,524	,155	152,342	5	,000	
6	,444	,169	177,665	6	,000	
7	,458	,179	204,842	7	,000	
8	,476	,188	234,409	8	,000	
9	,459	,198	262,217	9	,000	
10	,326	,207	276,325	10	,000	
11	,485	,211	307,868	11	,000	
12	,404	,220	329,949	12	,000	
13	,322	,226	344,120	13	,000	
14	,404	,230	366,665	14	,000	
15	,360	,236	384,720	15	,000	
16	,323	,240	399,425	16	,000	

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

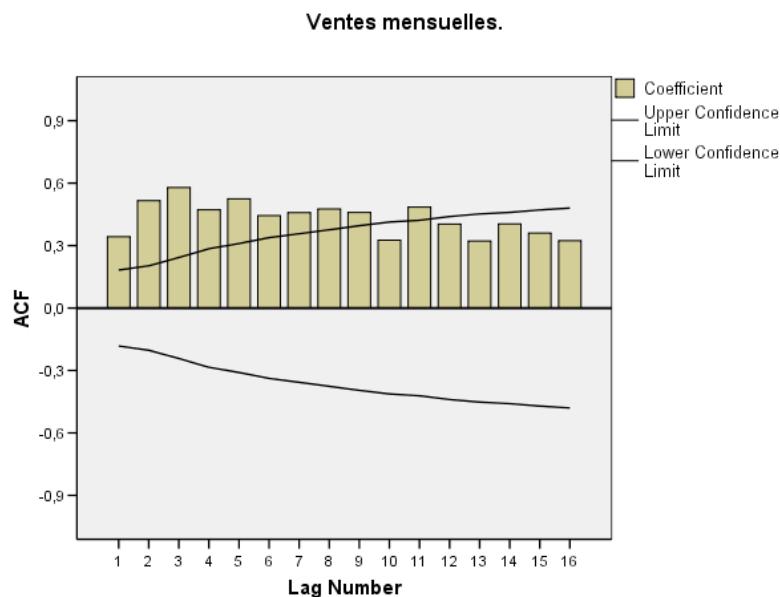


FIG. 13.4 – Les autocorrélations de la série originale

On peut visualiser ceci dans la deuxième sortie de la figure 13.4 : en effet, les bâtons du graphe illustrent les valeurs des autocorrélations, et les lignes représentent les valeurs $\pm 1,96 \cdot s_{r_k}$. Ainsi, si $|r_k| > 1,96 \cdot s_{r_k}$, on a alors $\left| \frac{r_k}{s_{r_k}} \right| > 1,96$, ce qui signifie que l'autocorrélation du *lag k* est significativement différente de 0, et dans ce cas le bâton correspondant à cette autocorrélation « dépasse » la ligne. Dans le graphe de la figure 13.4, on voit par exemple que la première autocorrélation qui n'est pas significative est r_{10} , puis r_{12} jusqu'à r_{16} .

Non seulement la fonction SAC nous aidera à identifier le modèle ARIMA qui convient à une série, mais elle permet aussi de détecter si une série est stationnaire ou pas. En effet, il peut être montré que si les autocorrélations décroissent rapidement, alors la série étudiée est stationnaire. Si, au contraire, la suite des autocorrélations décroît lentement, la série n'est pas stationnaire, et il faut alors la transformer pour obtenir la stationnarité. Dans notre exemple, on constate que les autocorrélations décroissent lentement, ce qui nous indique que la série des ventes n'est pas stationnaire. On avait en effet observé dans le graphe 13.1 que cette série présente une tendance à la hausse. On avait aussi vu dans l'exemple 13.2.2 qu'une différentiation semblait régler ce problème. Pour vérifier si la fonction SAC appuie ceci, il suffit de calculer les autocorrélations sur la série différentiée.

Les commandes sont les suivantes :

Menu SPSS :	\rightarrow Graphs
	\rightarrow Time Series
	\rightarrow Autocorrelations...
Dans la fenêtre Variables :	\rightarrow ventes
Dans la fenêtre Display :	\checkmark Autocorrelations (seulement)
Dans la fenêtre Transform :	\checkmark Difference : 1
Dans le bouton Options... :	Maximum Number of Lags : 16
	Standard Error Method
	\checkmark Bartlett's approximation

On obtient alors les sorties de la figure 13.5.

Autocorrelations						
Series: Ventes mensuelles.						
Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic			Sig. ^b
			Value	df	Sig.	
1	-.646	,092	50,905	1	,000	
2	,085	,124	51,801	2	,000	
3	,144	,125	54,362	3	,000	
4	-,150	,126	57,184	4	,000	
5	,122	,128	59,060	5	,000	
6	-,071	,128	59,697	6	,000	
7	-,012	,129	59,714	7	,000	
8	,029	,129	59,820	8	,000	
9	,101	,129	61,160	9	,000	
10	-,242	,130	68,897	10	,000	
11	,202	,133	74,327	11	,000	
12	-,020	,136	74,383	12	,000	
13	-,117	,136	76,227	13	,000	
14	,099	,137	77,560	14	,000	
15	-,007	,137	77,567	15	,000	
16	-,031	,137	77,700	16	,000	

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

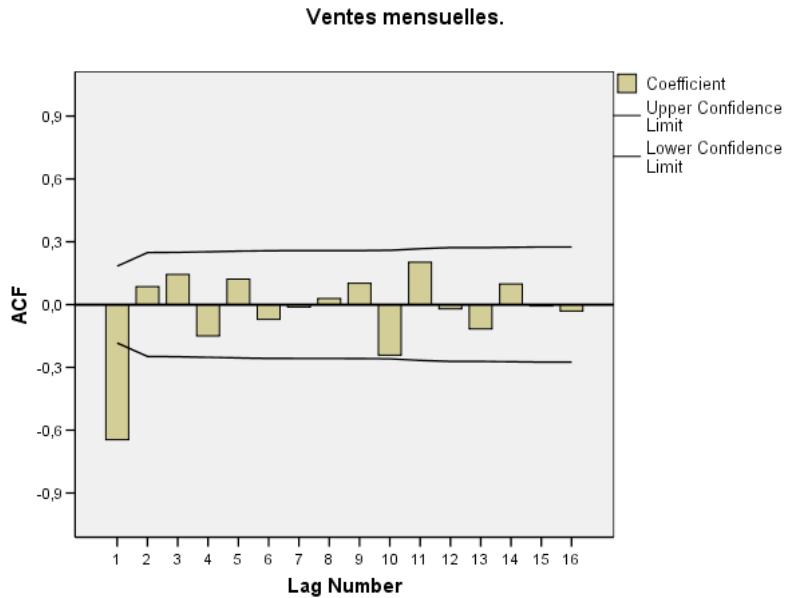


FIG. 13.5 – Les autocorrélations de la série différentiée

On constate que la série différentiée est effectivement stationnaire car ses autocorrélations décroissent rapidement (figure 13.5). On voit que seule la première autocorrélation est significativement différente de zéro. Puisque la série est maintenant stationnaire, il serait maintenant possible d'identifier le modèle ARIMA($p, 1, q$) qui donnerait les meilleures prédictions. Mais avant, nous devons présenter ce que sont les autocorrélations partielles et la fonction SPAC, qui serviront aussi à l'identification de ce modèle.

Définition de l'autocorrélation partielle

L'autocorrélation partielle r_{kk} peut être vue intuitivement comme étant l'autocorrélation entre Y_t et Y_{t-k} de laquelle on enlève l'effet des variables intermédiaires $Y_{t-1}, \dots, Y_{t-k+1}$. C'est donc le lien entre Y_t et Y_{t-k} qui ne transige pas par les variables $Y_{t-1}, \dots, Y_{t-k+1}$. L'autocorrélation partielle pour un retard de k périodes est donnée par :

$$r_{kk} = \begin{cases} r_1 & \text{si } k = 1; \\ \frac{r_k - \sum_{j=1}^{k-1} r_{k-1, j} \cdot r_{k-j}}{\sqrt{1 - \sum_{j=1}^{k-1} r_{k-1, j}^2}} & \text{si } k = 2, 3, \dots \end{cases}$$

où $r_{kj} = r_{k-1, j} - r_{kk} \cdot r_{k-1, k-j}$ pour $j = 1, 2, \dots, k-1$.

L'écart-type de la distribution d'échantillonnage de r_{kk} est donné par :

$$s_{r_{kk}} = \sqrt{\frac{1}{n}}.$$

La mesure r_{kk} varie elle aussi entre -1 et 1. Et tout comme pour l'autocorrélation, on peut juger si une autocorrélation partielle est significative ou pas en calculant le ratio $t_{kk} = \frac{r_{kk}}{s_{r_{kk}}}$. Ainsi, au seuil de 5 %, une autocorrélation partielle sera jugée significativement différente de zéro si $|t_{kk}| > 1,96$.

La **fonction SPAC** est simplement formée des valeurs que prennent les r_{kk} , et est la plupart du temps accompagnée d'un graphe représentant ces valeurs. Les commandes pour obtenir le graphe ainsi que les valeurs de la fonction SPAC dans SPSS sont les suivantes (toujours dans le contexte de l'exemple 13.2.2, et donc on demande une différentiation des ventes pour avoir une série stationnaire) :

Menu SPSS :	→ Graphs
	→ Time Series
	→ Autocorrelations...
Dans la fenêtre Variables :	→ ventes
Dans la fenêtre Display :	✓ Partial autocorrelations (seulement)
Dans la fenêtre Transform :	✓ Difference : 1
Dans le bouton Options... :	Maximum Number of Lags : 16

On obtient alors les sorties de la figure 13.6. Tout comme dans le SAC, une autocorrélation partielle qui « dépasse » les lignes du graphe (qui ici représentent les valeurs $\pm 1,96 \cdot \frac{1}{\sqrt{n}}$) sont considérées comme étant significativement différentes de zéro. Dans la figure 13.6, on voit que les cinq premières autocorrélations partielles sont significatives, et que r_{88} l'est aussi.

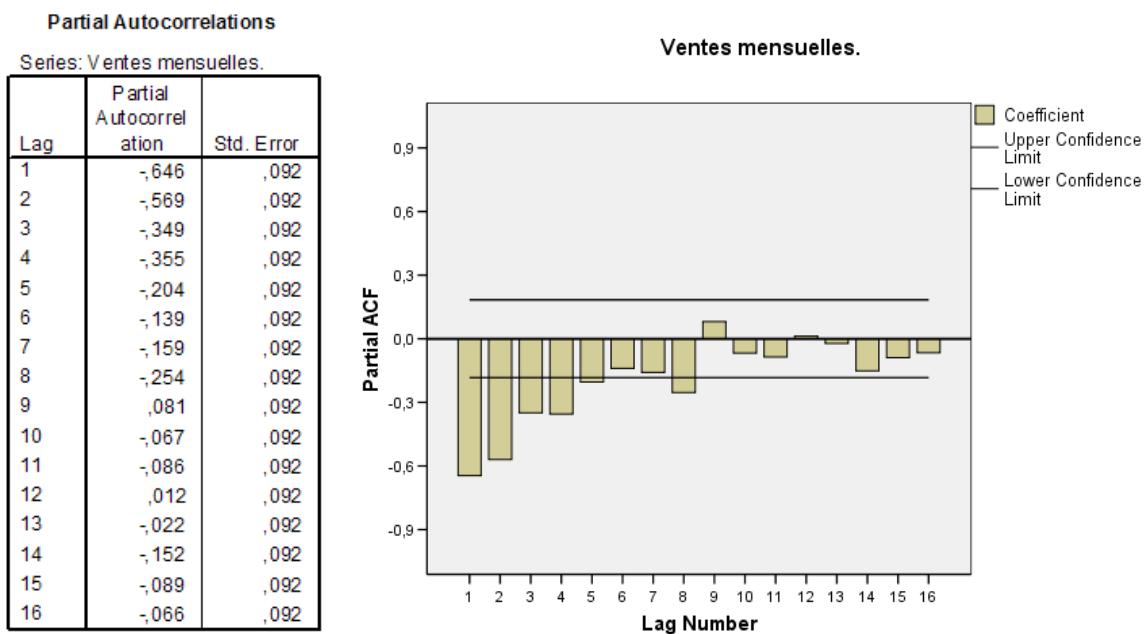


FIG. 13.6 – Les autocorrélations partielles de la série différentiée

13.5 Les comportements théoriques des SAC et SPAC

Les comportements des fonctions SAC et SPAC d'une série stationnaire nous permettront de déterminer les ordres p et q du modèle ARIMA(p, d, q) qui convient le mieux aux données. Nous présentons donc les différents comportements qui sont théoriquement possibles, puis il suffit ensuite de voir dans la pratique lequel de ces comportements semble être présent pour les données étudiées. Les comportements observés sont rarement identiques aux comportements théoriques, mais des similitudes sont toujours présentes. Lorsque plus d'un modèle semble possible, il suffit d'essayer... la méthodologie de Box-Jenkins nous permettra d'ailleurs de voir si le modèle choisi semble adéquat. Pour l'instant nous nous intéressons aux séries ne présentant pas d'effet saisonnier ; nous verrons dans une section ultérieure comment traiter les effets de saison.

Voici donc une brève description des comportements possibles, suivis dans les pages suivantes par une illustration de bon nombre de ceux-ci :

- La fonction SAC d'un modèle non stationnaire possède près d'une demi-douzaine (ou plus) d'autocorrélations significatives, et le déclin vers 0 est lent. Il est impératif de différentier la série avant de procéder à l'identification du modèle ARMA.
- La fonction SAC d'un processus AR décline vers 0 de manière exponentielle et présente un ou deux pics significatifs (c'est-à-dire une ou deux autocorrélations partielles significatives) dans la fonction SPAC. Le nombre de pics dans le SPAC détermine le degré p pour le modèle autorégressif.
- La fonction SPAC d'un processus MA décline vers 0 de manière exponentielle et présente un ou deux pics significatifs (c'est-à-dire une ou deux autocorrélations significatives) dans la fonction SAC. Le nombre de pics dans le SAC détermine le degré q pour le modèle de moyenne mobile.
- Un modèle mixte ARMA présente généralement un SAC et un SPAC qui déclinent vers 0 de manière exponentielle.

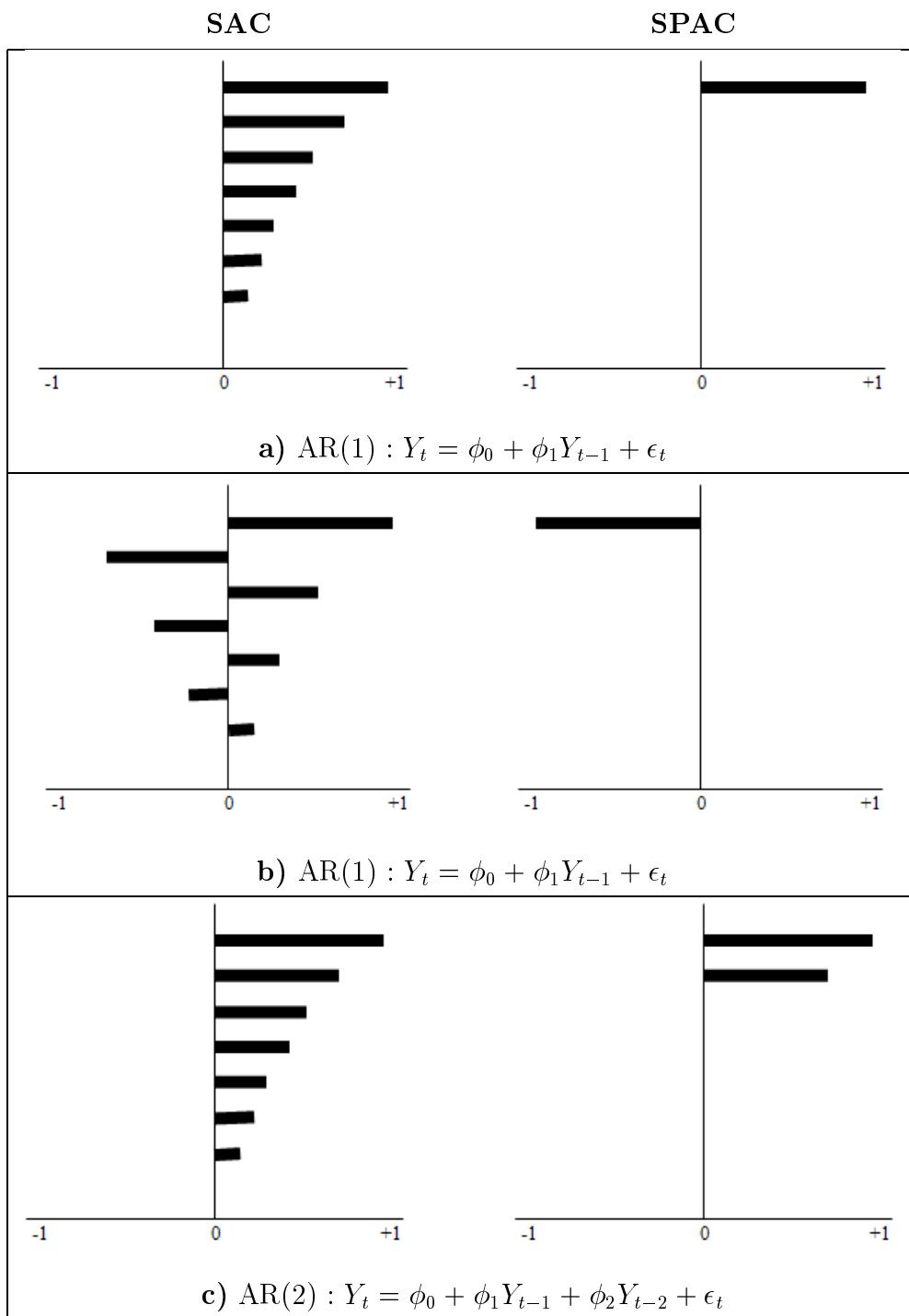


FIG. 13.7 – Comportements théoriques

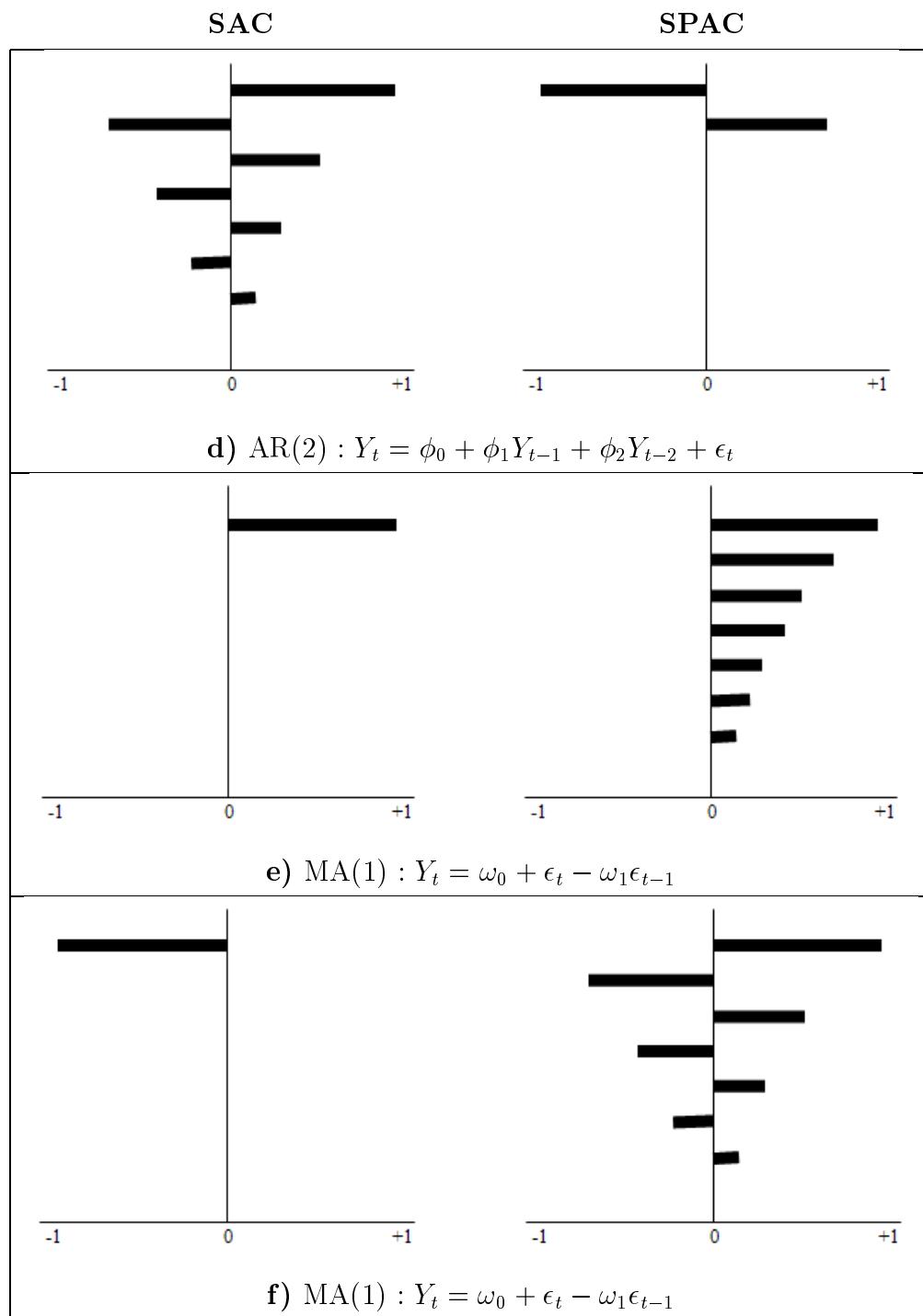


FIG. 13.8 – Comportements théoriques - suite

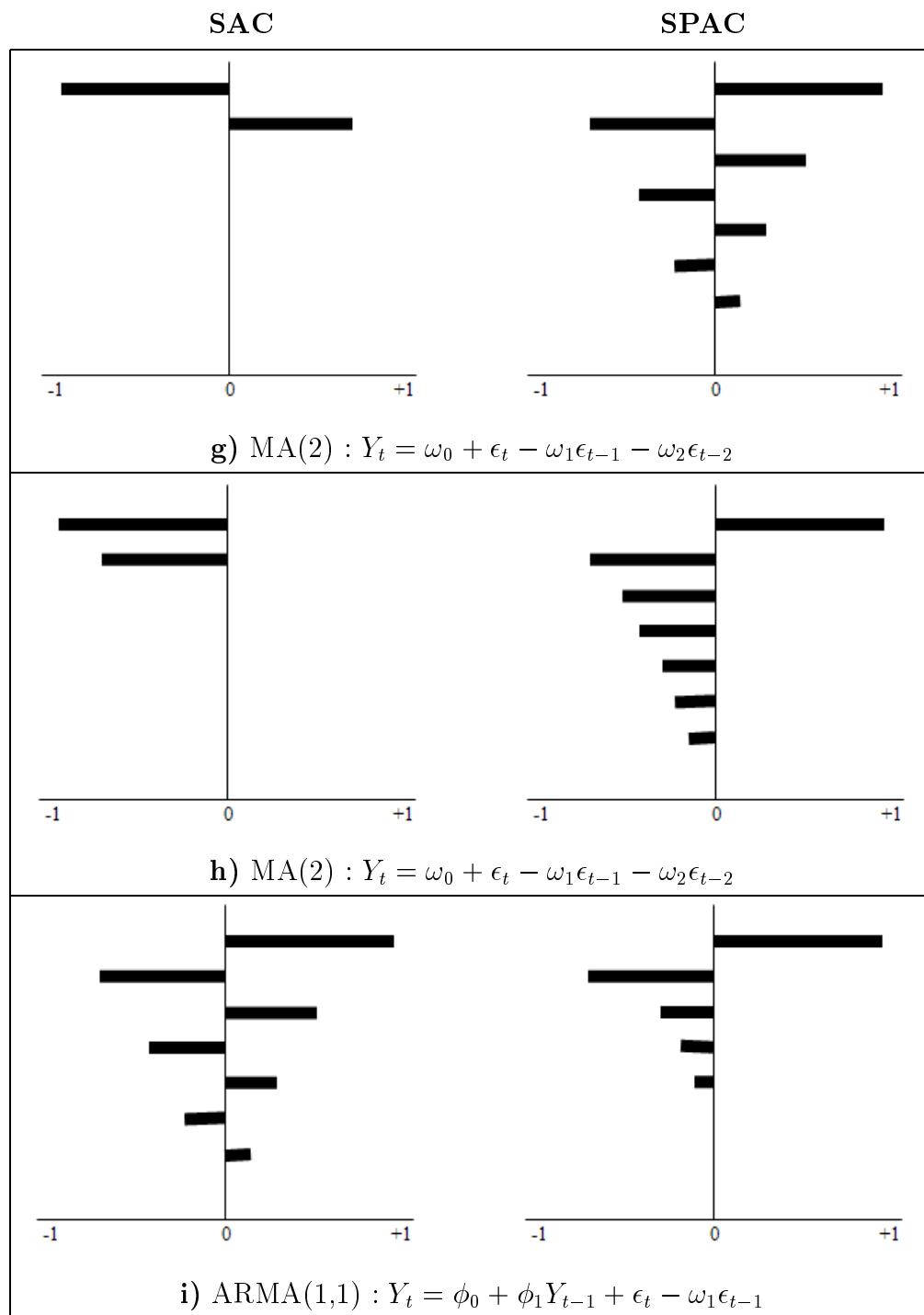


FIG. 13.9 – Comportements théoriques - suite

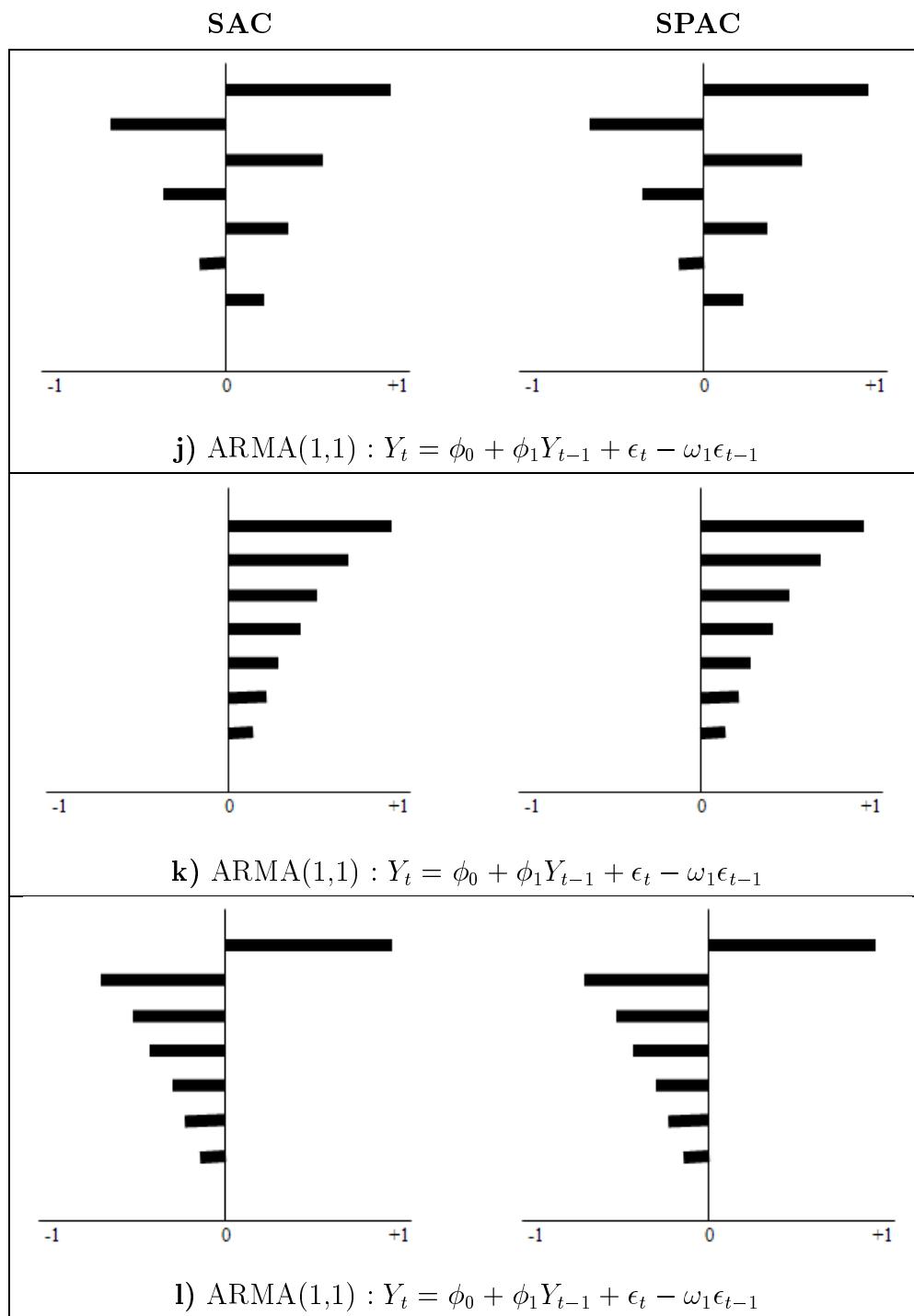


FIG. 13.10 – Comportements théoriques - suite

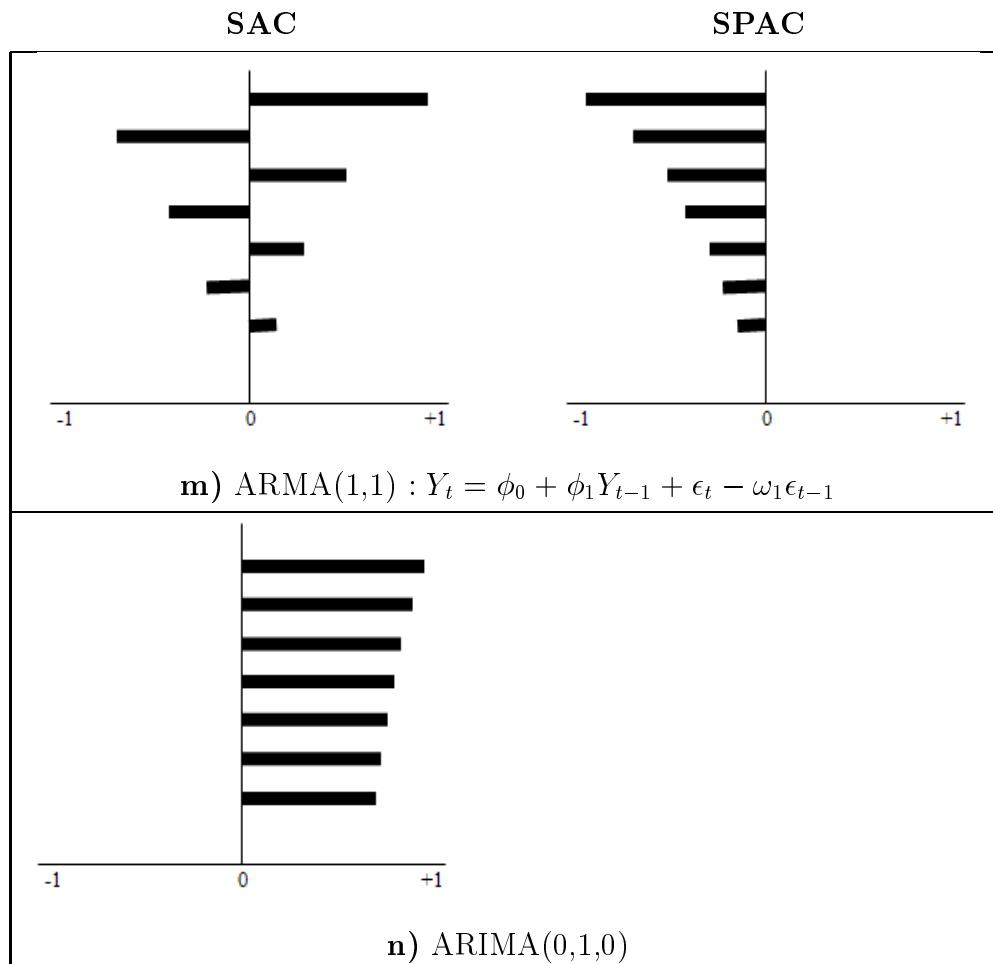


FIG. 13.11 – Comportements théoriques - suite et fin

Lors de l'identification du modèle, l'analyste n'a pas à s'inquiéter du signe des autocorrélations et autocorrélations partielles, pas plus que de l'ampleur du déclin exponentiel vers zéro des SAC et SPAC. Ces conditions seront reflétées par les signes et l'importance des coefficients des processus AR et MA du modèle. Une fois un modèle tenté, l'analyste doit analyser la distribution des résidus afin de vérifier la qualité du modèle. Plus précisément, la méthodologie de Box-Jenkins propose une approche constituée de quatre étapes pour l'identification des modèles ARIMA.

13.6 La méthodologie Box-Jenkins (sans saison)

L'approche proposée par Box-Jenkins est une méthodologie qui se décompose essentiellement en quatre étapes auxquelles est ajoutée, pour fin pédagogique, l'étape préliminaire de l'obtention de la stationnarité. Voici donc ces cinq étapes :

1. Vérification et obtention de la stationnarité, ce qui détermine l'ordre d de la différentiation.
2. Identification des ordres p et q du modèle ARMA à l'aide du comportement des graphiques SAC et SPAC.
3. Estimation des paramètres du modèle identifié à l'étape précédente à partir des observations de la série étudiée.
4. Diagnostique, validation et amélioration du modèle s'il y a lieu.
5. Calcul des estimations et prédictions.

Nous illustrons maintenant chacune de ces étapes à l'aide d'un exemple. La base de données `imprimante.sav` contient des observations issues d'un contrôle de production ; ce sont des mesures de la qualité d'impression d'une unité d'imprimante. Cette base de données contient 180 observations ; nous établirons le modèle à partir des 150 premières,

puis nous ferons des prédictions pour les trente dernières observations. Pour ce faire, il faut d'abord sélectionner les 150 premières données de façons à filtrer les 30 dernières. Les commandes sont les suivantes :

Menu SPSS : → Data
→ Select Cases...
Dans la fenêtre Select : Based on time or case range
Dans le bouton Range... : Observation : First Case 1 Last Case 150

13.6.1 Vérification et obtention de la stationnarité

La première étape consiste à s'assurer que la série à étudier est stationnaire, et si ce n'est pas le cas, à la transformer pour qu'elle le soit. La figure 13.12 présente le graphe séquentiel de la série. Elle ne semble pas présenter de tendance, et les variations sont assez semblables du début à la fin de la série, elle semble donc stationnaire.

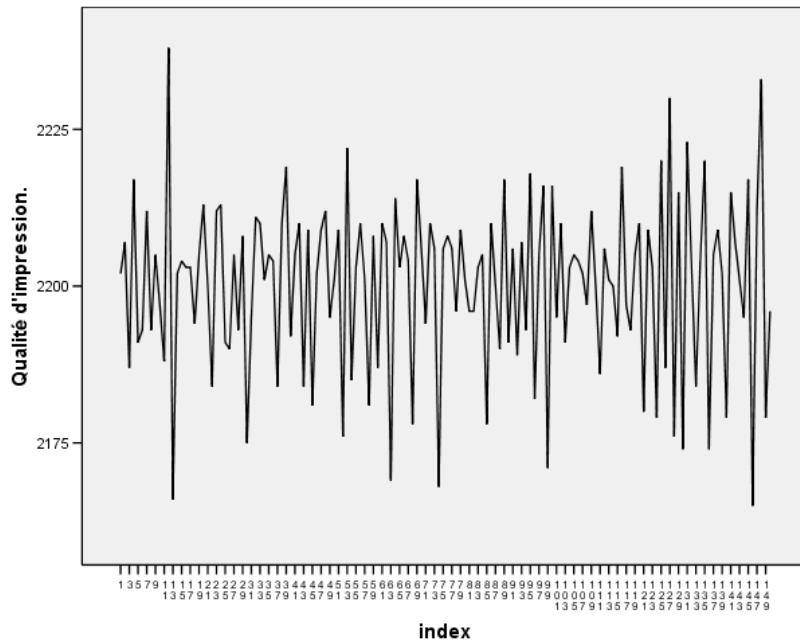


FIG. 13.12 – Le graphe de la série

Autocorrelations

Series: Qualité d'impression.

Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	-.516	,082	40,772	1	,000
2	-,035	,101	40,958	2	,000
3	,107	,101	42,724	3	,000
4	-,011	,102	42,743	4	,000
5	-,062	,102	43,353	5	,000
6	,039	,102	43,588	6	,000
7	,003	,102	43,589	7	,000
8	-,020	,102	43,653	8	,000
9	,026	,102	43,760	9	,000
10	-,061	,102	44,375	10	,000
11	,052	,103	44,826	11	,000
12	-,020	,103	44,895	12	,000
13	-,019	,103	44,955	13	,000
14	,018	,103	45,006	14	,000
15	-,015	,103	45,043	15	,000
16	,059	,103	45,641	16	,000

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

Qualité d'impression.

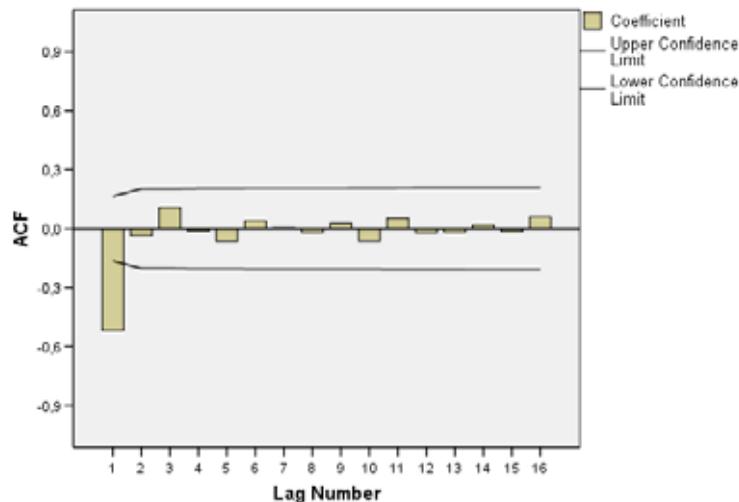


FIG. 13.13 – Les autocorrélations de la série

Le SAC de la série vient appuyer le fait qu'elle semble stationnaire : en effet, on voit dans la figure 13.13 que seule la première des autocorrélations est significative, toutes les autres tendent vers zéro. Il n'est donc pas nécessaire ici de différentier la série, donc le modèle sera de type $\text{ARMA}(p, q) = \text{ARIMA}(p, 0, q)$.

13.6.2 Identification des ordre p et q du modèle ARMA

Puisque la série est stationnaire, il est possible de procéder à l'identification des ordres des modèles autorégressif et de moyenne mobile à l'aide du SAC et du SPAC. On a déjà observé le SAC dans la figure 13.13 ; celui-ci présente une autocorrelation significative, puis toutes les autres sont non-significatives. La figure 13.14 présente le SPAC ; on voit que les autocorrélations partielles décline de façon exponentielle vers zéro. Ce comportement semble correspondre à la figure 13.8 e), et donc il semble que ce soit un modèle $\text{MA}(1) = \text{ARIMA}(0,0,1)$ qui soit adéquat pour cette série.

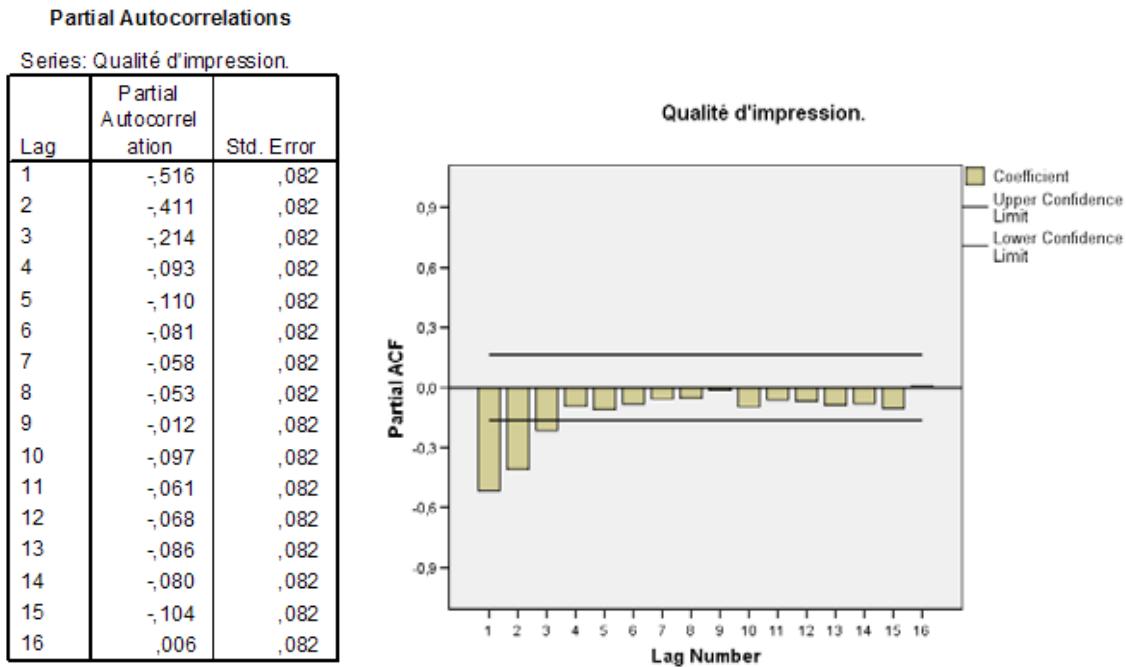


FIG. 13.14 – Les autocorrélations partielles de la série

13.6.3 Estimation des paramètres du modèle

Nous voulons ici estimer les paramètres du modèle ARIMA(0,0,1) pour nos mesures de qualité d'impression. Nous établirons aussi des estimations jusqu'à la donnée 180 afin de pouvoir juger de la qualité du modèle. Les commandes sont les suivantes :

Menu SPSS :	→ Analyse
	→ Time Series
	→ ARIMA...
Dans la fenêtre Dependent :	measure
Dans la fenêtre Model :	Autoregressive p : 0
	Difference d : 0
	Moving Average q : 1
Dans le bouton Save... :	Dans la fenêtre Predict Cases
	Predict through : Observation : 180

La figure 13.15 présente l'interface SPSS lors de ces commandes.

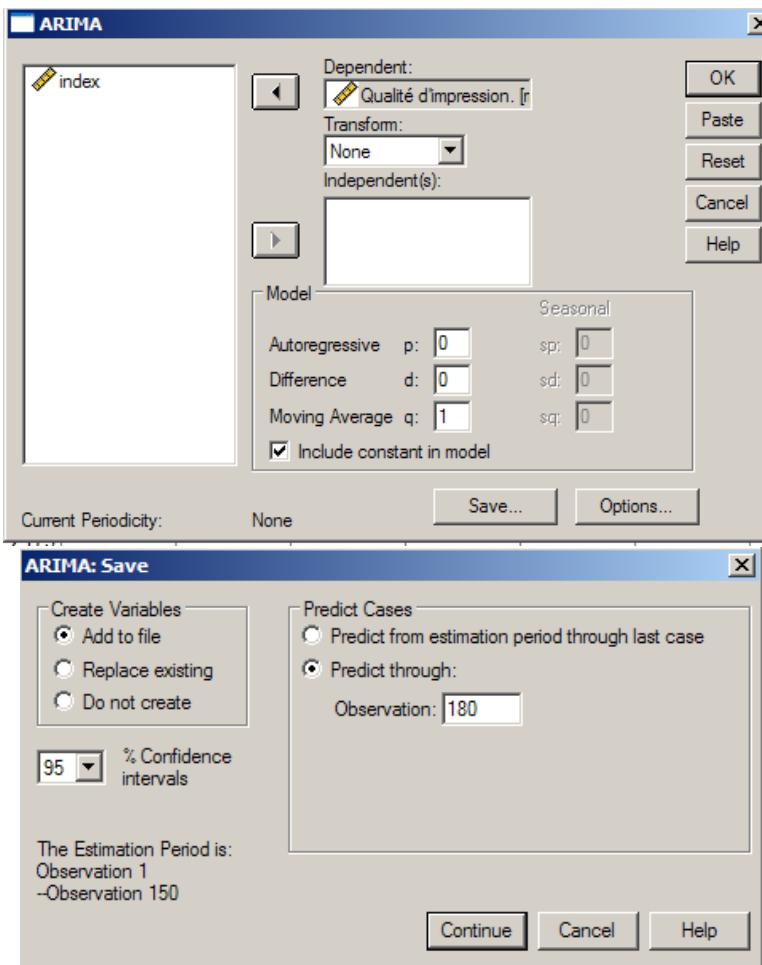


FIG. 13.15 – Pour obtenir le modèle ARIMA(0,0,1) et les prédictions

Les premières sorties sont celles des la figure 13.16. La sortie **Model Description** nous montre simplement que ce modèle ne contient pas de partie autorégressive, qu'il n'y a pas de différentiation et que l'ordre du modèle de moyenne mobile est 1.

La sortie **Iteration Termination Criteria** montre les critères qui font que le processus itératif pour l'estimation des paramètres se termine. Ceci correspond à ce qu'on indique dans le bouton **Options....** Par exemple, si d'une itération à une autre le changement dans la valeur des paramètres est de moins de 0,001, alors l'algorithme s'arrête. Il y a aussi une tolérance pour l'ampleur de la constante dans l'équation (1E+009). Un autre critère est le changement dans la somme des carrés des erreurs ; si d'une itération

à une autre elle baisse de moins de 0,001 %, alors l'algorithme s'arrête. Finalement, le nombre maximal d'itération est ici de 200 ; par défaut il est fixé à 10, mais ici je l'ai changé à 200 (ce qui était bien inutile d'ailleurs puisque l'algorithme s'est arrêté après une itération...).

Model Description ^a		
Model Name	MOD_4	
Dependent Series	Qualité d'impression.	
Transformation	None	
Constant	Included	
AR	None	0
Non-Seasonal Differencing		
MA	1	
Applying the model specifications from MOD_4		
a. Since there is no seasonal component in the model, the seasonality of the data will be ignored.		
Iteration Termination Criteria		
Maximum Parameter Change Less Than	,001	
Maximum Marquardt Constant Greater Than	1E+009	
Sum of Squares Percentage Change Less Than	,001%	
Number of Iterations Equal to	200	

Case Processing Summary		
Series Length		150
Number of Cases Skipped Due to Missing Values	At the Beginning of the Series	0
	At the End of the Series	0
Number of Cases with Missing Values within the Series		0 ^b
Number of Forecasted Cases		30
Number of New Cases Added to the Current Working File		0

a. Melard's Algorithm will be used for estimation.

FIG. 13.16 – Description du modèle

La sortie **Case Processing Summary** montre le nombre de données de la série, le nombre de données manquantes, et le nombre de prédictions.

La sortie **Iteration History** de la figure 13.17 nous montre que l'algorithme d'estimation s'est arrêté après une itération, car la somme du carré des erreurs n'a diminué que de 0,001 %.

La sortie **Residual Diagnostics** présente des mesures permettant de comparer ce modèle à un autre. En effet, si suite à l'examen du SAC et du SPAC plus d'un modèle

Requested Initial Configuration						
Non-Seasonal Lags		MA1	AUTO			
Constant			AUTO ^a			
a. The prior parameter value is invalid and is reset to 0.1.						
Iteration History						
	Non-Seasonal Lags	Constant	Adjusted Sum of Squares	Marquardt Constant		
0	,704	2200,309	16181,744	,001		
1	,843	2200,300	15687,994 ^a	,001		

Melard's algorithm was used for estimation.

a. The estimation terminated at this iteration, because the sum of squares decreased by less than ,001%.

Residual Diagnostics				
Number of Residuals				150
Number of Parameters				1
Residual df				148
Adjusted Residual Sum of Squares				15687,988
Residual Sum of Squares				16181,744
Residual Variance				105,125
Model Std. Error				10,253
Log-Likelihood				-561,593
Akaike's Information Criterion (AIC)				1127,186
Schwarz's Bayesian Criterion (BIC)				1133,207

Parameter Estimates				
	Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	MA1	,843	,045	18,718
Constant		2200,300	,136	16175,496

Melard's algorithm was used for estimation.

FIG. 13.17 – Estimation des paramètres

semble plausible, il est alors préférable d'estimer ces modèles puis de les comparer entre eux. Plus le Log-Likelihood est près de 0, meilleur est le modèle, et pour le AIC et le BIC, on cherche à en minimiser la valeur.

La sortie Parameter Estimates contient les estimés des paramètres, de même que leur cote-*t* et leur *p*-value. Ici on voit que le coefficient du terme e_{t-1} du modèle de moyenne mobile est significatif (*p*-value nulle, cote-*t* de 18,718). L'équation de notre modèle est

$$\hat{y}_t = 2\,200,3 - 0,843e_{t-1}.$$

Aussi, dans la base de données, cinq nouvelles variables sont créées (voir la figure 13.18) :

FIT_1 : Ce sont les estimations produites par le modèle. Donc par exemple la prédiction pour la période 147 est de 2 218,06363.

ERR_1 : Ce sont les erreurs d'estimations, qui sont obtenues en faisant la différence entre **mesure** et **FIT_1** (les observations et les estimations). Par exemple, $e_{146} = y_{146} - \hat{y}_{146} = 2\,165 - 2\,186,06076 = -21,06076$.

LCL_1 : Borne inférieure de l'intervalle de prédiction de niveau 95 % (il est possible de changer ce niveau dans le bouton **Save...**).

UCL_1 : Borne supérieure de l'intervalle de prédiction de niveau 95 %. Par exemple, l'intervalle de prédiction de niveau 95 % pour la période 147 est [2 197,7297, 2 238,3976].

SEP_1 : C'est l'erreur type associée à l'estimation (à partir de laquelle est construit l'intervalle de prédiction).

	index	mesure	FIT_1	ERR_1	LCL_1	UCL_1	SEP_1
146	146	2165	2186,06076	-21,0608	2165,7268	2206,3947	10,28982
147	147	2212	2218,06363	-6,06363	2197,7297	2238,3976	10,28982
148	148	2233	2205,41415	27,58585	2185,0802	2225,7481	10,28982
149	149	2179	2177,03220	1,96780	2156,6983	2197,3661	10,28982
150	150	2196	2198,63997	-2,63997	2178,3060	2218,9739	10,28982
151	151	2197	2202,52644	-5,52644	2182,1925	2222,8604	10,28982
152	152	2209	2200,29973	8,70027	2173,7922	2226,8072	13,41389
153	153	2188	2200,29973	-12,2997	2173,7922	2226,8072	13,41389
154	154	2196	2200,29973	-4,29973	2173,7922	2226,8072	13,41389
155	155	2227	2200,29973	26,70027	2173,7922	2226,8072	13,41389
156	156	2176	2200,29973	-24,2997	2173,7922	2226,8072	13,41389

FIG. 13.18 – Aperçu des nouvelles variables

Donc, par exemple,

$$\begin{aligned}
 \hat{y}_{147} &= 2\,200,3 - 0,843 \cdot e_{146} \\
 &= 2\,200,3 - 0,843 \cdot (-21,0608) \text{ (voir ligne 146 de la figure 13.18)} \\
 &= 2\,218,05 \approx 2\,218,06363.
 \end{aligned}$$

Mais avant d'utiliser les prédictions issues du modèle, il faut s'assurer que celui-ci est valide. S'il ne l'est pas, il faut recommencer les étapes précédentes afin d'identifier un meilleur modèle.

13.6.4 Validité du modèle

Une fois le modèle estimé, on vérifiera sa validité en étudiant ses résidus. Ceux-ci sont supposés être petits et aléatoirement distribués selon une loi normale. Si les résidus ne sont pas aléatoires, ceci signifie qu'il reste de l'information dans les résidus qui n'a pas été incluse dans le modèle. Il est alors possible d'améliorer le modèle, et la façon de l'améliorer nous est habituellement indiquée par le SAC et le SPAC des résidus.

La première étape consiste à étudier les statistiques de Box-Ljung qui sont données avec les autocorrélations (voir figure 13.19). On effectue un test du chi-deux (χ^2) basé sur cette statistique. Ce test évalue l'ampleur des autocorrélations en tant que groupe ; en effet, une autocorrélation pourrait être significative, mais si ce n'est qu'un effet du hasard la statistique de Box-Ljung ne réagira pas et nous indiquera que le modèle est adéquat.

Cette statistique est notée \mathcal{Q} et se calcule de la façon suivante :

$$\mathcal{Q}_m = n(n + 2) \sum_{k=1}^m \frac{r_k^2(e)}{n - k}$$

où

- r_k^2 est l'autocorrélation des résidus pour le *lag* k ;
- n est le nombre de résidus ;
- m est le nombre de *lags* à tester.

La statistique de Box-Ljung suit une loi du chi-deux ; on rejette l'adéquation du modèle si l'une des p -values associées aux \mathcal{Q}_m est plus petite que le seuil de signification α .

Si l'adéquation du modèle est rejetée, c'est l'examen des pics dans le SAC et le SPAC des résidus qui peuvent nous indiquer comment améliorer le modèle. Par exemple, s'il

reste un pic significatif au *lag* 6 du SPAC, on pourrait alors tenter d'ajouter un modèle autorégressif d'ordre 6 au modèle initialement choisi. Et même si l'adéquation du modèle n'est pas rejetée par la statistique de Box-Ljung, on peut quand même tenter d'améliorer le modèle en examinant le SAC et le SPAC des résidus.

Dans le cadre de l'exemple, on retrouve le SAC des résidus dans la figure 13.19, et le SPAC des résidus dans la figure 13.20.

Aucune autocorrélation et aucune autocorrélation partielle n'est significative, ce qui semble indiquer que les résidus sont aléatoires. Cette affirmation est appuyée par les *p*-values des Box-Ljung de la sortie 13.19 ; en effet, celles-ci sont toutes largement supérieures à 0,05, la plus petite étant de 0,384 au *lag* 3. L'adéquation du modèle n'est donc pas rejetée.

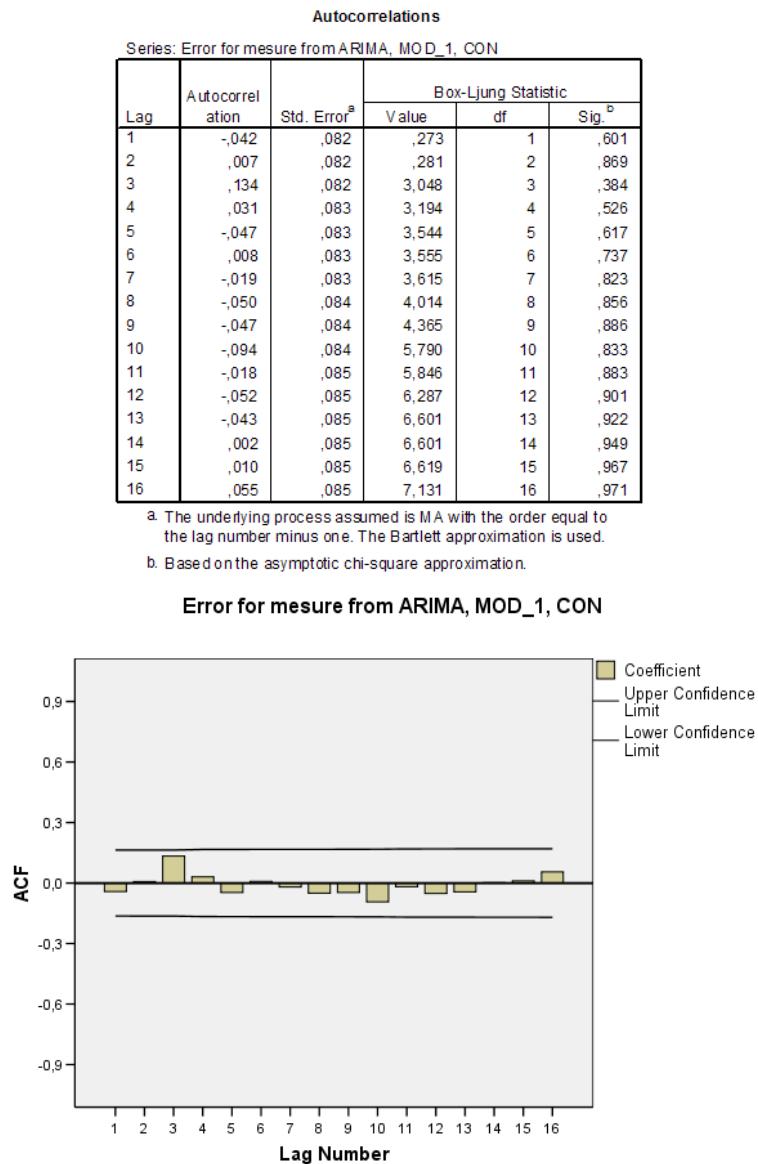


FIG. 13.19 – Le SAC des résidus

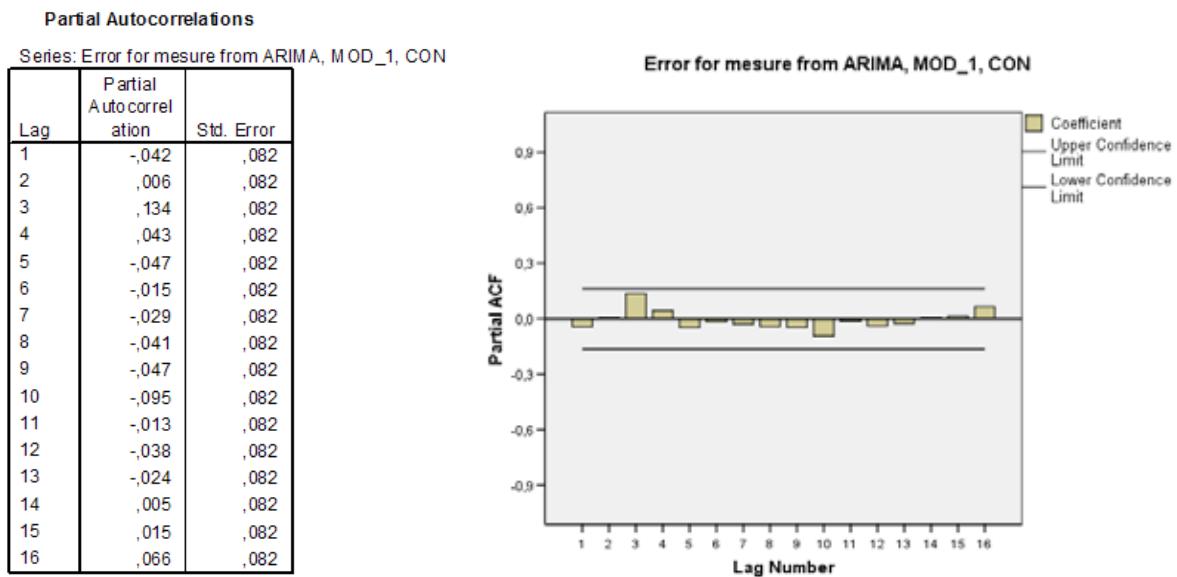


FIG. 13.20 – Le SPAC des résidus

Il reste à vérifier la normalité des résidus ; il est nécessaire que les résidus soient distribués selon une loi normale pour que les intervalles de confiance pour les prédictions soient valides. Rappelons les commandes pour effectuer un test de normalité :

Menu SPSS :	→ Analyse
	→ Descriptive Statistics
	→ Explore...
Dans la fenêtre Dependent List :	→ ERR_1
Display :	→ Plots
Dans le bouton Plots... :	✓ Normality plots with tests
	✓ Histogram

On obtient alors les sorties 13.21 et 13.22. Rappelons que les statistiques de Shapiro-Wilk et Kolmogorov-Smirnov nous permettent de résoudre le test suivant :

H_0 : Les résidus se distribuent selon une loi normale au niveau de la population.

H_1 : Les résidus ne se distribuent pas selon une loi normale au niveau de la population.

	Tests of Normality			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Error for measure from ARIMA, MOD_1, CON	,059	150	,200*	,994	150	,792

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. 13.21 – Test de normalité des résidus

Puisque 0,2 et 0,792 sont plus grandes que 0,05, on ne rejette pas H_0 . Ainsi au seuil $\alpha = 0,05$ on admet que les résidus suivent une loi normale.

Les sorties de la figure 13.22 viennent appuyer cette conclusion. En effet, l'histogramme des résidus nous montre clairement la forme de cloche de la distribution normale, et l'autre graphe illustre que les résidus se collent bien à la droite de la normalité théorique.

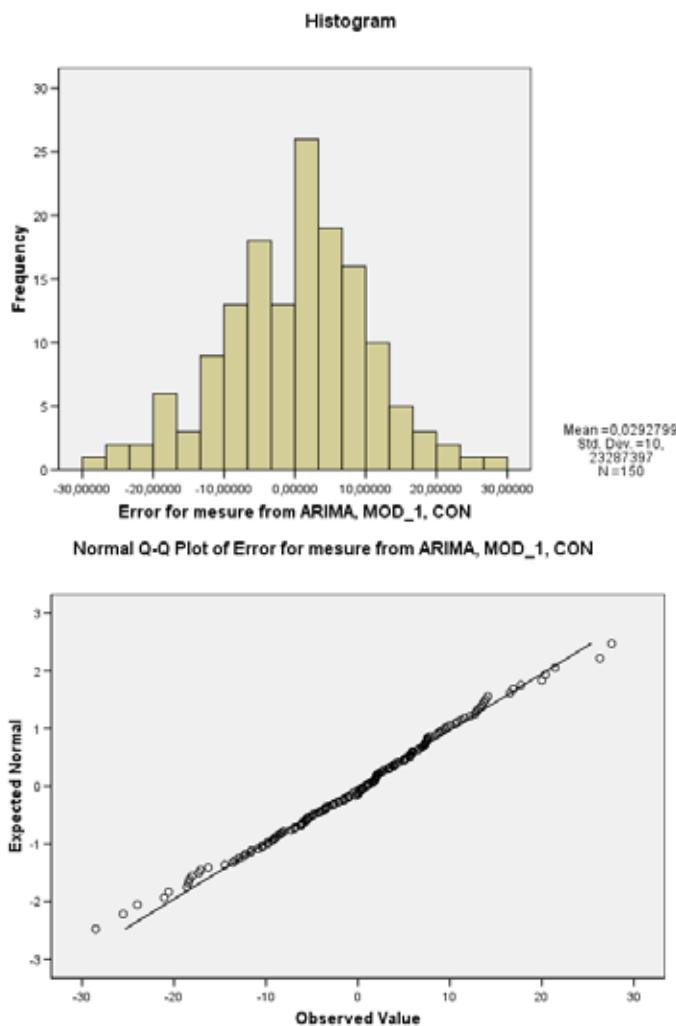


FIG. 13.22 – Histogramme et graphe de normalité des résidus

Finalement, la figure 13.23 montre la répartition des résidus en fonction du temps (les deux graphes sont équivalents, c'est juste que dans le graphe séquentiel les résidus consécutifs sont reliés). Ces graphes montrent que la répartition semble aléatoire, et il n'y a pas de *outliers* (si on refait le graphe mais avec les résidus standardisés, on peut voir que tous les points sont entre ± 3 écarts-type).

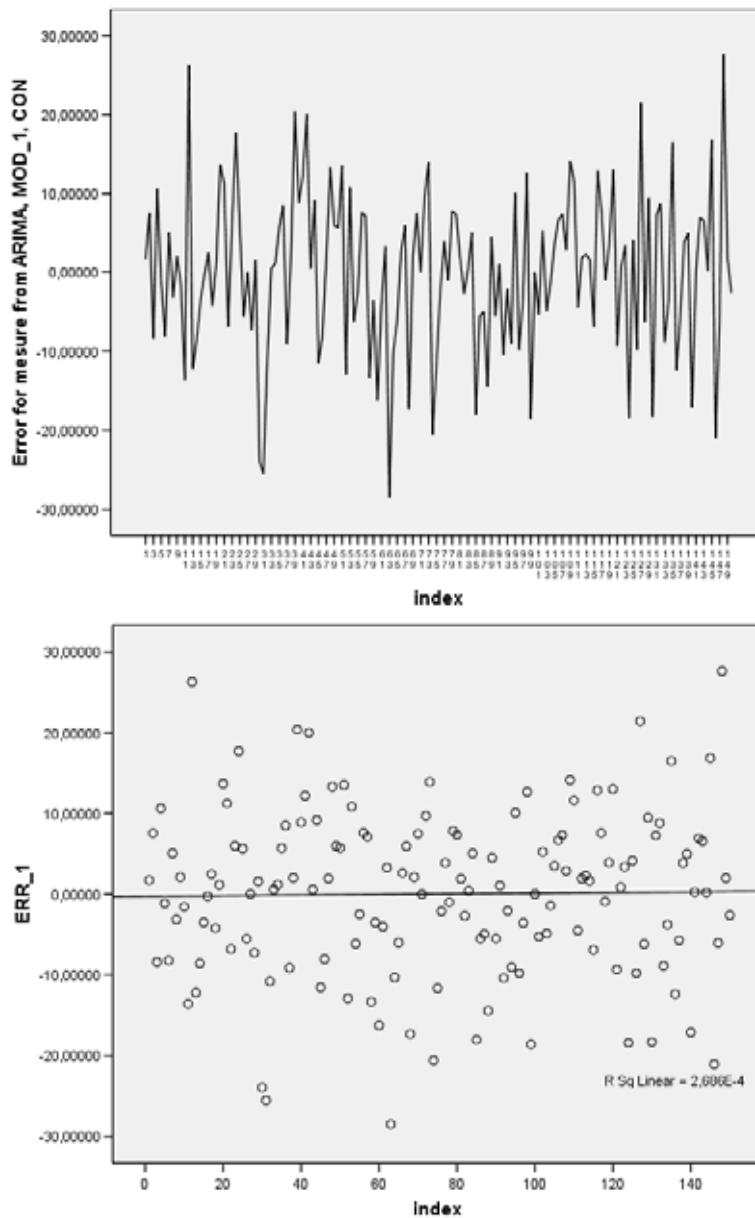


FIG. 13.23 – Répartition des résidus en fonction du temps

13.6.5 Calcul des prédictions

Les commandes effectuées à la sous-section 13.6.3 ont déjà généré les prédictions jusqu'à la période 180 ainsi que les intervalles de prédiction de niveau 95 %. La validité du modèle a été établie pour les 150 premières observations ; il est maintenant temps d'observer comment s'est comporté le modèle pour les observations 150 à 180.

D'abord, pour que SPSS considère à nouveau l'ensemble des données, il faut sélectionner **All cases** dans le menu **Select Cases...** du menu **Data**.

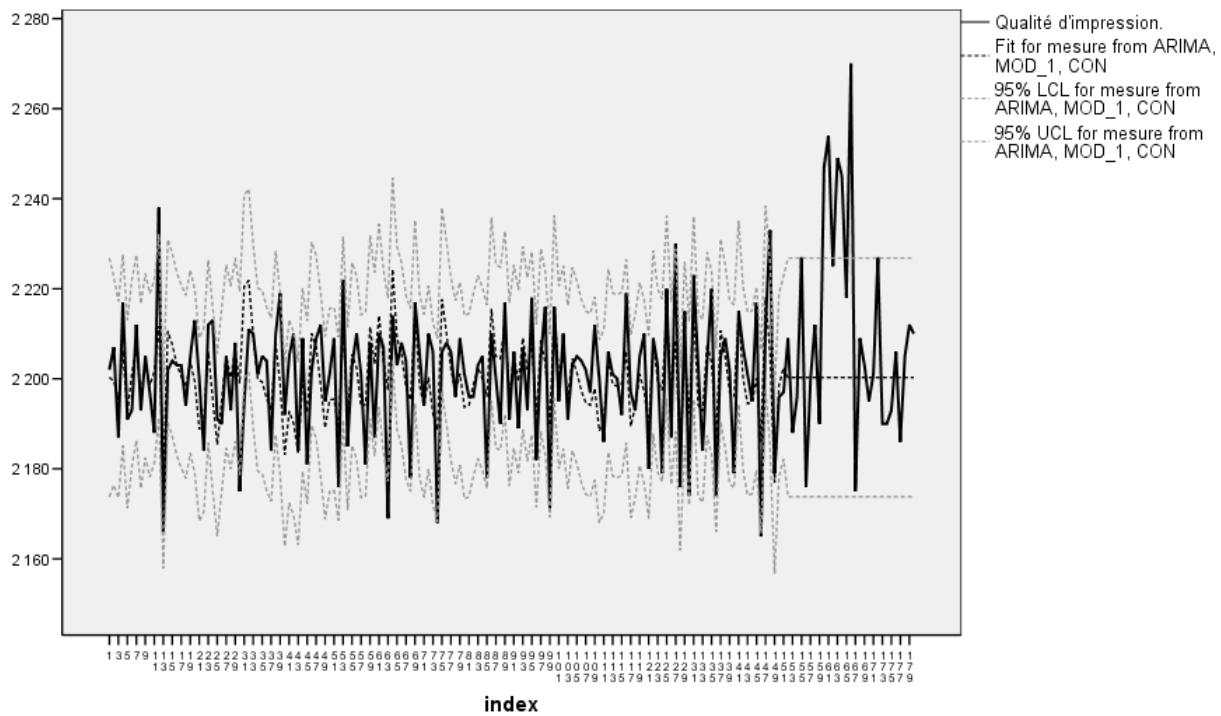


FIG. 13.24 – Les observations versus les estimations pour la série complète

La figure 13.24 présente les observations, les estimations et les bornes des intervalles de prédiction pour l'ensemble de la série. Pour mieux distinguer la période qui nous intéresse, on décide de sélectionner les données 145 à 180 puis de refaire le graphe séquentiel, ce qui donne la figure 13.25.

On observe d'abord que les prédictions pour les périodes 151 à 180 sont constantes (elles sont de 2 200,3), ce qui est toujours le cas avec un modèle n'ayant qu'une compo-

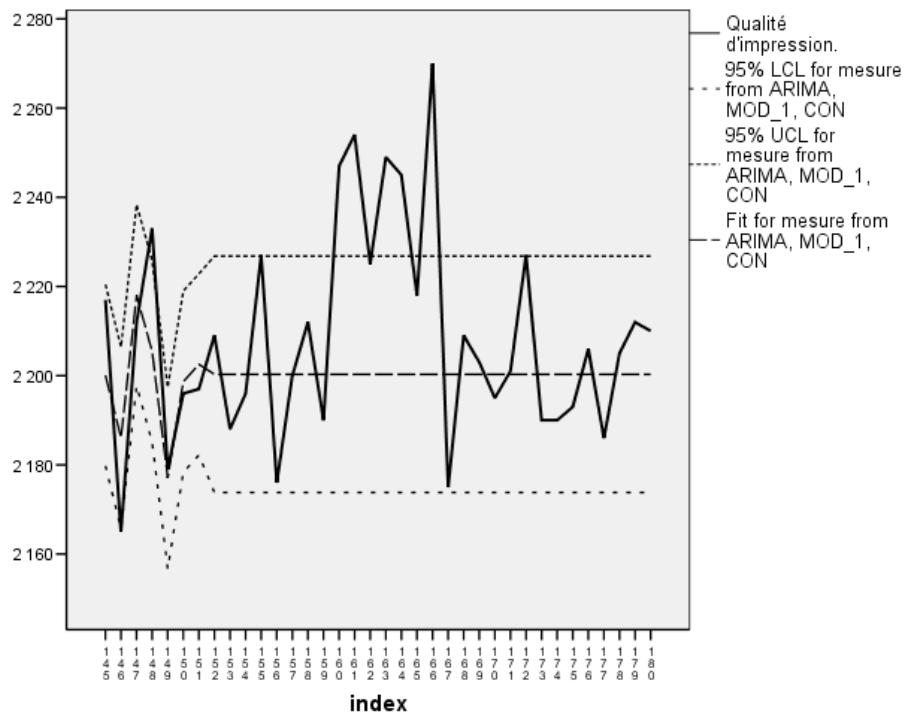


FIG. 13.25 – Les observations versus les estimations pour les périodes 145 à 180

sante de MA. Cette constante est simplement la constante ω_0 du modèle $Y_t = \omega_0 + \epsilon_t - \omega_1 \epsilon_{t-1}$. Ceci est dû au fait que les termes d'erreur étant inconnus, ils sont estimés à 0, qui est leur espérance mathématique.

On observe aussi que de la période 160 à 166 les observations dépassent la borne supérieure des intervalles de confiance. Donc si ce modèle est utilisé pour contrôler la qualité, on peut imaginer que ces « débordements » ont alerté les ingénieurs qui ont ensuite corrigé la situation. Cet exemple illustre comment un modèle ARIMA tient un processus sous contrôle.

Cependant, en général, l'analyste n'utilise pas le modèle ARIMA pour générer des bornes de confiance comme en contrôle de la qualité. L'analyste tente plutôt de produire les meilleures prédictions possible en temps pratiquement réel. Pour ce faire, lorsqu'une nouvelle donnée est observée, on peut alors calculer un nouveau terme d'erreur et l'utiliser pour refaire une prédition. Dans l'exemple, lorsque la 151^e observation a été disponible,

il était alors possible de calculer e_{151} , puis de l'utiliser pour calculer $\hat{y_{152}} = 2\,200,3 - 0,843e_{151}$.

Par contre cette façon de faire n'actualise pas le modèle dans son entier. Pour une prédiction encore plus « actuelle », il faudrait estimer à nouveau les paramètres du modèle en tenant compte de la nouvelle donnée.

Ainsi s'achève la présentation des étapes de la méthodologie de Box-Jenkins. Voici maintenant d'autres exemples.

Exemple 13.6.1 La base de données `entreprise.sav` contient un ensemble de données sur lesquelles nous voulons appliquer un modèle ARIMA. Ces données proviennent d'une entreprise qui désire, via un modèle mathématique, des prédictions pour la variable Y pour les périodes 96 à 100. On fixe les seuils à $\alpha = 0,05$.

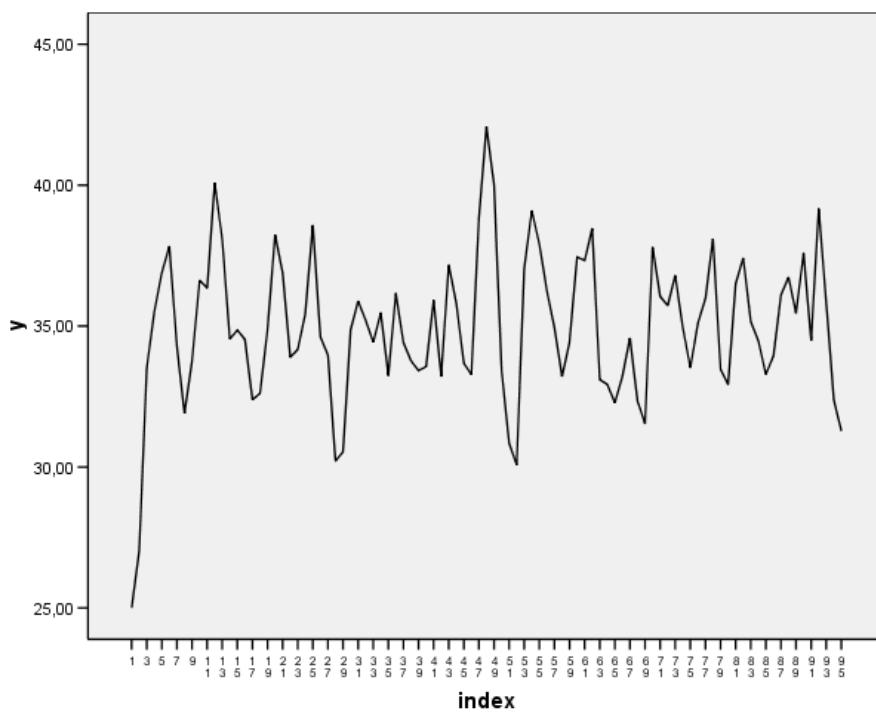


FIG. 13.26 – Graphe de la série

La figure 13.26 présente le graphe séquentiel de la série. Celle-ci ne semble pas présenter de tendance, et la variation est assez constante. La série semble donc stationnaire.

Autocorrelations					
		Box-Ljung Statistic			
Lag	Autocorrelation	Std. Error ^a	Value	df	Sig. ^b
1	,438	,103	18,827	1	,000
2	-,112	,121	20,069	2	,000
3	-,343	,122	31,839	3	,000
4	-,240	,132	37,696	4	,000
5	,003	,136	37,697	5	,000
6	,195	,136	41,642	6	,000
7	,115	,139	43,036	7	,000
8	-,047	,140	43,275	8	,000
9	-,136	,140	45,268	9	,000
10	-,110	,142	46,580	10	,000
11	-,032	,142	46,689	11	,000
12	,021	,143	46,737	12	,000
13	,039	,143	46,908	13	,000
14	,026	,143	46,987	14	,000
15	-,017	,143	47,021	15	,000
16	,010	,143	47,032	16	,000

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

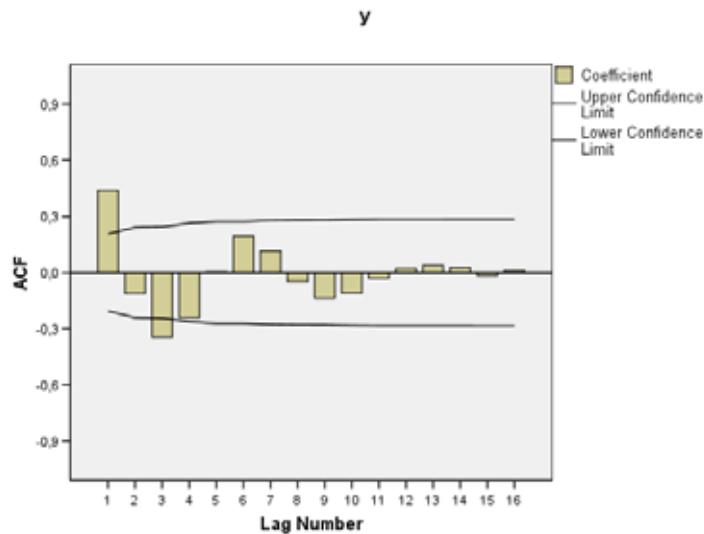


FIG. 13.27 – Le SAC

Cette affirmation est appuyée par les autocorrélations de la série que l'on retrouve dans le SAC (figure 13.27). En effet, les autocorrélations s'estompent suffisamment rapidement, on peut donc considérer que la série est stationnaire. Il n'est donc pas nécessaire d'appliquer une différentiation, ni de transformation logarithmique.

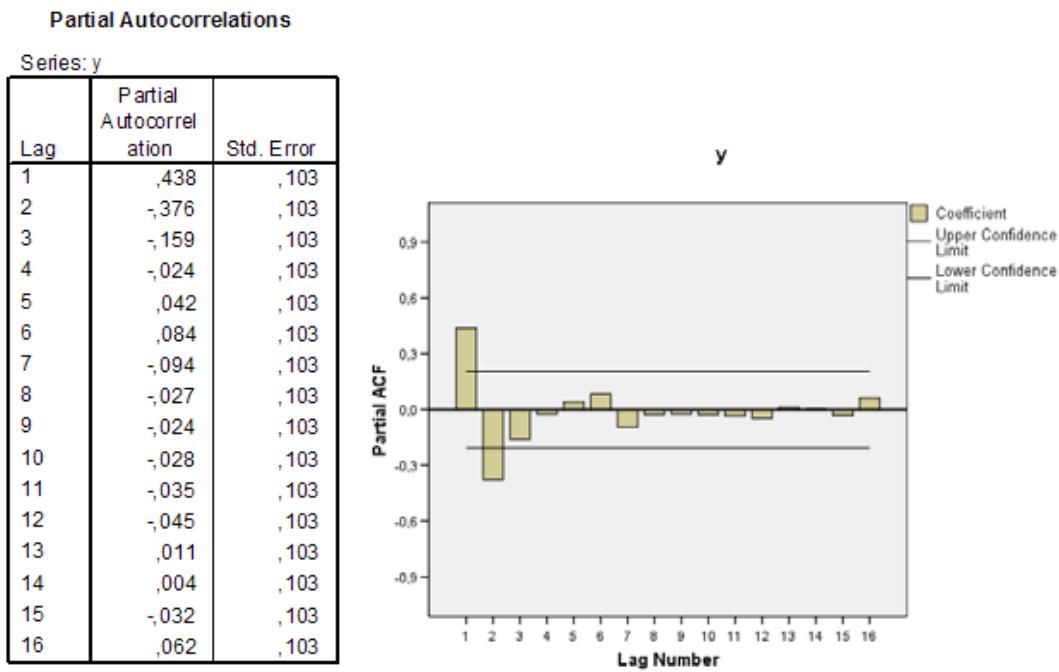


FIG. 13.28 – Le SPAC

La figure 13.28 présente le SPAC de la série. Puisque la série est considérée stationnaire, on peut procéder à l'identification du modèle ARIMA en examinant le SAC et le SPAC de la série.

Ici il semble que ce soient les autocorrélations (SAC) qui décroissent de façon exponentielle, et que ce soient les autocorrélations partielles (SPAC) qui, après deux pics significatifs, s'estompent brutalement. Ce comportement s'apparente à celui de la figure 13.8 d), ce qui suggère un modèle $AR(2) = ARIMA(2,0,0)$. On procède donc à l'estimation de ce modèle.

La première sortie de la figure 13.29 nous montre que l'algorithme a convergé en deux étapes. La deuxième sortie nous montre les mesures qui sont utiles pour comparer des

Iteration History					
	Non-Seasonal Lags		Constant	Adjusted Sum of Squares	Marquardt Constant
	AR1	AR2			
0	,603	-,376	34,942	437,833	,001
1	,681	-,429	34,945	434,558	,001
2	,682	-,432	34,946	434,550 ^a	,000

Melard's algorithm was used for estimation.

- a. The estimation terminated at this iteration, because the sum of squares decreased by less than ,001%.

Residual Diagnostics

Number of Residuals	95
Number of Parameters	2
Residual df	92
Adjusted Residual Sum of Squares	434,549
Residual Sum of Squares	437,833
Residual Variance	4,690
Model Std. Error	2,166
Log-Likelihood	-207,034
Akaike's Information Criterion (AIC)	420,069
Schwarz's Bayesian Criterion (BIC)	427,730

Parameter Estimates

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	AR1	,682	,098	6,944	,000
	AR2	-,433	,094	-4,595	,000
	Constant	34,946	,297	117,772	,000

Melard's algorithm was used for estimation.

FIG. 13.29 – Estimations des paramètres du modèles ARIMA(2,0,0)

modèles entre eux. Rappelons que plus le Log-Likelihood est près de 0, mieux c'est, et que ce sont de petites valeurs qui sont préférables pour le AIC et le BIC. La troisième sortie nous donne les estimations des paramètres. Les *p*-values étant nulles, on voit que les coefficients des termes y_{t-1} et y_{t-2} sont significatifs et ont une valeur de 0,682 et -0,433 respectivement.

	index	y	FIT_1	ERR_1
1	1	25,00	34,94621	-9,94621
2	2	27,00	30,21032	-3,21032
3	3	33,51	33,83177	-.32177
4	4	35,50	37,40783	-1,90783
5	5	36,90	35,94615	.95385

FIG. 13.30 – Aperçu des observations et estimations

Attention : lorsqu'il y a une partie autorégressive non-nulle dans un modèle ARIMA, SPSS ne donne pas la bonne valeur de la constante pour l'équation du modèle. La constante donnée correspond en fait à la prédiction pour la première observation de la série. Par contre, les prédictions établies par SPSS sont correctes. Pour obtenir la bonne valeur de la constante, observons d'abord (sortie 13.30) que nous avons

$$\begin{aligned}\hat{y}_3 &= 33,83177 = \phi_0 + 0,68239 \cdot y_2 - 0,43312 \cdot y_1 \\ &= \phi_0 + 0,68239 \cdot 27 - 0,43312 \cdot 25 \\ &= \phi_0 + 7,59653\end{aligned}$$

donc $\phi_0 = 33,83177 - 7,59653 = 26,23524$.

Ainsi l'équation du modèle est

$$\hat{y}_t = 33,83177 = 26,23524 + 0,68239y_{t-1} - 0,43312y_{t-2}.$$

On doit maintenant s'assurer que ce modèle est valide. L'examen du SAC des résidus ne révèle aucune autocorrélation significative, et toutes les statistiques de Box-Ljung sont non-significatives puisque la plus petite *p*-value est de 0,367.

Autocorrelations					
Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	-.091	,103	,815	1	,367
2	,079	,103	1,434	2	,488
3	-,045	,104	1,638	3	,651
4	-,073	,104	2,178	4	,703
5	-,036	,105	2,313	5	,804
6	,107	,105	3,492	6	,745
7	-,008	,106	3,498	7	,835
8	-,037	,106	3,642	8	,888
9	-,050	,106	3,908	9	,917
10	-,058	,106	4,276	10	,934
11	-,007	,107	4,282	11	,961
12	-,017	,107	4,316	12	,977
13	-,006	,107	4,320	13	,987
14	,048	,107	4,582	14	,991
15	-,059	,107	4,979	15	,992
16	,018	,107	5,019	16	,996

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

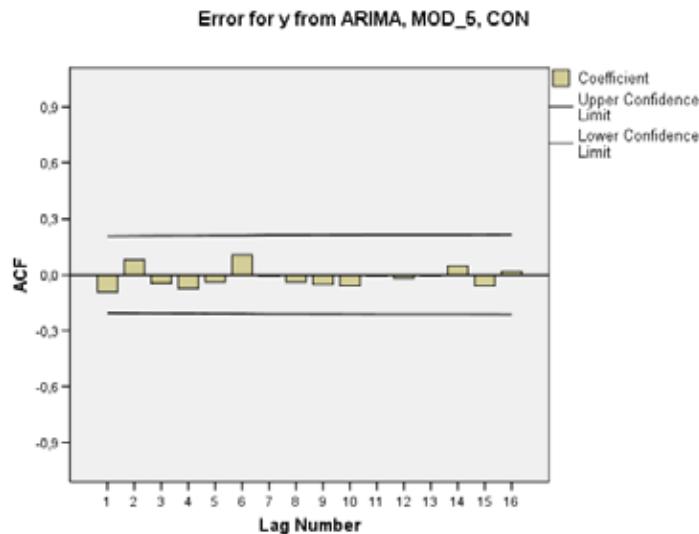


FIG. 13.31 – Le SAC des résidus

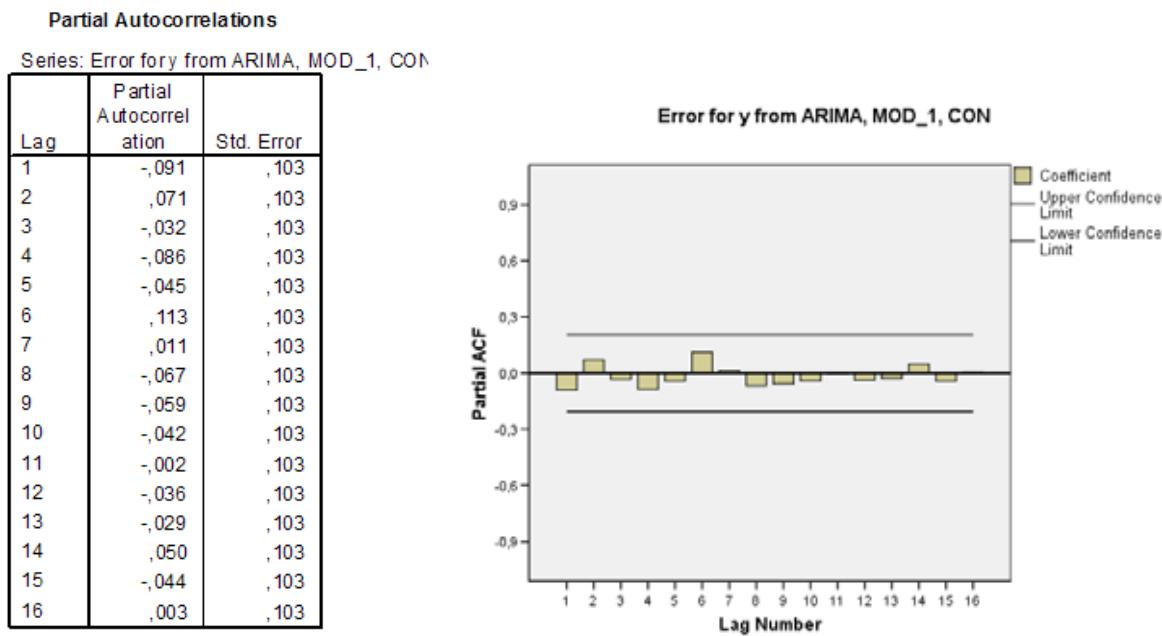


FIG. 13.32 – Le SPAC des résidus

De même, le SPAC (figure 13.32) ne révèle aucune autocorrélation partielle significative. Les résidus semblent donc être aléatoires ; mais le sont-ils selon une loi normale ?

La figure 13.33 nous montre les résultats des tests de normalité. On peut donc traiter le test suivant :

H_0 : Les résidus se distribuent selon une loi normale au niveau de la population.

H_1 : Les résidus ne se distribuent pas selon une loi normale au niveau de la population.

Le test de Kolmogorov-Smirnov ayant une p -value de 0,147, ce qui est plus grand que 0,05, on ne rejette pas H_0 , et ainsi on peut admettre que les résidus se distribuent selon une loi normale. Par contre on remarque que Shapiro-Wilk ne nous laissent pas arriver à cette même conclusion puisque la p -value est de $0,002 < 0,05$. En fait ce test est plus sensible aux *outliers*, et l'histogramme nous laisse voir qu'il y en a un (à gauche). L'autre graphe ne le montre pas, c'est un *bug* de la version 14 de SPSS.

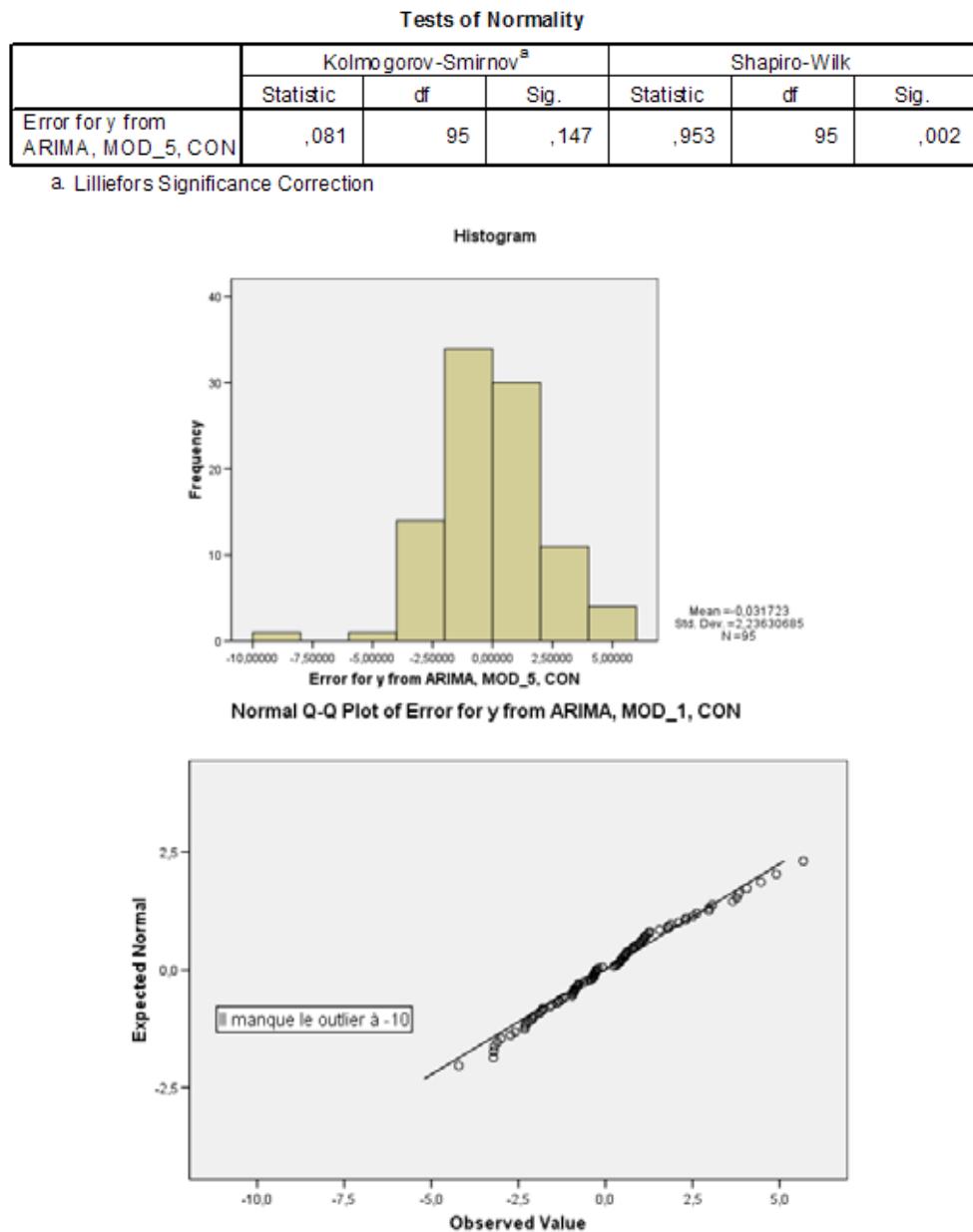


FIG. 13.33 – Test de normalité des résidus

Cette donnée aberrante se visualise bien dans le graphe 13.34. Mis à part cette donnée aberrante, l'histogramme a une forme de cloche, les résidus se collent assez bien à la droite de la normalité théorique et semblent répartis aléatoirement en fonction du temps.

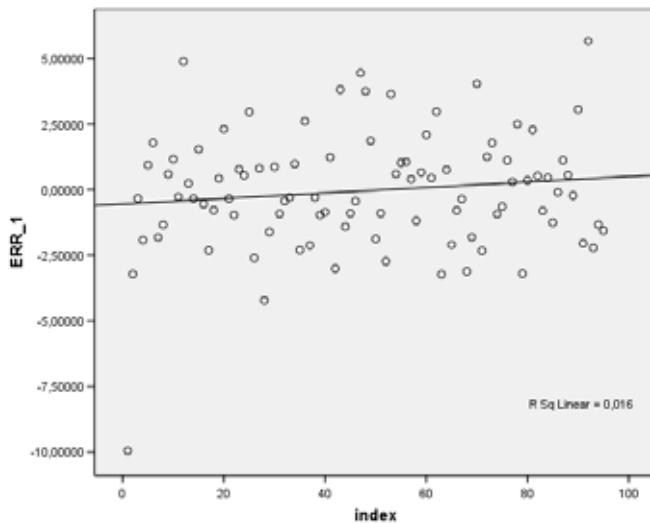


FIG. 13.34 – Répartition des résidus

L'analyste peut accepter ce modèle. Cependant, il est encouragé à refaire le modèle sans la première donnée qui fait office de *outlier* probable. L'analyste doit toujours noter les *outliers* dans une série car leur présence, comme en régression, peut grandement affecter le modèle. L'analyste peut alors comparer les deux modèles avec le Log-Likelihood et les statistiques AIC et BIC issues des deux modèles. Dans cet exemple, si on refait le même modèle mais en supprimant la première donnée de la série qui est la donnée aberrante, on note une amélioration pour ces trois mesures (qui passent à -201,915, 409,831 et 417,461 respectivement).

Il existe deux méthodes pour travailler en présence de *outliers* :

- L'analyste enlève la donnée aberrante et lance la procédure ARIMA. La procédure ARIMA supporte les données manquantes en utilisant une méthode itérative qui peut prendre un peu de temps pour converger.

- L'analyste remplace la donnée manquante à la moyenne ou encore il « intrapole » la valeur avec un modèle mathématique tel que présenté à la section 12.7.

Dans la littérature, plusieurs auteurs, même s'ils prônent que les résidus doivent se distribuer selon une loi normale, escamotent cette étape et ne base l'adéquation du modèle qu'essentiellement sur la statistique de Box-Ljung. La position de ce présent ouvrage consiste à rechercher le meilleur modèle qui ne viole pas, autant que possible, la normalité. En effet, la validité des intervalles de confiance repose sur l'hypothèse de normalité des résidus. Dans cet exemple, compte tenu que le test de Kolmogorov-Smirnov supporte la normalité, les résidus sont considérés adéquats.

On peut donc passer à l'examen des prédictions issues du modèle. La figure 13.35 montre entre autres les prédictions pour les périodes 96 à 100.

index	y	FIT_1	ERR_1	LCL_1	UCL_1	SEP_1
92	39,17	33,48982	5,68018	29,16596	37,81368	2,17708
93	35,82	38,02603	-2,20603	33,70217	42,34990	2,17708
94	32,39	33,71305	-1,32305	29,38919	38,03692	2,17708
95	31,28	32,82344	-1,54344	28,49958	37,14730	2,17708
96	.	33,55160	.	29,22774	37,87546	2,17708
97	.	35,58246	.	30,32238	40,84255	2,64847
98	.	35,98441	.	30,72041	41,24841	2,65044
99	.	35,37908	.	30,00105	40,75711	2,70785
100	.	34,79192	.	29,35478	40,22906	2,73762

FIG. 13.35 – Les prédictions

Les graphes de la figure 13.36 nous permettent de visualiser les observations versus les estimations et les bornes des intervalles de confiance (le deuxième graphe ne présente que les périodes 90 à 100). On constate qu'en général le modèle performe assez bien puisque les observations sont presque toujours entre les bornes des intervalles de confiance (mais pas toujours, voir l'observation 92 par exemple). On peut donc être assez confiant pour les prédictions des périodes 96 à 100.

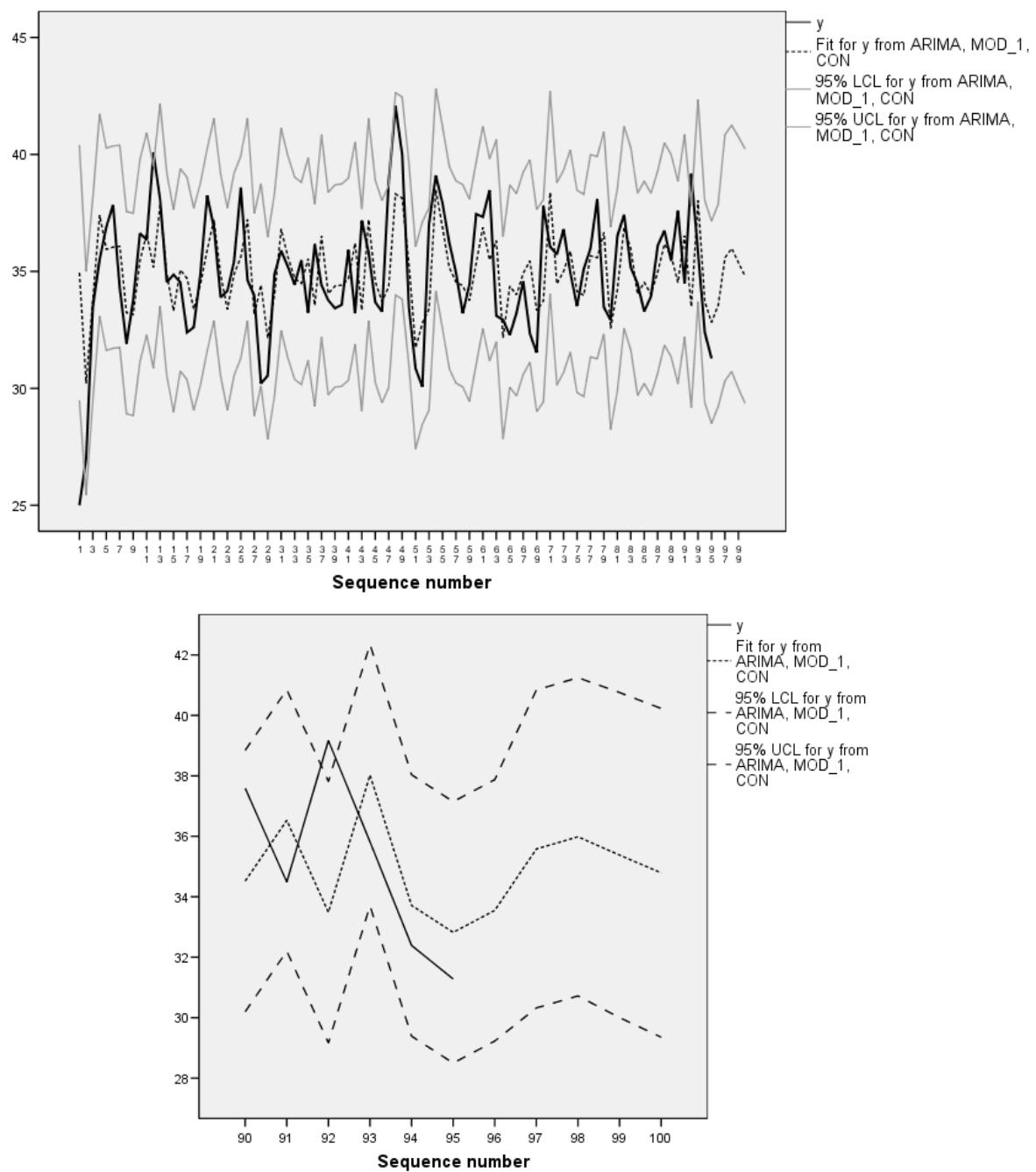


FIG. 13.36 – Graphes des observations et des estimations

Exemple 13.6.2 La base de données `dvd.sav` contient les ventes hebdomadaires d'un DVD (en milliers d'unités vendues) sur 161 semaines.

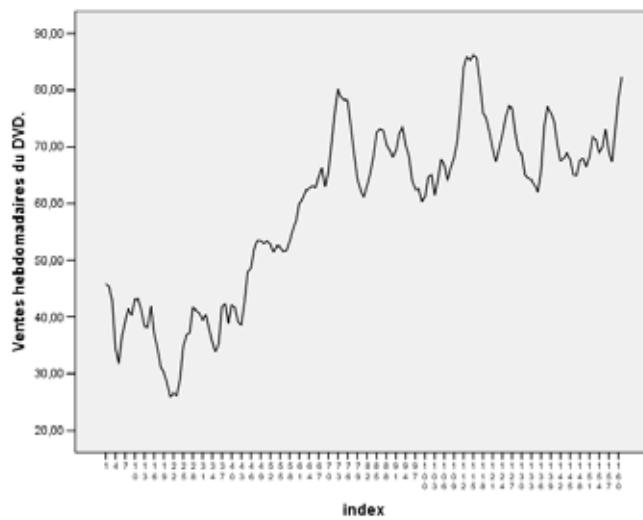


FIG. 13.37 – Les ventes hebdomadaires du DVD sur 161 semaines

On regarde d'abord le graphe séquentiel des ventes (figure 13.37). Une tendance est présente, mais aucun effet saisonnier ne semble être présent.

La figure 13.38 nous montre le SAC de la série originale. Puisqu'il y a plusieurs autocorrelations significatives, ceci confirme que la série a une tendance ; il faut donc la différentier.

Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	,974	,079	155,528	1	,000
2	,935	,134	299,771	2	,000
3	,899	,170	433,878	3	,000
4	,862	,197	558,205	4	,000
5	,825	,219	672,614	5	,000
6	,794	,238	779,331	6	,000
7	,777	,254	882,217	7	,000
8	,765	,268	982,639	8	,000
9	,753	,281	1080,655	9	,000
10	,745	,294	1177,121	10	,000
11	,739	,305	1272,719	11	,000
12	,730	,316	1366,696	12	,000
13	,715	,326	1457,264	13	,000
14	,695	,336	1543,453	14	,000
15	,676	,345	1625,499	15	,000
16	,654	,353	1702,986	16	,000

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

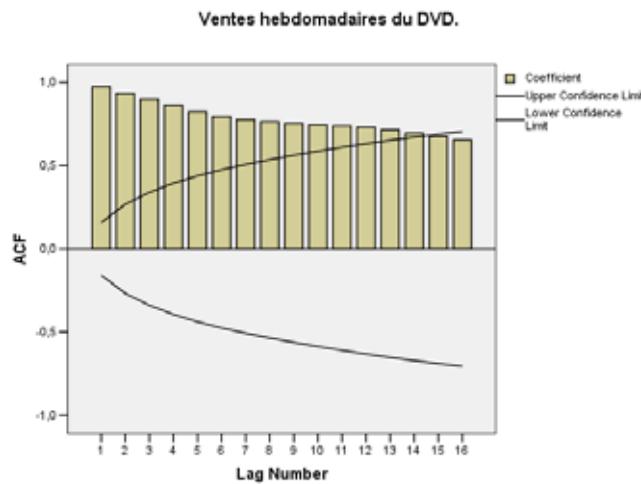


FIG. 13.38 – Les autocorrélations de la série originale (SAC)

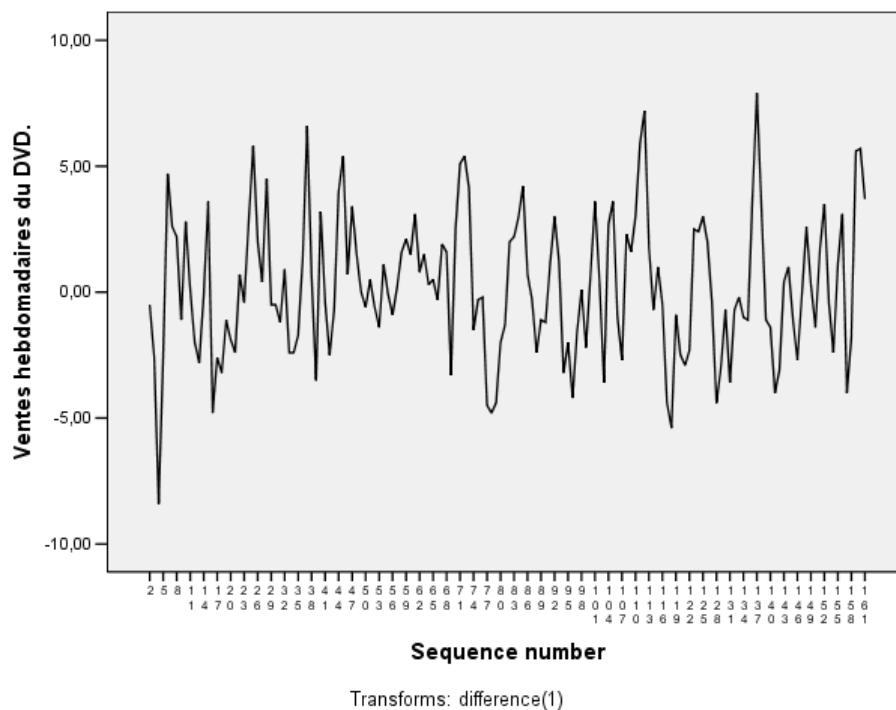


FIG. 13.39 – La série différentiée

La figure 13.39 nous montre la série différentiée. Elle semble stationnaire. Et effectivement, si on regarde le SAC de la série différentiée (figure 13.40), on voit qu'il ne reste que trois autocorrelations significatives (aux *lags* 1, 5 et 6).

On peut maintenant tenter d'identifier quel serait le modèle ARIMA approprié à l'aide du SAC et du SPAC de la série différentiée (figures 13.40 et 13.41).

Autocorrelations					
Lag	Autocorrelation	Std.Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	,435	,079	30,832	1	,000
2	-,008	,093	30,843	2	,000
3	,002	,093	30,844	3	,000
4	-,017	,093	30,893	4	,000
5	-,239	,093	40,420	5	,000
6	-,336	,097	59,370	6	,000
7	-,113	,104	61,515	7	,000
8	-,066	,104	62,260	8	,000
9	-,080	,105	63,361	9	,000
10	-,019	,105	63,420	10	,000
11	,092	,105	64,885	11	,000
12	,149	,106	68,799	12	,000
13	,120	,107	71,338	13	,000
14	,062	,108	72,014	14	,000
15	,004	,108	72,016	15	,000
16	-,103	,108	73,910	16	,000
17	-,076	,109	74,955	17	,000
18	-,033	,109	75,149	18	,000
19	-,018	,109	75,210	19	,000
20	-,079	,109	76,378	20	,000
21	-,025	,109	76,494	21	,000

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

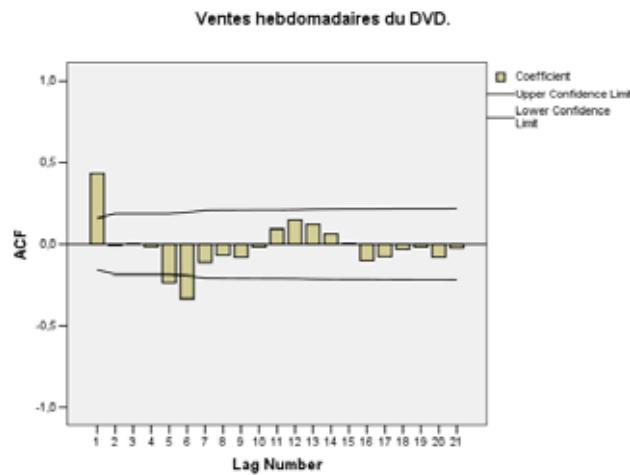


FIG. 13.40 – Les autocorrélations de la série différentiée (SAC)

La décroissance exponentielle semble plus plausible dans le SPAC. Ainsi on peut tenter un modèle de moyennes mobiles, et selon le SAC l'ordre approprié semble être 6, mais puisque dans les premières autocorrélations une seule est significative, on commençera par un modèle plus simple d'ordre 1. On tentera donc un modèle ARIMA(0,1,1), et l'analyse des résidus nous montrera si c'est suffisant.

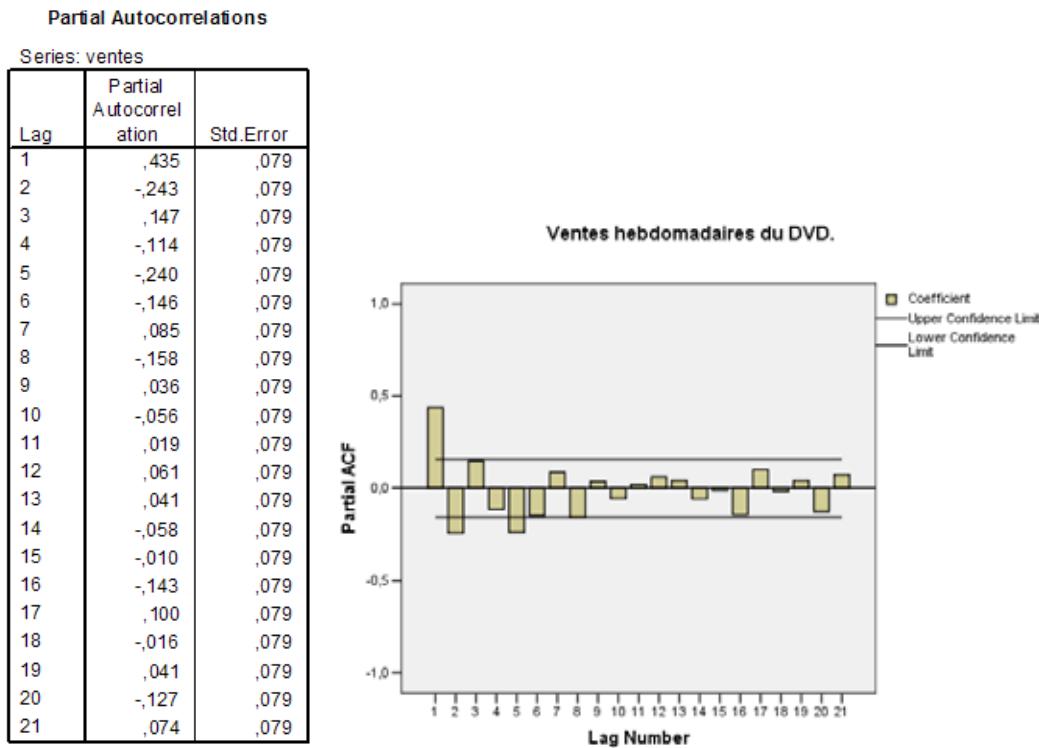


FIG. 13.41 – Les autocorrélations partielles de la série différentiée (SPAC)

Residual Diagnostics	
Number of Residuals	160
Number of Parameters	1
Residual df	158
Adjusted Residual Sum of Squares	945,345
Residual Sum of Squares	947,703
Residual Variance	5,968
Model Std. Error	2,443
Log-Likelihood	-369,144
Akaike's Information Criterion (AIC)	742,289
Schwarz's Bayesian Criterion (BIC)	748,439

Parameter Estimates				
		Estimates	Std Error	t
Non-Seasonal Lags	MA1	-,579	,065	-8,891
	Constant	,231	,304	,758

Melard's algorithm was used for estimation.

FIG. 13.42 – Les sorties du modèle ARIMA(0,1,1)

La figure 13.42 nous montre les sorties pertinentes pour l'interprétation de ce modèle. La première sortie contient les mesures qui seront utiles pour comparer ce modèle à un autre. La deuxième nous donne l'estimation des paramètres. On voit que le paramètre du terme ϵ_{t-1} est jugé significativement différent de 0 puisque sa *p*-value est nulle ($0 < 0,05$). L'équation du modèle est

$$\hat{z}_t = 0,231 + 0,579e_{t-1}$$

avec $z_t = y_t - y_{t-1}$ (où y_t représente la série originale).

La figure 13.43 nous montre les estimations des ventes et les ventes réelles (d'abord de l'index 1 à 80, puis de 81 à 161). On voit que le modèle semble donner de bonnes estimations, mais il faut analyser les résidus pour s'en assurer.

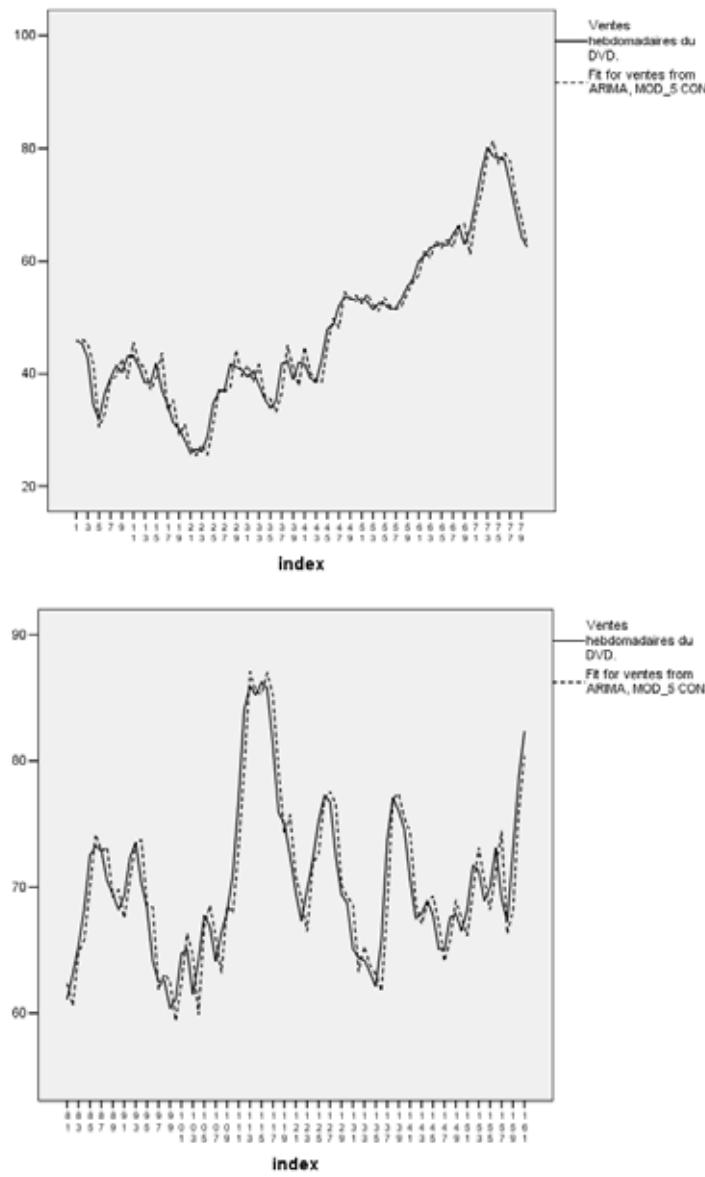


FIG. 13.43 – Visualisation des estimations

La figure 13.44 nous montre le SAC des résidus. Le modèle ne semble pas adéquat ; au lag 6 l'autocorrélation est significative, et la p -value de la statistique Box-Ljung est de $0,006 < 0,05$. Ceci nous indique que de l'information pertinente n'a pas été incluse dans le modèle. On remarque aussi que l'autocorrélation partielle du *lag* 6 est significative (SPAC, figure 13.45).

Autocorrelations						
Series: ERR_1			Box-Ljung Statistic			
Lag	Autocorrelation	Std. Error ^a	Value	df	Sig. ^b	
1	,008	,079	,011	1	,918	
2	-,002	,079	,011	2	,994	
3	-,012	,079	,036	3	,998	
4	,044	,079	,355	4	,986	
5	-,133	,079	3,317	5	,651	
6	-,298	,081	18,285	6	,006	
7	,050	,087	18,706	7	,009	
8	-,075	,087	19,675	8	,012	
9	-,033	,088	19,865	9	,019	
10	-,033	,088	20,056	10	,029	
11	,073	,088	20,993	11	,033	
12	,080	,088	22,101	12	,036	
13	,087	,089	23,425	13	,037	
14	,001	,089	23,425	14	,054	
15	,061	,089	24,089	15	,064	
16	-,132	,090	27,232	16	,039	

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

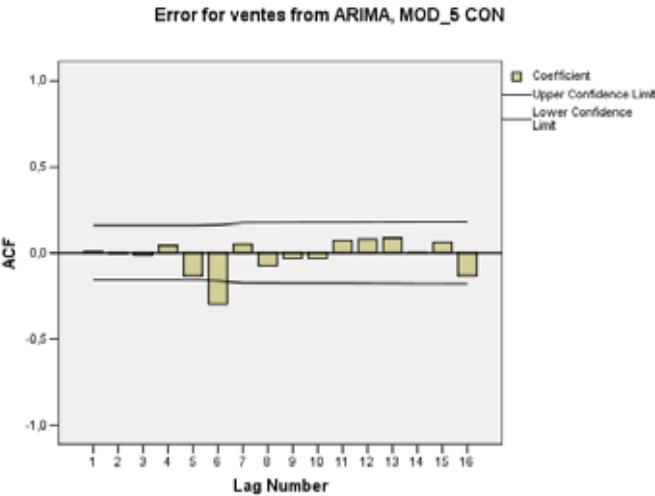


FIG. 13.44 – Le SAC des résidus du modèle ARIMA(0,1,1)

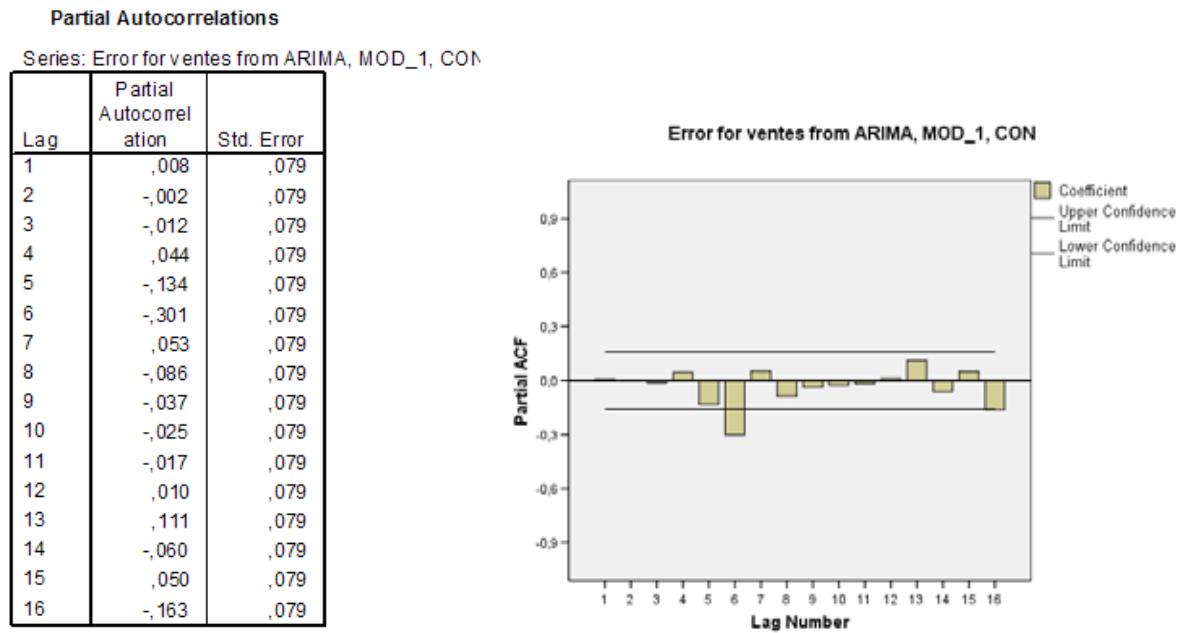


FIG. 13.45 – Le SPAC des résidus du modèle ARIMA(0,1,1)

En fait, étant donné qu'il y a une autocorrélation significative au lag 6, il faudrait considérer le modèle ARIMA(0,1,6) afin d'inclure e_{t-6} dans le modèle. On estime donc le modèle ARIMA(0,1,6) ; la figure 13.46 présente les sorties de ce modèle.

Residual Diagnostics		Parameter Estimates				
Number of Residuals	160					
Number of Parameters	6					
Residual df	153					
Adjusted Residual Sum of Squares	796,344					
Residual Sum of Squares	851,315	Non-Seasonal	MA 1	-,615	,2,104	-,292
Residual Variance	5,063	Lags	MA 2	,039	,,818	,,048
Model Std. Error	2,250		MA 3	-,010	,,735	-,,014
Log-Likelihood	-355,430		MA 4	,,053	,,715	,,074
Akaike's Information Criterion (AIC)	724,861		MA 5	,,167	,,606	,,275
Schwarz's Bayesian Criterion (BIC)	746,387		MA 6	,,449	,,961	,,468
		Constant		,,234	,,167	1,401
M elard's algorithm was used for estimation.						

FIG. 13.46 – Les sorties du modèle ARIMA(0,1,6)

Un avertissement est donné avec les sorties : le modèle est à la limite du respect de la condition d'invertibilité. Ceci est une hypothèse théorique concernant les racines d'un polynôme associé au modèle de moyenne mobile. Il faut faire attention à l'interprétation du modèle dans cette situation.

Si on compare le AIC et le BIC du modèle ARIMA(0,1,1) avec ceux du modèle ARIMA(0,1,6), on note une amélioration (en faveur du modèle ARIMA(0,1,6)) : le AIC est passé de 742,289 à 724,861, et le BIC de 748,439 à 746,387.

Ce qui est malheureux, c'est que puisque nous sommes aux limites de la condition

d'inversibilité, les écarts-types des paramètres du modèle ne sont pas bien évalués. Ceci a pour effet d'amener la fausse conclusion que les paramètres ne sont pas significatifs (les *p*-values sont toutes plus grandes que 0,05), et les intervalles de confiance pour les prévisions ne seront pas très fiables. C'est l'examen des résidus qui saura nous convaincre que ce modèle est meilleur que le précédent.

La figure 13.47 présente d'abord le SAC des résidus. Aucune autocorrélation n'est significative, et ceci est fortement appuyé par les *p*-values des Box-Ljung qui sont toutes largement supérieures à $\alpha = 0,05$. De même, on voit qu'aucune autocorrélation partielle n'est significative (figure 13.48).

Autocorrelations						
Series: ERR_2			Box-Ljung Statistic			
Lag	Autocorrelation	Std.Error ^a	Value	df	Sig. ^b	
1	,014	,078	,030	1	,862	
2	,003	,078	,032	2	,984	
3	-,027	,078	,152	3	,985	
4	,046	,078	,509	4	,973	
5	,000	,077	,509	5	,992	
6	-,016	,077	,553	6	,997	
7	-,069	,077	1,350	7	,987	
8	,015	,077	1,389	8	,994	
9	-,058	,076	1,959	9	,992	
10	-,032	,076	2,135	10	,995	
11	,050	,076	2,573	11	,995	
12	,074	,076	3,531	12	,990	
13	,058	,075	4,127	13	,990	
14	,027	,075	4,254	14	,994	
15	-,005	,075	4,258	15	,997	
16	-,108	,075	6,347	16	,984	

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

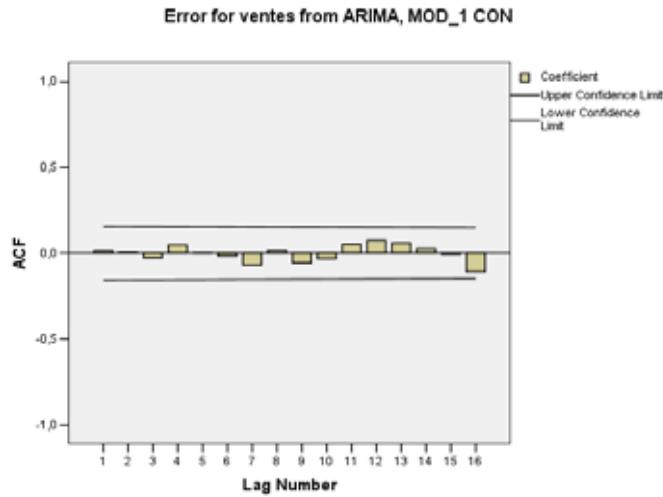


FIG. 13.47 – Le SAC des résidus du modèle ARIMA(0,1,6)

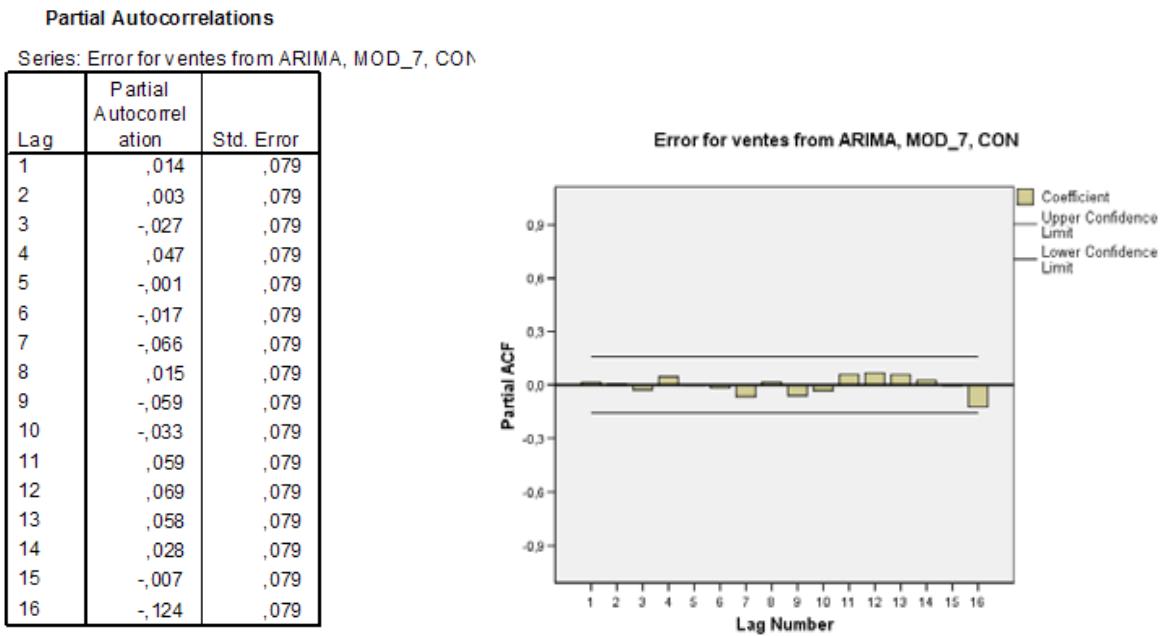


FIG. 13.48 – Le SPAC des résidus du modèle ARIMA(0,1,6)

La première sortie de la figure 13.49 nous permet de résoudre le test suivant :

H_0 : Les résidus se distribuent selon une loi normale au niveau de la population.

H_1 : Les résidus ne se distribuent pas selon une loi normale au niveau de la population.

Puisque les deux p -values sont supérieures à 0,05 (elles ont une valeur de 0,2 et 0,659), on ne rejette pas H_0 et on conclut donc que les résidus suivent une loi normale.

La dernière sortie de la figure 13.49 nous montre la répartition des résidus en fonction du temps, qui semble bien aléatoire.

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ERR_2	,044	160	,200*	,993	160	,659

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

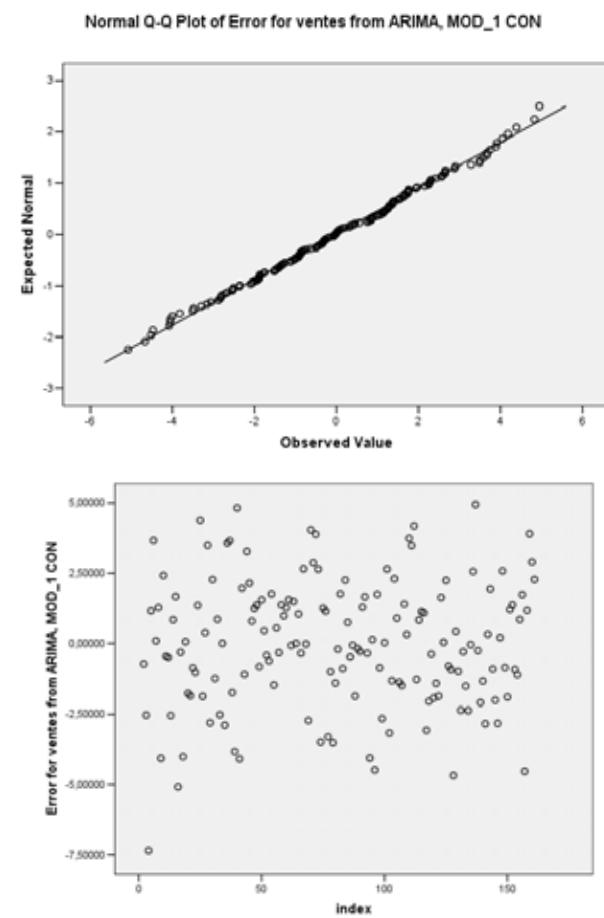


FIG. 13.49 – Normalité des résidus

Finalement, la figure 13.50 nous montre les estimations versus les ventes réelles. Non seulement le modèle semble produire de bonnes estimations, mais finalement les intervalles de prédiction semblent se comporter adéquatement, car les observations sont presque toujours comprises entre les bornes de ces intervalles. Il serait cependant souhaitable de tester ce modèle sur de nouvelles données avant de l'utiliser.

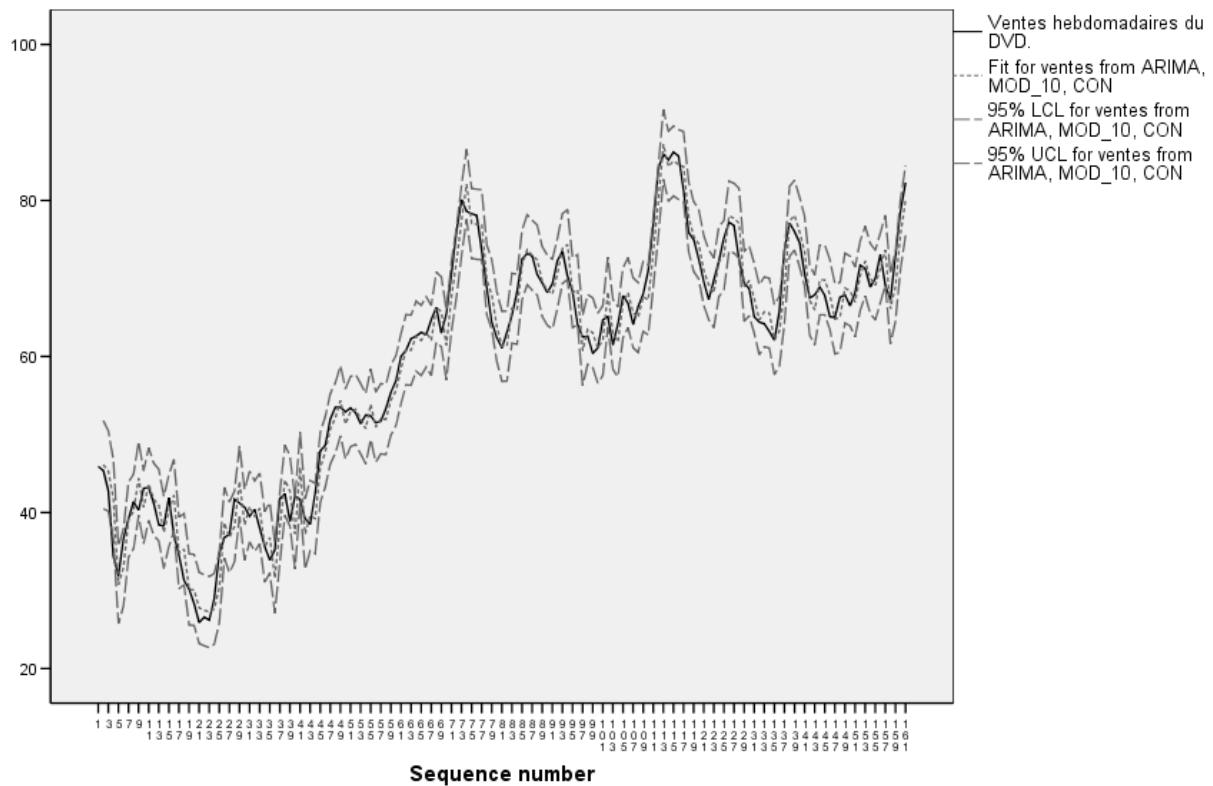


FIG. 13.50 – Visualisation des estimations

On conclut donc que c'est le modèle ARIMA(0,1,6) qui donne de meilleurs résultats, mais qu'il serait plus prudent de le tester sur de nouvelles données.

13.7 ARIMA sans saisons avec E-Views

Voici quelques indications pour l'utilisation d'E-Views pour les modèles ARIMA sans effets de saison. Reprenons l'exemple où on mesurait la qualité d'impression (voir le fichier `imprimante.wf1`). Tout d'abord, pour ne fonctionner qu'avec les 150 premières données, il faut taper dans le haut de la fenêtre `smp1 1 150`. Ensuite, après avoir ouvert la série `mesure`, on peut obtenir les fonctions SAC et SPAC en allant dans `View → Correlogram...` puis en laissant l'option `Level` (on n'a pas besoin de différentier cette série). On obtient alors la figure 13.51 qui nous montre à la fois la fonction SAC et la fonction SPAC de la série.

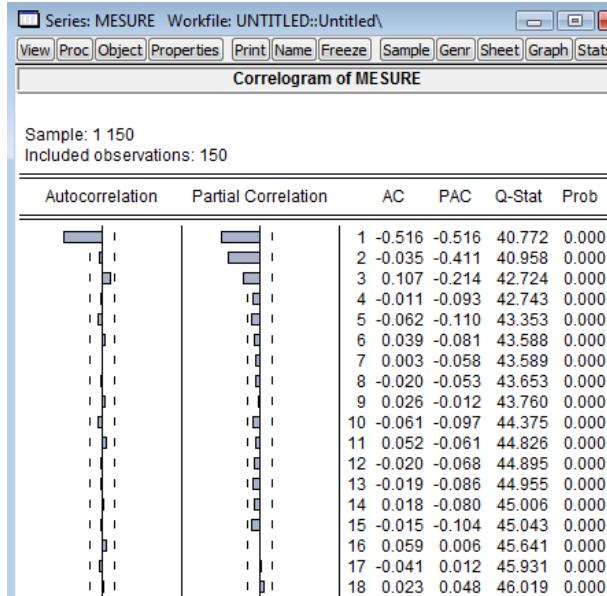


FIG. 13.51 – SAC et SPAC de la série `mesure`

Tel qu'on l'a vu précédemment, le modèle approprié pour cette série semble être un ARIMA(0,0,1). Pour estimer ce modèle, l'équation est

`mesure c ma(1).`

On obtient alors la sortie 13.52 qui nous donne les estimations de ce modèle. L'équation du modèle est

$$\hat{y}_t = 2200,297 - 0,853098e_{t-1}.$$

Il est important de noter que dans E-Views le signe devant le terme MA est celui qui apparaît dans l'équation, contrairement à ce qu'on avait dans SPSS.

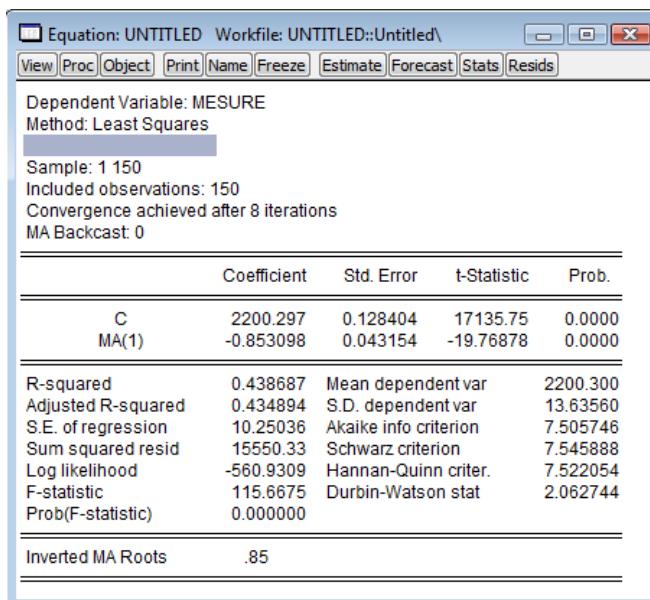


FIG. 13.52 – Modèle ARIMA(0,0,1)

De façon plus générale, les termes autorégressifs d'ordre k s'écrivent `ar(k)`, et si la série d'origine doit être différenciée l fois, il suffit d'écrire `d(y,1)` pour la variable dépendante (si $l = 1$ on peut l'omettre). Par exemple, pour une série y pour laquelle on voudrait un modèle ARIMA(2,1,1), on écrirait

`d(y) c ar(1) ar(2) ma(1).`

Voici une petite parenthèse pour vous présenter les formules associées aux principales mesures d'adéquation présentées par EVViews. Soit n la taille d'échantillon, et supposons qu'on a développé un modèle avec k variables explicatives.

R-squared

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Il est à noter que le R^2 peut être négatif lorsque les méthodes d'estimation (telle que celle utilisée pour les modèles ARCH) font que la décomposition de la variance, que l'on retrouve dans la table ANOVA d'une régression classique, ne tient plus.

Adjusted R-squared

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - k - 1}$$

C'est la même formule que ce qui a été vu dans le chapitre sur la régression linéaire multiple, c'est-à-dire que cette mesure tient compte du nombre de paramètres à estimer.

Log likelihood

$$l = -\frac{n}{2} \left(1 + \log(2\pi) + \log \left(\frac{\sum_{i=1}^n e_i^2}{n} \right) \right)$$

Mesure d'adéquation du modèle que l'on cherche à maximiser.

Akaike info criterion (AIC)

$$AIC = \frac{-2l}{n} + \frac{2(k+1)}{n}$$

Mesure d'adéquation du modèle que l'on cherche à minimiser.

Schwarz criterion (SC ou BIC)

$$SC = \frac{-2l}{n} + \frac{(k+1)\log n}{n}$$

Mesure d'adéquation du modèle que l'on cherche à minimiser.

Revenons à l'exemple. Pour voir le SAC et le SPAC des résidus du modèle il faut aller dans `View → Residual Tests → Correlogram - Q-Statistics`. On obtient alors la première sortie de la figure 13.53. On peut aussi générer le test de normalité des résidus tel qu'on l'a déjà vu, et on obtient la deuxième sortie de la figure 13.53. On voit que l'adéquation du modèle est bonne puisqu'il n'y a pas d'autocorrelations significatives et qu'on ne rejette pas la normalité ; on arrive aux mêmes conclusions qu'avec SPSS (fiou !). Et la figure 13.54 présente les résidus standardisés, qui semblent effectivement aléatoires.

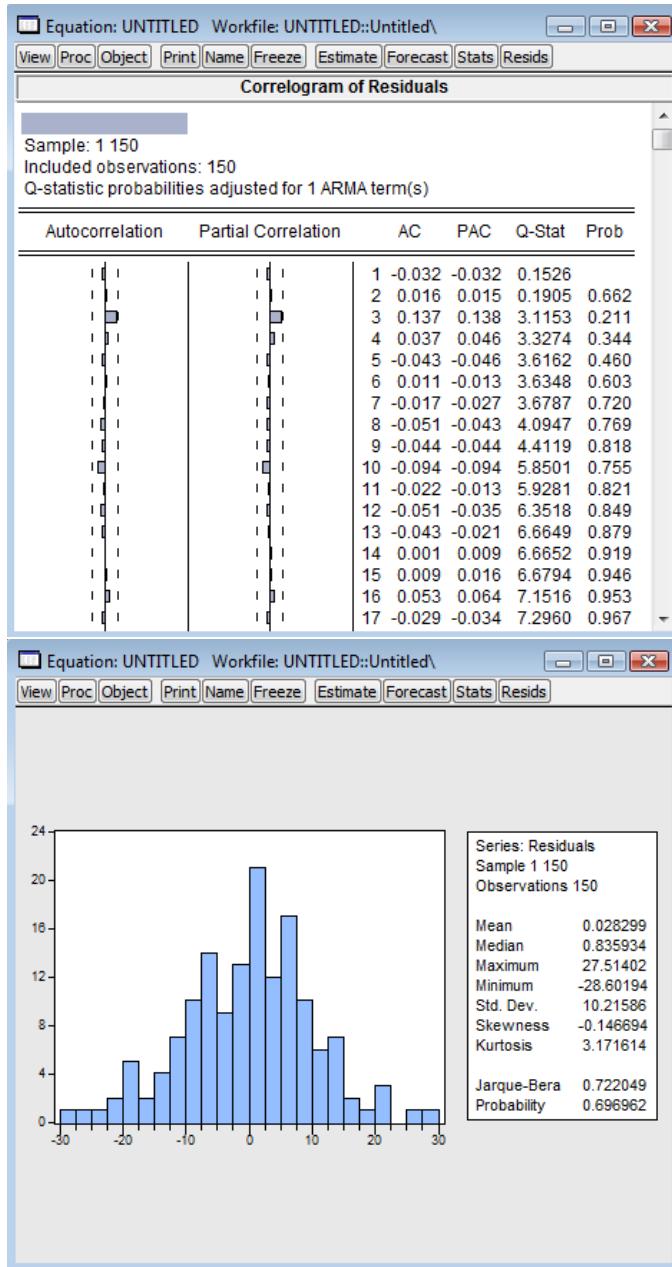


FIG. 13.53 – SAC et SPAC des résidus du modèle et test de normalité des résidus

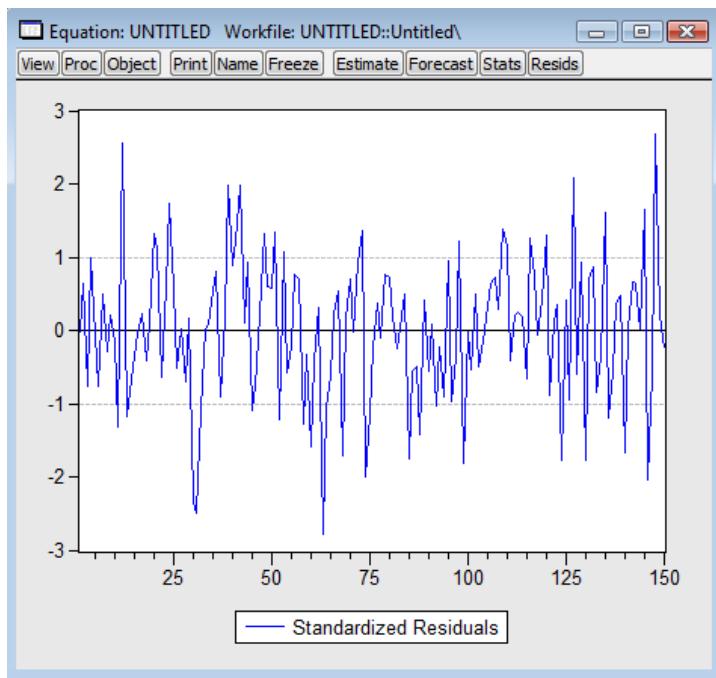


FIG. 13.54 – Résidus standardisés

Pour obtenir les prévisions, on va dans **Forecast**, puis on génère des prévisions statiques pour les 150 premières données (on utilise les véritables valeurs de la série observée), puis pour les 30 dernières données on fait des prévisions dynamiques, c'est-à-dire qu'on fait comme si on ne connaissait pas ces 30 dernières données. On peut voir les menus et les graphes des prévisions dans les figures 13.55 et 13.56.

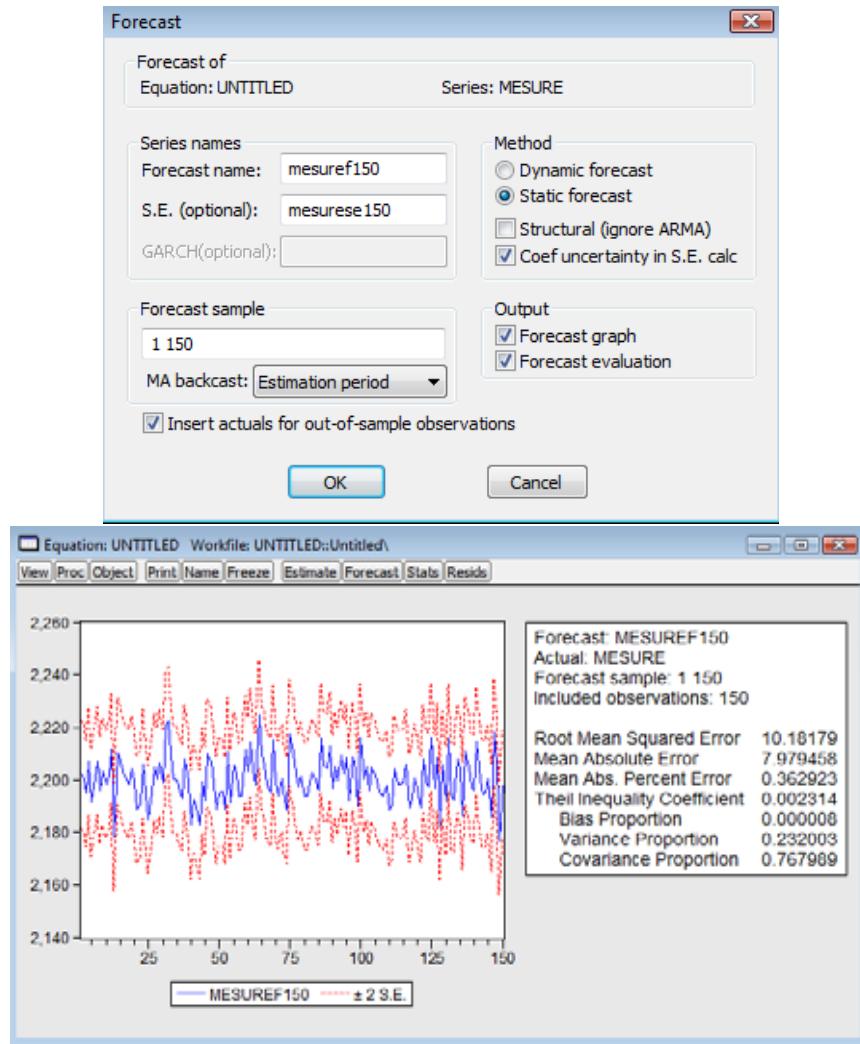


FIG. 13.55 – Prévisions statiques (sur les données connues)

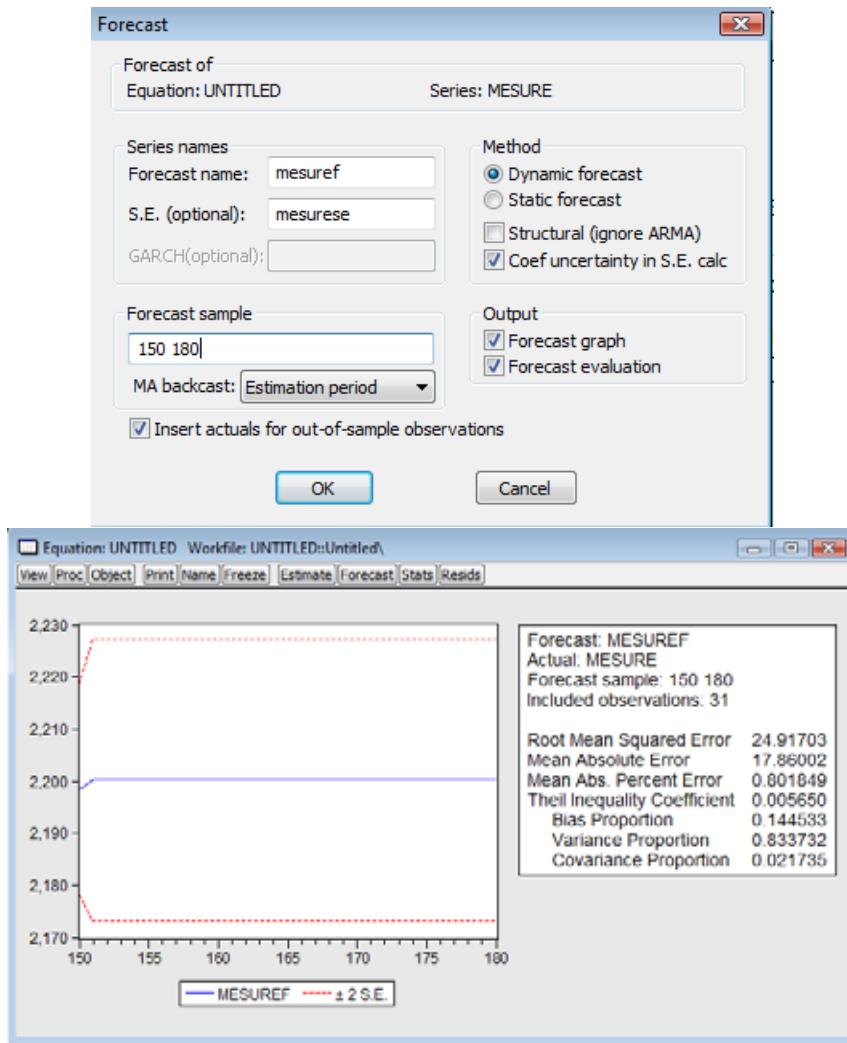


FIG. 13.56 – Prévisions dynamiques

Avec les séries sauvegardées (les prévisions et les erreurs types) on peut créer les intervalles de prédiction (haut de la figure 13.57). Ensuite en ouvrant simultanément toutes ces séries on peut visualiser le tout. Dans la figure 13.58 j'ai simplement restreint l'échantillon pour qu'on voit mieux ce qui se passe pour les prévisions dynamiques.

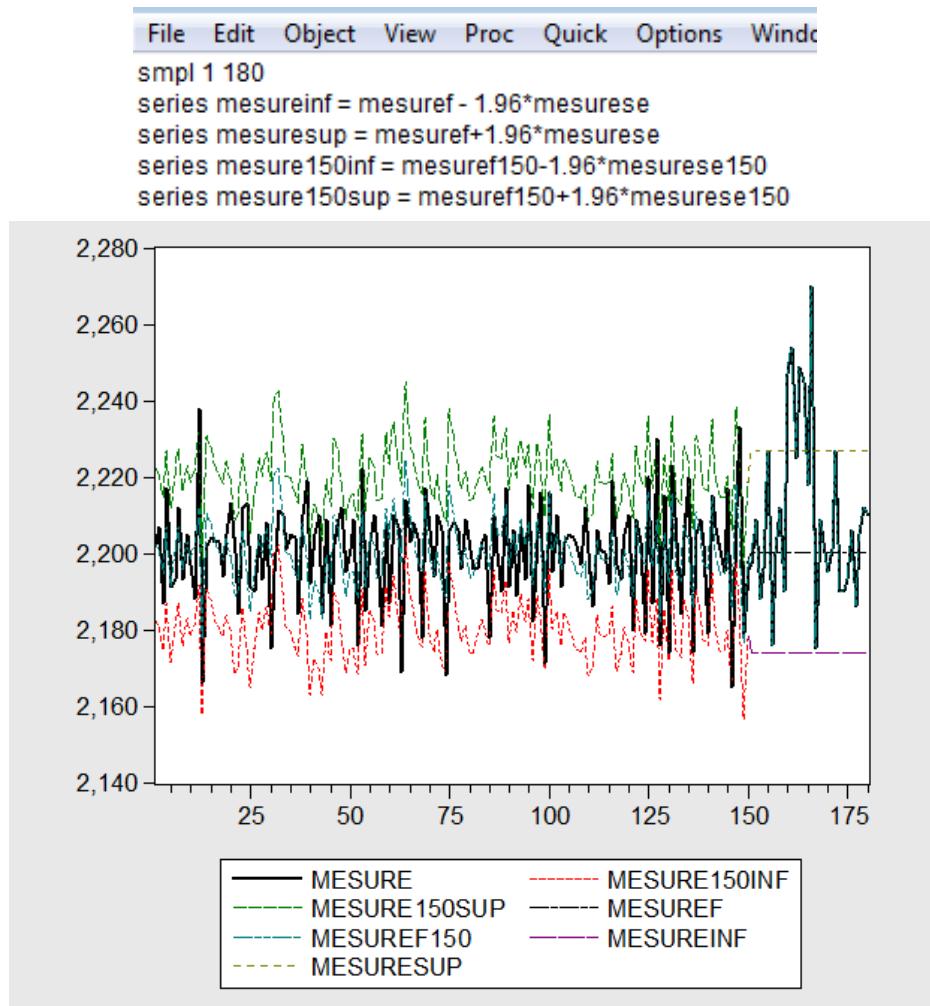


FIG. 13.57 – Visualisation des observations versus les prévisions

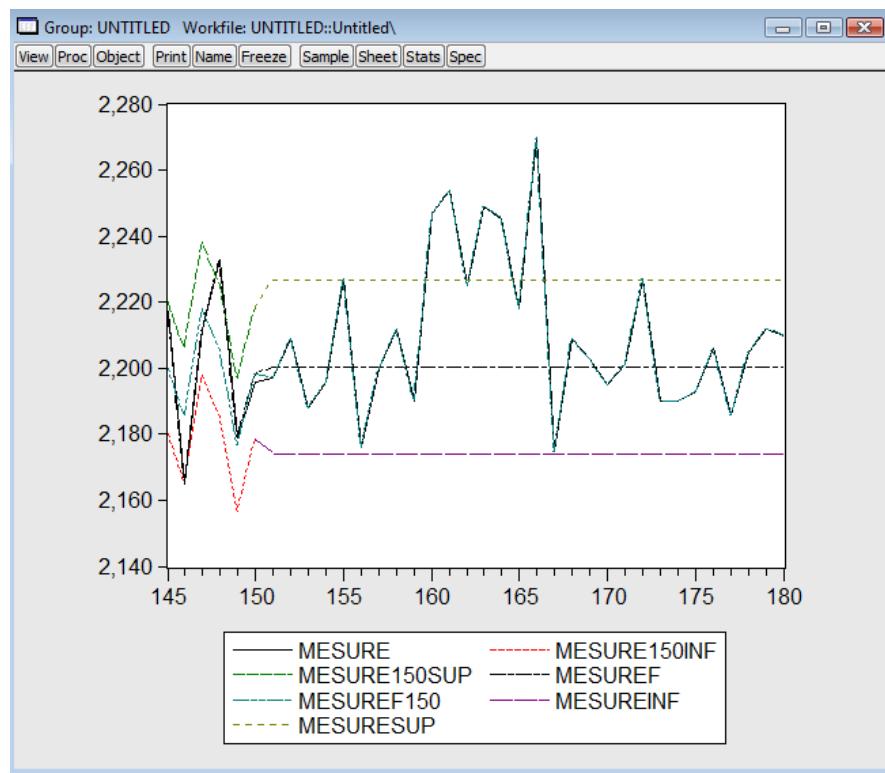


FIG. 13.58 – Les prévisions pour les périodes 145 à 180

13.8 Le modèle ARIMA et l'effet des saisons

Les modèles ARIMA permettent aussi de modéliser les effets saisonniers. Un tel modèle s'écrit $\text{ARIMA}(p, d, q)(P, D, Q)_L$ où :

- p est le paramètre associé à l'ordre du processus d'autorégression régulier à inclure dans le modèle.
- d est le paramètre qui se réfère à l'ordre de la différence régulière à faire pour obtenir une série stationnaire.
- q est le paramètre qui représente l'ordre du processus de moyenne mobile régulier à inclure dans le modèle.
- P est le paramètre associé à l'ordre du processus d'autorégression lié aux effets de saisons à inclure dans le modèle.
- sD est le paramètre qui se réfère à l'ordre de la différence saisonnière à faire pour obtenir une série stationnaire au sens saisonnier du terme.
- sQ est le paramètre qui représente l'ordre du processus de moyenne mobile lié aux effets de saisons à inclure dans le modèle.
- L représente la périodicité de l'effet saisonnier. En général, bien que d'autres périodicités soient possibles, on a $L = 12$ pour des données mensuelles, $L = 4$ pour des données trimestrielles, $L = 7$ pour des données journalières, etc.

L'identification d'un modèle ARIMA pour une série qui comporte un effet saisonnier est quelque peu plus complexe que pour un modèle sans effet de saison. En effet, les fonctions SAC et SPAC contiendront à la fois les effets $\text{AR}(p)$ et $\text{MA}(q)$ de période en période et les effets $\text{AR}(P)$ et $\text{MA}(Q)$ de « saison » en « saison ».

La partie du modèle pour les effets saisonniers s'identifie de la même façon que le modèle régulier, mais au lieu de considérer tous les *lags*, on ne considère que ceux correspondant à la périodicité de l'effet de saison. Par exemple, en présence de données trimestrielles, il faudrait examiner les autocorrélations et autocorrélations partielles aux

lags 4, 8, 12, 16, 20, etc... afin de repérer un des comportements théoriques décrits dans la section 13.5. Ceci exige un plus grand nombre de données que lorsqu'il n'y a pas d'effet saisonnier. Par exemple, si on ne dispose que de 36 données et que les données sont mensuelles, cela ne laisse que 3 autocorrélations et autocorrélations partielles à examiner pour l'effet de saison. Bien que des auteurs estiment que le nombre minimum de données nécessaires est de $3L$, il est préférable d'utiliser $7L$ ou $8L$ données pour bien identifier la partie saisonnière du modèle.

De plus, pour obtenir la stationnarité, il est probable que l'analyste doive différentier en terme de saisons, indépendamment de la différentiation régulière. Plus précisément, pour des données mensuelles, on soustrait le mois courant au même mois de l'année précédente. Contrairement à la différentiation standard où la base de données est réduite d'une seule observation, la différentiation saisonnière retranche L données.

Nous proposons ici de suivre la méthodologie de Box-Jenkins d'abord pour l'effet saisonnier, puis une fois cette partie du modèle identifiée, de reprendre la méthodologie sur les erreurs du modèle partiellement identifié pour finaliser l'identification avec la partie régulière. C'est ce que nous allons illustrer dans l'exemple qui suit.

Exemple 13.8.1 Les données de cet exemple portent sur les nouvelles connections et sur les arrêts de connections de l'entreprise publique de téléphone du Wisconsin (Wisconsin Telephone Company). Ces données ont été analysées par Thompson et Tiao en 1971. La base de données se nomme `telephone.sav`.

Les données sont constituées de 218 observations à partir de janvier 1951 à février 1969. L'objectif de l'étude était de prédire la croissance du ratio des connections sur les arrêts de connections. L'exemple est intéressant puisqu'il propose une solution pour contrôler l'hétéroscédasticité pour finalement obtenir la stationnarité.

Il faut d'abord définir les dates, et calculer le ratio connect/deconnect dans Compute...

La figure 13.59 présente un aperçu de la base de données suite à ces deux opérations.

	index	connect	disconnect	YEAR	MONTH	DATE	ratio
1	1	12093	8291	1951		1 JAN 1951	1,46
2	2	10564	7874	1951		2 FEB 1951	1,34
3	3	9860	8026	1951		3 MAR 1951	1,23
4	4	10280	8761	1951		4 APR 1951	1,17
5	5	11201	9379	1951		5 MAY 1951	1,19
6	6	12016	11018	1951		6 JUN 1951	1,09
7	7	11476	9838	1951		7 JUL 1951	1,17
8	8	12286	10976	1951		8 AUG 1951	1,12

FIG. 13.59 – Aperçu de la base de données

On décide ici de se garder les données de janvier 1967 à février 1969 pour vérifier la performance du modèle. Ce qui suit a donc été obtenu suite à une sélection filtrant ces dernières données.

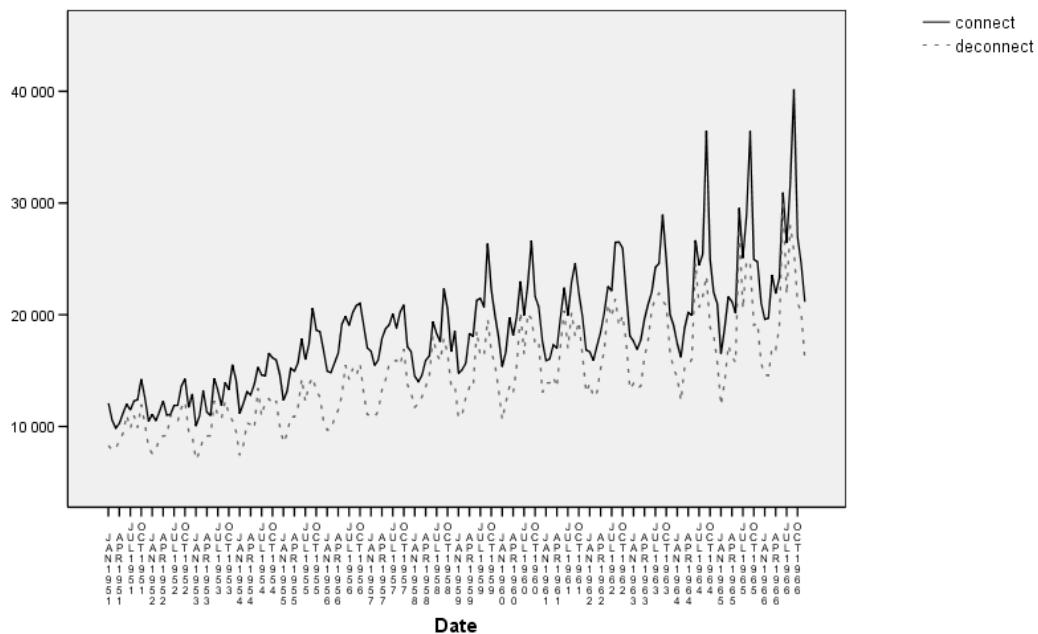


FIG. 13.60 – Séries connect et disconnect

La figure 13.60 présente l'évolution des séries `connect` et `deconnect`. On remarque que ces séries présentent des effets saisonniers marqués, une tendance à la hausse ainsi

qu'une croissance dans la variabilité (elles sont hétéroscédastiques).

Cependant, c'est la série **ratio** qui nous intéresse. On voit cette série dans la figure 13.61. Le graphe de la figure 13.60 illustrait que le nombre de connections est supérieur au nombre d'arrêt, laissant le ratio à modéliser supérieur à 1, ce que la figure 13.61 confirme.

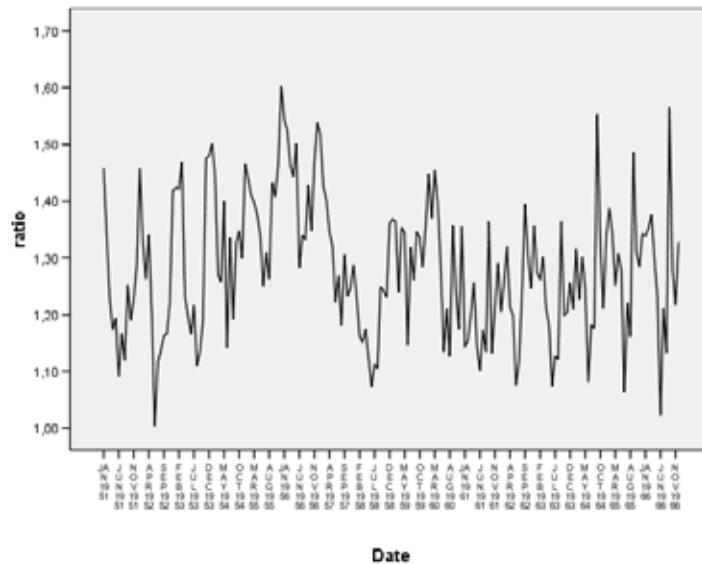


FIG. 13.61 – La série **ratio**

Tout comme les séries originales, la série **ratio** comporte un effet saisonnier. Aussi, elle semble plutôt hétéroscédastique. La différentiation ne corrige pas ce genre de problème.

Pour stabiliser la variance, l'analyste peut utiliser une transformation tel un logarithme ou une racine. Cette technique préliminaire porte le nom de pré-différenciation. SPSS permet à l'analyste d'utiliser la transformation logarithmique en tout temps sans avoir à transformer la série de base. Dans le cadre de cet exemple, nous utilisons la transformation logarithmique pour tenter de résoudre le problème d'hétéroscédasticité.

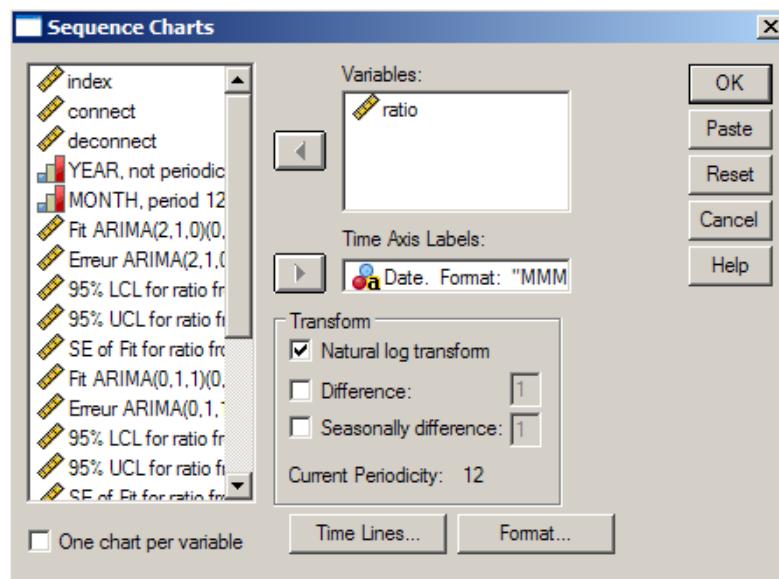


FIG. 13.62 – Pour appliquer un log à une série dans un graphe séquentiel

La figure 13.62 montre l’interface pour faire un graphe séquentiel ; on voit que la transformation logarithmique est sélectionnée. On obtient alors le premier graphe de la figure 13.63. On voit que le log a corrigé légèrement le problème d’hétérosécédasticité (on va s’en contenter... d’ici à ce qu’on voit les modèles GARCH). Le SAC de cette série montre qu’il semble il y avoir plusieurs autocorrélations significatives aux *lags* correspondant à l’effet saisonnier (12, 24, 36). Puisque ces autocorrélations décroissent lentement, il faudra probablement appliquer une différentiation saisonnière. Pour mieux comprendre, allons voir la figure 13.64.

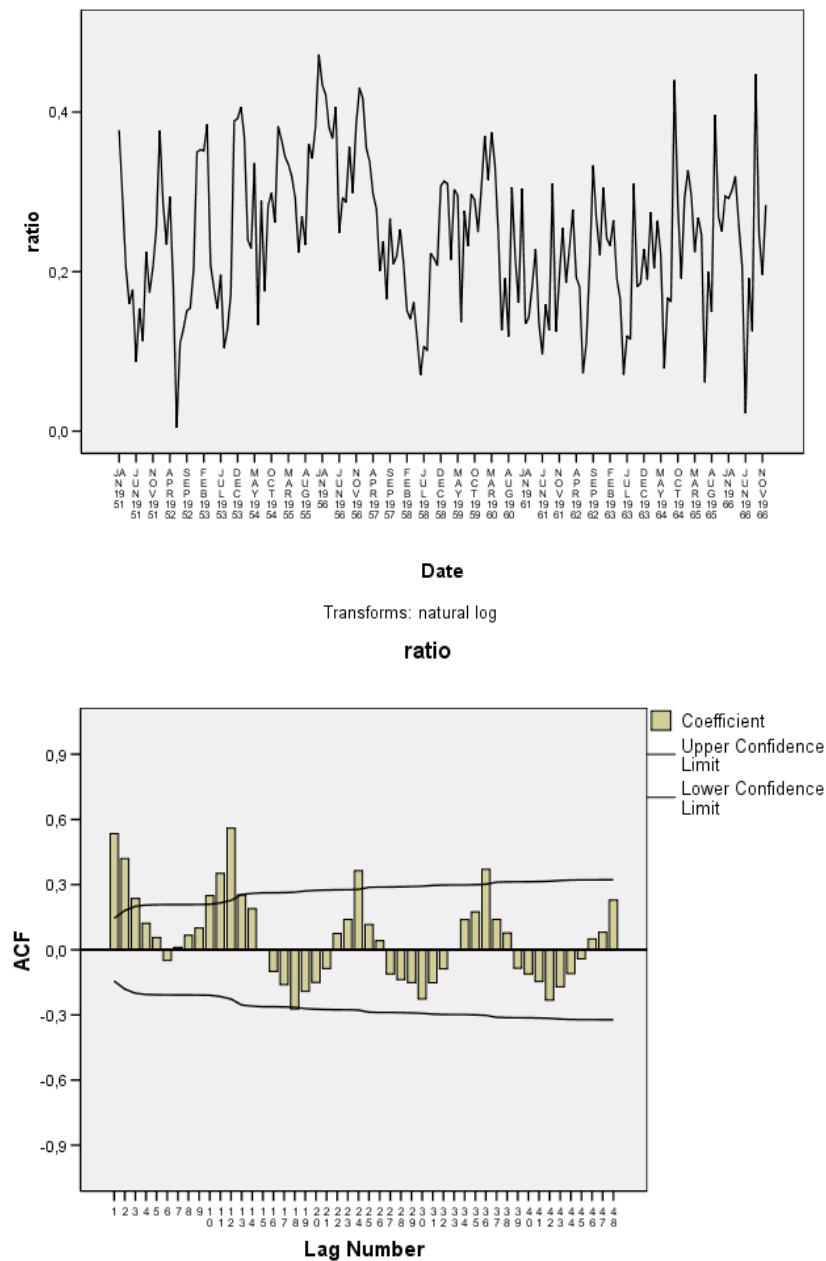


FIG. 13.63 – La série transformée et son SAC

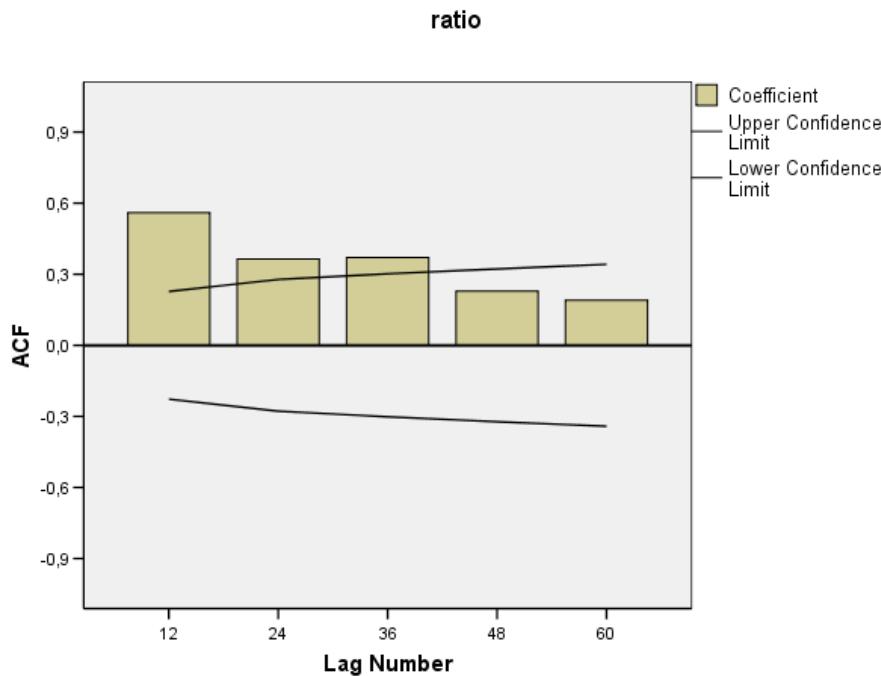


FIG. 13.64 – Le SAC saisonnier de la série transformée.

Dans le bouton **Options...** du menu pour obtenir le SAC, on a coché l'option **Display autocorrelations at periodic lags** et demandé **60 lags**, ce qui donne le SAC saisonnier de la figure 13.64. On n'a plus maintenant que les autocorrélations des *lags* correspondant à l'effet saisonnier. Puisque la décroissance est assez lente, on doit faire une différentiation saisonnière, tout comme on fait une différentiation régulière quand la décroissance des autocorrélations aux premiers *lags* est lente. On refait donc les commandes pour obtenir le SAC et le SPAC, mais en cochant **Seasonally difference** :

1. Ici la commande a été effectuée deux fois, la première fois sans l'option **Display autocorrelations at periodic lags** (figure 13.65), la deuxième fois avec cette option (figure 13.66).

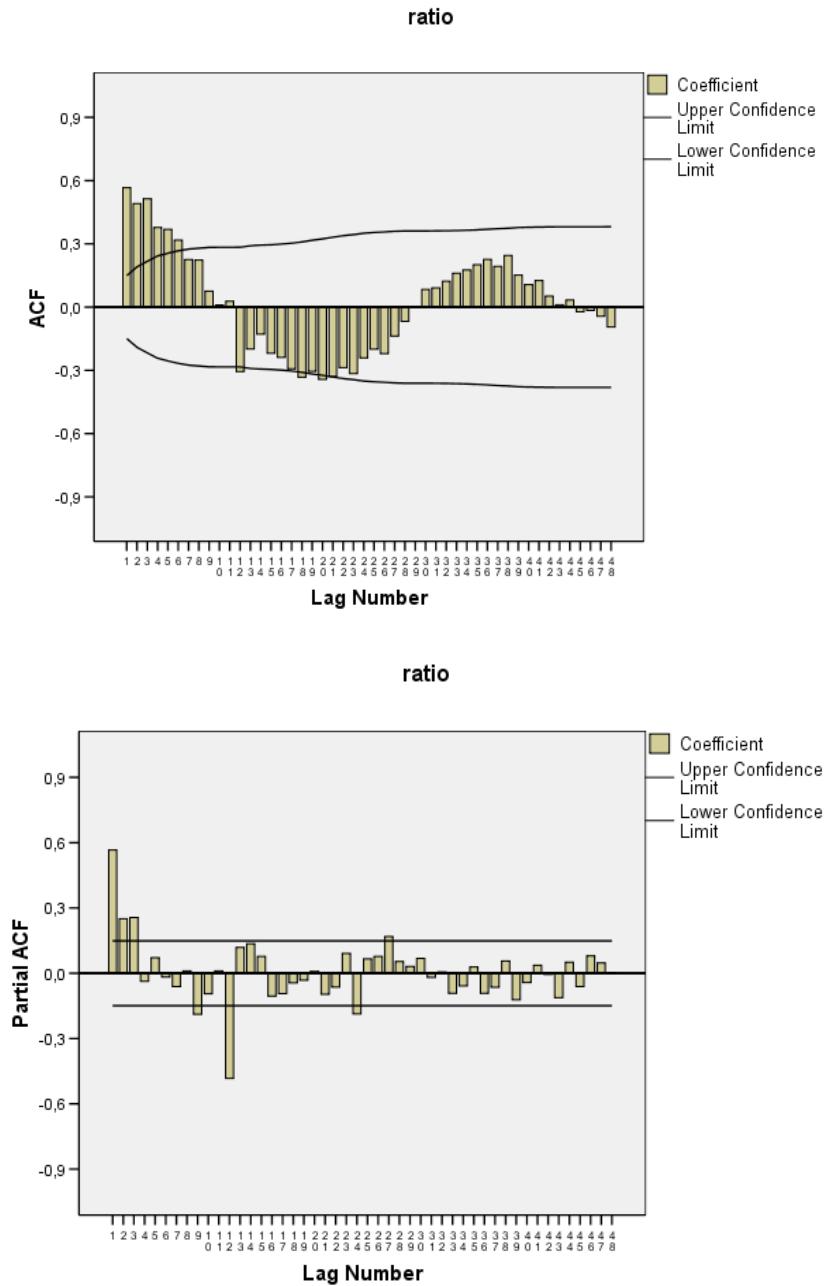


FIG. 13.65 – Le SAC et le SPAC de la série avec un log et une différentiation saisonnière

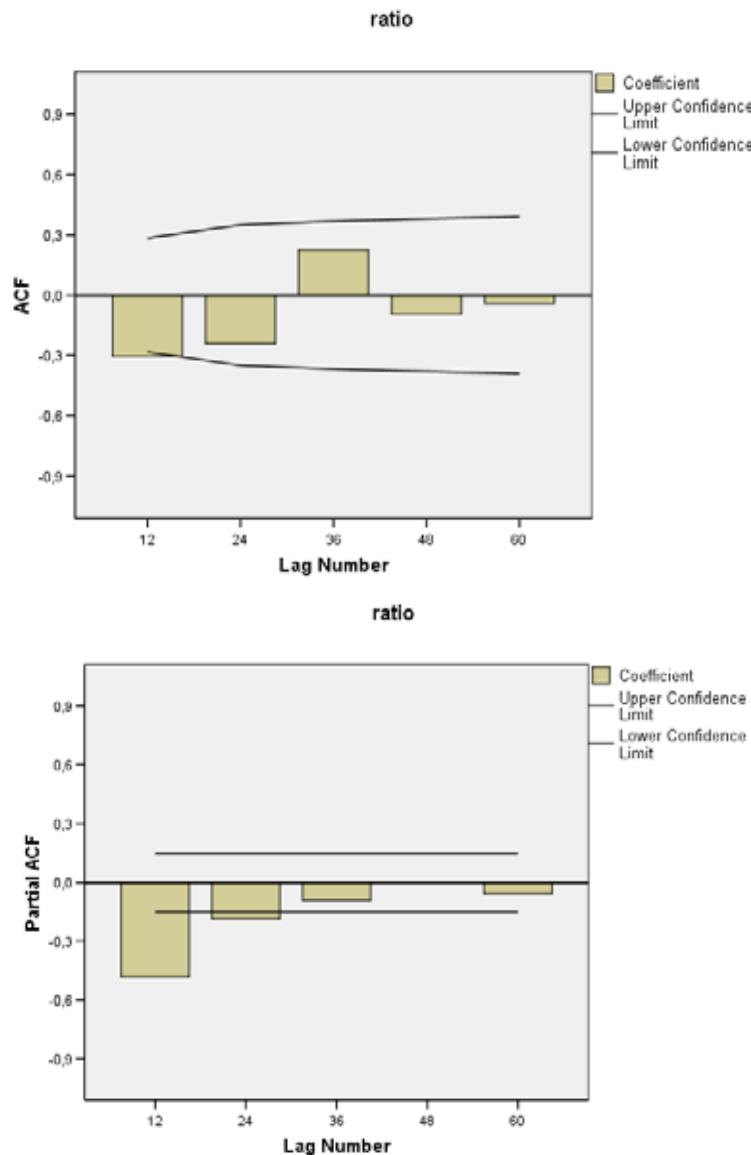


FIG. 13.66 – Le SAC et le SPAC saisonniers de la série avec un log et une différentiation saisonnière

La figure 13.65 ne nous est pas vraiment utile, sauf pour constater que la différentiation saisonnière a eu l'effet d'enlever les autocorrélations significatives des *lags* saisonniers. Il nous faut maintenant tenter d'identifier quel est le modèle requis pour modéliser l'effet saisonnier, et ce à l'aide du SAC et du SPAC saisonniers (figure 13.66). On se réfère alors aux comportements théoriques qui ont été présentés à la section 13.5. Ce n'est pas évident ici, mais on décide que la décroissance exponentielle est plus plausible dans le SPAC, et il y a une seule autocorrélation significative dans le SAC, donc le modèle saisonnier que l'on va tenter est un ARIMA(0,0,0)(0,1,1)₁₂, sans oublier la transformation logarithmique (voir la figure 13.67).

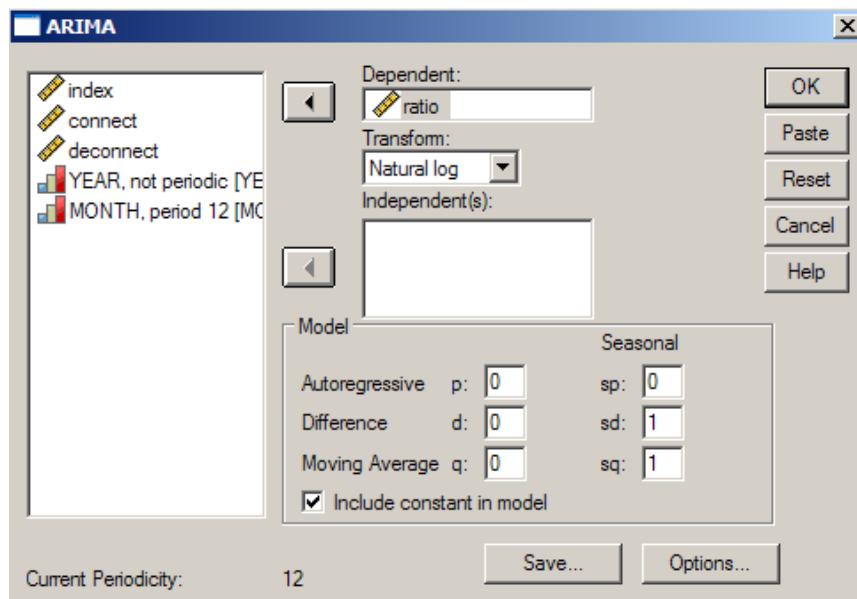


FIG. 13.67 – Pour estimer le modèle ARIMA(0,0,0)(0,1,1)₁₂.

Une fois ce modèle estimé, il nous reste à en examiner les résidus pour identifier le resstant du modèle. La figure 13.68 présente le SAC des résidus du modèle ARIMA(0,0,0)(0,1,1)₁₂. On voit qu'il y a beaucoup d'autocorrélations significatives, il faut donc appliquer une différentiation régulière avant de pouvoir identifier le modèle.

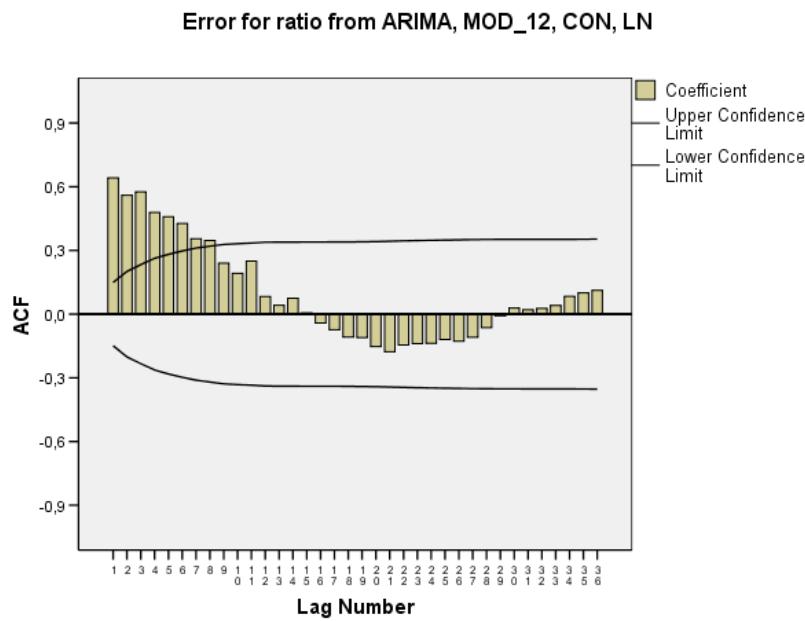


FIG. 13.68 – Le SAC des résidus du modèle ARIMA $(0,0,0)(0,1,1)_{12}$

La figure 13.69 présente donc le SAC et le SPAC des résidus différentiés. L'identification n'est pas facile ici ; on peut hésiter entre les modèles ARIMA $(2,1,0)(0,1,1)_{12}$ et ARIMA $(0,1,1)(0,1,1)_{12}$ selon que l'on considère que la décroissance exponentielle est dans le SAC ou le SPAC. On décide donc d'essayer ces deux modèles et de les comparer.

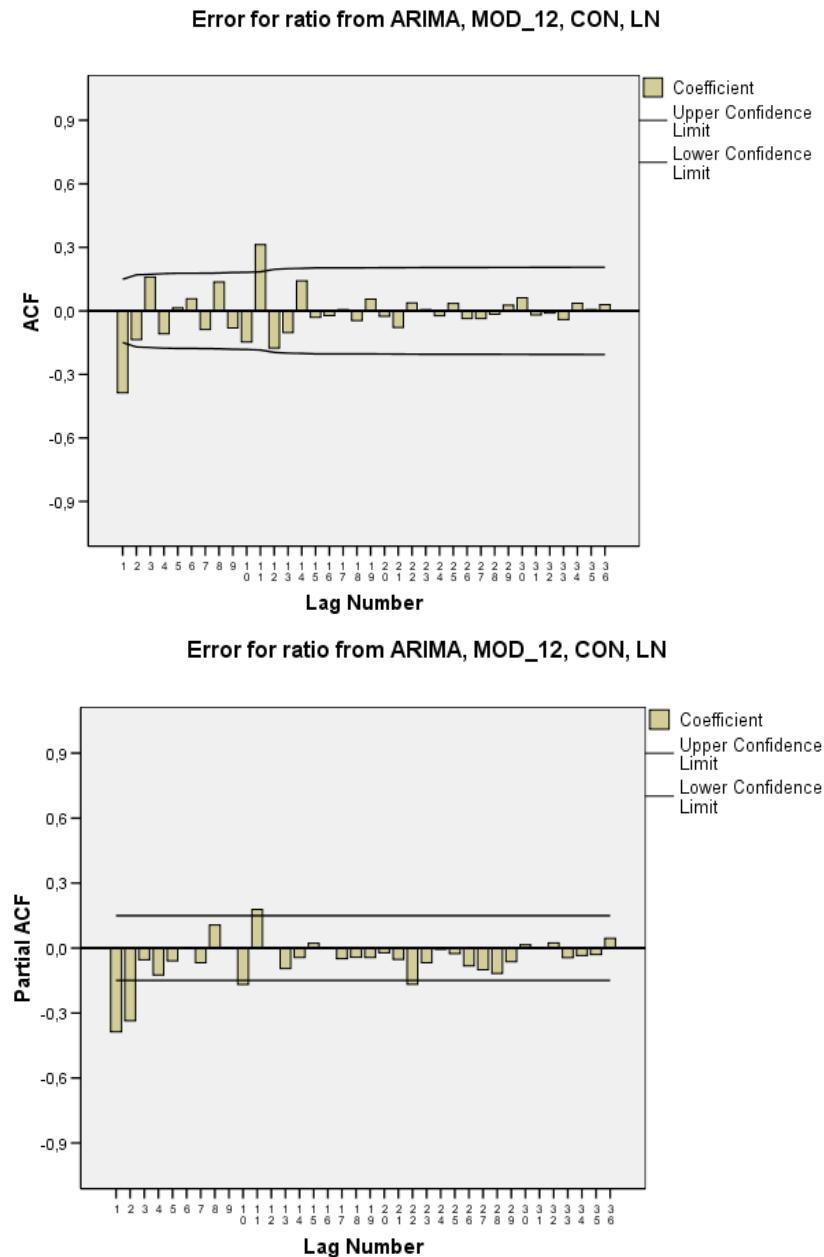


FIG. 13.69 – Le SAC et le SPAC des résidus différentiés

Residual Diagnostics	
Number of Residuals	179
Number of Parameters	3
Residual df	175
Adjusted Residual Sum of Squares	,546
Residual Sum of Squares	,548
Residual Variance	,003
Model Std. Error	,055
Log-Likelihood	264,417
Akaike's Information Criterion (AIC)	-520,835
Schwarz's Bayesian Criterion (BIC)	-508,085

Parameter Estimates				
		Estimates	Std Error	t
Non-Seasonal Lags	AR1	-,509	,072	-7,104
	AR2	-,316	,070	-4,495
Seasonal Lags	Seasonal MA1	,630	,070	8,974
Constant		-7,5E-005	,001	-,079

Melard's algorithm was used for estimation.

FIG. 13.70 – Les estimations du modèle ARIMA(2,1,0)(0,1,1)₁₂

On considère d'abord le modèle ARIMA(2,1,0)(0,1,1)₁₂ dont les sorties sont dans la figure 13.70. Les trois paramètres sont jugés significatifs puisque leurs *p*-values sont nulles, ce qui est bon signe. Les mesures de la première sortie nous seront utiles pour comparer ce modèle avec le ARIMA(0,1,1)(0,1,1)₁₂. Le AIC est de -520,835 et le BIC est de -508,085 ; il faudra voir si l'autre modèle a un AIC et un BIC encore plus petits (donc encore plus négatifs). Et comme je l'ai dit en classe, il faut un Log-Likelihood le plus grand possible (et non le plus près de 0 comme c'était écrit précédemment), et ici il a une valeur de 264,417.

La figure 13.71 présente le SAC des résidus. Il y a une autocorrélation significative au *lag* 11, mais ça ne semble pas important puisque toutes les *p*-values des Box-Ljung sont plus grande que 0,05. Celle du *lag* 11 est de 0,194, et la plus petite est de 0,149 (*lag* 13), ce n'est pas alarmant. La figure 13.72 présente le SPAC des résidus. Lui aussi a une autocorrélation (partielle) significative au *lag* 11, mais encore une fois ce n'est rien d'alarmant puisque les Box-Ljung soutiennent que le modèle est adéquat.

Autocorrelations						
Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic			Sig. ^b
			Value	df	Sig.	
1	-.019	.075	.065	1	,799	
2	-.077	.075	1,144	2	,564	
3	-.093	.075	2,731	3	,435	
4	-.078	.076	3,858	4	,426	
5	,011	.076	3,879	5	,567	
6	,050	.076	4,355	6	,629	
7	-.015	.076	4,398	7	,733	
8	,050	.077	4,875	8	,771	
9	-.069	.077	5,790	9	,761	
10	-.075	.077	6,882	10	,737	
11	,202	.077	14,745	11	,194	
12	-.097	.080	16,572	12	,166	
13	-.092	.081	18,226	13	,149	
14	,048	.082	18,687	14	,177	
15	,000	.082	18,687	15	,228	
16	-.026	.082	18,823	16	,278	
17	-.036	.082	19,076	17	,324	
18	-.043	.082	19,448	18	,365	
19	,009	.082	19,464	19	,427	
20	-.075	.082	20,598	20	,421	

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

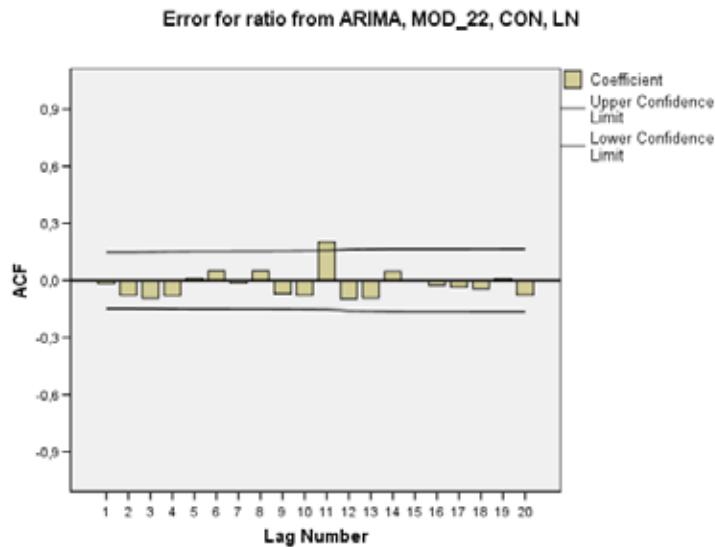
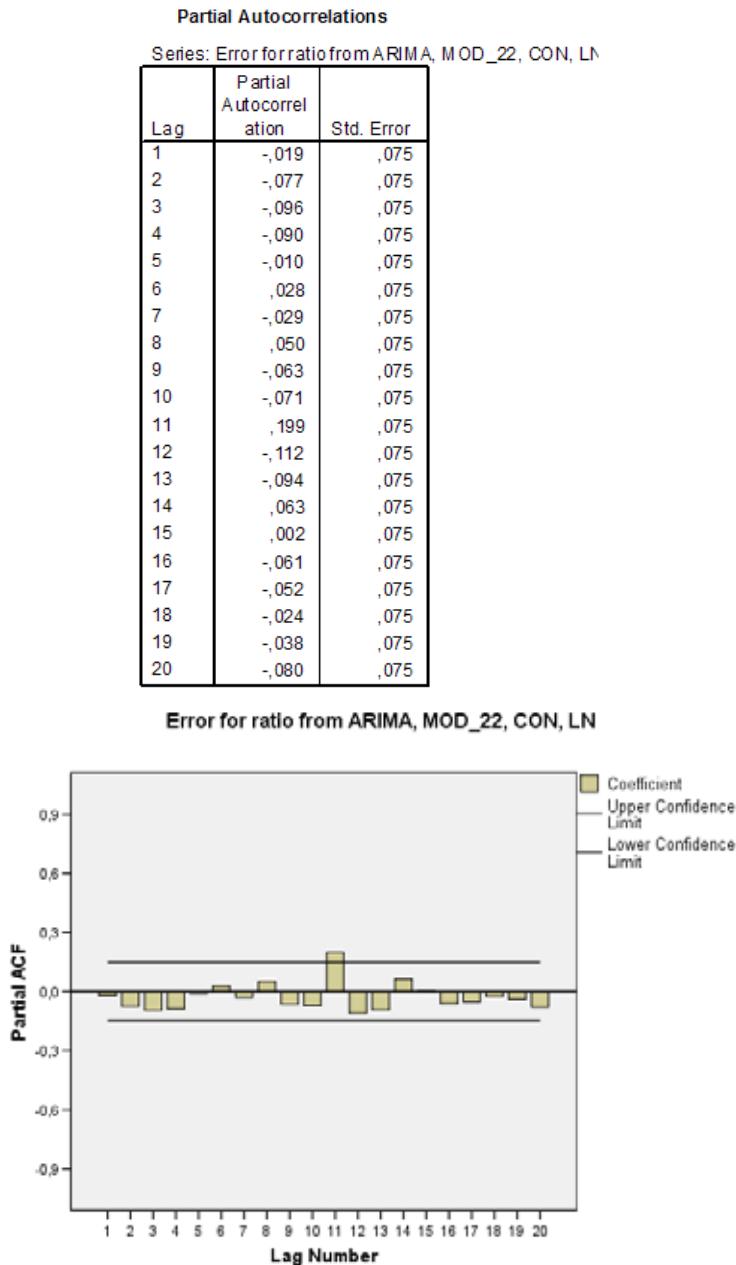


FIG. 13.71 – Le SAC des résidus du modèle ARIMA(2,1,0)(0,1,1)₁₂

FIG. 13.72 – Le SPAC des résidus du modèle ARIMA(2,1,0)(0,1,1)₁₂

Residual Diagnostics	
Number of Residuals	179
Number of Parameters	2
Residual df	176
Adjusted Residual Sum of Squares	,545
Residual Sum of Squares	,552
Residual Variance	,003
Model Std. Error	,055
Log-Likelihood	264,540
Akaike's Information Criterion (AIC)	-523,081
Schwarz's Bayesian Criterion (BIC)	-513,518

Parameter Estimates					
		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	MA 1	,582	,061	9,579	,000
Seasonal Lags	Seasonal MA 1	,651	,067	9,725	,000
Constant		-8,3E-005	,001	-,120	,904

Melard's algorithm was used for estimation.

FIG. 13.73 – Les estimations du modèle ARIMA(0,1,1)(0,1,1)₁₂

Regardons maintenant ce qu'il en est du modèle ARIMA(0,1,1)(0,1,1)₁₂. La deuxième sortie de la figure 13.73 nous montre que les deux paramètres du modèle sont jugés significatifs puisque leurs *p*-values sont nulles. On peut maintenant comparer les mesures de la première sortie avec celles qu'on avait obtenues pour le modèle ARIMA(2,1,0)(0,1,1)₁₂.

En fait toutes ces mesures favorisent (parfois très légèrement) le modèle ARIMA(0,1,1)(0,1,1)₁₂. En effet, c'est ce modèle qui a le plus grand Log-Likelihood (mais à peine : 264,54 > 264,417), et son AIC et son BIC sont plus petits (-523,081 < -520,835, et -513,518 < -508,085). Probablement que les deux modèles sont à peu près équivalents. Poursuivons l'analyse du modèle ARIMA(0,1,1)(0,1,1)₁₂, et il nous restera aussi à vérifier la normalité des résidus des deux modèles.

Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	,022	,075	,088	1	,767
2	-,095	,075	1,738	2	,419
3	,074	,075	2,734	3	,434
4	-,070	,076	3,646	4	,456
5	,003	,076	3,648	5	,601
6	,046	,076	4,048	6	,670
7	-,031	,076	4,225	7	,754
8	,072	,076	5,210	8	,735
9	-,090	,077	6,741	9	,664
10	-,090	,077	8,282	10	,601
11	,216	,078	17,316	11	,099
12	-,086	,081	18,747	12	,095
13	-,119	,082	21,509	13	,063
14	,066	,083	22,361	14	,071
15	-,026	,083	22,496	15	,095
16	-,058	,083	23,173	16	,109
17	-,044	,083	23,562	17	,132
18	-,067	,083	24,456	18	,141
19	-,013	,084	24,488	19	,178
20	-,092	,084	26,230	20	,158

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

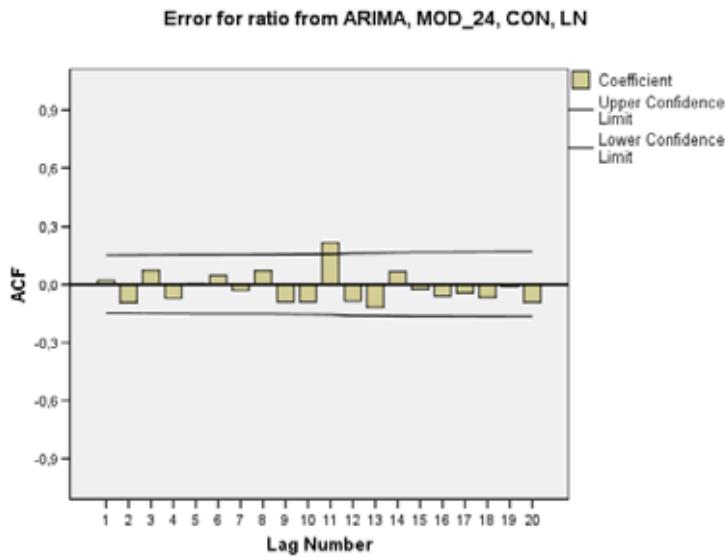
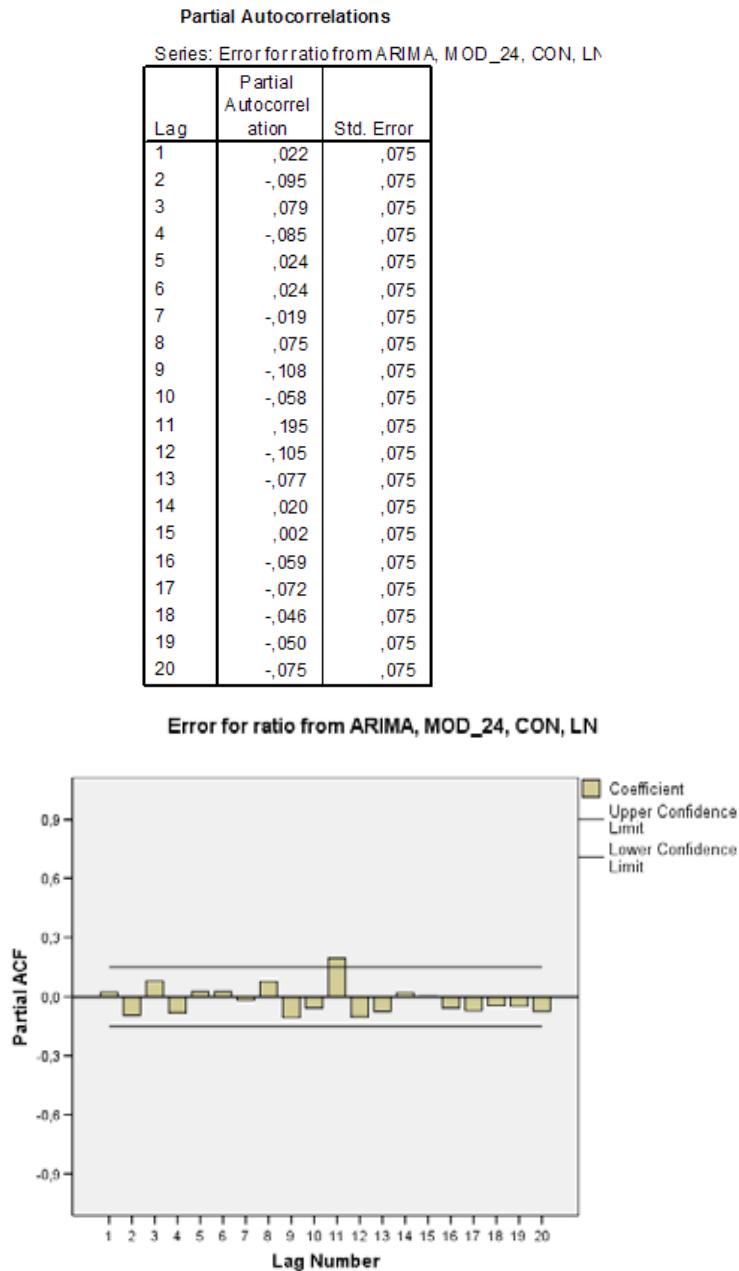


FIG. 13.74 – Le SAC des résidus du modèle ARIMA(0,1,1)(0,1,1)₁₂

FIG. 13.75 – Le SPAC des résidus du modèle ARIMA(0,1,1)(0,1,1)₁₂

L'examen du SAC et du SPAC des résidus du modèle ARIMA(0,1,1)(0,1,1)₁₂ (figures 13.74 et 13.75) nous révèle que le modèle est adéquat puisque toutes les *p*-values des Box-Ljung sont plus grandes que 0,05. Par contre on remarque comme pour l'autre modèle qu'au *lag* 11 les autocorrélations (normale et partielle) sont significatives, et que c'est plus inquiétant que dans le modèle ARIMA(2,1,0)(0,1,1)₁₂ car on remarque que certaines *p*-values à partir de ce *lag* sont relativement petites (la plus petite est de 0,063 au *lag* 13, ce qui est très près du 0,05). Donc ici le modèle ARIMA(2,1,0)(0,1,1)₁₂ est légèrement favorisé.

	Tests of Normality					
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Erreur ARIMA(2,1,0)(0,1,1)	,054	179	,200*	,988	179	,134
Erreur ARIMA(0,1,1)(0,1,1)	,053	179	,200*	,992	179	,460

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correlation

FIG. 13.76 – Test de normalité

La figure 13.76 nous permet de tester la normalité des résidus des deux modèles (il faudrait écrire le test formellement...). Dans les deux cas on ne rejette pas la normalité puisque les *p*-values sont plus grandes que 0,05. Elles sont de 0,2 et 0,134 pour le modèle ARIMA(2,1,0)(0,1,1)₁₂, et de 0,2 et 0,460 pour le modèle ARIMA(0,1,1)(0,1,1)₁₂. Les histogrammes de la figure 13.77 viennent appuyer ceci, de même que les graphes des résidus en fonction du temps (figure 13.79) qui nous montrent des répartitions aléatoires et la présence d'un seul *outlier* dans chacun des graphes, ce qui ne semble pas très inquiétant.

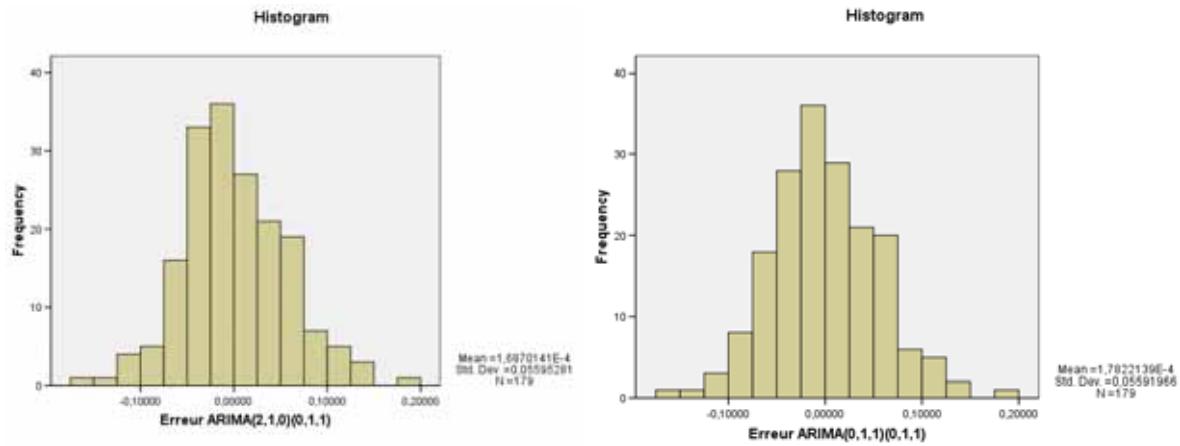


FIG. 13.77 – Les histogrammes des résidus

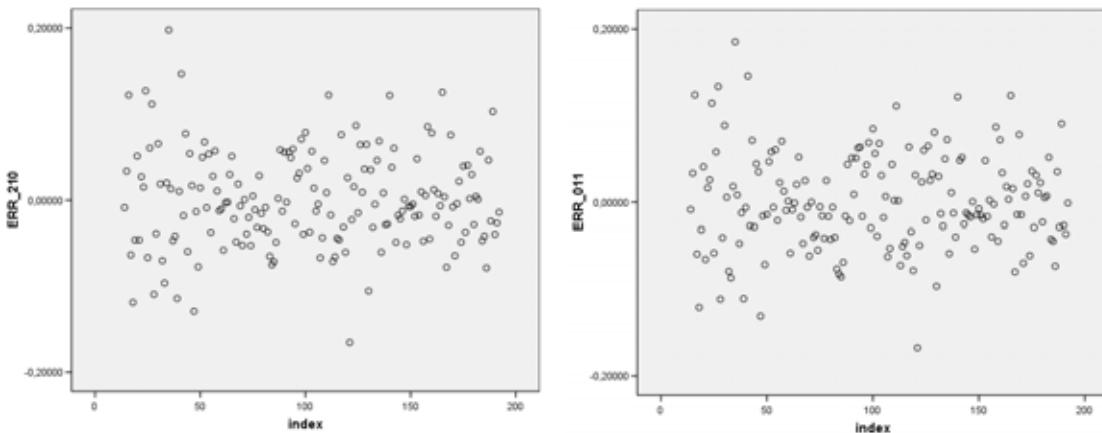


FIG. 13.78 – Répartition des résidus en fonction du temps

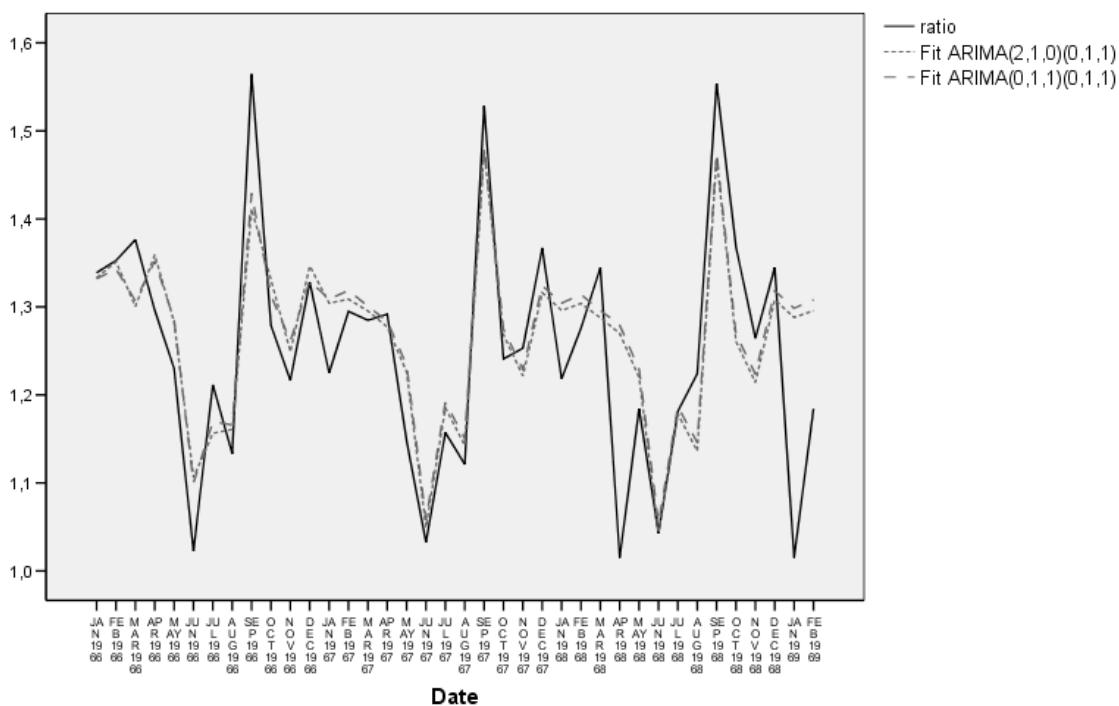


FIG. 13.79 – Estimations des deux modèles de janvier 1966 à février 1969

Finalement, la figure 13.79 présente les estimations pour janvier 1966 à février 1969 (rappelons que la plage d'essai est de janvier 1967 à février 1969). On voit que les deux modèles performent relativement bien. Il resterait à regarder les intervalles de prédiction (eux aussi se comportent de façon semblable).

En conclusion, les deux modèles semblent tout à fait acceptables.

Exemple 13.8.2 Considérons maintenant le nombre total mensuel des passagers (en milliers de passagers) pour les vols aériens internationaux de 1990 à 2000, et tentons de modéliser cette série avec un modèle ARIMA. La base de données se nomme `passagers.sav`.

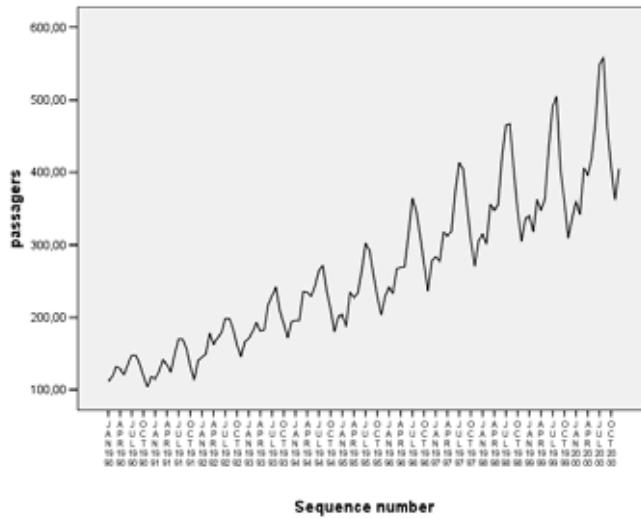


FIG. 13.80 – Série originale

La figure 13.80 nous présente la série originale sur les 11 années. On remarque que cette série n'est pas stationnaire, à la fois du point de vue de la tendance et de la variance constante. Aussi, on peut se douter qu'il y a un effet saisonnier car un patron semble se répéter dans la série.

On commence ici par tenter de remédier au fait que la variance n'est pas constante en appliquant un log naturel à la série. La figure 13.81 présente la série sur laquelle le log a été appliqué ; le résultat est satisfaisant, la variance est beaucoup plus constante.

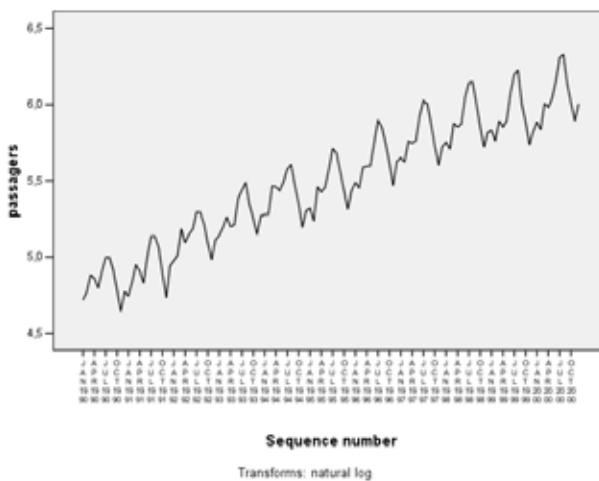


FIG. 13.81 – Série transformée par un log

Jetons maintenant un coup d’œil au SAC de la série sur laquelle le log a été appliqué (figure 13.82). On voit immédiatement qu’il est impératif d’appliquer une différentiation non-saisonnier (de la forme $z_t = y_t - y_{t-1}$) puisque plusieurs autocorrélations sont significatives et qu’il est impossible pour l’instant de constater un effet de saison (même en demandant plus de *lags* on ne peut distinguer l’effet de saison).

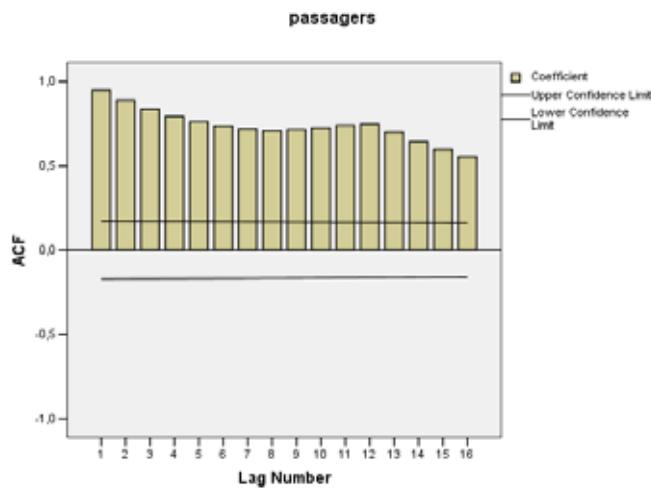


FIG. 13.82 – SAC de la série transformée par un log

La figure 13.83 nous permet de visualiser la série après la différentiation. Elle semble

maintenant tout à fait stationnaire du point de vue de la tendance et de la variance ; son SAC nous permettra sans doute de repérer l'effet saisonnier.

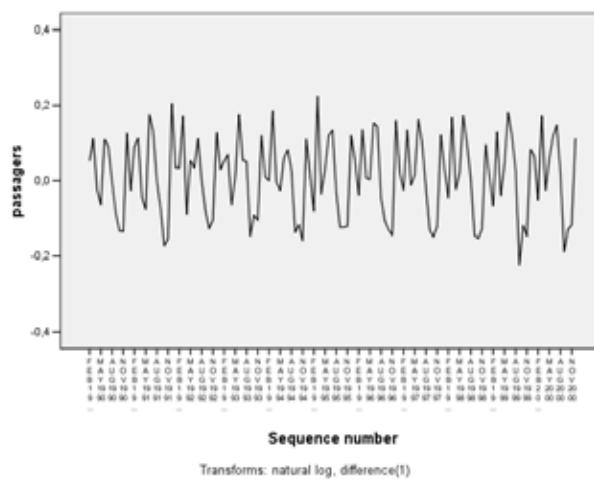


FIG. 13.83 – Série transformée par un log puis différentiée

La figure 13.84 nous donne le SAC de la série ; 48 lags sont présentés afin de bien cerner l'effet saisonnnier. Comme plusieurs autocorrélations sont significatives (à chaque lag multiple de 12), il faut effectuer une différentiation saisonnière avant de poursuivre.

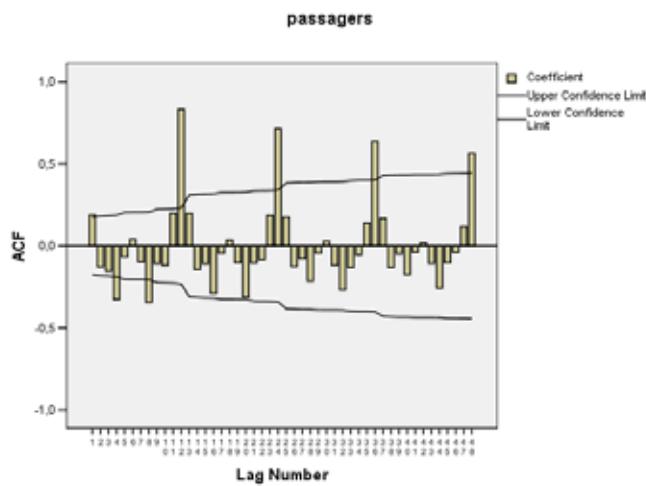


FIG. 13.84 – Le SAC de la série transformée par un log puis différentiée

La figure 13.85 nous présente le SAC et le SPAC **saisonniers** de la série **après** la

différentiation saisonnière.

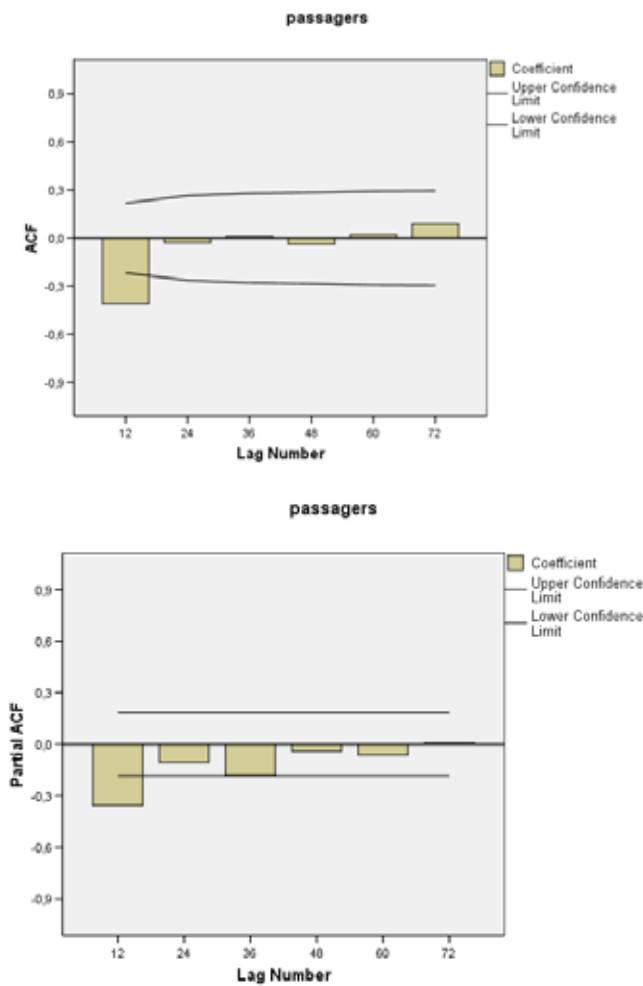


FIG. 13.85 – Le SAC et le SPAC saisonniers de la série après la différentiation saisonnière

La décroissance exponentielle semble plus plausible dans le SPAC, mais bizarrement en testant plusieurs modèles c'est en considérant la décroissance exponentielle dans le SAC que j'obtiens les meilleurs résultats. Il semble que ce soit le modèle ARIMA(0,1,0)(1,1,0)₁₂ qui soit approprié pour modéliser l'effet de saison de la série.

La figure 13.86 présente le SAC et le SPAC des résidus du modèle ARIMA(0,1,0)(1,1,0)₁₂. L'examen de ces figures vont nous permettre d'identifier le modèle approprié pour modéliser les fluctuations non-saisonnieres.

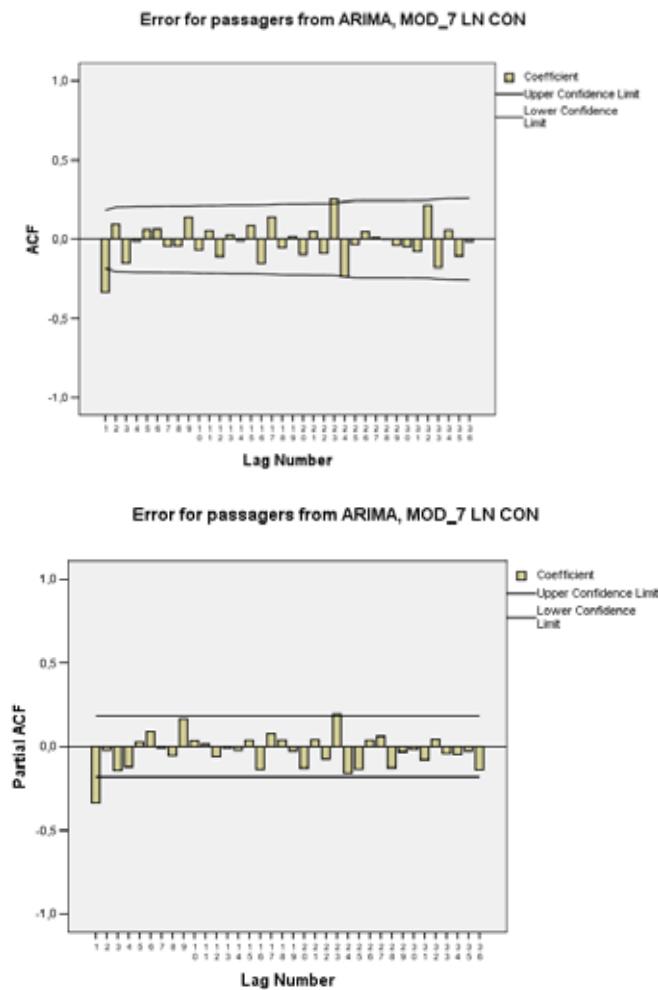


FIG. 13.86 – Le SAC et le SPAC des résidus du modèle ARIMA(0,1,0)(1,1,0)

La décroissance exponentielle semble ici plus plausible dans le SPAC, ce qui suggère un modèle de moyenne mobile. Et puisqu'il n'y a qu'un seul pic dans le SAC, l'ordre de ce modèle sera 1. Ainsi le modèle proposé pour modéliser la série est le modèle ARIMA(0,1,1)(1,1,0)₁₂. La figure 13.87 présente les sorties de ce modèle.

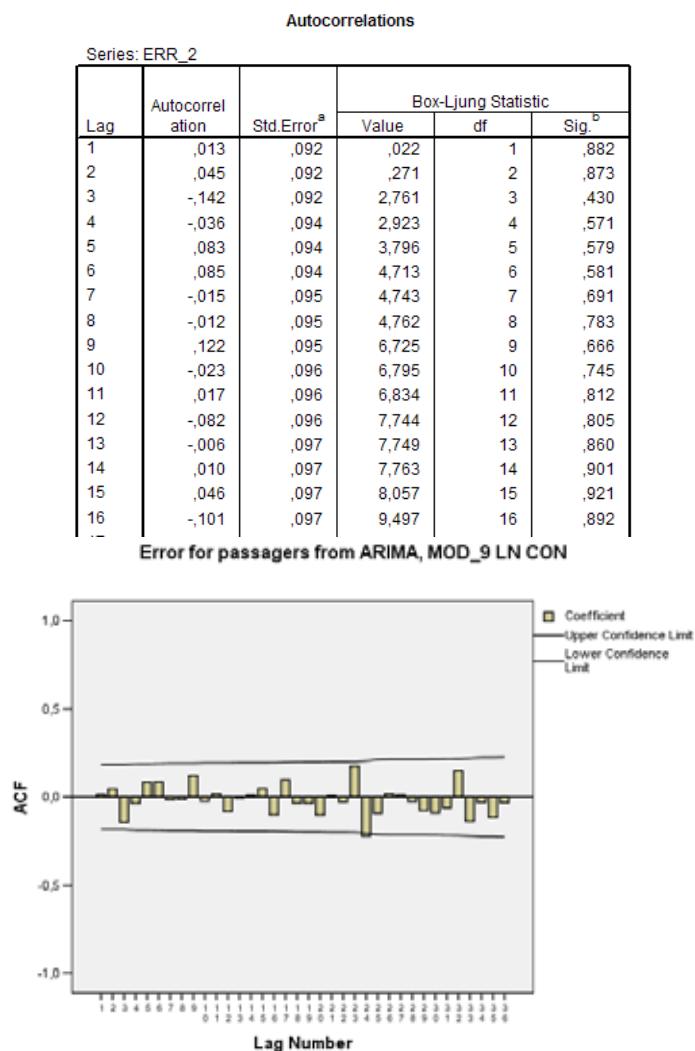
Residual Diagnostics	
Number of Residuals	119
Number of Parameters	2
Residual df	116
Adjusted Residual Sum of Squares	,172
Residual Sum of Squares	,175
Residual Variance	,001
Model Std. Error	,038
Log-Likelihood	220,396
Akaike's Information Criterion (AIC)	-434,791
Schwarz's Bayesian Criterion (BIC)	-426,454

Parameter Estimates					
		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	MA 1	,392	,086	4,553	,000
Seasonal Lags	Seasonal AR 1	-,467	,083	-5,616	,000
Constant		,000	,002	,125	,901

Melard's algorithm was used for estimation.

FIG. 13.87 – Les sorties du modèle ARIMA(0,1,1)(1,1,0)₁₂

On remarque que les deux paramètres sont significatifs. Aussi, le SAC des erreurs (figure 13.88) semble démontrer que le modèle est adéquat. Les *p*-values des Box-Ljung abondent dans ce sens puisqu'elles sont toutes largement supérieures à $\alpha = 0,05$.

FIG. 13.88 – Le SAC des erreurs du modèle ARIMA(0,1,1)(1,1,0)₁₂

De plus, la figure 13.89 nous permet de tester la normalité des résidus et de conclure qu'ils sont normaux.

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ERR_2	,054	119	,200*	,994	119	,918

*. This is a lower bound of the true significance.
^a. Lilliefors Significance Correction

FIG. 13.89 – Test de normalité des erreurs du modèle ARIMA(0,1,1)(1,1,0)₁₂

Seul petit bémol : la répartition des résidus (figure 13.90) ne semble pas aléatoire, la variance n'est pas constante (effet de cône). Mais un examen attentif de la série nous explique le pourquoi de cet effet : plus les années passent, et plus les fluctuations semblent régulières. Il est donc normal que le modèle arrive à mieux estimer le nombre de passagers au fil du temps.

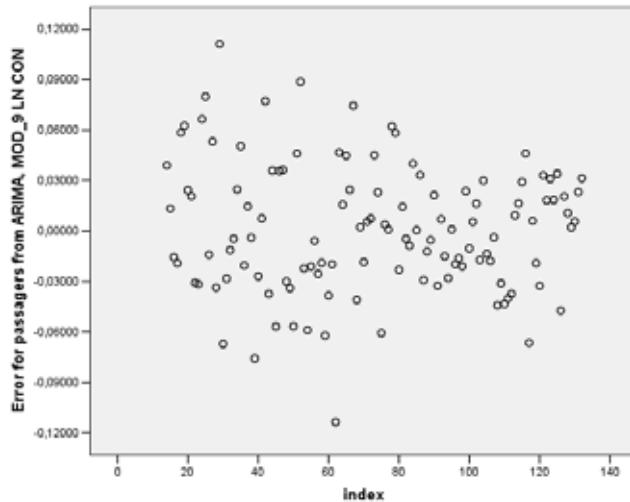


FIG. 13.90 – Répartition des erreurs du modèle ARIMA(0,1,1)(1,1,0)₁₂

13.9 ARIMA saisonnier avec E-Views

Pour illustrer les détails techniques relatifs à E-Views pour faire un modèle ARIMA saisonnier, reprenons l'exemple 13.8.2. La figure 13.91 nous montre tout d'abord dans le haut comment créer la série `logpass` qui est la série `passagers` sur laquelle on applique un log naturel. Ensuite, comme on l'avait vu dans l'exemple 13.8.2, on appliquait à cette série une différentiation saisonnière puis une différentiation ordinaire. La deuxième ligne dans le haut de la figure 13.91 nous montre comment appliquer à la fois le log et ces deux différentiations à la série originale. Il est à noter que dans la commande `d(x,k,l)`, k est l'ordre de la différentiation ordinaire, et l celui de la différentiation saisonnière. Le restant de la sortie nous montre le SAC et le SPAC de la série ainsi créée.

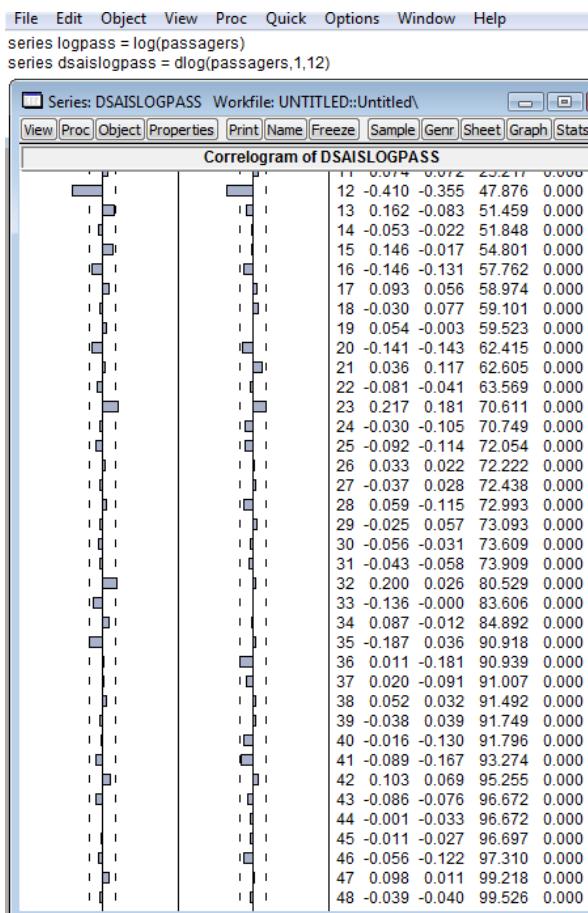


FIG. 13.91 – SAC et SPAC de la série `dsaislogpass`

Il est à noter que E-Views n'offre pas l'option de ne faire apparaître que les lags saisonniers dans le SAC et le SPAC. Il faut donc étudier les lags saisonniers parmi les lags ordinaires.

Ensuite, toujours dans l'exemple 13.8.2, pour modéliser l'effet de saison nous avons tenté un modèle ARIMA(0,1,0)(1,1,0)₁₂. La figure 13.92 nous montre comment spécifier ce modèle. Il est important d'effectuer les différentiations et autres transformations comme le log directement dans l'équation si on veut obtenir des prévisions par rapport aux unités de la série originale (ici en milliers de passagers). Autrement dit, il ne faut pas mettre la série `dsaislogpass` pour dépendante, mais plutôt reprendre la série `passagers` et lui appliquer les transformations.

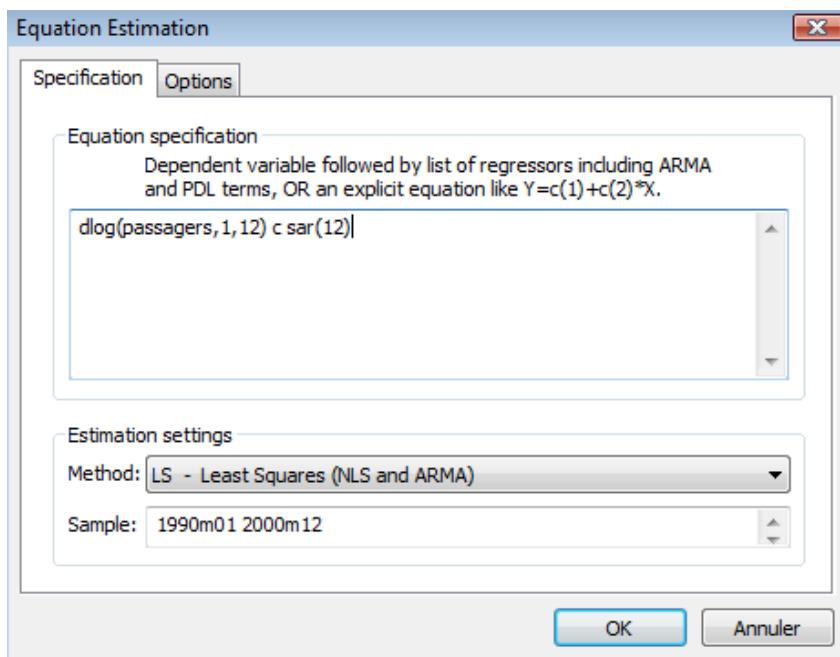


FIG. 13.92 – Pour évaluer le ARIMA(0,1,0)(1,1,0)₁₂

Ensuite, pour tout terme autorégressif ou de moyennes mobiles saisonnier, il suffit d'ajouter un `s` devant la notation habituellement utilisée. Par exemple, pour un terme saisonnier d'ordre 1 de moyennes mobiles, on aura `sma(12)` si les données sont mensuelles (pour des données trimestrielles ce serait `sma(4)`), et pour un terme saisonnier d'ordre

2 de moyennes mobiles, ce serait `sma(24)` pour des données mensuelles. Dans la figure 13.92 on voit qu'on a mis un terme saisonnier autorégressif d'ordre 1.

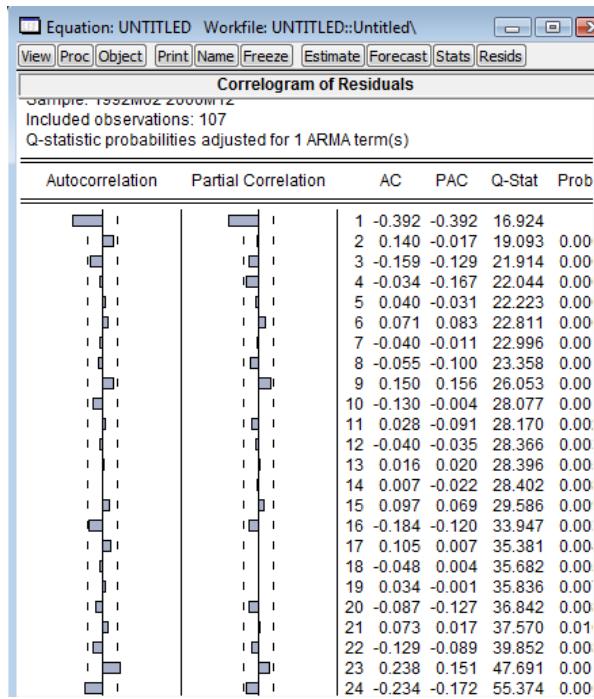
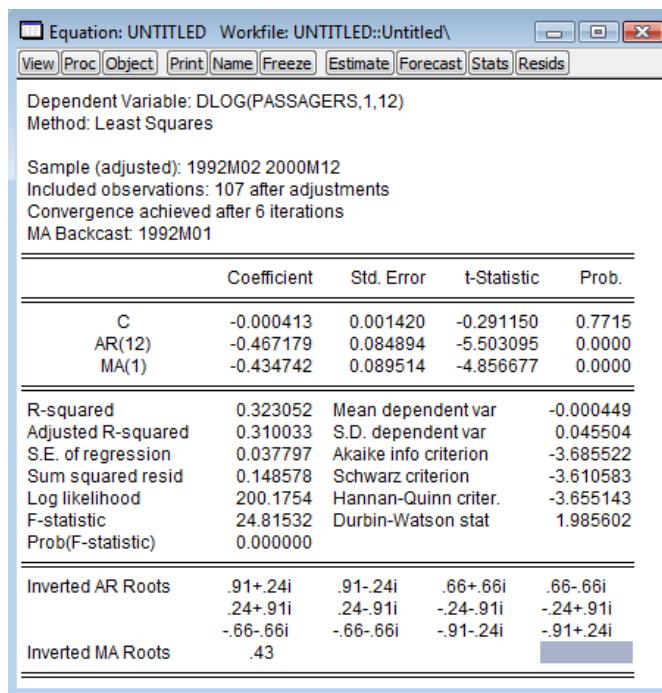


FIG. 13.93 – SAC et SPAC des résidus du modèle ARIMA(0,1,0)(1,1,0)₁₂

La figure 13.93 nous présente simplement le SAC et le SPAC des résidus du modèle ARIMA(0,1,0)(1,1,0)₁₂. Dans l'exemple 13.8.2, on avait alors identifié que pour l'effet non saisonnier c'était un modèle de moyennes mobiles d'ordre 1 qui semblait nécessaire. Pour spécifier ce modèle, il suffit d'ajouter le terme `ma(1)` à l'expression de la figure 13.92. L'estimation du modèle nous donne la sortie de la figure 13.94. Pour la suite de l'exemple les procédures sont déjà connues.

FIG. 13.94 – Modèle ARIMA(0,1,1)(1,1,0)₁₂

13.10 Quelques compléments

Il est possible d'inclure des variables indépendantes (dichotomiques ou continues) dans un modèle ARIMA. On obtient alors les estimations des paramètres pour ces variables, un peu comme dans une régression. Par contre, il est important de noter que dans SPSS, si une différentiation est effectuée sur la variable dépendante, alors les variables indépendantes seront également différencierées.

L'utilisation des variables binaires dans une procédure ARIMA est très intéressante. Elles permettent entre autre d'analyser l'effet d'une catastrophe ou d'une bonne nouvelle sur le parcours d'une série temporelle comme le cours des actions. La figure 13.95 illustre ce type de montée ou de descente.

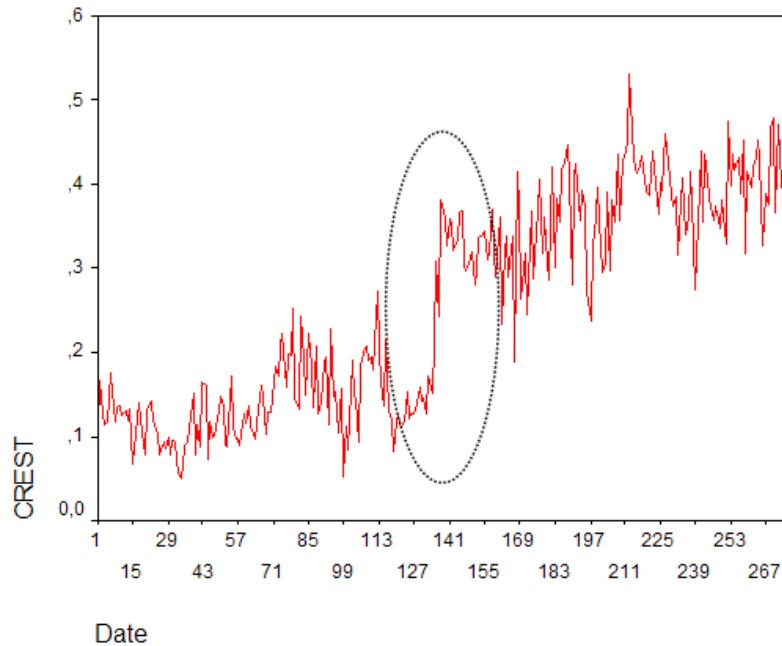


FIG. 13.95 – L'effet d'une bonne nouvelle sur le cours des actions de Crest

Il existe deux types de variables binaires : les *steps* et les *pulses* (parfois appelées *impulses*). Une fonction binaire *steps* est fixée à 0 jusqu'à la veille de la réalisation d'un événement, puis prend la valeur 1 à partir de la date précise de l'événement jusqu'à la

fin de la série. Une fonction binaire *pulse* est égale à 0 partout sauf à un endroit, soit la date précise de l'événement où elle est égale à 1. Le lien entre ces deux types de variable est simple : la différentiation d'une *step* produit une *pulse*.

Ainsi, une variable *step* introduite dans un modèle ARIMA($p,1,q$) devient une variable *pulse* (dans SPSS).

Si une variable *step* est significative dans le modèle étudié, alors son coefficient s'ajoute à la constante du modèle à partir de la date de l'événement. Ceci a pour effet de créer une montée ou une décente subite dans la série, qui pourrait être celle d'un titre boursier par exemple. Lorsque l'analyste désire étudier à l'aide d'un modèle mathématique l'impact d'un événement sur le changement de comportement d'une série chronologique, on parle alors d'*Intervention Analysis*. L'utilisation des variables binaires *pulses* et *steps* est bien courante dans ce type d'analyse.

La méthodologie de base liée aux analyses d'intervention est la suivante :

- Développer le modèle ARIMA avant l'événement.
- Ajouter une ou plusieurs variables binaires représentant l'intervalle d'intervention.
- Estimer le modèle à nouveau avec l'ensemble de la base de données.
- Interpréter les valeurs des variables binaires comme une mesure de l'effet de l'intervention.

Dans un autre ordre d'idée, supposons que l'analyste désire établir des prédictions à l'aide d'une régression. Une stratégie intéressante consiste alors à utiliser un modèle ARIMA pour effectuer des prédictions sur la variable indépendante X , pour ensuite effectuer de meilleures prédictions sur la variable dépendante Y avec la régression. Dans le

cadre d'une régression multiple, il est même possible d'utiliser un mélange de plusieurs modèles. L'utilisation de modèles mathématiques (ARIMA, régression, lissage, etc.) pour établir des prédictions sur des variables qui seront utilisées à l'intérieur d'un autre modèle forme un modèle général qui porte le nom de modèles de fonctions de transfert (*transfer functions model*). Il existe plusieurs types de modèles de fonctions de transfert et, dépendant du mélange, ils peuvent devenir très complexes.

Et lorsque l'analyste est intéressé à inclure des variables indépendantes dans un modèle, il peut être intéressant de détecter les *leading indicators*. Ces variables ont un impact déphasé sur la variable dépendante. Par exemple, le taux de chômage mensuel peut avoir un impact sur les ventes d'un produit deux mois plus tard.

Il est possible de découvrir un *leading indicator X* ainsi que le nombre de *lags* en avance que cet indice possède sur une variable *Y* en effectuant une analyse en corrélations croisées (*Cross Correlations*). Préalablement à l'analyse, il est impératif que les deux séries à l'étude soient stationnaires. On applique donc, si nécessaire, une ou des différentiations pour obtenir des séries stationnaires. Lorsque les ordres des différentiations sont les mêmes pour les deux séries, on peut effectuer celles-ci directement dans le menu des corrélations croisées.

Prenons la base de données `catalogue.sav` afin de présenter un exemple. Nous voulons effectuer une analyse de corrélations croisées entre la variable des ventes des vêtements pour femmes et la variable du nombre de catalogues postés. Les commandes sont les suivantes :

Menu SPSS :

- Graphs
- Time Series
- Cross-Correlations...

Dans la fenêtre Variables :

- femmes
- poste

Dans la fenêtre Transform :

- ✓ Difference 1

(car les deux séries présentent une tendance)

Dans le bouton Options... : Maximum Number of Lags : 12

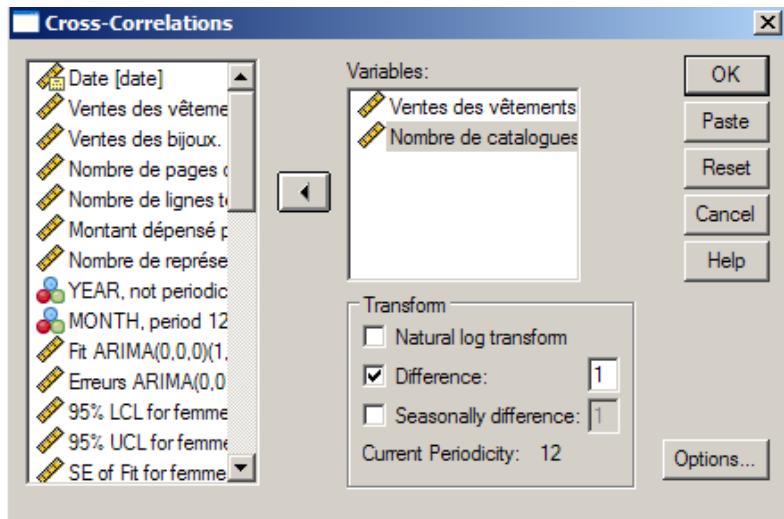


FIG. 13.96 – Analyse des corrélations croisées entre les variables `femmes` et `poste`

On obtient alors les sorties de la figure 13.97. On voit d'abord que la corrélation entre les deux variables différentierées est de 0,634 (au *lag* 0), c'est la plus forte. Pour interpréter les autres corrélations, l'ordre dans lequel les variables ont été présentées (ici `femmes` avant `poste`) est important. Cet ordre est rappelé dans les titres des figures. Le *lag* d'une corrélation correspond au *lag* qu'il faut considérer pour la deuxième variable.

Par exemple, au *lag* -1 on a une corrélation significative de -0,403 ; celle-ci correspond à la corrélation entre dfemmes_t (la variable **femmes** différentiée) et dposte_{t-1} (la variable **poste** différentiée et déphasée d'une période). D'autre part, la corrélation au *lag* 1 (qui a une valeur de 0,275) est la corrélation entre dfemmes_t et dposte_{t+1} (ou, si vous préférez, entre dfemmes_{t-1} et dposte_t).

Si le but est d'établir un modèle de prédiction pour les ventes de vêtements pour femmes, on s'intéressera alors aux *lags* négatifs qui nous montrent si les valeurs passées de la variable **poste** peuvent avoir un impact sur ces ventes. On voit par exemple que le nombre de catalogues postés un mois auparavant aura certainement un impact significatif sur les ventes du mois présent puisque la corrélation au *lag* -1 est significative.

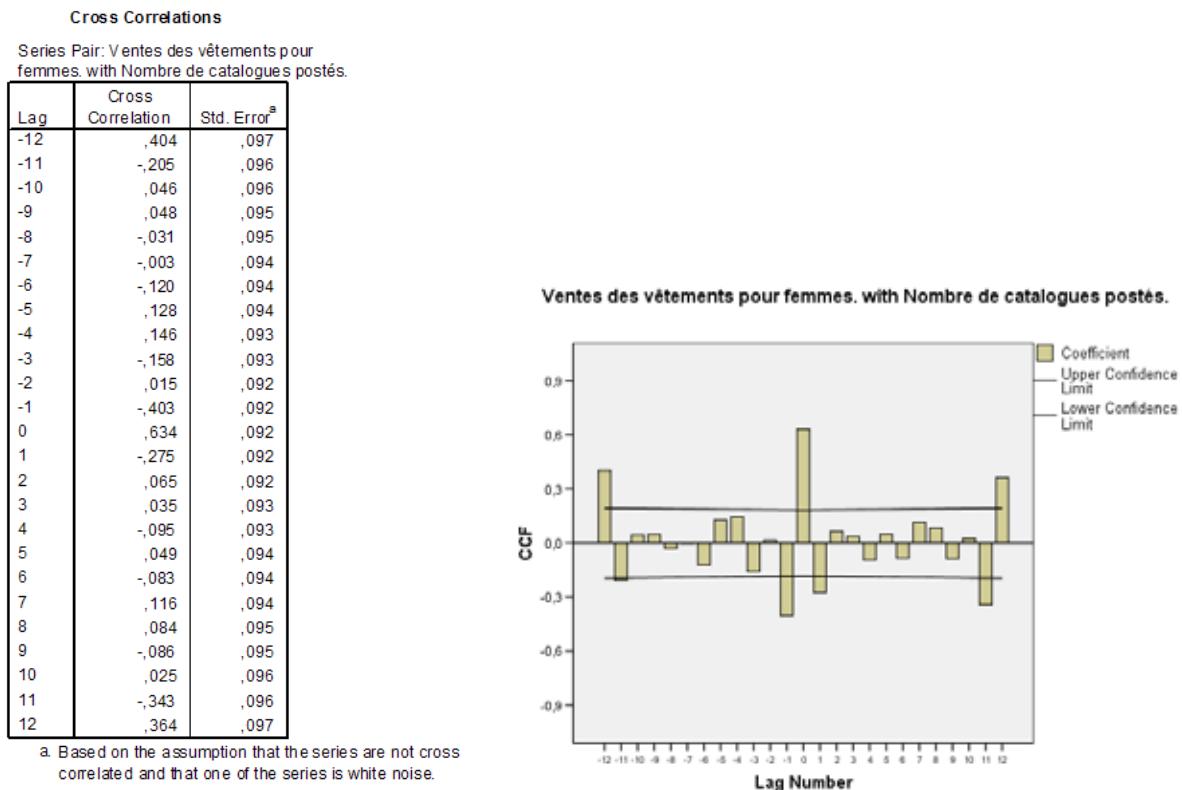


FIG. 13.97 – Sorties de l'analyse des corrélations croisées entre les variables **femmes** et **poste**

Voyons maintenant un exemple illustrant quelques-uns des concepts que l'on vient de décrire.

Exemple 13.10.1 Prenons la base de données `catalogue.sav`. On désire établir un modèle pour estimer les ventes des vêtements pour femmes. On se laisse l'année 1998 comme plage d'essai, et donc on établira le modèle sur les données de 1989 à 1997 inclusivement.

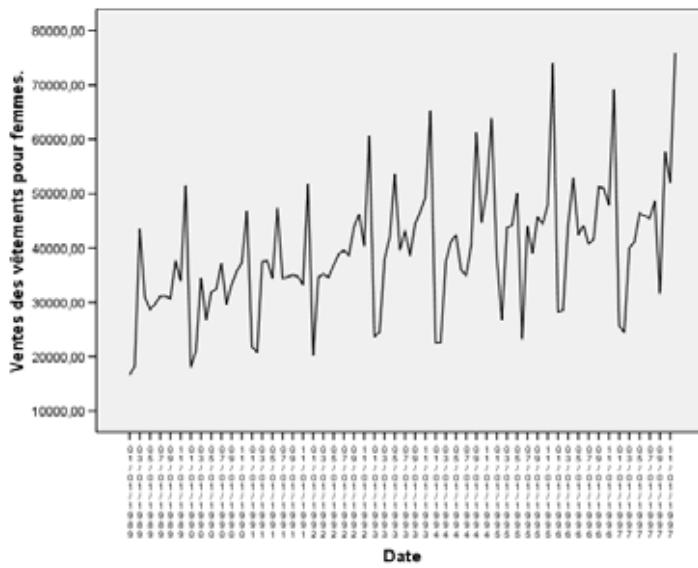


FIG. 13.98 – L'évolution des ventes des vêtements pour femmes

La figure 13.98 nous présente l'évolution des ventes jusqu'en 1997. Cette série présente une tendance et un effet saisonnier, ainsi qu'un peu d'hétéroscédasticité. Il aurait été bien d'essayer d'appliquer un log sur cette série.

La figure 13.99 présente le SAC de la série. Les autocorrélations des *lags* saisonniers étant fortement significatives et ayant une décroissance lente, on décide d'effectuer une différentiation saisonnière. La figure 13.100 présente la série avec une différentiation saisonnière ; on voit que cette différentiation a enlevé la tendance.

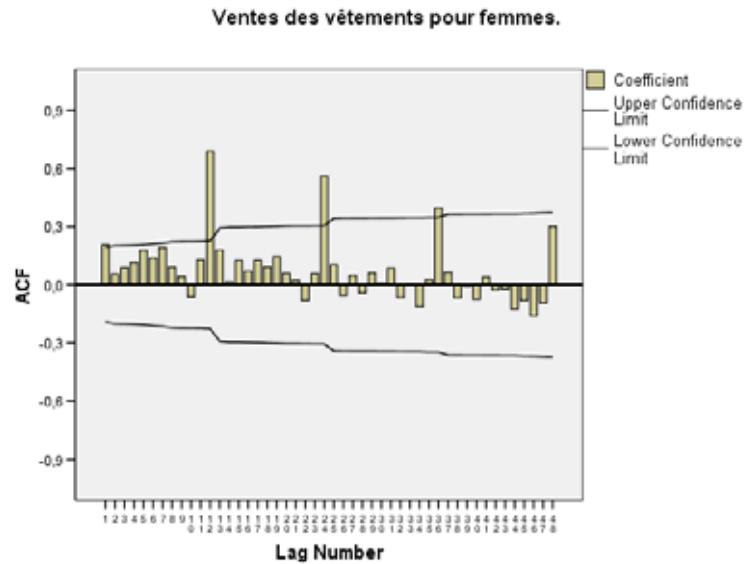


FIG. 13.99 – Le SAC de la série

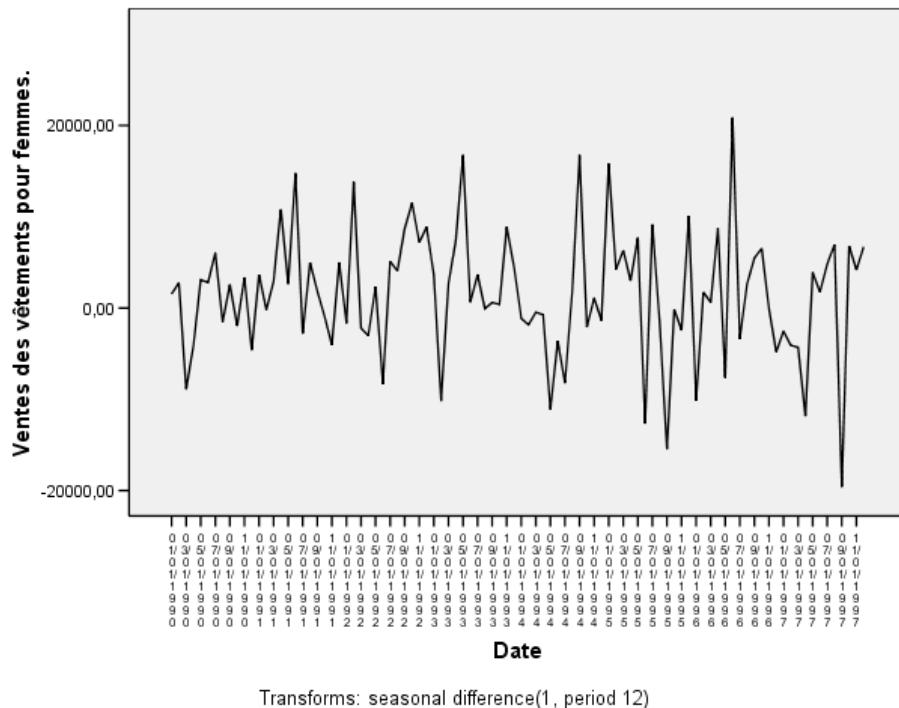


FIG. 13.100 – Série des ventes avec une différentiation saisonnière

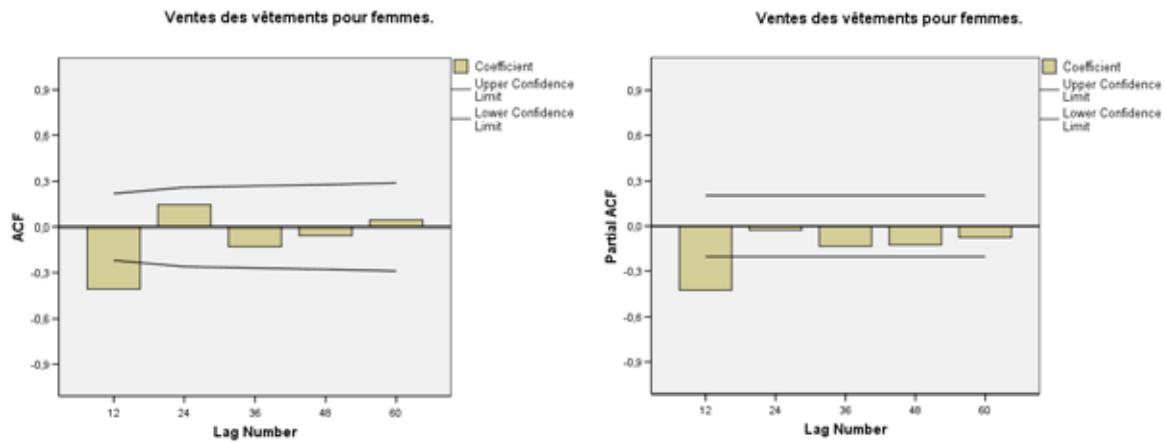


FIG. 13.101 – Le SAC et le SPAC saisonniers de la série avec une différentiation saisonnière

La figure 13.101 présente les SAC et SPAC saisonniers de la série différentiée. On tente à partir de ceux-ci d'établir le modèle ARIMA saisonnier. On décide ici de tenter un modèle autorégressif d'ordre 1 (on a donc considéré que la décroissance exponentielle est dans le SAC). Il aurait aussi été possible de tenter un modèle de moyenne mobile d'ordre 1.

On estime donc un modèle ARIMA(0,0,0)(1,1,0)₁₂; les sorties de ce modèle se retrouvent à la figure 13.102. On voit que le paramètre du modèle autorégressif saisonnier est jugé significatif puisque sa *p*-value est nulle. Remarquons aussi les valeurs du Log-Likelihood, du AIC et du BIC; nous les comparerons à celles d'un autre modèle tantôt.

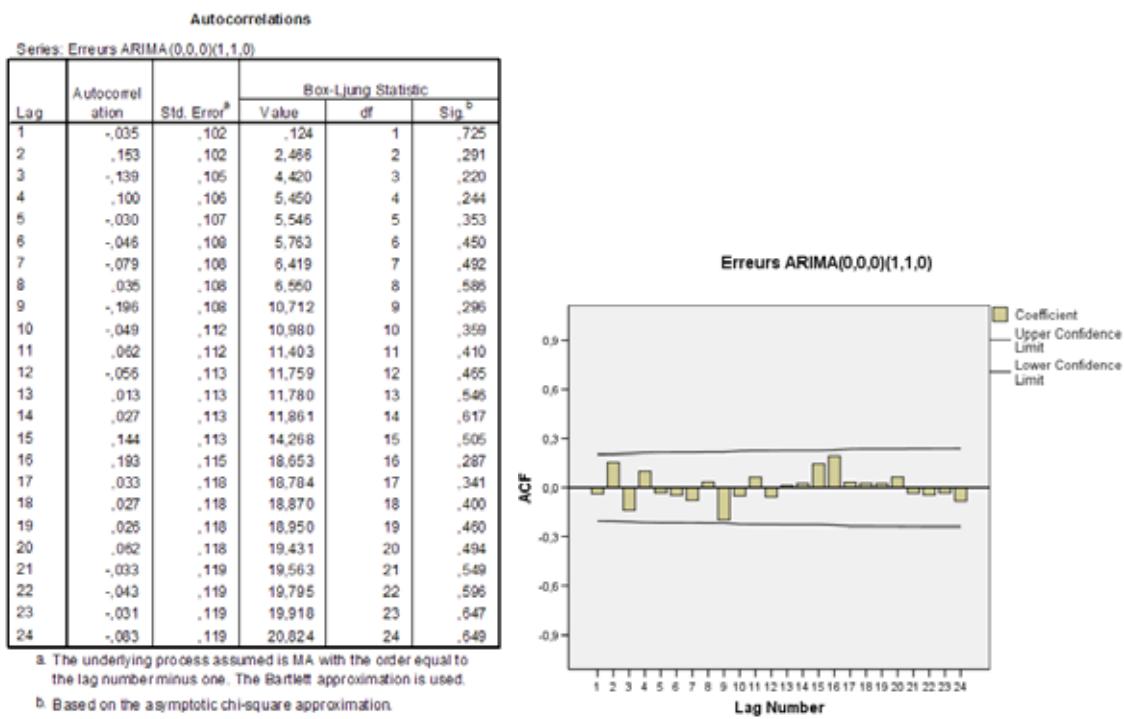
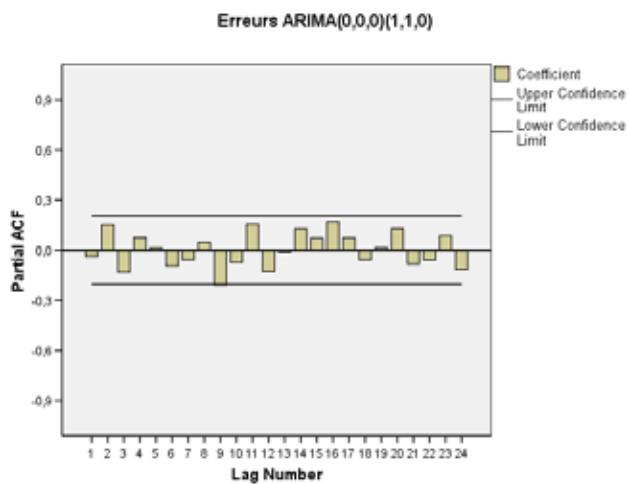
Residual Diagnostics	
Number of Residuals	96
Number of Parameters	1
Residual df	94
Adjusted Residual Sum of Squares	4E+009
Residual Sum of Squares	4E+009
Residual Variance	4E+007
Model Std. Error	6258,192
Log-Likelihood	-975,807
Akaike's Information Criterion (AIC)	1955,615
Schwarz's Bayesian Criterion (BIC)	1960,743

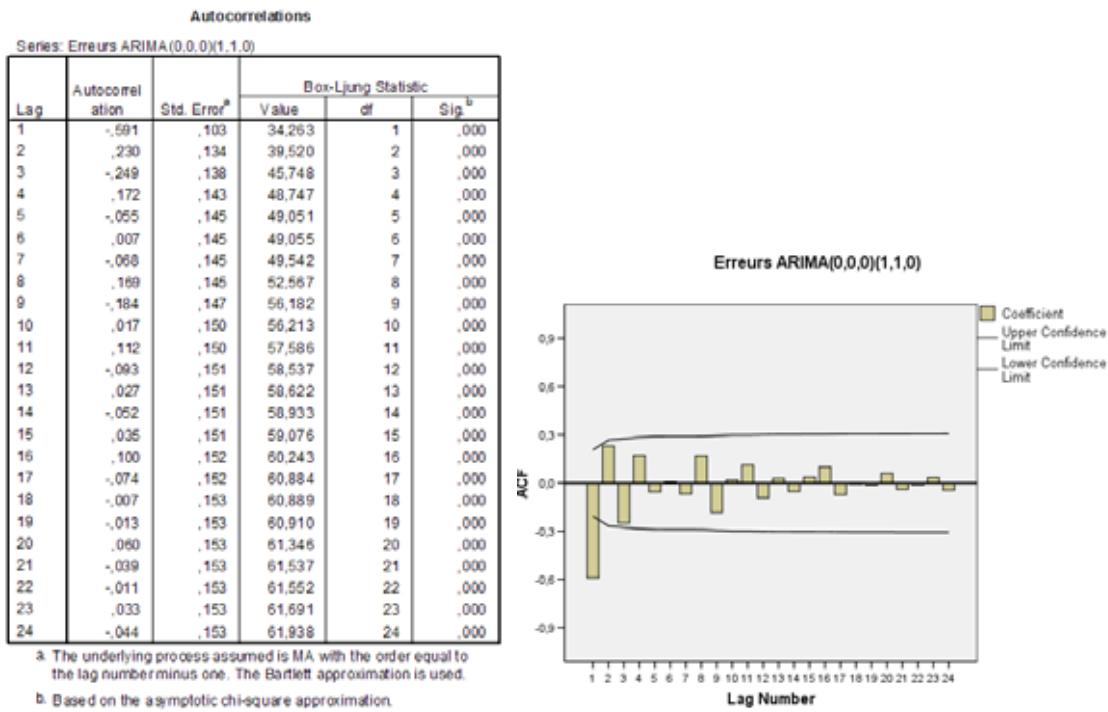
Parameter Estimates					
		Estimates	Std Error	t	Approx Sig
Seasonal Lags	Seasonal AR1	-,456	,099	-4,613	,000
Constant		1731,540	456,876	3,790	,000

Melard's algorithm was used for estimation.

FIG. 13.102 – Sorties pour le modèle ARIMA(0,0,0)(1,1,0)₁₂

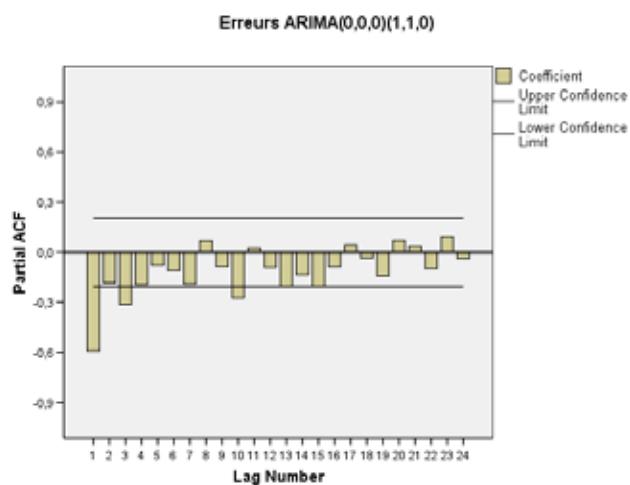
Les figures 13.103 et 13.104 présentent le SAC et le SPAC des résidus du modèle ARIMA(0,0,0)(1,1,0)₁₂. Aucune autocorrélation n'apparaît significative, et toutes les *p*-values des Box-Ljung sont supérieures à 0,05 (la plus petite est de 0,220 au *lag* 3). Donc le modèle semble adéquat. Mais dans le but de « débusquer » des informations permettant d'améliorer ce modèle, nous tentons ici une différentiation sur les résidus...

FIG. 13.103 – SAC des erreurs du modèle ARIMA(0,0,0)(1,1,0)₁₂FIG. 13.104 – SPAC des erreurs du modèle ARIMA(0,0,0)(1,1,0)₁₂

FIG. 13.105 – SAC des erreurs différentiées du modèle ARIMA(0,0,0)(1,1,0)₁₂

Les figures 13.105 et 13.106 présentent donc le SAC et le SPAC des résidus différentiés. Il semble clair ici qu'on a une décroissance exponentielle dans le SPAC, et une seule autocorrélation significative dans le SAC ; on tentera donc un modèle de moyenne mobile d'ordre 1 pour la partie régulière du ARIMA, ce qui donnera un ARIMA(0,1,1)(1,1,0)₁₂.

La figure 13.107 présente les sorties de ce modèle. On voit que les écart-types ne sont peut-être pas bien estimés. Par conséquent l'interprétation des *p*-values des paramètres n'est peut-être pas valide. Par contre on peut voir que les mesures d'adéquation sont meilleures que celles du modèle ARIMA(0,0,0)(1,1,0)₁₂. En effet, le Log-Likelihood est passé de -975,807 à -968,022, le AIC de 1955,615 à 1947,045, puis le BIC de 1960,743 à 1949,707. De plus, le SAC et le SPAC des résidus de ce modèle (figures 13.108 et 13.109) montrent que le modèle est adéquat puisque toutes les *p*-values des Box-Ljung sont supérieures à 0,05.

FIG. 13.106 – SPAC des erreurs différentiées du modèle ARIMA(0,0,0)(1,1,0)₁₂**Warnings**

Our tests have determined that the estimated model lies close to the boundary of the invertibility region. Although the moving average parameters are probably correctly estimated, their standard errors and covariances should be considered suspect.

Residual Diagnostics

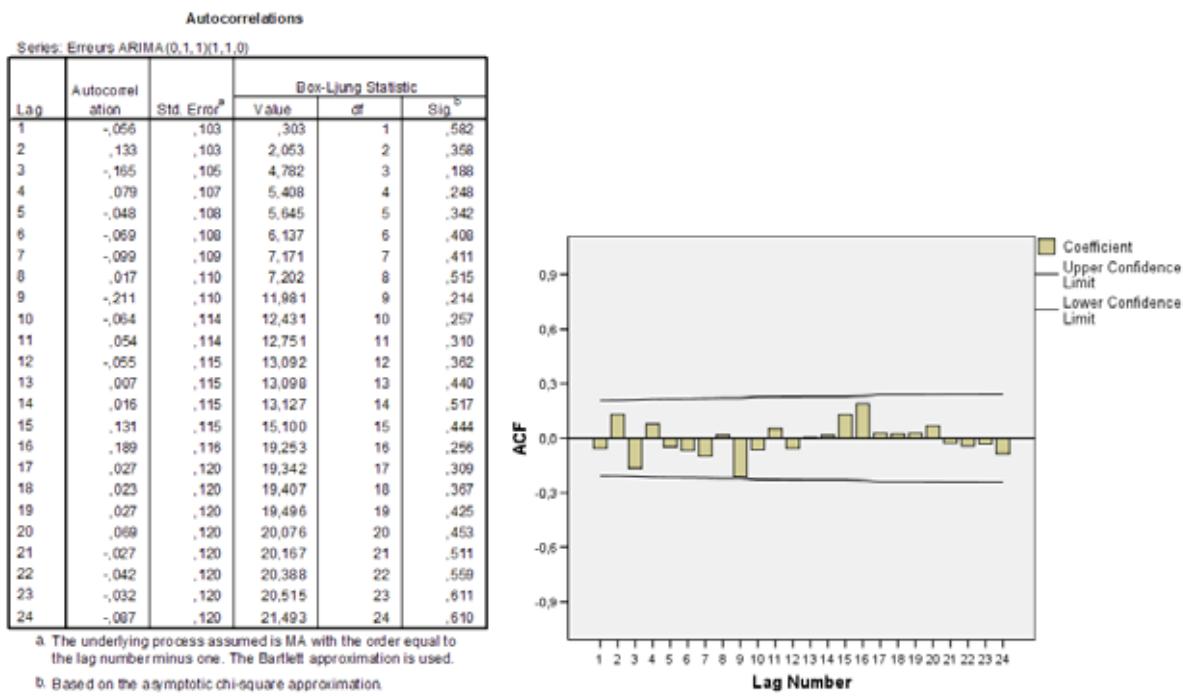
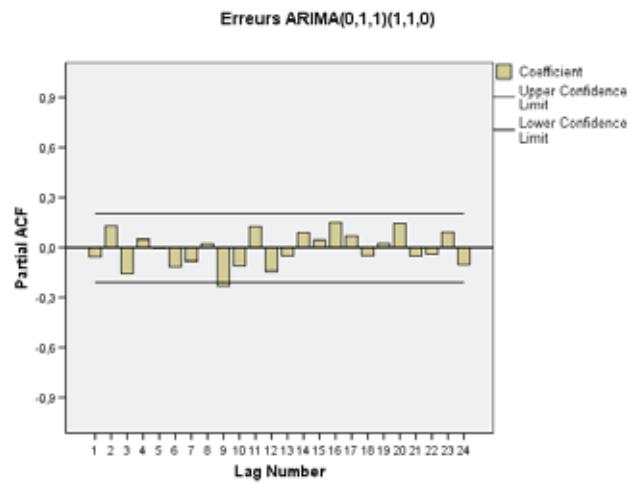
Number of Residuals	95
Number of Parameters	2
Residual df	92
Adjusted Residual Sum of Squares	4E+009
Residual Sum of Squares	4E+009
Residual Variance	4E+007
Model Std. Error	6276,795
Log-Likelihood	-968,022
Akaike's Information Criterion (AIC)	1942,045
Schwarz's Bayesian Criterion (BIC)	1949,707

Parameter Estimates

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	MA 1	,997	,483	2,063	,042
Seasonal Lags	Seasonal AR1	-,464	,098	-4,732	,000
Constant		-22,602	17,897	-1,263	,210

Melard's algorithm was used for estimation.

FIG. 13.107 – Sorties du modèle ARIMA(0,1,1)(1,1,0)₁₂

FIG. 13.108 – SAC des erreurs du modèle ARIMA(0,1,1)(1,1,0)₁₂FIG. 13.109 – SPAC des erreurs du modèle ARIMA(0,1,1)(1,1,0)₁₂

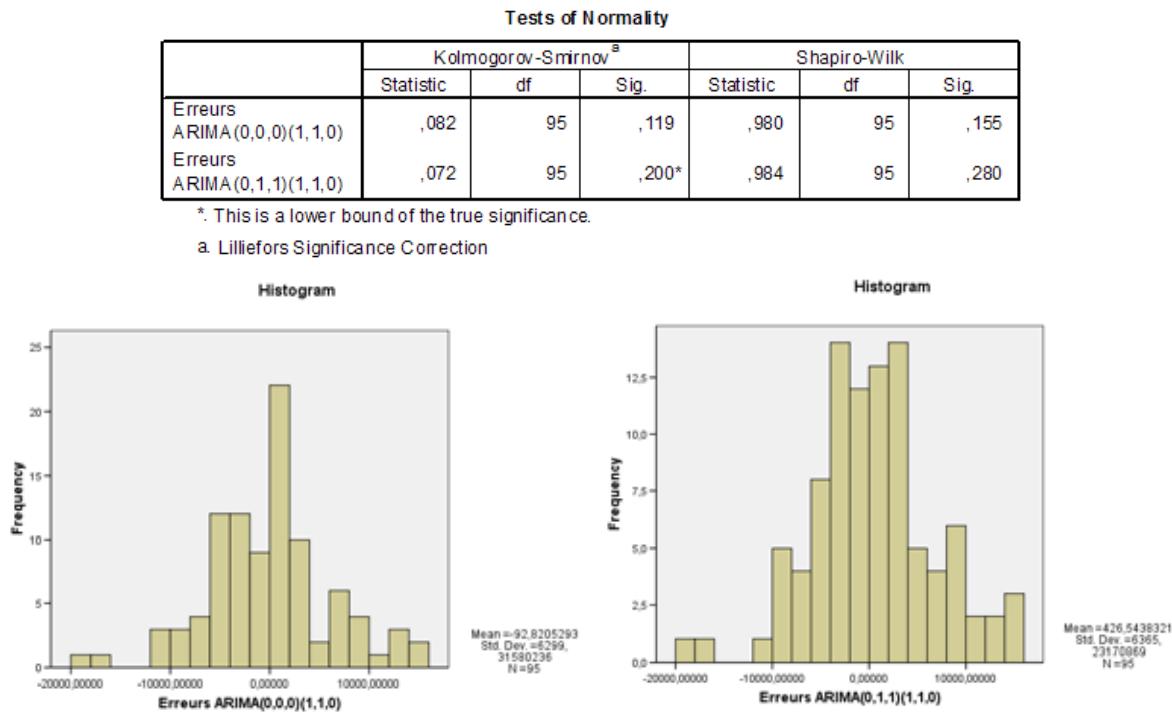


FIG. 13.110 – Normalité des erreurs des deux modèles

Vérifions maintenant la normalité des résidus des deux modèles (il faudrait énoncer le test d'hypothèses). Pour le modèle ARIMA(0,0,0)(1,1,0)₁₂, les *p*-values sont plus grandes que 0,05 (elles sont de 0,119 et 0,155). Il en est de même pour le modèle ARIMA(0,1,1)(1,1,0)₁₂ (les *p*-values sont de 0,2 et 0,28). Ainsi pour les deux modèles on peut conclure que les résidus se distribuent selon une loi normale.

La figure 13.111 montre la répartition des résidus des deux modèles. La répartition est assez aléatoire, mais il semble il y avoir quelques outliers, il faudrait investiguer davantage de ce côté.

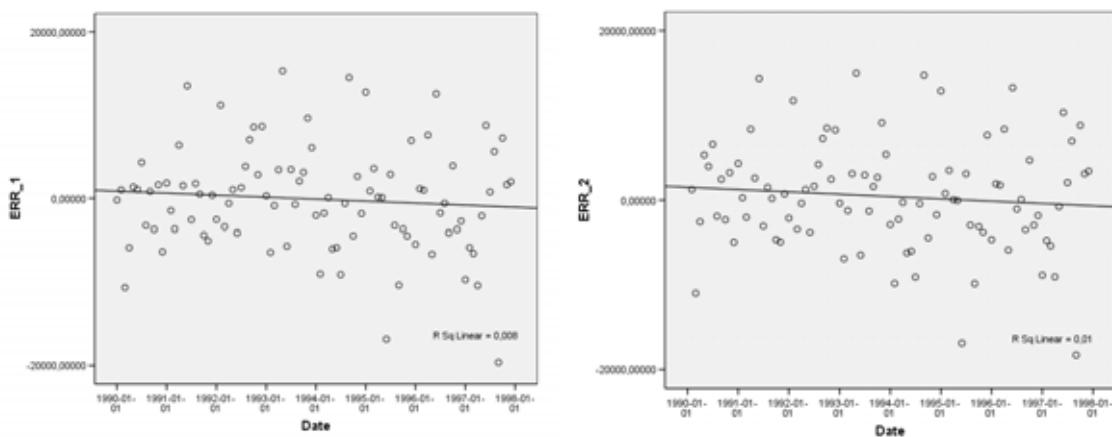
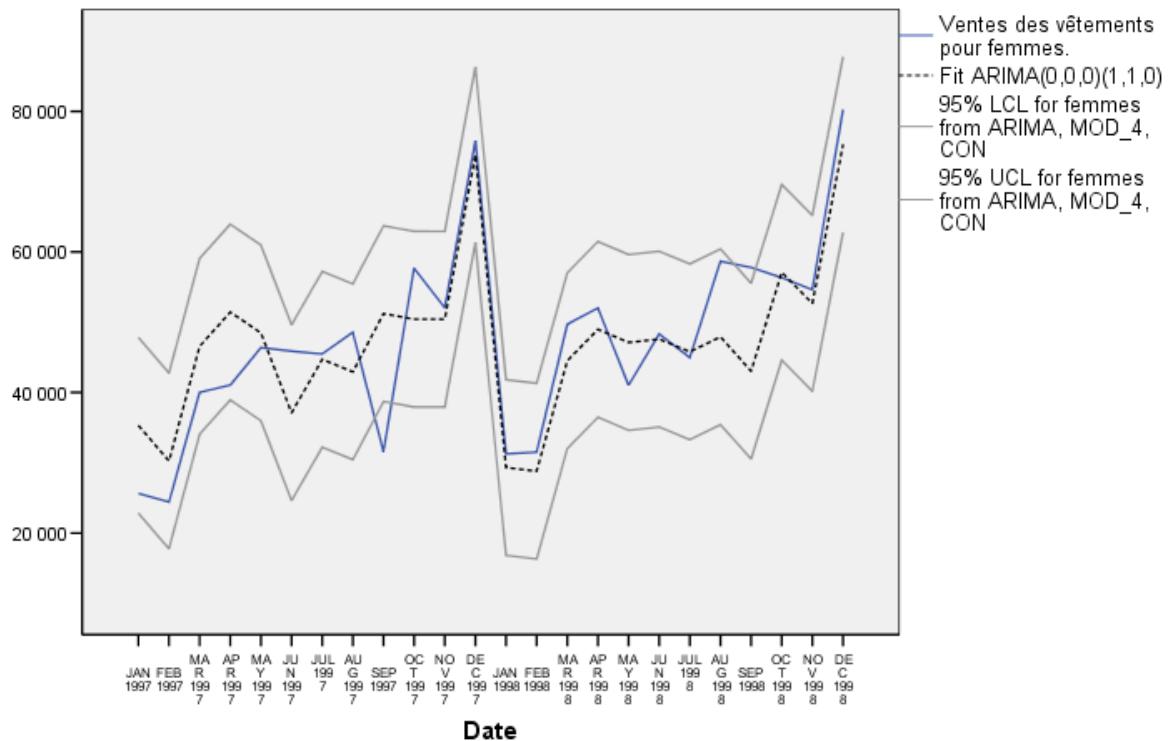


FIG. 13.111 – Répartition des erreurs des deux modèles

FIG. 13.112 – Estimations du modèle ARIMA(0,0,0)(1,1,0)₁₂ en 1997 et 1998

Finalement, les figures 13.112 et 13.113 montrent les estimations des deux modèles en 1997 et 1998 (il aurait été indiqué de regarder toutes les estimations). On voit que les deux modèles performent de façon très semblable. Les deux arrivent à suivre assez bien les observations (n'oublions pas que 1998 est la plage d'essai), et les intervalles de prédiction n'arrivent pas à contenir les véritables observations que deux fois pendant ces deux années, et c'est en septembre les deux fois. Pour parfaire la comparaison on pourrait utiliser des mesures de la section 12.8.

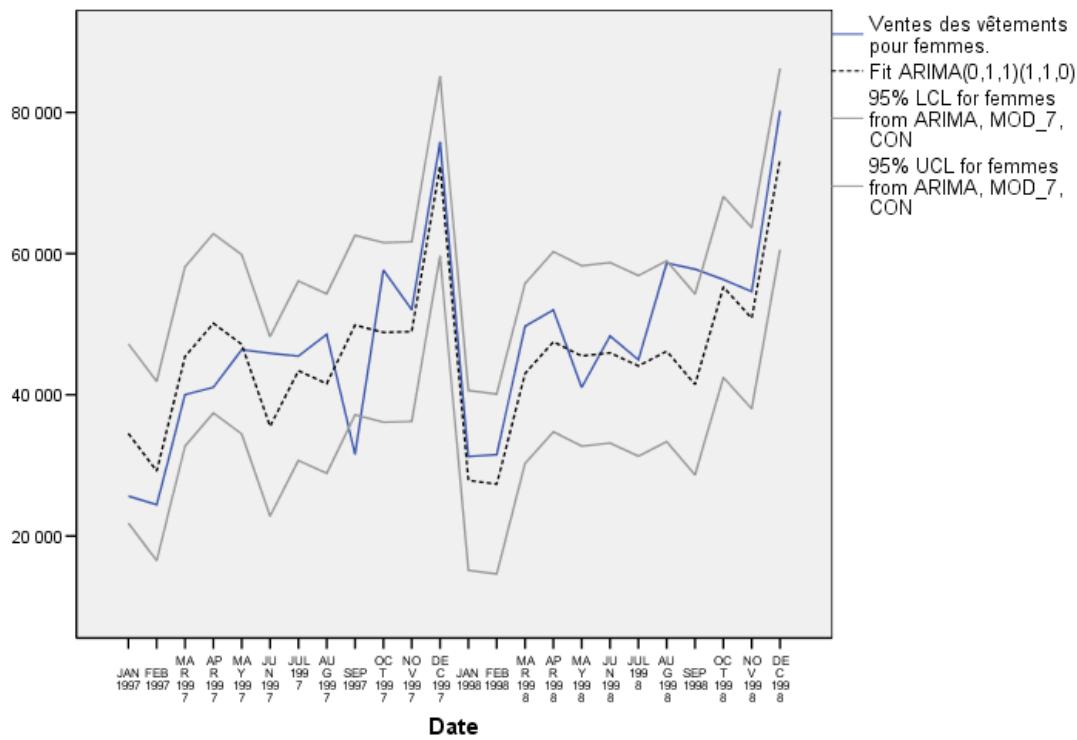


FIG. 13.113 – Estimations du modèle ARIMA(0,1,1)(1,1,0)₁₂ en 1997 et 1998

Une façon d'améliorer les modèles précédents serait d'ajouter des variables indépendantes. Les sorties qui suivent présentent le modèle ARIMA(0,1,1)(1,1,0)₁₂ auquel on a ajouté les variables **poste** et **service** (la figure 13.114 montre cet ajout). Faites l'analyse du modèle.

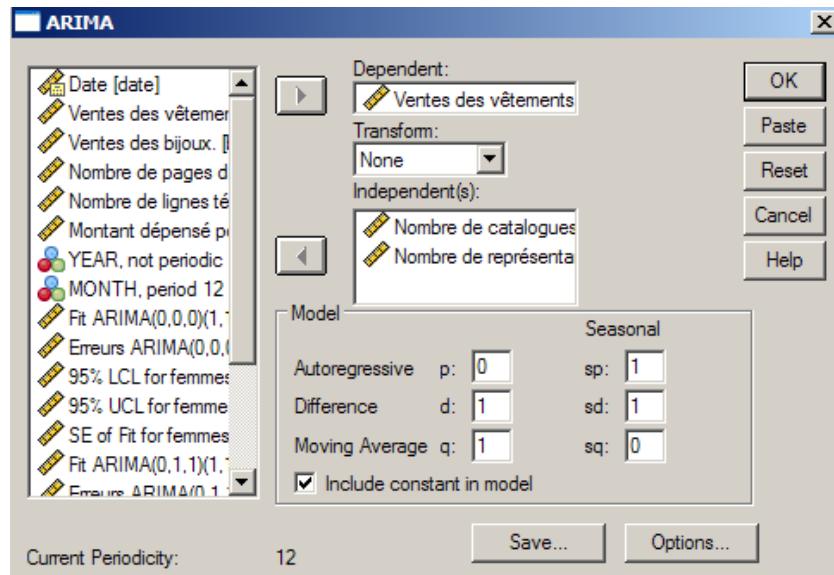


FIG. 13.114 – Ajout de variables explicatives au modèle

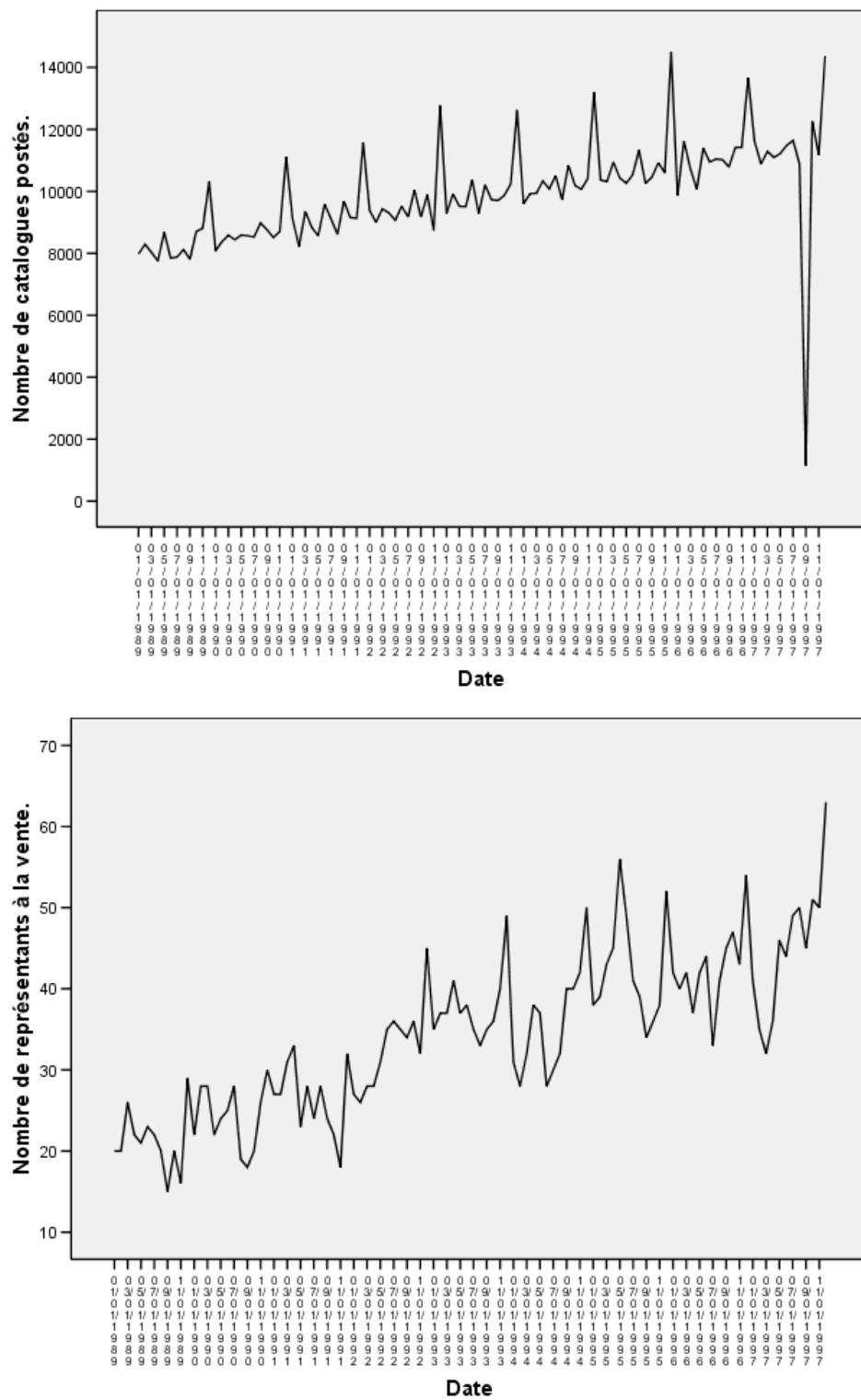


FIG. 13.115 – Graphes séquentiels des variables indépendantes

Warnings

Our tests have determined that the estimated model lies close to the boundary of the invertibility region. Although the moving average parameters are probably correctly estimated, their standard errors and covariances should be considered suspect.

Residual Diagnostics

Number of Residuals	95
Number of Parameters	2
Residual df	90
Adjusted Residual Sum of Squares	3E+009
Residual Sum of Squares	3E+009
Residual Variance	3E+007
Model Std. Error	5490,117
Log-Likelihood	-954,139
Akaike's Information Criterion (AIC)	1918,278
Schwarz's Bayesian Criterion (BIC)	1931,047

Parameter Estimates

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	MA1	,997	,581	1,717	,089
Seasonal Lags	Seasonal AR1	-,453	,094	-4,805	,000
Regression Coefficients	Nombre de catalogues postés.	2,036	,505	4,030	,000
	Nombre de représentants à la vente.	382,760	100,223	3,819	,000
Constant		-3,851	16,106	-,239	,812

Melard's algorithm was used for estimation.

FIG. 13.116 – Sorties du modèle ARIMA(0,1,1)(1,1,0)₁₂ avec des variables indépendantes

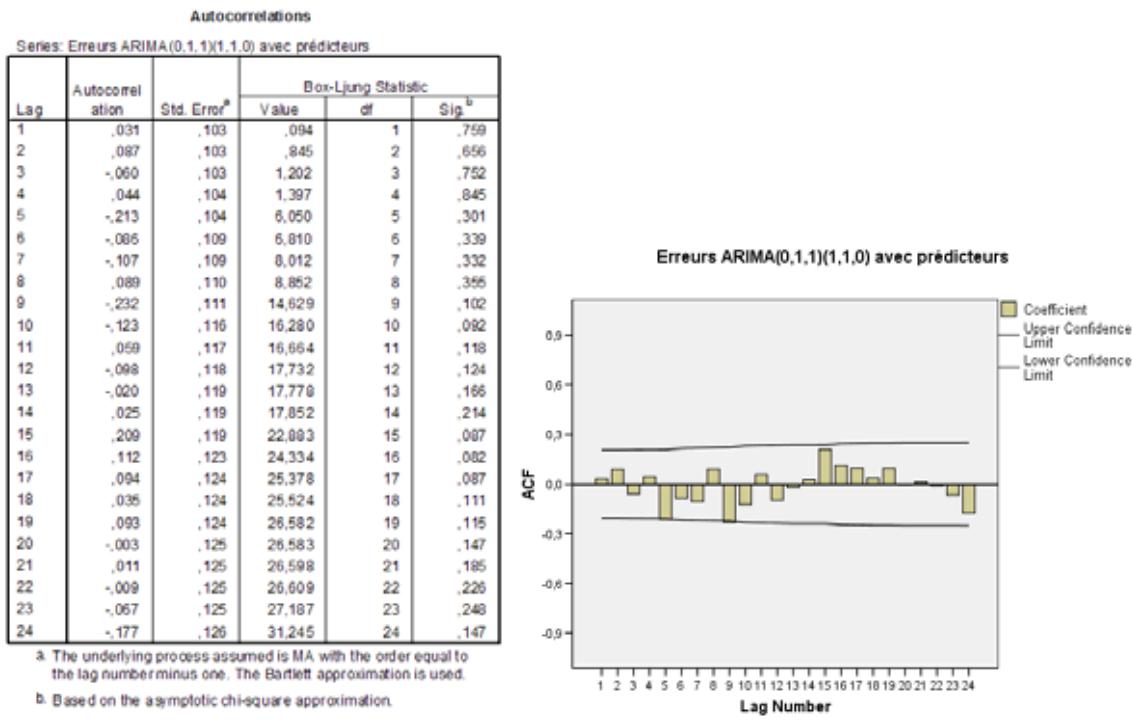


FIG. 13.117 – SAC des erreurs du modèle ARIMA(0,1,1)(1,1,0)₁₂ avec des variables indépendantes

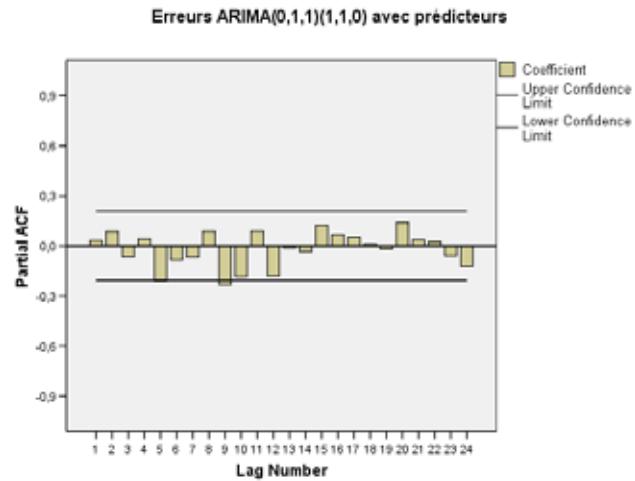


FIG. 13.118 – SPAC des erreurs du modèle ARIMA(0,1,1)(1,1,0)₁₂ avec des variables indépendantes

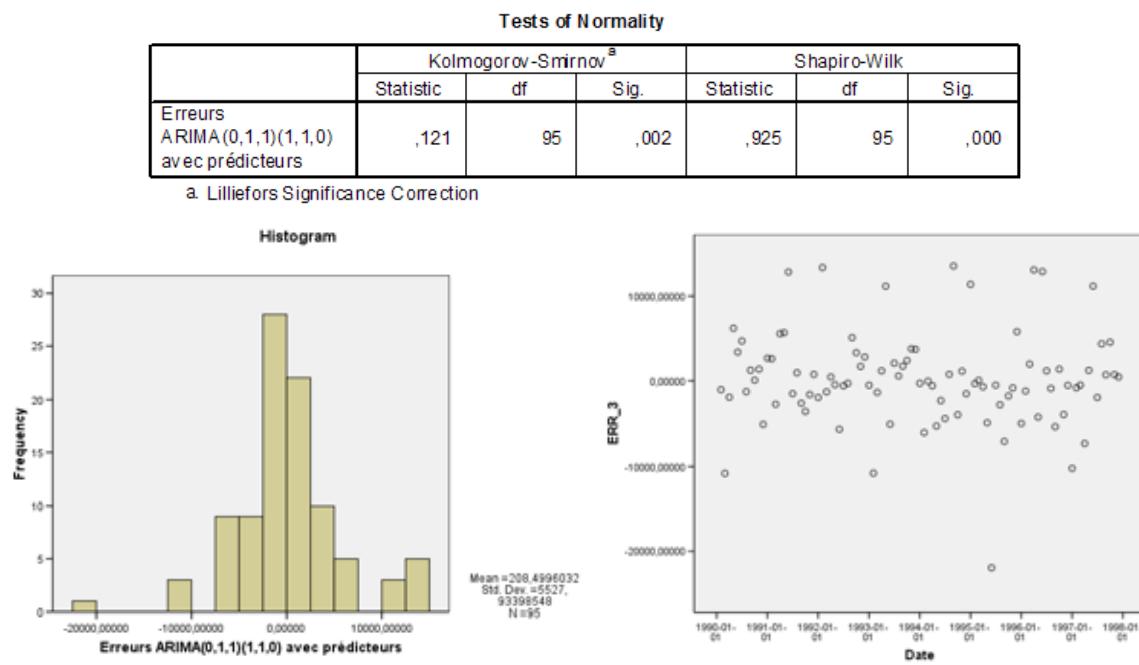


FIG. 13.119 – Test de normalité et répartition des erreurs du modèle ARIMA(0,1,1)(1,1,0)₁₂ avec des variables indépendantes

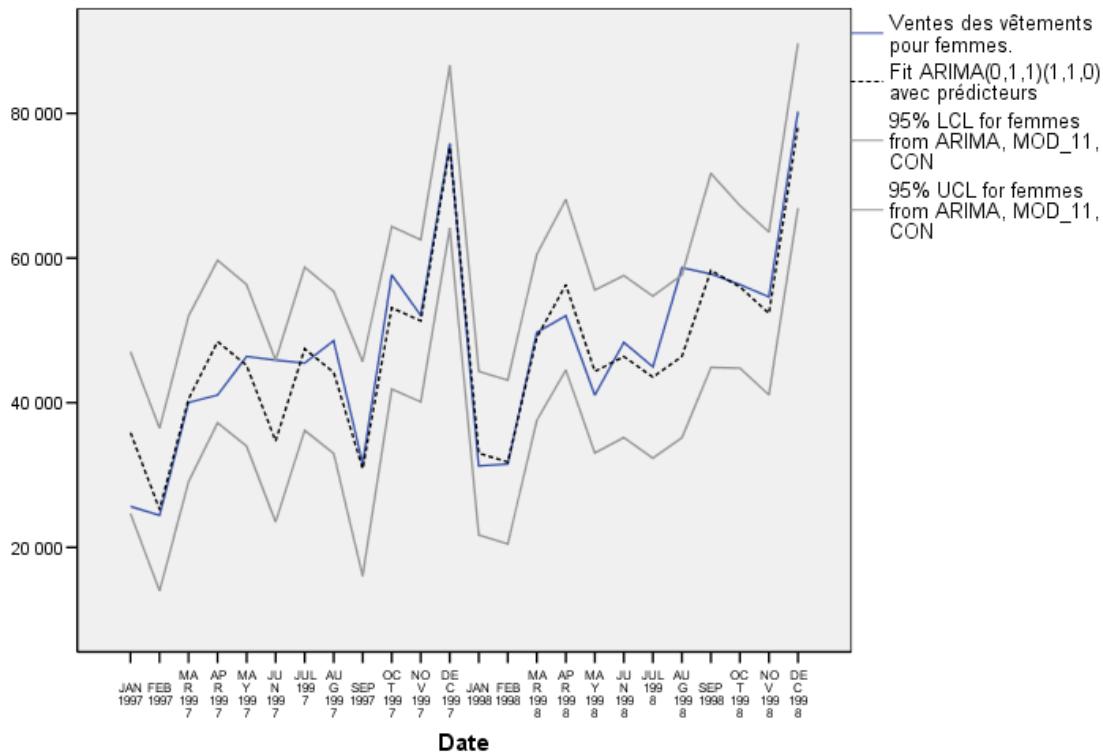


FIG. 13.120 – Estimations du modèle ARIMA $(0,1,1)(1,1,0)_{12}$ avec des variables indépendantes en 1997 et 1998

Les sorties 13.121 et 13.122 nous présentent maintenant une analyse de corrélations croisées, et un nouveau modèle. Analysez-les.

Cross Correlations
 Series Pair: Ventes des vêtements pour femmes.
 with Nombre de représentants à la vente.

Lag	Cross Correlation	Std. Error ^a
-12	,485	,097
-11	-,129	,096
-10	-,018	,096
-9	,007	,095
-8	-,086	,095
-7	-,006	,094
-6	,021	,094
-5	,044	,094
-4	,063	,093
-3	,001	,093
-2	-,058	,092
-1	-,470	,092
0	,699	,092
1	-,142	,092
2	-,034	,092
3	-,045	,093
4	-,013	,093
5	,023	,094
6	,062	,094
7	-,052	,094
8	,021	,095
9	,090	,095
10	-,101	,096
11	-,340	,096
12	,433	,097

a. Based on the assumption that the series are not cross correlated and that one of the series is white noise.

Ventes des vêtements pour femmes. with Nombre de représentants à la vente.

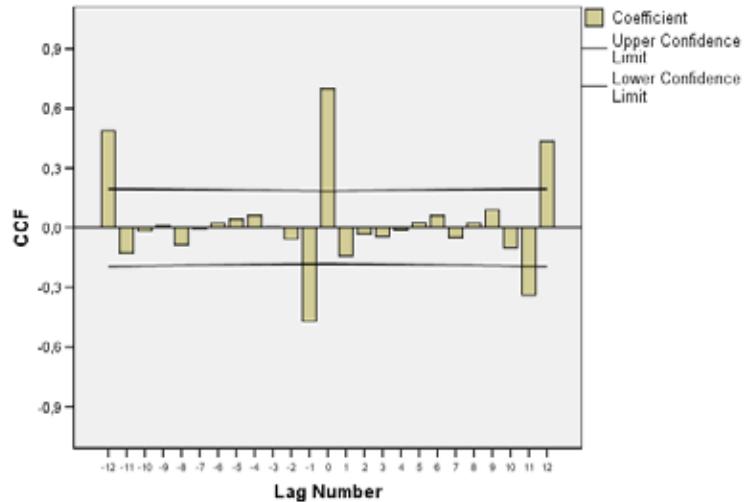


FIG. 13.121 – *Cross Correlations* entre les ventes des vêtements pour femmes et le nombre de représentants à la vente

Warnings

Our tests have determined that the estimated model lies close to the boundary of the invertibility region. Although the moving average parameters are probably correctly estimated, their standard errors and covariances should be considered suspect.

Residual Diagnostics

Number of Residuals	94
Number of Parameters	2
Residual df	88
Adjusted Residual Sum of Squares	3E+009
Residual Sum of Squares	3E+009
Residual Variance	3E+007
Model Std. Error	5479,445
Log-Likelihood	-943,212
Akaike's Information Criterion (AIC)	1898,423
Schwarz's Bayesian Criterion (BIC)	1913,683

Parameter Estimates

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	MA1	,993	,213	4,668	,000
Seasonal Lags	Seasonal AR1	-,457	,097	-4,697	,000
Regression Coefficients	Nombr de catalogues postés.	1,901	,510	3,727	,000
	Nombr de représentants à la vente.	499,788	122,481	4,081	,000
	LAGS(service, 1)	-205,726	124,628	-1,651	,102
Constant		-6,750	16,681	-,405	,687

Melard's algorithm was used for estimation.

FIG. 13.122 – Sorties du modèle ARIMA(0,1,1)(1,1,0)₁₂ avec des variables indépendantes, dont une déphasée

Chapitre 14

Les modèles ARCH et GARCH

La méthodologie ARCH (AutoRegressive Conditionnal Heteroscedasticity) présentée par Engle en 1982 a le mérite d'être la toute première approche en mesure de maîtriser efficacement une série temporelle aux prises avec une variance hétéroscléastique. Quelques années plus tard, Bollerslev (1986) présenta le modèle GARCH (Generalized AutoRegressive Conditionnal Heteroscedasticity) qui n'est qu'une évolution logique du modèle ARCH.

Dans le cadre de ce chapitre, le lecteur sera introduit aux philosophies ARCH et GARCH. Il sera amené à voir le lien de parenté entre les modèles ARIMA présentés au chapitre 11 et les modèles ARCH et GARCH. Un exemple illustrera l'utilisation du logiciel EViews.

14.1 Formulation du modèle ARCH

La classe des modèles ARCH et GARCH a pour objet de pallier les insuffisances des modèles ARIMA traditionnels non adaptés aux problématiques financières. Les séries

financières sont en effet caractérisées par une volatilité variable et par des phénomènes d'asymétries qui ne peuvent être pris en compte par les modélisations de type ARIMA.

En effet, on a vu que pour les modèles ARIMA la variance de la série est supposée constante. Or il est évident que bien des séries ne respectent pas cette condition ; elles montrent souvent des périodes de volatilité importante, suivies de périodes plus tranquilles. Dans de telles circonstances, supposer l'homocédasticité est inadéquat. Il est plus sensé de vouloir prédire la variance conditionnelle d'une série ; en effet, pourquoi ne pas modéliser cette variance qui révèle des informations importantes à propos du comportement de la série ?

Comme il n'est pas simple de choisir une variable externe pour expliquer la volatilité d'une série, Engle a proposé de procéder de façon autorégressive : d'abord, on modélise les résidus hétéroscédastiques d'un modèle de la manière suivante :

$$\epsilon_t = \nu_t \sqrt{h_t}$$

où $\nu_t \sim N(0, 1)$ est une variable aléatoire représentant le bruit de fond ou encore le bruit blanc qui évolue de façon indépendante de la fonction h_t , qui elle représente la composante hétéroscédastique conditionnelle de la variance du résidu (souvent appelée innovation). C'est la composante h_t qui sera modélisée de façon autorégressive.

En effet, un modèle ARCH(q) est un processus hétéroscédastique conditionnel autorégressif d'ordre q si la variance conditionnelle du résidu de l'équation temporelle h_t peut être modélisée ainsi :

$$h_t = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + \cdots + \alpha_q \epsilon_{t-q}^2.$$

14.2 Formulation du modèle GARCH

Bollerslev (1986) a généralisé le travail de Engel en développant une technique qui modélise la variance conditionnelle avec un processus ARIMA. On suppose d'abord que les

résidus d'un modèle temporel peut être modélisé par un modèle multiplicatif $\epsilon_t = \nu_t \sqrt{h_t}$ tel que proposé par Engle. Un modèle GARCH(p, q) modélise la variance conditionnelle du résidu de l'équation temporelle h_t de la façon suivante :

$$h_t = \alpha_0 + \beta_1 h_{t-1} + \cdots + \beta_p h_{t-p} + \alpha_1 \epsilon_{t-2}^2 + \cdots + \alpha_q \epsilon_{t-q}^2$$

où tous les coefficients sont positifs. On voit donc qu'un modèle GARCH(0, q) est un ARCH(q).

14.3 Procédure de création d'un modèle

Voici les étapes principales à suivre pour l'élaboration d'un modèle GARCH(p, q).

- Tout d'abord, créer un modèle de base avec par exemple une régression ou un modèle ARIMA.
- Étudier la présence d'hétéroscédasticité dans les résidus ; s'il y en a, c'est qu'un processus ARIMA devrait être appliqué aux résidus au carré. Il suffit donc d'étudier le SAC et le SPAC des résidus au carré pour voir si l'on doit appliquer un modèle GARCH.
- Si la conclusion à l'étape précédente indique qu'il faut appliquer un modèle GARCH, on tente d'identifier l'ordre du modèle GARCH avec le SAC et le SPAC des résidus au carré. Toujours commencer par les modèles les plus simples possibles.
- Un modèle est adéquat lorsque les résidus au carré ne révèlent plus de trace d'hétéroscédasticité, et autant que possible lorsque les résidus sont normaux.

14.4 Un exemple

On tentera ici de modéliser les données mensuelles du « S&P Composite index returns » de janvier 1954 à septembre 2001. Le fichier se nomme `s&pdata.wf1`. Les variables

sont `inf` (« consumer price index (CPI) inflation rate »), `dt_bill` (« three-month Treasury bill (T-bill) rate ») et `ret` pour le « S&P Composite index returns ».

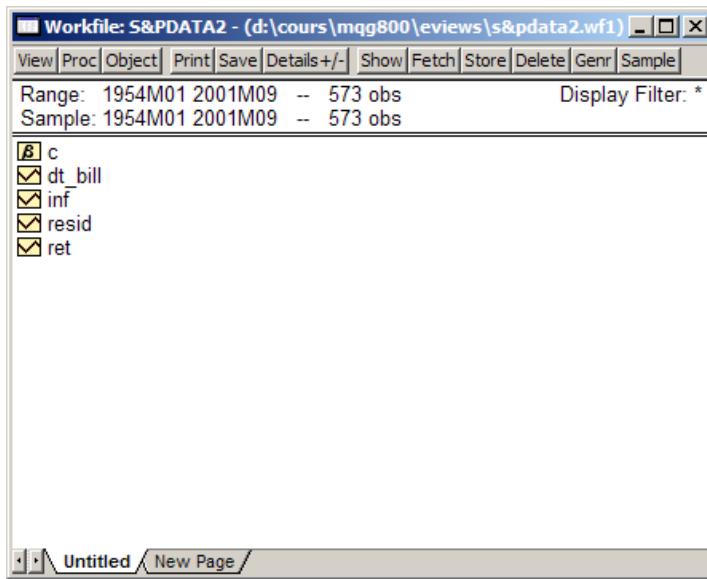


FIG. 14.1 – Aperçu de la feuille de travail dans EViews

Visualisons d'abord les séries des trois variables. Pour ce faire, il faut sélectionner la variable, puis cliquer avec le bouton droit pour pouvoir sélectionner `open`. On peut alors visualiser les données de cette variable (figure 14.2).

En cliquant sur le bouton `View` (en haut à gauche), plusieurs options s'offrent à nous. En sélectionnant `Graph` puis `Line`, on peut visualiser la série. On peut voir les trois séries dans les figures 14.3 et 14.4.

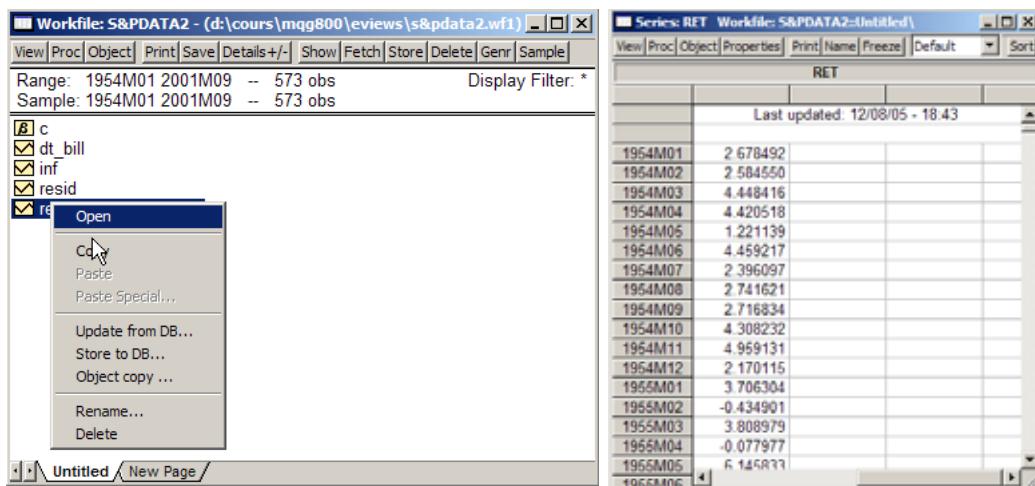
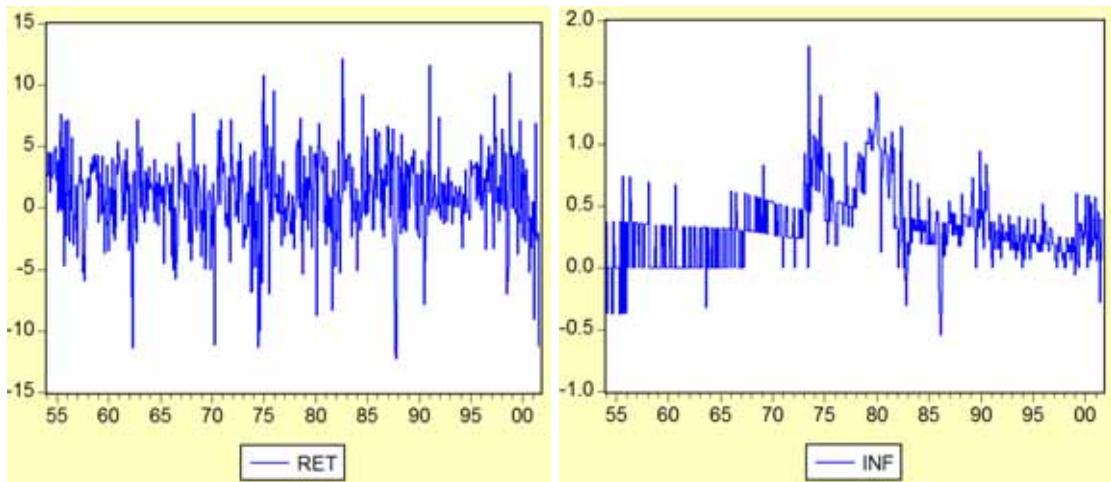
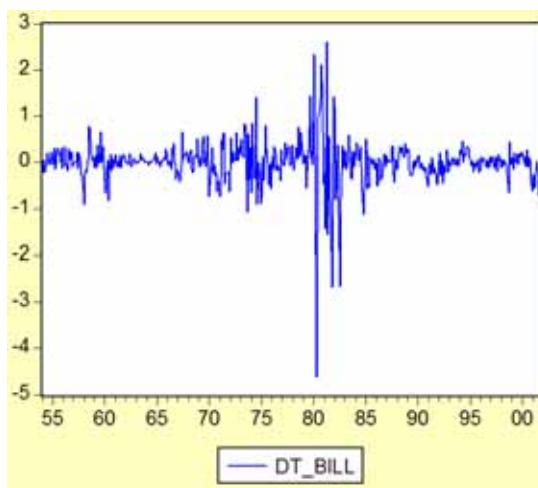


FIG. 14.2 – Sélectionner une variable

FIG. 14.3 – Visualisation des variables *ret* et *inf*

FIG. 14.4 – Visualisation de la variable `dt_bill`

On commencera ici par modéliser la variable `ret` par les variables `inf` et `dt_bill` déphasées de 1 mois. Pour indiquer ceci à EViews, il faut aller dans le menu **Quick**, puis **Estimate Equation...**. Ensuite, il faut écrire l'équation qui nous intéresse en donnant d'abord le nom de la dépendante, puis on indique `c` si on veut une constante, puis le nom des variables indépendantes. Pour indiquer une variable déphasée d'une période, on inscrit `(-1)` à la suite du nom de la variable. Voir la figure 14.5.

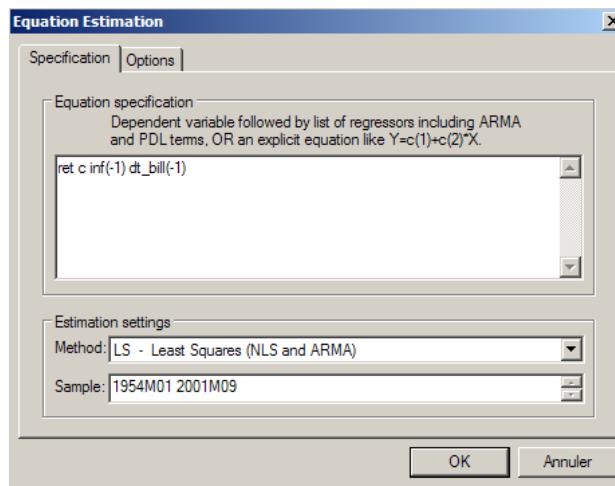


FIG. 14.5 – L'écriture du modèle

On obtient alors la sortie de la figure 14.6.

Equation: UNTITLED Workfile: S&PDATA2::Untitled\				
View	Proc	Object	Print	Name
Freeze	Estimate	Forecast	Stats	Resids
Dependent Variable: RET Method: Least Squares Date: 12/08/05 Time: 20:46 Sample (adjusted): 1954M02 2001M08 Included observations: 571 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.502641	0.204291	7.355382	0.0000
INF(-1)	-1.521212	0.452477	-3.361968	0.0008
DT_BILL(-1)	-1.414550	0.295769	-4.782609	0.0000
R-squared	0.060304	Mean dependent var	0.994713	
Adjusted R-squared	0.056995	S.D. dependent var	3.42856	
S.E. of regression	3.329417	Akaike info criterion	5.248712	
Sum squared resid	6296.290	Schwarz criterion	5.271553	
Log likelihood	-1495.507	F-statistic	18.22539	
Durbin-Watson stat	1.554574	Prob(F-statistic)	0.000000	

FIG. 14.6 – Les estimations du modèle

En examinant les cotes-*t* des variables ainsi que leurs *p*-values (Prob.), on voit que les deux paramètres sont significatifs. On voit aussi que la *p*-value de la statistique *F* est nulle, ce qui signifie que le modèle est significatif (comme dans une table ANOVA).

En allant dans le bouton *View*, on peut faire plusieurs manipulations pour évaluer si ce modèle est adéquat. En allant dans *Actual,Fitted,Residuals* puis dans *Actual,Fitted,Residuals Graph*, on obtient un graphe contenant à la fois les valeurs réelles, les estimations et les résidus. Voir la figure 14.7; la série du bas représente les résidus.

On voit que les estimations ne suivent pas de très près toutes les fluctuations. En allant voir le SAC et le SPAC des erreurs de ce modèles, il sera possible de voir si on peut améliorer ce modèle en ajoutant des composantes d'un modèle ARIMA. Pour obtenir ceci il faut aller dans *View, Residual Tests* puis *Correlogram - Q-statistics*. On obtient alors la figure 14.8.

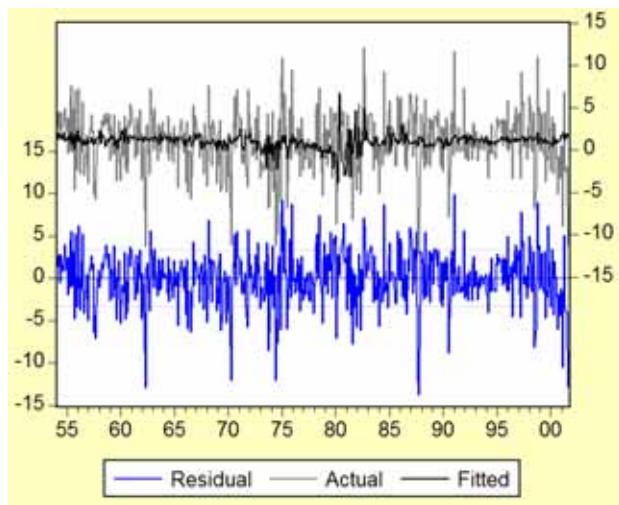


FIG. 14.7 – Les valeurs réelles, les estimations et les résidus

Date: 12/08/05 Time: 21:04
 Sample: 1954M02 2001M08
 Included observations: 571

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
1	1	1	0.209	0.209	25.170 0.000
2	2	2	0.019	-0.026	25.369 0.000
3	3	3	0.033	0.037	26.013 0.000
4	4	4	0.026	0.013	26.415 0.000
5	5	5	0.066	0.061	28.943 0.000
6	6	6	-0.030	-0.060	29.460 0.000
7	7	7	-0.053	-0.036	31.106 0.000
8	8	8	0.050	0.068	32.567 0.000
9	9	9	0.008	-0.018	32.606 0.000
10	10	10	-0.042	-0.043	33.649 0.000
11	11	11	0.064	0.091	36.077 0.000
12	12	12	-0.021	-0.054	36.334 0.000

FIG. 14.8 – Extrait du SAC et du SPAC des erreurs

La figure 14.8 nous indique qu'un modèle ARIMA(1,0,0) ou ARIMA(0,0,1) pourrait améliorer notre modèle. Nous tenterons ici d'ajouter une partie autorégressive d'ordre 1 à notre modèle. Il suffit pour ceci d'ajouter le terme ar(1) au modèle. Pour des moyennes mobiles il aurait fallu ajouter ma(1). Pour ajuster l'équation du modèle il suffit d'aller dans Proc puis Specify/Estimate. On peut alors modifier le modèle (voir la figure 14.9).

La sortie associée à ce nouveau modèle est donnée dans la figure 14.10.

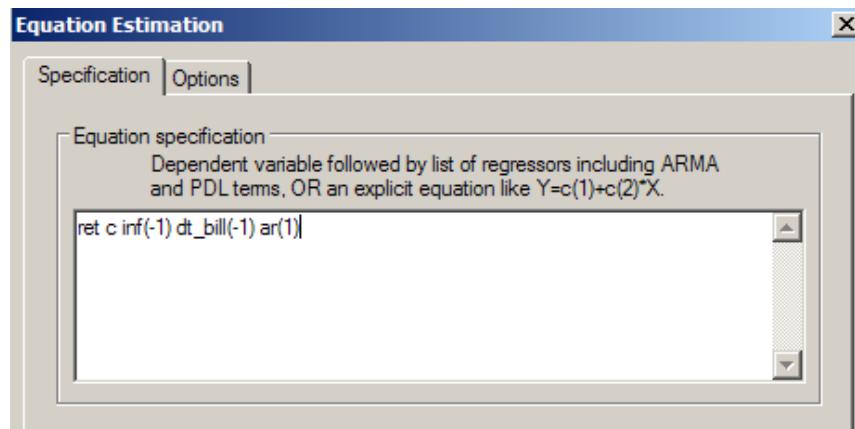


FIG. 14.9 – Modification du modèle

Dependent Variable: RET				
Method: Least Squares				
Date: 12/08/05 Time: 21:14				
Sample (adjusted): 1954M03 2001M08				
Included observations: 570 after adjustments				
Convergence achieved after 4 iterations				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.463743	0.240354	6.089943	0.0000
INF(-1)	-1.432679	0.501672	-2.855808	0.0045
DT_BILL(-1)	-1.315660	0.304721	-4.317587	0.0000
AR(1)	0.216253	0.041753	5.179365	0.0000
R-squared	0.102786	Mean dependent var	0.991923	
Adjusted R-squared	0.098030	S.D. dependent var	3.430919	
S.E. of regression	3.258415	Akaike info criterion	5.207352	
Sum squared resid	6009.375	Schwarz criterion	5.237848	
Log likelihood	-1480.095	F-statistic	21.61386	
Durbin-Watson stat	1.964422	Prob(F-statistic)	0.000000	
Inverted AR Roots	.22			

FIG. 14.10 – Estimation du modèle avec une composante autorégressive d'ordre 1

Ce modèle semble meilleur que le précédent (on peut comparer le AIC (Aikake info criterion) et le BIC (Schwarz criterion)), et le paramètre de la composante autorégressive est significatif. Pour voir s'il reste une partie des résidus à modéliser avec une composante autorégressive ou de moyennes mobiles il suffit d'examiner le SAC et le SPAC de ces résidus (figure 14.11).

Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.005	0.005	0.0168	
		2 -0.035	-0.035	0.7087	0.400
		3 0.027	0.027	1.1147	0.573
		4 0.009	0.008	1.1625	0.762
		5 0.071	0.073	4.0685	0.397
		6 -0.037	-0.038	4.8640	0.433
		7 -0.059	-0.054	6.8920	0.331
		8 0.064	0.058	9.2373	0.236
		9 0.007	0.003	9.2647	0.320
		10 -0.060	-0.059	11.383	0.250
		11 0.084	0.089	15.460	0.116
		12 -0.037	-0.038	16.247	0.132
		13 0.035	0.031	16.947	0.152

FIG. 14.11 – Le SAC et le SPAC des résidus du modèle avec une composante AR(1)

On voit qu'il ne reste plus rien à modéliser à ce niveau (aucune corrélation significative, et les p -values des Q-Stat (Box-Ljung) sont supérieures à 0,05). Mais il semble clair que la variance de la série n'est pas constante (visuellement). Peut-être faudrait-il modéliser cette variance. C'est dans le SAC et le SPAC des résidus au carré qu'on peut détecter l'effet d'hétérosécédasticité. Pour l'obtenir, il suffit d'aller dans **View, Residual Tests** puis **Correlogram Squared Residuals**. On obtient alors la figure 14.12.

Autocorrelation		Partial Correlation		AC	PAC	Q-Stat	Prob
				1	0.101	0.101	5.8199
				2	0.016	0.006	5.9746 0.015
				3	0.052	0.050	7.5166 0.023
				4	0.015	0.004	7.6404 0.054
				5	0.053	0.051	9.2627 0.055
				6	0.108	0.097	16.050 0.007
				7	0.031	0.010	16.616 0.011
				8	0.012	0.002	16.698 0.019
				9	0.014	0.002	16.810 0.032
				10	0.006	-0.001	16.830 0.051
				11	0.027	0.016	17.245 0.069
				12	0.034	0.018	17.929 0.083
				13	-0.018	-0.028	18.117 0.112
				14	-0.024	-0.025	18.466 0.141
				15	0.038	0.040	19.318 0.153

FIG. 14.12 – Le SAC et le SPAC des résidus **au carré**

Ah! ah! On avait bien raison de penser que les résidus n'ont pas une variance constante ! Il y a des autocorrélations significatives, et des p -values de la Q-stat en-dessous de 0,05 ! Il faut donc considérer un modèle GARCH. Il suffit pour ceci de retourner à notre équation, et de sélectionner ARCH dans la fenêtre Method. On conserve la même équation pour la modélisation usuelle, puis il faut décider de l'ordre du modèle GARCH pour estimer la variance. Ici ce n'est pas facile de cerner une décroissance exponentielle d'un côté ou de l'autre. Dans l'indécision, il est recommandé de commencer par un modèle ARCH d'ordre 1, c'est-à-dire un GARCH(0, 1) (voir la figure 14.13). Si ce n'est pas suffisant, le petit pic au lag 6 signifie sûrement qu'il faudra se rendre à l'ordre 6. La figure 14.14 présente les coefficients du modèle GARCH(0, 1).

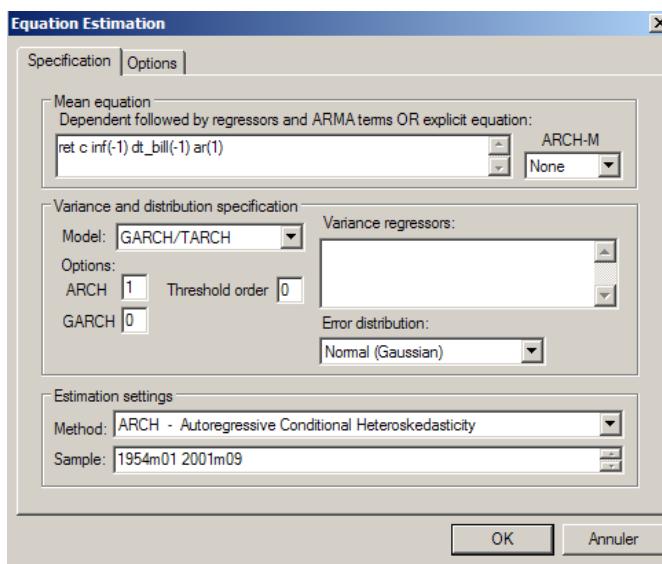


FIG. 14.13 – Pour obtenir le modèle GARCH(0, 1)

Sample (adjusted): 1954M03 2001M08				
Included observations: 570 after adjustments				
Convergence achieved after 19 iterations				
Variance backcast: ON				
GARCH = C(5) + C(6)*RESID(-1)^2				
	Coefficient	Std. Error	z-Statistic	Prob.
C	1.638655	0.254794	6.431304	0.0000
INF(-1)	-1.643869	0.448520	-3.665097	0.0002
DT_BILL(-1)	-1.387110	0.260063	-5.333739	0.0000
AR(1)	0.175633	0.050628	3.469106	0.0005
Variance Equation				
C	9.131751	0.599468	15.23310	0.0000
RESID(-1)^2	0.142484	0.049696	2.867133	0.0041
R-squared	0.100366	Mean dependent var	0.991923	
Adjusted R-squared	0.092390	S.D. dependent var	3.430919	
S.E. of regression	3.268587	Akaike info criterion	5.198290	
Sum squared resid	6025.585	Schwarz criterion	5.244033	
Log likelihood	-1475.513	F-statistic	12.58428	
Durbin-Watson stat	1.887668	Prob(F-statistic)	0.000000	
Inverted AR Roots	.18			

FIG. 14.14 – Coefficients du modèle GARCH(0, 1)

Le modèle ne semble pas s'être amélioré de beaucoup, mais tous les coefficients sont significatifs. Il faut examiner le SAC et le SPAC des résidus au carré pour pouvoir voir si la variance a été suffisamment modélisée (figure 14.15).

On remarque alors un problème au lag 6. En essayant un GARCH(0, 6), il est alors clair que la variance a été modélisée correctement (figures 14.16 et 14.17), mais ce modèle est plus complexe.

Sample: 1954M03 2001M08
 Included observations: 570
 Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 -0.009	-0.009	0.0424	
		2 0.014	0.014	0.1614	0.688
		3 0.055	0.055	1.8977	0.387
		4 0.013	0.014	1.9921	0.574
		5 0.052	0.051	3.5608	0.469
		6 0.116	0.115	11.405	0.044
		7 0.006	0.006	11.424	0.076
		8 0.024	0.016	11.751	0.109
		9 0.018	0.005	11.935	0.154
		10 0.011	0.005	12.008	0.213
		11 0.027	0.013	12.433	0.257
		12 0.034	0.019	13.112	0.286

FIG. 14.15 – Le SAC et le SPAC des résidus **au carré** du modèle GARCH(0, 1)

Convergence achieved after 298 iterations				
Variance backcast: ON				
GARCH = C(5) + C(6)*RESID(-1)^2 + C(7)*RESID(-2)^2 + C(8)*RESID(-3)^2 + C(9)*RESID(-4)^2 + C(10)*RESID(-5)^2 + C(11)*RESID(-6)^2				
	Coefficient	Std. Error	z-Statistic	Prob.
C	1.518617	0.232611	6.528564	0.0000
INF(-1)	-1.004706	0.432078	-2.325290	0.0201
DT_BILL(-1)	-0.978698	0.258220	-3.790168	0.0002
AR(1)	0.221846	0.044692	4.963893	0.0000
Variance Equation				
C	5.377712	0.970537	5.540965	0.0000
RESID(-1)^2	0.111404	0.044831	2.484974	0.0130
RESID(-2)^2	0.017167	0.038088	0.450718	0.6522
RESID(-3)^2	0.130440	0.050168	2.600085	0.0093
RESID(-4)^2	-0.003316	0.043910	-0.075521	0.9398
RESID(-5)^2	0.136328	0.056617	2.407885	0.0160
RESID(-6)^2	0.153470	0.052047	2.948662	0.0032
R-squared	0.097539	Mean dependent var	0.991923	
Adjusted R-squared	0.081395	S.D. dependent var	3.430919	
S.E. of regression	3.288326	Akaike info criterion	5.180712	
Sum squared resid	6044.516	Schwarz criterion	5.264575	
Log likelihood	-1465.503	F-statistic	6.041751	
Durbin-Watson stat	1.957924	Prob(F-statistic)	0.000000	
Inverted AR Roots	.22			

FIG. 14.16 – Le modèle GARCH(0, 6)

Q-statistic probabilities adjusted for 1 ARMA term(s)						
Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
1	1	1	0.010	0.010	0.0541	
2	1	2	0.007	0.007	0.0832	0.773
3	1	3	-0.018	-0.018	0.2672	0.875
4	1	4	-0.014	-0.014	0.3854	0.943
5	1	5	-0.019	-0.019	0.6032	0.963
6	1	6	-0.015	-0.015	0.7294	0.981
7	1	7	-0.001	-0.001	0.7303	0.994
8	1	8	0.017	0.017	0.9019	0.996
9	1	9	0.029	0.027	1.3799	0.995
10	1	10	-0.011	-0.013	1.4555	0.997
11	1	11	0.022	0.022	1.7331	0.998
12	1	12	0.016	0.017	1.8756	0.999
13	1	13	-0.045	-0.045	3.0621	0.995

FIG. 14.17 – Le SAC et le SPAC des résidus au carré du modèle GARCH(0, 6)

En fait, un modèle GARCH(1, 1) permet aussi de modéliser correctement la variance et est moins complexe, c'est donc sur ce modèle que s'arrêtera notre choix (figures 14.18 et 14.19).

Q-statistic probabilities adjusted for 1 ARMA term(s)						
Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
1	1	1	0.005	0.005	0.0165	
2	1	2	-0.040	-0.040	0.9537	0.329
3	1	3	-0.001	-0.000	0.9540	0.621
4	1	4	-0.042	-0.043	1.9473	0.583
5	1	5	0.011	0.012	2.0188	0.732
6	1	6	0.092	0.089	6.9384	0.225
7	1	7	-0.022	-0.022	7.2158	0.301
8	1	8	-0.029	-0.023	7.6958	0.360
9	1	9	0.003	0.003	7.7025	0.463
10	1	10	-0.024	-0.019	8.0475	0.529
11	1	11	0.028	0.025	8.5111	0.579
12	1	12	0.018	0.006	8.7030	0.649
13	1	13	-0.048	-0.042	10.061	0.611

FIG. 14.18 – Le SAC et le SPAC des résidus au carré du modèle GARCH(1, 1)

Variance backcast: ON GARCH = C(5) + C(6)*RESID(-1)^2 + C(7)*GARCH(-1)				
	Coefficient	Std. Error	z-Statistic	Prob.
C	1.604605	0.233151	6.882254	0.0000
INF(-1)	-1.569422	0.419001	-3.745631	0.0002
DT_BILL(-1)	-1.225877	0.270030	-4.539778	0.0000
AR(1)	0.185509	0.048568	3.819557	0.0001
Variance Equation				
C	1.134264	0.521984	2.172984	0.0298
RESID(-1)^2	0.108629	0.033660	3.227250	0.0012
GARCH(-1)	0.791771	0.067767	11.68372	0.0000
R-squared	0.101147	Mean dependent var	0.991923	
Adjusted R-squared	0.091567	S.D. dependent var	3.430919	
S.E. of regression	3.270068	Akaike info criterion	5.182334	
Sum squared resid	6020.354	Schwarz criterion	5.235702	
Log likelihood	-1469.965	F-statistic	10.55893	
Durbin-Watson stat	1.902835	Prob(F-statistic)	0.000000	
Inverted AR Roots	.19			

FIG. 14.19 – Le modèle GARCH(1, 1)

Malheureusement, les résidus de ce modèle ne sont pas normaux. Pour le vérifier, il suffit d'aller dans **View**, **Residual Tests** puis **Histogram - Normality Test**. On obtient alors la figure 14.20. La *p*-value nulle nous indique que la normalité est rejetée.

Il est aussi possible de faire des prévisions avec le menu **Forecast**. Étant donné qu'ici nous avons des variables indépendantes dont la valeur est inconnue pour les périodes à venir (plus de 1 mois), il faudrait estimer les valeurs futures de ces variables pour faire des prévisions à plus long terme.

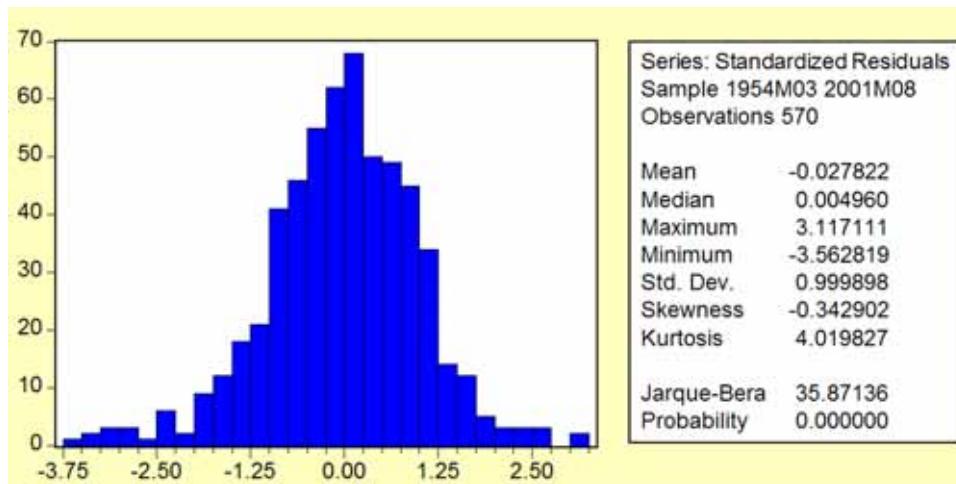


FIG. 14.20 – Pour tester la normalité des résidus

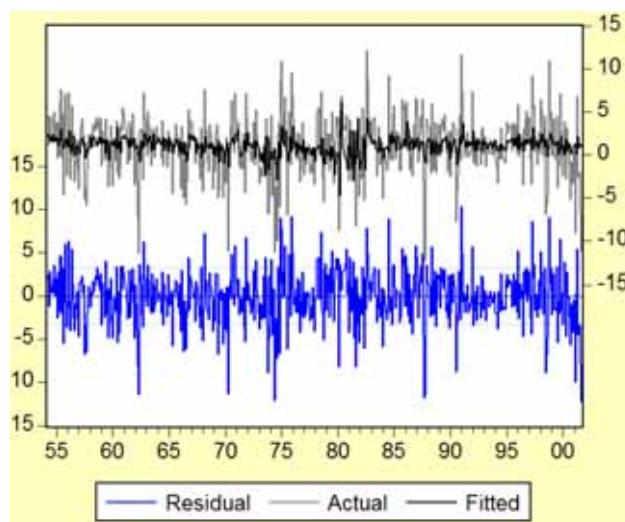


FIG. 14.21 – Les valeurs réelles, les estimations et les résidus

14.5 Autres modèles ARCH

Rappelons d'abord qu'avant de modéliser la variance conditionnelle, on développe d'abord un modèle pour les estimations usuelles, dont l'équation est appelée *Mean equation* (équation du niveau moyen). On note celle-ci de la façon suivante :

$$Y_t = X'_t \theta + \epsilon_t.$$

On a vu qu'on peut exprimer les résidus hétéroscédastiques de la façon suivante :

$$\epsilon_t = \nu_t \sqrt{h_y}$$

où $\nu_t \sim N(0, 1)$ est une variable aléatoire représentant le bruit de fond ou encore le bruit blanc qui évolue de façon indépendante de la fonction h_t , qui elle représente la composante hétéroscédastique conditionnelle de la variance du résidu.

On s'intéresse donc à modéliser h_t , et l'équation du modèle pour h_t est la *variance equation* (équation pour la variance). On a vu le modèle GARCH(p, q) , qui s'écrit de la façon suivante :

$$h_t = \alpha_0 + \sum_{j=1}^p \beta_j h_{t-j} + \sum_{i=1}^q \alpha_i \epsilon_{t-i}^2.$$

14.5.1 Modèle GARCH-M

Si on introduit la variance conditionnelle dans la *mean equation*, on obtient le modèle « GARCH-in-Mean » (GARCH-M) (Engle, Lilien et Robins, 1987) :

$$Y_t = X'_t \theta + \lambda h_t + \epsilon_t.$$

Les modèles ARCH-M sont utilisés dans les applications financières où le rendement espéré d'un actif est relié au risque espéré. Le coefficient associé au risque espéré est alors une mesure du risque par rapport au rendement en regard des échanges (*risk-return tradeoff*).

Il existe deux autres variantes du modèle GARCH-M, selon qu'on utilise l'écart-type conditionnel ou le logarithme de la variance conditionnelle au lieu de la variance conditionnelle :

$$Y_t = X'_t \theta + \lambda \sqrt{h_t} + \epsilon_t,$$

$$Y_t = X'_t \theta + \lambda \log(h_t) + \epsilon_t.$$

14.5.2 Variables indépendantes dans l'équation de la variance

Il est possible d'inclure des variables indépendantes dans l'équation de la variance :

$$h_t = \alpha_0 + \sum_{j=1}^p \beta_j h_{t-j} + \sum_{i=1}^q \alpha_i \epsilon_{t-i}^2 + Z'_t \pi$$

Il est à noter que les variances estimées par ce modèle ne seront pas nécessairement positives, tout dépend des valeurs que prennent les variables indépendantes. Il est préférable de transformer les variables indépendantes de façon à ce qu'elles ne prennent que des valeurs positives, ce qui réduit le risque de produire des estimations négatives.

14.5.3 Modèle TARCH

Les modèles Threshold ARCH (TARCH) et Threshold GARCH ont été introduits indépendamment par Zakoïan (1994) et Glosten, Jagannathan et Runkle (1993). L'équation de la variance de ce modèle est

$$h_t = \alpha_0 + \sum_{j=1}^p \beta_j h_{t-j} + \sum_{i=1}^q \alpha_i \epsilon_{t-i}^2 + \sum_{k=1}^r \gamma_k \epsilon_{t-k}^2 I_{t-k}^-$$

où $I_t^- = 1$ si $\epsilon_t < 0$ et 0 sinon, et $r = \text{Threshold order}$.

Dans ce modèle, les bonnes nouvelles ($\epsilon_{t-i} > 0$) et les mauvaises nouvelles ($\epsilon_{t-i} < 0$) ont des effets différents sur la variance conditionnelle. L'impact des bonnes nouvelles est de α_i , tandis que l'impact des mauvaises nouvelles est de $\alpha_i + \gamma_i$. Si $\gamma_i > 0$, les mauvaises nouvelles augmentent la volatilité, et on dit alors qu'il y a un effet de levier d'ordre i . Dès que $\gamma_i \neq 0$, on dit que l'impact des nouvelles est asymétrique.

14.5.4 Modèle EGARCH

Le modèle GARCH Exponentiel (EGARCH) a été proposé par Nelson (1991). L'équation de la variance est alors

$$\log(h_t) = \alpha_0 + \sum_{j=1}^p \beta_j \log(h_{t-j}) + \sum_{i=1}^q \alpha_i \left| \frac{\epsilon_{t-i}}{\sqrt{h_{t-i}}} \right| + \sum_{k=1}^r \gamma_k \frac{\epsilon_{t-k}}{\sqrt{h_{t-k}}}$$

où $r = \text{Asymmetric order}$.

Cette équation a comme variable dépendante le logarithme de la variance conditionnelle, ce qui fait que l'effet de levier est exponentiel au lieu d'être quadratique, et donc les estimations de la variance conditionnelle sont nécessairement positives. Il y a présence d'asymétrie si $\gamma_i \neq 0$.

Exemple 14.5.1 La base de données `series.wf1` contient des séries s'étendant de janvier 1999 à décembre 2003 (gracieusement fournies par Jean Desrochers). Nous essayons ici de modéliser la série ABX ; fixons le seuil à $\alpha = 0,05$. On voit le graphe séquentiel de la série ABX dans la figure 14.22. Elle semble présenter de l'hétéroscléasticité.

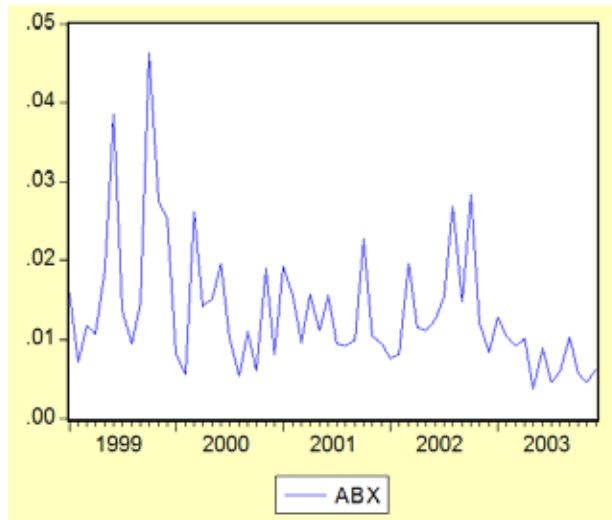


FIG. 14.22 – Série ABX

Commençons par développer un modèle ARIMA, et si l y a de l'hétéroscléasticité dans les résidus nous ferons un modèle GARCH. La figure 14.23 présente d'abord le SAC et le SPAC de la série. Il n'est pas facile ici d'identifier un modèle ; on décide de tenter un ARIMA(1,0,0) (en fait j'ai tenté plusieurs modèles, et celui-ci donne de bons résultats combiné avec un GARCH). La sortie de ce modèle est dans la deuxième sortie de la figure 14.23.

On voit que le terme autorégressif est significatif ($0,0366 < 0,05$). Est-ce que ce modèle est adéquat ? Examinons le SAC et le SPAC de ses résidus.

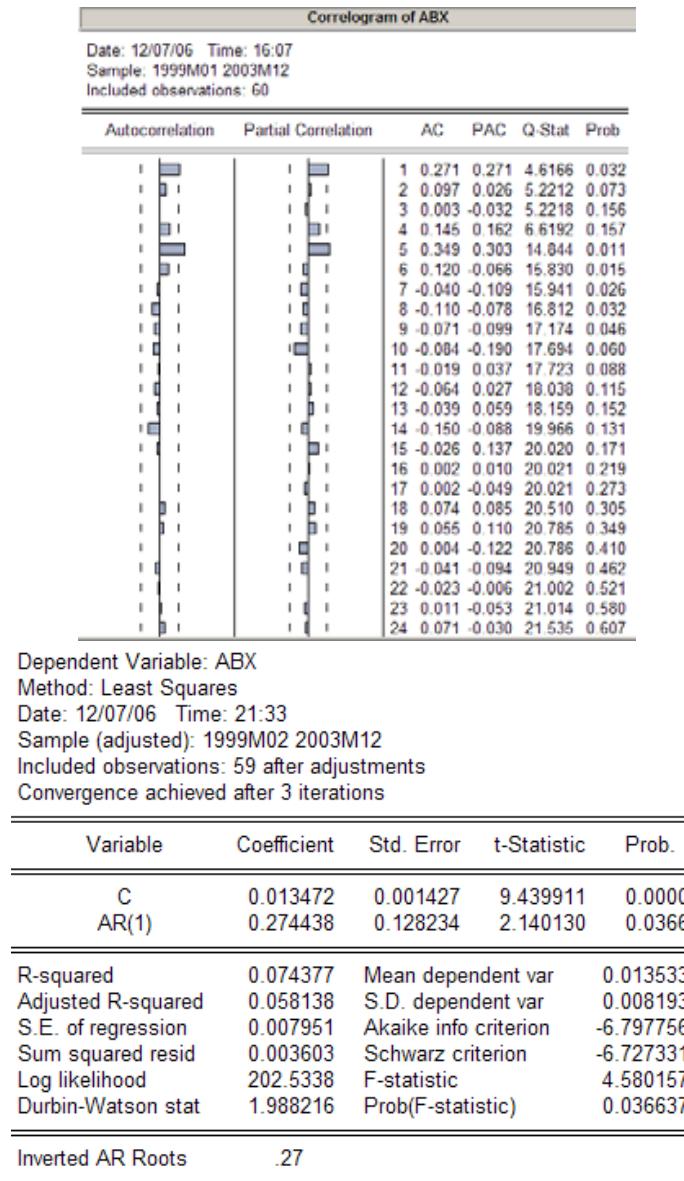


FIG. 14.23 – SAC et SPAC de la série ABX, puis sortie du modèle ARIMA(1,0,0)

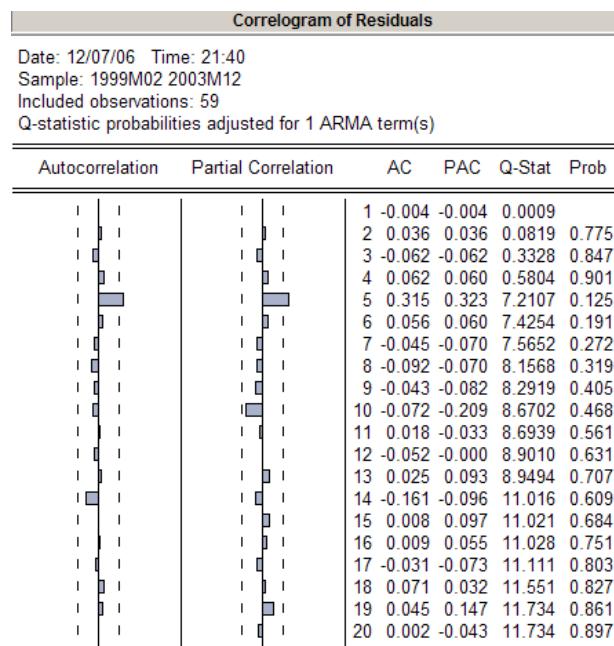


FIG. 14.24 – SAC et SPAC des résidus du modèle ARIMA(1,0,0)

Les autocorrélations au *lag* 5 semblent significatives, mais les *p*-values des Box-Ljung sont toutes supérieures à 0,05. Celle au *lag* 5 est un peu « achalante » (à 0,125), mais poursuivons tout de même avec l'examen du SAC et SPAC des résidus au carré (figure 14.25).

On voit au *lag* 4 que la *p*-value des Box-Ljung est de 0,044, ce qui est plus petit que 0,05. Les résidus semblent donc présenter de l'hétérosécédasticité. On peut le constater dans le graphe séquentiel des résidus (figure 14.26).

Correlogram of Residuals Squared						
	Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
1	-0.062	-0.062	0.2377			
2	-0.027	-0.031	0.2847	0.594		
3	-0.027	-0.031	0.3329	0.847		
4	0.345	0.342	8.1083	0.044		
5	0.112	0.172	8.9479	0.062		
6	-0.073	-0.039	9.3100	0.097		
7	-0.054	-0.056	9.5129	0.147		
8	-0.022	-0.176	9.5457	0.216		
9	0.066	-0.064	9.8623	0.275		
10	-0.000	0.028	9.8623	0.362		
11	-0.055	0.013	10.091	0.433		
12	0.005	0.102	10.093	0.522		
13	-0.004	0.021	10.094	0.608		
14	0.004	-0.034	10.095	0.686		
15	-0.011	-0.019	10.105	0.754		
16	-0.034	-0.083	10.204	0.807		
17	-0.016	-0.040	10.226	0.855		
18	-0.031	-0.022	10.309	0.890		
19	-0.005	0.001	10.311	0.921		
20	-0.045	0.002	10.496	0.940		
21	-0.017	0.007	10.523	0.958		

FIG. 14.25 – SAC et SPAC des résidus au carré du modèle ARIMA(1,0,0)

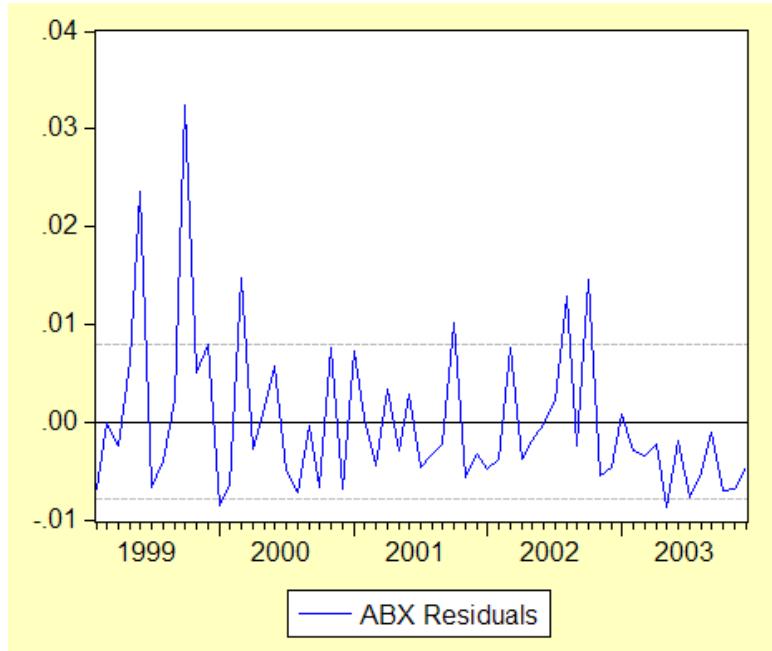


FIG. 14.26 – Graphe séquentiel des résidus du modèle ARIMA(1,0,0)

On décide donc de tenter un modèle GARCH. J'ai tenté les modèles GARCH(0,1) et GARCH(1,0), ils n'ont pas donné de bons résultats. J'ai donc tenté un GARCH(1,1) ; on en retrouve la sortie dans la figure 14.27.

Dependent Variable: ABX
 Method: ML - ARCH (Marquardt) - Normal distribution
 Date: 12/07/06 Time: 21:52
 Sample (adjusted): 1999M02 2003M12
 Included observations: 59 after adjustments
 Convergence achieved after 16 iterations
 Variance backcast: ON
 GARCH = C(3) + C(4)*RESID(-1)^2 + C(5)*GARCH(-1)

	Coefficient	Std. Error	z-Statistic	Prob.
C	0.011477	0.001086	10.56524	0.0000
AR(1)	0.250233	0.119897	2.087064	0.0369
Variance Equation				
C	-9.77E-07	2.77E-07	-3.521341	0.0004
RESID(-1)^2	-0.045516	0.002085	-21.82901	0.0000
GARCH(-1)	1.067101	0.000608	1754.251	0.0000
R-squared	0.039628	Mean dependent var	0.013533	
Adjusted R-squared	-0.031511	S.D. dependent var	0.008193	
S.E. of regression	0.008321	Akaike info criterion	-6.974396	
Sum squared resid	0.003739	Schwarz criterion	-6.798333	
Log likelihood	210.7447	F-statistic	0.557050	
Durbin-Watson stat	1.869191	Prob(F-statistic)	0.694750	

FIG. 14.27 – Sortie du modèle ARIMA(1,0,0) - GARCH(1,1)

On voit d'abord que tous les paramètres sont jugés significatifs : le terme AR(1) a une p -value de $0,0369 < 0,05$, et les termes e_{t-1}^2 et h_{t-1} ont des p -values nulles, ce qui est bien sûr bon signe. De plus, les mesures d'adéquation se sont améliorées par rapport au modèle sans GARCH : le Log likelihood est passé de 202,534 à 210,745, le AIC est passé de -6,798 à -6,974, et le SC (BIC) de -6,727 à -6,798. Allons maintenant examiner le SAC et le SPAC des résidus (figure 14.28).

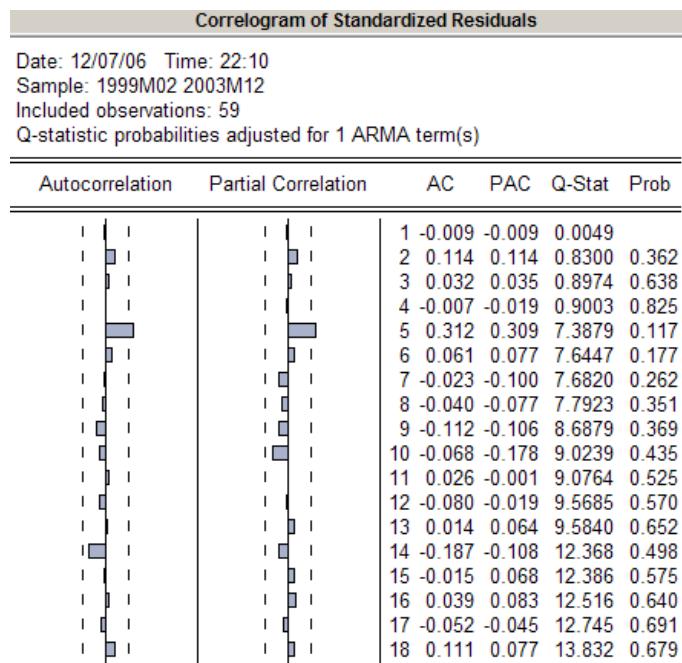


FIG. 14.28 – SAC et SPAC des résidus du modèle ARIMA(1,0,0) - GARCH(1,1)

On voit que tout est « presque » beau ; toutes les *p*-values des Box-Ljung sont supérieures à 0,05, mais la *p*-value du *lag* 5 est encore « achalante » avec une valeur de 0,117, mais il n'y a pas de quoi s'arracher les cheveux. On poursuit donc avec l'examen du SAC et du SPAC des résidus au carré (figure 14.29).

On constate que toutes les *p*-values des Box-Ljung sont supérieures à 0,05 (la plus petite dans cette sortie est de 0,273, ce n'est vraiment pas inquiétant). On semble donc avoir modélisé l'hétérosécédasticité. Par contre on n'a pas la normalité des résidus puisque la *p*-value du test de Jarque-Bera est de 0,001903 < 0,05 (figure 14.30).

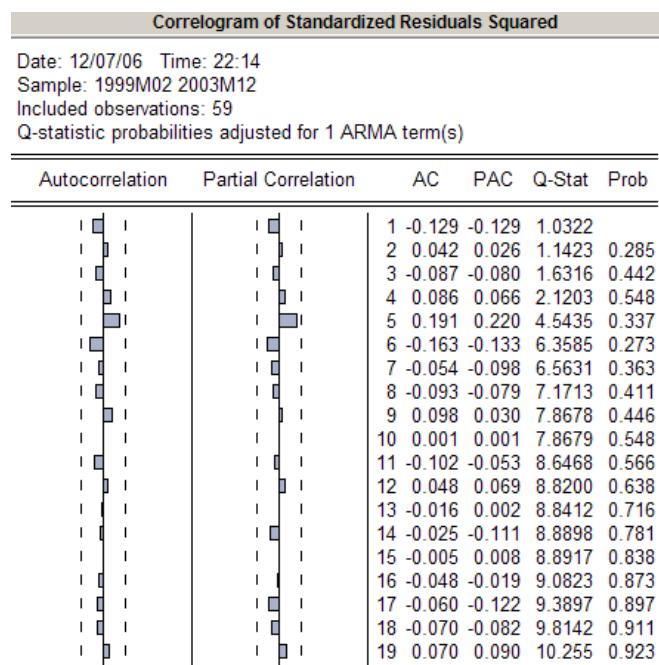


FIG. 14.29 – SAC et SPAC des résidus au carré du modèle ARIMA(1,0,0) - GARCH(1,1)

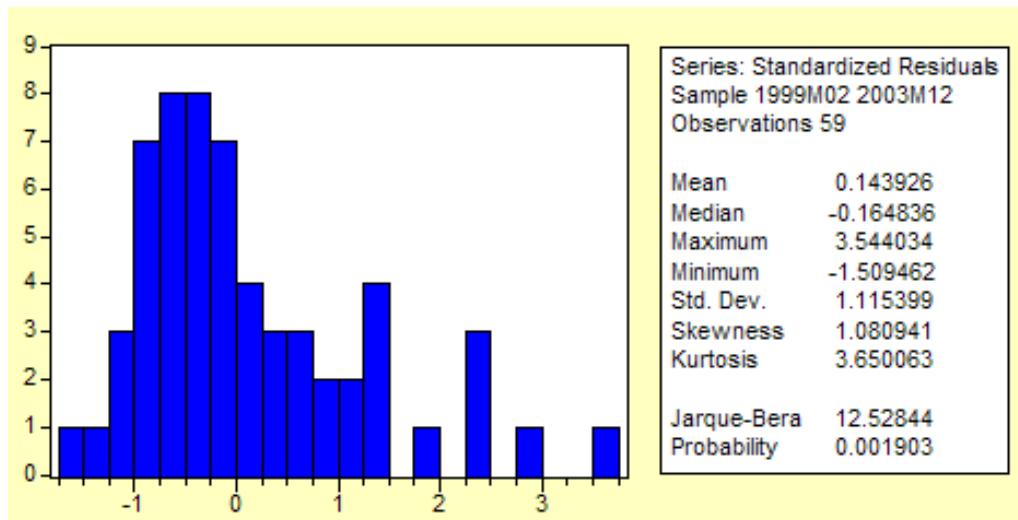


FIG. 14.30 – Test de normalité des résidus

Tentons maintenant de voir s'il est possible d'améliorer le modèle avec un autre GARCH. Je teste d'abord s'il y a un effet asymétrique d'ordre 1 au niveau de la variance ; on obtient alors la sortie de la figure 14.31. Non seulement le terme d'asymétrie n'est pas significatif (sa *p*-value est de $0,5469 > 0,05$), mais les mesures d'adéquation révèle que ce modèle est moins bon que le précédent. On conclut donc qu'il n'y a pas d'effet asymétrique.

Dependent Variable: ABX				
Method: ML - ARCH (Marquardt) - Normal distribution				
Date: 12/07/06 Time: 22:25				
Sample (adjusted): 1999M02 2003M12				
Included observations: 59 after adjustments				
Convergence achieved after 17 iterations				
Variance backcast: ON				
GARCH = C(3) + C(4)*RESID(-1)^2 + C(5)*RESID(-1)^2*(RESID(-1)<0)				
+ C(6)*GARCH(-1)				
	Coefficient	Std. Error	z-Statistic	Prob.
C	0.011578	0.001467	7.891576	0.0000
AR(1)	0.272404	0.124814	2.182476	0.0291
Variance Equation				
C	1.50E-06	2.35E-06	0.639022	0.5228
RESID(-1)^2	-0.021336	0.023175	-0.920652	0.3572
RESID(-1)^2*(RESID(-1)<0)	-0.217916	0.361742	-0.602408	0.5469
GARCH(-1)	1.013024	0.039520	25.63311	0.0000
R-squared	0.045555	Mean dependent var	0.013533	
Adjusted R-squared	-0.044487	S.D. dependent var	0.008193	
S.E. of regression	0.008373	Akaike info criterion	-6.875281	
Sum squared resid	0.003716	Schwarz criterion	-6.664006	
Log likelihood	208.8208	F-statistic	0.505929	
Durbin-Watson stat	1.924126	Prob(F-statistic)	0.770452	

FIG. 14.31 – Test de l'effet d'asymétrie (Treshold d'ordre 1)

On peut aussi tenter un GARCH-M ; ici j'inclus l'écart-type conditionnel dans l'équation du ARIMA. On obtient alors la sortie de la figure 14.32. Malheureusement, l'ajout de cet écart-type n'améliore pas notre modèle ARIMA(1,0,0) - GARCH(1,1), comme le montrent les mesures d'adéquation. J'ai aussi tenté un EGARCH, ce n'est pas concluant non plus.

Dependent Variable: ABX				
Method: ML - ARCH (Marquardt) - Normal distribution				
Date: 12/07/06 Time: 22:28				
Sample (adjusted): 1999M02 2003M12				
Included observations: 59 after adjustments				
Failure to improve Likelihood after 52 iterations				
Variance backcast: ON				
GARCH = C(4) + C(5)*RESID(-1)^2 + C(6)*GARCH(-1)				
	Coefficient	Std. Error	z-Statistic	Prob.
@SQRT(GARCH)	-1.966441	1.299218	-1.513557	0.1301
C	0.028651	0.010397	2.755673	0.0059
AR(1)	0.067782	0.264454	0.256308	0.7977
	Variance Equation			
C	4.38E-05	4.22E-05	1.037327	0.2996
RESID(-1)^2	-0.057377	0.054618	-1.050503	0.2935
GARCH(-1)	0.334492	0.730756	0.457734	0.6471
R-squared	0.061654	Mean dependent var	0.013533	
Adjusted R-squared	-0.026869	S.D. dependent var	0.008193	
S.E. of regression	0.008302	Akaike info criterion	-6.711179	
Sum squared resid	0.003653	Schwarz criterion	-6.499904	
Log likelihood	203.9798	F-statistic	0.696475	
Durbin-Watson stat	1.940818	Prob(F-statistic)	0.628436	

FIG. 14.32 – GARCH-M avec l'écart-type conditionnel

Examinons maintenant les estimations produites par le modèle ARIMA(1,0,0) - GARCH(1,1). La figure 14.33 montre les estimations et les résidus de ce modèle. On voit que les résultats ne sont quand même pas extraordinaires. Les estimations présentées sont les estimations **statiques** : à chaque étape, on calcule la prédition de la prochaine période avec les observations réelles. Il est aussi possible d'obtenir les estimation **dynamiques** : cette méthode fait ses prédictions à partir des prédictions précédentes, en partant de la première observation, comme si on avait voulu prédire dès le début de la série avec ce modèle. On peut voir les estimations dynamiques avec les intervalles de confiance dans la figure 14.34.

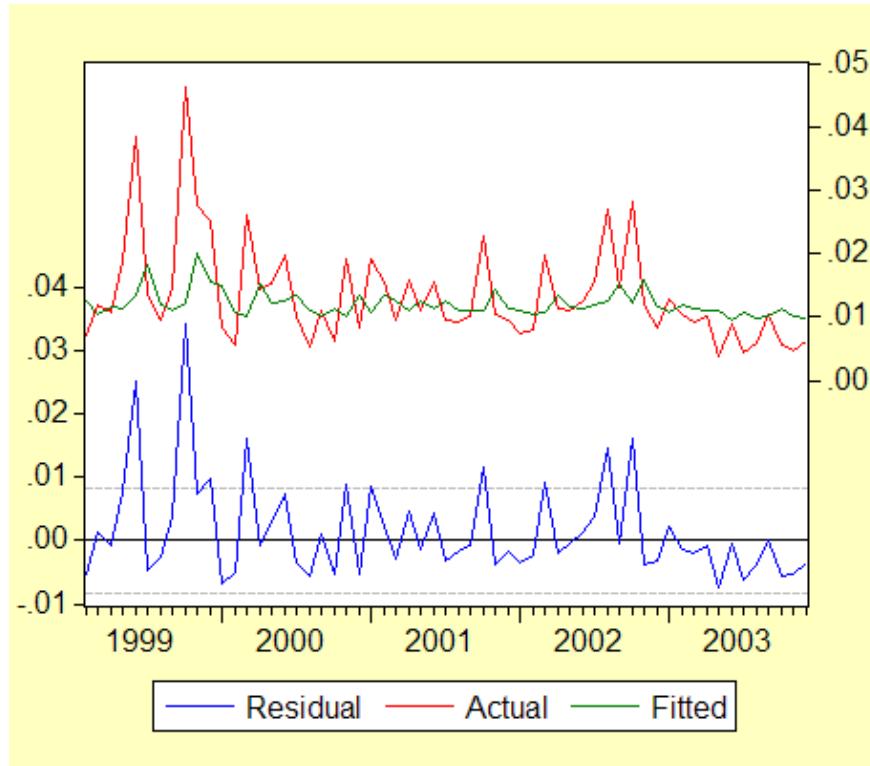


FIG. 14.33 – Estimations et résidus du modèle ARIMA(1,0,0) - GARCH(1,1)

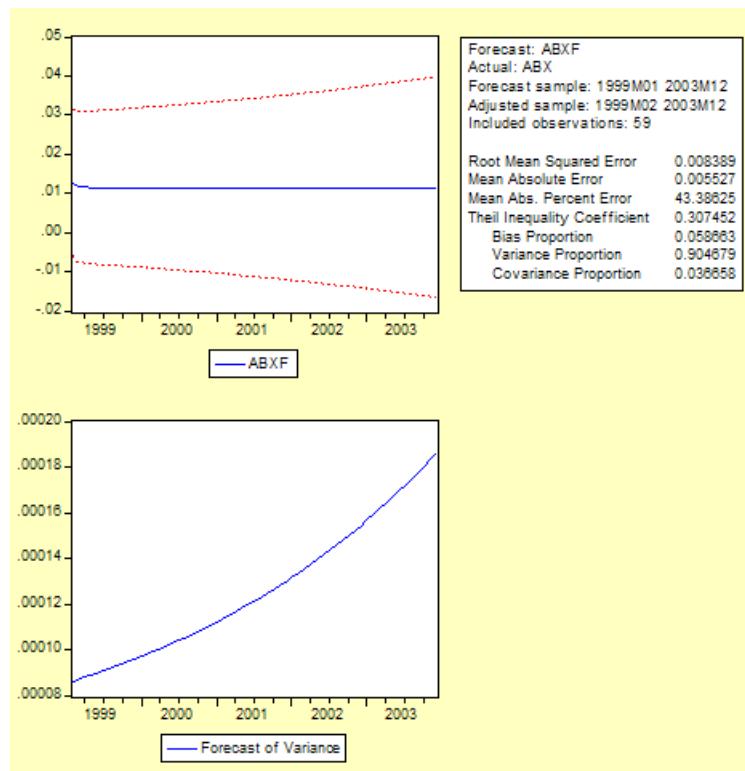


FIG. 14.34 – Estimations dynamiques

Finalement, la figure 14.35 présente les estimations statiques des modèles ARIMA(1,0,0) et ARIMA(1,0,0) - GARCH(1,1). Le GARCH semble avoir induit un certain décalage, mais n'améliore pas de façon flagrante les estimations.

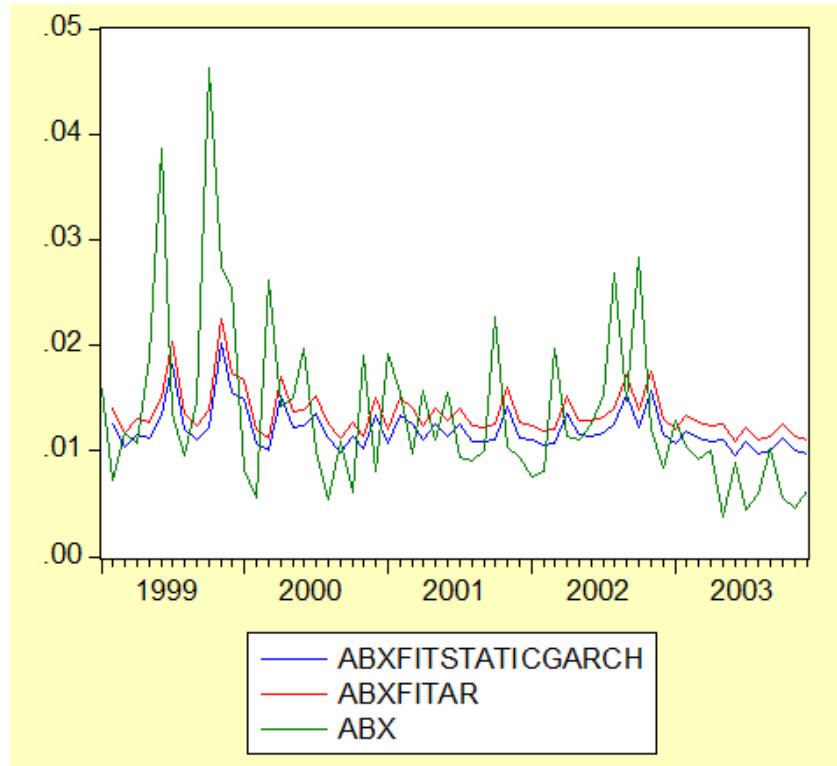


FIG. 14.35 – Estimations statiques des modèles ARIMA(1,0,0) et ARIMA(1,0,0) - GARCH(1,1)

J'ai essayé quelques autres modèles. Un AR(1) - MA(5) - GARCH(2,2) donne des estimations qui me semblent quelque peu améliorées, et d'ailleurs le AIC et le Log likelihood sont meilleurs.

Dependent Variable: ABX				
Method: ML - ARCH (Marquardt) - Normal distribution				
Date: 12/07/06 Time: 23:02				
Sample (adjusted): 1999M02 2003M12				
Included observations: 59 after adjustments				
Convergence achieved after 10 iterations				
MA backcast: 1998M09 1999M01, Variance backcast: ON				
GARCH = C(4) + C(5)*RESID(-1)^2 + C(6)*RESID(-2)^2 + C(7)				
*GARCH(-1) + C(8)*GARCH(-2)				
	Coefficient	Std. Error	z-Statistic	Prob.
C	0.013054	0.002013	6.485619	0.0000
AR(1)	0.145630	0.143813	1.012633	0.3112
MA(5)	0.485435	0.167357	2.900602	0.0037
Variance Equation				
C	2.99E-06	3.00E-06	0.997405	0.3186
RESID(-1)^2	-0.054288	0.015862	-3.422491	0.0006
RESID(-2)^2	0.141469	0.066887	2.115037	0.0344
GARCH(-1)	0.563499	0.904927	0.622701	0.5335
GARCH(-2)	0.238284	0.873162	0.272898	0.7849
R-squared	0.266928	Mean dependent var	0.013533	
Adjusted R-squared	0.166310	S.D. dependent var	0.008193	
S.E. of regression	0.007480	Akaike info criterion	-7.006739	
Sum squared resid	0.002854	Schwarz criterion	-6.725039	
Log likelihood	214.6988	F-statistic	2.652888	
Durbin-Watson stat	1.816697	Prob(F-statistic)	0.020304	

FIG. 14.36 – Sortie du modèle AR(1) - MA(5) - GARCH(2,2)

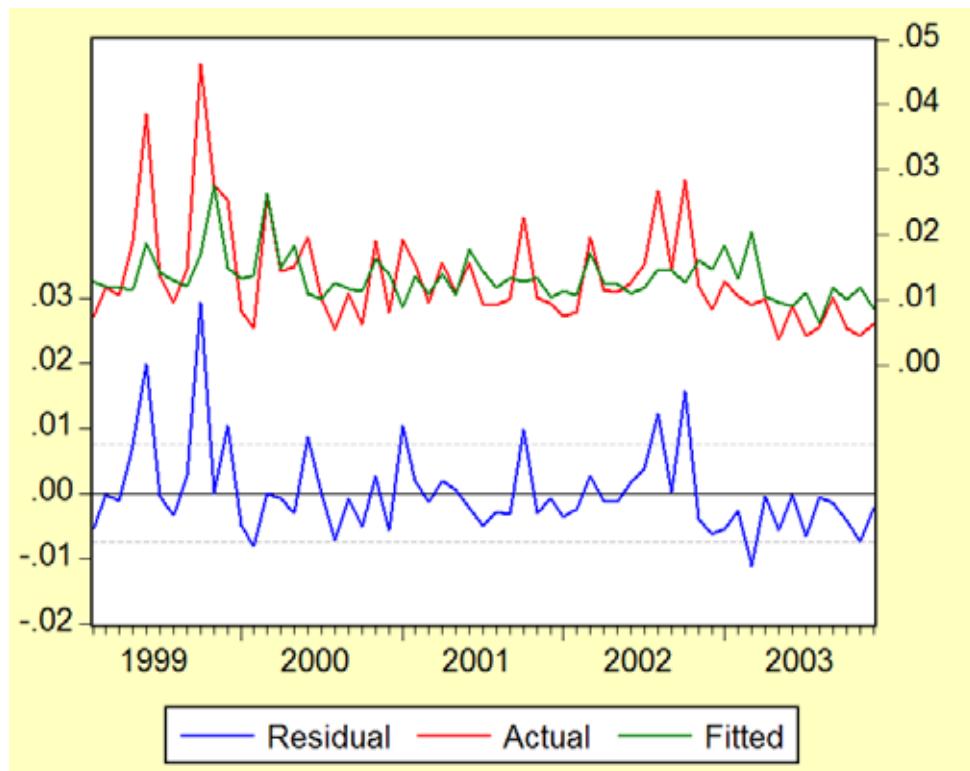


FIG. 14.37 – Estimations et résidus du modèle AR(1) - MA(5) - GARCH(2,2)

14.6 Le test de White

Le test de White permet de détecter l'hétéroscédasticité d'un modèle qui utilise des variables indépendantes. Supposons par exemple qu'on a le modèle de régression suivant :

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \epsilon_t.$$

Une fois le modèle estimé, on développe un autre modèle (une régression *auxiliaire*) ayant comme variable dépendante les résidus au carré :

$$\epsilon_t^2 = \beta'_0 + \beta'_1 X_{1t} + \beta'_2 X_{2t} + \beta_3 X_{1t}^2 + \beta_4 X_{2t}^2 + \beta_5 X_{1t} X_{2t} + \nu_t.$$

On voit que les variables indépendantes de cette régression auxiliaire sont les variables indépendantes d'origine, ces mêmes variables au carré et le terme multiplicatif des variables indépendantes d'origine (s'il y avait eu 3 variables indépendantes, il y aurait eu 3 termes multiplicatifs).

On obtient alors le r^2 de cette régression auxiliaire. Sous l'hypothèse nulle qu'il n'existe pas d'hétéroscédasticité, on peut montrer que la taille de l'échantillon multipliée par le r^2 suit une loi du khi-deux. Lorsque le khi-deux observé devient trop grand (et alors la p -value du test passe sous le seuil de signification), on conclut qu'il y a présence d'hétéroscédasticité. Sinon on conclut qu'il n'y a pas d'hétéroscédasticité, ce qui revient à dire que les paramètres de la régression auxiliaires sont considérés nuls ($\beta'_1 = \beta'_2 = \beta_3 = \beta_4 = \beta_5 = 0$).

Si on reprend le premier exemple de ce chapitre, on se rappelle qu'on avait développé un modèle avec deux variables indépendantes déphasées et un terme autorégressif (voir figure 14.38). On sait aussi qu'on avait détecté de l'hétéroscédasticité dans nos résidus. Regardons ce que nous dit le test de White.

Pour accéder au test de White, il faut à partir de la fenêtre de la figure 14.38 aller dans `View → Residual Tests → White Heteroskedasticity (cross terms)`. On obtient alors la sortie de la figure 14.39.

Dependent Variable: RET				
Method: Least Squares				
Date: 12/08/05 Time: 21:14				
Sample (adjusted): 1954M03 2001M08				
Included observations: 570 after adjustments				
Convergence achieved after 4 iterations				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.463743	0.240354	6.089943	0.0000
INF(-1)	-1.432679	0.501672	-2.855808	0.0045
DT_BILL(-1)	-1.315660	0.304721	-4.317587	0.0000
AR(1)	0.216253	0.041753	5.179365	0.0000
R-squared	0.102786	Mean dependent var	0.991923	
Adjusted R-squared	0.098030	S.D. dependent var	3.430919	
S.E. of regression	3.258415	Akaike info criterion	5.207352	
Sum squared resid	6009.375	Schwarz criterion	5.237848	
Log likelihood	-1480.095	F-statistic	21.61386	
Durbin-Watson stat	1.964422	Prob(F-statistic)	0.000000	
Inverted AR Roots	.22			

FIG. 14.38 – Modèle utilisé pour modéliser la variable `ret`

White Heteroskedasticity Test:				
F-statistic	3.044935	Prob. F(5,564)	0.010098	
Obs*R-squared	14.98221	Prob. Chi-Square(5)	0.010439	
Test Equation:				
Dependent Variable: RESID^2				
Method: Least Squares				
Sample: 1954M03 2001M08				
Included observations: 570				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.288244	1.302206	6.364773	0.0000
DT_BILL(-1)	-8.569502	3.514336	-2.438442	0.0151
DT_BILL(-1)^2	-0.413143	0.816690	-0.505875	0.6131
DT_BILL(-1)*INF(-1)	8.175812	4.533262	1.803516	0.0718
INF(-1)	5.008716	5.435252	0.921524	0.3572
INF(-1)^2	3.122185	5.254563	0.594186	0.5526
R-squared	0.026285	Mean dependent var	10.54276	
Adjusted R-squared	0.017652	S.D. dependent var	19.66326	
S.E. of regression	19.48894	Akaike info criterion	8.788042	
Sum squared resid	214217.7	Schwarz criterion	8.833785	
Log likelihood	-2498.592	F-statistic	3.044935	
Durbin-Watson stat	1.755509	Prob(F-statistic)	0.010098	

FIG. 14.39 – Test de White

Cette sortie nous décrit la régression auxiliaire. La p -value associée au khi-deux étant de 0,010439 (deuxième ligne en haut de la sortie, à droite), ce qui est sous le seuil de 5 %, on rejette au risque de se tromper 1 fois sur 20 l'hypothèse stipulant qu'il n'y a pas d'hétéroscédasticité, ce qui est cohérent avec ce qu'on avait vu auparavant. D'ailleurs on voit que dans la régression auxiliaire il y a une des variables qui est significative au seuil de 5 % (`dt_bill(-1)`), ce qui nous indique que les erreurs au carré sont reliées à cette variable, d'où la conclusion que la variance des erreurs n'est certainement pas constante dans le temps.

Chapitre 15

Analyse factorielle (analyse en composantes principales)

Dans le cadre d'une étude, plusieurs questions sont posées aux individus sélectionnés, ce qui crée les variables statistiques que nous pouvons analyser.

Les méthodes factorielles cherchent à réduire le nombre de variables à l'étude en les résumant en un petit nombre de composantes. Selon qu'on travaille avec un tableau de variables numériques ou qualitatives, on utilisera l'analyse en composantes principales ou l'analyse des correspondances. Dans le cadre de ce cours, nous supposerons que les échelles utilisées peuvent être interprétées de façon continue et seule l'analyse en composantes principales sera présentée.

15.1 Généralités

Qu'est ce que l'amour ? Contrairement à l'âge, le poids ou la température, l'amour ne se mesure pas précisément sur une échelle. L'amour est un concept. On pourrait être amené à conclure qu'il y a présence d'amour entre deux personnes lorsqu'une réponse du type « totalement en accord » est donnée à tous les points tels :

- Il (elle) m'envoie des fleurs ;
- Il (elle) écoute mes problèmes ;
- Nous ne pouvons plus vivre sans l'autre ;
- Il (elle) m'embrasse tendrement ;
- Nous ne formons qu'un seul être.

L'amour n'est pas directement observable et n'est pas le fruit d'une seule variable. L'amour est un tout, un construit dont les dérivés peuvent être mesurés par le biais de variables observables. En fait, tous conviendront que la présence de quelque chose appelé l'amour entre deux personnes pourrait bien expliquer un ensemble de corrélations entre plusieurs variables.

Contrairement à l'analyse en régression, une bonne analyse en composantes principales repose sur le fait qu'il existe beaucoup de corrélation entre les variables à l'étude. Pas de corrélation, pas d'analyse factorielle (pas moyen de faire de résumé!).

L'analyse en composantes principales est une méthode descriptive qui permet de résumer un ensemble de variables en composantes synthétiques, appelées composantes principales. Plus précisément, ce type d'analyse tente d'expliquer les corrélations entre les variables.

L'identification des composantes principales (aussi appelées **facteurs**) peut grandement simplifier l'interprétation de phénomènes complexes. L'hypothèse de base de ce type d'analyse réside dans le fait qu'il existe toujours des dimensions cachées, présentement inconnues de l'expérimentateur, qui expliquent les corrélations entre deux ou plusieurs variables. Une bonne ACP dégage justement ces dimensions souvent inconnues.

Dans le cadre de ce chapitre, les principales étapes d'une analyse en composantes principales seront illustrées à l'aide d'un exemple. On désire prévoir les ventes d'automobiles à l'aide d'un ensemble de variables. Or, ces variables semblent être corrélées entre elles, et il semble adéquat de vouloir regrouper celles qui partagent de l'information. Ceci permettra de bien résumer quelles sont les facettes qui influencent les ventes.

La base de données se nomme `ventesautos.sav`. Les variables à l'étude sont les suivantes :

	Name	Type	Width	Decimals	Label
1	manufact	String	13	0	Manufacturier
2	modele	String	17	0	Modèle
3	ventes	Numeric	11	3	Nombre de ventes en milliers
4	revente	Numeric	11	3	Valeur de revente après 4 ans
5	type	Numeric	11	0	Type de véhicule
6	prix	Numeric	11	3	Prix en milliers de dollars
7	taillemot	Numeric	11	1	Taille du moteur
8	chevaux	Numeric	11	0	Puissance du moteur (en chevaux)
9	empatt	Numeric	11	1	Empattement
10	largeur	Numeric	11	1	Largeur
11	longueur	Numeric	11	1	Longueur
12	poids	Numeric	11	3	Poids à vide
13	gazcap	Numeric	11	1	Capacité du réservoir à essence
14	consomm	Numeric	11	0	Consommation d'essence (mpg)

FIG. 15.1 – Les variables

Ici nous allons tenter de résumer les variables `revente`, `prix`, `taillemot`, `chevaux`, `empatt`, `largeur`, `longueur`, `poids`, `gazcap`, `consomm`.

15.2 Les étapes d'une ACP

L'objectif d'une bonne ACP est de résumer avec le moins de facteurs synthétiques possible les interrelations entre les variables. Une bonne ACP est à la fois simple et surtout interprétable.

Pour obtenir les sorties SPSS des pages suivantes, il faut effectuer les commandes suivantes :

Menu SPSS :	→ Analyse
	→ Data Reduction
	→ Factor...
Dans la fenêtre Variable(s) :	→ revente, prix, taillemot, chevaux, empatt, largeur, longueur, poids, gazcap, consomm
Dans le bouton Descriptives... :	✓ KMO and Bartlett's test of sphericity
Dans le bouton Extraction... :	✓ Scree plot
Dans le bouton Rotation... :	✓ Varimax
Dans le bouton Options... :	Coefficient Display Formay ✓ Sorted by size

15.2.1 La mesure de Kaiser-Meyer-Olkin (KMO)

La première étape consiste à vérifier si les variables que l'on considère sont assez corrélées pour que l'ACP donne de bons résultats. La mesure de Kaiser-Meyer-Olkin (KMO) donne une bonne idée de l'efficacité éventuelle d'une ACP. Cette mesure varie entre 0 et 1. Plus une mesure KMO s'approche de 1, meilleure sera l'ACP. L'interprétation est la suivante :

0,9 < KMO ≤ 1,0	L'ACP sera incroyable
0,8 < KMO ≤ 0,9	L'ACP sera méritoire
0,7 < KMO ≤ 0,8	L'ACP sera moyenne
0,6 < KMO ≤ 0,7	L'ACP sera médiocre
0,5 < KMO ≤ 0,6	L'ACP sera misérable
0,0 < KMO ≤ 0,5	L'ACP sera inacceptable

Dans le cadre de l'exemple on obtient le KMO dans la sortie 15.2. La valeur de celui-ci est de 0,819, et donc l'ACP de cet exemple sera méritoire, tout va bien.

KMO and Bartlett's Test			
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.			,819
Bartlett's Test of Sphericity	df	Approx. Chi-Square	1466,298
	Sig.		,000

FIG. 15.2 – Le KMO

Dans le cas où le KMO n'est pas satisfaisant, une solution consiste à examiner la matrice des corrélations et à enlever de l'analyse les variables qui sont peu corrélées aux autres.

15.2.2 L'extraction de facteurs

L'extraction s'effectue de façon à ce que la première composante principale explique la plus grande part de la variation totale. La seconde composante à être extraite est orthogonale à la première (indépendante) et explique la seconde plus grande part de la variation. Cette deuxième part de la variation est plus petite que la première et sera plus grande que la troisième, et ainsi de suite.

La figure 15.3 présente les *communalities* de chaque variable. Celles-ci sont les % de variation de chaque variable expliquée par les facteurs. La colonne **Initial** donne toujours des *communalities* de 1 car au départ (avant l'extraction) chaque variable est elle-même un facteur. La colonne **Extraction** donne les *communalities* une fois que les facteurs sont extraits. Ainsi, par exemple, la variation de la variable **revente** est expliquée à 90,6 % par les facteurs de cette ACP. Étant donné que toutes les *communalities* sont assez élevées, on voit que l'ACP sera bonne.

C'est à partir du tableau de la figure 15.4 que l'on aura une première idée du nombre

Communalities		
	Initial	Extraction
revente	1,000	,906
prix	1,000	,931
taillermot	1,000	,805
chevaux	1,000	,880
empatt	1,000	,837
largeur	1,000	,758
longueur	1,000	,783
poids	1,000	,870
gazcap	1,000	,749
consomm	1,000	,705

Extraction Method: Principal Component Analysis

FIG. 15.3 – Les communalités

de facteurs qui seront extraits. Dans la partie **Initial Eigenvalues**, on a les informations suivantes : dans la colonne **Total**, on retrouve la portion de la variance totale des variables expliquée par le facteur correspondant à la ligne (ce sont les **valeurs propres**). Par exemple, le premier facteur explique 6,024 de la variance totale. Dans la seconde colonne, on retrouve le % de la variance totale expliquée par le facteur. Ainsi le premier facteur explique à 60,239 % la variance totale. Dans la colonne **Cumulative %**, on retrouve simplement les % cumulés. Ainsi les deux premiers facteurs expliquent 82,247 % de la variance totale.

Compo nent	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6,024	60,239	60,239	6,024	60,239	60,239	4,572	45,720	45,720
2	2,201	22,008	82,247	2,201	22,008	82,247	3,653	36,527	82,247
3	,645	6,450	88,697						
4	,429	4,292	92,989						
5	,255	2,552	95,542						
6	,144	1,441	96,982						
7	,131	1,308	98,290						
8	,094	,939	99,230						
9	,055	,547	99,777						
10	,022	,223	100,000						

Extraction Method: Principal Component Analysis.

FIG. 15.4 – Les valeurs propres

Habituellement, ce sont les valeurs propres qui nous indiquent combien de facteurs nous devons garder : on garde ceux dont la valeur propre est supérieure à 1. Donc dans le cadre de l'exemple nous garderons deux facteurs pour expliquer l'ensemble des variables, ce qui se reflète dans la partie **Extraction Sums of Squared Loadings** du tableau 15.4.

La dernière partie du tableau 15.4 (**Rotation Sums of Squared Loadings**) redonne les valeurs expliquées auparavant après une **rotation**. La rotation est une opération qui consiste à mieux répartir la variance parmi les facteurs, et on l'effectue toujours (la rotation est effectuée lorsque l'option **Varimax** est cochée dans le bouton **Rotation...**).

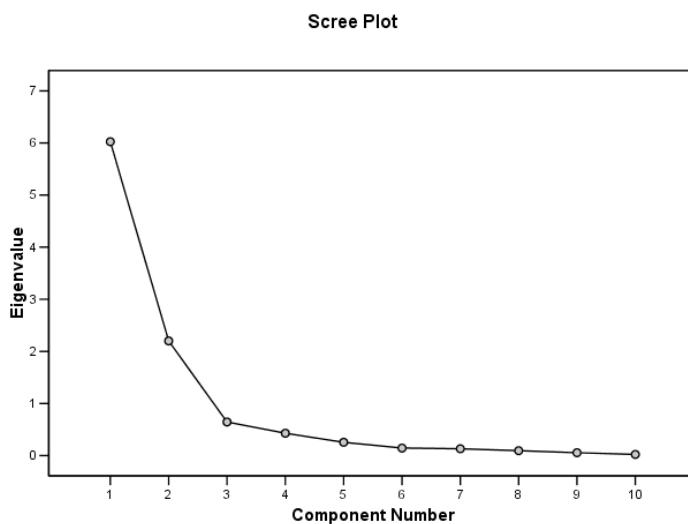


FIG. 15.5 – Le Scree plot

Le graphe de la figure 15.5 permet simplement de visualiser l'information contenue dans le tableau 15.4. Il permet de voir quelles sont les valeurs propres qui se détachent des autres.

15.2.3 Interprétation des facteurs

L'étape la plus importante de l'analyse est l'interprétation des facteurs : on voit à cette étape le lien entre chacune des variables et les facteurs, et on tente ensuite de bien

interpréter chacun des facteurs.

Cette interprétation se fait à l'aide de la matrice des facteurs obtenue après la rotation (**Rotated Component Matrix**). Celle de l'exemple est donnée dans la figure 15.6. Cette matrice donne la corrélations entre les variables et les facteurs.

	Component	
	1	2
Empattement	,909	-,104
Longueur	,884	,025
Largeur	,842	,221
Poids à vide	,829	,427
Capacité du réservoir à essence	,793	,348
Consommation d'essence (mpg)	-,676	-,498
Prix en milliers de dollars	,115	,958
Valeur de revente après 4 ans	-,035	,951
Puissance du moteur (en chevaux)	,343	,873
Taille du moteur	,590	,676

Extraction Method: Principal Component Analysis.
 Rotation Method: Varimax with Kaiser Normalization.
 a. Rotation converged in 3 iterations.

FIG. 15.6 – La matrice des facteurs

Dans une bonne ACP, la matrice des facteurs est telle que

- Chaque facteur est fortement corrélé avec un nombre restreint de variables et peu corrélé avec les autres.
- Chaque variable est fortement corrélée avec un seul facteur principal (après rotation bien sûr).
- Toutes les composantes principales peuvent être définies.

L'interprétation d'un facteur se fait à partir des variables qui sont fortement corrélées avec celui-ci. Dans notre exemple, le facteur 1 est en forte corrélation avec les variables **empatt**, **longueur**, **largeur**, **poids**, **gazcap**, **consomm**, tandis que le facteur 2 est fortement corrélé avec les variables **prix**, **revente**, **chevaux**, **taillemot**.

Ainsi, les variables qui sont corrélées avec le facteur 1 ont en commun le « physique » de l'auto, tandis que celles qui sont corrélées avec le facteur 2 ont en commun la valeur et la puissance de l'automobile.

Si on décide d'utiliser ces deux facteurs pour résumer les variables originales, nous ne perdront que 17,75 % de l'information, et de plus ces facteurs ne sont pas corrélés entre eux, ce qui élimine la multicolinéarité s'ils sont utilisé dans une régression linéaire multiple. L'objectif de la prochaine section est de voir comment créer ces facteurs.

15.2.4 Les scores factoriels

Un score factoriel est une variable que l'on crée à partir de l'ACP pour représenter les facteurs extraits. Pour les créer, il suffit de refaire les commandes vues au début du chapitre, en ajoutant cette fois-ci l'étape suivante :

Dans le bouton Factor... : Save as variables
(laisser l'option Regression telle quelle)

Ces commandes permettent de créer de nouvelles variables **FAC1_1** et **FAC1_2**, qu'il est recommandé de renommer selon leur signification. Ces scores sont souvent utilisés pour produire des cartes perceptuelles.

15.2.5 Un autre exemple et cartes perceptuelles

Une étude a été menée auprès de familles situées dans quatre villes (Granby, Montréal, Québec et Sherbrooke). Le questionnaire était formé de 23 questions visant à faire ressortir la satisfaction par rapport au domicile de la famille (leur maison). Les 23 questions ont donné les variables suivantes :

Nom	Description
appa	APPARENCE EXTERIEURE
entr	FACILITE ENTRETIEN
chau	COUT CHAUFFAGE
prix	PRIX D'ACHAT
isol	ISOLATION
enso	ENSOLEILLEMENT
disp	DISPOSITION PIECES
plom	PLOMBERIE
elec	COUT ELECTRICITE
repa	REPARATIONS PREALABLES
rang	ESPACES RANGEMENT
vois	VOISINAGE
cuis	COMMODITE CUISINE
grob	GARDE-ROBES
taxe	TAXES
serv	PROXIMITE SERVICES
fini	FINITION PROPRETE
soli	SOLIDITE CONSTRUCTION
paie	MODE PAIEMENT
loca	EMPLACEMENT DANS LOCALITE
fene	QUALITE FENETRES
gran	GRANDEUR PIECES
taux	TAUX D'INTERET

Ces variables mesurent la satisfaction des gens sur une échelle de 1 à 9 avec la signification suivante :

- 1 très insatisfait
- 5 neutre
- 9 très satisfait

Cette échelle a été interprétée comme étant continue.

On décide de faire une ACP avec ces 23 questions afin de trouver quelles sont les principales facettes de la satisfaction qui ont été mesurées avec ces questions. On obtient

d'abord la sortie 15.7. On voit que le KMO a une valeur de 0,83, ce qui nous indique que l'ACP sera méritoire.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,830
Bartlett's Test of Sphericity	Approx. Chi-Square	1028,102
	df	153
	Sig.	,000

FIG. 15.7 – Le KMO

Communalities		
	Initial	Extraction
appa	1,000	,925
chau	1,000	,833
cuis	1,000	,794
disp	1,000	,782
elec	1,000	,733
enso	1,000	,927
entr	1,000	,804
fene	1,000	,790
fini	1,000	,939
gran	1,000	,944
grob	1,000	,774
isol	1,000	,800
loca	1,000	,735
paie	1,000	,663
plom	1,000	,694
prix	1,000	,781
rang	1,000	,635
repa	1,000	,807
serv	1,000	,847
soli	1,000	,795
taux	1,000	,759
taxe	1,000	,714
vois	1,000	,942

Extraction Method: Principal Component Analysis

FIG. 15.8 – Les communalités

On voit qu'effectivement l'ACP semble être bonne en jetant un coup d'œil aux *communalities* de la figure 15.8. La plus basse est à 0,663 (pour la variable *paie*), ce qui n'est pas mauvais du tout (ceci signifie que 66,3 % de la variance de la variable *paie* est expliquée par les facteurs de cette ACP). On voit aussi, par exemple, que la variance de la variable *gran* est expliquée à 94,4 % par les facteurs.

Justement, qu'en est-il de ces facteurs ? La figure 15.9 nous donne plus d'informations sur ceux-ci. Premièrement, étant donné qu'il y a 3 valeurs propres ayant une valeur supérieure à 1, on en déduit qu'il y aura 3 facteurs. On voit aussi dans les colonnes **Cumulative %** que ces 3 facteurs expliquent 80 % de la variance totale des 23 variables. Après rotation, on voit que le premier facteur explique 40,2 % de cette variance totale, que le deuxième en explique 25,6 % et le troisième 14,2 %.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	11,661	50,700	50,700	11,661	50,700	50,700	9,246	40,200	40,200
2	4,699	20,400	71,100	4,699	20,400	71,100	5,888	25,600	65,800
3	2,036	8,900	80,000	2,036	8,900	80,000	3,266	14,200	80,000
4	,640	2,800	82,800						
5	,567	2,500	85,200						
6	,463	2,000	87,300						
7	,457	2,000	89,200						
8	,419	1,800	91,100						
9	,324	1,400	92,500						
10	,317	1,400	93,800						
11	,290	1,300	95,100						
12	,273	1,200	96,300						
13	,168	,700	97,000						
14	,159	,700	97,700						
15	,123	,500	98,200						
16	,108	,500	98,700						
17	,080	,300	99,100						
18	,068	,300	99,400						
19	,057	,200	99,600						
20	,030	,100	99,700						
21	,023	,100	99,800						
22	,022	,100	99,900						
23	,015	,100	100,000						

Extraction Method: Principal Component Analysis.

FIG. 15.9 – Les valeurs propres

Il faut maintenant interpréter la matrice des facteurs obtenue après rotation (**Rotated Component Matrix**, figure 15.10) pour pouvoir définir les trois facteurs que l'on vient d'extraire.

	Component		
	1	2	3
appa	,348	,872	,209
chau	-,089	,032	,908
cuis	,817	,353	-,050
disp	,822	,274	,175
elec	,089	-,009	,851
enso	,326	,905	,040
entr	,859	,251	-,052
fene	,886	,032	,060
fini	,422	,863	,131
gran	,384	,891	,047
grob	,837	,260	-,075
isol	,814	,371	,012
loca	,833	,203	,001
paie	-,029	,211	,786
plom	,762	,255	-,219
prix	,030	,077	,868
rang	,773	,195	-,018
repa	,851	,236	-,165
serv	,836	,382	-,049
soli	,813	,363	-,051
taux	-,231	-,106	,833
taxe	-,016	,141	,833
vois	,382	,889	,072

a.

FIG. 15.10 – La matrice des facteurs après rotation

On remarque tout d'abord que 12 variables sont fortement corrélées avec le premier facteur et peu corrélées avec les deux autres. On peut donc se servir de ces variables pour définir le facteur 1. Ces variables sont les suivantes : **fene**, **cuis**, **disp**, **entr**, **grob**, **isol**, **loca**, **plom**, **rang**, **repa**, **serv**, **soli**. Ces variables semblent avoir en commun la **fonctionnalité** de la maison.

On voit aussi que 5 variables sont fortement corrélées avec le deuxième facteur et peu corrélées avec les deux autres. On peut donc se servir de ces variables pour définir le facteur 2. Ces variables sont les suivantes : **vois**, **appa**, **enso**, **fini**, **gran**. Ces variables semblent avoir en commun la **apparence** de la maison.

Finalement, on voit que 6 variables sont fortement corrélées avec le troisième facteur et peu corrélées avec les deux autres. Ces variables sont les suivantes : `paie`, `prix`, `taux`, `taxe`, `chau`, `elec`. Ces variables semblent avoir en commun les **coûts** reliés à la maison.

Ainsi, ce qui avait été initialement mesuré par les 23 questions se résume par la fonctionnalité, l'apparence et les coûts de la maison (mais avec une perte d'information de 20 %).

Il est possible de sauvegarder les scores factoriels de cette ACP (voir le chapitre sur l'ACP). Ici nous décidons de les renommer, et ces trois nouvelles variables portent maintenant les noms suivants : `fonction`, `apparence`, `cout`.

15.2.6 Les cartes perceptuelles

Une carte perceptuelle est un graphe à deux dimensions dont chacun des axes est associé à une variable continue. On choisit ensuite une variable discrète, et on calcule la valeur moyenne de chacune des variables continues pour chacune des modalités de la variable discrète, ce qui donne les points de la carte perceptuelle.

Dans le cadre d'une ACP, il est courant d'utiliser les scores factoriels pour construire des cartes perceptuelles. Pour illustrer nos propos, nous ferons ici des cartes perceptuelles à partir des facteurs de l'exemple (`fonction`, `apparence`, `cout`), et nous utiliserons comme variable discrète la variable `ville` (Granby, Montréal, Québec ou Sherbrooke).

Il faut donc calculer les moyennes des facteurs pour chacune de ces villes. Il faut effectuer les commandes ci-dessous pour dégager ces moyennes :

Menu SPSS :

→ Data

→ Split File...

✓ Compare groups

Dans la fenêtre Groups based on : → ville (la variable discrète)

puis

-
- | | |
|-------------------------------|--|
| Menu SPSS : | → Analyse |
| | → Descriptive Statistics |
| | → Descriptives... |
| Dans la fenêtre Variable(s) : | → fonction, apparence, cout (les facteurs) |
| Dans le bouton Options... : | <input checked="" type="checkbox"/> Mean (seulement) |
-

On obtient alors les moyennes de chaque facteur par ville, et à partir de ces données on crée une nouvelle base de données, ce qui dans le cadre de l'exemple donne ceci :

	ville	fonction	apparence	cout
1	Sherbrooke	-,49	,31	-1,02
2	Québec	1,03	-1,19	-1,35
3	Montréal	-1,14	-1,46	1,10
4	Granby	,55	1,37	,17
5				

FIG. 15.11 – La nouvelle base de données pour les cartes

On peut ensuite générer les cartes. Les commandes sont les suivantes (à effectuer pour chacune des cartes) :

-
- | | |
|--|--------------------------------|
| Menu SPSS : | → Graphs |
| | → Legacy Dialogs |
| | → Scatter/Dot... |
| Sélectionnez Simple Scatter, puis appuyez sur Define | |
| Dans la fenêtre Y Axis : | → fonction (un des facteurs) |
| Dans la fenêtre X Axis : | → apparence (un autre facteur) |
| Dans la fenêtre Set Markers by : | → ville (la variable discrète) |
-

En faisant les combinaisons possibles on obtient les cartes 15.12, 15.13 et 15.14. Voici leurs interprétations.

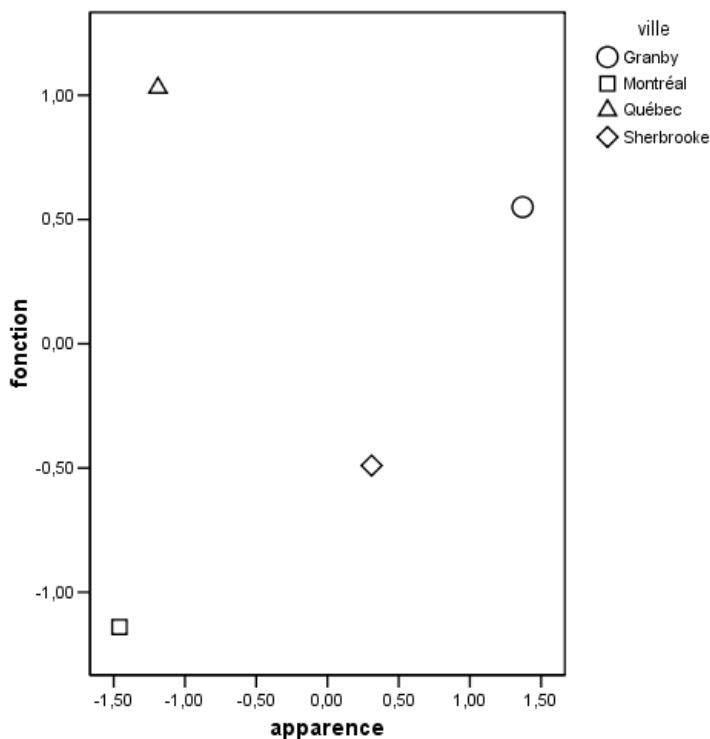


FIG. 15.12 – Carte perceptuelle pour fonction et apparence

Dans la carte 15.12, il ressort que les maisons de Québec sont plus fonctionnelles que la moyenne, mais moins belles en apparence. Les maisons les plus belles et les plus fonctionnelles sont situées à Granby, tandis que les moins belles et les moins fonctionnelles sont à Montréal.

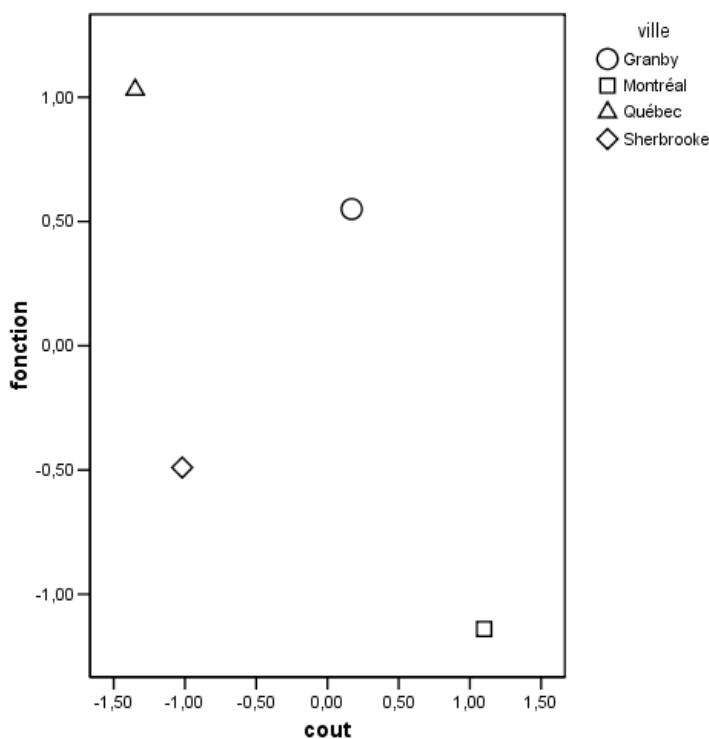


FIG. 15.13 – Carte perceptuelle pour fonction et cout

La deuxième carte (figure 15.13) met en relief les facteurs **fonction** et **cout**. Ce sont les familles de Québec qui se plaignent le plus des coûts, mais elles sont les plus satisfaites au niveau de la fonctionnalité. Bizarrement, les familles de Sherbrooke se plaignent plus des coûts que les familles de Montréal... mais celles-ci sont les moins satisfaites pour la fonctionnalité.

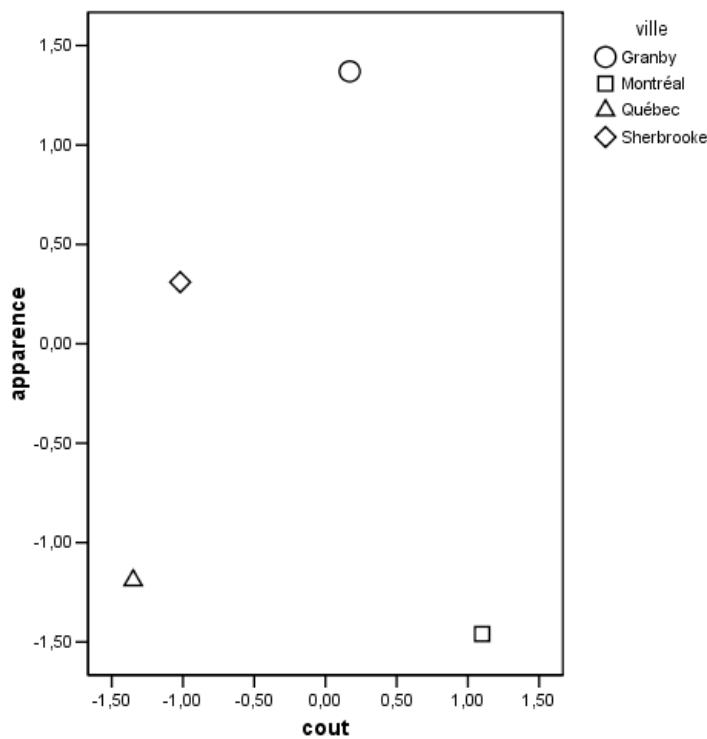


FIG. 15.14 – Carte perceptuelle pour apparence et cout

Finalement, la carte 15.14 concerne les facteurs **cout** et **apparence**. Encore une fois plusieurs choses ressortent. On peut dire entre autres que les familles de Montréal sont celles qui se plaignent le plus au niveau de l'apparence, mais sont les plus satisfaites pour les coûts (on voit que l'information commence à se répéter)...

15.3 Exercices du chapitre

Exercice 1 Un fournisseur de services de télécommunications aimerait mieux comprendre la façon dont ses services sont utilisés par ses abonnés de façon à leur offrir des forfaits attrayants. Il dispose de la base de données `telco.sav`, laquelle contient plusieurs informations à propos de l'utilisation des services des abonnés.

On décide donc de faire une ACP à propos des variables concernant les services de façon à mieux comprendre l'usage de ceux-ci par les abonnés, et à voir les liens possibles entre les facteurs qui seront dégagés. Plus précisément, faites l'ACP avec les variables `longdistm`, `sfracism`, `equipm`, `cartem`, `sansfilm`, `lignmult`, `boite`, `pagette`, `internet`, `affich`, `apellatt`, `transfert`, `conf`, `efact`, et analysez les résultats. Sauvegardez les scores factoriels, ils vous seront utiles pour un exercice du chapitre sur les clusters.

Exercice 2 Reprenons la base de données `satisfactiontravail.sav`. Il y 9 variables de la participation, et celles-ci semblent se regrouper naturellement de la façon suivante : les 3 premières, les 3 suivantes et finalement les 3 dernières. Faites une ACP avec ces 9 variables, et comparez les facteurs obtenus avec ces regroupements.

Chapitre 16

Analyse de la fidélité

Dès l'instant de notre naissance, le monde commence déjà à nous mesurer. En effet, une minute après la naissance, la santé du nouveau-né est mesurée sur l'échelle en dix points d'Apgar, suivie, cinq minutes plus tard, de la seconde échelle d'Apgar. Plus tard, d'autres échelles tenteront de mesurer notre intelligence, nos aptitudes, nos habiletés à conduire une automobile, etc...

En général, une échelle est constituée de plusieurs items. Le but de ce chapitre est de vérifier si un ensemble d'items donnés constituent une bonne échelle.

16.1 Fidélité et validité

Mesurer l'âge d'un individu est chose facile, mais mesurer le niveau de satisfaction au travail ou l'intelligence d'un individu n'est généralement pas si simple. Ainsi, la mesure d'un concept, telles la satisfaction ou l'intelligence, ne s'obtient pas directement et elle nécessite l'intervention d'un ensemble de variables (appelés items), chaque variable (question, dessin, etc...) mesurant, l'espère-t-on, l'intensité d'une facette du concept étudié.

Un outil de mesure est généralement composé d'un ensemble de variables. La réponse à chaque variable (item) peut être graduée, et ensuite, additionnée pour résulter d'un unique nombre, appelé score, et ce, pour chaque individu. Un bon outil possède la caractéristique suivante : plus la présence du concept étudié est forte chez un individu, plus son score sera élevé ; de même, moins l'intensité du concept est présente chez un individu, moins son score sera élevé. Mais justement, l'échelle obtenue est-elle adéquate ? Dans l'optique de répondre à la question, il est important de comprendre quelques caractéristiques des échelles de mesure.

Lorsqu'un professeur construit un examen afin de mesurer si les étudiants ont bien acquis les connaissances transmises, il sélectionne, parmi la population des questions possibles, un sous-ensemble limité de questions. Ainsi, l'outil de mesure ne contient qu'un échantillon de questions, et pourtant, le professeur désire tirer des conclusions sur le niveau de maîtrise des connaissances de chacun des étudiants. Dans les faits, le professeur espère que les scores obtenus par les étudiants auraient été en grande relation avec les scores que ces mêmes étudiants auraient pu obtenir en passant d'autres examens similaires. Il désire en fait que son examen soit **fidèle** et **valide**.

Un bon test fournit toujours des résultats stables, et ce, indépendamment des circonstances (par exemple, peu importe le professeur ou les conditions d'administration). Un tel outil est dit **fidèle** (*reliable*) et sa performance est dite **répétable**.

Pour être utile, un outil doit être fidèle, mais ce n'est pas suffisant. L'outil doit aussi être **valide** ; c'est-à-dire qu'un outil doit mesurer ce que l'analyste désire mesurer. Ainsi, un examen de statistique peut être un outil fidèle, mais mesure très mal les compétences des étudiants à faire du théâtre et ne serait donc pas valide dans ce contexte.

Dans le cadre de ce cours, nous étudions la fiabilité (la validité est difficilement quantifiable).

16.2 Analyse de la fidélité

Pour illustrer les différentes étapes de l'analyse de la fiabilité, nous considérons un ensemble de variables de la base de données `satisfactiontravail.sav` (les noms des variables ont été changés pour une meilleure compréhension, mais vous pouvez les retrouver par leurs *labels*). Le but est de savoir si ces cinq variables forment une bonne échelle (mesurent-elles un même concept ?) :

<code>info_sup</code>	Suis-je satisfait de l'information que partage mon patron avec moi ?
<code>estime</code>	Suis-je satisfait de l'estime qu'on me témoigne pour un travail bien fait ?
<code>entente</code>	Suis-je satisfait de l'entente qui existe entre mon supérieur et ses employés ?
<code>comp_sup</code>	Suis-je satisfait de la compétence technique de mon supérieur lorsqu'il prend des décisions ?
<code>sec_emploi</code>	Suis-je satisfait de ma sécurité d'emploi ?

Les quatre premières variables semblent mesurer la qualité de la relation qui réside entre les employés et leur patron. En effet, on se doute de la présence d'une relation positive entre la qualité de cette relation et le niveau de satisfaction global, d'où l'intérêt d'étudier cette facette. La dernière variable qui traite de la sécurité d'emploi semble légèrement détachée des autres variables. Ainsi, avant de construire une échelle de mesure basée sur ces cinq variables, il est impératif d'effectuer une étude de la fidélité.

Ces cinq variables ont été mesurées sur une échelle continue oscillant entre 0 (insatisfaction totale) à 12 (satisfaction absolue). Pour obtenir l'ensemble des sorties SPSS nécessaires à l'analyse de fidélité, il faut effectuer les commandes suivantes :

Menu SPSS :	→ Analyse
	→ Scale
	→ Reliability Analysis...
Dans la fenêtre Items :	→ info_sup, estime, entente, comp_sup, sec_emploi (les variables à l'étude)
Dans le bouton Statistics... :	→ Descriptive for ✓ Item ✓ Scale ✓ Scale if item deleted → Summaries ✓ Means ✓ Variances ✓ Correlations → Inter-Item ✓ Correlations

16.2.1 Description des résultats

Lorsque l'analyste désire étudier une échelle, il s'intéresse à plusieurs choses : les caractéristiques de chacune des variables, les relations entre chacun des items (variables), et les caractéristiques de l'ensemble de l'échelle finale.

La sortie 16.1 contient un ensemble de statistiques descriptives de chacun des cinq items. L'analyste observe que les moyennes des scores pour chacune des variables (items) varient entre 5,3877 et 6,8662 et que les écarts types (Std. Deviation) varient entre 2,58561 et 2,79313. La variable **estime** est celle qui possède le plus de variation. Compte tenu que les échelles sont les mêmes pour toutes les variables à l'étude, il apparaît que les

individus se diffèrentient plus sur la variable `estime` que sur les autres. Cette variable doit être une source importante expliquant la variation de la satisfaction chez les individus.

Item Statistics

	Mean	Std. Deviation	N
comp_sup	6,8662	2,58561	721
entente	6,7939	2,61605	721
sec_emploi	5,3877	2,77657	721
estime	5,9215	2,79313	721
info_sup	6,1000	2,70381	721

FIG. 16.1 – Statistiques descriptives

La sortie 16.1 contient la matrice des corrélations inter-items. Rappelons que l'analyste cherche à mesurer l'intensité d'un concept avec une intention du type : « Un score élevé indique la forte présence du concept étudié ». C'est pourquoi il est important que l'ensemble des items soient en corrélation. Si un item ne l'est pas, alors il ne va pas dans le même sens que les autres et ne mesure pas la même entité. Dans la sortie 16.1, l'analyste conataste que la variable `sec_emploi` possède bien peu de corrélation avec les autres variables, la corrélation maximale étant 0,236 avec la variable `estime`.

Inter-Item Correlation Matrix

	comp_sup	entente	sec_emploi	estime	info_sup
comp_sup	1,000	,710	,198	,530	,610
entente	,710	1,000	,232	,575	,632
sec_emploi	,198	,232	1,000	,236	,202
estime	,530	,575	,236	1,000	,617
info_sup	,610	,632	,202	,617	1,000

The covariance matrix is calculated and used in the analysis.

FIG. 16.2 – Matrice des corrélations

La sortie 16.3 contient différentes statistiques à propos de l'échelle résultante. Comme pour un examen standard, l'échelle (*scale*) est construite à partir de la somme des cinq variables.

La moyenne de l'échelle résultante est de 31,0692 avec un écart-type de 10,08578. La moyenne des cinq items est de 6,214 avec une variance de 0,386, le minimum des moyennes des cinq items est de 5,388 tandis que le maximum est de 6,866. Ces statistiques peuvent être validées à partir de la sortie 16.1. Il en va de même pour les variances et les corrélations.

Summary Item Statistics							
	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	6,214	5,388	6,866	1,479	1,274	,386	5
Item Variances	7,270	6,685	7,802	1,116	1,167	,250	5
Inter-Item Correlations	,454	,198	,710	,512	3,582	,042	5

The covariance matrix is calculated and used in the analysis.

Scale Statistics			
Mean	Variance	Std. Deviation	N of Items
31,0692	101,723	10,08578	5

FIG. 16.3 – Description de l'échelle

L'analyste est maintenant prêt à étudier les relations entre chacun des items et l'échelle résultante. La première et la seconde colonne de la sortie 16.4 contiennent la moyenne et la variance de l'échelle résultante si l'item de la ligne lue était enlevé. Par exemple, si l'item `comp_sup` est enlevé des analyses, alors la moyenne et la variance de l'échelle résiduelle seraient respectivement de 24,2031 et de 66,396.

La troisième colonne de la sortie 16.4 contient la corrélation de Pearson entre les valeurs de l'item en question et l'échelle résultante sans cet item. Une grande corrélation est généralement associée à un bon item. Par exemple, la corrélation entre les valeurs de l'item `sec_emploi` et d'une échelle basée sur la somme des quatre autres item est de seulement 0,258, ce qui est faible.

La quatrième colonne contient le coefficient de détermination r^2 d'une régression

Item-Total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
comp_sup	24,2031	66,396	,680	,553	,738
entente	24,2753	64,561	,721	,589	,724
sec_emploi	25,6816	81,102	,258	,070	,862
estime	25,1477	64,877	,646	,450	,747
info_sup	24,9692	64,583	,686	,524	,734

FIG. 16.4 – Liens entre les items et l'échelle

multiple. Ainsi, l'analyste peut voir que 55,3 % de la variation de la variable `comp_sup` est expliquée par les autres variables. De même, seulement 7,0 % de la variation de l'item `sec_emploi` est expliquée par les quatre autres variables.

La dernière colonne de la sortie 16.4 contient l'alpha de Cronbach de l'échelle si l'item est enlevé. L'alpha de Cronbach est souvent appelé le coefficient de « consistance interne » d'un test. Cette statistique varie entre -1 et +1 mais elle ne s'interprète que pour ses valeurs positive. Notez que des valeurs négatives pour l'alpha peuvent être obtenues ; ces valeurs surviennent lorsque les items ne sont pas positivement corrélés ensemble, et par conséquent dans ce cas la fidélité n'est pas bonne.

L'alpha peut être vu comme étant la corrélation entre l'échelle obtenue et toutes les autres échelles qui auraient pu être construite à partir de cinq items provenant de la population de tous les items possibles mesurant le concept à l'étude. Plus la corrélation est grande, meilleure est notre échelle de mesure.

Mentionnons finalement que l'alpha de Cronbach est aussi interprété comme étant le coefficient de détermination r^2 entre le résultat obtenu par une personne au présent test (le score observé) et le résultat que cette personne aurait pu obtenir en répondant à tous les items possibles (le vrai score).

L'alpha de Cronbach se calcule de la façon suivante :

$$\alpha = \frac{k \times \overline{\text{cov}}/\overline{\text{var}}}{1 + (k - 1) \times \overline{\text{cov}}/\overline{\text{var}}}$$

où k représente le nombre d'items, $\overline{\text{cov}}$ représente la moyenne de toutes les covariances entre les items et $\overline{\text{var}}$ représente la moyenne des variances des items.

Si les items ont été préalablement normalisés, l'alpha de Cronbach s'écrit plus simplement de la façon suivante :

$$\alpha = \frac{k \times \bar{r}}{1 + (k - 1) \times \bar{r}}$$

où la statistique \bar{r} représente la moyenne des coefficients de corrélation de Pearson entre les items. L'observation de cette équation illustre que l'alpha de Cronbach dépend à la fois des corrélations et du nombre de questions dans l'outil. Ceci peut amener certaines controverses ; en effet, si la moyenne des corrélations est de 0,20 avec 10 questions, l'alpha sera de 0,71, et si le nombre d'items passe à 25, l'alpha sera de 0,86 ! En somme, si le nombre de questions est assez élevé, il est possible d'obtenir un grand coefficient même si les corrélations entre les items sont faibles ! Dans le cadre de cet exemple, l'alpha de Cronbach est de 0,803 (sortie 16.6). Pour construire une échelle, il est recommandé d'obtenir un alpha supérieur à 0,7.

0,6 ≤ α ≤ 0,7	médiocre
0,7 < α ≤ 0,8	moyen
0,8 < α ≤ 0,9	très bien
0,9 < α ≤ 1	excellent

FIG. 16.5 – Interprétation de l'alpha de Cronbach

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,803	,806	5

FIG. 16.6 – L'alpha de Cronbach de lexemple

Revenons à la quatrième colonne de la sortie 16.4 qui contient la valeur de l'alpha si l'item est enlevé. Par exemple, si l'item `sec_emploi` est enlevé, l'alpha de Cronbach passe de 0,803 à 0,862, ce qui améliore la consistance globale de l'échelle de mesure. Dans tous les autres cas, le coefficient se détériore. Ainsi, sans l'item `sec_emploi`, notre échelle est davantage fidèle.

Chapitre 17

Méthode de classification (*clusters analysis*)

En se basant sur les ressemblances naturelles entre les choses, un vieux dicton stipule que : « Qui s'assemble se ressemble ! ». En effet, par l'observation des caractéristiques communes, il est possible de classer des unités statistiques (individus, autos, bières, animaux, etc...) en groupes. Par exemple, en examinant les comportements des oiseaux, on peut les classifier selon ceux-ci (les oiseaux de proies, les oiseaux migrateurs, etc...). On pourrait aussi classifier les étudiants selon leurs comportements d'étude (les studieux, les procrastinateurs, les indisciplinés, les efficaces, etc...). Il est possible de former de tels groupes d'éléments similaires en utilisant une méthode de classification.

En microbiologie, les analyses de classification (*cluster analysis*) sont utilisées pour identifier à quelle souche connue appartient une nouvelle bactérie. En marketing, ces analyses sont utilisées pour identifier les consommateurs qui adoptent des comportements similaires. Ces groupes sont inconnus et sont à identifier. Que le nombre de groupements soit connu à l'avance ou non, l'objectif d'une analyse de classification est d'identifier les

groupes homogènes (appelés en anglais *clusters*).

17.1 Généralités

Il est pertinent de se poser les questions suivantes avant de se lancer dans une analyse de classification :

- Sur quelles variables se baser pour former les *clusters* ?
- Comment calculer les distances ou la similarité entre les éléments ?
- Quel critère utiliser pour combiner deux groupes d'éléments (deux *clusters*) ?

La sélection des variables à inclure dans les analyses est l'étape la plus importante. En effet, ce sont sur ces variables que seront basés les critères de groupement des éléments. Ainsi, si d'importantes variables ne sont pas incluses, il est fort probable que les résultats soient erronés. Par exemple, dans une étude de classement des oiseaux, oublier une variable tel le poids moyen des oiseaux pourrait amener à grouper des oiseaux de poids extrêmement différents, ce qui peut mener à des groupements insatisfaisants.

17.1.1 La distance et la similarité

Les concepts de distance et de similarité sont importants dans ce type d'analyse. La distance mesure à quel point deux objets sont distants l'un de l'autre, tandis que la similarité mesure la proximité de ceux-ci. En fait, deux objets sont semblables si la distance entre eux est petite ou encore si la similarité est grande.

Le choix de l'utilisation d'une mesure de distance ou de similarité dépend du choix de l'échelle utilisée. Si les variables sont continues (échelle *scale*), les mesures utilisant le concept de distance sont appropriées. Cependant, si les variables sont discrètes (échelle nominale ou ordinale), le concept de similarité est plus adapté. Dans ce chapitre, nous traiterons le cas où les échelles sont continues.

Pour illustrer la simplicité des calculs des distances, prenons l'exemple de trois bières dont nous connaissons la valeur des variables `calories` et `prix` (prix à l'unité d'une bière d'une caisse de 24 bières). Le tableau 17.1 contient les valeurs de l'exemple. Mesurons la distance entre les bières afin de savoir quelles sont celles qui se ressemblent le plus en utilisant la distance Euclidienne. La distance Euclidienne est un outil extrêmement utilisé, qui correspond simplement à la racine de la somme des différences au carré.

Marque	calories	prix
Budweiser	144	0,43
Coors	140	0,44
Lowenbraw	157	0,48

FIG. 17.1 – Valeurs des variables `calories` et `prix`

$$d(\text{Bud}, \text{Coors}) = \sqrt{(144 - 140)^2 + (0,43 - 0,44)^2} = \sqrt{16,0001} \cong 4$$

$$d(\text{Bud}, \text{Low}) = \sqrt{(144 - 157)^2 + (0,43 - 0,48)^2} = \sqrt{169,0025} \cong 13$$

$$d(\text{Low}, \text{Coors}) = \sqrt{(157 - 140)^2 + (0,48 - 0,44)^2} = \sqrt{289,0016} \cong 17$$

D'après ces mesures on voit que la Budweiser et la Coors sont assez proches, tandis que la Lowenbraw est plus loin. Ainsi, en considérant les variables `calories` et `prix`, on peut dire que les bières Budweiser et Coors sont semblables.

Cependant, la distance Euclidienne possède le désavantage de se laisser influencer par l'unité des variables. Ainsi, si les variables sont mesurées par des échelles différentes, il est possible que l'une des variables contribue plus au calcul de distance que l'autre. Dans notre exemple, il est clair que la variable `calories` a été plus influente que la variable `prix`.

Pour contourner la problématique, il est courant de normaliser (standardiser) les variables soumises aux analyses. De cette façon, toutes les variables ont une moyenne de 0 et un écart-type de 1 et contribuent de façon « égale » aux calculs de distance.

Dans le cadre de ce chapitre, nous verrons deux méthodes de classification et utiliseront la base de données `Bières.sav` pour illustrer les analyses. Voici un extrait de cette base de données.

bière	origine	prix	calories	sodium	alcool
Budweiser Light	Américaine	2,63	113	8	3,70
Coors Light	Américaine	2,73	102	15	4,10
Michelob Light	Américaine	2,99	135	11	4,20
Miller Light	Américaine	2,55	99	10	4,30
Olympia Gold Light	Américaine	2,75	72	6	2,90
Pabst Extra Light	Américaine	2,29	68	15	2,30
Schlitz Light	Américaine	2,79	97	7	4,20
Anchor Steam	Américaine	7,19	154	17	4,70
Augsberger	Américaine	2,39	175	24	5,50
Becks	Allemande	4,55	150	19	4,70
Blatz	Américaine	1,79	144	13	4,80
Budweiser	Américaine	2,59	144	15	4,70
Coors	Américaine	2,65	140	18	4,80
Dos Equis	Mexique	4,22	145	14	4,50
Hamms	Américaine	2,59	136	19	4,40
Heilmans Old Style	Américaine	2,59	144	24	4,90
Heineken	Hollandaise	4,59	152	11	5,00
Henry Weinhard	Américaine	3,65	149	7	4,70
Kirin	Japonnaise	4,75	149	6	5,00
Kronenbourg	France	4,39	170	7	5,20

FIG. 17.2 – Un aperçu des données

Nous verrons deux types d'analyse pour faire des groupements : la **méthode hiérarchique** et la **méthode des nuées dynamiques**.

Avec la méthode hiérarchique, lorsque deux unités sont regroupées, elles ne peuvent être séparées ultérieurement. Avec les nuées dynamiques les regroupements se font d'abord de façon plutôt aléatoire, ce qui fait que deux unités qui faisaient d'abord partie d'un même groupe peuvent se retrouver séparées en cours d'analyse.

17.2 La méthode hiérarchique

La formation des *clusters* (groupes) avec cette méthode se divise en plusieurs étapes. On forme des groupes de plus en plus importants jusqu'à ce que toutes les unités soient regroupées au sein d'un seul et même *cluster*. Pour expliquer ces étapes, supposons que

nous avons k unités (observations) dans la base de données.

Étape 0 : Toutes les unités sont considérées comme étant des *clusters* distincts.

Ainsi, il y a autant de *clusters* que d'observations. Par exemple, si nous voulons classer 20 bières, il y aura au départ 20 clusters.

Étape 1 : Les deux *clusters* ayant la plus petite distance entre eux sont regroupés, formant ainsi un nouveau *cluster* contenant deux éléments. Dans l'exemple du classement des 20 bières, il y a 19 clusters à la fin de cette étape.

Étape 2 : à cette étape, soit un troisième *cluster* est combiné au *cluster* de l'étape 1, soit deux autres *clusters*(observations solitaires) sont combinés pour former un nouveau *cluster* contenant deux observations.

Étape 3 : ...

Étape k : Toutes les observations ont été regroupées au sein d'un seul et même *cluster*.

Une fois un *cluster* formé, il ne peut être décomposé en plusieurs *clusters*, il ne peut qu'être combiné avec un autre *cluster*. Dans notre exemple, si deux bières sont groupées à une étape, elles le seront jusqu'à la fin du processus. Évidemment, ultérieurement d'autres bières pourront être ajoutées à celles-ci. Cette stratégie d'agglomération se reflète dans le nom que lui donnent les anglophones : *agglomerative hierarchical cluster analysis*.

Le processus arrête lorsque tous les *clusters* sont combinés en un seul et même *cluster*. Ainsi, au départ, il y avait autant de *clusters* que d'observations ; à la fin du processus il ne reste qu'un seul *cluster*. L'analyste doit alors choisir à partir de quelle étape les groupements des *clusters* deviennent incohérents, ce qui lui permettra de connaître combien de *clusters* il conservera.

17.2.1 Critères pour combiner deux *clusters*

Il existe plusieurs critères pour combiner deux *clusters*. Trois techniques sont plus courantes que les autres. Pour les illustrer, supposons que nous avons deux *clusters* (I et II), que le *cluster* I contient l'unité *A* et que le *cluster* II contient les unités *B* et *C*. On suppose de plus que $d(A, B) = 5$ et $d(A, C) = 7$.



Voici les trois techniques :

- La *nearest neighbor technique* porte aussi le nom de *single linkage*. Selon cette technique, la distance entre deux *clusters* est le minimum des distances que l'on peut calculer avec toutes les paires d'observations formées à partir d'une observation dans un des *clusters* et d'une observation choisie dans l'autre. Par exemple, à l'étape 1, la distance entre deux *clusters*, qui sont des observations solitaires, est la distance entre ces deux observations. Ainsi les *clusters* regroupés sont ceux dont la distance les séparant est la plus petite, formant ainsi un nouveau *cluster*. Ensuite, la distance entre ce nouveau *cluster* et un autre se définit comme étant la plus petite distance entre les observations de ce nouveau *cluster* et l'observation de l'autre *cluster*. En somme, la distance entre deux *clusters* est la distance entre les deux observations les plus proches. Dans l'exemple, la distance entre le *cluster* I et le *cluster* II serait ainsi de 5.
- La *furthest neighbor technique* porte aussi le nom de *complete linkage*. Selon cette technique, la distance entre deux *clusters* est le maximum des distances que l'on peut calculer avec toutes les paires d'observations formées à partir d'une observation dans un des *clusters* et d'une observation choisie dans l'autre. En somme, la distance entre deux *clusters* est la distance séparant les deux observations les plus éloignées. Dans l'exemple, la distance entre le *cluster* I et le *cluster* II serait ainsi de 7.
- La *average linkage between groups method* est la méthode par défaut que propose

SPSS (*between-groups linkage*), et se retrouve à mi-chemin entre les deux techniques précédentes. Selon cette technique, la distance entre deux *clusters* est la moyenne des distances que l'on peut calculer avec toutes les paires d'observations formées à partir d'une observation dans un des *clusters* et d'une observation choisie dans l'autre. Dans l'exemple, la distance entre le *cluster* I et le *cluster* II serait ainsi de 6.

Illustrons maintenant la méthode hiérarchique à l'aide d'un exemple (*Bières.sav*). Tout d'abord, il est important de comprendre que cette technique s'utilise habituellement avec peu de données (50 ou moins), sinon elle est plutôt lourde à interpréter. Par contre, elle a l'avantage de présenter toutes les solutions possibles alors que pour la méthode des nuées dynamiques il faut indiquer dès le départ le nombre de *clusters* désirés. Ainsi, on utilise souvent la méthode hiérarchique avec un sous-échantillon pour se donner une idée du nombre de *clusters* requis, puis on utilise la méthode des nuées dynamiques avec un nombre de *clusters* inspiré de la méthode hiérarchique.

Ainsi, même si dans l'exemple la taille de l'échantillon n'est que de 35, nous choisirons 20 données pour la méthode hiérarchique, et nous prendrons l'échantillon complet pour l'autre méthode.

Habituellement, on choisit de façon aléatoire le sous-échantillon ; afin d'y arriver il faut effectuer les commandes suivantes :

Menu SPSS : → Data

→ Select Cases...

Sélectionnez Random sample of cases, puis appuyez sur Sample...

Faites Exactly 20 cases from the first 35 cases

De cette façon SPSS sélectionne 20 observations au hasard parmi tout l'échantillon.

Ainsi, nous sommes prêts à utiliser la méthode hiérarchique de classification. Ici on

s'intéresse à classifier les bières selon les variables **prix**, **calories**, **sodium** et **alcool**. Les commandes à effectuer sont les suivantes :

Menu SPSS :	<ul style="list-style-type: none"> → Analyse → Descriptive Statistics → Descriptives...
Dans la fenêtre Variable(s) :	<ul style="list-style-type: none"> → prix, calories, sodium, alcool (les variables de classification, elles sont continues) <input checked="" type="checkbox"/> Save standardized values as variables (ceci sert à standardiser les variables pour qu'elles aient toutes la même portée)

Menu SPSS :	<ul style="list-style-type: none"> → Analyse → Classify → Hierarchical Cluster...
Dans la fenêtre Variable(s) :	<ul style="list-style-type: none"> → Zprix, Zcalories, Zsodium, Zalcool (les variables de classification standardisées)

La table d'agglomération (figure 17.3) présente les résultats de chacune des étapes de la formations des *clusters*.

Au départ, on considère qu'il y a 20 *clusters*. Le *stage* 1 (première ligne) de la table nous indique que pour passer de 20 à 19 *clusters*, on a combiné la bière 27 avec la bière 34, ce qui signifie que ces deux bières étaient les plus semblables d'après la distance Euclidienne (parmi les 20 bières choisies). Cette distance est donnée dans la colonne **Coefficients**. Ainsi la distance entre les bières 27 et 34 est de 0,075.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
	Stage	Stage	Stage	Stage	Stage	
1	27	34	,075	0	0	7
2	13	15	,121	0	0	9
3	10	33	,139	0	0	10
4	21	25	,142	0	0	8
5	28	29	,191	0	0	6
6	28	31	,332	5	0	16
7	16	27	,393	0	1	12
8	21	22	,393	4	0	11
9	12	13	,415	0	2	11
10	10	14	1,036	3	0	15
11	12	21	1,157	9	8	14
12	9	16	2,289	0	7	14
13	5	6	2,660	0	0	19
14	9	12	2,718	12	11	17
15	10	19	3,660	10	0	18
16	1	28	4,289	0	6	17
17	1	9	5,779	16	14	18
18	1	10	7,549	17	15	19
19	1	5	20,868	18	13	0

FIG. 17.3 – La table d’agglomération

Les deux colonnes de *Stage Cluster First Appears* indiquent à quel *stage* les *clusters* avec plus de deux observations apparaissent. Ainsi, au *stage* 6, le 5 dans la colonne **Cluster 1** nous indique que la bière 28 avait déjà été combinée à la bière 29 au *stage* 5. Le nombre 16 dans la colonne **Next Stage** indique que ce nouveau *cluster* (formé des bières 28, 29 et 31) sera impliqué dans un regroupement au *stage* 16.

En examinant les coefficients de la colonne **Coefficients**, il est possible de se faire une idée sur la similarité des *clusters* qui viennent d’être combinés. De petites valeurs de distance indiquent que les éléments combinés sont homogènes. De grandes valeurs indiquent que les éléments combinés sont plutôt différents. Rappelons que ces valeurs dépendent de la mesure choisie.

Ces valeurs peuvent être utilisées pour savoir combien de *clusters* seront gardés à la fin pour bien représenter les données. À titre de règle non officielle, il est bien d’arrêter l’agglomération lorsque les coefficients effectuent un bond important entre deux *stages* successifs. Par exemple, ici il y a un bond important entre la distance de la solution comportant 4 *clusters* et celle de la solution comportant 3 *clusters* ; en effet, la distance passe

de 4,289 à 5,779 (dans les stages 16 et 17), montrant ainsi le manque d'homogénéité. Ce qui indique qu'il serait préférable de garder quatre *clusters* ou plus pour bien représenter les groupements naturels dans les données. Pour savoir combien de *clusters* il faut garder, il faut aussi consulter les autres sorties.

La prochaine sortie à interpréter est celle du **Vertical Icicle** (sortie 17.4).

Number of clusters	Vertical Icicle																					
	Case																					
6	5	4	19	14	33	10	22	25	21		15	13	12	34	27	16	9	31		29	28	1
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
2	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
3	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
4	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
5	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
6	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
7	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
8	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
9	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
10	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
11	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
12	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
13	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
14	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
15	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
16	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
17	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
18	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
19	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

FIG. 17.4 – Le Vertical Icicle

Les colonnes dans la figure **Vertical Icicle** correspondent aux éléments à classer ; ici, les bières. Contrairement aux représentations usuelles, la figure doit être lue de bas en haut. La ligne 19 représente la première étape et la ligne 1 représente la dernière étape où tous les éléments sont groupés dans un seul et même *cluster*. L'étape 0, ou la ligne 20, n'est pas illustrée. À cette étape, il y a autant de *clusters* que d'éléments.

Dans notre exemple, l'étape 0, non illustrée, correspondrait à la ligne 20, à 20 bières et à 20 *clusters*. À l'étape 1, la ligne 19, les bières les plus proches (selon le critère choisi) sont groupées et forment un nouveau *cluster*. Dans cet exemple, les bières 27 et 34 sont regroupées. Il reste donc 19 *clusters* à la fin de la première étape, et ainsi de suite. Le déroulement peut être lu dans la table d'agglomération. En somme, le numéro de la ligne correspond au nombre de *clusters* résultant.

La ligne 5 correspond ainsi à la présence de cinq *clusters*. Les bières 5 et 6 forment un des ces *clusters*, les bières 19, 14, 33, 10 en forment un autre, et ainsi de suite.

L'objectif consiste à déterminer le nombre de *clusters* qui représentera bien la tendance naturelle des données à se grouper. Selon la statistique **Coefficient** issue de la table d'agglomération, le nombre de *clusters* devait être supérieur ou égal à quatre. En général, il est de mise de scruter les solutions se situant près de 4. Étudions les solutions avec 4, 5 ou 6 *clusters*.

Les solutions à 5 et 6 *clusters* ont le défaut d'avoir un *cluster* ne contenant qu'une seule bière (la bière 1). La solution à 4 *clusters* semble donc meilleure pour la répartition. C'est donc celle-ci que nous allons interpréter. Pour ce faire, il faut à nouveau effectuer des commandes SPSS :

Menu SPSS :	→ Analyse
	→ Classify
	→ Hierarchical Cluster...
Dans la fenêtre Variable(s) :	→ Zprix, Zcalories, Zsodium, Zalcool (les variables de classification standardisées)
Dans le bouton Statistics... :	✓ Single solution Number of clusters : 4
Dans le bouton Save... :	✓ Single solution Number of clusters : 4

Ces commandes auront comme effets (nouveaux) de générer la sortie 17.5 et de créer la variable **CLU4_1**, qui dans les deux cas servent à montrer clairement à quel *cluster* appartiennent les observations.

Il est ensuite possible de générer le tableau des moyennes standardisées. Ce tableau est très important puisque c'est sur celui-ci que repose l'interprétation des *clusters*. Voici comment le générer :

Cluster Membership	
Case	4 Clusters
1	1
5	2
6	2
9	3
10	4
12	3
13	3
14	4
15	3
16	3
19	4
21	3
22	3
25	3
27	3
28	1
29	1
31	1
33	4
34	3

FIG. 17.5 – Appartenance aux clusters

Menu SPSS : → Data
→ Split File...
✓ Compare Groups

Dans la fenêtre Groups Based on : : → CLU4_1

ensuite

Menu SPSS :
 → Analyse
 → Descriptive Statistics
 → Descriptives...

Dans la fenêtre Variable(s) : → Zprix, Zcalories, Zsodium, Zalcool

Dans le bouton Options... : ✓ Mean (seulement)

On remarque d'abord que le *cluster* 1 contient 4 bières, le *cluster* 2 en a 2, le *cluster* 3 en a 10 et le *cluster* 4 en a 4.

CLU4_1	Zprix		Zcalories		Zsodium		Zalcool		Valid N
	N	Mean	N	Mean	N	Mean	N	Mean	
1	4	-,8645796	4	-,0057322	4	-1,16383	4	-,0334116	4
2	2	-,5141520	2	-2,64254	2	-,7637626	2	-2,57269	2
3	10	-,2682782	10	,4108066	10	,6619276	10	,4209859	10
4	4	1,7923511	4	,2999844	4	-,1091089	4	,2672926	4

FIG. 17.6 – Le tableau des moyennes standardisées

De ce tableau, on remarque que les bières incluses dans le *cluster 3* se distinguent par leur plus grande quantité de sodium. Les bières incluses dans le *cluster 4* se distinguent par leur prix plus élevé (bières importées). Le *cluster 1* contient des bières moyennes ne se distinguant pas vraiment des autres. Le *cluster 2* regroupe des bières en plus faible teneur en alcool et en calories, on pourrait interpréter ces bières comme étant des bières légères.

Dans notre exemple, les solutions sont assez distinctes. Si ce n'est pas le cas, il faut alors observer d'autres solutions, par exemple la solution avec 5 ou 6 *clusters*, et demander autant de tableaux de moyennes que de solutions analysées. La meilleure solution est souvent davantage une question d'interprétation.

17.3 La méthode des nuées dynamiques

Il est clair que la méthode de classification hiérarchique est intéressante pour classer de petits échantillons ayant au maximum 50 données. Cependant, cette méthode est moins adaptée aux études comportant, par exemple, 200 unités à classer. Cette section présente la méthode de classification en nuées dynamiques, appelée en anglais *K-Means Cluster*, qui a été conçue pour classer un nombre important d'éléments.

Contrairement à la méthode hiérarchique, qui présente plusieurs solutions, chacune comportant un nombre différents de *clusters*, les nuées dynamiques ne fournissent qu'une

seule solution pour un nombre prédéterminé de *clusters*. C'est l'analyste qui détermine le nombre de *clusters* désirés. Aussi, avec la méthode hiérarchique, deux éléments combinés à une étape ne seront jamais dissociés ; avec les nuées dynamiques, des éléments associés au cours d'une étape peuvent être dissociés dès l'étape suivante.

Deux approches sont possibles pour spécifier le nombre de *clusters* voulus : l'essai-erreur et la méthode hiérarchique. En effet, une méthode hiérarchique peut être utilisée sur un sous-échantillon aléatoire pour déterminer un nombre probable de *clusters*, par exemple, 4, 5 ou 6 *clusters*.

Le processus utilisé par les nuées dynamiques repose sur le centroïde le plus rapproché. Un élément est assigné au *cluster* le plus près.

Plus précisément, supposons que nous voulons obtenir k *clusters*. L'algorithme se résume alors ainsi :

- Initialement, la procédure tire au hasard k éléments de la base de données. Ces éléments sont imposés à titre de k centroïdes temporaires.
- Les distances de tous les éléments par rapport à chacun des k centroïdes sont calculées.
- Chaque élément est associé au centroïde le plus rapproché, formant ainsi k groupes (les *clusters*).
- Pour chaque groupe, la moyenne des valeurs de chacune des unités est calculée, formant un point fictif moyen. Ces points fictifs sont appelés les **centres de gravité**. Il y en a donc un par groupe. Ces points moyens deviennent les nouveaux centroïdes. Le processus réitère.

Cette méthode est donc itérative et converge vers une solution optimale. Le processus est arrêté lorsque les centroïdes ne bougent plus ou lorsque le nombre d'étapes fixées à l'avance est atteint.

Utilisons l'exemple du classement des bières selon quatre variables en utilisant l'ensemble des 35 bières. En s'inspirant de l'exemple précédent, on décide de fixer à 4 le nombre de *clusters*.

Premièrement, il faut reprendre l'exemple de zéro. C'est-à-dire qu'il faut re-sélectionner toutes les données et les standardiser à nouveau (ici nous supposons que les anciennes variables standardisées ont été effacées de la base de données). On effectue donc les commandes suivantes :

Menu SPSS :	→ Data
	→ Select Cases...

Sélectionnez All Cases

Menu SPSS :	→ Analyse
	→ Descriptive Statistics
	→ Descriptives...

Dans la fenêtre Variable(s) :	→ prix, calories, sodium, alcool (les variables de classification, elles sont continues) ✓ Save standardized values as variables (ceci sert à standardiser les variables pour qu'elles aient toutes la même portée)
-------------------------------	--

On peut ensuite effectuer les commandes pour l'analyse avec les nuées dynamiques :

Menu SPSS :	→ Analyse
	→ Classify
	→ K-Means Cluster...

Dans la fenêtre Variable(s) :	→ Zprix, Zcalories, Zsodium, Zalcool (les variables de classification standardisées)
-------------------------------	---

Number of clusters : 4

Dans le bouton Save... : ✓ Cluster membership

Dans le bouton Options... : ✓ Initial cluster centers
✓ ANOVA table

À la première étape, l'algorithme sélectionne au hasard quatre bières dans la liste des 35 bières. La sortie 17.7 contient justement ces centres initiaux.

		Initial Cluster Centers			
		1	2	3	4
Zprix		-,65641	-,21134	,3,70524	-,56740
Zcalories		-2,93581	-1,74956	,58202	1,44102
Zsodium		,05579	-1,24607	,38126	1,52039
Zalcool		-3,77651	-,62547	,20375	1,53050

FIG. 17.7 – Les centres initiaux

D'itération en itération, les centres de gravité se déplacent vers la solution finale, d'où le nom de nuées dynamiques. À chaque étape, la distance entre les centres de gravité de cette étape et celle d'avant devient de plus en plus petite. L'algorithme arrête lorsque la distance entre les centres de gravité est nulle (0 est l'option par défaut dans le bouton **Iterate...**) ou lorsque le nombre maximum d'itération est atteint (nombre fixé par défaut à 10 dans **Iterate...**). Le tableau 17.8 contient ces distances.

		Iteration History ^a			
		Change in Cluster Centers			
Iteration		1	2	3	4
1		,912	1,526	2,045	1,679
2		,000	,247	,144	,061
3		,000	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 3. The minimum distance between initial centers is 3,637.

FIG. 17.8 – Les distances d'étape en étape

Ainsi, à partir de la solution initiale, qui dépend du hasard, on obtient la solution finale. Cette solution finale est présentée dans le tableau *Final Cluster Centers* (figure 17.9). Cette solution finale est composée des centres de gravité des quatre *clusters*; ce

sont des moyennes. Ces moyennes sont justement les moyennes standardisées servant à interpréter les *clusters*.

	Final Cluster Centers			
	1	2	3	4
Zprix	-,45168	-,43744	,66429	-,39827
Zcalories	-2,85400	-,47333	,50021	,43374
Zsodium	-,67651	-,85551	-,17668	,69656
Zalcool	-3,27898	-,27720	,41698	,40069

FIG. 17.9 – Le tableau des moyennes standardisées (centres finaux)

Ici, on voit que c'est le 3e *cluster* qui se distingue au niveau du prix, et les bières de ce *cluster* ont plus de calories et d'alcool en moyenne que les bières des autres *clusters*. On peut interpréter ce groupe comme étant le groupe des bières importées.

Le *cluster* 1, quant à lui, se distingue au niveau des calories et de l'alcool : ce sont probablement les bières légères.

Le *cluster* 4 contient des bières qui ont plus de sodium. Ces bières correspondent sûrement aux bières américaines.

Finalement, le *cluster* 2 contient les bières qui ont le moins de sodium.

Pour valider si une analyse de groupement a été efficace, il est utile d'analyser les données incluses dans la table ANOVA (figure 17.10). Cette table contient les carrés moyens utilisés pour examiner l'hétérogénéité entre les *clusters*.

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zprix	8,083	3	,315	31	25,696	,000
Zcalories	7,764	3	,345	31	22,479	,000
Zsodium	5,405	3	,574	31	9,422	,000
Zalcool	8,686	3	,256	31	33,902	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

FIG. 17.10 – Table ANOVA

Compte tenu que nous voulons des *clusters* hétérogènes entre eux et que chacun des *clusters* contient des éléments homogènes, de grandes valeurs pour les carrés moyens des *clusters* et de petites valeurs pour les autres carrés moyens (**Error**) sont espérées. Ainsi, de grandes valeurs pour les quotients F ($ClusterMS/ErrorMS = F$) ou de petites valeurs pour les *p*-values sont associées à une bonne analyse. Cependant, compte tenu que l'analyse en nuées dynamiques n'a pas été conçue pour minimiser les carrés moyens, il faut prendre les données dans ce tableau à titre descriptif seulement. Ici, toutes les *p*-values étant à 0,000 et les F étant grands, on peut prétendre que le groupement est bon.

Le dernier tableau nous informe sur la répartition des bières dans les *clusters*.

Number of Cases in each Cluster		
Cluster	1	2,000
2		10,000
3		7,000
4		16,000
Valid		35,000
Missing		,000

FIG. 17.11 – Répartition des bières

On peut parfois préférer une solution à une autre (en terme de nombre de *clusters*) à cause de la répartition. On préfère habituellement que la répartition ne soit pas trop déséquilibrée (par exemple ce n'est habituellement pas intéressant d'avoir un *cluster* avec un seul élément).

Exemple 17.3.1 La base de données `ventesautos.sav` contient des informations à propos de plusieurs modèles d'automobiles et de camions. On s'intéresse ici à les classifier selon leurs caractéristiques physiques et leur valeur.

On décide d'explorer les solutions à 3 et 4 clusters avec les nuées dynamiques. De plus, dans le bouton `Save...`, on a coché `Distance from cluster center` afin d'évaluer visuellement la variation dans chacun des clusters, et dans le bouton `Options...` on a

coché *Cluster information for each case*. Nous verrons les conséquences de ceci en détail plus loin dans l'exemple.

Iteration History ^a			
Iteration	Change in Cluster Centers		
	1	2	3
1	4,565	3,899	3,975
2	,699	,197	1,823
3	,389	,182	,601
4	,193	,091	,597
5	,078	,090	,351
6	,000	,141	,443
7	,000	,035	,117
8	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 8. The minimum distance between initial centers is 10,326.

Final Cluster Centers			
	Cluster		
	1	2	3
Zprix	-,58809	,12022	1,34528
Ztaillemot	-,86360	,23733	1,74226
Zchevaux	-,79218	,24963	1,42275
Zempatt	-,66949	,28632	1,01185
Zlargeur	-,80908	,29864	1,33192
Zlongueur	-,73877	,34766	,93731
Zpoids	-,86234	,33622	1,47717
Zgaz cap	-,75885	,22229	1,57030
Zconsomm	,74852	-,30752	-1,25325

FIG. 17.12 – Le nombre d'itérations et les centres finaux (3 clusters)

Commençons par analyser la solution à 3 clusters. La figure 17.12 nous montre que l'algorithme a convergé complètement en 8 itérations, ce qui est rassurant ; la solution est stable.

La deuxième sortie de la figure 17.12 nous montre les centres finaux, c'est à partir de ceci qu'on peut interpréter les clusters. On voit que dans le premier groupe, tous les véhicules sont en bas de la moyenne pour tous les aspects exception faite de la consommation. Il semble donc que ce soit les véhicules les moins chers, les plus petits, les moins puissants, et qui consomment le moins d'essence. On peut les identifier comme étant les **petits véhicules**.

Le deuxième groupe se retrouve entre les clusters 1 et 3 pour tous les aspects. Donc ce sont des véhicules plus gros, plus puissants et consommant plus que les véhicules du premier cluster, mais moins gros, moins puissants et consommant moins que ceux du 3e cluster. On va donc les identifier comme étant les **véhicules moyens**.

Et donc avec une analyse semblable on en vient à identifier le 3e groupe comme étant

celui des **gross véhicules**.

Distances between Final Cluster Centers						
Cluster	1	2	3			
1		3,104	6,369			
2		3,104		3,331		
3	6,369		3,331			

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zpri	30,387	2	,615	149	49,408	,000
Ztaillemot	57,271	2	,255	149	224,832	,000
Zchevaux	43,110	2	,439	149	98,166	,000
Zempatt	27,649	2	,663	149	41,722	,000
Zlargeur	42,256	2	,454	149	93,126	,000
Zlongueur	30,492	2	,610	149	49,979	,000
Zpoids	50,179	2	,360	149	139,556	,000
Zgazcap	45,711	2	,426	149	107,323	,000
Zconsomm	37,356	2	,522	149	71,497	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal

Number of Cases in each Cluster		
Cluster	1	2
1	63,000	
2		68,000
3		21,000
Valid		152,000
Missing		5,000

FIG. 17.13 – Distance entre les centres finaux, table ANOVA et répartition

La première sortie de la figure 17.13 nous montre la distance entre les centres finaux de chacun des clusters (sortie qui est générée lorsqu'on coche **Cluster information for each case** dans le bouton **Options...**). Ainsi on voit que c'est la distance entre les clusters 1 et 3 qui est la plus grande (6,369), le deuxième cluster est à mi-chemin entre les deux (distance de 3,104 du premier cluster et de 3,331 du troisième cluster) ; ceci va dans le même sens que notre interprétation.

Les *p*-values de la table ANOVA sont toutes nulles, ce qui nous indique que la solution est bonne ; on pourra comparer les *F* avec ceux de la solution à 4 clusters. Aussi, les

variables ayant les plus grands F sont celles qui varient le plus d'un cluster à l'autre. Ici on voit que c'est la taille du moteur qui varie le plus, suivie du poids puis de la capacité du réservoir à essence.

Finalement, la dernière sortie nous indique la répartition. Il y a 63 petits véhicules, 68 véhicules moyens et 21 gros véhicules. Cette répartition est satisfaisante en ce sens qu'il n'y a pas de très gros ou très petit cluster.

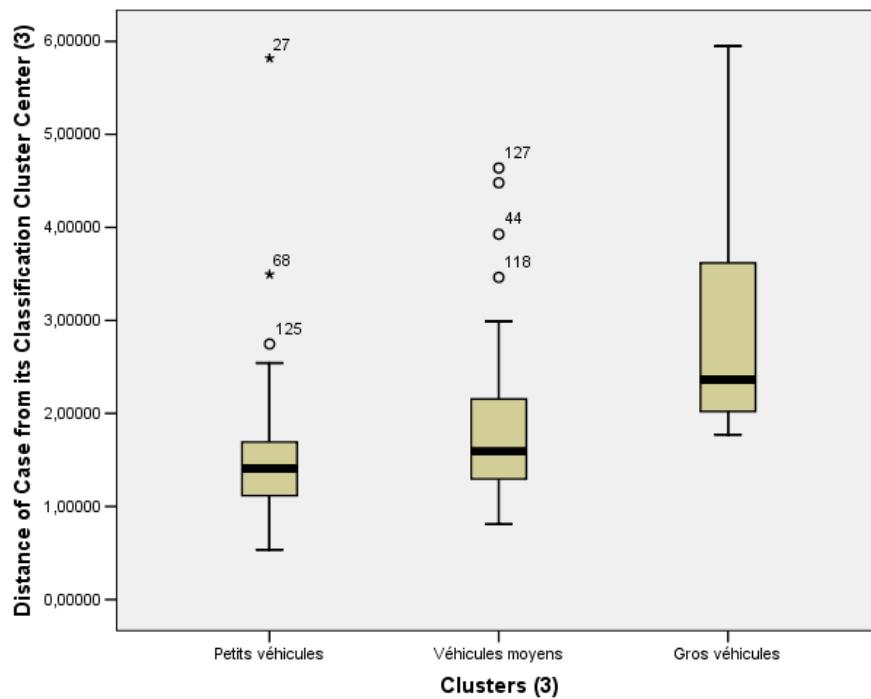


FIG. 17.14 – Variation dans les clusters

La sortie de la figure 17.14 contient des boxplots qui illustrent la variation dans chacun des clusters et les véhicules qui s'éloignent beaucoup des centres (les *outliers*). Pour obtenir cette sortie, il suffit d'effectuer les commandes suivantes :

Menu SPSS :

- Graphs
- Boxplot...
- Define

Dans la fenêtre Variable :

- la variable représentant la distance entre chaque cas et le centre du cluster

Dans la fenêtre Category Axis :

- la variable des clusters

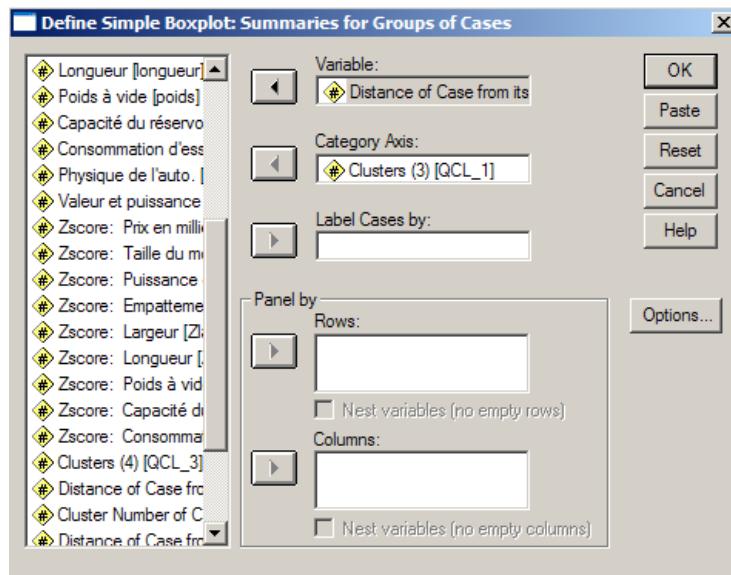


FIG. 17.15 – Pour obtenir les Boxplots

	manufact	modele	ventes	revente	type	prix	taillemot	chevaux
27	Chevrolet	Metro	21,855	5,160	Automobile	9,235	1,0	55

FIG. 17.16 – Le véhicule 27

C'est dans le cluster des petits véhicules qu'il semble il y avoir le moins de variation, mais il possède le plus grand outlier ; c'est le véhicule de la ligne 27, la Chevrolet Metro. Si on regarde en détail les valeurs de cette voiture, on voit qu'elle se distingue de la

moyenne des petites voitures sur plus d'un aspect : elle est plus petite, et surtout ne consomme que très peu (son Z_{consomm} est à 4,94).

On voit aussi que c'est dans le cluster des gros véhicules qu'il y a le plus de variation.

Iteration	Change in Cluster Centers			
	1	2	3	4
1	3,452	2,991	3,112	1,729
2	1,207	,745	1,383	,116
3	,471	,161	,000	,182
4	,263	,000	,000	,154
5	,173	,000	,000	,113
6	,129	,088	,000	,117
7	,065	,121	,000	,073
8	,058	,000	,000	,048
9	,000	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 9. The minimum distance between initial centers is 7,628.

	Final Cluster Centers			
	1	2	3	4
Z_{prix}	-,66501	,44694	2,67343	-,00565
$Z_{\text{taillemot}}$	-,96216	1,21527	1,61158	,08670
Z_{chevaux}	-,89687	,63122	2,40144	,10958
Z_{empatt}	-,73747	1,49016	-,48690	,10291
Z_{largeur}	-,86687	1,43526	,29775	,09841
Z_{longueur}	-,85513	1,33978	-,18937	,17850
Z_{poids}	-,96511	1,48359	,35506	,18270
$Z_{\text{gaz cap}}$	-,85323	1,54554	,59537	,05511
Z_{consomm}	,88566	-1,11064	-,74178	-,20451

FIG. 17.17 – Le nombre d'itérations et les centres finaux (4 clusters)

Passons maintenant à la solution à 4 clusters. L'algorithme a convergé en 9 itérations, alors la solution est stable. La deuxième sortie de la figure 17.17 nous montre les centres finaux. Le premier cluster est assez semblable au premier cluster de la solution à 3 clusters : ce groupe a les valeurs les plus basses pour tous les aspects sauf pour la consommation, pour laquelle il a la valeur la plus élevée. Ce sont donc les petits véhicules

Distances between Final Cluster Centers					
Cluster	1	2	3	4	
1		6,255	6,083	2,930	
2	6,255		4,255	3,398	
3	6,083	4,255		3,986	
4	2,930	3,398	3,986		

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zpri	30,639	3	,409	148	74,953	,000
Ztaille mot	35,325	3	,314	148	112,380	,000
Zchevaux	34,550	3	,324	148	106,513	,000
Zempatt	27,398	3	,485	148	56,438	,000
Zlargeur	29,288	3	,434	148	67,457	,000
Zlongueur	27,244	3	,474	148	57,471	,000
Zpoids	34,154	3	,348	148	98,211	,000
Zgaz cap	32,064	3	,397	148	80,858	,000
Zconsomm	25,652	3	,511	148	50,214	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal

Number of Cases in each Cluster		
Cluster	1	2
1	52,000	
2		23,000
3		9,000
4		68,000
Valid		152,000
Missing		5,000

FIG. 17.18 – Distance entre les centres finaux, table ANOVA et répartition

qui consomment peu ; on les appelle ici les **compactes**. Ce cluster contient 52 véhicules.

Le deuxième cluster contient les véhicules qui consomment le plus et qui sont le plus gros. Ce sont les deuxièmes pour le prix et la puissance. On les nomme donc les **gros** véhicules ; il y en a 23.

Le troisième cluster contient les véhicules les plus chers et les plus puissants. Ça ne semble pas être de gros véhicules (deuxième valeur la plus faible pour l'empattement et la longueur). Ce groupe sera désigné par les **puissantes**. C'est le plus petit cluster avec 9 véhicules.

Finalement, le dernier cluster a des valeurs près de la moyenne pour tous les aspects.

On les nomme donc les **moyens**. C'est d'ailleurs le plus grand des 4 clusters (68 véhicules).

En analysant les distances de la première sortie de la figure 17.18 on voit que le premier cluster semble très différent des clusters 2 et 3. On voit que c'est le cluster 4 qui est le plus près des autres clusters, ce qui n'est pas étonnant puisque c'est le cluster « moyen ».

Pour ce qui est de la table ANOVA, une fois de plus toutes les p -values sont nulles. Si on compare les F avec ceux de la solution à 3 clusters, on constate que certains ont diminué tandis que d'autres ont augmentés. En fait on voit que l'importance donnée aux variables pour distinguer les groupes est mieux répartie. C'est encore la taille du moteur qui varie le plus d'un groupe à l'autre, mais cette fois-ci c'est le nombre de chevaux qui arrive deuxième, suivi du poids.

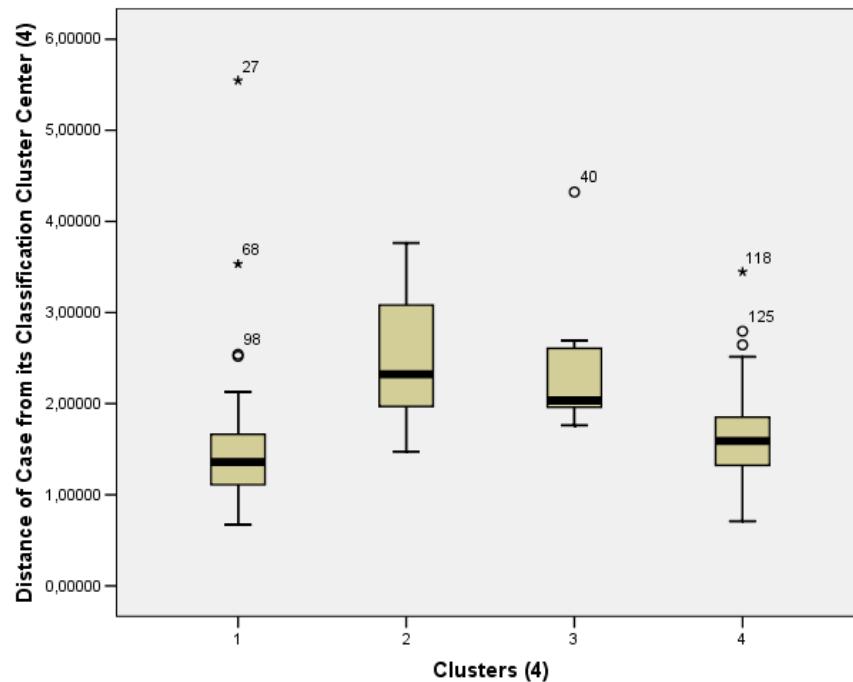


FIG. 17.19 – Variation dans les clusters

La figure 17.19 nous permet de visualiser la variation dans les clusters. Le fait d'avoir un quatrième clusters a permis de rendre les groupes plus homogènes, on voit qu'il y a

moins de variation que dans la solution à 3 clusters. Dans le cluster 3 (les puissantes) il y a très peu de variation, ce qui n'est pas étonnant puisque ce groupe ne compte que 9 véhicules. Il y a quand même le véhicule de la ligne 40 (Dodge Viper) qui se distingue ; ceci se comprend quand on voit qu'elle a un moteur de 8 litres et 450 chevaux...

Il semble que la solution à 4 clusters soit meilleure ; en effet, les clusters se distinguent bien les uns des autres et apportent plus de nuances que la solution à trois clusters. La solution à 5 clusters (pas présentée ici) a le défaut d'avoir un cluster ne contenant qu'un seul véhicule (c'est la Dodge Viper qui se retrouve seule).

Faire une segmentation à l'aide des clusters crée une nouvelle variable discrète qu'il est possible d'utiliser pour faire des analyses avec d'autres variables. Ici nous regarderons (de façon très informelle) le lien entre le type de véhicule (auto ou camion) et les clusters.

			type		Total	
			Automobile	Camion		
QCL_1	Petits véhicules	Count	57	6	63	
		Expected Count	46,4	16,6	63,0	
		% within type	50,9%	15,0%	41,4%	
		Std. Residual	1,6	-2,6		
	Véhicules moyens	Count	44	24	68	
		Expected Count	50,1	17,9	68,0	
		% within type	39,3%	60,0%	44,7%	
		Std. Residual	-.9	1,4		
	Gros véhicules	Count	11	10	21	
		Expected Count	15,5	5,5	21,0	
		% within type	9,8%	25,0%	13,8%	
		Std. Residual	-1,1	1,9		
Total		Count	112	40	152	
		Expected Count	112,0	40,0	152,0	
		% within type	100,0%	100,0%	100,0%	

FIG. 17.20 – Type de véhicule ⇒ clusters (3)

Les figures 17.20 et 17.21 présentent cette relation avec la solution à 3 clusters (qualifiée d'intéressante avec un Cramer's V de 0,333). On voit que les camions se retrouvent

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16,903 ^a	2	,000
Likelihood Ratio	18,217	2	,000
Linear-by-Linear Association	16,035	1	,000
N of Valid Cases	152		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 5,53.

Symmetric Measures		
	Value	Approx. Sig.
Nominal by Nominal	.333	,000
Nominal	Cramer's V	,333
N of Valid Cases	152	

- a. Not assuming the null hypothesis.
- b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 17.21 – Type de véhicule ⇒ clusters (3)

surtout dans les moyens et gros véhicules, ce qui est tout à fait logique.

L'analyse est plus intéressante avec la solution à 4 clusters (figures 17.22 et 17.23). D'ailleurs le Cramer's V a augmenté à 0,437. On voit que les « puissantes » sont des autos, et que les gros véhicules sont de façon marquée des camions.

Il serait intéressant d'examiner de plus près ces résultats (par exemple, quels sont les 5 camions qui se retrouvent avec les compactes ?).

			type		Total	
			Automobile	Camion		
QCL_3	Compactes	Count	47	5	52	
		Expected Count	38,3	13,7	52,0	
		% within type	42,0%	12,5%	34,2%	
		Std. Residual	1,4	-2,3		
Gros	Gros	Count	8	15	23	
		Expected Count	16,9	6,1	23,0	
		% within type	7,1%	37,5%	15,1%	
		Std. Residual	-2,2	3,6		
Puissantes	Puissantes	Count	9	0	9	
		Expected Count	6,6	2,4	9,0	
		% within type	8,0%	,0%	5,9%	
		Std. Residual	,9	-1,5		
Moyens	Moyens	Count	48	20	68	
		Expected Count	50,1	17,9	68,0	
		% within type	42,9%	50,0%	44,7%	
		Std. Residual	-,3	,5		
Total		Count	112	40	152	
		Expected Count	112,0	40,0	152,0	
		% within type	100,0%	100,0%	100,0%	

FIG. 17.22 – Type de véhicule ⇒ clusters (4)

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	28,980 ^a	3	,000
Likelihood Ratio	30,176	3	,000
Linear-by-Linear Association	2,056	1	,152
N of Valid Cases	152		

a. 1 cells (12,5%) have expected count less than 5. The minimum expected count is 2,37.

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	,437	,000
Nominal	Cramer's V	,437	,000
N of Valid Cases		152	

- a. Not assuming the null hypothesis.
- b. Using the asymptotic standard error assuming the null hypothesis.

FIG. 17.23 – Type de véhicule ⇒ clusters (4)

17.4 Exercices du chapitre

Exercice 1 Prenons la base de données `satisfactiontravail.sav`. Dans le cadre de l'étude de satisfaction au sein de cette entreprise, 18 questions ont été posées à propos de la satisfaction. On désire classifier les employés selon leur satisfaction. Pour ce faire utilisez d'abord la méthode hiérarchique avec une cinquantaine d'individus pour vous donner une idée du nombre de *clusters* qu'il vous faudra considérer. Ensuite, lorsque vous utiliserez les nuées dynamiques, ajoutez l'étape suivante aux commandes que vous devez effectuer : dans le bouton `Iterate...`, changez la valeur de `Maximum Iterations` : à 50 (au lieu de 10).

Exercice 2 Un important fournisseur industriel a mené une étude auprès de ses clients ; les résultats sont dans la base de données `fournisseur.sav`. Faites une classification des clients avec les variables continues, puis explorez les liens entre cette classification et les autres variables.

Exercice 3 Reprenons l'exercice 1 du chapitre 15. Le fournisseur de services de télécommunications aimerait maintenant cerner quels sont les consommateurs qui se désabonnent (par rapport aux services qu'ils utilisent). Pour ce faire, faites d'abord une analyse de classification (*clusters*) en utilisant les scores factoriels de l'exercice 1 du chapitre 15. Vous pouvez ainsi classifier les consommateurs selon les services qu'ils utilisent. Pour voir lesquels de ces groupes ont plus tendance à se désabonner, étudiez la relation *clusters* ⇒ `desabon` (la variable *clusters* est celle que vous créez suite à l'analyse de classification, et la variable `desabon` est une variable discrète dichotomique qui indique si le client s'est désabonné ou non au mois dernier).

Chapitre 18

Analyse discriminante

18.1 Introduction

Lire dans les boules de cristal n'appartient pas seulement aux médiums de ce monde. En effet, les banquiers, les investisseurs, les scientifiques et autres professionnels doivent être en mesure d'allouer des marges de crédit adéquates, de prédire le succès ou l'échec d'une entreprise, de classifier de nouvelles bactéries, etc... Souvent, ces prédictions, ou ces classements, sont effectués subjectivement suivant l'expérience de la personne en charge de solutionner le problème. Cependant, lorsque le problème grandit en complexité et lorsque les conséquences d'une mauvaise décision sont plus importantes, une méthode objective devient nécessaire.

L'analyse discriminante est une méthode objective qui permet de déterminer à quel groupe un objet est le plus susceptible d'appartenir. Une stratégie intuitive consiste à comparer les caractéristiques d'un nouvel objet avec des objets similaires déjà existants pour lesquels nous connaissons déjà le classement. Basée sur les similarités et les différences, une préiction (dans les faits, un classement) peut être établie. À titre d'exemple,

l'analyse discriminante est utilisée pour évaluer les nouvelles demandes de crédit. Ainsi, on qualifie le risque de crédit d'une nouvelle demande selon le profil de celle-ci ; si le profil de la nouvelle demande ressemble aux profils des demandes ayant déjà été classées comme étant un mauvais risque de crédit, la nouvelle demande sera classée comme étant un mauvais risque de crédit. À l'opposé, si le profil ressemble plutôt aux profils des demandes ayant déjà été classées comme étant un bon risque de crédit, la nouvelle demande sera classée comme étant un bon risque de crédit.

Pour établir le profil de chaque unité de l'échantillon et ainsi pouvoir la classer, il faut pouvoir disposer de certaines informations. Ces informations sont véhiculées par des variables continues, qui constitueront les variables indépendantes de l'analyse discriminante. Les groupes, eux, sont représentés par les modalités d'une variable discrète, qui constitue la variable dépendante. Par conséquent, la variable dépendante Y forme autant de groupes qu'elle possède de modalités (par exemple bon risque de crédit, moyen risque de crédit, mauvais risque de crédit) et la combinaison de l'information provenant des variables indépendantes est utilisée afin de faire correspondre chaque unité de l'échantillon à la modalité avec laquelle il a le plus d'affinités.

Pour arriver à faire cette classification, un ou des **scores discriminants** doivent être calculés à l'aide de **fonctions discriminantes** qui ont la forme suivante :

$$Z_{jk} = a + W_{j1}X_{1k} + W_{j2}X_{2k} + \cdots + W_{jm}X_{mk}$$

où

Z_{jk} = le score discriminant de la fonction j pour l'unité k ;

W_{ji} = le poids discriminant pour la variable indépendante X_i dans la fonction j ;

X_{ik} = valeur de la variable indépendante X_i pour l'unité k .

Ainsi, on calcule un score Z_{jk} à l'aide d'une expression formée des variables indépendantes et de coefficients (c'est une **combinaison linéaire** des variables indépendantes). On voit que cette équation ressemble à celle d'une régression linéaire. Les grandes différences résident ici dans le fait qu'au lieu d'établir des prédictions sur une variable

continue, on calcule des scores pour classifier les unités de l'échantillon dans les groupes formés par la variable discrète Y , et que les coefficients de la fonction sont calculés de façon à maximiser la différence entre les scores des unités n'appartenant pas à un même groupe. Par exemple, on veut que le score d'un individu qui a un mauvais risque de crédit soit très différent du score d'un individu qui a un bon risque de crédit.

La prochaine section présente deux exemples pour illustrer ce qu'est une analyse discriminante. Les sections subséquentes permettront de voir en détails quelles sont toutes les étapes d'une telle analyse.

18.2 Illustration du concept

Exemple 18.2.1 Voici un premier exemple pour illustrer le concept de l'analyse discriminante. Supposons qu'une entreprise veut déterminer si un produit sera un succès commercial ou non (un système de conservation sous vide). On aimerait entre autres identifier quels sont les consommateurs qui seraient intéressés à acheter ce nouveau produit. On fait donc une étude, et on demande la perception qu'ont les consommateurs à propos de trois facettes du produit : la facilité d'utilisation, la performance et le style, et s'ils sont intéressés à acheter ou non le produit. Les résultats sont dans le tableau de la figure 18.1.

Plusieurs questions se posent : est-ce que la connaissance de la perception des consommateurs par rapport à la facilité d'utilisation, la performance et le style du produit permet de savoir s'il est intéressé ou non à acheter le produit ? Si oui, laquelle ou lesquelles de ces perceptions permettent le plus de faire la différence entre un acheteur et un non-acheteur ?

On voit que 5 consommateurs ont indiqué vouloir acheter ce produit, et les 5 autres ne sont pas intéressés. Comparer les moyennes des réponses dans chacun des groupes pour chacune des perceptions permet de voir laquelle ou lesquelles des perceptions permet d'identifier les acheteurs. Tout d'abord, on voit que les acheteurs ont une moyenne de

Groupes selon l'intention d'achat	Facilité d'utilisation	Performance	Style
Groupe 1 : veulent acheter			
Sujet 1	8	9	6
Sujet 2	6	7	5
Sujet 3	10	6	3
Sujet 4	9	4	4
Sujet 5	4	8	2
Moyennes	7.4	6.8	4.0
Groupe 2 : ne veulent pas acheter			
Sujet 6	5	4	7
Sujet 7	3	7	2
Sujet 8	4	5	2
Sujet 9	2	4	3
Sujet 10	2	2	2
Moyennes	3.2	4.4	3.8
Différences entre les moyennes	4.2	2.4	0.2

Les variables ont des valeurs entre 1 = très pauvre et 10 = excellent

FIG. 18.1 – Les résultats

7,4 pour la facilité d'utilisation, tandis que les non-acheteurs ont une moyenne de 3,2 ; c'est la perception pour laquelle il y a la plus grande différence entre les moyennes, c'est donc elle qui semble pouvoir permettre une bonne discrimination entre les deux groupes. Pour la performance, les moyennes sont 6,8 et 4,4 respectivement ; il y a encore une discrimination, mais moins qu'avec la facilité d'utilisation. Finalement, pour le style, la différence entre les deux moyennes est petite (0,2 seulement de différence), il semble donc que cette perception ne permettra pas de bien identifier quels seront les acheteurs. Mais il faudra vérifier ceci de façon formelle ; une grande différence entre deux moyennes peut

en fait être non-significative s'il y a beaucoup de variation à l'intérieur de l'un ou des deux groupes.

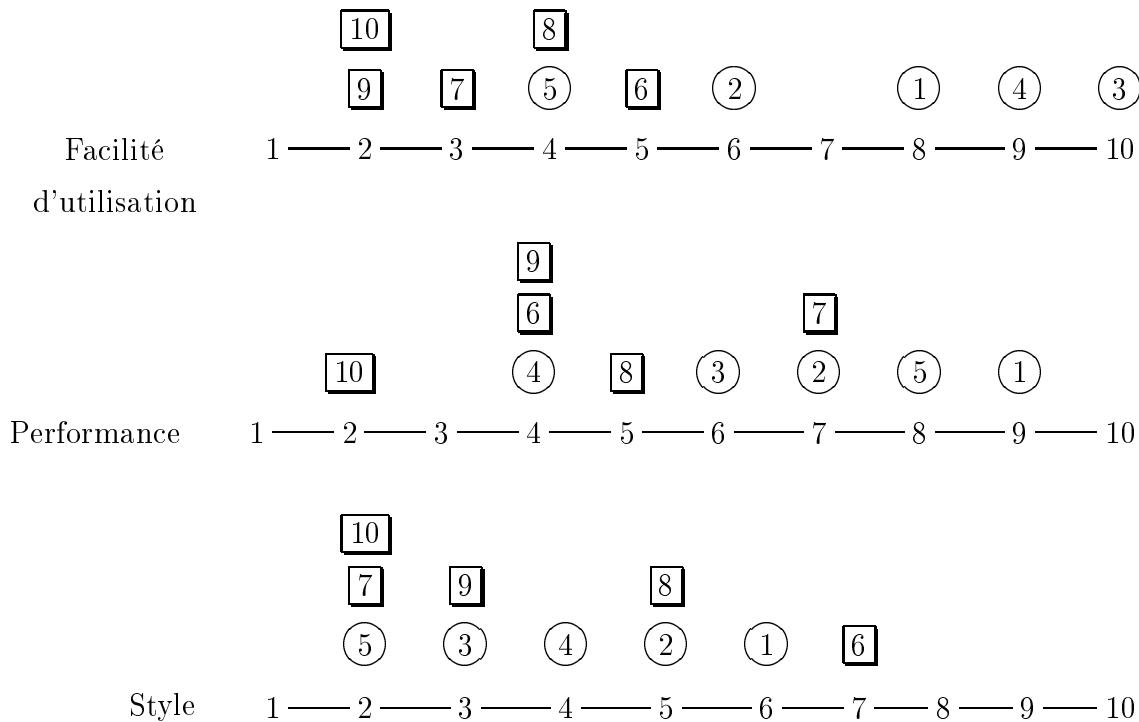


FIG. 18.2 – Représentation graphique des perceptions

Puisque nous sommes ici en présence d'un petit échantillon, on peut regarder les données dans un graphe pour visualiser l'effet des perceptions (figure 18.2). Les consommateurs intéressés à acheter le produit sont représentés par des cercles, tandis que les non-intéressés sont représentés par des carrés.

En regardant la représentation pour la facilité d'utilisation (celle qui a la plus grande différence de moyennes entre les deux groupes), on voit que l'on pourrait presque parfaitement séparer les deux groupes en n'utilisant que cette variable. Si on choisissait la valeur 5,5 pour séparer les deux groupes avec cette variable, il n'y a que l'individu 5 qui ne serait pas bien classé.

En regardant la représentation pour la performance, on voit que la séparation entre les deux groupes est moins claire. Par contre, l'individu 5 serait certainement classé cor-

rectement avec cette variable. Donc une combinaison des valeurs de ces deux perceptions (facilité d'utilisation et performance) permettrait sûrement de classer parfaitement tous les individus. D'où l'idée d'utiliser des fonctions discriminantes. La fonction discriminante pour cet exemple pourrait ne contenir que les deux premières variables puisque la variable du style ne semble pas faire de discrimination entre les groupes.

En fait la procédure pour calculer la ou les fonctions discriminantes fonctionne selon le principe ici illustré : elle identifie les variables permettant de faire une bonne discrimination entre les groupes, et calcule des coefficients pour faire une fonction qui reflète ces différences.

Lorsque la variable discrète ne possède que deux modalités, une seule fonction discriminante est calculée (comme dans l'exemple précédent avec les acheteurs et les non-acheteurs). Lorsque la variable discrète possède plus de deux modalités, disons l modalités, alors il y a habituellement $l - 1$ fonctions discriminantes (sauf si le nombre de variables indépendantes m est plus petit que $l - 1$, à ce moment il y aura m fonctions discriminantes). L'exemple qui suit illustre ce qui se passe avec une classification à trois groupes.

Exemple 18.2.2 Un fournisseur a fait une petite étude (en fait un pré-test) auprès de 15 clients d'un compétiteur. On leur demande d'évaluer leur satisfaction par rapport aux prix et au service du compétiteur, et d'indiquer s'il ont l'intention de changer de fournisseur (oui, indécis, non). Les résultats sont dans le tableau de la figure 18.3.

On se demande s'il est possible de prédire l'intention de changer ou non de fournisseur selon la satisfaction par rapport au prix et par rapport au service. Puisqu'on a trois groupes pour la classification, une analyse discriminante produirait 2 fonctions discriminantes.

Mais commençons par examiner ce qui se passe dans chacun des groupes. Si on regarde les moyennes, la satisfaction par rapport aux prix semble distinguer le groupe 1 (ceux qui veulent changer) des deux autres ; la différence entre ceux qui sont indécis et ceux qui ne

Groupes selon leur intention	Prix	Service
Groupe 1 : veulent changer		
Sujet 1	2	2
Sujet 2	1	2
Sujet 3	3	2
Sujet 4	2	1
Sujet 5	2	3
Moyennes	2.0	2.0
Groupe 2 : indécis		
Sujet 6	4	2
Sujet 7	4	3
Sujet 8	5	1
Sujet 9	5	2
Sujet 10	5	3
Moyennes	4.6	2.2
Groupe 3 : ne veulent pas changer		
Sujet 11	2	6
Sujet 12	3	6
Sujet 13	4	6
Sujet 14	5	6
Sujet 15	5	7
Moyennes	3.8	6.2

Les variables ont des valeurs entre 0 = très insatisfait et 10 = très satisfait

FIG. 18.3 – Les résultats

veulent pas changer n'est pas très grande, et ne va pas dans le sens auquel on pourrait s'attendre.

Pour ce qui est de la satisfaction par rapport au service, elle permet de distinguer ceux qui ne veulent pas changer des deux autres groupes. Une représentation graphique de ceci est donnée dans la figure 18.4. Ceux qui veulent changer sont représentés par des cercles, les indécis par des carrés, et ceux qui ne veulent pas changer par des carrés avec un cadre double.

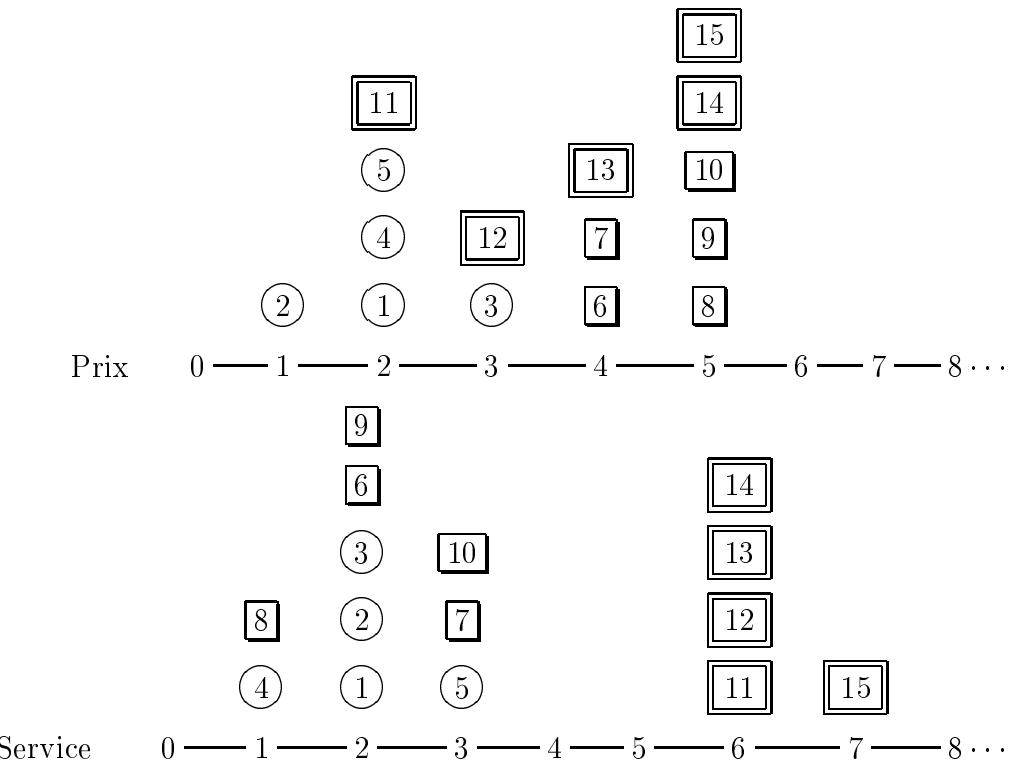


FIG. 18.4 – Représentations graphiques

À des fins d'illustration, supposons que les deux fonctions discriminantes pour cet exemple sont les suivantes :

$$Z_1 = 1 \cdot X_{\text{Prix}} + 0 \cdot X_{\text{Service}}$$

$$Z_2 = 0 \cdot X_{\text{Prix}} + 1 \cdot X_{\text{Service}}$$

On obtient alors la représentation graphique de la figure 18.5. Le fait de faire la discrimination avec deux scores au lieu d'un seul rend les choses très claires ici; pour faire la classification il suffirait de trancher selon les lignes pointillés.

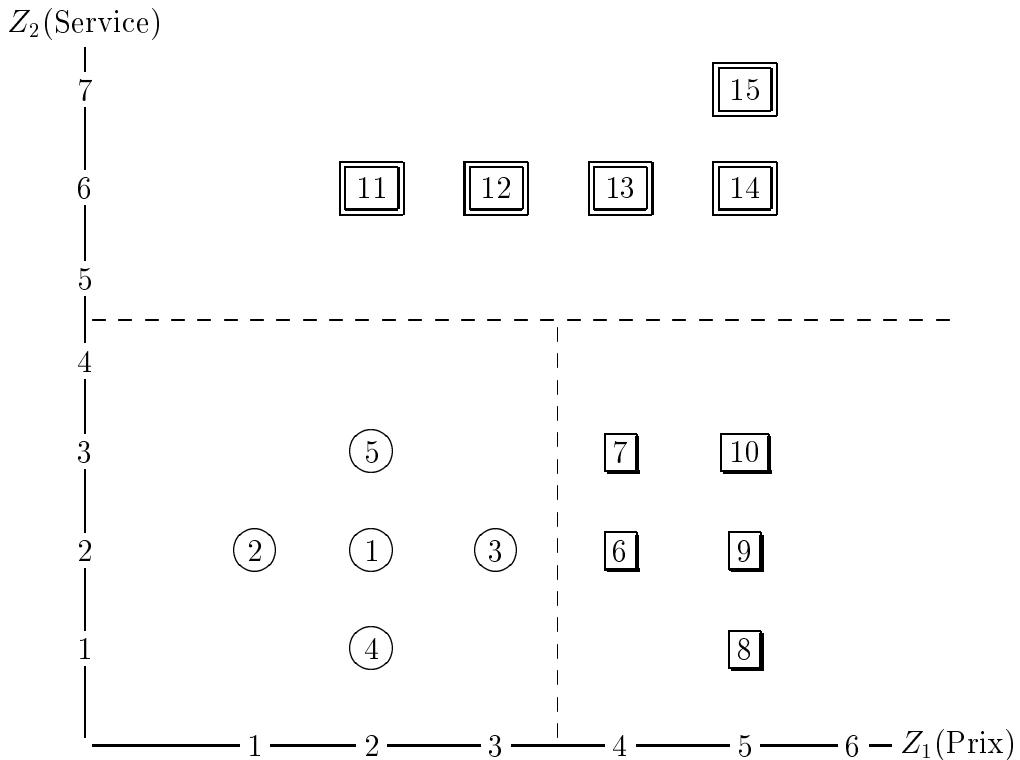


FIG. 18.5 – Représentation avec les deux fonctions

Dans la pratique l'analyse discriminante est bien entendu plus complexe que ceci. Les prochaines sections sont consacrées à l'étude de toutes les étapes nécessaires pour une bonne analyse discriminante.

18.3 Un exemple à deux groupes

Pour voir en détails les étapes d'une analyse discriminante, nous prendrons la base de données `telco.sav` qui a déjà été utilisée dans les chapitres précédents. Nous appliquerons

rons ici l'analyse discriminante pour tenter d'établir le profil des consommateurs qui se désabonnent versus ceux qui ne se désabonnent pas. Si l'analyse donne de bons résultats, la fonction discriminante pourra être utilisée pour identifier, dans le futur, les clients qui sont plus à risque de se désabonner.

Donc la variable dépendante Y est ici `desabon`, une variable binaire qui indique si le client s'est désabonné au dernier mois ou non.

Les variables indépendantes choisies sont `longdistm` (longue distance dernier mois), `sfracism` (numéro sans frais dernier mois), `equipm` (équipement dernier mois), `cartem` (carte d'appel dernier mois) et `sansfilm` (sans fil dernier mois). Pour s'assurer que chacune de ces variables ait le même poids dans l'analyse, elles seront d'abord standardisées.

Il est fortement recommandé, lorsque la taille de l'échantillon le permet, de diviser l'échantillon en deux : une partie pour faire l'analyse (**échantillon d'analyse**), et une autre partie pour valider celle-ci (**échantillon de validation**). Sans être une règle formelle, il est recommandé d'avoir une dizaine d'individus par variable à l'étude pour l'échantillon d'analyse. Ainsi, pour notre étude, il y a six variables en cause. Donc, un échantillon d'analyse d'une soixantaine d'unités serait suffisant. La base de données `telco.sav` contient les informations sur 1 000 individus, donc nous sommes largement au-dessus du minimum.

Cependant, il faut faire attention à ne pas verser dans l'excès d'individus. En effet, lorsque les tailles d'échantillons sont très grandes, il s'y présente un phénomène de corrélation significative un peu partout alors qu'il n'y en a pas vraiment. Ce phénomène se répercute dans l'analyse discriminante, qui utilise une matrice de covariances, qui est, malheureusement, influencée par les corrélations entre les variables. À titre d'ordre de grandeur, nous recommandons d'utiliser une taille d'échantillon d'analyse variant de 8 à 30 individus pour chacune des variables incluses dans le modèle. Dans le cadre de cet exemple, nous divisons l'échantillon de la façon suivante : 70 % pour l'échantillon d'analyse, 30 % pour l'échantillon de validation. Un bon exercice consisterait à refaire l'analyse sur un plus petit échantillon d'analyse et à comparer les résultats.

Voici comment procéder pour diviser l'échantillon en deux : il faut d'abord créer une variable **test** qui de façon aléatoire vaudra 1 pour les individus faisant partie de l'échantillon d'analyse, et 0 pour les autres. Il faut d'abord faire les commandes suivantes :

Menu SPSS : → **Transform**
→ **Random Number Generators...**

Dans la fenêtre **Active**

Generator Initialization : **Set Starting Point**
 Fixed Value puis tapez
9191972 dans la fenêtre **Value**.

Ces commandes vont vous permettre de créer exactement la même variable **test** que celle de l'exemple. Pour la créer, allez dans **Transform** puis **Compute...**, puis créez la variable **test** en tapant la fonction **rv.bernoulli(0.7)** (voir la figure 18.6).

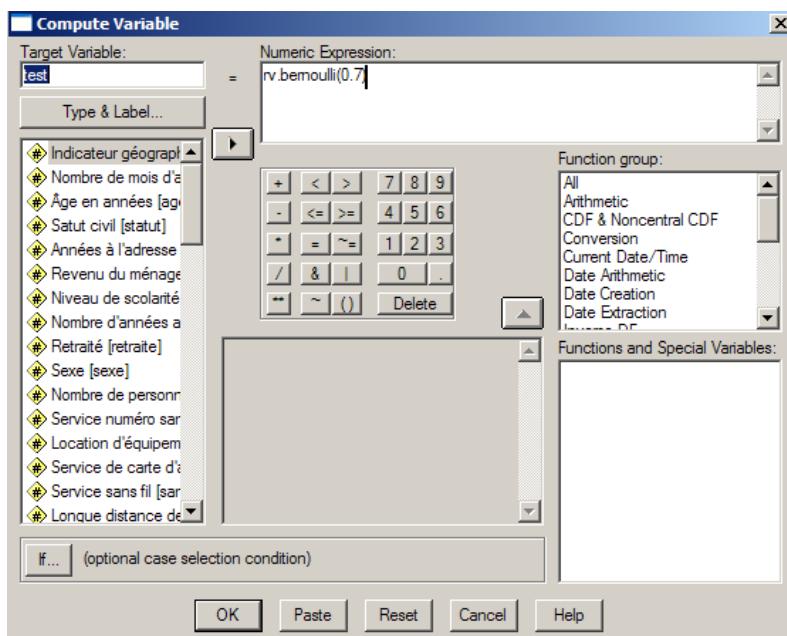


FIG. 18.6 – Pour créer la variable **test**

Ensuite, pour obtenir les sorties présentées pour cet exemple, les commandes sont :

Menu SPSS :	→ Analyse
	→ Classify
	→ Discriminant...
Dans la fenêtre Grouping Variable :	→ desabon
Dans le bouton Define Range... :	Minimum : 0 Maximum : 1
Dans la fenêtre Independents :	→ Zlongdistm, Zsfraism, Zequipm, Zcartem, Zsansfilm
Dans le bouton Statistics... :	Descriptives ✓ Means ✓ Univariate ANOVAs ✓ Box's M Function Coefficients ✓ Unstandardized Matrices ✓ Within-groups correlation
Dans le bouton Classify... :	Prior Probabilities ✓ Compute from group sizes Display ✓ Casewise results ✓ Summary table Use Covariance Matrix ✓ Within-groups
Dans le bouton Save... :	✓ Predicted group membership ✓ Discriminant scores ✓ Probabilities of group membership
Dans la fenêtre Selection Variable :	→ test
Puis dans le bouton Value... indiquez 1 dans Value for Selection Variable	

18.3.1 Survol des données

La première étape d'une analyse discriminante consiste à faire un survol des données à inclure dans l'analyse. En effet, les logiciels statistiques écartent des analyses les individus pour lesquels il manque une quelconque valeur sur l'une des variables à l'étude. Ainsi, si un fichier présente plusieurs valeurs manquantes, l'analyse sera effectuée sur un nombre plus restreint d'individus ayant toutes leurs valeurs.

Il est clair que si les valeurs manquantes appartiennent toutes à une catégorie importante d'individus, l'analyse discriminante ne tiendra pas compte de cette différence dans les calculs. C'est pourquoi il est très important de survoler l'ensemble des données et d'apporter une attention toute particulière aux individus ayant des données manquantes. Aussi, si une variable possède trop de valeurs manquantes, il faudra considérer la possibilité de l'éliminer des analyses. Cependant, enlever une variable peut parfois être critique si elle se situe au cœur de la problématique. Une autre stratégie consiste à imputer les valeurs manquantes par la moyenne. Cette stratégie possède l'avantage de récupérer l'ensemble des individus. Elle possède cependant l'inconvénient de réduire la variance des variables qui ont beaucoup de valeurs manquantes, ce qui devient très « nocif ». En effet, plus il y a de variance, plus la discrimination est effective.

Analysis Case Processing Summary			
Unweighted Cases		N	Percent
Valid		702	70,2
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Unselected	298	29,8
	Total	298	29,8
	Total	1000	100,0

FIG. 18.7 – Première sortie de l'analyse discriminante

Justement, la première sortie de l'analyse (figure 18.7) montre s'il y a des données

manquantes. Ici on voit qu'il n'y a aucune donnée manquante. On voit aussi que 702 cas sont utilisés pour établir l'analyse, et 298 ne sont pas sélectionnés et serviront à valider l'analyse.

18.3.2 Analyse des différences entre les groupes

Puisqu'on espère que les moyennes des variables indépendantes soient différentes d'un groupe à l'autre, il est tout à fait logique de jeter un œil aux statistiques descriptives (figure 18.8) pour avoir une première idée de ce qui se passe entre les groupes.

Group Statistics					
desabon		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
Non	Zlongdistm	,1566040	,1,06461304	508	508,000
	Zsfraism	,0524160	,99936783	508	508,000
	Zequipm	-,1690347	,96194485	508	508,000
	Zcartem	,1087366	1,01037531	508	508,000
	Zsansfilm	-,0441650	,99563232	508	508,000
Oui	Zlongdistm	-,4535310	,42030359	194	194,000
	Zsfraism	-,1622153	,78949333	194	194,000
	Zequipm	,3928744	,98268163	194	194,000
	Zcartem	-,2539035	,98325845	194	194,000
	Zsansfilm	,1503106	1,02457366	194	194,000
Total	Zlongdistm	-,0120088	,97104293	702	702,000
	Zsfraism	-,0068981	,95035334	702	702,000
	Zequipm	-,0137493	,99917619	702	702,000
	Zcartem	,0085198	1,01531008	702	702,000
	Zsansfilm	,0095790	1,00674783	702	702,000

FIG. 18.8 – Statistiques descriptives

Tout d'abord, on voit que sur 702 individus, 194 se sont désabonnés au dernier mois. Ceux qui se sont désabonnés au dernier mois ont utilisés moins que les autres les longues distances, les numéros sans frais et les cartes d'appel, mais ont de plus grands montants pour l'équipement et le sans fil. Les écarts les plus grands se retrouvent avec les variables Zcartem, Zlongdistm et Zequipm.

Il est possible de voir si ces différences sont significatives à l'aide de la sortie de la figure 18.9. En effet, cette sortie contient la statistique F qui est le quotient des carrés moyens associés à la décomposition de la variance à un facteur (*One-Way ANOVA*) de

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
Zlongdistm	,921	60,097	1	700	,000
Zsfraism	,990	7,224	1	700	,007
Zequipm	,937	47,334	1	700	,000
Zcartem	,974	18,353	1	700	,000
Zsansfilm	,993	5,271	1	700	,022

FIG. 18.9 – Statistiques descriptives

chacune des variables par rapport aux deux groupes, et la *p*-value du test d'hypothèses sur l'égalité des moyennes. Dans le cas présent on voit qu'au seuil de 5 % toutes les moyennes sont considérées comme étant significativement différentes pour ceux qui se sont désabonnés et ceux qui ne se sont pas désabonnés.

De plus, en regardant les valeurs des *F*, il semble que ce soit la variable des longues distances qui différencie le mieux les groupes, tandis que la variable pour le service sans fil est celle qui différencie le moins.

La statistique *Wilks' Lambda* représente le pourcentage de la variation de la variable qui n'est pas expliquée par la variable **desabon**. Ces pourcentages variant entre 92,1 % et 99,3 %, on voit que le lien entre la variable **desabon** et les variables continues de cette analyse prises individuellement n'est pas très fort. Mais il faut se rappeler que c'est le lien entre toutes les variables continues et la variable discrète qui nous intéresse, on ne peut donc pas se borner à l'examen des ANOVA individuelles pour se faire une idée complète de la qualité de l'analyse.

Comme l'interdépendance entre les variables affecte toute analyse multivariée, il est intéressant, en analyse discriminante, d'examiner de plus près les corrélations entre les variables indépendantes. Lorsque deux variables sont en très grande corrélation, elles affectent la grandeur et le signe des coefficients de la fonction discriminante. Compte tenu que l'analyse discriminante étudie la séparation entre les groupes formés par la variable

dépendante, une matrice de corrélations appelée **Pooled Within-Groups Matrices** est justement utilisée pour étudier l'importance des corrélations.

Pooled Within-Groups Matrices						
	Zlongdistm	Zsfraism	Zequipm	Zcartem	Zsansfilm	
Correlation	Zlongdistm	1,000	,230	,067	,385	,159
	Zsfraism	,230	1,000	,103	,340	,485
	Zequipm	,067	,103	1,000	,108	,559
	Zcartem	,385	,340	,108	1,000	,306
	Zsansfilm	,159	,485	,559	,306	1,000

FIG. 18.10 – Les moyennes des corrélations calculées selon les deux groupes de `desabon`

La **Pooled Within-Groups Matrices** est obtenue en deux étapes. Dans un premier temps, des matrices de corrélations sont calculées, une pour chaque groupe de la variable dépendante `desabon`, donc ici deux matrices. Ces deux matrices sont ensuite combinées en une seule matrice, justement la **Pooled Within-Groups Matrices**, en calculant, pour chacune des paires de variables dans notre cas, la moyenne des corrélations des deux matrices. La figure 18.10 contient la matrice **Pooled Within-Groups Matrices**.

En observant les corrélations, il semble que les variables `equipm` et `sansfilm` sont celles qui corrèlent le plus entre elles (0,559). Cette valeur peut paraître petite, mais en sciences humaines, une telle valeur est intéressante. Mentionnons que cette matrice combinée peut être très différente de la matrice des corrélations totales usuelles. En effet, la matrice des corrélations totales calcule les corrélations de Pearson entre les variables indépendantes lorsque toutes les observations sont prises en considération en même temps (les groupes de la variable dépendante n'interviennent pas). La figure 18.11 présente la matrice des corrélations totales usuelles entre les variables indépendantes. Rappelons que cette matrice n'est pas utilisée par l'analyse discriminante.

On voit qu'il y a certaines différences entre les deux matrices ; certaines variables corrèlent mieux entre elles quand on ne tient pas compte des groupes (par exemple la variable des longues distances corrèle mieux avec trois des quatre autres variables

Correlations					
	Zlongdistrm	Zsfraism	Zequipm	Zcartem	Zsansfilm
Pearson Correlation	1 .248** .009 .410** .128**	,248** 1 .074 .350** .472**	-,009 .074 1 .063 .561**	,410** .350** .063 1 .287**	,128** .472** .561** .287** 1
Sig. (2-tailed)	Zlongdistrm .000 .814 .000 .001	,000 .051 .000 .000	,814 .051 .097 .000	,000 .000 .097 .000	,001 .000 .000 .000
N	Zlongdistrm 702 Zsfraism 702 Zequipm 702 Zcartem 702 Zsansfilm 702	702 702 702 702 702	702 702 702 702 702	702 702 702 702 702	702 702 702 702 702

**. Correlation is significant at the 0.01 level (2-tailed).

FIG. 18.11 – Les corrélations calculées sur tout l'échantillon d'analyse

lorsqu'on ne tient pas compte des groupes), tandis qu'on a le phénomène inverse pour d'autres paires de variables.

Compte tenu que l'analyse discriminante tente de maximiser la séparation entre les deux groupes, la matrice **Pooled Within-Groups Matrices** est utilisée dans les calculs de l'analyse.

18.3.3 Estimation des coefficients de la fonction discriminante

Les statistiques descriptives et les analyses ANOVA univariées sont utiles pour identifier les variables qui, seules, « séparent » bien les groupes. Cependant, dans l'analyse discriminante, l'emphase est mise sur l'analyse des variables combinées et non prises séparément. En considérant les variables simultanément, l'analyste sera en mesure d'inclure de l'information importante concernant leurs interrelations.

Lorsqu'on effectue une analyse discriminante, une combinaison linéaire des variables indépendantes est calculée et cette combinaison est ensuite utilisée pour assigner les observations à l'un des groupes de la variable dépendante. De cette façon, toute l'informa-

mation contenue dans les variables indépendantes est résumée par le score discriminant z (qui estime le vrai score Z). Par exemple, en formant une combinaison pondérée des variables indépendantes `Zlongdistm`, `Zsfraism`, `Zequipm`, `Zcartem` et `Zsansfilm`, nous espérons obtenir un score z qui distingue les individus qui se désabonnent de ceux qui ne se désabonnent pas. Tout comme pour la régression multiple, les coefficients de pondération W_j sont estimés (par les w_j) pour assurer la meilleure séparation entre les groupes. Dans l'exemple présent, il n'y a qu'une seule fonction discriminante, et elle s'écrit

$$z = a + w_1 X_{\text{Zlongdistm}} + w_2 X_{\text{Zsfraism}} + w_3 X_{\text{Zequipm}} + w_4 X_{\text{Zcartem}} + w_5 X_{\text{Zsansfilm}}.$$

Si cette fonction linéaire discriminante est en mesure de bien distinguer les deux groupes de la variable dépendante `desabon`, le score moyen dans les groupes devrait être significativement différent. À la différence des coefficients d'une régression, les w_j sont calculés de façon à maximiser la différence des valeurs de la fonction discriminante dans les deux groupes afin de séparer le mieux possible ces deux groupes. En fait les coefficients w_j maximisent le ratio de la variation entre les groupes sur la variation à l'intérieur des groupes :

$$\frac{\text{Variation entre les groupes}}{\text{Variation à l'intérieur des groupes}}.$$

Toute autre combinaison linéaire des variables dépendantes aura un ratio plus petit. La figure 18.12 contient les coefficients de la fonction discriminante de l'exemple.

Canonical Discriminant Function Coefficients	
	Function
	1
<code>Zlongdistm</code>	,708
<code>Zsfraism</code>	,135
<code>Zequipm</code>	-,694
<code>Zcartem</code>	,181
<code>Zsansfilm</code>	-,056
(Constant)	-,001

Unstandardized coefficients

FIG. 18.12 – Les coefficients de la fonction discriminante

Ainsi la fonction s'écrit

$$z = -0,001 + 0,708X_{\text{Zlongdistm}} + 0,135X_{\text{Zsfraism}} - 0,694X_{\text{Zequipm}} + 0,181X_{\text{Zcartem}} \\ - 0,056X_{\text{Zsansfilm}}.$$

D'après cette équation, c'est la variable `Zlongdistm` qui favorise le plus la discrimination entre les groupes, et c'est la variable `Zsansfilm` qui fait le moins la différence. En fait l'ordre ici respecte celui observé avec les ANOVA individuelles, mais ce n'est pas toujours le cas. L'interprétation des ces coefficients n'est pas toujours fiable, surtout s'il y a de la multicolinéarité.

Si les variables n'ont pas été standardisées au préalable, il est alors préférable d'interpréter les coefficients standardisés.

Il existe une autre façon d'évaluer l'importance des variables qui est moins influencée par la multicolinéarité, et elle se fait à l'aide de la sortie suivante :

Structure Matrix	
	Function
	1
<code>Zlongdistm</code>	,706
<code>Zequipm</code>	-,626
<code>Zcartem</code>	,390
<code>Zsfraism</code>	,245
<code>Zsansfilm</code>	-,209

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

FIG. 18.13 – Les *loadings*

Les *loadings* sont simplement les corrélations entre les variables indépendantes et la fonction discriminante. Par exemple, la variable `Zlongdistm` a une corrélation de 0,706 avec la fonction discriminante. Ainsi les *loadings* reflètent la variance que les variables indépendantes partagent avec la fonction discriminante. Les *loadings* sont de plus en plus utilisés pour l'interprétation de la fonction discriminante.

18.3.4 La classification

Pour chaque individu j , un score z_j est calculé à l'aide de la fonction discriminante en remplaçant simplement chaque variable indépendante par la valeur correspondant à cet individu. Dans la base de données, les scores discriminants apparaissent dans la colonne de la variable `Dis1_1`.

En se basant sur les scores discriminants, il est possible de dégager une règle de classification des observations dans les groupes. Une technique utilisée est la règle de Bayes, une formule pour calculer une probabilité conditionnelle dans une situation où l'on connaît d'autres probabilités conditionnelles. Plus précisément, la probabilité qu'une observation j appartienne au groupe G_l sachant que son score discriminant est Z_j est déterminée par :

$$P(G_l|Z_j) = \frac{P(Z_j|G_l)P(G_l)}{\sum_{i=1}^g P(Z_j|G_i)P(G_i)}.$$

où g représente le nombre de groupes créés par la variable dépendante (donc dans le cadre de l'exemple on a $g = 2$).

La probabilité a priori, représentée par $P(G_i)$, est une estimation de la probabilité qu'un individu appartienne au groupe i lorsqu'aucune information sur cet individu n'est encore disponible. Par exemple, si 30 % des enfants ayant des problèmes respiratoires infantiles décèdent, alors, a priori, un nouveau-né qui connaît ce type de problème a 30 % de risque de mourir.

La probabilité a priori peut être évaluée de différentes manières. Par exemple, si l'échantillon est représentatif de la population, les proportions observées dans l'échantillon peuvent servir à titre de probabilité a priori. C'est ce qu'on a fait dans l'exemple, on a coché `Compute from group sizes` dans `Prior Probabilities`; on voit ainsi dans la sortie 18.14 que d'après l'échantillon, la probabilité d'appartenir au groupe des consommateurs qui ne se sont pas désabonnés est de 72,4 %, et celle d'appartenir à l'autre groupe est donc de 27,6 %.

Prior Probabilities for Groups			
desabon	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Non	,724	508	508,000
Oui	,276	194	194,000
Total	1,000	702	702,000

FIG. 18.14 – Les probabilités a priori

Cependant, dans le cas de maladies rares, disons un cas sur 1000, il est clair qu'un échantillon représentatif aboutira à un nombre très petit de cas à étudier. Ainsi, l'analyste peut inclure le même nombre de cas dans chaque groupe. Dans cette situation, la probabilité a priori peut être estimée différemment, soit à l'aide des archives d'un hôpital, d'un expert, etc. Il n'est cependant pas possible d'indiquer à SPSS une telle probabilité a priori.

Quand les groupes sont égaux, ou qu'aucune information n'est disponible sur le partage des observations dans les groupes, l'option **All groups equal** dans **Prior Probabilities** peut être sélectionnée. Dans tous les cas, comme toutes les observations doivent appartenir à un seul groupe, la somme des probabilités doit être égale à 1 (obtenant ainsi une loi de probabilité).

La probabilité conditionnelle $P(Z_j|G_i)$ représente la probabilité que l'individu j obtienne un score Z_j lorsqu'on sait qu'il appartient au groupe i . En se basant sur les observations, et en supposant que les valeurs Z_j sont distribuées normalement dans chaque groupe, la probabilité $P(Z_j|G_i)$ peut être estimée.

Finalement, La probabilité $P(G_l|Z_j)$ représente la probabilité qu'une observation j appartienne au groupe l lorsqu'on connaît son score discriminant Z_j .

C'est la probabilité a posteriori ; elle représente la probabilité qu'un individu appar-

tienne à un groupe lorsqu'on a une certaine information à son sujet ; ici cette information est véhiculée par le score discriminant. Ainsi, une observation est classée dans le groupe pour laquelle sa probabilité a posteriori est la plus grande.

Casewise Statistics											
Case Num ber	Actual Group	Predicted Group	Highest Group				Second Highest Group			Discriminant Scores	
			P(D>d G=g)		P(G=g D=d)	Squared Mahalanobis Distance to Centroid	Group	P(G=g D=d)	Squared Mahalanobis Distance to Centroid	Function 1	
			p	df							
Original	1 ^u	1	0**	.659	1	.728	.195	1	.272	.236	-.185
	2	1	0**	.818	1	.765	.053	1	.235	.486	.026
	3 ^u	0	0	.326	1	.909	.964	1	.091	3.644	1.238
	4	1	0**	.884	1	.779	.021	1	.221	.611	.111
	5	0	0	.718	1	.742	.130	1	.258	.321	-.104
	6	0	0	.732	1	.847	.117	1	.153	1.612	.599
	7	1	0**	.947	1	.811	.004	1	.189	.987	.322

FIG. 18.15 – Extrait de la sortie de classification

Les probabilités a posteriori sont sauvegardées dans la base de données (variables `Dis1_2` et `Dis2_2` lorsqu'il y a deux groupes ; ce sont les probabilités d'appartenir au premier groupe et au deuxième groupe respectivement), et apparaissent aussi dans la sortie de classification (voir figure 18.15). On voit par exemple dans la colonne `Actual Group` que l'individu 3 appartient au groupe 0 (donc il ne s'est pas désabonné au dernier mois). Cet individu a été correctement classé puisque c'est le groupe 0 qui apparaît dans la colonne `Predicted Group`. S'il avait été mal classé, deux astérisques (**) apparaîtraient à côté de la prédiction. La colonne `P(D>d | G=g)` contient une des probabilités conditionnelles qui sont estimées pour calculer la probabilité a posteriori d'appartenir à un groupe lorsqu'on connaît le score discriminant.

Cette probabilité apparaît dans la colonne `P(G=g | D=d)`. Ainsi, lorsqu'on connaît son score discriminant, on voit que la probabilité que l'individu 3 appartienne au groupe 0 a été estimée à 90,9 %. Ainsi la probabilité qu'il appartienne au groupe 0 a été estimée à 9,1 % (cette probabilité apparaît dans la colonne `P(G=g | D=d)` du `Second Highest Group`).

On note aussi que lorsqu'un individu fait partie de l'échantillon de validation, alors

un petit `u` apparaît à côté de son `Case Number`; c'est le cas par exemple des individus 1 et 3.

Les colonnes `Squared Mahalanobis Distance to Centroid` contiennent une distance calculée entre l'individu et le centre moyen des groupes calculés à partir de la fonction discriminante. Ainsi l'individu 3 est à une distance de 0,964 du centre moyen du groupe 0, et à une distance de 3,644 du centre moyen du groupe 1.

Finalement, la dernière colonne (`Discriminant Scores`) contient le score discriminant de chaque individu.

18.3.5 Le sommaire de la classification

La sortie de classification peut être résumée par la matrice de classification, parfois appelée matrice de confusion (figure 18.16). Cette matrice carrée compare le nombre d'observations bien classées au nombre mal classées. Il va de soi que plus le % d'observations bien classées est élevé, plus l'analyse est efficiente. Mais, tout comme le coefficient r^2 d'une régression linéaire multiple, le pourcentage d'observations bien classées issu de l'échantillon d'analyse surestime la performance réelle de la fonction de discrimination. C'est pourquoi il est important de valider la performance du modèle sur l'échantillon de validation.

En somme, si nous utilisons un échantillon d'analyse et un échantillon de validation, l'analyste obtiendra deux matrices de classification (dans une seule sortie comme la 18.16) et donc deux pourcentages d'observations bien classées. Ces valeurs permettront à l'analyste de se faire une idée plus précise de la performance de son modèle de discrimination.

Dans le cadre de l'exemple, on voit que sur l'échantillon d'analyse, 75,5 % des cas ont été bien classés, tandis que dans l'échantillon de validation on a une performance de 75,8 %. Il semble donc ici que la performance n'a pas été surestimée sur l'échantillon d'analyse.

				Classification Results ^{a,b}		Total	
				Predicted Group Membership			
				Non	Oui		
Cases Selected	Original	Count	Non	474	34	508	
			Oui	138	56	194	
		%	Non	93,3	6,7	100,0	
	Count	Oui	71,1	28,9	100,0		
		Non	199	19	218		
		Oui	53	27	80		
	%	Non	91,3	8,7	100,0		
		Oui	66,3	33,8	100,0		

a. 75,5% of selected original grouped cases correctly classified.

b. 75,8% of unselected original grouped cases correctly classified.

FIG. 18.16 – La matrice de classification

Est-ce que ces pourcentages sont satisfaisants ? Examinons plus en détails les matrices de classification. Sur l'échantillon d'analyse, on voit dans la dernière colonne qu'il y a en réalité 508 individus qui ne se sont pas désabonnés, et il y en a 194 qui se sont désabonnés. Sur les 508 qui ne se sont pas désabonnés, 474 ont été correctement classés, ce qui représente 93,3 % de bien classés pour ce groupe, ce qui est excellent. Le hic, c'est que pour le groupe des individus qui se sont désabonnés, seulement 56 sur 194 ont été bien classés, ce qui ne représente que 28,9 % de réussite pour ce groupe. Or pour le fournisseur de services de télécommunications, ce groupe est d'une grande importance ; l'analyse ne semble donc pas satisfaisante.

En regardant la matrice de l'échantillon de validation on retrouve sensiblement le même phénomène : 91,3 % des individus du groupe des non-désabonnés ont été correctement classés, tandis que pour les désabonnés ce pourcentage tombe à 33,8 %.

Étant donné qu'au départ, 72,4 % des individus ne se sont pas désabonnés (probabilités a priori qui sont simplement les proportions présentes dans l'échantillon), l'analyse qui consisterait à classer tout le monde dans ce groupe obtiendrait un pourcentage de classification de 72,4 %, ce qui n'est pas beaucoup moins que les pourcentages obtenus. Ce raisonnement est un peu exagéré en ce sens qu'on obtient alors 0 % de réussite pour

le groupe des désabonnés, mais montre qu'il faut justement se méfier du résultat global lorsqu'un des deux groupes est vraiment plus grand que l'autre.

				Classification Results ^{a,b}		
				Predicted Group Membership		
				Non	Oui	Total
Cases Selected	Original	Count	Non	362	146	508
			Oui	67	127	194
		%	Non	71,3	28,7	100,0
	Cases Not Selected	Original	Oui	34,5	65,5	100,0
			Non	145	73	218
		%	Oui	25	55	80
		%	Non	66,5	33,5	100,0
			Oui	31,3	68,8	100,0

a. 69,7% of selected original grouped cases correctly classified.

b. 67,1% of unselected original grouped cases correctly classified.

FIG. 18.17 – La matrice de classification lorsque les probabilités a priori sont fixées à 0,5

En effet, lorsqu'un des groupes de la variable dépendante est petit par rapport à l'autre (ou aux autres) groupe(s), il est fort probable que l'analyse discriminante, qui minimise le pourcentage de mauvaises classifications, classe très bien les observations du grand groupe et très mal celles du petit groupe. Ceci peut être gênant lorsque le groupe d'intérêt est justement le petit groupe, comme dans le cas de l'exemple. Une façon de contourner le problème consiste à considérer la distribution des valeurs discriminantes Z_j ; si les valeurs du petit groupe sont dans l'une des extrémités de la distribution, il suffit de fixer manuellement un point de classification à une valeur plus intéressante. De façon équivalente, une autre méthode consiste à travailler avec des groupes égaux et à ajuster les probabilités a priori à la hausse pour représenter le plus petit groupe. Ces deux techniques s'effectuent au prix de l'augmentation de la mauvaise classification du grand groupe. Par exemple, la figure 18.17 présente la matrice de classification que l'on aurait obtenue si au lieu de prendre les probabilités issues des proportions de l'échantillon on avait plutôt coché l'option **All groups equal**. Ceci a comme effet de considérer qu'un individu a autant de chances de se retrouver dans le groupe des désabonnés que dans l'autre groupe. On voit

que le pourcentage global de réussite a baissé, mais dans l'échantillon de validation le pourcentage d'individus bien classés pour le groupe des désabonnés est passé de 33,8 % à 68,8 %. Par contre le pourcentage d'individus bien classés a baissé pour l'autre groupe (il est passé de 91,3 % à 66,5 %).

S'il n'est pas possible d'avoir un échantillon de validation, une autre technique pour valider l'analyse consiste à utiliser la *Leave-one-out classification*. Cette technique met de côté tour à tour un individu, fait à chaque fois une analyse discriminante basée sur les $n - 1$ observations qui restent, et classe cet individu avec la fonction discriminante issue de cette analyse. L'estimation de la performance est alors moins biaisée que celle obtenue avec une analyse discriminante sans groupe de validation. La sortie 18.18 présente la matrice de classification de l'exemple lorsqu'on ne prend pas d'échantillon de validation (ce qui en fait serait ridicule ici étant donné la taille de l'échantillon) et qu'on a coché l'option **Leave-one-out classification** dans le bouton **Classify**. Il y a une petite différence dans les pourcentages de classification.

		Predicted Group Membership		Total
		Non	Oui	
Original	Count	682	44	726
	Oui	206	68	274
	%	93,9	6,1	100,0
	Oui	75,2	24,8	100,0
Cross-validated ^a	Count	679	47	726
	Oui	210	64	274
	%	93,5	6,5	100,0
	Oui	76,6	23,4	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 75,0% of original grouped cases correctly classified.

c. 74,3% of cross-validated grouped cases correctly classified.

FIG. 18.18 – La matrice de classification sans échantillon de validation, mais avec la *Leave-one-out classification*

18.3.6 Une bonne analyse discriminante

Une bonne analyse discriminante est bien entendu une analyse qui permet de faire une bonne classification. Mais à partir de quand peut-on dire qu'une classification est bonne ? La réponse n'est pas toujours évidente, mais nous donnons ici quelques éléments de réponse.

Tout d'abord, pour que l'analyse soit bonne, il est nécessaire que la fonction de discrimination réussisse à bien distinguer les groupes. Ainsi, lorsqu'on regarde les valeurs moyennes de la fonction dans chacun des groupes, on espère qu'elles soient significativement différentes. Ces moyennes sont appelées les **centroïdes**, et apparaissent dans la sortie 18.19.

Functions at Group Centroids	
	Function
desabon	1
Non	,256
Oui	-,671

Unstandardized canonical discriminant functions evaluated at group means

FIG. 18.19 – Les centres moyens de la fonction par groupe

Dans le cadre de l'exemple on voit donc que les centroïdes sont effectivement différents : la fonction discriminante a une valeur moyenne de 0,256 dans le groupe des non-désabonnés, et de -0,671 dans le groupe des désabonnés. Mais cette différence entre les centroïdes est-elle significative ?

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,853	110,964	5	,000

FIG. 18.20 – Le Wilks' Lambda de la fonction

C'est la figure 18.20 qui permet de répondre à cette question. Tout d'abord, la valeur

du Wilks' Lambda représente la proportion de la variation des scores discriminants qui n'est pas expliquée par la différence entre les groupes. Donc ici 85,3 % de la variation dans les scores n'est pas expliquée par la différence entre les désabonnés et les non-désabonnés. La fonction ne semble donc pas très performante, ce qui ne fait que confirmer ce qu'on a constaté avec la matrice de classification.

La *p*-value de la sortie 18.20 permet de résoudre le test d'hypothèses suivant à l'aide de la loi du χ^2 :

$$H_0 : \bar{Z}_{\text{Non-désabonnés}} = \bar{Z}_{\text{Désabonnés}}$$

$$H_1 : \bar{Z}_{\text{Non-désabonnés}} \neq \bar{Z}_{\text{Désabonnés}}$$

Fixons le seuil à $\alpha = 0,05$. Puisque la *p*-value est nulle, elle est plus petite que le seuil, et donc on rejette l'égalité des centroïdes au risque de se tromper une fois sur 20. Donc les centroïdes sont significativement différents ; ceci est nécessaire pour une bonne analyse, mais malheureusement pas suffisant, comme nous l'avons constaté. De plus, une grande taille d'échantillon peut parfois rendre significative une différence qui ne le serait pas sur un plus petit échantillon ; il est donc préférable de ne pas avoir un échantillon d'analyse trop grand.

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,172 ^a	100,0	100,0	,384

a. First 1 canonical discriminant functions were used in the analysis.

FIG. 18.21 – La corrélation canonique

Il y a aussi la sortie 18.21 qui nous renseigne sur la performance de la fonction discriminante. Elle contient tout d'abord la valeur propre (*Eigenvalue*) qui est simplement le ratio de la variation entre les groupes sur la variation à l'intérieur des groupes ; une bonne analyse discriminante produit une grande valeur propre. En fait les coefficients

de la fonction discriminante ont été calculés de façon à maximiser cette valeur propre. On peut calculer ce ratio à partir de la sortie 18.22 qui est l'ANOVA faite sur les variables `Dis1_1` (scores discriminants) et `desabon`. Il suffit de prendre les valeurs `Between` et `Within` de la colonne `Sum of Squares` : $120,709/700 = 0,172$. Il est aussi possible, à partir de cette ANOVA, de calculer la statistique Wilks' Lambda vue dans la figure 18.20 : $\text{Wilks' Lambda} = 0,853 = 700/820,709$. (**Attention** : pour obtenir cette table ANOVA, il faut prendre l'échantillon d'analyse seulement.)

ANOVA					
Dis1_1	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	120,709	1	120,709	120,709	,000
Within Groups	700,000	700	1,000		
Total	820,709	701			

FIG. 18.22 – L'ANOVA des scores discriminants selon les groupes de `desabon`

Mais revenons à la figure 18.21 ; on a vu que la valeur propre de cette analyse est de 0,172, ce qui n'est pas très grand. Aussi, la dernière colonne de cette sortie contient la corrélation canonique ; elle a ici une valeur de 0,384. Cette corrélation est en fait le ETA qu'on pourrait calculer avec l'ANOVA : $\sqrt{\frac{120,709}{820,709}} = 0,384$. Elle mesure le degré d'association entre les scores discriminants et les groupes. Dans une situation de deux groupes, la corrélation canonique est simplement le coefficient de corrélation de Pearson entre les scores discriminants Z_j et les groupes lorsqu'ils sont codés 0 et 1. Ici la corrélation est assez faible.

Ainsi, les statistiques que nous venons de voir nous donnent une idée de la performance de l'analyse discriminante, mais l'analyse de la performance se base avant tout sur la classification elle-même. Et il n'est pas toujours évident de savoir si le pourcentage de données bien classées est satisfaisant ou non, surtout lorsque les groupes sont de

tailles inégales. Mais disons qu'à la base, il faut au moins que le pourcentage de bonnes classifications dépasse ce qui serait obtenu par simple chance ; lorsqu'il y a deux groupes, ceci revient à exiger de dépasser 50 % de cas bien classés. Mais comme on l'a vu, il faut aussi regarder les pourcentages pour chacun des groupes, et évaluer dans le contexte ce qui est satisfaisant ou pas.

18.3.7 Les pré-requis

Comme pour toute technique d'analyse statistique, pour obtenir une analyse discriminante optimale qui minimise la probabilité de mauvaise classification, certaines hypothèses doivent être rencontrées. Dans le cadre d'une analyse discriminante, trois hypothèses sont à vérifier :

- Chaque groupe doit provenir d'une population multinormale.
- Les matrices de covariances doivent toutes être égales.
- La multicolinéarité doit être tolérable ($VIF < 10$).

La normalité

Une variété de tests d'hypothèses existe pour vérifier la multinormalité. Une première tactique, simpliste, consiste à vérifier la normalité de la distribution de chacun des groupes des variables indépendantes. Si de façon marquée l'une des variables ne se distribue pas de façon normale, l'analyste est en droit de se douter que l'hypothèse de multinormalité est violée. L'inverse est cependant faux. Le fait que toutes les variables se distribuent de façon normale ne peut mener à la conclusion que l'ensemble des variables se distribuent de façon multinormale. Dans le cadre de ce cours, on se contente de vérifier la normalité pour chaque variable indépendante dans chacun des groupes.

Il faut donc traiter le test d'hypothèses suivant pour chaque groupe et chaque variable indépendante :

H_0 : Les données de la population se répartissent selon une loi normale.

H_1 : Les données de la population ne se répartissent pas selon une loi normale.

Dans le cadre de cet exemple, on obtient la sortie 18.23 (avec l'échantillon d'analyse). Fixons le seuil à $\alpha = 0,05$. Les p -values étant nulles, on rejette la normalité pour toutes les variables, dans les deux groupes, et ce au risque de se tromper une fois sur 20. Ceci a pu causer certains problèmes lors de l'estimation des coefficients de la fonction discriminante ; il est parfois préférable de se tourner vers la régression logistique (qui sera vue au chapitre suivant) lorsque la normalité ne tient pas.

Tests of Normality						
desabon	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Zlongdistm	Non ,161	508	,000	,782	508	,000
	Oui ,132	194	,000	,893	194	,000
Zsfraism	Non ,309	508	,000	,802	508	,000
	Oui ,337	194	,000	,777	194	,000
Zequipm	Non ,440	508	,000	,638	508	,000
	Oui ,253	194	,000	,866	194	,000
Zcartem	Non ,141	508	,000	,872	508	,000
	Oui ,285	194	,000	,752	194	,000
Zsansfilm	Non ,442	508	,000	,611	508	,000
	Oui ,367	194	,000	,734	194	,000

a. Lilliefors Significance Correction

FIG. 18.23 – Les tests de normalité

Égalité des matrices de covariances

Pour vérifier l'égalité des matrices de covariances, SPSS utilise le test appelé *Box's M Test*. Le test utilise une statistique basée sur le déterminant des matrices formées par chacun des groupes de la variable dépendante Y . Ce test est cependant fragile à l'absence de multinormalité. En effet, ce test d'hypothèses tend à rejeter l'hypothèse nulle lorsque l'hypothèse de multinormalité n'est pas rencontrée, ce qui n'est pas mauvais en soi.

Le test d'hypothèses s'écrit de la façon suivante :

H_0 : Les matrices de covariances sont égales au niveau de la population.

H_1 : Les matrices de covariances ne sont pas égales au niveau de la population.

La figure 18.24 contient la sortie pour résoudre ce test. On rejette H_0 si la p -value est plus petite que le seuil α . C'est le cas ici puisque la p -value est nulle ; ainsi au risque de se tromper une fois sur 20 (si le seuil est de 5 %), on admet que dans l'exemple, les matrices de covariances ne sont pas égales. Ce n'est pas étonnant puisque la normalité a été rejetée. Il faut donc émettre des réserves quant à la qualité du modèle, ce qui a déjà été fait suite à la classification peu satisfaisante...

Test Results		
Box's M		230,541
F	Approx.	15,213
df1		15
df2		560668,7
Sig.		,000

Tests null hypothesis of equal population covariance matrices.

FIG. 18.24 – Le test de Box's M

Si les matrices de covariances ne sont pas égales mais que la multinormalité est rencontrée, la littérature soutient que la meilleure règle de classification est la règle quadratique. Soulignons que cette règle peut mal fonctionner en présence d'une petite taille d'échantillon. Malheureusement, SPSS ne permet pas encore de calculer la règle de classification quadratique optimale. Si les matrices de covariances ne sont pas trop différentes, Welsh (1977) soulève que l'analyse discriminante est passablement efficace, surtout en présence d'échantillons de petites tailles. Dans le cas de la classification en deux groupes, la régression logistique, qui n'exige pas la multinormalité, peut être une solution de rechange à l'analyse discriminante. Cependant, la régression logistique possède aussi ses restrictions sur les valeurs des variables indépendantes.

Mentionnons finalement que, dans une situation où les variables indépendantes sont toutes des variables binaires (oui-non, homme-femme, ...), ou encore une combinaison de

variables discrètes et continues, l'analyse discriminante n'est pas optimale. La littérature souligne qu'elle peut offrir tout de même de bonnes analyses.

Multicolinéarité tolérable

Il n'est pas indispensable de vérifier la multicolinéarité parmi les variables indépendantes. Il faut cependant être conscient que si les VIF sont élevés et qu'on désire utiliser la méthode stepwise pour savoir quelles variables sont les plus importantes pour l'analyse discriminante, il se peut que la solution proposée ne soit pas la meilleure. En effet, les tests effectués pour comparer les variables sont sensibles à la multicolinéarité.

Pour vérifier les VIF, il suffit de faire les commandes comme si on voulait une régression linéaire avec les variables de l'analyse. Dans le cadre de l'exemple, on obtient la sortie 18.25. Puisque le plus grand VIF a une valeur de 2, il n'y aurait pas de problème si on voulait utiliser un stepwise avec ces variables.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	,274	,013		20,864	,000		
Zlongdistm	-,101	,014	-,227	-7,037	,000	,831	1,203
Zsfraism	-,018	,016	-,041	-1,174	,241	,700	1,428
Zequipm	,107	,017	,239	6,458	,000	,633	1,579
Zcartem	-,030	,015	-,066	-1,965	,050	,761	1,315
Zsansfilm	,007	,019	,015	,367	,714	,500	2,000

a. Dependent Variable: desabon

FIG. 18.25 – Les VIF des variables indépendantes

18.4 Un autre exemple

Exemple 18.4.1 On prend ici la base de données `fournisseur.sav`. On aimerait comprendre ce qui distingue les entreprises qui lors de leurs achats veulent connaître en

détail le prix de chaque produit et service (*specification buying*) de celles qui optent plutôt pour un prix estimé pour l'ensemble de ce qu'elles désirent (*total value analysis*). Pour ce faire, on fait une analyse discriminante avec 7 variables indépendantes : **livraison**, **prix**, **flexibilite**, **image**, **service**, **forcevente** et **produit**. La méthode stepwise sera utilisée pour déterminer lesquelles de ces variables devraient vraiment être incluses dans le modèle. Il est à noter que pour le stepwise de cet exemple, on a coché l'option **Use probability of F** dans le bouton **Method...** (et on a laissé les probabilités par défaut). Les seuils sont fixés à $\alpha = 0,05$.

Analysis Case Processing Summary			
Unweighted Cases		N	Percent
Valid		78	78,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Unselected	22	22,0
	Total	22	22,0
	Total	100	100,0

FIG. 18.26 – Les données incluses dans l'analyse

La figure 18.26 montre que l'échantillon d'analyse se compose ici de 78 observations, et celui de validation de 22 observations.

La figure 18.27 contient les statistiques descriptives pour chacun des groupes. On remarque certaines différences au niveau des moyennes ; on peut confirmer lesquelles sont significatives avec les ANOVA individuelles.

Group Statistics					
		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
Specification Buying	livraison	2,524	1,1122	33	33,000
	prix	3,106	1,1121	33	33,000
	flexibilite	6,652	,8171	33	33,000
	image	5,291	,8240	33	33,000
	service	2,782	,9534	33	33,000
	forcevente	2,645	,5740	33	33,000
	produit	8,479	,8192	33	33,000
Total Value Analysis	livraison	4,069	1,0083	45	45,000
	prix	2,049	1,1034	45	45,000
	flexibilite	8,491	1,2908	45	45,000
	image	5,287	1,3252	45	45,000
	service	3,033	,5954	45	45,000
	forcevente	2,671	,8387	45	45,000
	produit	6,109	1,2710	45	45,000
Total	livraison	3,415	1,2981	78	78,000
	prix	2,496	1,2191	78	78,000
	flexibilite	7,713	1,4375	78	78,000
	image	5,288	1,1339	78	78,000
	service	2,927	,7720	78	78,000
	forcevente	2,660	,7342	78	78,000
	produit	7,112	1,6095	78	78,000

FIG. 18.27 – Les statistiques descriptives

Les p -values des ANOVA individuelles (figure 18.28) nous montrent qu'il y a une différence significative entre les moyennes des deux groupes pour les variables **livraison**, **prix**, **flexibilite** et **produit**. Et il semble que ce soit la variable **produit** qui discrimine le plus d'un groupe à l'autre (elle a le plus grand F et le plus petit Wilks' Lambda). N'oublions pas cependant que ce sont les apports individuels.

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
livraison	,650	40,944	1	76	,000
prix	,814	17,360	1	76	,000
flexibilite	,595	51,717	1	76	,000
image	1,000	,000	1	76	,987
service	,974	2,048	1	76	,156
forcevente	1,000	,023	1	76	,880
produit	,464	87,805	1	76	,000

FIG. 18.28 – Les ANOVA individuelles

Variables Entered/Removed ^{a,b,c,d}										
Step	Entered	Wilks' Lambda				Exact F				Sig.
		Statistic	df1	df2	df3	Statistic	df1	df2		
1	produit	,464	1	1	76,000	87,805	1	76,000	,000	
2	flexibilite	,340	2	1	76,000	72,898	2	75,000	,000	
3	forcevente	,306	3	1	76,000	56,024	3	74,000	,000	

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

- a. Maximum number of steps is 14.
- b. Maximum significance of F to enter is .05.
- c. Minimum significance of F to remove is .10.
- d. F level, tolerance, or VIN insufficient for further computation.

FIG. 18.29 – Les variables gardées par le stepwise

Les figures 18.29, 18.30 et 18.31 nous montrent les étapes du stepwise ; la première variable à être entrée dans le modèle est **produit**, ce qui n'est pas étonnant étant donné les résultats de l'ANOVA. Ensuite c'est la variable **flexibilite** qui entre ; pas de surprise encore une fois. Par contre la troisième étape est plus surprenante : la dernière variable à intégrer le modèle est la variable **forcevente** qui pourtant ne semblait pas faire une bonne discrimination d'un groupe à l'autre. De plus les autres variables qui semblaient faire une bonne discrimination ne sont pas incluses dans le modèle. Il se peut que ce phénomène soit simplement dû au fait que ces trois variables combinées ensemble forment une bonne fonction discriminante, ou bien il y a de la multicolinéarité qui est venue brouiller les pistes. Il faudra vérifier ceci.

Variables in the Analysis				
Step	Tolerance	Sig. of F to Remove	Wilks' Lambda	
1 produit	,1,000	,000		
2 produit flexibilite	,997 ,997	,000 ,000	,595 ,464	
3 produit flexibilite forcevente	,872 ,967 ,857	,000 ,000 ,005	,588 ,431 ,340	

FIG. 18.30 – Les étapes du stepwise

Variables Not in the Analysis				
Step	Tolerance	Min. Tolerance	Sig. of F to Enter	Wilks' Lambda
0	livraison	1,000	,000	,650
	prix	1,000	,000	,814
	flexibilite	1,000	,000	,595
	image	1,000	,987	1,000
	service	1,000	,156	,974
	forcevente	1,000	,880	1,000
	produit	1,000	,000	,464
1	livraison	,988	,000	,393
	prix	,952	,147	,451
	flexibilite	,997	,000	,340
	image	,915	,058	,442
	service	,991	,120	,449
	forcevente	,883	,018	,431
2	livraison	,956	,022	,316
	prix	,814	,630	,339
	image	,882	,014	,313
	service	,971	,050	,322
	forcevente	,857	,005	,306
3	livraison	,926	,083	,293
	prix	,801	,914	,306
	image	,381	,609	,305
	service	,916	,208	,299

FIG. 18.31 – Les variables exclues par le stepwise

Les sorties 18.32 et 18.33 nous donnent les Wilks' Lambda des fonctions discriminantes. Plus précisément, la figure 18.32 analyse les fonctions discriminantes à chaque étape du stepwise. Ainsi, si on fait une analyse discriminante avec la variable **produit** comme seule variable indépendante, le Wilks' Lambda de la fonction est alors de 0,464, et les centres moyens de la fonction dans chaque groupe sont significativement différents puisque la *p*-value est nulle.

Step	Number of Variables	Wilks' Lambda						Exact F			
		Lambda	df1	df2	df3	Statistic	df1		df2		Sig.
							1	2	3	4	
1	1	,464	1	1	76	87,805	1	76,000	1	,000	
2	2	,340	2	1	76	72,898	2	75,000	2	,000	
3	3	,306	3	1	76	56,024	3	74,000	3	,000	

FIG. 18.32 – Les Wilks' Lambda des fonctions

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,306	88,296	3	,000

FIG. 18.33 – Le Wilks' Lambda de la fonction avec 3 variables

On voit qu'aux étapes 2 et 3 le Wilks' lambda diminue (il passe de 0,464 à 0,340 puis à 0,306), ce qui veut dire que l'ajout des variables **flexibilite** et **forcevente** améliore la performance de la fonction discriminante.

La sortie 18.33 donne le Wilks' Lambda et la *p*-value de la meilleure solution selon le stepwise. Ainsi il y a 30,6 % de la variation des scores discriminants qui n'est pas expliquée par le passage d'un groupe à l'autre, ceci semble assez satisfaisant. On peut résoudre le test suivant :

$$H_0 : \bar{Z}_{\text{specification buying}} = \bar{Z}_{\text{total value analysis}}$$

$$H_1 : \bar{Z}_{\text{specification buying}} \neq \bar{Z}_{\text{total value analysis}}$$

Puisque la p -value = 0 < 0,05, on rejette H_0 . Ainsi, au risque de se tromper une fois sur 20, on admet que les centroïdes sont différents. La figure 18.34 nous donne justement les valeurs de ces centres : le centroïde du groupe *specification buying* a une valeur de -1,737, tandis que celui de *total value analysis* a une valeur de 1,274.

Functions at Group Centroids	
	Function
x11	1
Specification Buying	-1,737
Total Value Analysis	1,274

Unstandardized canonical discriminant
functions evaluated at group means

FIG. 18.34 – Les centres moyens de la fonction par groupe

La figure 18.35 nous donne la valeur propre ainsi que la corrélation canonique de la fonction discriminante. La valeur propre est de 2,271, ce qui est bien (c'est un ratio de la variation expliquée sur l'inexpliquée, on espère donc que ce soit plus grand que 1). La corrélation canonique de 0,833 nous montre qu'il y a une bonne association entre les scores discriminants et les groupes.

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2,271 ^a	100,0	100,0	,833

a. First 1 canonical discriminant functions were used in the analysis.

FIG. 18.35 – La corrélation canonique

La figure 18.36 contient les sorties des loadings et des coefficients de la fonction discriminante. La variable qui a le plus grand loading (en valeur absolue) est la variable *produit*; on peut ainsi dire que cette variable partage 71,3 % de la variance de la fonction discriminante, et que c'est donc elle qui a le plus d'impact dans le calcul des scores

discriminants. On interprète de façon semblable les autres loadings. Remarquez que les loadings sont donnés pour toutes les variables, même celles qui n'ont pas été retenues par le stepwise. Lorsqu'une variable qui n'est pas retenue par le stepwise a un loading intéressant, ceci révèle que cette variable a un lien linéaire avec les variables retenues pour l'analyse (elles partagent de l'information). Ce qui est par contre surprenant, c'est que la variable **forcevente** qui a été retenue pour l'analyse a un loading très faible de 0,012. On commence à se douter qu'il doit il y avoir un phénomène de multicolinéarité.

Structure Matrix	
	Function
	1
produit	-,713
flexibilite	,547
prix ^a	-,331
livraison ^a	,249
service ^a	-,062
image ^a	-,046
forcevente	,012

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

a. This variable not used in the analysis.

Canonical Discriminant Function Coefficients	
	Function
	1
flexibilite	,589
forcevente	,555
produit	-,807
(Constant)	-,278

Unstandardized coefficients

FIG. 18.36 – Les loadings et les coefficients de la fonction discriminante

La deuxième sortie de la figure 18.36 permet d'écrire la fonction discriminante :

$$z = -0,278 + 0,589X_{\text{flexibilite}} + 0,555X_{\text{forcevente}} - 0,807X_{\text{produit}}.$$

Quelle interprétation peut-on donner aux coefficients de la fonction et aux loadings ?

En fait ce sont eux qui nous permettent de comprendre les différences entre les deux groupes (en supposant que la fonction fait une bonne discrimination). Le groupe de *specification buying* a un centroïde négatif, et la variable **produit** est la seule à avoir un coefficient négatif. Donc lorsqu'une entreprise est très satisfaite de la qualité du produit, son score discriminant se retrouve diminué, ce qui l'amène à être plus près des scores discriminants du groupe *specification buying*. On observe la situation inverse pour le groupe *total value analysis* et les variables **flexibilite** et **forcevente**. Et ces interprétations vont dans le même sens que ce qu'on avait constaté dans les ANOVA individuelles. Le but de l'analyse étant surtout de comprendre les différences entre ces deux groupes, l'interprétation est très importante. Il reste à vérifier sa validité avec les autres étapes de l'analyse (si la classification n'est pas satisfaisante, on se doutera alors que cette analyse n'est pas très fiable).

Prior Probabilities for Groups			
x11	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Specification Buying	,423	33	33,000
Total Value Analysis	,577	45	45,000
Total	1,000	78	78,000

FIG. 18.37 – Les probabilités a priori

La figure 18.37 nous montre les probabilités a priori qui ont été utilisées pour le classement. On voit que 42,3 % des entreprises sont de type *specification buying* tandis que 57,7 % sont de type *total value analysis*.

Finalement, la sortie 18.38 nous montre le sommaire de la classification. On voit que sur l'échantillon d'analyse, 94,9 % des cas ont été bien classés, ce qui semble excellent. De plus, ce pourcentage est assez bien réparti entre les deux groupes : 97 % du groupe *specification buying* ont été bien classés, et 93,3 % pour l'autre groupe. Le pourcentage global pour l'échantillon de validation est un peu plus bas : 86,4 % des cas ont été bien

				Classification Results ^{a,b}			
				Predicted Group Membership		Total	
Cases Selected	Original	Count	x11	Specification	Total Value Analysis		
			Specification Buying	32	1	33	
			Total Value Analysis	3	42	45	
Cases Not Selected	Original	Count	%	Specification Buying	97,0	3,0	100,0
			Total Value Analysis	6,7	93,3	100,0	
			%	Specification Buying	57,1	42,9	100,0
				Total Value Analysis	,0	100,0	100,0

a. 94,9% of selected original grouped cases correctly classified.

b. 86,4% of unselected original grouped cases correctly classified.

FIG. 18.38 – La matrice de classification

classés. Le hic est au niveau du groupe *specification buying* : seulement 4 sur 7 ont été bien classés. Il faudrait investiguer plus de ce côté pour tenter de comprendre ce phénomène. Peut-être serait-il pertinent d'inclure d'autres variables pour faire l'analyse.

Test Results		
Box's M		19,680
F	Approx.	3,134
df1		6
df2		32727,978
Sig.		,005

Tests null hypothesis of equal population covariance matrices.

FIG. 18.39 – Le test de Box's M

Jetons maintenant un œil du côté des pré-requis. La sortie 18.39 nous permet de résoudre le test sur les covariances :

H_0 : Les matrices de covariances sont égales au niveau de la population.

H_1 : Les matrices de covariances ne sont pas égales au niveau de la population.

Puisque la *p*-value est de $0,005 < 0,05$, on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet que les matrices de covariances ne sont pas égales. Ainsi il faut être prudent quant aux conclusions tirées de cette analyse.

x11		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
flexibilite	Specification Buying	,113	33	,200*	,981	33	,822
	Total Value Analysis	,148	45	,015	,878	45	,000
forcevente	Specification Buying	,129	33	,180	,965	33	,361
	Total Value Analysis	,128	45	,063	,960	45	,118
produit	Specification Buying	,128	33	,183	,971	33	,500
	Total Value Analysis	,118	45	,126	,969	45	,261

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. 18.40 – Les tests de normalité

La sortie 18.40 nous présente les tests de normalité (on a sélectionné au préalable les données telles que `test = 1`). On traite le test suivant pour les trois variables retenues et pour les deux groupes :

H_0 : Les données de la population se répartissent selon une loi normale.

H_1 : Les données de la population ne se répartissent pas selon une loi normale.

On voit que toutes les p -values sont plus grandes que le seuil $\alpha = 0,05$, sauf celles pour la variable `flexibilite` dans le groupe *total value analysis*. Et encore là, on pourrait dire que c'est tolérable car la p -value issue du test de Kolmogorov-Smirnov permettrait de ne pas rejeter la normalité à un seuil de 0,01, et de plus l'analyse discriminante est assez robuste à la violation de la normalité. On peut donc dire que les résultats sont assez satisfaisant de ce côté.

Par contre, l'examen de la figure 18.41 nous montre qu'il y a présence d'une forte multicolinéarité (trois des VIF sont plus grands que 10). Ainsi le stepwise peut avoir donné des résultats erronés. Il serait donc recommandé de recommencer cette analyse avec une autre combinaison de variables ; cet exercice vous est proposé à la fin du chapitre.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	,126	,329		,382	,703		
livraison	,141	,133	,379	1,062	,291	,028	35,747
prix	,055	,138	,134	,400	,690	,032	31,597
flexibilité	,118	,027	,331	4,329	,000	,608	1,645
image	-,013	,044	-,030	-,298	,766	,347	2,879
service	-,100	,259	-,152	-,385	,701	,023	43,834
forcevente	,089	,063	,139	1,414	,161	,371	2,697
produit	-,137	,023	-,443	-5,851	,000	,623	1,606

a. Dependent Variable: x11

FIG. 18.41 – Les VIF pour toutes les variables indépendantes

18.5 Un exemple à trois groupes

Jusqu'à maintenant, l'utilisation de l'analyse discriminante s'est limitée à distinguer deux groupes. Cependant l'analyse discriminante révèle sa véritable puissance dans le fait qu'elle permet de discriminer plus de deux groupes. Même si les analyses de base sont les mêmes que celles proposées dans l'analyse discriminante à deux groupes, plusieurs éléments additionnels sont à prendre en considération.

Cette section présente un exemple d'analyse discriminante où la variable dépendante a 3 modalités. La base de données utilisée est **fournisseurplus.sav**. Elle est très semblable à la base données **fournisseur.sav**; elle est issue d'une étude qu'un fournisseur industriel a menée auprès de ses clients.

On s'intéresse ici à classer les clients en trois groupes selon la durée pendant laquelle ils ont été (ou sont) clients de ce fournisseur : moins d'un an, entre un an et 5 ans, et plus de 5 ans. On veut faire la classification à l'aide de 13 variables qui mesurent la perception du client sur différents aspects du fournisseur. Les 13 variables sont mesurées sur une échelle continue entre 0 et 10 (0 = pauvre, 10 = excellente). Le tableau qui suit donne leurs descriptions.

qualite	Qualité perçue du produit
web	Image perçue du site web du fournisseur
support	Qualité perçue du support technique
probres	Perception de la manière dont sont résolus les problèmes
pub	Perception de la publicité
variete	Perception de la variété des produits (pour rencontrer les besoins)
forcevente	Image de la force de vente
prix	Perception : prix compétitifs
garanties	Respect des garanties
nouveau	Perception : développement et vente de nouveaux produits
commandes	Perception : commandes et facturations traitées correctement
flexibilite	Perception de la latitude de négociation des prix sur tous les types d'achats
livraison	Rapidité de livraison

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		83	83,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Unselected	17	17,0
	Total	17	17,0
	Total	100	100,0

FIG. 18.42 – Première sortie de l'analyse

Étant donné le grand nombre de variables et la petite taille d'échantillon (100), il serait préférable de ne sélectionner que quelques unes de ces variables (en se guidant avec les ANOVA, ou avec un stepwise). Ici nous gardons toutes les variables, et avons même

divisé l'échantillon en deux : on a gardé 83 observations pour l'échantillon d'analyse (voir figure 18.42). Ce n'est cependant pas idéal de procéder ainsi ; il vous est proposé à la fin du chapitre de reprendre cet exemple avec moins de variables.

La figure 18.43 présente les statistiques descriptives. Certaines des variables semblent permettre de différencier le premier groupe des deux derniers (comme par exemple `livraison`), les deux premiers du dernier (comme par exemple `qualite`), mais aucune ne semble pouvoir distinguer les trois groupes à la fois (c'est `variete` qui varie le plus sur les trois groupes à la fois, mais les différences ne semblent pas très grandes). Il semble donc vraiment nécessaire de calculer deux fonctions discriminantes pour pouvoir faire une bonne classification.

En présence de deux groupes, il est possible de calculer une unique fonction discriminante qui maximise le ratio *between* sur *within sum of squares*. Avec trois groupes (ou plus), et en assumant qu'il y a un nombre suffisant de variables indépendantes (au moins le nombre de groupes (g) moins un), la technique produira $g - 1$ fonctions discriminantes. S'il y a trois groupes comme dans l'exemple, il y a alors deux fonctions discriminantes. La première fonction, comme dans le cas de deux groupes, présentera le ratio *between* sur *within sum of squares* le plus important. La seconde fonction, orthogonale à la première, produira le second ratio *between* sur *within sum of squares* le plus élevé. Dans le cas général, les $g - 1$ fonctions sont toutes orthogonales les unes par rapport aux autres et maximisent le ratio *between* sur *within sum of squares* moyennant les contraintes d'orthogonalité.

		Group Statistics			
type		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
Moins d'un an	qualité	7,103	1,0381	31	31,000
	web	3,684	,7095	31	31,000
	support	4,981	1,5806	31	31,000
	probres	4,352	,9486	31	31,000
	pub	3,700	1,0188	31	31,000
	varieté	4,781	1,0303	31	31,000
	forcevente	4,868	,9669	31	31,000
	prix	7,565	1,2330	31	31,000
	garanties	5,835	,8995	31	31,000
	nouveau	5,203	1,6506	31	31,000
	commandes	3,555	,8869	31	31,000
	flexibilité	4,284	,9771	31	31,000
1 à 5 ans	livraison	3,168	,6247	31	31,000
	qualité	7,193	1,3303	28	28,000
	web	3,757	,5859	28	28,000
	support	5,261	1,5293	28	28,000
	probres	5,911	,8954	28	28,000
	pub	4,275	1,1884	28	28,000
	varieté	5,475	1,0102	28	28,000
	forcevente	5,382	,8650	28	28,000
	prix	7,582	1,4241	28	28,000
	garanties	5,946	,8185	28	28,000
	nouveau	4,811	1,3500	28	28,000
	commandes	4,639	,7955	28	28,000
Plus de 5 ans	flexibilité	5,621	1,1764	28	28,000
	livraison	4,225	,5732	28	28,000
	qualité	9,017	,7032	24	24,000
	web	3,550	,6587	24	24,000
	support	5,517	1,3101	24	24,000
	probres	5,863	1,0890	24	24,000
	pub	3,892	1,2086	24	24,000
	varieté	7,017	,9083	24	24,000
	forcevente	4,917	1,1586	24	24,000
	prix	5,938	1,3484	24	24,000
	garanties	6,179	,7138	24	24,000
	nouveau	5,142	1,5234	24	24,000
Total	commandes	4,475	,6180	24	24,000
	flexibilité	3,933	,7305	24	24,000
	livraison	4,196	,4759	24	24,000
	qualité	7,687	1,3584	83	83,000
	web	3,670	,6525	83	83,000
	support	5,230	1,4877	83	83,000
	probres	5,314	1,2189	83	83,000
	pub	3,949	1,1468	83	83,000
	varieté	5,661	1,3411	83	83,000
	forcevente	5,055	1,0101	83	83,000
	prix	7,100	1,5134	83	83,000
	garanties	5,972	,8242	83	83,000

FIG. 18.43 – Les statistiques descriptives

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
qualite	,605	26,162	2	80	,000
web	,984	,657	2	80	,521
support	,978	,884	2	80	,417
probres	,623	24,184	2	80	,000
pub	,954	1,935	2	80	,151
variete	,531	35,268	2	80	,000
forcevente	,946	2,297	2	80	,107
prix	,757	12,837	2	80	,000
garanties	,971	1,203	2	80	,306
nouveau	,986	,549	2	80	,579
commandes	,711	16,258	2	80	,000
flexibilite	,646	21,912	2	80	,000
livraison	,549	32,836	2	80	,000

FIG. 18.44 – Les ANOVA individuelles

La figure 18.44 présente les ANOVA individuelles. On peut ainsi constater que prises individuellement, ce sont les variables **qualite**, **probres**, **variete**, **prix**, **commandes**, **flexibilite** et **livraison** qui varient de façon significative d'un groupe à l'autre. C'est la variable **variete** qui a le plus grand ratio *F* et le plus petit Wilks' Lambda, c'est donc cette variable qui varie le plus entre les groupes.

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2,291 ^a	75,6	75,6	,834
2	,741 ^a	24,4	100,0	,652

a. First 2 canonical discriminant functions were used in the analysis.

FIG. 18.45 – Les corrélations canoniques

Pour chacune des deux fonctions, la valeur propre est le rapport *between group* sur *within group sum of squares* où les groupes sont ceux de la variable dépendante. Rapelons que l'algorithme optimise ces rapports tout en préservant l'orthogonalité entre les fonctions discriminantes. À cet effet, à partir de la sortie 18.46 qui présente les ANOVA séparées entre les deux fonctions discriminantes et la variable dépendante, il est possible

de voir que la valeur propre de la première fonction est $2,291 = 183,318 / 80$, et que la valeur propre de la seconde fonction est $0,741 = 59,309 / 80$. On voit aussi que c'est la première fonction qui a la plus grande corrélation canonique, ce qui est en fait toujours le cas.

ANOVA					
		Sum of Squares	df	Mean Square	F
Dis1_1	Between Groups	183,318	2	91,659	91,659
	Within Groups	80,000	80	1,000	
	Total	263,318	82		
Dis2_1	Between Groups	59,309	2	29,655	29,655
	Within Groups	80,000	80	1,000	
	Total	139,309	82		

FIG. 18.46 – Les ANOVA entre les fonctions discriminantes et la variable dépendante (3 groupes)

Lorsque deux fonctions discriminantes ou plus sont présentes dans un modèle, il est intéressant pour le praticien de comparer leur mérite. Pour ce faire, il faut remarquer qu'à travers le processus d'optimisation, les variances résiduelles *Within Groups* sont toujours égales à 1, et ce, peu importe la fonction discriminante considérée ; on peut le constater dans la figure 18.46 (colonne *Mean Square*). Dans ce cas, les fonctions discriminantes ne diffèrent que dans leurs sommes de carrés respectives entre les groupes (*Between Group Sum of Squares*). Rappelons que par construction, la première fonction détient toujours le plus grand ratio *Between* sur *Within*, donc le plus important *Between*. La seconde fonction discriminante détient le second ratio le plus important *Between* sur *Within*, donc le second plus important *Between*, et ainsi de suite si plusieurs fonctions discriminantes sont concurrentes. De la colonne *% of Variance* de la figure 18.45, il est possible de voir que la première fonction discriminante explique 75,6 % de la variation entre les groupes, tandis que la seconde compte pour les 24,4 % de la variation restante entre les groupes.

Un test d'hypothèses vérifiant l'égalité des moyennes des fonctions discriminantes dans

chacune des populations peut être basé sur le Wilks' Lambda. Comme plusieurs fonctions discriminantes indépendantes doivent être considérées simultanément, le Wilks' Lambda est justement le produit des Wilks' Lambda univariés. Plus précisément, en prenant les sommes des carrés de la sortie 18.46, le Wilks' Lambda pour les fonctions 1 et 2 prises simultanément est égal à :

$$\Lambda = \left(\frac{80}{263,318} \right) \left(\frac{80}{139,309} \right) = 0,174.$$

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,174	129,204	26	,000
2	,574	41,046	12	,000

FIG. 18.47 – Les Wilks' Lambda des fonctions

À l'aide de la transformation du Lambda en statistique du Chi-deux et par la *p*-value ainsi obtenue (*Sig.* = ,000), il est possible de conclure, au risque de se tromper une fois sur 20, que l'hypothèse nulle suivant laquelle les moyennes des deux fonctions sont égales dans les trois populations peut être rejetée. Ceci montre qu'au moins une des deux fonctions possède un pouvoir de discrimination intéressant.

La seconde ligne de la sortie 18.47 illustre le Lambda univarié de la seconde fonction en solitaire, c'est-à-dire une fois la première fonction enlevée. Il est alors possible de voir que la seconde fonction est effective. En résumé, lorsque plusieurs fonctions discriminantes sont présentes dans un modèle, il est possible d'utiliser la sortie 18.47 afin de tester d'abord si les fonctions ont un pouvoir discriminant lorsque prises simultanément ; on teste ensuite les sous-groupes de fonctions discriminantes, enlevant une fonction à la fois.

Une manière intéressante d'étudier la contribution de chacune des variables indépendantes sur les fonctions discriminantes consiste à regarder les corrélations (*loadings*) entre les variables et les scores des fonctions discriminantes. La sortie 18.48 montre que ce sont

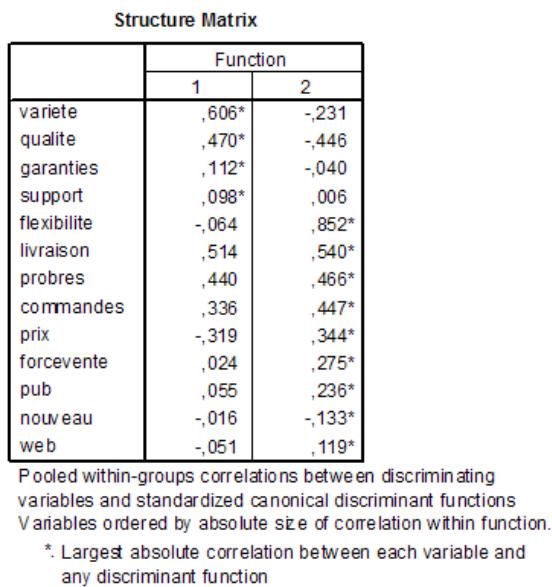


FIG. 18.48 – Les loadings

les variables **variete**, **qualite**, **garanties** et **support** qui sont en grande corrélation avec la première fonction discriminante. Ceci illustre que ce sont ces variables, qui, prisent ensemble, permettent la plus grande discrimination entre les trois groupes de la variable dépendante. De même, il est possible de dégager les variables ayant une plus grande corrélation avec le second facteur. À cet effet, SPSS groupe les plus grandes corrélations ensemble et place une étoile afin d'identifier avec quelle fonction la corrélation est la plus élevée. Cela ne signifie pas que l'association est significative ; il suffit de regarder la variable **support** pour le constater.

La figure 18.49 présente les coefficients des fonctions. Ainsi les fonctions s'écrivent

$$\begin{aligned}
 z_1 = & -7,541 + 0,695X_{\text{qualite}} - 0,865X_{\text{web}} - 0,035X_{\text{support}} - 0,081X_{\text{probres}} \\
 & + 0,015X_{\text{pub}} - 0,351X_{\text{variete}} + 0,542X_{\text{forcevente}} - 0,304X_{\text{prix}} + 0,145X_{\text{garanties}} \\
 & - 0,066X_{\text{nouveau}} - 0,157X_{\text{commandes}} - 0,446X_{\text{flexibilite}} + 2,495X_{\text{livraison}} \\
 z_2 = & -6,003 - 0,095X_{\text{qualite}} - 0,699X_{\text{web}} + 0,102X_{\text{support}} + 0,045X_{\text{probres}} \\
 & - 0,085X_{\text{pub}} + 1,155X_{\text{variete}} + 0,483X_{\text{forcevente}} + 0,034X_{\text{prix}} - 0,161X_{\text{garanties}} \\
 & - 0,249X_{\text{nouveau}} + 0,265X_{\text{commandes}} + 1,950X_{\text{flexibilite}} - 2,169X_{\text{livraison}}
 \end{aligned}$$

Canonical Discriminant Function Coefficients		
	Function	
	1	2
qualite	,695	-,095
web	-,865	-,699
support	-,035	,102
probres	-,081	,045
pub	,015	-,085
variete	-,351	1,155
forcevente	,542	,483
prix	-,304	,034
garanties	,145	-,161
nouveau	-,066	-,249
commandes	-,157	,285
flexibilite	-,446	1,950
livraison	2,495	-2,169
(Constant)	-7,541	-6,003

Unstandardized coefficients

FIG. 18.49 – Les coefficients de la fonction discriminante

C'est horrible, je sais. ;-)

Functions at Group Centroids		
	Function	
	1	2
type		
Moins d'un an	-1,674	-,541
1 à 5 ans	,140	1,182
Plus de 5 ans	1,999	-,681

Unstandardized canonical discriminant functions evaluated at group means

FIG. 18.50 – Les centres moyens de la fonction par groupe

L'interprétation des fonctions discriminantes se fait en regardant le signe de chaque coefficient et les valeurs des centroïdes (figure 18.50). Par exemple, la fonction 1 permet de différencier les trois groupes, mais surtout le premier du dernier, tandis que la deuxième fonction permet surtout de distinguer le deuxième groupe des deux autres. Ensuite, on peut observer par exemple que pour la variable **livraison**, une valeur élevée (donc perception positive) tend à classer un client dans le groupe 3 ; en effet, la valeur élevée augmente la valeur de la fonction 1 puisque le coefficient est positif, et diminue la valeur de la fonction 2 puisque le coefficient est négatif. Or une valeur élevée pour la fonction

1 et une valeur basse pour la fonction 2 correspond au groupe des plus de 5 ans si on se fie aux centroïdes. On peut ainsi regarder l'effet de chacune des variables pour bien comprendre ce qui distingue chacun des groupes.

Prior Probabilities for Groups			
type	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Moins d'un an	,373	31	31,000
1 à 5 ans	,337	28	28,000
Plus de 5 ans	,289	24	24,000
Total	1,000	83	83,000

FIG. 18.51 – Les probabilités a priori

La figure 18.51 nous montre qu'a priori, il y a 37,3 % des clients dans le premier groupe, 33,7 % dans le deuxième et 28,9 % dans le troisième.

Le graphe de la figure 18.52 ne peut être obtenu qu'avec une analyse qui compte au moins deux fonctions discriminantes. Elle permet simplement de visualiser la position de chaque individu selon ses scores discriminants. Chaque individu est identifié par son groupe.

Finalement, la matrice de classification nous montre que 86,7 % des cas ont été bien classés pour l'échantillon d'analyse. Ce pourcentage est assez bien réparti dans les trois groupes, ce qui donne l'impression que la classification est très bonne (la probabilité de bien classer chacun des cas par chance est de 33,3 % ici). Il faut cependant vérifier avec l'échantillon de validation.

Le pourcentage global descend à 76,5 % dans l'échantillon de validation. Ceci est dû au fait que dans le deuxième groupe, seulement 4 cas sur 7 ont été bien classés. Il faudrait investiguer davantage de ce côté...

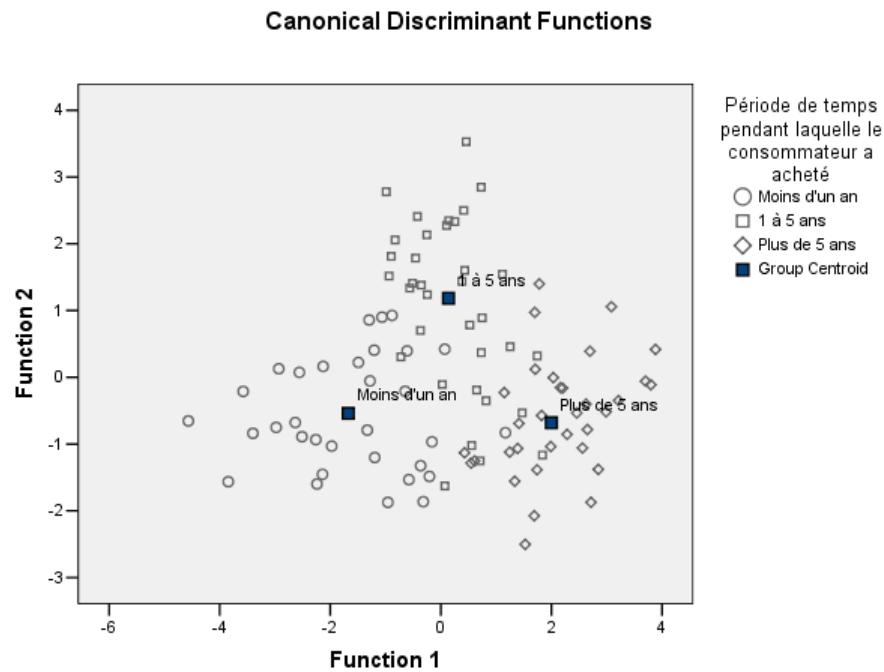


FIG. 18.52 – Le graphe des scores discriminants

Classification Results ^{a,b}							
Cases Selected	Original	Count	type	Predicted Group Membership			Total
				Moins d'un an	1 à 5 ans	Plus de 5 ans	
Cases Selected	Original	Count	Moins d'un an	26	4	1	31
			1 à 5 ans	2	23	3	28
			Plus de 5 ans	0	1	23	24
	%		Moins d'un an	83,9	12,9	3,2	100,0
			1 à 5 ans	7,1	82,1	10,7	100,0
			Plus de 5 ans	,0	4,2	95,8	100,0
Cases Not Selected	Original	Count	Moins d'un an	1	0	0	1
			1 à 5 ans	0	4	3	7
			Plus de 5 ans	0	1	8	9
	%		Moins d'un an	100,0	,0	,0	100,0
			1 à 5 ans	,0	57,1	42,9	100,0
			Plus de 5 ans	,0	11,1	88,9	100,0

a. 86,7% of selected original grouped cases correctly classified.

b. 76,5% of unselected original grouped cases correctly classified.

FIG. 18.53 – La matrice de classification

Je vous laisse vérifier les pré-requis formellement avec les sorties qui suivent.

Test Results		
Box's M		330,664
F	Approx.	1,384
df1		182
df2		15588,538
Sig.		,001

Tests null hypothesis of equal population covariance matrices.

FIG. 18.54 – Le test de Box's M

Model	Coefficients ^a			t	Sig.	Collinearity Statistics	
	Unstandardized Coefficients		Standardized Coefficients			Tolerance	VIF
	B	Std. Error	Beta				
1	(Constant)	-,962	1,027	-,937	,352		
	qualite	,263	,052	,439	5,062	,000	,586 1,708
	web	-,330	,132	-,265	-2,500	,015	,393 2,544
	support	-,013	,069	-,023	-,186	,853	,277 3,605
	probres	-,031	,098	-,046	-,313	,755	,205 4,887
	pub	,005	,058	,008	,092	,927	,656 1,524
	variete	-,130	,239	-,214	-,543	,589	,028 35,151
	forcevente	,207	,097	,257	2,137	,036	,305 3,282
	prik	-,115	,046	-,214	-2,511	,014	,604 1,654
	garanties	,054	,127	,055	,427	,670	,266 3,763
	nouveau	-,026	,038	-,048	-,676	,501	,878 1,138
	commandes	-,059	,104	-,067	-,568	,572	,321 3,120
	flexibilite	-,163	,244	-,244	-,669	,506	,033 30,112
	livraison	,940	,460	,873	2,041	,045	,024 41,540

a. Dependent Variable: type

FIG. 18.55 – Les VIF

		Tests of Normality					
type		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
qualité	Moins d'un an	,114	31	,200*	,967	31	,441
	1 à 5 ans	,163	28	,055	,906	28	,016
	Plus de 5 ans	,130	24	,200*	,948	24	,241
web	Moins d'un an	,168	31	,025	,939	31	,079
	1 à 5 ans	,149	28	,111	,928	28	,056
	Plus de 5 ans	,119	24	,200*	,971	24	,683
support	Moins d'un an	,093	31	,200*	,968	31	,473
	1 à 5 ans	,131	28	,200*	,955	28	,269
	Plus de 5 ans	,124	24	,200*	,952	24	,303
probres	Moins d'un an	,083	31	,200*	,978	31	,741
	1 à 5 ans	,073	28	,200*	,991	28	,995
	Plus de 5 ans	,103	24	,200*	,965	24	,536
pub	Moins d'un an	,127	31	,200*	,937	31	,068
	1 à 5 ans	,078	28	,200*	,981	28	,882
	Plus de 5 ans	,133	24	,200*	,927	24	,084
variété	Moins d'un an	,112	31	,200*	,976	31	,682
	1 à 5 ans	,074	28	,200*	,973	28	,677
	Plus de 5 ans	,119	24	,200*	,945	24	,206
forcevente	Moins d'un an	,158	31	,046	,957	31	,247
	1 à 5 ans	,145	28	,137	,924	28	,042
	Plus de 5 ans	,123	24	,200*	,963	24	,507
prix	Moins d'un an	,126	31	,200*	,955	31	,215
	1 à 5 ans	,168	28	,042	,937	28	,095
	Plus de 5 ans	,197	24	,017	,924	24	,071
garanties	Moins d'un an	,116	31	,200*	,970	31	,529
	1 à 5 ans	,085	28	,200*	,985	28	,951
	Plus de 5 ans	,101	24	,200*	,955	24	,348
nouveau	Moins d'un an	,083	31	,200*	,971	31	,543
	1 à 5 ans	,098	28	,200*	,977	28	,776
	Plus de 5 ans	,178	24	,049	,934	24	,118
commandes	Moins d'un an	,111	31	,200*	,963	31	,346
	1 à 5 ans	,192	28	,010	,902	28	,013
	Plus de 5 ans	,128	24	,200*	,977	24	,828
flexibilité	Moins d'un an	,112	31	,200*	,951	31	,163
	1 à 5 ans	,167	28	,043	,928	28	,054
	Plus de 5 ans	,141	24	,200*	,925	24	,075
livraison	Moins d'un an	,072	31	,200*	,982	31	,863
	1 à 5 ans	,107	28	,200*	,979	28	,836
	Plus de 5 ans	,095	24	,200*	,955	24	,353

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. 18.56 – Les tests de normalité

18.6 Exercices du chapitre

Exercice 1 Reprenez l'exemple de la section 18.5. Maintenant que vous avez un bon aperçu de la situation, tentez de produire une analyse discriminante moins lourde (avec moins de variables) que celle de l'exemple, mais qui produit quand même des résultats satisfaisants.

Exercice 2 Dans un exercice, vous aviez à créer des clusters dans la base de données `Telco.sav`, vous permettant ainsi d'identifier les profils des consommateurs. À l'aide d'une analyse discriminante, tentez de voir si vous arrivez à classer les clients selon leur profil en utilisant les variables démographiques disponibles. Une autre alternative est de remplacer la variable des clusters avec la variable `catcons` qui représente la catégorie du consommateur.

Exercice 3 Reprenez l'exemple 18.4.1. Il y avait un problème de multicolinéarité, et on se doute donc qu'il est possible de faire une meilleure analyse discriminante. Faites-le directement avec la base de données `fournisseur.sav`, ou à partir des sorties qui suivent.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,188	,309	,607	,545		
	livraison	,100	,027	,268	3,629	,000	,659 1,517
	prix	,010	,030	,024	,335	,738	,683 1,464
	flexibilité	,119	,027	,335	4,398	,000	,618 1,619
	produit	-,129	,023	-,416	-5,601	,000	,649 1,541

a. Dependent Variable: `x11`

FIG. 18.57 – Les VIF pour les quatre variables qui avaient les meilleurs résultats aux ANOVA individuelles

Tests of Normality						
x11	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
flexibilite	Specification Buying	,113	33	,200*	,981	33
	Total Value Analysis	,148	45	,015	,878	45
produit	Specification Buying	,128	33	,183	,971	33
	Total Value Analysis	,118	45	,126	,969	45
livraison	Specification Buying	,076	33	,200*	,992	33
	Total Value Analysis	,090	45	,200*	,972	45
prix	Specification Buying	,095	33	,200*	,974	33
	Total Value Analysis	,104	45	,200*	,941	45

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. 18.58 – Les tests de normalité

Analysis Case Processing Summary			
Unweighted Cases		N	Percent
Valid		78	78,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Unselected	22	22,0
	Total	22	22,0
	Total	100	100,0

FIG. 18.59 – Première sortie de l'analyse avec les quatre variables

Test Results		
Box's M	35,851	
F	Approx.	3,374
	df1	10
	df2	22357,862
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

FIG. 18.60 – Le test du Box's M pour les quatre variables

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,315	85,456	4	,000

FIG. 18.61 – Le Wilks' Lambda de la fonction avec les 4 variables

Structure Matrix	
	Function
	1
produit	-,729
flexibilite	,560
livraison	,498
prix	-,324

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

FIG. 18.62 – Les loadings des 4 variables

Canonical Discriminant Function Coefficients	
	Function
	1
livraison	,309
prix	,069
flexibilite	,513
produit	-,674
(Constant)	-,394

Unstandardized coefficients

FIG. 18.63 – Les coefficients de la fonction

Functions at Group Centroids	
	Function
x11	1
Specification Buying	-1,699
Total V alue Analysis	1,246

Unstandardized canonical discriminant functions evaluated at group means

FIG. 18.64 – Les centres moyens de la fonction par groupe

			Classification Results ^{a,b}		
			Predicted Group Membership		Total
x11			Specification Buying	Total Value Analysis	
Cases Selected	Original	Count	Specification Buying	32	1
			Total V alue Analysis	8	37
	%		Specification Buying	97,0	3,0
			Total V alue Analysis	17,8	82,2
Cases Not Selected	Original	Count	Specification Buying	4	3
			Total V alue Analysis	0	15
	%		Specification Buying	57,1	42,9
			Total V alue Analysis	,0	100,0

a. 88,5% of selected original grouped cases correctly classified.

b. 86,4% of unselected original grouped cases correctly classified.

FIG. 18.65 – La matrice de classification

Classification Results ^{a,b}					
x11			Predicted Group Membership		Total
			Specification Buying	Total Value Analysis	
Cases Selected	Original	Count	Specification Buying	32	33
			Total Value Analysis	3	45
			%	97,0	100,0
Cases Not Selected	Original	Count	Specification Buying	5	7
			Total Value Analysis	0	15
			%	71,4	100,0
			Total Value Analysis	,0	100,0

a. 94,9% of selected original grouped cases correctly classified.

b. 90,9% of unselected original grouped cases correctly classified.

FIG. 18.66 – La matrice de classification avec les sept variables

Chapitre 19

Régression logistique

La régression logistique est une analyse qui nous permet d'expliquer une variable dépendante discrète Y dichotomique à l'aide de variables indépendantes continues et discrètes. Lorsque Y n'est pas dichotomique, il faut se tourner vers la régression logistique multinomiale, sujet du prochain chapitre.

Puisque la variable dépendante est ici une variable discrète dichotomique, nous verrons dans la section 19.1 que la régression linéaire classique n'est pas adaptée dans ce cas. Il ne sera alors pas possible d'utiliser le principe des moindres carrés pour trouver les coefficients de l'équation de la régression logistique. Il faudra plutôt utiliser le principe du maximum de vraisemblance, que nous abordons dans la section 19.2. Ensuite la section 19.3 présente la fonction logistique sur laquelle se base le modèle de régression logistique. Finalement, la section 19.4 présente un exemple complet d'analyse d'une régression logistique, suivie de la dernière section qui présente un autre exemple.

19.1 Le modèle linéaire

Considérons le modèle de régression linéaire simple :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

où Y_i est une variable discrète dichotomique qui prend les valeurs 0 et 1. La valeur espérée $E(Y_i)$ a une interprétation particulière dans ce contexte. Tout d'abord, puisque $E(\epsilon_i) = 0$, on a

$$E(Y_i) = \beta_0 + \beta_1 X_i.$$

On peut considérer Y_i comme étant une variable aléatoire de Bernoulli dont la loi de probabilité s'écrit comme suit :

Y_i	Probabilité
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Donc π_i est la probabilité que $Y_i = 1$, et $1 - \pi_i$ est la probabilité que $Y_i = 0$. Par définition de l'espérance d'une variable aléatoire, on obtient

$$E(Y_i) = 1 \cdot \pi_i + 0 \cdot (1 - \pi_i) = \pi_i$$

et donc

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i.$$

La moyenne (valeur espérée) de Y_i donnée par l'équation est donc simplement la probabilité que $Y_i = 1$ lorsque la valeur de la variable indépendante est X_i . Cette interprétation tient lorsque l'équation pour prédire Y est celle d'une régression linéaire, simple ou multiple.

Mais certains problèmes se présentent lorsqu'on utilise un modèle linéaire pour une variable indépendante dichotomique. Voici trois de ces problèmes :

1. Erreurs qui ne suivent pas une loi normale. Puisque Y_i ne prend que deux valeurs, il en va de même pour $\epsilon_i (= Y_i - (\beta_0 + \beta_1 X_i))$:

$$\text{Lorsque } Y_i = 1 : \epsilon_i = 1 - \beta_0 - \beta_1 X_i;$$

$$\text{Lorsque } Y_i = 0 : \epsilon_i = -\beta_0 - \beta_1 X_i.$$

2. Variance non constante de l'erreur. Un autre problème avec les erreurs ϵ_i , c'est qu'elles ne sont pas de même variance lorsque la variable dépendante est dichotomique. En effet, tout d'abord, on a

$$\sigma^2(Y_i) = E(Y_i^2) - [E(Y_i)]^2 = \pi_i - \pi_i^2 = \pi_i(1 - \pi_i).$$

La variance de ϵ_i est la même que celle de Y_i puisque $\epsilon_i = Y_i - \pi_i$ et π_i est une constante. Donc

$$\sigma^2(\epsilon_i) = \pi_i(1 - \pi_i).$$

Ainsi $\sigma^2(\epsilon_i)$ dépend de X_i , et donc la variance changera selon les valeurs de X . Par conséquent la méthode des moindres carrés ne sera pas optimale dans ce contexte.

3. Contraintes sur les valeurs prédictes. Puisque les valeurs prédictes représentent des probabilités lorsque Y est une variable dichotomique, les valeurs prédictes par l'équation devraient être comprises entre 0 et 1 ; la plupart des équations ne respecteront pas automatiquement cette contrainte.

C'est le troisième problème (que les valeurs prédictes soient entre 0 et 1) qui est le plus sérieux. Il sera possible de le résoudre en utilisant une fonction en forme de S (fonction logistique) plutôt que le modèle linéaire classique. Et au lieu d'utiliser la méthode des moindres carrés pour estimer les paramètres du modèle, c'est le principe du maximum de vraisemblance qui sera utilisé.

19.2 Principe du maximum de vraisemblance

Dans une analyse, lorsqu'on veut estimer des paramètres, plus d'une méthode s'offre à nous. Par exemple, pour estimer les paramètre de l'équation d'une régression linéaire, c'est la méthode des moindres carrés qui donne les meilleurs résultats. Une autre méthode pour estimer des paramètres est celle du maximum de vraisemblance ; cette méthode revient simplement à estimer les paramètres de façon à « coller » le mieux possible à l'échantillon. L'exemple qui suit illustre de façon très simple ce principe.

Exemple 19.2.1 Considérons un cas où les données se distribuent selon une loi normale au niveau de la population et dont l'écart-type est $\sigma = 10$. On recueille un échantillon de 3 données : $y_1 = 250$, $y_2 = 265$ et $y_3 = 259$, et on veut estimer la valeur de μ qui s'ajuste le mieux à l'échantillon. Supposons que $\mu = 230$. La figure 19.1 montre la distribution d'une loi normale de paramètres $\mu = 230$ et $\sigma = 10$ ainsi que la position des trois valeurs de l'échantillon par rapport à cette distribution.

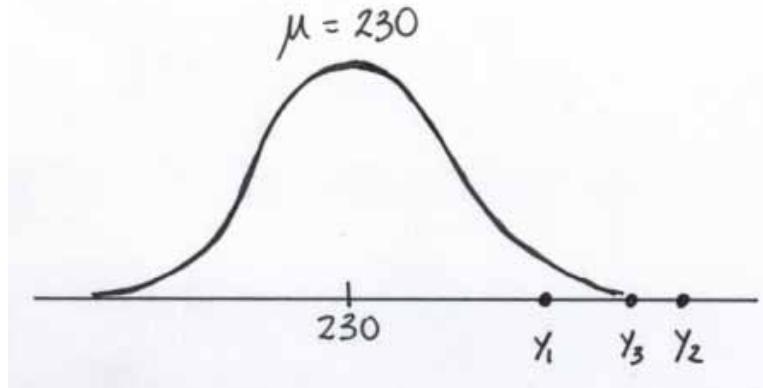


FIG. 19.1 –

On remarque que les valeurs de l'échantillon se retrouvent dans la queue droite de la distribution, ce qui montre que si la moyenne est de 230, la probabilité de tomber sur ces valeurs n'est pas grande. Donc $\mu = 230$ ne concorde pas bien avec cet échantillon.

La figure 19.2 montre ce qui se passe si on prend plutôt $\mu = 259$. Les observations de l'échantillon se retrouvent alors au centre de la distribution, et donc $\mu = 259$ s'ajuste

mieux à l'échantillon.

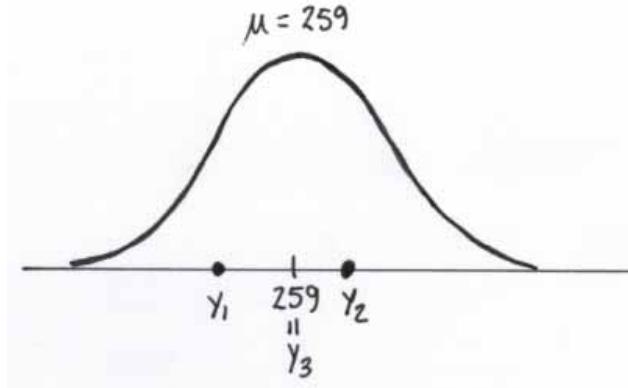


FIG. 19.2 –

La méthode du maximum de vraisemblance utilise la fonction de densité de la distribution évaluée sur les observations y_i de l'échantillon (ce qui correspond à la hauteur de la courbe aux points y_i) pour mesurer à quel point l'estimation du paramètre s'ajuste bien aux données. Dans l'exemple, si on prend y_1 , on voit que la hauteur de la courbe lorsque $\mu = 230$ est beaucoup plus petite que lorsque $\mu = 259$, ce qui signifie que $\mu = 259$ s'ajuste mieux à cette donnée que $\mu = 230$. De façon plus précise, on peut calculer les densités f au point y_1 selon le paramètre μ (σ est fixé à 10) :

$$\mu = 230, \quad y_1 = 250 : \quad f(y_1; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \cdot 10} \cdot e^{-\frac{1}{2} \left(\frac{250-230}{10} \right)^2} = 0,005399.$$

$$\mu = 259, \quad y_1 = 250 : \quad f(y_1; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \cdot 10} \cdot e^{-\frac{1}{2} \left(\frac{250-259}{10} \right)^2} = 0,026609.$$

Les densités pour les trois observations de l'échantillon pour les deux valeurs de μ sont les suivantes :

	$\mu = 230$	$\mu = 259$
$f(y_1; \mu, \sigma)$	0,005399	0,026609
$f(y_2; \mu, \sigma)$	0,000087	0,033322
$f(y_3; \mu, \sigma)$	0,000595	0,039894

Ainsi, pour chacune des observations prises séparément, on voit que c'est $\mu = 259$ qui donne les meilleures densités. Mais il faut aussi obtenir une mesure pour tout l'échantillon, pour voir à quel point l'estimé concorde bien avec l'ensemble des observations. Pour ce faire on prend le produit des densités ; ce produit est appelé la **fonction de vraisemblance** et est noté $L(\mu, \sigma)$. Lorsque les paramètres concordent bien avec l'échantillon, les densités sont plus élevées, et donc la valeur de la fonction de vraisemblance aussi. Ainsi le principe du maximum de vraisemblance vise simplement à trouver les estimés des paramètres qui maximisent la fonction de vraisemblance.

Dans le cadre de l'exemple, on a

$$L(\mu = 230, \sigma = 10) = 0,005399 \cdot 0,000087 \cdot 0,000595 = 0,279 \times 10^{-9}$$

$$L(\mu = 259, \sigma = 10) = 0,026609 \cdot 0,033322 \cdot 0,039894 = 0,0000354$$

Puisque $L(\mu = 230, \sigma = 10) < L(\mu = 259, \sigma = 10)$, on conclut comme précédemment que c'est $\mu = 259$ qui concorde le mieux avec l'échantillon. Mais est-ce la meilleure valeur ? La meilleure valeur est celle qui permet à la fonction de vraisemblance d'atteindre son maximum. Dans le cas d'une loi normale (comme dans cet exemple) c'est la moyenne échantillonnale \bar{y} qui estime le mieux μ au sens du maximum de vraisemblance. Donc dans le cadre de l'exemple c'est $\mu = 258 = \bar{y}$ qui maximise la fonction de vraisemblance. On a $L(\mu = 258, \sigma = 10) = 0,0000359$, ce qui est un tout petit plus grand que ce qu'on avait obtenu avec $\mu = 259$.

Si au lieu de fixer σ^2 on avait voulu l'estimer selon le maximum de vraisemblance, on aurait trouvé que le meilleur estimateur est

$$\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2,$$

qui n'est pas la variance échantillonnale

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Cet estimateur ($\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$) est biaisé pour σ^2 ; en effet, rappelons que s^2 est sans biais pour σ^2 . La méthode du maximum de vraisemblance ne garantit pas des estimateurs sans biais. Cependant, ils sont généralement convergents.

De façon plus générale, si les paramètres à estimer sont $\theta_1, \theta_2, \dots, \theta_m$, alors la fonction de vraisemblance s'écrit

$$\begin{aligned} L(\theta_1, \theta_2, \dots, \theta_m) &= f(y_1, y_2, \dots, y_n; \theta_1, \theta_2, \dots, \theta_m) \\ &= f(x_1; \theta_1, \theta_2, \dots, \theta_m) \cdots f(x_n; \theta_1, \theta_2, \dots, \theta_m) \\ &= \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_m). \end{aligned}$$

19.3 La fonction logistique

On a vu à la section 19.1 que certains problèmes se posent si on tente d'utiliser le modèle de régression linéaire avec une variable dépendante dichotomique. Le problème le plus sérieux est que l'on désire que les valeurs prédites se retrouvent toujours entre 0 et 1; le modèle de régression logistique permet de résoudre ce problème. Ce modèle découle de la distribution logistique, une loi de probabilité très semblable à celle de la loi normale. Ce modèle est le suivant :

$$\ln \left(\frac{P(Y = 1)}{P(Y = 0)} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon.$$

La quantité

$$\ln \left(\frac{P(Y = 1)}{P(Y = 0)} \right)$$

porte le nom de Logit, tandis que la quantité

$$\frac{P(Y = 1)}{P(Y = 0)} = \frac{\text{Probabilité événement}}{\text{Probabilité non événement}} = \frac{P(Y = 1)}{1 - P(Y = 1)}$$

s'appelle un *odds*.

Il faut comprendre qu'un *odds* n'est pas une probabilité. Par exemple, si on pique une carte dans un jeu de cartes conventionnel et qu'on désire ne pas piger une carte de cœur,

le *odds* de piger autre chose que du cœur est de 3 contre 1 :

$$\text{odds} (\text{ne pas piger du coeur}) = \frac{P(\text{ne pas piger du coeur})}{P(\text{piger du coeur})} = \frac{3/4}{1/4} = \frac{3}{1}.$$

Ainsi, dans le modèle de régression logistique, lorsqu'une des variables indépendantes X_i augmente d'une unité, on ne peut interpréter l'impact directement sur Y avec le paramètre β_i comme on le fait avec un modèle linéaire. En effet, lorsque X_i augmente d'une unité, c'est le logarithme naturel (aussi appelé logarithme népérien) du *odds* qui augmente ou diminue (selon le signe de β_i).

Si on veut plutôt parler en terme de *odds*, il suffit d'appliquer l'exponentielle à l'équation et on obtient alors

$$\text{odds} = \frac{P(Y = 1)}{P(Y = 0)} = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon} = e^{\beta_0} \cdot e^{\beta_1 X_1} \dots e^{\beta_k X_k} \cdot e^{\epsilon}.$$

Ainsi, lorsqu'une des variables X_i augmente d'une unité, l'effet sur le *odds* est multiplicatif puisque $e^{\beta_i(X_i+1)} = e^{\beta_i X_i + \beta_i} = e^{\beta_i X_i} \cdot e^{\beta_i}$. En outre, la fonction exponentielle prend une valeur plus grande que 1 lorsque son exposant est positif, et entre 0 et 1 lorsque l'exposant est négatif. Donc lorsque β_i est positif, le *odds* augmentera lorsque X_i augmente ; inversement, lorsque β_i est négatif, le *odds* diminuera lorsque X_i augmentera. Si le paramètre est nul, on a $e^0 = 1$, et donc le *odds* ne changera pas.

Il est également possible de transformer l'équation de façon à travailler avec la probabilité de l'événement plutôt qu'avec le *odds*. Rappelons d'abord que

$$\text{odds} = \frac{P(Y = 1)}{P(Y = 0)} = \frac{\text{Probabilité événement}}{\text{Probabilité non événement}} = \frac{P(Y = 1)}{1 - P(Y = 1)}$$

donc

$$\text{odds} = \frac{P(Y = 1)}{1 - P(Y = 1)} = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon}.$$

$$\begin{aligned} \text{Alors } P(Y = 1) &= e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon} \cdot (1 - P(Y = 1)) \\ &= e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon} - (e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon} \cdot P(Y = 1)) \end{aligned}$$

$$\text{ainsi } P(Y = 1) \cdot (1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon}) = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon}$$

$$\begin{aligned} \text{et donc } P(Y = 1) &= \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon}} \\ &= \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon)}} \\ &= (1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon)})^{-1}. \end{aligned}$$

Ce modèle nous assure que la probabilité de l'événement $P(Y = 1)$ sera bien estimée en ce sens que ce modèle prend toujours des valeurs entre 0 et 1. De plus, tout comme la régression linéaire, les variables indépendantes de ce modèle peuvent être continues ou discrètes, en autant que ces dernières soient codifiées de façon binaire.

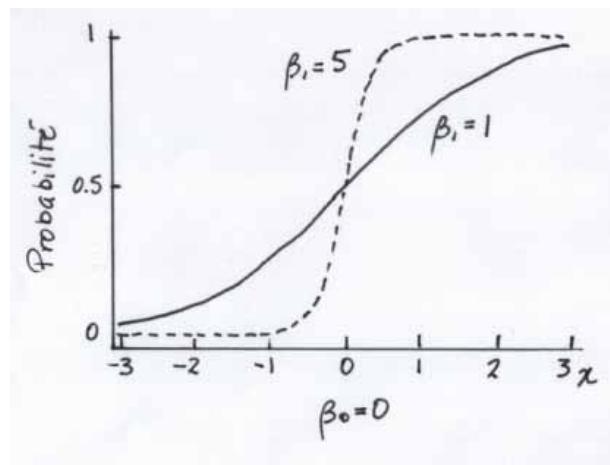
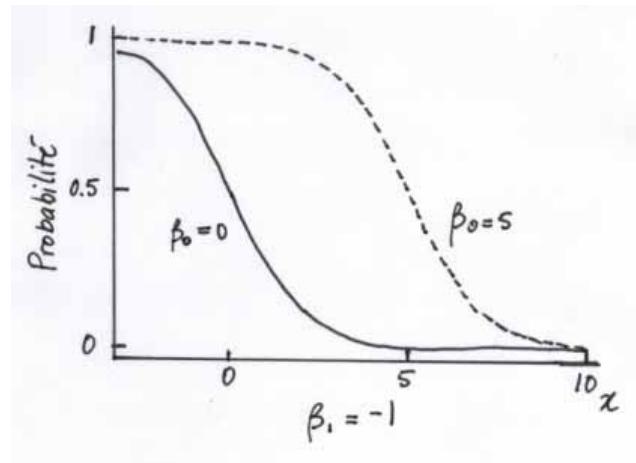


FIG. 19.3 – Fonctions logistiques avec $\beta_0 = 0$

Les figures 19.3 et 19.4 montrent les graphes de fonctions logistiques selon des valeurs de β_0 et de β_1 .

FIG. 19.4 – Fonctions logistiques avec $\beta_1 = -1$

19.4 Un exemple complet

Nous présentons maintenant un exemple illustrant toutes les étapes d'une analyse en régression logistique.

Exemple 19.4.1 Prenons la base de données `fournisseurplus.sav`. On aimerait comprendre ce qui fait qu'une entreprise cliente du fournisseur peut vouloir ou pas considérer une alliance stratégique. Pour ceci on utilisera certaines des variables mesurant la satisfaction des entreprises sur différents aspects, ainsi qu'une variable binaire indiquant si l'entreprise est située en Amérique du Nord ou ailleurs. Voici la description des variables :

alliance	Voudrait bien considérer une alliance stratégique (<code>non</code> = 0, <code>oui</code> = 1)
region	Région de la firme (<code>Amérique du Nord</code> = 0, <code>En-dehors de l'Amérique du Nord</code> = 1)
qualite	Qualité perçue du produit*
probres	Perception de la manière dont sont résolus les problèmes*
forcevente	Image de la force de vente*
nouveau	Perception : développement et vente de nouveaux produits*

*Pauvre = 0, Excellente = 10

Fixons les seuils à $\alpha = 0,05$. Les commandes pour obtenir les sorties de cet exemple sont les suivantes :

Menu SPSS :	→ Analyse → Regression → Binary Logistic...
Dans la fenêtre Dependent :	→ alliance (la variable dépendante binaire)
Dans la fenêtre Covariates :	→ region, qualite, probres, forcevente, nouveau (les variables indépendantes)
Dans le bouton Categorical :	Dans la fenêtre Categorical Covariates : → region (variables indépendantes qui sont discrètes) Laisser les paramètres par défaut : Contrast : Indicator Reference Category : <input checked="" type="checkbox"/> Last
Dans le bouton Save... :	Predicted Values <input checked="" type="checkbox"/> Probabilities <input checked="" type="checkbox"/> Group membership Influence <input checked="" type="checkbox"/> Cook's <input checked="" type="checkbox"/> Leverage values Residuals <input checked="" type="checkbox"/> Unstandardized <input checked="" type="checkbox"/> Standardized
Dans le bouton Options... :	<input checked="" type="checkbox"/> Classification plots <input checked="" type="checkbox"/> Iteration history

La figure 19.5 montre d'abord qu'il y a 100 données, qu'elles sont toutes valides et qu'il n'y a pas d'échantillon de validation (pas de données non sélectionnées)). On voit aussi la codification de la variable dépendante (**alliance**) ; le **non** est codé 0, et le **oui** est codé 1.

Case Processing Summary		
Unweighted Cases ^a	N	Percent
Selected Cases	Included in Analysis	100 100,0
	Missing Cases	0 ,0
	Total	100 100,0
Unselected Cases		0 ,0
Total		100 100,0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding	
Original Value	Internal Value
non	0
oui	1

Categorical Variables Codings		
	Frequency	Parameter
region Amérique du Nord	39	1,000
En-dehors de l'Amérique du Nord	61	,000

FIG. 19.5 – Les données et codification des variables discrètes

Ceci est important puisque la probabilité calculée par le modèle est celle de l'événement codé 1, donc ici la probabilité que l'entreprise considère une alliance stratégique.

L'autre sortie montre la codification de la variable indépendante discrète `region`, ainsi que la fréquence associée à chaque modalité. Il y a 39 entreprises qui sont en Amérique du Nord, et 61 qui sont ailleurs.

ATTENTION : on voit que l'Amérique du Nord est codifiée 1 pour les calculs, mais dans la base de données elle est codifiée 0... pourquoi faire simple quand on peut faire compliqué !

19.4.1 Le modèle constant

Pour commencer, un premier modèle est calculé sans aucune variable indépendante. D'une certaine façon ce modèle représente la situation où aucune des variables indépendantes n'a d'influence sur la variable dépendante. Si les variables indépendantes aident vraiment à expliquer la dépendante, alors le modèle calculé avec celles-ci devrait être meilleur que le modèle avec seulement une constante.

Iteration History ^{a,b,c}		
Iteration	-2 Log likelihood	Coefficients
		Constant
Step 1	137,628	-.200
0 2	137,628	-.201

- a. Constant is included in the model.
- b. Initial -2 Log Likelihood: 137,628
- c. Estimation terminated at iteration number 2 because parameter estimates changed by less than ,001.

FIG. 19.6 – Le -2LL du modèle sans variables indépendantes

Pour pouvoir comparer ces deux modèles, on utilise la valeur **-2 Log likelihood**. Le **likelihood**, appelé la vraisemblance en français, est la probabilité d'obtenir les résultats observés étant donné les paramètres estimés. Cette probabilité est transformée en lui appliquant un logarithme puis en multipliant celui-ci par -2 (on appellera cette quantité -2LL) ; l'intérêt de cette transformation réside dans le fait que sous certaines conditions, on obtient alors une statistique dont la distribution est connue (par exemple, si X suit une loi uniforme, alors $-2 \log X$ suit la loi χ_1^2).

La statistique -2LL est utilisée dans le modèle de régression logistique comme une mesure d'ajustement du modèle aux données observées. Un bon modèle obtient une grande probabilité de vraisemblance, ce qui se traduit par de petites valeurs de la statistique -2LL. Un modèle parfait obtiendrait une vraisemblance de 1 (**likelihood = 1**), et donc une valeur -2LL = 0. La sortie 19.6 montre que le -2LL du modèle initial (sans variables indépendantes) est de 137,628. On espère donc que l'ajout des variables indépendantes donnera un modèle dont le -2LL sera plus petit que 137,628.

Le modèle initial n'ayant pas de variables indépendantes, ses prévisions seront les mêmes pour toutes les données. Donc toutes les observations seront classées dans le même groupe. Puisque l'algorithme du maximum de vraisemblance donne le modèle qui colle le mieux aux données, celles-ci seront toutes classées dans le groupe le plus volumineux, maximisant ainsi le taux de classifications correctes. La figure 19.7 nous montre que dans

Observed		Predicted		Percentage Correct	
		Voudrait bien considérer une alliance stratégique			
		non	oui		
Step 0	Voudrait bien considérer une alliance stratégique	non	55	100,0	
		oui	45	,0	
	Overall Percentage			55,0	

a. Constant is included in the model.

b. The cut value is ,500

FIG. 19.7 – La table de classification du modèle sans variables indépendantes

le cadre de l'exemple, il y a 55 non et 45 oui, et donc toutes les observations ont été classées dans le groupe du non, donnant ainsi un taux de réussite de 55 %.

Variables not in the Equation				
Step	Variables	Score	df	Sig.
0	region(1)	.034	1	.853
	qualité	10.551	1	.001
	probres	26.876	1	.000
	forcevente	12.631	1	.000
	nouveau	.648	1	.421
	Overall Statistics	44.912	5	.000

FIG. 19.8 – L'apport des variables prises individuellement

La figure 19.8 contient des scores qui indiquent à quel point chacune des variables prise individuellement a un impact sur la variable dépendante. Ici c'est la variable probres qui a le score le plus élevé (il a une valeur de 26,876), et donc le plus grand impact. Si on appliquait une méthode de type *forward*, ce serait la première variable à entrer dans le modèle. Et puisque ces scores dépendent de l'apport **individuel** de chacune des variables, une variable avec un grand score peut quand même se retrouver non significative en présence d'autres variables.

19.4.2 Le modèle est-il bon ?

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	76.808 ^a	.456	.610

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

FIG. 19.9 – Le -2LL du modèle avec les variables indépendantes

La figure 19.9 donne le -2LL du modèle avec toutes les variables indépendantes. Il a ici une valeur de 76,808, ce qui est plus petit que le -2LL = 137,628 du modèle sans variables indépendantes. Il semble donc que ces variables améliorent le classement.

D'autres statistiques apparaissent dans la sortie 19.9 : le r^2 de Cox & Snell et le r^2 de Nagelkerke. Ces deux statistiques tentent de mesurer la proportion de la variation expliquée par la régression. Plus précisément, la mesure de Cox & Snell s'exprime de la manière suivante :

$$r_{\text{Cox \& Snell}}^2 = 1 - \left(\frac{L(0)}{L(B)} \right)^{\frac{2}{n}}$$

avec $L(0)$ la vraisemblance du modèle ne contenant que la constante, $L(B)$ la vraisemblance du modèle étudié à cette étape et n la taille de l'échantillon (ne pas confondre la vraisemblance avec -2LL). Le problème avec cette statistique provient du fait que la valeur maximale qu'elle peut prendre est un peu en-dessous de 1, alors qu'un « vrai » r^2 doit pouvoir atteindre le 1. En 1991, Nagelkerke proposa d'ajuster la statistique précédente de façon à corriger ce problème. Cet ajustement donne la statistique suivante :

$$r_{\text{Nagelkerke}}^2 = \frac{r_{\text{Cox \& Snell}}^2}{\max r_{\text{Cox \& Snell}}^2}.$$

Ainsi, c'est habituellement le r^2 de Nagelkerke que l'on interprète. Dans le cadre de cet exemple, il semble que 61 % de la variation de la variable dépendante est expliquée par le présent modèle de régression logistique. Ainsi le modèle semble significatif ; il est

possible de tester ceci formellement avec la p -value de la sortie 19.10. En effet, le test omnibus permet de confronter les hypothèses suivantes :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \text{Au moins un des } \beta_j \neq 0 \ (1 \leq j \leq k)$$

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	60.820	5	.000
	Block	60.820	5	.000
	Model	60.820	5	.000

FIG. 19.10 – Le test omnibus

La valeur du χ^2 présentée dans la sortie 19.10 est simplement la différence entre $-2\text{LL}_{\text{modèle constant}}$ et $-2\text{LL}_{\text{modèle complet}}$. Et effectivement, on constate que

$$-2\text{LL}_{\text{modèle constant}} - (-2\text{LL}_{\text{modèle complet}}) = 137,628 - 76.808 = 60,820.$$

Plus cette différence est grande, plus le modèle semble significatif, et on sera alors porté à rejeter H_0 . C'est le cas ici puisque la p -value est nulle, ce qui est plus petit que $\alpha = 0,05$. Donc au risque de se tromper une fois sur 20 on rejette H_0 et on conclut que le modèle est significatif. Ceci signifie qu'au moins une des variables indépendantes a un impact sur le fait qu'une entreprise voudrait bien considérer une alliance stratégique ou pas.

Si une méthode *forward* ou *backward* est utilisée, la ligne **Step** permet alors de tester si la différence entre les modèles d'une étape à une autre est significative. Si la différence est significative, cela veut dire que l'ajout de la dernière variable dans le modèle est significative, et ce, malgré la présence des autres variables déjà incluses dans le modèle à l'étape précédente. Il est aussi possible d'entrer des variables en « bloc », en utilisant le bouton **Next block** dans la fenêtre principale de la régression logistique. La statistique du Chi-deux mesurera si, de bloc en bloc, -2LL diminue significativement (sur la ligne **Block**), illustrant ainsi l'amélioration du modèle de régression logistique.

		Predicted		Percentage Correct
		Voudrait bien considérer une alliance stratégique	Oui	
Observed	non	47	8	85.5
	Oui	7	38	84.4
Overall Percentage				85.0

a. The cut value is .500

FIG. 19.11 – La table de classification du modèle complet

Mais ce qui permettra avant tout de juger si un modèle performe bien, ce sont bien entendu les résultats de la classification. Ceux-ci sont contenus dans la table de classification (figure 19.11). Ainsi on voit que sur les 55 entreprises qui ne veulent pas considérer une alliance stratégique, 47 ont été bien classées, donnant un pourcentage de 85,5 % de bien classées. Pour l'autre groupe, 38 sur 45 ont été bien classées, donnant 84,4 % de bonnes classifications pour ce groupe. Dans l'ensemble, 85 % des cas ont été bien classés. Cette performance semble très bonne ; on a vu qu'avec le modèle constant le taux de réussite global était de 55 %.

L'histogramme de la figure 19.12 permet de visualiser la classification. Rappelons d'abord comment fonctionne celle-ci : pour chaque entreprise, le modèle calcule une probabilité, qui ici est la probabilité que cette entreprise veuille considérer une alliance stratégique (puisque c'est le oui qui est codé 1). Lorsque cette probabilité dépasse 50 %, l'entreprise est classée dans le groupe du oui, et sinon dans le groupe du non (car la **Classification cutoff** est fixée à 0,5 par défaut, mais il est possible de changer cette valeur). Si le modèle fait une bonne classification, la grande majorité des entreprises qui ont répondu oui devrait se retrouver à droite de la valeur 0,5. Inversement, les non devraient se retrouver à gauche. Ici l'examen de l'histogramme permet simplement de constater ce que la table de classification nous avait déjà révélé, c'est-à-dire que la majorité des oui et des non sont bien classés.

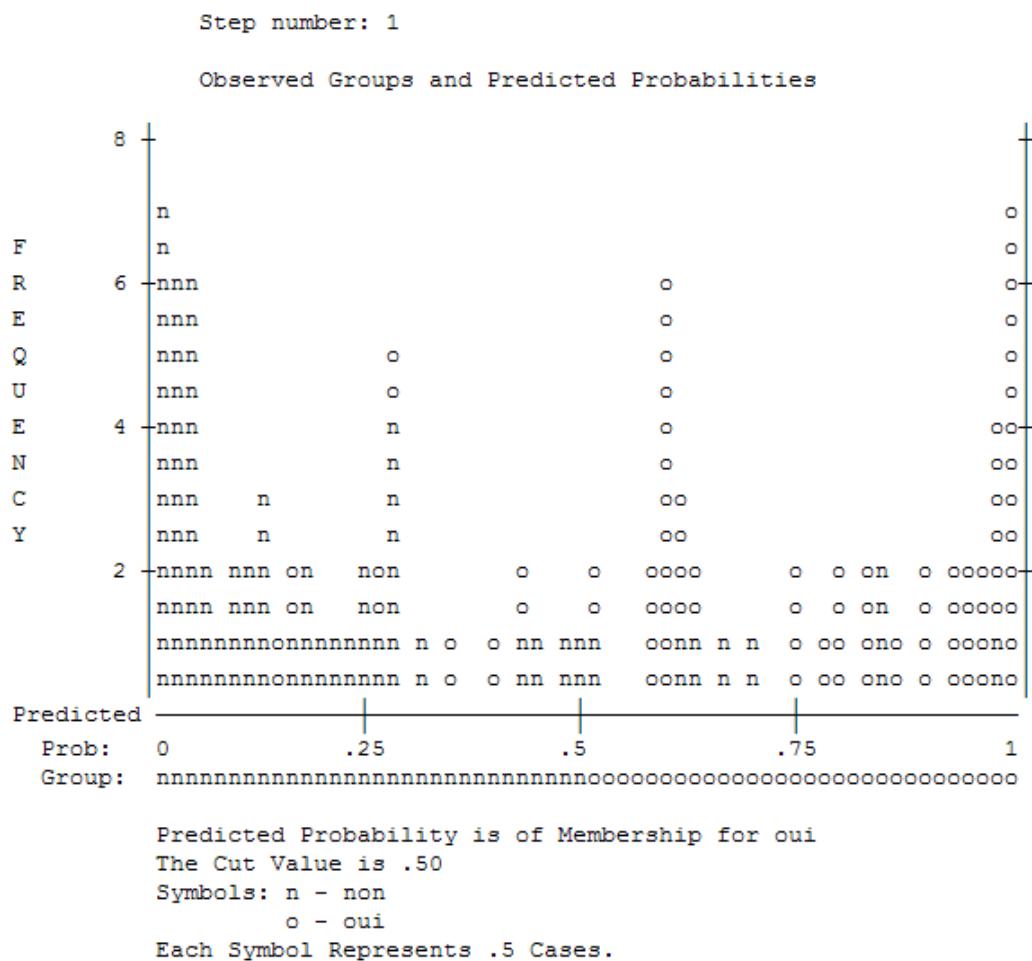


FIG. 19.12 – Histogramme des probabilités prédictées

Dans un problème donné, si les conséquences liées à une mauvaise classification ne sont pas les mêmes pour les deux groupes, il est possible de changer la borne de coupure (tel que mentionné précédemment). Par exemple, une telle approche est grandement utilisée en médecine où, en déplaçant le point de coupure, le médecin déclare des patients comme étant potentiellement cancéreux au moindre doute. Cette tactique permet d'aiguiller plus d'individus vers des examens plus approfondis. Ceci a certainement comme effet d'inquiéter des patients qui en réalité n'ont pas le cancer, mais cette conséquence est considérée moins problématique que de déclarer sain un individu qui ne l'est pas.

19.4.3 Les détails du modèle

La figure 19.13 permet d'écrire l'équation du modèle. Les coefficients de chacune des variables se retrouvent dans la colonne B.

Variables in the Equation						
Step	B	S.E.	Wald	df	Sig.	Exp(B)
1	region(1) -0.411	.796	.266	1	.606	.663
	qualite 1.054	.311	11.471	1	.001	2.870
	probres 1.427	.339	17.684	1	.000	4.168
	forcevente 1.133	.383	8.736	1	.003	3.104
	nouveau -0.365	.209	3.052	1	.081	.694
	Constant -20.110	4.502	19.949	1	.000	.000

a. Variable(s) entered on step 1: region, qualite, probres, forcevente, nouveau.

FIG. 19.13 – La table des coefficients

On a donc comme équation (abrégée, l'important c'est de comprendre le concept !) :

$$P(\text{alliance stratégique}) = \left(1 + e^{(-20,110 - 0,411x_{\text{region}} + 1,054x_{\text{qualité}} + \dots - 0,365x_{\text{nouveau}})}\right)^{-1}.$$

Cette équation permet donc de calculer la probabilité pour une entreprise qu'elle ait répondu oui à la question à propos de l'alliance stratégique. Par exemple, si une entreprise en Amérique du Nord a répondu 6 à toutes les questions, la probabilité qu'elle ait répondu oui se calcule de la façon suivante :

$$P(\text{alliance stratégique}) = \left(1 + e^{(-20,110 - 0,411 \cdot 1 + 1,054 \cdot 6 + \dots - 0,365 \cdot 6)}\right)^{-1} = 0,2637.$$

Donc on estime qu'une telle entreprise n'aurait que 26,37 % de chance de vouloir une alliance stratégique. Il est à noter qu'il est possible de faire des prédictions comme on le fait avec un modèle linéaire ; elles se retrouvent dans la colonne PRE_1. (Mais attention : il faut indiquer correctement que l'entreprise est en Amérique du Nord en mettant un 0 dans la colonne region et non pas un 1...)

La figure 19.13 permet aussi de faire les tests suivants sur les paramètres β_j :

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

La résolution de ce test se base sur la statistique de Wald. Pour des tailles d'échantillons supérieures ou égales à 30 éléments, la statistique de Wald se distribue suivant une loi du Chi-deux. Lorsque la statistique de Wald ne possède qu'un seul degré de liberté (c'est le cas pour les variables continues), elle est simplement le carré du rapport de la statistique b_j sur son écart-type. Pour les variables discrètes, la statistique de Wald possède un degré de liberté de moins que le nombre de modalités de cette variable, d'où l'importance de laisser le logiciel effectuer les recodifications des variables discrètes via le bouton `categorical` (plus de détails seront donnés sur la recodification dans un autre exemple).

Par exemple, pour la variable `qualite`, on a $b_{\text{qualité}} = 1,054$ et son écart-type est de 0,311; on peut donc calculer la statistique de Wald :

$$\left(\frac{1,054}{0,311}\right)^2 = 11,49 \cong 11,471.$$

Et puisque la p -value est de 0,001 (et donc plus petite que $\alpha = 0,05$), on rejette H_0 et on conclut donc que le coefficient est significativement différent de 0.

Au seuil $\alpha = 0,05$, on voit que les variables jugées significatives sont `qualite`, `probres` et `forcevente`. Ces résultats sont ici très similaires à ce que nous révélait la sortie 19.8 sur chacune des variables prises individuellement. Il est à noter que ce n'est pas toujours le cas, surtout lorsqu'il y a de la multicolinéarité.

En fait, la statistique de Wald possède l'inconvénient suivant : lorsqu'un coefficient en valeur absolue devient grand, l'écart-type estimé a tendance à devenir très important, ayant pour conséquence de ne pas rejeter H_0 alors qu'elle devrait l'être. Ainsi, lorsque le coefficient d'une variable semble trop grand, il est préférable de ne pas se fier sur la statistique de Wald pour effectuer le test d'hypothèses. Il est préférable pour l'analyste de comparer le modèle complet au modèle réduit (sans ladite variable) en regardant le changement sur -2LL (Hauck & Donner, 1977).

Pour interpréter les paramètres de la droite de régression, il est utile de rappeler

l'équation suivante :

$$odds = \frac{P(Y = 1)}{P(Y = 0)} = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon} = e^{\beta_0} \cdot e^{\beta_1 X_1} \dots e^{\beta_k X_k} \cdot e^{\epsilon}.$$

Dans le cadre de l'exemple, on a

$$\begin{aligned} odds &= \frac{P(\text{alliance stratégique})}{P(\text{pas alliance stratégique})} \\ &= e^{-20,110} \cdot e^{-0,411x_{\text{region}}} \cdot e^{1,054x_{\text{qualité}}} \dots e^{-0,365x_{\text{nouveau}}}. \end{aligned}$$

Rappelons que la quantité e^{β_j} représente le facteur multiplicatif avec lequel le *odds* changera pour une augmentation unitaire de la variable indépendante X_j . On retrouve ces facteurs dans la colonne Exp(B) de la figure 19.13. Si β_j est positif, le facteur sera plus grand que 1. Si β_j est négatif, le facteur sera plus petit que 1. Si $\beta_j = 0$, le facteur sera égal à 1, et n'entraînera donc aucun changement dans le *odds*. Dans le cadre de cet exemple, on voit par exemple que lorsque la variable `qualité` augmente d'une unité, alors le *odds* est multiplié par 2,87, et donc la probabilité d'une alliance augmente. Inversement, lorsque la variable `nouveau` augmente d'une unité, le *odds* est multiplié par 0,694, diminuant ainsi le *odds* de 30,6 %.

Parfois il peut être intéressant de calculer le *odds* d'une observation en particulier. Reprenons l'exemple de l'entreprise située en Amérique du Nord et ayant octroyé un score de satisfaction de 6 à toutes les questions. Le *odds* qu'elle appartienne au groupe du oui se calcule de la manière suivante :

$$\begin{aligned} odds &= \frac{P(\text{alliance stratégique})}{P(\text{pas alliance stratégique})} \\ &= e^{-20,110} \cdot e^{-0,411 \cdot 1} \cdot e^{1,054 \cdot 6} \dots e^{-0,365 \cdot 6} \\ &= 0,3581. \end{aligned}$$

Si le *odds* est supérieur à 1, les mises seront en faveur du fait que l'entreprise veut considérer une alliance stratégique. Si le résultat est inférieur à 1, les mises favoriseront l'état inverse. Un *odds* égal à 1 indique que le présent modèle ne vaut pas mieux que la chance pour prédire le groupe d'appartenance de cet individu. Donc ici on voit que cette entreprise n'appartiendra probablement pas au groupe du oui, ce qu'on savait déjà puisqu'on avait calculé sa probabilité d'appartenir à ce groupe.

19.4.4 Validité du modèle

Lors de la construction d'un modèle utilisant une forme de régression, il est toujours important d'effectuer l'analyse des résidus afin de mesurer l'adéquation générale du modèle résultant. Il faut également regarder les VIF pour s'assurer qu'il n'y a pas un problème de multicolinéarité.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	-2.026	.357		-5.678	.000		
region	.073	.100	.072	.732	.466	.613	1.631
qualite	.130	.032	.364	4.029	.000	.717	1.395
probres	.175	.033	.422	5.289	.000	.922	1.085
forcevente	.135	.040	.289	3.341	.001	.782	1.279
nouveau	-.044	.026	-.132	-1.708	.091	.975	1.026

a. Dependent Variable: alliance

FIG. 19.14 – Les VIF

Tout d'abord, la figure 19.14 nous présente les VIF des variables indépendantes (obtenus avec les commandes pour une régression linéaire). On voit que tous les VIF sont bas, ce qui nous rassure quant à la validité de notre modèle.

	alliance	PRE_1	PGR_1	COO_1	LEV_1	RES_1	ZRE_1
1	oui	.92423	oui	.00340	.03983	.07577	.28633
2	non	.61168	oui	.28645	.15387	-.61168	-1.2551
3	oui	.92238	oui	.00445	.05021	.07762	.29009
4	non	.01343	non	.00020	.01483	-.01343	-.11666
5	non	.26699	non	.02022	.05259	-.26699	-.60353
6	non	.00125	non	.00000	.00312	-.00125	-.03544
7	non	.01947	non	.00058	.02839	-.01947	-.14093
8	non	.06321	non	.00188	.02712	-.06321	-.25976
9	oui	.63021	oui	.05139	.08052	.36979	.76601

FIG. 19.15 – Les variables créées

Les options cochées dans le bouton **Save...** ont fait apparaître de nouvelles variables dans la base de données ; on en a un aperçu dans la figure 19.15. Voici ce qu'elles repré-

sentent :

PRE_1 : c'est la probabilité correspondant à l'événement codé 1 (de la variable dépendante). Par exemple, l'entreprise de la première ligne a une probabilité de 92,42 % d'avoir répondu oui pour l'alliance stratégique.

PGR_1 : classification prédite par le modèle. Par exemple, puisque $92,42\% > 50\%$, le modèle prédit que l'entreprise de la ligne 1 appartient au groupe des oui, ce qui est effectivement le cas.

RES_1 : le résidu est la différence entre le code du groupe auquel l'observation appartient réellement et la probabilité calculée par le modèle. Par exemple, pour l'entreprise de la première ligne, le résidu est $1 - 0,92423 = 0,07577$. Si l'entreprise avait plutôt appartenu au groupe du non, le résidu aurait été $0 - 0,92423 = -0,92423$. Plus le résidu est grand, plus la prédiction est loin de l'état réel de l'observation. Si toutes les prédictions sont bonnes (dans le cas où la coupure est à 0,5), alors tous les résidus seront plus petit que 0,5 en valeur absolue. Dans tous les cas, la plus grande valeur absolue que peut prendre un résidu non-standardisé est 1.

ZRE_1 : le résidu standardisé de chacun des individus prend la forme suivante :

$$\text{ZRE}_{1i} = \frac{\text{RES}_{1i}}{\sqrt{\text{RES}_{1i} \cdot (1 - \text{RES}_{1i})}}.$$

Si la taille d'échantillon est grande, les résidus standardisés devraient avoir une moyenne de 0 et une variance de 1. Rappelons à la base que la distribution de l'erreur est par définition binomiale, d'où la forme de l'écart-type du résidu.

LEV_1 : La mesure *leverage* est utilisée pour détecter les valeurs qui ont un impact important sur les coefficients de la régression logistique. Si une donnée i n'a pas d'influence sur le modèle obtenu, la mesure LEV_{1i} sera petite, idéalement nulle. En régression logistique, la statistique *leverage* varie entre 0 et 1. La moyenne de ces statistiques est k/n où k est le nombre de paramètres estimés dans le modèle et où n est la taille de l'échantillon.

C00_1 : la distance de Cook est une mesure de l'influence de chacune des données sur le modèle. Si une donnée i n'a pas d'influence sur le modèle obtenu, la mesure C00_{1i}

sera petite, idéalement nulle. Cette mesure illustre à quel point le fait d'enlever la donnée i change la valeur résiduelle de cette dernière illustrant ainsi un changement potentiel dans les coefficients de la régression. Voici la formule de cette mesure :

$$COO_1_i = \frac{ZRES_1_i \cdot LEV_1_i}{(1 - LEV_1_i)^2}.$$

Voici quelques graphiques utiles ainsi que leurs interprétations. Avant de faire les graphiques, il est important que les données du fichier soient classées en ordre croissant de la variable d'identification.

Tout d'abord, pour détecter les données qui ont un bras de levier important et donc une large influence sur le modèle, on peut regarder les graphes séquentiels avec les variables LEV_1 et COO_1 en axe des Y (figures 19.16 et 19.17).

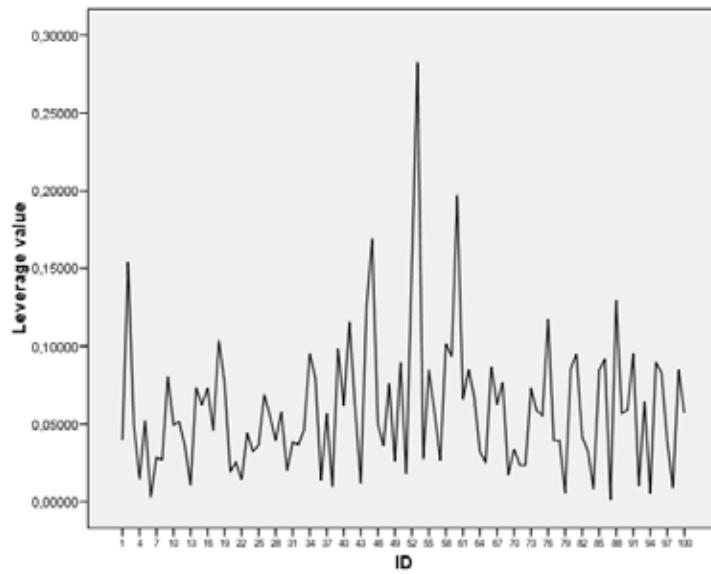


FIG. 19.16 – Les *leverage* en fonction du numéro d'identification

On voit que certaines observations ont un grand *leverage* (le plus grand a une valeur près de 0,5) ou une grande mesure de *Cook's*; il est conseillé d'étudier ces cas afin de s'assurer d'abord qu'il n'y a pas d'erreur d'entrées de données, et de voir si ces observations, selon leurs profils, devraient vraiment faire partie de cette étude.

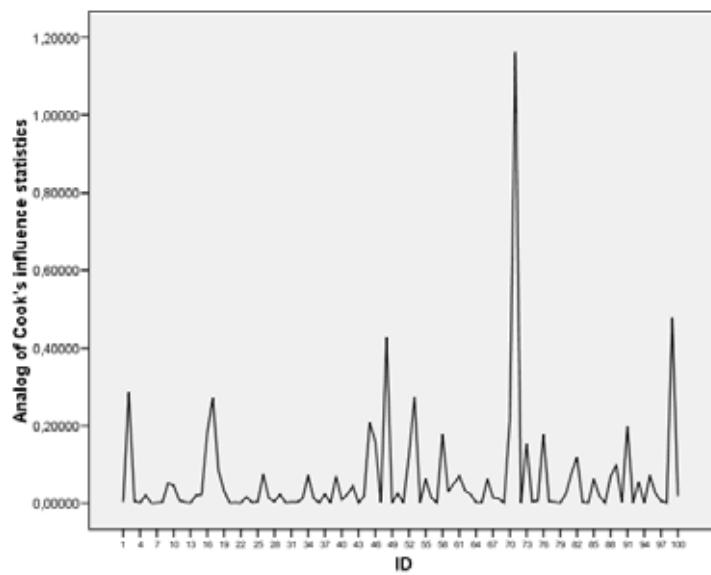
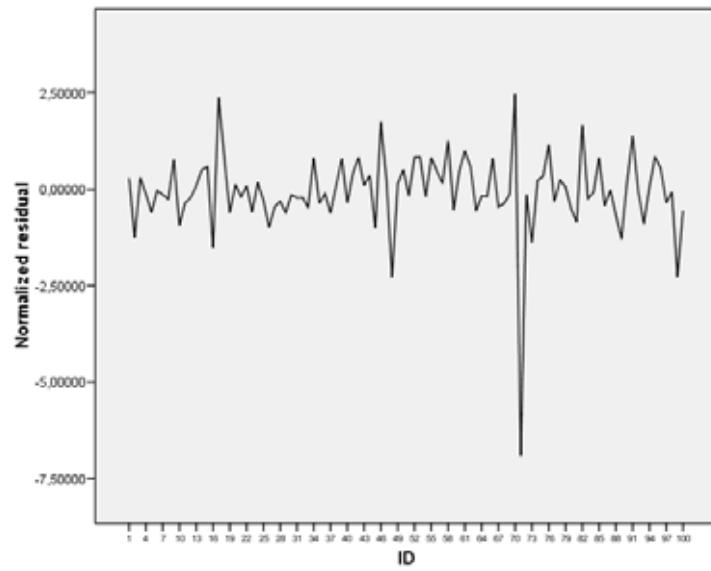
FIG. 19.17 – Les mesures de *Cook's* en fonction du numéro d'identification

FIG. 19.18 – Les résidus standardisés en fonction du numéro d'identification

De façon semblable, on peut examiner les résidus avec un graphe séquentiel (figure 19.18).

Il est flagrant ici qu'un des résidus est très élevé relativement aux autres. C'est le résidu de l'entreprise 71 ; le modèle a évalué à 97,96 % la probabilité qu'elle veuille considérer une alliance stratégique, alors qu'elle a répondu non.

Si suite à l'examen des graphes vu précédemment il est décidé que certaines observations devraient être corrigées ou enlevées de l'étude, il faut alors refaire le modèle de régression logistique.

19.4.5 Les courbes ROC

La régression logistique classe chaque individu dans un groupe en fonction de la probabilité que l'événement étudié survienne. La valeur de coupe permettant le classement des individus est fixée à 0,5 par défaut. Cette valeur joue un rôle central dans la pré-diction de classement. Mais cette valeur de coupe est-elle la meilleure en tout temps ? Dans certains cas, 0,5 n'est pas la valeur optimale et l'analyste utilise les courbes ROC (*Receiver Operating Characteristic*) pour découvrir la valeur de coupe de type optimal.

		Predicted		Percentage Correct	
		Voudrait bien considérer une alliance stratégique			
		non	oui		
Step 1	Voudrait bien considérer une alliance stratégique	non	47	85.5	
		oui	7	84.4	
	Overall Percentage			85.0	

a. The cut value is .500

FIG. 19.19 – La table de classification du modèle complet

La figure 19.19 est la table de classification que nous avons observée précédemment.

De cette table on peut tirer le tableau suivant, qui contient les probabilités a posteriori du classement avec la valeur de coupe 0.5 :

Résultat	Probabilité
Oui bien classé (<i>True positive</i>)	0,844
Non bien classé (<i>True negative</i>)	0,855
Oui mal classé (<i>False positive</i>)	0,156
Non mal classé (<i>False negative</i>)	0,145

Dans les études cliniques, la probabilité liée au résultat *True positive* est appelée la **sensibilité** (*sensitivity*) de la règle de classement. La probabilité liée au résultat *True negative* est appelée la **spécificité** (*specificity* dans le sens de précision) de la règle utilisée. Plus ces probabilités sont grandes, plus le modèle est performant. Ces valeurs sont importantes puisqu'elles résument à quel point le modèle est efficace.

Il peut être intéressant pour l'analyste d'apprécier comment ces probabilités varient en fonction des changements dans la règle de coupe. En d'autres termes, quelle serait la performance de classement de la régression logistique si la règle de coupure était autre que 0,5 ? Il serait manuellement possible pour le praticien de voir les changements en changeant le lieu de la coupure, en reconstruisant la table de classification et en recalculant les probabilités de sensibilité et de spécification. À cet égard, SPSS propose au praticien une procédure automatisée et les résultats sont résumés dans les courbes ROC. Les commandes pour obtenir la courbe ROC de l'exemple sont les suivantes :

Menu SPSS :	→ Analyse
	→ ROC Curve...
Dans la fenêtre Test Variable :	→ PRE_1
Dans la fenêtre State Variable :	→ alliance (la variable dépendante)
	Value of State Variable : 1
Dans la fenêtre Display :	✓ ROC Curve
	✓ With diagonal reference line
	✓ Coordinates points of the ROC Curve

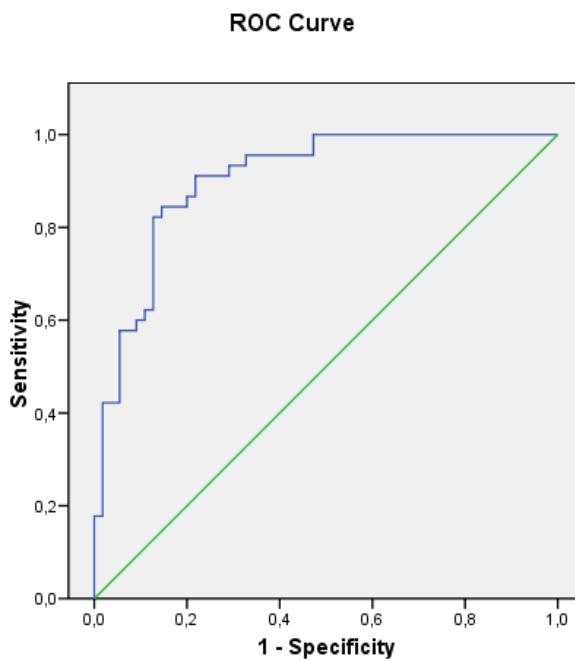


FIG. 19.20 – La courbe ROC

La figure 19.20 présente la courbe ROC de l'exemple. L'axe vertical est la sensibilité (*true positive*) et l'axe horizontal présente le complément de la spécificité : 1 - la spécificité (*false negative*). La diagonale représente une ligne de référence montrant ce qui se passerait si les deux mesures augmentaient au même rythme (comme si le modèle tirait à pile ou face).

L'analyste sera intéressé aux endroits où la courbe ROC croît rapidement démontrant un accroissement important de la sensibilité pour de petits changements au niveau des mauvaises prédictions.

La figure 19.21 nous permet de trouver la règle de classification optimale. La première colonne de cette sortie contient différents points de coupure en ordre croissant de 0 à 1. Afin de se familiariser avec la table, l'analyste peut regarder la performance du modèle avec la coupure par défaut fixée à 0,5. La valeur 0,5025458 est assez près de 0,5 et donne les mêmes résultats vus précédemment en termes de sensibilité et de spécificité : la sensibilité est de 0,844 et la spécificité est de 0,855 (1,0-0,145).

Par exemple, si la coupure est plutôt fixée à 0,4832876, la sensibilité serait encore de 0,844 tandis que la spécificité serait de 0,836 (1,0-0,164). Ici il n'est pas possible d'améliorer la sensibilité sans faire baisser la spécificité. Et inversement, on pourrait vouloir améliorer la spécificité, mais ceci se ferait au détriment de la sensibilité.

.3305939	.911	.218
.3681762	.889	.218
.4063936	.867	.218
.4264321	.867	.200
.4368733	.844	.200
.4553024	.844	.182
.4832876	.844	.164
.5025458	.844	.145
.5057374	.822	.145
.5425503	.822	.127
.5803876	.800	.127
.5855738	.778	.127
.5910654	.756	.127
.5940430	.733	.127
.5962783	.711	.127
.5973348	.689	.127
.5979341	.667	.127
.6027271	.644	.127
.6094069	.622	.127
.6121861	.622	.109
.6171821	.600	.109
.6259430	.600	.091
.6449257	.578	.091
.6789078	.578	.073
.7179690	.578	.055

FIG. 19.21 – Extrait de la table des coordonnées

19.5 Un autre exemple

La section précédente nous a permis de voir toutes les étapes d'une analyse en régression logistique. On présente maintenant un autre exemple qui a la particularité de présenter un modèle dont l'une des variables indépendantes est une variable discrète avec plus de deux modalités.

Exemple 19.5.1 La base de données `supermarches.sav` contient les données d'une étude qui a portée sur trois bannières d'épicerie : ProfitGros, JaunePartout et SuperSelect. Les variables à l'étude sont les suivantes :

<code>arrondissement</code>	Dans quel arrondissement habitez-vous ?
<code>auto</code>	Avez-vous une automobile ?
<code>etudiant</code>	Êtes-vous étudiant ?
<code>emplacement</code>	Quelle est votre satisfaction par rapport à l'emplacement du supermarché ?
<code>prix</code>	Quelle est votre satisfaction par rapport aux prix ?
<code>variete</code>	Quelle est votre satisfaction par rapport à la variété des produits ?
<code>qualite</code>	Quelle est votre satisfaction par rapport à la qualité des produits ?
<code>service</code>	Quelle est votre satisfaction par rapport au service ?
<code>rapidite</code>	Quelle est votre satisfaction par rapport à la rapidité aux caisses ?
<code>presentation</code>	Quelle est votre satisfaction par rapport à la présentation des produits ?
<code>proprete</code>	Quelle est votre satisfaction par rapport à la propreté du supermarché ?
<code>régulier</code>	Allez-vous ou seriez-vous intéressé à aller régulièrement à ce supermarché ?

Les variables sur la satisfaction varient entre 0 et 10, 0 signifiant très insatisfait et 10 signifiant très satisfait.

On décide de faire une régression logistique avec toutes ces variables pour voir ce qui distingue un client qui veut aller régulièrement au même supermarché d'un client qui ne le veut pas. Donc ici la variable dépendante est **régulier**. Fixons les seuils à $\alpha = 0,05$.

Case Processing Summary		
Unweighted Cases ^a	N	Percent
Selected Cases	Included in Analysis	211 100,0
	Missing Cases	0 ,0
	Total	211 100,0
Unselected Cases		0 ,0
Total		211 100,0

^a If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding	
Original Value	Internal Value
non	0
oui	1

Categorical Variables Codings				
	Frequency	Parameter coding		
		(1)	(2)	(3)
arrondissement	1	64	1,000	,000
	2	49	,000	1,000
	3	52	,000	,000
	4	46	,000	,000
etudiant	non	126	1,000	
	oui	85	,000	
auto	non	56	1,000	
	oui	155	,000	

FIG. 19.22 – Premières sorties

Dans la figure 19.22 on voit tout d'abord qu'il y a 211 individus, et seulement un échantillon d'analyse. La variable dépendante **régulier** est codée de la façon suivante : **non** = 0, **oui** = 1.

La dernière sortie de la figure 19.22 nous montre comment sont codées les variables indépendantes discrètes. Ainsi on voit que c'est l'arrondissement 4 qui sert de référence pour la variable **arrondissement** qui est représentée par trois variables binaires. On peut également voir les fréquences associées à chaque modalité. Par exemple, il y a 64 individus provenant de l'arrondissement 1, et il y a 85 étudiants.

Il est à noter qu'il y a plus d'une façon de coder les variables discrètes ayant plus de deux modalités (la codification des variables discrètes se nomme le **contraste**). Ici on a pris la méthode par défaut, qui est **Indicator** (dans le bouton **Categorical...**). Il est

important de procéder de cette façon (codification faite par SPSS et non pas manuellement) pour que les statistiques soient correctement calculées.

Iteration History ^{a,b,c}		
Iteration	-2 Log likelihood	Coefficients
		Constant
Step 0	288,510	,275
2	288,510	,277
3	288,510	,277

a. Constant is included in the model.
 b. Initial -2 Log Likelihood: 288,510
 c. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.

FIG. 19.23 – Le -2LL du modèle constant

La figure 19.23 nous montre que le -2LL « à battre » est de 288,510. Donc on s'attend à ce que le modèle avec les variables indépendantes ait un -2LL inférieur à 288,510.

Observed		Predicted		Percentage Correct	
		Allez-vous ou seriez-vous intéressé à aller régulièrement à ce supermarché?			
		non	oui		
Step 0	Allez-vous ou seriez-vous intéressé à aller régulièrement à ce supermarché?	non	0	91 ,0	
		oui	0	120 100,0	
	Overall Percentage			56,9	

a. Constant is included in the model.
 b. The cut value is ,500

FIG. 19.24 – La table de classification du modèle constant

La figure 19.24 nous montre que le modèle constant a, comme prévu, classé tous les individus dans le groupe le plus volumineux qui est celui du « oui » avec 120 individus. Le pourcentage de réussite est donc $120/211 = 56,9\%$.

La figure 19.25 nous montre qu'individuellement, ce sont les variables **emplacement** et **prix** qui semblent avoir le plus d'impact sur l'idée d'aller régulièrement ou pas au même

Variables not in the Equation				
Step	Variables	Score	df	Sig.
0	arrondissement	13,936	3	,003
	arrondissement(1)	12,308	1	,000
	arrondissement(2)	2,567	1	,109
	arrondissement(3)	4,496	1	,034
	auto(1)	2,330	1	,127
	etudiant(1)	,897	1	,344
	emplacement	34,744	1	,000
	prix	24,089	1	,000
	variete	,027	1	,871
	qualite	1,752	1	,186
	service	1,128	1	,288
	rapidite	5,821	1	,016
	presentation	,567	1	,451
	proprete	,030	1	,862
Overall Statistics		82,670	13	,000

FIG. 19.25 – Les variables prises individuellement

supermarché. En effet, ces deux variables ont les plus hauts scores (34,744 et 24,089) et leurs *p*-values sont nulles.

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	185,589 ^a	,386	,518

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

FIG. 19.26 – Le -2LL du modèle et le pseudo r^2

On voit que le -2LL du modèle est de 185,589 (figure 19.26), ce qui est meilleur que celui du modèle constant (qui est de 288,510). Et le modèle explique environ 51,8 % de la variation de la variable dépendante.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	102,920	13	,000
	Block	102,920	13	,000
	Model	102,920	13	,000

FIG. 19.27 – Test omnibus : les modèle est-il significatif?

La figure 19.27 nous permet de résoudre le test suivant :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \text{Au moins un des } \beta_j \neq 0 \ (1 \leq j \leq k)$$

Puisque la p -value = 0 < 0,05 = α , on rejette H_0 , et donc au risque de se tromper une fois sur 20 on conclut que le modèle est significatif. Donc au moins une des variables indépendantes a de l'impact sur la dépendante.

Variables in the Equation

Step		B	S.E.	Wald	df	Sig.	Exp(B)
1	arrondissement			7,833	3	,050	
	arrondissement(1)	1,090	,898	1,474	1	,225	2,975
	arrondissement(2)	,207	,781	,070	1	,791	1,230
	arrondissement(3)	-,960	,679	2,000	1	,157	,383
	auto(1)	-,102	,565	,033	1	,856	,903
	etudiant(1)	-,262	,573	,210	1	,647	,769
	emplacement	,831	,182	20,983	1	,000	2,297
	prix	,755	,160	22,234	1	,000	2,127
	variete	-,141	,191	,548	1	,459	,868
	qualite	,395	,154	6,526	1	,011	1,484
	service	-,055	,183	,090	1	,764	,947
	rapidite	,433	,116	14,030	1	,000	1,542
	presentation	-,081	,138	,340	1	,560	,923
	proprete	-,098	,198	,243	1	,622	,907
	Constant	-13,333	2,636	25,577	1	,000	,000

a. Variable(s) entered on step 1: arrondissement, auto, etudiant, emplacement, prix, variete, qualite, service, rapidite, presentation, proprete.

FIG. 19.28 – Les coefficients

La figure 19.28 nous permet de tester lesquelles des variables sont significatives. Le

test est

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Au seuil $\alpha = 0,05$, on voit que seules les variables `emplacement`, `prix`, `qualite` et `rapidite` sont significatives puisque ce sont celles qui ont une *p*-value plus petite que 0,05. La variable discrète `arrondissement` a une *p*-value de 0,05 ; on peut la considérer comme étant significative, d'autant plus que la différence entre le -2LL du modèle avec et celui sans cette variable est significative (figure 19.29).

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8,555	3	,036
	Block	8,555	3	,036
	Model	102,920	13	,000

FIG. 19.29 – La différence entre les -2LL des modèles avec et sans `arrondissement`

Revenons à la figure 19.28 ; elle nous permet aussi d'écrire l'équation du modèle :

$$\begin{aligned} odds &= \frac{P(\text{aller régulièrement})}{P(\text{pas aller régulièrement})} \\ &= e^{-13,333} \cdot e^{1,090x_{\text{arr}(1)}} \cdot e^{0,207x_{\text{arr}(2)}} \cdots e^{-0,081x_{\text{presentation}}} \cdot e^{-0,098x_{\text{proprete}}}. \end{aligned}$$

Rappelons qu'un coefficient positif augmente les chances que l'individu veuille aller régulièrement au supermarché lorsque la valeur de la variable augmente, et qu'au contraire un coefficient négatif diminue les chances que l'individu veuille aller régulièrement au supermarché lorsque la valeur de la variable augmente. Dans le cas présent, toutes les variables significatives ont des coefficients positifs. Par exemple, plus un individu sera satisfait des prix, plus il sera porté à vouloir aller régulièrement à ce supermarché.

Les coefficients associés à la variable `arrondissement` s'interprètent quelque peu différemment. Il faut garder à l'idée que l'arrondissement 4 constitue la référence. Si un individu provient de l'arrondissement 4, alors les trois variables `arrondissement(1)`, `arrondissement(2)` et `arrondissement(3)` seront nulles, et le *odds* sera simplement

multiplié par 1. Pour comprendre l'effet de chacun des arrondissement, il faut les comparer à cet effet neutre. Par exemple, un individu provenant de l'arrondissement 1 verra son *odds* multiplié par 2,975, ce qui augmente les chances qu'il veuille aller régulièrement à ce supermarché comparativement à quelqu'un de l'arrondissement 4. Au contraire, quelqu'un qui provient de l'arrondissement 3 verra son *odds* multiplié par 0,383, ce qui diminuera donc les chances qu'il veuille aller régulièrement à ce supermarché.

		Predicted		Percentage Correct	
		Allez-vous ou seriez-vous intéressé à aller régulièrement à ce supermarché?			
		non	oui		
Step 1	Allez-vous ou seriez-vous intéressé à aller régulièrement à ce supermarché?	71	20	78,0	
	oui	15	105	87,5	
	Overall Percentage			83,4	

a. The cut value is .500

FIG. 19.30 – La table de classification

La figure 19.30 nous montre que 83,4 % des individus ont été bien classés par ce modèle. Plus précisément, 78 % de ceux qui ont répondu non (aller régulièrement au supermarché) ont été classés correctement, et 87,5 % de ceux qui ont répondu oui ont été bien classés. Cette performance est tout à fait acceptable, elle dépasse de loin la performance globale de 56,9 % du modèle constant.

La figure 19.31 illustre ce classement. On voit que la coupure est assez nette en ce sens que la majorité des o se retrouvent à droite de la coupure de 0,5, tandis que la majorité des n sont à gauche.

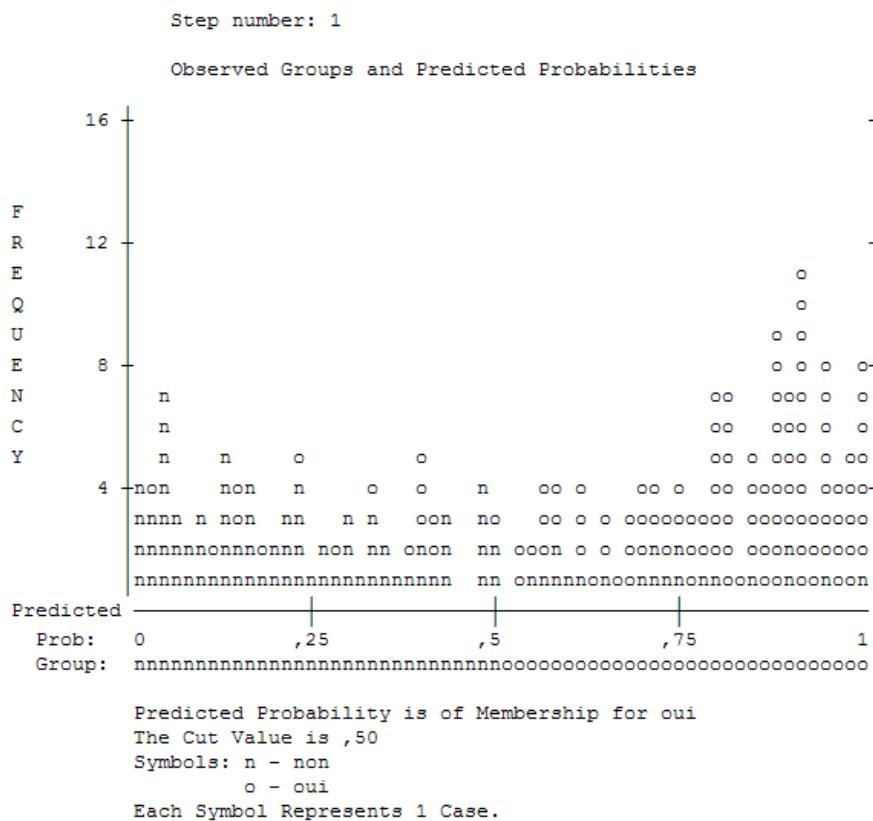


FIG. 19.31 – Histogramme des probabilités prédictes

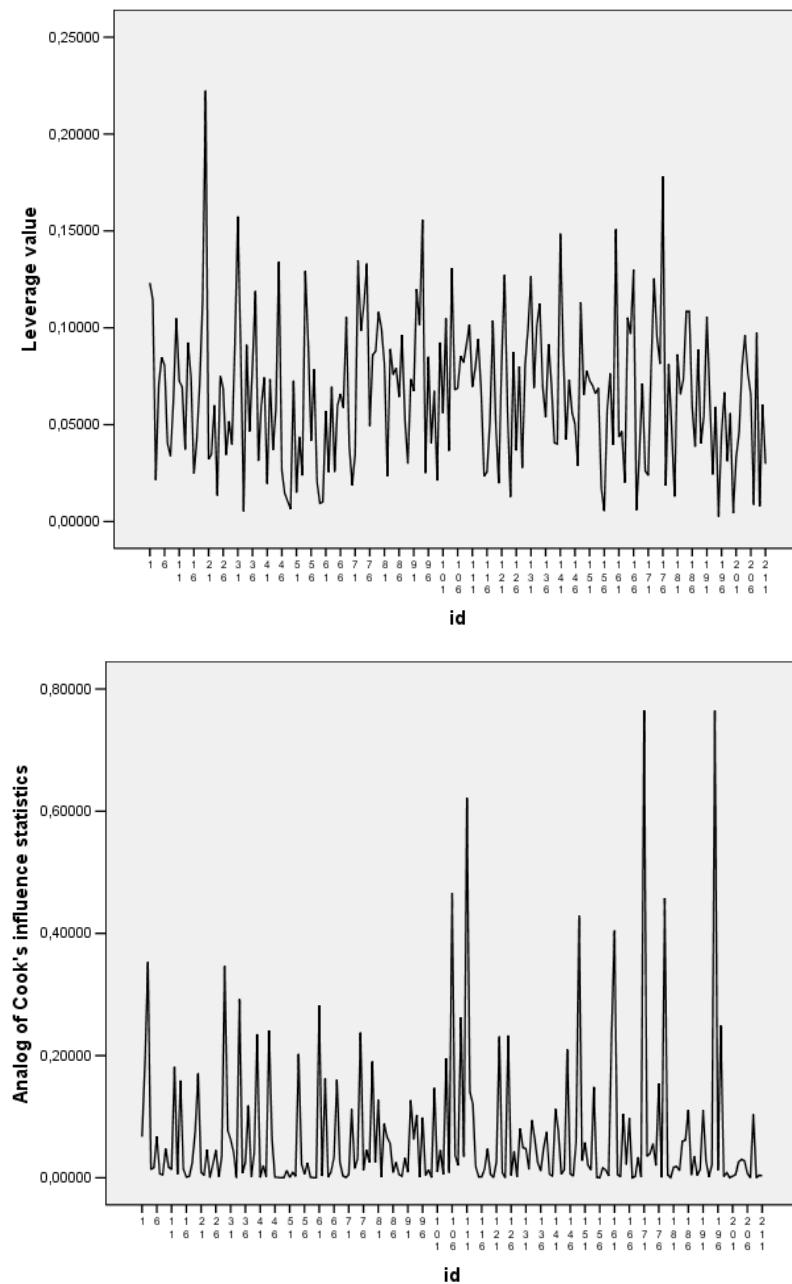
La figure 19.32 nous montre qu'il n'y a aucun problème de multicolinéarité puisque le plus grand VIF est de 5,188. Donc l'interprétation tirée des coefficients de l'équation est valable.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	-1,556	,358		-4,350	,000		
auto	,015	,083	,013	,176	,860	,564	1,774
etudiant	,038	,091	,037	,415	,678	,382	2,615
emplacement	,122	,022	,328	5,407	,000	,840	1,191
prix	,119	,022	,421	5,508	,000	,528	1,894
variete	-,020	,029	-,067	-,678	,498	,319	3,140
qualite	,061	,022	,217	2,801	,006	,515	1,940
service	-,008	,028	-,021	-,273	,785	,512	1,954
rapidite	,061	,016	,256	3,747	,000	,662	1,510
presentation	-,014	,020	-,070	-,703	,483	,315	3,171
proprete	-,011	,029	-,029	-,391	,696	,567	1,762
arr1	,129	,136	,120	,949	,344	,193	5,188
arr2	-,009	,119	-,008	-,075	,940	,298	3,361
arr3	-,163	,098	-,142	-1,660	,098	,424	2,358

a. Dependent Variable: régulier

FIG. 19.32 – Les VIF

L'examen des figures 19.33 et 19.34 nous permet de voir s'il y a des individus qui influencent grandement le modèle ou qui sont aberrants (les aberrants influencent souvent le modèle). Par exemple, l'individu 195 a le plus grand résidu standardisé en valeur absolue (-17,05), et une des plus grandes mesures de Cook's (la 2e plus grande en fait). Par contre son *leverage* est petit. Le modèle a estimé à 0,99657 la probabilité que cet individu veuille aller régulièrement à ce supermarché, alors qu'il a répondu non à cette question. Par contre l'examen des données relatives à cet individu ne laisse pas croire à une erreur d'entrée de données (il faudrait quand même vérifier sur le questionnaire), et rien n'indique qu'il ne devrait pas faire partie de cette étude. Il n'y a donc pas vraiment d'action à prendre à propos de cet individu, c'est seulement quelqu'un « hors-normes » par rapport à notre modèle. Il y a sûrement une raison qui fait que cet individu n'est pas intéressé à aller dans cette épicerie régulièrement, mais elle n'a pas été détectée par les questions à l'étude. Il est d'ailleurs utile d'avoir une section « Commentaires » à la fin du questionnaire, ceux-ci pouvant aider à comprendre ce type de problème.

FIG. 19.33 – Les *leverage* et les mesures de Cook's

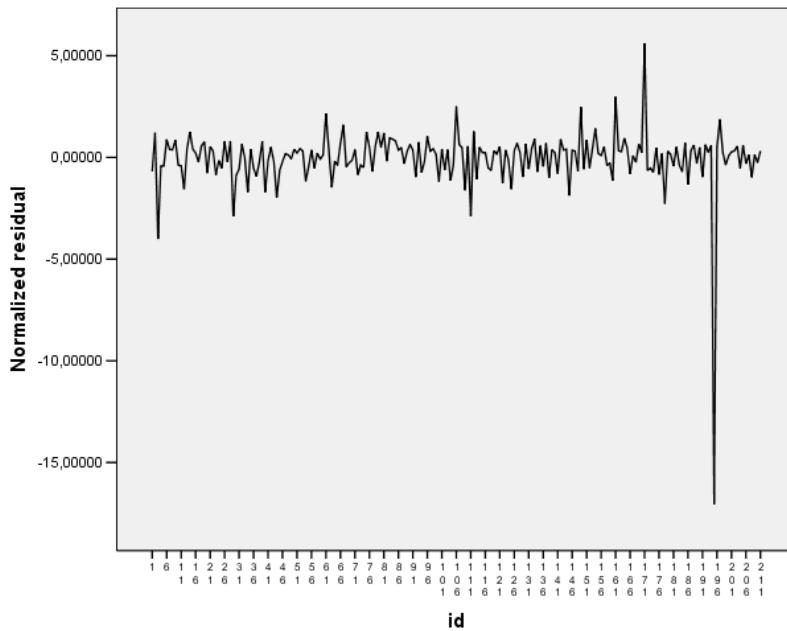


FIG. 19.34 – Les résidus

L'étude de ce cas est presque finalisée ; on a un bon modèle entre les mains, et on peut en tirer une interprétation valable. Mais la coupure à 0,5 pour le classement est-elle la meilleure ? On peut répondre à cette question en examinant les figures 19.35 et 19.36.

On voit qu'à la coupure 0,5117 on obtient les performances du modèle présent. Si on diminue un peu (à 0,4940), on n'améliore pas la sensibilité (elle demeure à 0,875), et on diminue la spécificité (elle passe de 0,78 à $1 - 0,231 = 0,769$), donc on diminue la performance. Si on augmente la coupure à 0,5273, la sensibilité diminue à 0,867 et la spécificité ne bouge pas. Donc il n'y a pas moyen d'améliorer de façon globale ce modèle en changeant la valeur de la coupure.

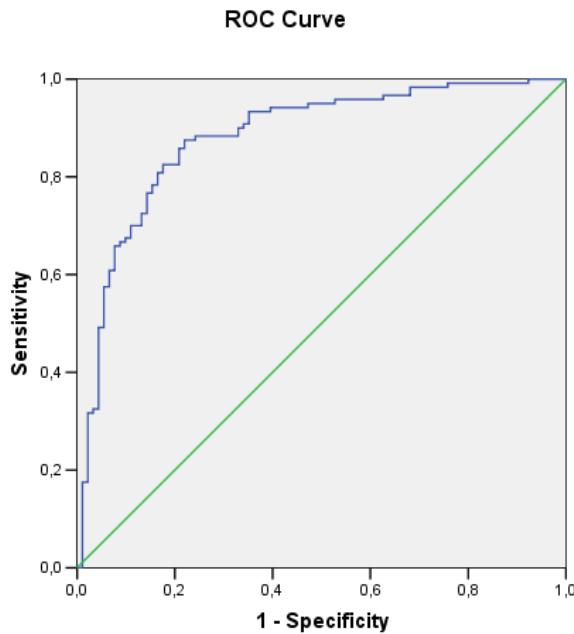


FIG. 19.35 – La courbe ROC

,3940997	,900	,341
,4004982	,900	,330
,4076297	,892	,330
,4128424	,883	,330
,4173522	,883	,319
,4193712	,883	,308
,4260977	,883	,297
,4496224	,883	,286
,4731655	,883	,275
,4808068	,883	,264
,4826243	,883	,253
,4841911	,883	,242
,4873879	,875	,242
,4940169	,875	,231
,5117148	,875	,220
,5272778	,867	,220
,5338985	,858	,220
,5413845	,858	,209
,5494771	,850	,209
,5575110	,842	,209
,5624215	,833	,209
,5647284	,825	,209
,5666785	,825	,198
,5725681	,825	,187

FIG. 19.36 – Extrait de la table des coordonnées

Chapitre 20

Régression logistique multinomiale

Il est possible de généraliser les concepts du dernier chapitre afin d'utiliser la régression logistique avec une variable dépendante qui a plus de deux modalités. On obtient alors une régression logistique multinomiale. Encore une fois, l'avantage de celle-ci par rapport à l'analyse discriminante est qu'elle n'exige pas vraiment de pré-requis.

Et tout comme l'analyse discriminante, le fait que la dépendante ait plus de deux groupes entraîne que nous devrons travailler avec plus d'une équation. Celles-ci seront de la même forme que l'équation de la régression logistique ; cette représentation étant plus complexe que les fonctions discriminantes, certains analystes préfèrent travailler avec l'analyse discriminante.

20.1 Les principes de base

On a vu, dans le chapitre précédent, que le modèle de la régression logistique binaire prend la forme suivante :

$$\ln \left(\frac{P(\text{événement})}{P(\text{pas événement})} \right) = \ln \left(\frac{P(Y=1)}{P(Y=0)} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

où les X_i sont les variables indépendantes et les β_i sont les paramètres de la régression logistique estimés suivant le principe du maximum de vraisemblance.

Rappelons aussi que la quantité

$$\ln \left(\frac{P(Y=1)}{P(Y=0)} \right)$$

porte le nom de Logit, tandis que la quantité

$$\frac{P(Y=1)}{P(Y=0)} = \frac{P(\text{événement})}{P(\text{pas événement})}$$

s'appelle un *odds*.

Lorsque la variable dépendante n'a que deux modalités, une seule équation suffit. En effet, le seul autre Logit qui existe dans ce cas est

$$\ln \left(\frac{P(\text{pas événement})}{P(\text{événement})} \right)$$

et celui-ci peut être calculé à partir du modèle existant.

Cependant, lorsque la variable dépendante possède j modalités avec $j > 2$, il faudra $j - 1$ équations afin de calculer tous les Logit. Il faudra alors choisir une des modalités de la variable dépendante comme étant la catégorie de référence. Par défaut, SPSS choisit la dernière modalité (donc la modalité j), et alors les équations s'écrivent ainsi :

$$\ln \left(\frac{P(\text{modalité}_1)}{P(\text{modalité}_j)} \right) = \beta_{0,1} + \beta_{1,1} X_1 + \cdots + \beta_{k,1} X_k + \epsilon_1$$

$$\ln \left(\frac{P(\text{modalité}_2)}{P(\text{modalité}_j)} \right) = \beta_{0,2} + \beta_{1,2} X_1 + \cdots + \beta_{k,2} X_k + \epsilon_2$$

...

$$\ln \left(\frac{P(\text{modalité}_{j-1})}{P(\text{modalité}_j)} \right) = \beta_{0,j-1} + \beta_{1,j-1} X_1 + \cdots + \beta_{k,j-1} X_k + \epsilon_{j-1}$$

De cette façon toutes les modalités sont comparées à la modalité de référence j . Il y aura $j - 1$ ensembles de paramètres à estimer, toujours selon le maximum de vraisemblance. La régression de référence est nulle ; en effet,

$$\ln \left(\frac{P(\text{modalité}_j)}{P(\text{modalité}_j)} \right) = \ln(1) = 0.$$

Pour chaque unité de l'échantillon, il sera possible de calculer la probabilité que cette unité appartienne à un des groupes formés par la dépendante, et ce pour tous les groupes. Donc pour chaque unité il y aura j probabilités de calculées, une pour chaque groupe. Cette unité sera classée dans le groupe pour lequel elle a obtenu la plus grande probabilité.

20.2 L'exemple du chapitre

On reprend ici le contexte de l'exemple 19.5.1 ; attention par contre, la base de données a été légèrement modifiée pour cet exemple, elle se nomme `supermarchesmod.sav`. On reprend les variables indépendantes de l'exemple 19.5.1, mais cette fois-ci la variable dépendante est le nom de l'épicerie (ProfitGros, JaunePartout et SuperSelect). L'analyse qui suit nous permettra entre autres de voir si la satisfaction varie d'un supermarché à l'autre, et de quelle façon. On tient aussi compte de l'arrondissement de l'individu, s'il est étudiant ou non et s'il a une auto. Fixons les seuils à $\alpha = 0,05$.

Voici les commandes pour obtenir les sorties de cet exemple :

Menu SPSS :	→ Analyse
	→ Regression
	→ Multinomial Logistic...
Dans la fenêtre Dependent :	→ nom (la variable dépendante)
Dans la fenêtre Factor(s) :	→ arrondissement, auto, etudiant (les indépendantes discrètes)
Dans la fenêtre Covariate(s) :	emplACEMENT, prix, variete, qualite, service, rapidite, presentation, proprete (les indépendantes continues)
Dans le bouton Statistics... :	✓ Case processing summary ✓ Pseudo R-square ✓ Step summary ✓ Model fitting information ✓ Classification table ✓ Estimates
Dans le bouton Save... :	✓ Estimated response probabilities ✓ Predicted category

La figure 20.1 nous montre les fréquences associées aux modalités des variables discrètes du modèle, dont la variable dépendante. Toutes les techniques d'analyses de dépendance où la variable dépendante est discrète sont sujettes à mieux classer les groupes les plus imposants au détriment des petits. Plus la répartition de la variable dépendante est uniforme, moins cet effet est nocif. En regardant les % des cas bien classés, qui est une mesure de performance du modèle, il est facile de détecter l'impact nocif de la taille des groupes sur le classement. Si un des groupes est vraiment imposant, il peut être nécessaire de rééchantillonner les groupes de manière à mieux balancer les tailles des groupes, puis de recommencer l'analyse.

Ici la distribution des fréquences de la variable dépendante est relativement uniforme

Case Processing Summary			
		N	Marginal Percentage
nom	Profitgros	74	35,1%
	JaunePartout	78	37,0%
	SuperSelect	59	28,0%
arrondissement	1	64	30,3%
	2	49	23,2%
	3	52	24,6%
	4	46	21,8%
auto	non	56	26,5%
	oui	155	73,5%
etudiant	non	126	59,7%
	oui	85	40,3%
Valid		211	100,0%
Missing		0	
Total		211	
Subpopulation		211 ^a	

a. The dependent variable has only one value observed in 211 (100,0%) subpopulations.

FIG. 20.1 – La distribution des fréquences des variables discrètes (74, 78 et 59).

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
		-2 Log Likelihood	Chi-Square	df
Intercept Only	460,687			
Final	93,315	367,372	26	,000

FIG. 20.2 – Les -2LL (modèle constant et modèle complet)

Tout comme en régression logistique binaire, le -2LL est utilisé pour mesurer l'ajustement du modèle aux données. La figure 20.2 présente deux -2LL : celui du modèle constant (c'est-à-dire sans aucune variable indépendante) et celui du modèle complet. Le -2LL du modèle constant est la mesure « à battre ». Ici il a une valeur de 460,687. Le -2LL du modèle complet étant de 93,315, il semble qu'il est de loin meilleur que le modèle

constant.

Cette même sortie contient une p -value qui permet de tester si les paramètres des équations (ici on a trois groupes, donc deux équations) sont tous nuls. La valeur du χ^2 est simplement la différence entre le -2LL du modèle constant et celui du modèle complet ; ici il a une valeur de 367,372. Plus cette valeur est grande, plus on sera porté à rejeter l'hypothèse selon laquelle tous les paramètres sont nuls.

Ici, puisque la p -value est nulle et donc plus petite que $\alpha = 0,05$, on rejette la nullité des paramètres. Donc au moins une des équations contient au moins une variable qui a un impact significatif pour le classement des épiceries.

Pseudo R-Square	
Cox and Snell	,825
Nagelkerke	,929
McFadden	,797

FIG. 20.3 – Les pseudo r^2

La figure 20.3 présente des mesures qui donnent une idée de la performance globale du modèle. Les mesures de Cox and Snell et de Nagelkerke ont été présentées au chapitre précédent. Ici, d'après Nagelkerke, 92,9 % de la variation de la variable dépendante est expliquée par le modèle, ce qui est excellent.

La mesure de McFadden représente la proportion de la vraisemblance expliquée par les variables du modèle complet.

La figure 20.4 présente les coefficients des équations. On a

$$\ln \left(\frac{P(\text{Profitgros})}{P(\text{SuperSelect})} \right) = 40,543 - 0,657X_{\text{emplACEMENT}} + \dots - 1,920X_{\text{etudiant}}.$$

$$\ln \left(\frac{P(\text{JaunePartout})}{P(\text{SuperSelect})} \right) = 55,615 - 1,385X_{\text{emplACEMENT}} + \dots - 1,808X_{\text{etudiant}}.$$

		Parameter Estimates							
		B	Std. Error	Wald	df	Sig.	Exp(B)	95 % Confidence Interval for Exp(B)	
nom ^a								Lower Bound	Upper Bound
Profitgros	Intercept	40,543	14,833	7,471	1	,006			
	emplacement	-,657	,612	1,151	1	,283	,518	,156	1,722
	prix	2,210	,757	8,535	1	,003	9,120	2,070	40,183
	variete	-1,241	,620	4,005	1	,045	,289	,086	,975
	qualite	-2,540	,792	10,290	1	,001	,079	,017	,372
	service	-,315	,536	,345	1	,557	,730	,255	2,088
	rapidite	-,768	,305	6,341	1	,012	,464	,255	,843
	presentation	,052	,494	,011	1	,916	1,054	,400	2,775
	proprete	-1,221	,755	2,619	1	,106	,295	,067	1,294
	[arrondissement=1]	-3,128	2,830	1,221	1	,269	,044	,000	11,241
	[arrondissement=2]	-1,190	2,495	,227	1	,633	,304	,002	40,479
	[arrondissement=3]	2,215	2,059	1,158	1	,282	9,161	,162	517,948
	[arrondissement=4]	0 ^b			0				
	[auto=0]	-1,871	1,739	1,158	1	,282	,154	,005	4,648
	[auto=1]	0 ^b			0				
	[etudiant=0]	-1,920	1,741	1,217	1	,270	,147	,005	4,445
	[etudiant=1]	0 ^b			0				
JaunePartout	Intercept	55,615	15,347	13,132	1	,000			
	emplacement	-1,385	,671	4,259	1	,039	,250	,067	,933
	prix	2,936	,829	12,551	1	,000	18,843	3,713	95,626
	variete	-1,432	,694	4,255	1	,039	,239	,061	,931
	qualite	-2,888	,879	10,804	1	,001	,056	,010	,312
	service	-,252	,665	,143	1	,705	,778	,211	2,864
	rapidite	-,835	,365	5,244	1	,022	,434	,212	,887
	presentation	-,530	,562	,890	1	,346	,589	,196	1,770
	proprete	-2,119	,850	6,217	1	,013	,120	,023	,636
	[arrondissement=1]	-7,719	3,384	5,204	1	,023	,000	6E-007	,337
	[arrondissement=2]	-4,375	2,951	2,198	1	,138	,013	4E-005	4,089
	[arrondissement=3]	1,725	2,337	,545	1	,460	5,611	,058	546,965
	[arrondissement=4]	0 ^b			0				
	[auto=0]	-3,393	1,989	2,911	1	,088	,034	,001	1,657
	[auto=1]	0 ^b			0				
	[etudiant=0]	-1,808	2,144	,711	1	,399	,164	,002	10,953
	[etudiant=1]	0 ^b			0				

a. The reference category is: SuperSelected.

b. This parameter is set to zero because it is redundant.

FIG. 20.4 – Les coefficients des équations

L'interprétation des Logits n'est pas facile et, généralement, il est plus simple d'interpréter les *odds* associés. Les statistiques de Wald s'interprètent de la même manière que dans le cas de la régression logistique binaire, et l'effet des coefficients aussi.

Par exemple, dans la première équation, on voit que la variable **prix** a un impact significatif puisque sa *p*-value est de $0,003 < 0,05$. Le coefficient de cette variable dans cette équation est de 2,210 ; donc plus la satisfaction par rapport aux prix augmente, plus le *odds* augmente (il est multiplié par 9,120 pour chaque unité qui s'ajoute), favorisant Profitgros. Autrement dit, si un individu est satisfait des prix, la probabilité qu'il soit allé

au Profitgros est plus grande que celle qu'il soit allé au SuperSelect. Remarquez qu'on a le phénomène inverse pour la variable `qualite` puisque son coefficient est négatif; plus un individu est satisfait de la qualité des produits, plus la probabilité qu'il soit allé au SuperSelect augmente.

Observed	Predicted				Percent Correct
	Profitgros	JaunePartout	SuperSelect		
Profitgros	64	6	4		86,5%
JaunePartout	4	74	0		94,9%
SuperSelect	6	0	53		89,8%
Overall Percentage	35,1%	37,9%	27,0%		90,5%

FIG. 20.5 – La table de classification

La figure 20.5 présente la table de classification. Le modèle semble excellent ; 86,5 % de ceux qui sont allés au Profitgros ont été bien classés, 94,9 % pour ceux qui sont allés au JaunePartout et 89,8 % pour ceux qui sont allés au SuperSelect.

nom	EST1_1	EST2_1	EST3_1	PRE_1
Profitgros	,78	,22	,00	Profitgros
Profitgros	,95	,05	,00	Profitgros
Profitgros	,93	,01	,06	Profitgros
Profitgros	,97	,03	,00	Profitgros
Profitgros	,96	,04	,00	Profitgros
Profitgros	,99	,01	,00	Profitgros
Profitgros	,59	,41	,00	Profitgros
Profitgros	,34	,01	,65	SuperSelect
Profitgros	,99	,01	,01	Profitgros
JaunePartout	,27	,73	,00	JaunePartout

FIG. 20.6 – Les variables sauvegardées

Finalement, la figure 20.6 présente un extrait de la base de données où l'on peut voir les variables sauvegardées suite à cette analyse. Les variables `EST1_1`, `EST2_1` et `EST3_1` contiennent les probabilités que les individus soient allés au Profitgros, au JaunePartout

et au SuperSelect respectivement. Par exemple, sur la première ligne, on voit que le modèle a estimé à 78 % la probabilité que l'individu soit allé au Profitgros, à 22 % la probabilité qu'il soit allé au JaunePartout et à 0 % la probabilité qu'il soit allé au SuperSelect. La colonne PRE_1 contient simplement le nom correspondant à la plus grande probabilité. Donc pour l'individu de la première ligne on a Profitgros, et on peut constater que c'est effectivement là qu'il est allé (colonne nom).

Exemple 20.2.1 Une compagnie qui se spécialise dans les céréales et les déjeuners a fait une enquête auprès de 880 individus ; ceux-ci ont goûté à trois produits (barres de céréales, gruau d'avoine et céréales) et ont ensuite indiqué lequel des trois ils ont préféré. On a aussi noté leur âge, sexe, statut marital et s'ils ont ou non une vie active (ils sont considérés actifs si ils font de l'exercice au moins deux fois par semaine). Ces informations se retrouvent dans la base de données `céréales.sav`. On veut utiliser la régression logistique multinomiale pour tenter de déterminer un profil type de consommateur pour chacun des produits.

La variable dépendante sera donc `dejeuner` (le produit préféré), et les indépendantes sont `agecat`, `sexe`, `statut` et `actif`. Un examen de la base de données nous montre que les quatre variables indépendantes sont discrètes. Dans ce cas, il est possible de faire un test de plus pour l'ajustement du modèle ; il suffit de cocher `Cell probabilities` et `Goodness-of-fit` dans le bouton `Statistics`.

Le fait de travailler avec des variables indépendantes discrètes nous permet d'utiliser un test du χ^2 un peu comme dans les tableaux croisés. Effectivement, puisque nous sommes en présence de variables discrètes, il est possible de construire un (immense !) tableau avec toutes les modalités des variables discrètes indépendantes. La figure 20.15 nous en donne un aperçu. La différence d'avec les tableaux croisés usuels, c'est que les fréquences théoriques sont ici les prédictions issues du modèle. On espère donc que les fréquences théoriques soient près des fréquences observées.

Pour que le test soit valide, il faut que les cellules soit « assez bien remplies » ; ici la figure 20.7 nous indique que 11,4 % des cellules ont des fréquences nulles. Puisque ce

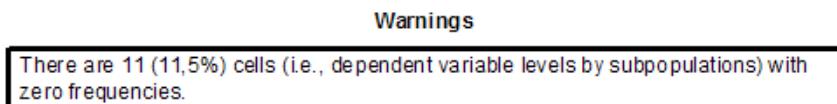


FIG. 20.7 – Avertissement

pourcentage n'est pas très grand, on peut être confiant dans le test du χ^2 . On reviendra à ce test un peu plus loin.

		Case Processing Summary	
		N	Marginal Percentage
dejeuner	Barre de céréales	231	26,3%
	Gruau (avoine)	310	35,2%
	Céréales	339	38,5%
agecat	Moins de 31 ans	181	20,6%
	31 à 45 ans	206	23,4%
	46 à 60 ans	231	26,3%
	Plus de 60 ans	262	29,8%
sexe	masculin	424	48,2%
	féminin	456	51,8%
statut	Pas marié	303	34,4%
	Marié	577	65,6%
actif	Inactif	474	53,9%
	Actif	406	46,1%
Valid		880	100,0%
Missing		0	
Total		880	
Subpopulation		32	

FIG. 20.8 – La distribution des fréquences des variables

La figure 20.8 nous montre simplement les fréquences observées pour chacune des modalités des variables. On voit aussi qu'il y a 32 sous-populations, c'est-à-dire qu'il y a 32 combinaisons possibles avec les modalités des variables discrètes indépendantes. On a donc un tableau avec 32 cellules.

La figure 20.9 nous présente les -2LL des modèles constant et complet. Celui du modèle constant est de 615,157, et celui du modèle complet est de 212,813. La différence entre les deux est significative puisque la p -value est nulle ; la valeur à battre est donc battue.

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	615,157			
Final	212,813	402,344	12	,000

FIG. 20.9 – Les -2LL

Pseudo R-Square	
Cox and Snell	,367
Nagelkerke	,414
McFadden	,210

FIG. 20.10 – Les pseudo r^2

La figure 20.10 nous montre qu'environ 41,4 % de la variation de la dépendante est expliquée par le modèle. Ce n'est pas extraordinaire, mais il ne faut pas s'attendre à des miracles avec un profil aussi sommaire des consommateurs.

La figure 20.11 est très intéressante car elle compare, pour chaque variable, la performance du modèle complet avec la performance du modèle auquel on a enlevé cette variable. On regarde alors la différence entre les -2LL des deux modèles ; lorsque la différence est grande, l'apport de la variable est jugé significatif et alors sa p -value est nulle. Ici on voit par exemple que seule la variable `sexe` est jugée non-significative. Et c'est la variable de l'âge qui semble avoir le plus d'impact car lorsqu'on l'enlève du modèle le -2LL passe de 212,813 à 533,697.

Effect	Model Fitting Criteria -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests		
		Chi-Square	df	Sig.
Intercept	212,813 ^a	,000	0	,
agecat	533,697	320,885	6	,000
sexe	213,408	,595	2	,743
statut	238,520	25,707	2	,000
actif	237,568	24,755	2	,000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

FIG. 20.11 – L'apport de chacune des variable

La figure 20.12 contient les coefficients des équations du modèle. On a

$$\ln \left(\frac{P(\text{Barre de céréales})}{P(\text{Céréales})} \right) = -1,167 + 0,993X_{\text{-31 ans}} + 1,316X_{31-45 \text{ ans}} + 0,552X_{46-60 \text{ ans}} \\ -0,135X_{\text{hommes}} + 0,840X_{\text{pas marié}} - 0,793X_{\text{inactif}}.$$

$$\ln \left(\frac{P(\text{Grau})}{P(\text{Céréales})} \right) = 1,136 - 4,272X_{\text{-31 ans}} - 2,531X_{31-45 \text{ ans}} - 1,191X_{46-60 \text{ ans}} \\ -0,006X_{\text{hommes}} - 0,26X_{\text{pas marié}} + 0,185X_{\text{inactif}}.$$

Les *p*-values des coefficients confirment que seule la variable **sexe** n'est pas significative car c'est la seule qui a de grandes *p*-values pour les deux équations.

L'interprétation se fait de la même façon que dans l'exemple précédent. Par exemple, le fait d'être inactif va faire augmenter les chances de préférer les céréales aux barres de céréales puisque le coefficient est négatif. Par contre le fait d'être actif ou pas ne fait pas de différence entre le grau et les céréales puisque la *p*-value est de 0,324.

		Parameter Estimates							
		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
dejeuner ^a								Lower Bound	Upper Bound
Barre de céréales	Intercept	-1,167	,322	13,105	1	,000			
	[agecat=1]	,993	,318	9,747	1	,002	2,699	1,447	5,034
	[agecat=2]	1,316	,322	16,671	1	,000	3,730	1,983	7,018
	[agecat=3]	,552	,342	2,602	1	,107	1,736	,888	3,393
	[agecat=4]	0 ^b			0				
	[sexe=0]	-,135	,180	,561	1	,454	,874	,614	1,244
	[sexe=1]	0 ^b			0				
	[statut=0]	,840	,194	18,808	1	,000	2,315	1,584	3,383
	[statut=1]	0 ^b			0				
	[actif=0]	-,793	,183	18,693	1	,000	,452	,316	,648
	[actif=1]	0 ^b			0				
Gruau (avoine)	Intercept	1,136	,238	22,790	1	,000			
	[agecat=1]	-4,272	,534	64,121	1	,000	,014	,005	,040
	[agecat=2]	-2,531	,282	80,424	1	,000	,080	,046	,138
	[agecat=3]	-1,191	,218	29,713	1	,000	,304	,198	,466
	[agecat=4]	0 ^b			0				
	[sexe=0]	-,006	,183	,001	1	,973	,994	,695	1,422
	[sexe=1]	0 ^b			0				
	[statut=0]	-,260	,214	1,477	1	,224	,771	,507	1,173
	[statut=1]	0 ^b			0				
	[actif=0]	,185	,188	,972	1	,324	1,204	,833	1,740
	[actif=1]	0 ^b			0				

a. The reference category is: Céréales.

b. This parameter is set to zero because it is redundant.

FIG. 20.12 – Les coefficients

La classification (figure 20.13) est assez moyenne dans l'ensemble (57,4 % de réussite), mais elle est quand même meilleure que celle du modèle constant qui aurait donné un % de réussite de 38,5 % (plus grand groupe). Par groupe, elle dépasse aussi le 33,3 % qu'on aurait obtenu par chance. Et pour le gruau on obtient tout de même un 77,1 % de réussite, ce qui n'est pas mal du tout. Le profil pour le gruau est donc assez fiable : c'est avant tout les gens plus âgés qui préfèrent ce type de céréales.

Tel que discuté précédemment, un test supplémentaire s'offre à nous lorsque les variables indépendantes sont discrètes. Un tableau est alors construit, et les fréquences prédites de ce tableau sont issues du modèles (la figure 20.15 donne un aperçu du tableau). Par exemple, on voit que pour un individu inactif, pas marié, masculin et de moins de 31 ans, il est prédit que 42,5 % de ceux-ci vont préférer les barres de céréales,

Observed	Classification			
	Barre de céréales	Gruau (avoine)	Céréales	Percent Correct
Barre de céréales	116	30	85	50,2%
Gruau (avoine)	19	239	52	77,1%
Céréales	81	108	150	44,2%
Overall Percentage	24,5%	42,8%	32,6%	57,4%

FIG. 20.13 – La table de classification

2,2 % vont préférer le gruau et 55,3 % vont préférer les céréales. Ces probabilités sont données pour toutes les combinaisons possibles. Lorsque ces probabilités sont près des pourcentages observés, on peut alors conclure que le modèle s'ajuste bien aux données.

C'est la sortie 20.14 qui permet de tester l'ajustement du modèle. Ici l'hypothèse nulle est que l'ajustement est bon ; on espère donc avoir une grande p -value pour ne pas rejeter H_0 .

Goodness-of-Fit			
	Chi-Square	df	Sig.
Pearson	36,684	50	,920
Deviance	44,250	50	,702

FIG. 20.14 – Qualité d'ajustement du modèle

Deux tests sont effectués : un à partir de la statistique du χ^2 usuelle (ligne Deviance), et l'autre à partir des résidus de Pearson (ligne Pearson). Ces deux statistiques mesurent l'écart entre les fréquences prédites par le modèle et les fréquences observées et suivent une loi du χ^2 . Lorsque l'écart n'est pas trop grand, on conclut que le modèle est bon. C'est le cas ici puisque les deux p -values sont grandes (0,920 et 0,702). Donc notre modèle est jugé adéquat.

					Observed and Predicted Frequencies				
actif	statut	sexe	agecat	dejeuner	Frequency			Percentage	
					Observed	Predicted	Pearson Residual	Observed	Predicted
Inactif	Pas marié	masculin	Moins de 31 ans	Barre de céréales	7	7,650	-,310	38,9%	42,5%
				Gruau (avoine)	0	,399	-,639	,0%	2,2%
				Céréales	11	9,951	,497	61,1%	55,3%
		31 à 45 ans		Barre de céréales	6	4,637	,864	60,0%	46,4%
				Gruau (avoine)	0	,998	-1,053	,0%	10,0%
				Céréales	4	4,365	-,232	40,0%	43,6%
		46 à 60 ans		Barre de céréales	0	1,044	-1,149	,0%	20,9%
				Gruau (avoine)	2	1,844	,144	40,0%	36,9%
				Céréales	3	2,112	,804	60,0%	42,2%
		Plus de 60 ans		Barre de céréales	1	2,124	-,799	3,2%	6,9%
				Gruau (avoine)	21	21,421	-,164	67,7%	69,1%
				Céréales	9	7,456	,649	29,0%	24,1%
féminin		Masculin	Moins de 31 ans	Barre de céréales	7	6,873	,066	46,7%	45,8%
				Gruau (avoine)	0	,315	-,567	,0%	2,1%
				Céréales	8	7,812	,097	53,3%	52,1%
		31 à 45 ans		Barre de céréales	4	5,965	-1,135	33,3%	49,7%
				Gruau (avoine)	1	1,129	-,127	8,3%	9,4%
				Céréales	7	4,906	1,229	58,3%	40,9%
		46 à 60 ans		Barre de céréales	4	4,398	-,216	21,1%	23,1%
				Gruau (avoine)	10	6,830	1,516	52,6%	35,9%
				Céréales	5	7,772	-1,294	26,3%	40,9%
		Plus de 60 ans		Barre de céréales	4	3,788	,114	8,2%	7,7%
				Gruau (avoine)	33	33,591	-,182	67,3%	68,6%
				Céréales	12	11,621	,127	24,5%	23,7%
Marié		Masculin	Moins de 31 ans	Barre de céréales	5	3,359	1,027	35,7%	24,0%
				Gruau (avoine)	0	,526	-,739	,0%	3,8%
				Céréales	9	10,116	-,666	64,3%	72,3%

FIG. 20.15 – Aperçu du tableau

20.3 Exercice du chapitre

La base de données `hotels.sav` contient les variables d'une étude qui visait à savoir si certaines variables socio-démographiques peuvent avoir une influence sur le choix d'une chaîne d'hôtels lors d'un voyage lié au travail. Les variables sont les suivantes :

Nom	Label	Précision
ID	Numéro d'identification	
HOTELPR	Choix de l'hôtel préféré. 3 chaînes d'hôtels seulement *	1=Chaîne A (***) 2=Chaîne B (*****) 3=Chaîne C (*)
PROTRA	Proche du lieu de travail	0=Pas important à 9=Très important
PROACT	Proche des restaurants et sorties, magasins	0=Pas important à 9=Très important
AGE	Âge du répondant	1=(18-34), 2=(35-49), 3=(50 et +)
SEXE	Sexe du répondant	0=Femme, 1=Homme
NSCOLA	Niveau de scolarité du répondant	1=post universitaire, 2=universitaire, 3=collégial, 4=technique, 5=secondaire
FONCT	Fonction du répondant	1=haute direction, 2=direction, 3=professionnel
REVENU	Revenu annuel du répondant	1=moins de 50 milles, 2=50-74 milles, 3=75 milles et +
LANGUE	Langue du répondant	1=anglais, 2=français

Pour répondre à cette question, faites une régression logistique multinomiale avec la variable `hotelpr` comme variable dépendante.

Chapitre 21

Modèles d'équations structurelles

(MES)

21.1 Introduction

Les modèles d'équations structurelles (MES, en anglais *SEM : Structural Equations Model*) combinent plus d'une technique afin de bien modéliser les liens qui existent entre des concepts. Elles permettent non seulement de modéliser plusieurs liens consécutifs et parallèles, mais aussi de considérer que ces concepts ne sont pas nécessairement mesurés parfaitement.

Pour mieux comprendre on peut comparer un modèle d'équations structurelles à un modèle de régression linéaire multiple classique. La figure 21.1 illustre un modèle de régression linéaire multiple : toutes les variables explicatives (ici X_1 , X_2 et X_3) ont un accès direct à la variable dépendante Y . Une variable est indépendante ou dépendante, elle ne peut être les deux à la fois.

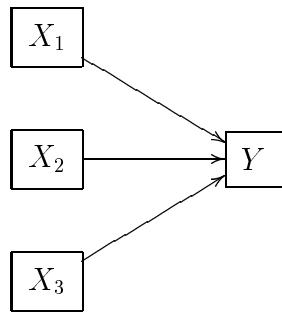


FIG. 21.1 – Un modèle de régression linéaire multiple

La figure 21.2 présente un modèle d'équations structurelles (mais quelque peu simplifié pour l'instant). Les cercles représentent des concepts, et on voit que plusieurs liens sont étudiés. Par exemple, ce modèle évaluera l'impact de ξ_1 sur η_1 , puis l'impact de η_1 sur η_4 . On remarque donc que η_1 est à la fois un concept explicatif et expliqué, ce qu'on ne retrouve pas dans une régression linéaire classique. Et ici, l'effet que peut avoir ξ_1 sur η_4 transige par η_1 et η_2 .

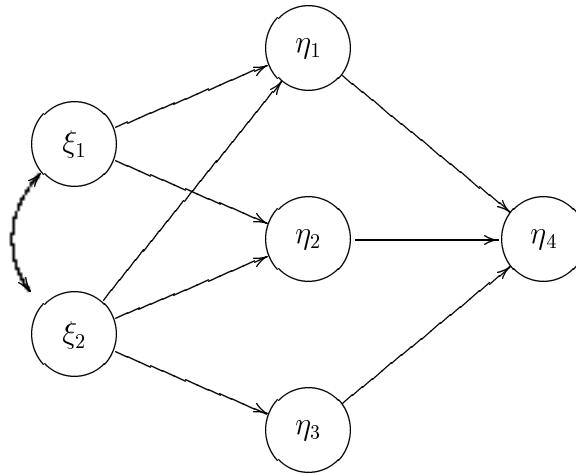


FIG. 21.2 – Un modèle d'équations structurelles (avec les concepts latents seulement)

Les concepts ξ_1 et ξ_2 sont dit **exogènes** car ils ne sont le but d'aucune flèche de dépendance. Les autres concepts, qui eux sont le but d'au moins une flèche de dépendance, sont dit **endogènes**.

La figure 21.3 présente un autre exemple d'un diagramme représentant un modèle d'équations structurelles, complet celui-ci. Les concepts latents sont ξ et η ; ξ est exogène

tandis que η est endogène. ξ a deux indicateurs, soit x_1 et x_2 . Les indicateurs sont les variables observables ; par exemple, un concept latent pourrait être la perception d'utilité d'une nouvelle technologie, et les indicateurs de ce concept seraient les questions que l'on pose pour mesurer cette perception d'utilité. Comme les questions que l'on pose ne peuvent mesurer de façon parfaite la perception d'utilité, on leur associe à chacune un terme d'erreur. Par exemple, dans le diagramme, le terme d'erreur de x_1 est δ_1 .

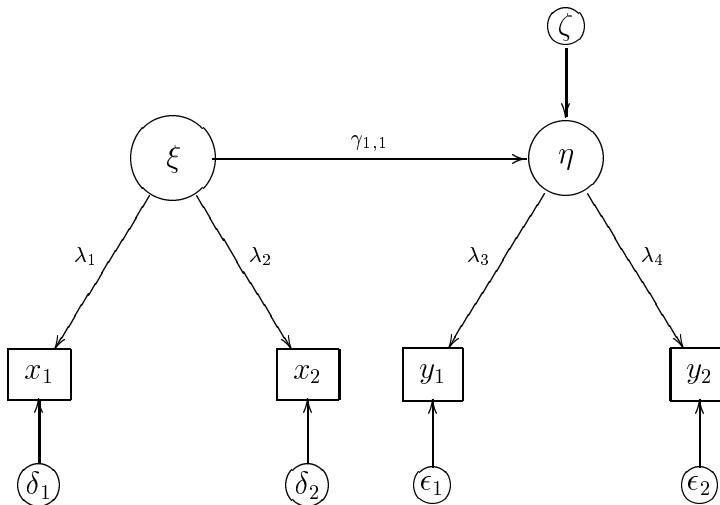
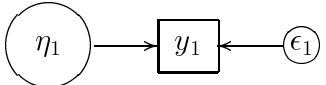
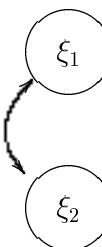


FIG. 21.3 – Exemple d'un diagramme

Le concept η a lui aussi deux indicateurs, y_1 et y_2 . De plus, η étant endogène, un terme d'erreur ζ lui est associé. Ce terme permet de mesurer à quel point le modèle explique ce concept endogène, ce qui nous donnera une mesure qui est l'équivalent du r^2 en régression linéaire multiple.

Le tableau qui suit présente les différents symboles utilisés dans les diagrammes d'équations structurelles.

Symboles utilisés dans les diagrammes

	Un rectangle désigne une variable observée.
	Un cercle ou une ellipse désigne une variable latente (non observée).
	Les petits cercles désignent les termes d'erreur, et une flèche droite signifie que la variable à la base de la flèche « cause » la variable vers laquelle la flèche pointe.
	Une flèche à double-tête et courbée désigne une association non analysée entre deux variables exogènes.

La figure 21.14 montre un diagramme d'équations structurelles conçu dans AMOS. On reconnaît les éléments décrit précédemment, avec des noms significatifs cette fois. Par exemple, le concept exogène **Attitudes Cow** représente le concept de *Attitudes Toward Coworkers*. Ce concept a été mesuré avec quatre questions (AC1, AC2, AC3 et AC4). Nous reviendrons à cet exemple plus loin dans le chapitre. Nous verrons alors que non seulement les MES nous permettent d'évaluer rapidement tous les liens d'un modèle, mais aussi que le fait de travailler avec des variables latentes nous permet de beaucoup mieux estimer les liens qu'avec un modèle de régression linéaire classique.

Ce chapitre se veut une introduction aux MES. La section 21.2 présente l'analyse des chemins, qui sont à l'origine des MES. La section 21.3 présente brièvement la méthode du maximum de vraisemblance, qui est souvent utilisée pour estimer les paramètres des MES. La section 21.4 présente les mesures statistiques qui permettent de vérifier si un MES s'ajuste bien à l'échantillon. Ensuite la section 21.5 présente les concepts de construit

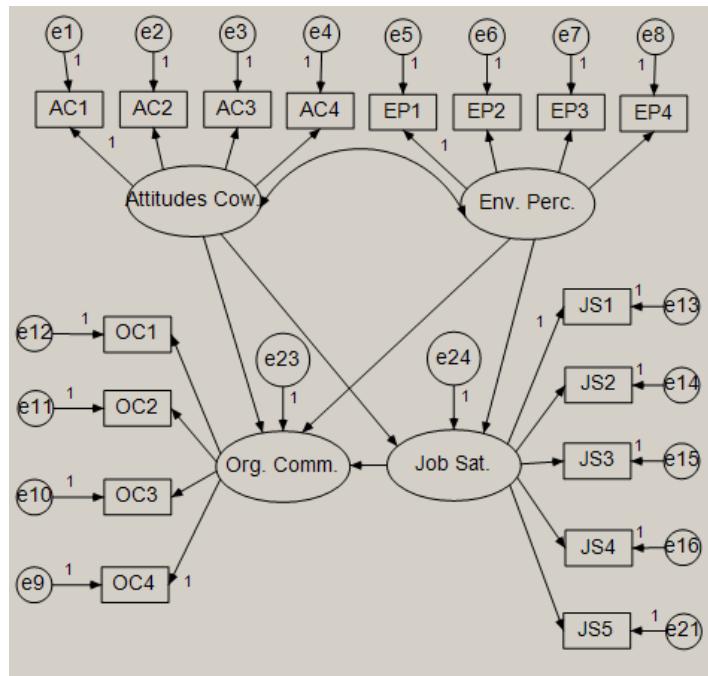


FIG. 21.4 – Exemple de représentation d'un modèle dans AMOS

formatif et réflexif. Les sections 21.6 et 21.7 illustrent les différentes étapes d'analyse d'un modèle de mesure et structurel. Les dernières sections abordent les concepts de médiation et le logiciel AMOS.

21.2 Les origines : *Path analysis*

Les modèles d'équations structurelles ont pris leur origine en exploitant et en généralisant une technique de base appelée « *Path Analysis* ». Inspirée des réalités perçues de la vie, cette technique tente de modéliser les liens de causalités entre des variables par l'entremise d'un graphique fléché appelé le « *Path Diagram* » (diagramme des chemins). La figure 21.5 illustre un tel diagramme.

Suivant le tracé d'un modèle de causalités proposé par le chercheur, le *Path Analysis* estime la puissance des liens unissant ces variables. Pour ce faire, cette technique utilise

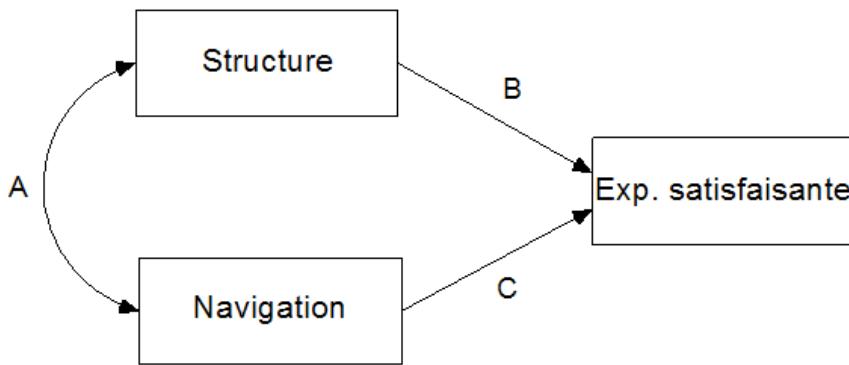


FIG. 21.5 – Modèle simplifié de la satisfaction de l'expérience

Variables	Exp. Satisfaisante	Structure	Navigation
Exp. Satisfaisante	1		
Structure	0,490	1	
Navigation	0,476	0,631	1

FIG. 21.6 – Les corrélations

la matrice des corrélations entre les variables du modèle à l'étude. Supposons qu'un chercheur tente de déterminer si la structure ainsi que les éléments facilitant la navigation sur un site Internet d'achats en ligne ont une influence sur la satisfaction de l'expérience d'achat du consommateur. La figure 21.5 présente le diagramme des chemins du modèle proposé tandis que le tableau 21.6 associé présente les corrélations standard dites de Pearson entre les trois variables. Mentionnons simplement que les trois variables utilisées sont des construits obtenus en effectuant la moyenne de différentes variables leur étant associées.

Dans une analyse de type *Path analysis*, l'effet total entre deux variables est décomposé, selon le modèle étudié, en effet direct et indirect.

Tout d'abord, les corrélations du tableau 21.6 représentent les effets totaux, et se décomposent de la façon suivante, selon les chemins du diagramme :

$$\begin{aligned}
 r_{\text{struct, navig}} &= A \\
 r_{\text{struct, exp sat}} &= B + AC \\
 r_{\text{navig, exp sat}} &= C + AB
 \end{aligned}$$

Ainsi, par exemple, B représente l'effet direct entre la structure et l'expérience satisfaisante, tandis que AC est l'effet indirect entre ces deux concepts. Cet effet indirect transige par la navigation ; ceci signifie que si la perception par rapport à la structure augmente, alors la perception par rapport aux éléments facilitant la navigation va augmenter elle aussi, et à son tour cette augmentation entraînera l'augmentation de la perception de la satisfaction par rapport à l'expérience.

On a donc les équations suivantes :

$$\begin{aligned}
 0,631 &= A \\
 0,49 &= B + AC \\
 0,476 &= C + AB
 \end{aligned}$$

En les résolvant, on trouve $A = 0,631$, $B = 0,316$ et $C = 0,277$. La figure 21.7 présente ces valeurs reportées sur le diagramme étudié.

De cette manière, en suivant un modèle proposé et en décortiquant les effets directs et indirects, l'analyste est en mesure d'évaluer des modèles simples à la main. Cependant, lorsque le modèle devient complexe, certains principes s'appliquent et le système d'équation en cause est plus difficile à dégager. Il faut alors appliquer la règle de Duncan qui permet essentiellement de naviguer sur les chemins comme une araignée sur sa toile. Pour plus de détails sur cette technique manuelle, voir Maruyama Geoffrey M. : *Basics of Structural Equation Modeling*, Sage, London (1997).

Ici, puisque le modèle comporte, dans sa partie exogène, une relation d'interdépendance entre les variables indépendantes, et que nos variables sont observées et non pas

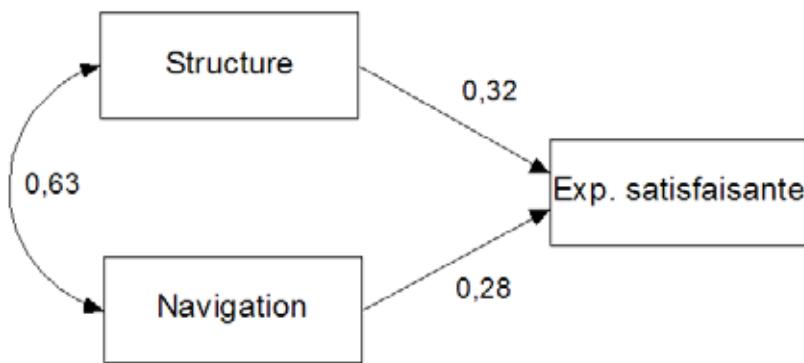


FIG. 21.7 – Modèle avec les coefficients

latentes, il se trouve qu'une régression linéaire multiple aurait donné les mêmes résultats. Cependant, ce sont les coefficients standardisés qui correspondent à ces résultats.

Pour comprendre ce que sont les coefficients standardisés, prenons justement le tableau issu de la régression linéaire multiple du modèle $Y_{\text{exp sat}} = \beta_0 + \beta_1 X_{\text{struct}} + \beta_2 X_{\text{navig}} + \epsilon_t$ (figure 21.8).

Model	Coefficients ^a				
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	3,569	,40307	8,855	,000
	Indice de la qualité de la structure du site	,297	,07038	,316	,000
	Indice de la facilité de la navigation	,256	,06934	,277	,000

a. Dependent Variable: Indice de satisfaction lié à l'expérience

FIG. 21.8 – Les estimations issues de la régression

L'équation usuelle est alors la suivante :

$$\hat{y}_{\text{exp sat}} = 3,569 + 0,297x_{\text{struct}} + 0,256x_{\text{navig}}$$

et on trouve $r^2 = 0,287$ et $r^2_{\text{ajusté}} = 0,28$. L'interprétation des coefficients se fait de la même façon que celle vue au chapitre sur la régression linéaire. Donc par exemple, si la perception de la structure augmente d'une unité et que la perception de la navigation est

constante, alors la perception de la satisfaction par rapport à l'expérience augmente en moyenne de 0,297 unité.

Les coefficients standardisés qui nous intéressent ici sont obtenus à partir d'une base de données où les trois variables auraient été préalablement standardisées (de moyennes 0 et de variances 1). Voici l'équation que l'analyste aurait obtenue à partir de variables standardisées :

$$\hat{y}_{\text{zexp sat}} = 0 + 0,316x_{\text{zstruct}} + 0,277x_{\text{znavig}}.$$

L'interprétation de ces coefficients se fait un peu de la même façon que les coefficients non-standardisés, mais au lieu de parler en termes d'unités, on parle en terme d'écart-type. Par exemple, si la perception de la structure augmente d'un écart-type et que la perception de la navigation est constante, alors la perception de la satisfaction par rapport à l'expérience augmente en moyenne de 0,316 écarts-type. Ce type d'interprétation se transpose au diagramme des chemins ainsi qu'aux coefficients d'un modèle de régressions structurelles.

Rappelons que toute droite de régression passe nécessairement par les moyennes des variables ; or les moyennes de variables standardisées sont nulles. Ainsi un modèle standardisé passe toujours par l'origine (le point (0,0, ...0)), qui donc constitue nécessairement l'ordonnée à l'origine du modèle, et donc la constante de l'équation est toujours nulle.

21.3 Le maximum de vraisemblance (MV)

Pour évaluer les coefficients d'une régression linéaire classique, la technique du moindre carré est habituellement utilisée. Cette technique minimise la variance des résidus et par conséquent la variance sur l'unique variable dépendante. Dans les méthodes d'équations structurelles, la technique la plus utilisée pour évaluer les coefficients est plutôt la technique du maximum de vraisemblance.

Mais qu'est-ce que le maximum de vraisemblance ? Essentiellement, cette technique tente d'évaluer les paramètres de la régression qui auraient pu, vraisemblablement, produire un tel échantillon. Même que dans certaines circonstances, la formule algébrique d'une statistique issue d'une telle technique change en fonction de ce qui se retrouve dans l'échantillon. Ce qui n'est pas le cas pour les statistiques usuelles.

Il existe d'autres méthodes d'estimation permettant d'évaluer des équations structurelles. Nous ne faisons ici que les mentionner :

- Unweighted Least Square (ULS) ;
- Genelarized Least Squares (GLS) ;
- Asymptotically Distribution Free (ADF) ;
- Scale-Free Least Squares (SFLS).

Les méthodes autres que le MV ont été développées pour compenser le problème théorique du MV face à l'absence de multinormalité des données. Cependant, la pratique a montré que, dans bien des cas, le MV fournit des résultats plus fiables que les autres techniques, et ce, même lorsque la mutinormalité n'est pas respectée.

Selon la complexité du modèle et la performance du modèle de mesure, il est conseillé d'avoir entre 150 et 450 observations pour estimer un modèle d'équations structurelles sous le MV.

Mentionnons finalement que l'analyse des résidus permettant de valider les hypothèses sous-jacentes à une régression linéaire classique est, en pratique, impossible en MES. En effet, au lieu d'utiliser les données brutes pour ses calculs, les MES utilisent la matrice des corrélations ou la matrice de variances-covariances. Le principe est de calculer la véritable matrice des variances-covariances sur toutes les variables impliquées dans le modèle, puis d'estimer la matrice des variances-covariances associée au modèle. En somme, l'analyse de l'efficacité d'une régression structurale sera davantage fastidieuse que dans le cadre d'une régression multiple, car il faut évaluer à quel point ces deux matrices sont semblables, mais avec certaines nuances.

Mentionnons finalement que les MES ne sont pas conçues pour établir des prédictions

ou des estimations, donc si tel est le but il vaut mieux utiliser une régression linéaire classique.

21.4 Les mesures d'adéquation

Pour un MES, répondre à la légitime question : « Est-ce que le modèle s'ajuste bien aux données ? » est définitivement plus complexe que pour un modèle de régression linéaire. En effet, en régression linéaire, l'appréciation du coefficient de détermination ajusté et l'analyse des résidus permettent de se faire une idée de la performance de la régression. Rappelons simplement que l'analyse des résidus permet de vérifier les hypothèses de base liées à une bonne régression, soit la linéarité de la relation, l'homoscédasticité de la dispersion des résidus, la normalité des résidus et l'indépendance des résidus et des variables indépendantes. Cependant, pour obtenir les résidus en question, il est nécessaire de travailler avec les données brutes.

Au lieu d'utiliser les données brutes, les logiciels de MES utilisent les matrices de variances-covariances et/ou de corrélations, qui sont des données agrégées. Ainsi l'analyse classique des résidus est de cette manière impossible. Cependant, la problématique de l'évaluation de l'adéquation du modèle est toujours présente et se fera par l'entremise d'une batterie de statistiques.

Le tableau 21.9 présente l'ensemble des statistiques liées à l'évaluation d'un modèle donné. La performance d'un modèle peut être analysée suivant trois angles différents : l'angle de la performance globale ou absolue, l'angle de la performance relative à des modèles théoriques et finalement, l'angle de la parcimonie. Soulignons que les statistiques présentées dans le tableau 21.9 ainsi que les valeurs clefs ne sont valides que si les estimations ont été produites sous le maximum de vraisemblance.

Malgré la panoplie de statistiques disponibles, aucune n'a donné en solitaire des résultats satisfaisant et il est plutôt recommandé d'utiliser un panel de statistiques afin

Indices	Valeur Clef (<i>threshold</i>)
Indices absolus	
χ^2 ou Scaled χ^2 *	Aucune (<i>p-value</i>)
GFI*, AGFI, Gamma 1 et 2	> 0,9
PNI	Le plus faible possible
PNNI	> 0,95
Hoelter's Critical. N	> 200
RMR et RMSR*	Le plus proche de 0 et fixé par le chercheur
RMSEA	< 0,08 et si possible < 0,05
Indices incrémentaux	
Type I	
NFI* et BL86	> 0,9
Type II	
TLI* (NNFI) et IFI (BL89)	> 0,9
Type III	
CFI et BFI (RNI)	> 0,9
Indices de parcimonie	
χ^2 normé*	Le plus faible entre 1 et 3, voire 5
AIC, CAIC, CAK et ECVI	Le plus faible possible en comparaison
BCC, BIC et MECVI	Le plus faible possible en comparaison
PNFI et PGFI	Le plus fort possible en comparaison

* Indices recommandés lors d'une analyse de type factorielle confirmatoire.

FIG. 21.9 – Les statistiques d'ajustement sur le MV

de se faire une idée de l'adéquation du modèle. Roussel (2002) propose d'utiliser deux indices absous, deux indices incrémentaux (si possible un de type II et un de type III, ceux de type I étant à éviter (Hu et Bentler (1995)) et finalement un ou deux indices de parcimonie. Des sous-sections s'attarderont plus précisément à chacune de ces familles.

Il est important de noter que les valeurs clefs du tableau 21.9 sont des bornes généralement admises comme représentant un bon ajustement et non pas comme étant la seule règle à partir de laquelle un modèle est rejeté. Le bon sens du chercheur et la tolérance sont parfois au rendez-vous car, tout comme en régression, il est préférable de travailler sur un modèle qui s'interprète bien qu'un modèle qui obtient le meilleur ajustement.

Par exemple, si le modèle est plus complexe, le chercheur peut être moins sévère envers les bornes critiques car le modèle a définitivement plus de chances d'être rejeté. Roussel *et al.* (2002) soutiennent que si une analyse factorielle confirmatoire précède le test d'un modèle, il est préférable d'être plus strict lors de l'analyse factorielle que lors du test du modèle global. Ils proposent de fixer le seuil des indices d'ajustement GFI, AGFI et autres à 0.95 au lieu de 0.90 pour l'analyse de la validité des mesures et à 0.90 pour l'analyse de l'adéquation du modèle global.

Le même type de souplesse peut s'appliquer lorsque deux catégories d'indices sont acceptables (absous et incrémentaux par exemple) tandis que la troisième catégorie est médiocre ; le chercheur peut accepter le modèle moyennant la présence d'une discussion à cet effet dans son article.

Hoyles et Panter (1995) précisent qu'avant de présenter les résultats d'un modèle structurel il faut :

- Préciser et justifier la liste des indices qui seront utilisés.
 - Donner une définition et une interprétation de chacun des indices, surtout si le résultat de l'un d'entre eux n'est pas bon.
 - Indiquer les valeurs critiques retenues et justifier ces bornes si ces dernières sont inférieures aux valeurs généralement admises.
-

21.4.1 Les indices de mesure absolus

Les indices de mesure absolus permettent de se faire une idée de l'adéquation générale du modèle aux données. Ces statistiques évaluent le modèle dans sa globalité qui contient à la fois le modèle structurel et le modèle de mesure. En d'autres termes, avec ces statistiques, il ne sera pas possible de dégager la source potentielle d'un mauvais ajustement.

Il est possible, la plupart du temps, de comparer les statistiques de mesures absolues avec celles obtenues sur le modèle saturé et sur le modèle indépendant. Le modèle saturé est celui qui contient tous les chemins possibles entre toutes les variables ; ce modèle est toujours le meilleur théoriquement, mais il n'est pas intéressant du point de vue de l'interprétation. Le modèle indépendant est celui où toutes les variables sont isolées les unes des autres, il ne contient aucun chemin. Ce dernier a toujours le pire ajustement.

Ainsi le modèle saturé est un modèle sans contraintes où tous les paramètres sont libres et donc estimés. Ce modèle reproduit entièrement la matrice de variances-covariances en corrélant tout sur son passage. Le chercheur désire que son modèle s'en approche. À l'opposé, le modèle indépendant suppose que toutes les corrélations entre les variables sont nulles.

21.4.2 Les indices incrémentaux

Les mesures incrémentales évaluent la qualité de l'ajustement du modèle par rapport au modèle indépendant, qui agit à titre d'échelon de référence représentant le cas ayant le pire des ajustements possibles. On peut aussi comparer ces mesures à celles du modèle saturé.

Compte tenu du nombre important d'indices incrémentaux, Marsh textitet al. (1988) proposèrent une typologie. Ils classèrent les statistiques incrémentales en trois types. Essentiellement, les indices de type I utilisent le moins d'informations et sont par ce fait moins riches. Les indices de type II utilisent davantage d'informations et exploitent une

loi du chi-deux dite non centralisée. Les indices de type III sont les plus riches en terme d'informations puisqu'elles sont un amalgame des deux premiers groupes.

21.4.3 Les indices de parcimonie

Pour améliorer les mesures d'ajustement, le chercheur peut être tenté de libérer un paramètre de sa contrainte (c'est-à-dire ajouter un chemin), l'amenant à être estimé et améliorant ainsi le degré d'ajustement du modèle. Les indices de parcimonie tiennent justement compte du nombre de paramètres à estimer. Cette stratégie amènera à étudier l'amélioration relative d'ajustement par paramètre estimé.

Essentiellement, les indices de parcimonie sont des dérivés des indices absolus ou incrémentaux ajustés au nombre de degrés de liberté de manière à tenir compte du nombre de paramètres estimés.

21.5 Concept réflexif ou formatif ?

Jusqu'à présent, dans la littérature, presque toutes les échelles de mesures présentées sont traitées comme étant réflexives, alors que ce n'est pas toujours le cas. Il arrive parfois qu'elles soient formatives, et alors l'ajustement du modèle qui utilise une telle échelle peut être mauvais si elle a été traitée comme étant réflexive. Mais avant de poursuivre, examinons le tableau 21.10 qui montre les différences entre les deux types de concept.

On voit que les indicateurs d'un concept réflexif sont très semblables, ce sont pratiquement des synonymes. Par conséquent, si un individu a une perception très positive pour l'un de ces indicateurs, alors il devrait en être de même pour les autres indicateurs. C'est en fait ce qu'on vérifie quand on mesure la cohérence interne d'un construit (avec l'alpha de Cronbach par exemple). On voit aussi que les indicateurs sont des manifestations du concept réflexif; on pourrait les comparer aux symptômes d'une maladie. Quand on relie

Concept formatif	Concept réflexif
Le sens de causalité va des indicateurs au concept.	Le sens de causalité va du construit aux indicateurs.
Les indicateurs sont des caractéristiques de définition du concept.	Les indicateurs sont des manifestations du concept.
Des changements dans les indicateurs devraient causer des changements dans le concept.	Des changements dans les indicateurs ne devraient pas causer des changements dans le concept.
Les indicateurs ne sont pas interchangeables en général, ils sont habituellement complémentaires.	Les indicateurs doivent être interchangeables, ils sont donc des synonymes.
Supprimer un des indicateurs peut altérer le concept lui-même.	Supprimer une des indicateurs ne devrait pas altérer le concept.
Il n'est pas nécessaire que les indicateurs soient corrélés entre eux.	Les indicateurs sont supposés être fortement corrélés entre eux.

FIG. 21.10 – Différences entre un concept formatif et un concept réflexif

les indicateurs réflexifs au concept latent, les flèches vont du concept vers les indicateurs, exprimant ainsi que les indicateurs émanent du concept (figure 21.11).

D'autre part, les indicateurs d'un concept formatif sont plutôt complémentaires que synonymes. Ils mesurent des aspects variés du concept, et n'obtiennent pas nécessairement la même perception chez un même individu. Il est ainsi insensé de vouloir mesurer la cohérence interne d'un concept formatif (mais la mauvaise cohérence interne d'un concept peut être un indice que celui-ci est formatif). Les indicateurs formatifs causent le concept ; on pourrait les comparer aux habitudes de vie qui peuvent causer une maladie. Quand on relie les indicateurs formatifs au concept latent, les flèches vont des indicateurs vers le concept, exprimant ainsi que les indicateurs causent le concept (figure 21.12). Le terme d'erreur est alors associé au concept plutôt qu'à chacun des indicateurs, et les indicateurs

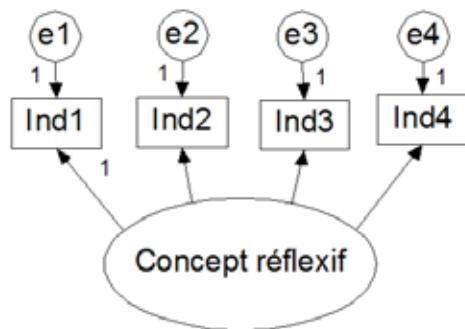


FIG. 21.11 – Un concept réflexif

doivent être reliés entre eux par des flèches de corrélation.

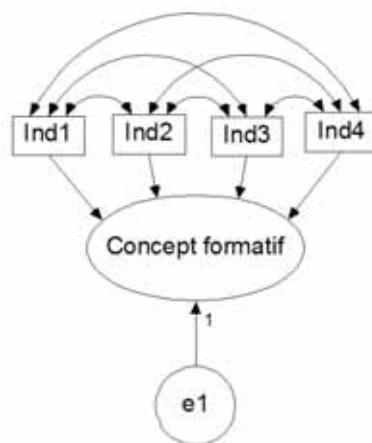


FIG. 21.12 – Un concept formatif

Dans ce qui suit, les exemples ne traitent que de concepts réflexifs, et les étapes proposées (surtout pour l'analyse factorielle confirmatoire) ne sont pas toujours adaptées pour les concepts formatifs.

Il est à noter que pour chaque concept latent réflexif présent dans un modèle, on recommande un minimum de trois indicateurs. Lorsque cette condition n'est pas respectée, il arrive que le logiciel soit incapable d'identifier une solution.

Les conditions sur les concepts formatifs latents pour permettre l'identification d'une solution sont un peu plus complexes. On propose, entre autres, soit d'ajouter deux indi-

cateurs réflexifs, soit d'avoir un lien de dépendance du construit formatif vers un concept réflexif.

21.6 L'analyse factorielle confirmatoire

Il est possible d'utiliser les MES pour faire une analyse factorielle confirmatoire, qui est en fait une étape préalable à l'étude du modèle structurel en tant que tel. En effet, le modèle structurel permet d'étudier les liens entre des variables latentes ; or à quoi sert-il d'étudier ces liens si on ne s'assure pas d'abord que nos variables latentes ont été mesurées correctement ? Il est vrai que lors de l'étude du modèle structurel le modèle de mesure est implicitement mesuré, mais il est préférable de commencer par étudier celui-ci en détail puisque tous les résultats reposent sur les mesures de nos concepts. C'est ce que permet de faire l'analyse factorielle confirmatoire en examinant la validité de nos mesures.

Nous verrons comment nous assurer que notre modèle de mesure est bon en vérifiant les mesures d'ajustement du modèle, puis la validité convergente, discriminante et nomologique. Nous illustrons la façon de faire avec un exemple.

21.6.1 Un exemple

Nous présentons ici un exemple tiré du livre *Multivariate Data Analysis*, Hair *et al.*, 2006, mais légèrement modifié. La base de données se nomme `hbatssemmod.sav`. L'entreprise HBAT souhaite comprendre quels sont les facteurs qui peuvent contribuer à la rétention des employés. Suite à une exploration de la littérature et quelques entrevues préliminaires avec des employés, on décide de considérer les cinq construits réflexifs suivants :

JS	<i>Job Satisfaction</i>
OC	<i>Organizational Commitment</i>
SI	<i>Staying Intentions</i>
EP	<i>Environmental Perceptions</i>
AC	<i>Attitudes Toward Coworkers</i>

Ces cinq construits auront comme variables indicatrices les suivantes :

Item	Description
JS1	<i>All things considered, I feel very satisfied when I think about my job.</i>
JS2	<i>When you think of your job, how satisfied you feel ?</i>
JS3	<i>How satisfied are you with your current job at HBAT ?</i>
JS4	<i>How satisfied are you with HBAT as an employer ?</i>
JS5	<i>Indicate your satisfaction with your current job at HBAT by placing a percentage in the blank, with 0 % = Not satisfied at all, and 100 % = Highly satisfied.</i>
OC1	<i>My work at HBAT gives me a sense of accomplishment.</i>
OC2	<i>I am willing to put in a great deal of effort beyond that normally expected to help HBAT be successful.</i>
OC3	<i>I have a sense of loyalty to HBAT.</i>
OC4	<i>I am proud to tell others that I work for HBAT.</i>
EP1	<i>I am comfortable with my physical work environment at HBAT.</i>
EP2	<i>The place I work in is designed to help me do my job better.</i>
EP3	<i>There are few obstacles to make me less productive in my workplace.</i>
EP4	<i>What term best describe your work environment at HBAT ? Too hectic _____ Very soothing</i>
AC1	<i>How happy are you with the work of your coworkers ?</i>
AC2	<i>How do you feel about your coworkers ?</i>
AC3	<i>How often do you do things with your coworkers on your days off ?</i>
AC4	<i>Generally, how similar are your coworkers to you ?</i>

- | | |
|-----|--|
| SI1 | <i>I am not actively searching for another job.</i> |
| SI2 | <i>I seldom look at the job listings on monster.com.</i> |
| SI3 | <i>I have no interest in searching for a job in the next year.</i> |
| SI4 | <i>How likely is it that you will be working at HBAT one year from today ?</i> |

Nous voulons tester le modèle de mesure avec l'analyse factorielle confirmatoire ; s'il semble bon, nous pourrons alors préciser le modèle structurel. La figure 21.13 présente ce modèle. Les construits étant réflexifs, on voit que ce sont eux qui pointent vers les variables indicatrices. Aussi, puisque nous nous intéressons pour l'instant au modèle de mesure, aucun lien de causalité n'est présent ; on a plutôt toutes les corrélations possibles entre les construits. Ceux-ci sont tous des variables latentes exogènes ; certaines deviendront endogènes lorsque le modèle structurel sera précisé.

On voit que le minimum de trois items par construit est respecté : chaque variable latente a quatre ou cinq items, il devrait donc être possible d'identifier une solution.

On a recueilli un échantillon ayant une taille de 400, avec seulement deux réponses comportant une valeur manquante. On utilise ici l'échantillon de taille 398 sans données manquantes ; cette taille est suffisante pour utiliser le maximum de vraisemblance. On aurait aussi pu imputer les valeurs manquantes, ce qui est la procédure habituelle lorsque seulement quelques réponses manquent pour un même individu. Il est important de ne pas avoir de données manquantes pour pouvoir obtenir toutes les mesures d'ajustement avec AMOS.

On peut maintenant procéder à l'évaluation de l'ajustement du modèle. Fixons les seuils à $\alpha = 0,05$.

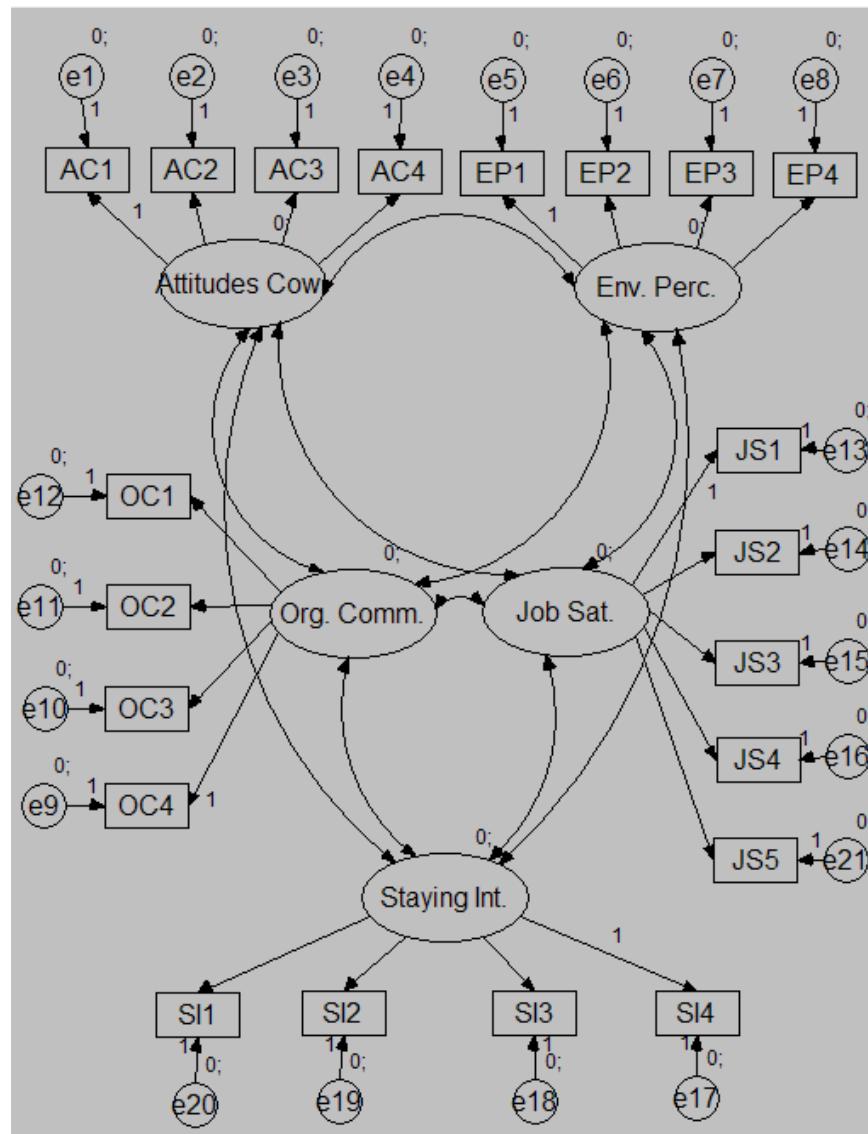


FIG. 21.13 – Le modèle de mesure

La sortie 21.14 présente le modèle avec les estimations. Par exemple, on voit que les charges factorielles (*loadings*) pour le construit OC sont 0,58, 0,89, 0,66 et 0,84 respectivement. Ces mêmes charges se retrouvent dans la sortie 21.20. On voit aussi, par exemple, que l'item OC4 est expliqué à 70 % par le construit OC. La plus basse charge a une valeur de 0,58 ; c'est en bas de la barre du 0,7, qui est le seuil que l'on espère respecter car sinon l'indicateur correspondant est expliqué à moins de 50 % par le concept. Ici l'item correspondant à la charge de 0,58 n'est expliqué qu'à $0,58^2 = 34\%$ par le construit OC. Si le modèle présente des faiblesses, il se peut que ce soit relié à cet item.

On voit aussi les corrélations entre les construits. Par exemple, les construits AC et EP ont une corrélation de 0,25.

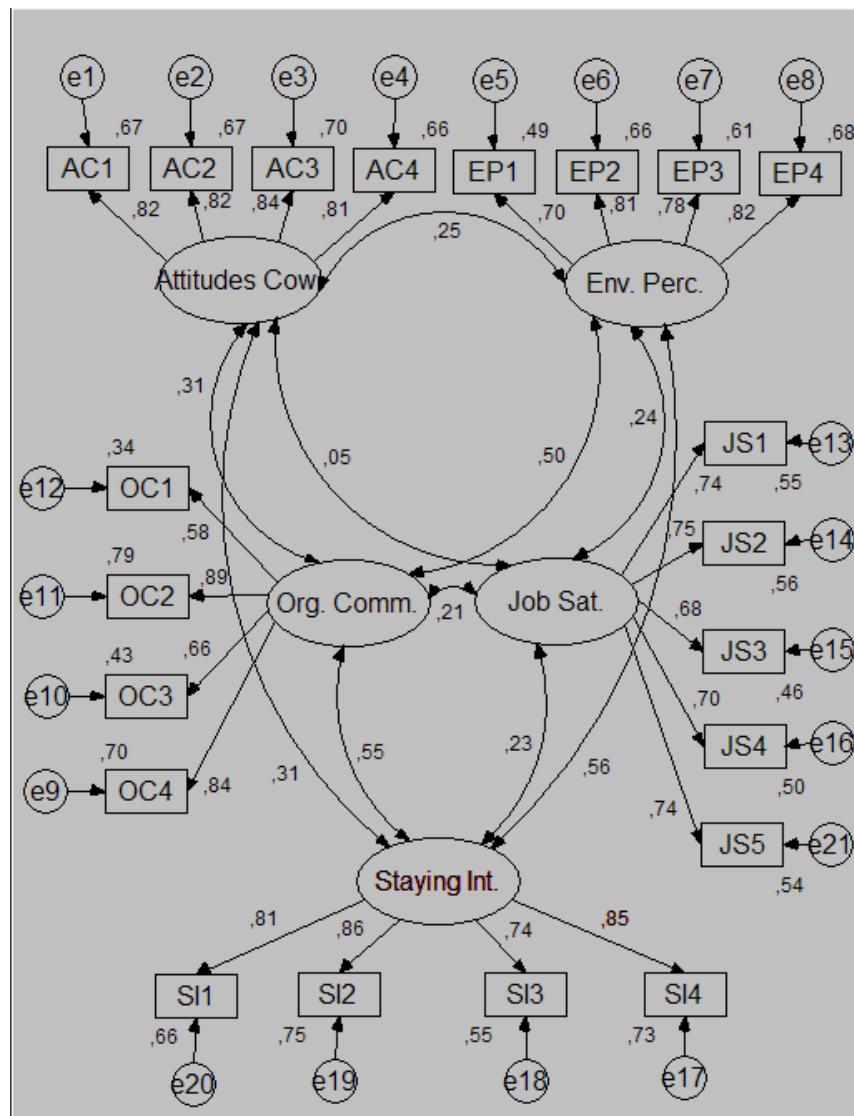


FIG. 21.14 – Le modèle de mesure avec les estimations standardisées

La figure 21.15 nous montre qu'il y a 52 paramètres à estimer. Comme la matrice des covariances contient 231 éléments distincts, il y a $231 - 52 = 179$ degrés de liberté. Le modèle est donc sur-identifié (condition pour pouvoir identifier une solution), et puisque la taille d'échantillon semble suffisante, le modèle devrait produire des résultats fiables.

Notes for Model (Default model)

Computation of degrees of freedom (Default model)

Number of distinct sample moments: 231

Number of distinct parameters to be estimated: 52

Degrees of freedom (231 - 52): 179

Result (Default model)

Minimum was achieved

Chi-square = 229,689

Degrees of freedom = 179

Probability level = .006

FIG. 21.15 – Nombre de paramètres à estimer

La figure 21.16 nous montre la première partie des mesures d'ajustement. Le χ^2 a une valeur de 229,689, et sa *p*-value est de $0,006 < 0,05$. Ainsi le bon ajustement du modèle aux données est rejeté (la matrice observée des covariances est jugée différente de la matrice estimée des covariances). Cependant, cette statistique rejette facilement l'adéquation du modèle, et ce d'autant plus lorsque la taille d'échantillon est grande, ce qui est le cas ici. On poursuit donc l'analyse avec les autres mesures d'ajustement.

Une règle du pouce dit d'examiner environ deux indices absolus et deux indices incrémentaux, qu'il est habituellement préférable de choisir à l'avance. Pour les indices absolus, il y a entre autres le GFI, AGFI, RMR (figure 21.16) et le RMSEA (figure 21.17). Ici, on voit que le GFI est de 0,949, ce qui est bien puisque l'on admet une bonne adéquation lorsqu'il dépasse 0,9. Le RMSEA, que l'on espère le plus petit possible, si possible plus petit que 0,05, a une valeur de 0,027. Cette mesure vient donc elle aussi appuyer l'adéquation du modèle. De plus, si on regarde l'intervalle de confiance de niveau 90 % qui

Model Fit Summary**CMIN**

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	52	229,689	179	,006	1,283
Saturated model	231	,000	0		
Independence model	21	4439,239	210	,000	21,139

RMR, GFI

Model	RMR	GFI	AGFI	PGFI
Default model	,391	,949	,934	,735
Saturated model	,000	1,000		
Independence model	2,365	,341	,276	,310

Baseline Comparisons

Model	NFI	RFI	IFI	TLI	CFI
	Delta1	rho1	Delta2	rho2	
Default model	,948	,939	,988	,986	,988
Saturated model	1,000		1,000		1,000
Independence model	,000	,000	,000	,000	,000

Parsimony-Adjusted Measures

Model	PRATIO	PNFI	PCFI
Default model	,852	,808	,842
Saturated model	,000	,000	,000
Independence model	1,000	,000	,000

FIG. 21.16 – Les mesures d'ajustement

lui est associé ([0,015, 0,036]), on voit que même la borne supérieure est en-dessous de 0,05. D'ailleurs, la valeur de la PCLOSE est de 1, ce qui fait qu'on ne rejette surtout pas l'hypothèse selon laquelle le RMSEA est plus petit ou égal à 0,05 (on rejetterait cette hypothèse si la PCLOSE était plus petite que 0,05, notre seuil de signification).

Pour ce qui est des indices incrémentaux, il est préférable d'en prendre un de type II et un autre de type III. Le TLI est un indice de type II (figure 21.16) ; il a ici une valeur de 0,986, ce qui dépasse la valeur clé de 0,9. D'autre part, le CFI est un indice de type III ; il a ici une valeur de 0,988, ce qui dépasse aussi le 0,9. Donc les deux indices incrémentaux viennent eux aussi appuyer l'adéquation du modèle.

NCP

Model	NCP	LO 90	HI 90
Default model	50,689	15,760	93,749
Saturated model	,000	,000	,000
Independence model	4229,239	4016,211	4449,537

FMIN

Model	FMIN	F0	LO 90	HI 90
Default model	,579	,128	,040	,236
Saturated model	,000	,000	,000	,000
Independence model	11,182	10,653	10,116	11,208

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	,027	,015	,036	1,000
Independence model	,225	,219	,231	,000

AIC

Model	AIC	BCC	BIC	CAIC
Default model	333,689	339,791	540,985	592,985
Saturated model	462,000	489,104	1382,870	1613,870
Independence model	4481,239	4483,703	4564,954	4585,954

FIG. 21.17 – Les mesures d'ajustement (suite)

Finalement, au niveau de la parcimonie, le χ^2 normé (CMIN/DF) a une valeur de 1,283, ce qui est très bon puisque compris entre 1 et 2.

ECVI

Model	ECVI	LO 90	HI 90	MECVI
Default model	,841	,753	,949	,856
Saturated model	1,164	1,164	1,164	1,232
Independence model	11,288	10,751	11,843	11,294

HOELTER

Model	HOELTER	
	.05	.01
Default model	366	391
Independence model	22	24

FIG. 21.18 – Dernières sorties des mesures d'ajustement

Regression Weights: (Group number 1 - Default model)

	Estimate	S.E.	C.R.	P	Label
AC1 <--- Attitudes Cow.	1,000				
AC2 <--- Attitudes Cow.	1,238	,068	18,283	***	par_1
AC3 <--- Attitudes Cow.	1,036	,055	18,745	***	par_2
AC4 <--- Attitudes Cow.	1,145	,063	18,121	***	par_3
EP1 <--- Env. Perc.	1,000				
EP2 <--- Env. Perc.	1,035	,072	14,374	***	par_4
EP3 <--- Env. Perc.	,805	,058	13,887	***	par_5
EP4 <--- Env. Perc.	,897	,062	14,510	***	par_6
OC4 <--- Org. Comm.	1,000				
OC3 <--- Org. Comm.	,670	,048	13,884	***	par_7
OC2 <--- Org. Comm.	1,124	,058	19,344	***	par_8
OC1 <--- Org. Comm.	,852	,071	11,945	***	par_9
JS1 <--- Job Sat.	1,000				
JS2 <--- Job Sat.	1,031	,075	13,704	***	par_10
JS3 <--- Job Sat.	,900	,072	12,520	***	par_11
JS4 <--- Job Sat.	,908	,070	12,965	***	par_12
SI4 <--- Staying Int.	1,000				
SI3 <--- Staying Int.	,912	,054	16,817	***	par_13
SI2 <--- Staying Int.	,919	,044	20,967	***	par_14
SI1 <--- Staying Int.	,857	,045	19,179	***	par_15
JS5 <--- Job Sat.	15,205	1,125	13,514	***	par_16

FIG. 21.19 – Les estimations non standardisées

La figure 21.19 nous montre les charges factorielles non standardisées. Puisque les CR (*Critical Ratio*, un écart-réduit) sont tous supérieurs (et de loin) à 1,96, toutes les charges factorielles sont significativement différentes de 0. La figure 21.20 nous montre les charges factorielles standardisées, ce sont les mêmes que celles qu'on a vu sur le modèle (figure 21.14).

Standardized Regression Weights: (Group number 1 - Default model)

	Estimate
AC1 <--- Attitudes Cow.	,821
AC2 <--- Attitudes Cow.	,820
AC3 <--- Attitudes Cow.	,837
AC4 <--- Attitudes Cow.	,814
EP1 <--- Env. Perc.	,699
EP2 <--- Env. Perc.	,813
EP3 <--- Env. Perc.	,779
EP4 <--- Env. Perc.	,823
OC4 <--- Org. Comm.	,840
OC3 <--- Org. Comm.	,658
OC2 <--- Org. Comm.	,888
OC1 <--- Org. Comm.	,582
JS1 <--- Job Sat.	,743
JS2 <--- Job Sat.	,748
JS3 <--- Job Sat.	,680
JS4 <--- Job Sat.	,705
SI4 <--- Staying Int.	,852
SI3 <--- Staying Int.	,741
SI2 <--- Staying Int.	,864
SI1 <--- Staying Int.	,811
JS5 <--- Job Sat.	,736

FIG. 21.20 – Les charges factorielles standardisées

L'ajustement du modèle étant vérifié, on peut procéder à l'analyse des validités convergente, discriminante et nomologique des construits.

Validité convergente Les variables indicatrices d'un même construit devraient converger, c'est-à-dire se comporter d'une façon semblable et donc partager une grande proportion de variance commune. On a déjà une bonne idée de la validité convergente des construits en regardant les charges factorielles qui leur sont associées. En effet, si celles-ci sont toutes de 0,7 ou plus, alors chaque variable indicatrice a sa variation expliquée à au moins 50 % par le concept du construit, ce qui assure une certaine convergence. Dans l'exemple, on voit que deux construits ont certaines de leurs charges factorielles en bas de 0,7 : *Organizational Commitment* et *Job Satisfaction*. Par contre le construit de *Job Satisfaction* ne génère pas beaucoup d'inquiétude pour la convergence car une seule charge factorielle est en-dessous du 0,7, et elle en est très près (0,68). Le construit du *Organizational Commitment* est un peu plus inquiétant, il a des charges factorielles de 0,58 et 0,66.

Une façon de mesurer la convergence est de calculer le pourcentage moyen de **variance extraite (VE)** pour chacun des construits. Cette mesure se calcule de la façon suivante :

$$\text{VE} = \frac{\sum_{i=1}^k \lambda_i^2}{k}$$

où les λ_i sont les charges factorielles et k est le nombre d'items (indicateurs) dans le construit. Puisque λ_i^2 représente le pourcentage de la variation de la variable indicatrice qui est expliquée par le concept, le VE est simplement une moyenne de ces pourcentages. Une règle du pouce suggère que le VE soit d'au moins 0,5, sinon on peut dire qu'en moyenne l'erreur est plus grande que ce qui est expliqué par le concept.

Pour les calculs on utilise les valeurs des charges factorielles que l'on retrouve dans la figure 21.20. Par exemple, pour le construit de *Job Satisfaction*, on obtient

$$\text{VE}_{\text{Job Sat.}} = \frac{0,743^2 + 0,748^2 + 0,68^2 + 0,705^2 + 0,736^2}{5} = 0,5225.$$

Voici les VE pour les cinq construits :

Concept	VE
<i>Job Satisfaction</i>	0,5225
<i>Organizational Commitment</i>	0,5665
<i>Staying Intentions</i>	0,6698
<i>Environmental Perceptions</i>	0,6084
<i>Attitudes Toward Coworkers</i>	0,6774

Les VE sont tous au-dessus de 0,5, ce qui appuie la validité convergente de ces construits.

Une autre façon de mesurer la validité de convergence est d'analyser la cohérence interne des construits à l'aide du rho de Joreskog. Cette mesure est préférable à l'alpha de Cronbach puisqu'elle tient compte des charges factorielles de chacune des variables indicatrices dans son calcul. Par contre elle se laisse influencer par le nombre de variables indicatrices ; en effet, tout comme l'alpha de Cronbach, le rho de Joreskog tend à devenir plus grand lorsqu'il y a un plus grand nombre de variables indicatrices pour un construit.

Voici la formule du rho de Joreskog :

$$\rho_{\text{Joreskog}} = \frac{\left(\sum_{i=1}^k \lambda_i \right)^2}{\left(\sum_{i=1}^k \lambda_i \right)^2 + \sum_{i=1}^k (1 - \lambda_i^2)}$$

où les λ_i sont les charges factorielles et k est le nombre d'items (variables indicatrices) dans le construit.

La pratique courante est d'interpréter ce coefficient avec la même charte que celle de l'alpha de Cronbach (figure 21.21).

On calcule donc le rho de Joreskog de chacun des construits, en utilisant les valeurs des charges factorielles que l'on retrouve dans la figure 21.20. Par exemple, pour le construit

$0,6 \leq \rho_{\text{Joreskog}} \leq 0,7$	médiocre
$0,7 < \rho_{\text{Joreskog}} \leq 0,8$	moyen
$0,8 < \rho_{\text{Joreskog}} \leq 0,9$	très bien
$0,9 < \rho_{\text{Joreskog}} \leq 1$	excellent

FIG. 21.21 – Interprétation du rho de Joreskog

de *Job Satisfaction*, le calcul est le suivant :

$$\rho_{\text{Job Sta.}} = \frac{(0,743+0,748+0,68+0,705+0,736)^2}{(0,743+0,748+0,68+0,705+0,736)^2 + (1-0,743^2) + (1-0,748^2) + (1-0,68^2) + (1-0,705^2) + (1-0,736^2)}$$

$$= 0,8453.$$

On retrouve les résultats pour chacun des construits dans le tableau suivant :

Concept	rho de Joreskog
<i>Job Satisfaction</i>	0,8453
<i>Organizational Commitment</i>	0,8355
<i>Staying Intentions</i>	0,8899
<i>Environmental Perceptions</i>	0,8609
<i>Attitudes Toward Coworkers</i>	0,8933

Le construit qui nous inquiétait le plus, soit celui du *Organizational Commitment*, est celui qui a le rho de Joreskog le plus bas, mais est quand même très bien puisqu'il a une valeur de 0,8355. Donc nos échelles de mesure semblent avoir une bonne cohérence interne, ce qui vient appuyer la convergence. Il est conseillé de présenter aussi les valeurs des alpha de Cronbach, car cette mesure est encore très présente dans la littérature ; je vous laisse le faire pour cet exemple ;-). On peut donc passer à la validité discriminante.

Validité discriminante Les différents construits utilisés sont supposés mesurer des concepts distincts, et c'est ce dont on veut s'assurer lorsqu'on tente de mesurer la validité discriminante. De bons construits devraient partager plus d'information avec leurs variables indicatrices qu'entre eux. Pour évaluer si deux construits sont assez distincts entre eux, on compare le r^2 mesuré entre ces deux construits avec les VE de chacun ; on considérera que la discrimination est valide si

$$r^2_{\text{construit1, construit2}} < \min\{VE_{\text{construit1}}, VE_{\text{construit2}}\}.$$

Correlations: (Group number 1 - Default model)

		Estimate
Attitudes Cow. <-->	Env. Perc.	,253
Attitudes Cow. <-->	Org. Comm.	,305
Env. Perc. <-->	Job Sat.	,241
Attitudes Cow. <-->	Job Sat.	,050
Env. Perc. <-->	Org. Comm.	,495
Org. Comm. <-->	Job Sat.	,209
Attitudes Cow. <-->	Staying Int.	,308
Env. Perc. <-->	Staying Int.	,562
Org. Comm. <-->	Staying Int.	,552
Job Sat. <-->	Staying Int.	,230

FIG. 21.22 – Les corrélations entre les variables latentes

On retrouve justement dans la figure 21.22 les corrélations entre les construits (elle fait partie des sorties générées par AMOS), et on a déjà calculé les VE. Les résultats sont résumés dans le tableau 21.23.

En fait, tous les VE étant supérieurs aux r^2 , on voit rapidement qu'ici la discrimination est validée.

Concepts en relation	r^2	$<$, $>$, $=$	minimum des deux VE
<i>Job Satisfaction</i> et <i>Organizational Commitment</i>	$0,209^2 = 0,04$	$<$	0,5225
<i>Job Satisfaction</i> et <i>Staying Intentions</i>	$0,23^2 = 0,05$	$<$	0,5225
<i>Job Satisfaction</i> et <i>Environmental Perceptions</i>	$0,241^2 = 0,06$	$<$	0,5225
<i>Job Satisfaction</i> et <i>Attitudes Toward Coworkers</i>	$0,05^2 = 0,003$	$<$	0,5225
<i>Organizational Commitment</i> et <i>Staying Intentions</i>	$0,552^2 = 0,30$	$<$	0,5665
<i>Organizational Commitment</i> et <i>Environmental Perceptions</i>	$0,495^2 = 0,25$	$<$	0,5665
<i>Organizational Commitment</i> et <i>Attitudes Toward Coworkers</i>	$0,305^2 = 0,09$	$<$	0,5665
<i>Environmental Perceptions</i> et <i>Staying Intentions</i>	$0,562^2 = 0,32$	$<$	0,6084
<i>Environmental Perceptions</i> et <i>Attitudes Toward Coworkers</i>	$0,253^2 = 0,06$	$<$	0,6084
<i>Staying Intentions</i> et <i>Attitudes Toward Coworkers</i>	$0,308^2 = 0,09$	$<$	0,6698

FIG. 21.23 – Comparaison entre les r^2 et les VE

Validité nomologique On veut ici vérifier que les relations entre les construits sont celles auxquelles on s'attend ; ici on s'attend à une relation positive entre chaque construit. La figure 21.22 confirme ceci, car toutes les corrélations sont positives. Le seul hic, c'est que la corrélation entre les concepts de *Attitudes Toward Coworkers* et *Job Satisfaction* semble très faible, elle n'est que de 0,05. Et effectivement, si on regarde la figure 21.24 qui contient les covariances, on voit que la covariance entre ces deux concepts est jugée nulle (p -value de 0,39 et C.R. = $0,859 < 1,96$). Cet accroc est mineur étant donné que toutes les autres corrélations appuient nos hypothèses. On considère donc que la validité

nomologique est appuyée.

Covariances: (Group number 1 - Default model)

		Estimate	S.E.	C.R.	P	Label
Attitudes Cow.	<--> Env. Perc.	,370	,088	4,178	***	par_17
Attitudes Cow.	<--> Org. Comm.	,602	,119	5,063	***	par_18
Env. Perc.	<--> Job Sat.	,306	,079	3,877	***	par_19
Attitudes Cow.	<--> Job Sat.	,056	,066	,859	,390	par_20
Env. Perc.	<--> Org. Comm.	1,093	,154	7,120	***	par_21
Org. Comm.	<--> Job Sat.	,358	,103	3,468	***	par_22
Attitudes Cow.	<--> Staying Int.	,291	,056	5,160	***	par_23
Env. Perc.	<--> Staying Int.	,594	,076	7,830	***	par_24
Org. Comm.	<--> Staying Int.	,787	,096	8,214	***	par_25
Job Sat.	<--> Staying Int.	,189	,049	3,833	***	par_26

FIG. 21.24 – Les covariances entre les variables latentes

On peut donc conclure que ce modèle de mesure se comporte bien. On peut donc passer à l'étape suivante, qui consiste à spécifier le modèle structurel, c'est-à-dire les liens qui sont supposés unir nos différents concepts.

21.7 Le modèle structurel

Lorsqu'on teste le modèle de mesure, aucun chemin causal n'est spécifié dans le modèle. Or le but de l'étude étant de comprendre ce qui favorise la rétention (Staying Intentions), on passe maintenant au modèle structurel en spécifiant de quelle façon les quatre autres construits sont supposés influencer la rétention. Suite à une revue de la littérature et à un raffinement (il aurait été possible de considérer plus de cinq construits, mais pour toutes sortes de raisons pratiques on s'est contenté de cinq), les hypothèses retenues sont les suivantes :

-
- H_1 : *Environmental Perceptions* influence positivement la *Job Satisfaction*.
- H_2 : *Environmental Perceptions* influence positivement le *Organizational Commitment*.
- H_3 : *Attitudes Toward Coworkers* influence positivement la *Job Satisfaction*.
- H_4 : *Attitudes Toward Coworkers* influence positivement le *Organizational Commitment*.
- H_5 : *Job Satisfaction* influence positivement le *Organizational Commitment*.
- H_6 : *Job Satisfaction* influence positivement les *Staying Intentions*.
- H_7 : *Organizational Commitment* influence positivement les *Staying Intentions*.
-

FIG. 21.25 – Les hypothèses du modèle

Il est possible d'exprimer visuellement ces hypothèses, c'est ce qu'on retrouve dans le modèle structurel de la figure 21.26. Les estimations sont déjà présentées dans cette figure.

On peut voir que l'on a deux variables exogènes dans ce modèle, soit les construits *Attitudes Toward Coworkers* et *Environmental Perceptions*. C'est d'ailleurs pourquoi une flèche incurvée apparaît entre ces deux construits, car on mesure toujours les corrélations entre les variables exogènes. Aussi, le fait que ces construits soient des variables exogènes signifie que ces variables latentes ne sont prédites par aucun autre concept dans ce modèle.

Les trois autres construits sont des variables endogènes car des chemins causals pointent vers eux. On voit aussi que ces construits ont maintenant chacun une erreur qui leur est associée ; en effet, maintenant que l'on tente d'expliquer ces construits, il faut mesurer l'erreur, c'est-à-dire ce qui n'est pas expliqué par les hypothèses (chemins).

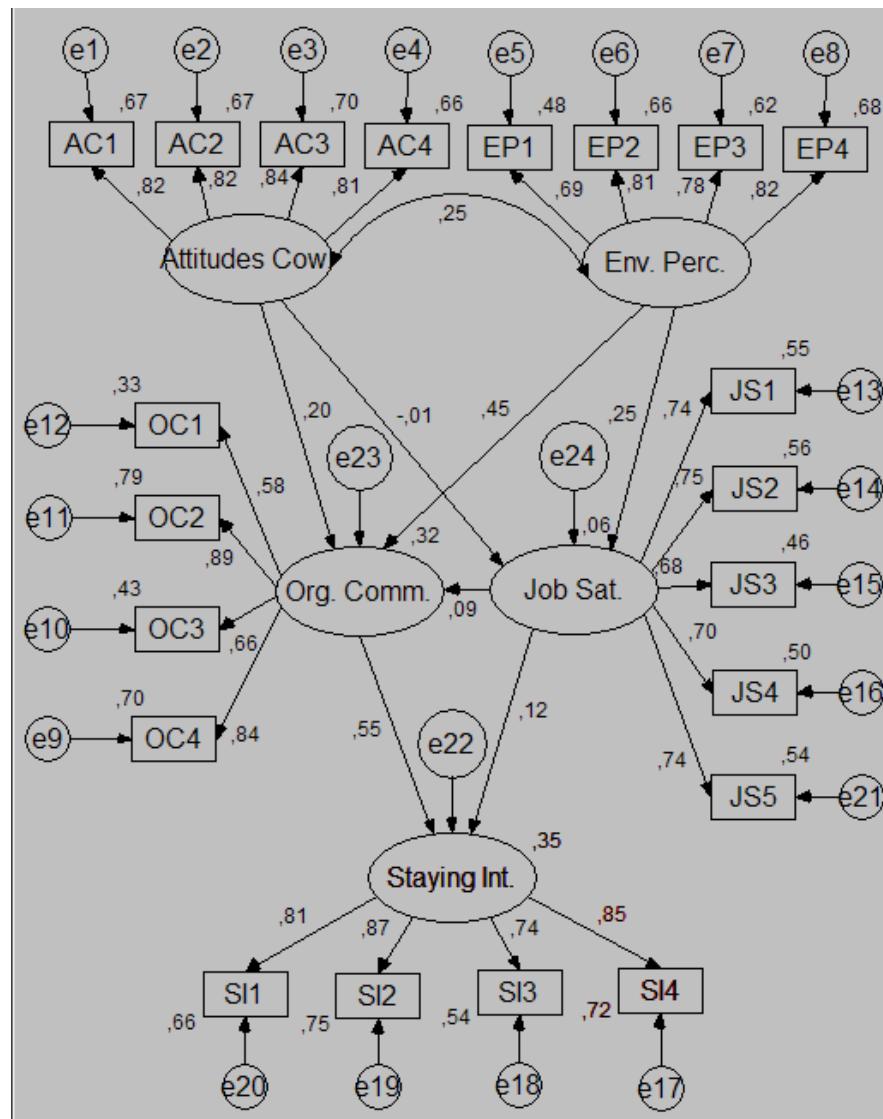


FIG. 21.26 – Le modèle structurel

Notes for Model (Default model)**Computation of degrees of freedom (Default model)**

Number of distinct sample moments:	231
Number of distinct parameters to be estimated:	50
Degrees of freedom (231 - 50):	181

Result (Default model)

Minimum was achieved
Chi-square = 276,369
Degrees of freedom = 181
Probability level = ,000

FIG. 21.27 – Nombre de paramètres à estimer

La figure 21.27 nous montre qu'il y a maintenant 181 degrés de liberté ; en effet, il y a moins de paramètres à estimer que dans le modèle de mesure, c'est pourquoi on est passé de 179 à 181 (deux flèches de moins entre les construits). Par contre le χ^2 a augmenté, et ce sera toujours le cas lorsqu'on passe du modèle de mesure à un modèle structurel **récursif** (dans un modèle récursif il n'est pas possible de partir d'un concept et d'y revenir en suivant les chemins), car il y a moins de paramètres à estimer. Ceci entraîne qu'il y a plus de paramètres fixés à 0, ce qui fait que la matrice des covariances estimée sera moins proche de la matrice des covariances observée que dans le cas du modèle de mesure. Il est à noter que si le modèle n'est pas récursif, ceci peut causer certains problèmes d'identification.

Les figures 21.28 et 21.29 nous montrent les estimations des paramètres, d'abord non standardisées, puis standardisées comme celles qui apparaissent sur le graphe (figure 21.26). Dans la figure 21.28, on retrouve d'abord les mesures des sept chemins causals qui correspondent à nos hypothèses. En regardant les C.R. et les *p*-values (colonne P), on voit que deux de nos hypothèses ne sont pas supportées : le lien entre A.C. et J.S. et entre J.S. et O.C. ne sont pas significatifs au seuil de 5 %. Au moins, le lien entre J.S. et O.C. est positif, donc il respecte le sens de l'hypothèse lui correspondant. Notre modèle semble donc partiellement supporté.

Regression Weights: (Group number 1 - Default model)

		Estimate	S.E.	C.R.	P	Label
Job Sat.	<--- Env. Perc.	,196	,049	4,024	***	par_18
Job Sat.	<--- Attitudes Cow.	-,009	,051	-,167	,867	par_20
Org. Comm.	<--- Attitudes Cow.	,301	,078	3,867	***	par_17
Org. Comm.	<--- Env. Perc.	,608	,082	7,384	***	par_19
Org. Comm.	<--- Job Sat.	,149	,091	1,629	,103	par_24
Staying Int.	<--- Org. Comm.	,264	,026	10,017	***	par_21
Staying Int.	<--- Job Sat.	,100	,042	2,381	,017	par_22
AC1	<--- Attitudes Cow.	1,000				
AC2	<--- Attitudes Cow.	1,238	,068	18,276	***	par_1
AC3	<--- Attitudes Cow.	1,036	,055	18,721	***	par_2
AC4	<--- Attitudes Cow.	1,146	,063	18,126	***	par_3
EP1	<--- Env. Perc.	1,000				
EP2	<--- Env. Perc.	1,043	,074	14,165	***	par_4
EP3	<--- Env. Perc.	,818	,059	13,794	***	par_5
EP4	<--- Env. Perc.	,905	,063	14,311	***	par_6
OC4	<--- Org. Comm.	1,000				
OC3	<--- Org. Comm.	,672	,049	13,847	***	par_7
OC2	<--- Org. Comm.	1,127	,058	19,518	***	par_8
OC1	<--- Org. Comm.	,847	,072	11,786	***	par_9
JS1	<--- Job Sat.	1,000				
JS2	<--- Job Sat.	1,034	,076	13,685	***	par_10
JS3	<--- Job Sat.	,902	,072	12,503	***	par_11
JS4	<--- Job Sat.	,910	,070	12,942	***	par_12
SI4	<--- Staying Int.	1,000				
SI3	<--- Staying Int.	,913	,055	16,605	***	par_13
SI2	<--- Staying Int.	,929	,045	20,845	***	par_14
SI1	<--- Staying Int.	,863	,045	19,066	***	par_15
JS5	<--- Job Sat.	15,246	1,130	13,495	***	par_23

FIG. 21.28 – Les estimations non standardisées

Pour ce qui est des charges factorielles non standardisées (restant de la figure 21.28), elles sont restées significatives (par rapport aux estimations du modèle de mesure). Mais examinons celles-ci plus attentivement avec les mesures standardisées de la figure 21.29.

Si le modèle de mesure est effectivement bon, les charges factorielles devraient être stables et donc ne pas avoir beaucoup changé en précisant le modèle structurel. En comparant les charges factorielles de la figure 21.29 avec celles de la figure 21.20, on voit qu'il y a très peu de changements, ce qui nous rend d'autant plus confiants envers le modèle de mesure.

Standardized Regression Weights: (Group number 1 - Default model)

		Estimate
Job Sat.	<--- Env. Perc.	,250
Job Sat.	<--- Attitudes Cow.	-,010
Org. Comm.	<--- Attitudes Cow.	,201
Org. Comm.	<--- Env. Perc.	,449
Org. Comm.	<--- Job Sat.	,086
Staying Int.	<--- Org. Comm.	,552
Staying Int.	<--- Job Sat.	,121
AC1	<--- Attitudes Cow.	,821
AC2	<--- Attitudes Cow.	,820
AC3	<--- Attitudes Cow.	,836
AC4	<--- Attitudes Cow.	,815
EP1	<--- Env. Perc.	,693
EP2	<--- Env. Perc.	,812
EP3	<--- Env. Perc.	,785
EP4	<--- Env. Perc.	,823
OC4	<--- Org. Comm.	,836
OC3	<--- Org. Comm.	,657
OC2	<--- Org. Comm.	,886
OC1	<--- Org. Comm.	,576
JS1	<--- Job Sat.	,741
JS2	<--- Job Sat.	,748
JS3	<--- Job Sat.	,680
JS4	<--- Job Sat.	,705
SI4	<--- Staying Int.	,848
SI3	<--- Staying Int.	,738
SI2	<--- Staying Int.	,869
SI1	<--- Staying Int.	,813
JS5	<--- Job Sat.	,737

FIG. 21.29 – Les charges factorielles standardisées

CMIN

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	50	276,369	181	,000	1,527
Saturated model	231	,000	0		
Independence model	21	4439,239	210	,000	21,139

RMR, GFI

Model	RMR	GFI	AGFI	PGFI
Default model	,388	,940	,923	,736
Saturated model	,000	1,000		
Independence model	2,365	,341	,276	,310

Baseline Comparisons

Model	NFI	RFI	IFI	TLI	CFI
	Delta1	rho1	Delta2	rho2	
Default model	,938	,928	,978	,974	,977
Saturated model	1,000		1,000		1,000
Independence model	,000	,000	,000	,000	,000

Parsimony-Adjusted Measures

Model	PRATIO	PNFI	PCFI
Default model	,862	,808	,842
Saturated model	,000	,000	,000
Independence model	1,000	,000	,000

FIG. 21.30 – Les mesures d'ajustement

On peut maintenant regarder l'adéquation du modèle. Tel que dit précédemment, le χ^2 a augmenté, il a maintenant une valeur de 276,369. La *p*-value étant nulle, l'adéquation du modèle est rejetée. Rappelons cependant qu'avec une taille d'échantillon de 398 ceci n'est pas étonnant, et on poursuit donc avec les autres mesures d'ajustement.

Commençons par les indices absolus ; on voit que le GFI a une valeur de 0,94, et le AGFI a une mesure de 0,923, ce qui dans les deux cas est au-dessus du seuil de 0,9, donc à première vue l'ajustement semble bon. De même, le RMSEA a une valeur de 0,036, ce qui est en-dessous de 0,05, ce qui est très bon. Même la borne supérieure de son intervalle de confiance est en-dessous de ce seuil (0,045), et la *p*-value de 0,997 nous indique de ne surtout pas rejeter l'hypothèse selon laquelle RMSEA \leq 0,05.

NCP

Model	NCP	LO 90	HI 90
Default model	95,369	54,455	144,242
Saturated model	,000	,000	,000
Independence model	4229,239	4016,211	4449,537

FMIN

Model	FMIN	F0	LO 90	HI 90
Default model	,696	,240	,137	,363
Saturated model	,000	,000	,000	,000
Independence model	11,182	10,653	10,116	11,208

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	,036	,028	,045	,997
Independence model	,225	,219	,231	,000

AIC

Model	AIC	BCC	BIC	CAIC
Default model	376,369	382,235	575,691	625,691
Saturated model	462,000	489,104	1382,870	1613,870
Independence model	4481,239	4483,703	4564,954	4585,954

FIG. 21.31 – Les mesures d'ajustement (suite)

ECVI

Model	ECVI	LO 90	HI 90	MECVI
Default model	,948	,845	1,071	,963
Saturated model	1,164	1,164	1,164	1,232
Independence model	11,288	10,751	11,843	11,294

HOELTER

Model	HOELTER .05	HOELTER .01
Default model	307	328
Independence model	22	24

FIG. 21.32 – Les mesures d'ajustement (fin)

Pour ce qui est des indices incrémentaux, le TLI (type I) a une valeur de 0,974, et le CFI (type II) a une valeur de 0,977, ce qui dans les deux cas est amplement satisfaisant (on dépasse le seuil de 0,9).

S'il y avait un autre modèle on pourrait comparer les indices de parcimonie. Pour l'instant on peut se contenter du χ^2 normé (CMIN/DF) ; il a une valeur de 1,527, ce qui est très bien puisqu'on espère la plus faible valeur possible entre 1 et 3.

Donc malgré le fait que nos hypothèses ne sont que partiellement supportées, l'ajustement du modèle semble très bien.

Regression Weights: (Group number 1 - Default model)

		M.I.	Par Change
Staying Int.	<--- Env. Perc.	25,156	,156
Staying Int.	<--- Attitudes Cow.	7,805	,095
SI2	<--- JS4	4,430	-,043
SI2	<--- JS3	4,753	-,043
SI4	<--- Env. Perc.	6,199	,062
SI4	<--- EP3	4,699	,049
SI4	<--- EP2	11,103	,061
JS1	<--- SI3	6,073	-,121
JS1	<--- EP1	5,593	-,064
OC1	<--- Staying Int.	7,414	-,372
OC1	<--- SI1	4,744	-,266
OC1	<--- SI2	6,297	-,304
OC1	<--- SI3	5,476	-,245
OC1	<--- SI4	8,193	-,315
OC1	<--- EP3	7,328	-,218
OC1	<--- AC1	5,030	-,172
OC2	<--- EP4	5,355	-,109
OC2	<--- EP1	4,766	-,078
OC2	<--- AC1	4,910	,105
OC3	<--- Env. Perc.	4,147	,119
OC3	<--- SI1	5,010	-,178
OC3	<--- EP4	6,417	,126
OC4	<--- AC1	5,142	-,107
EP4	<--- SI1	5,630	,126
EP3	<--- OC1	6,762	-,047
EP2	<--- SI4	8,263	,163
EP1	<--- SI2	4,749	,175

FIG. 21.33 – Les indices de modification

Même si l'ajustement semble bien, on peut examiner les *modification index*. Ceux-ci nous indiquent s'il est possible d'améliorer l'ajustement du modèle en ajoutant des chemins. On retrouve ces indices dans la figure 21.33. Le plus grand indice est sur la

première ligne, il a une valeur de 25,156 et concerne le lien entre E.P. et S.I. Ceci nous indique que le χ^2 diminuera d'au moins 25,156 si on ajoute une flèche allant de E.P. vers S.I.

Il faut faire très attention quant à l'utilisation de ces indices ; lorsqu'on mène une étude exploratoire ils peuvent être très utiles, mais dans le cas d'une analyse confirmatoire, ce n'est habituellement pas suffisant d'avoir un grand indice de modification pour pouvoir justifier la modification d'un modèle.

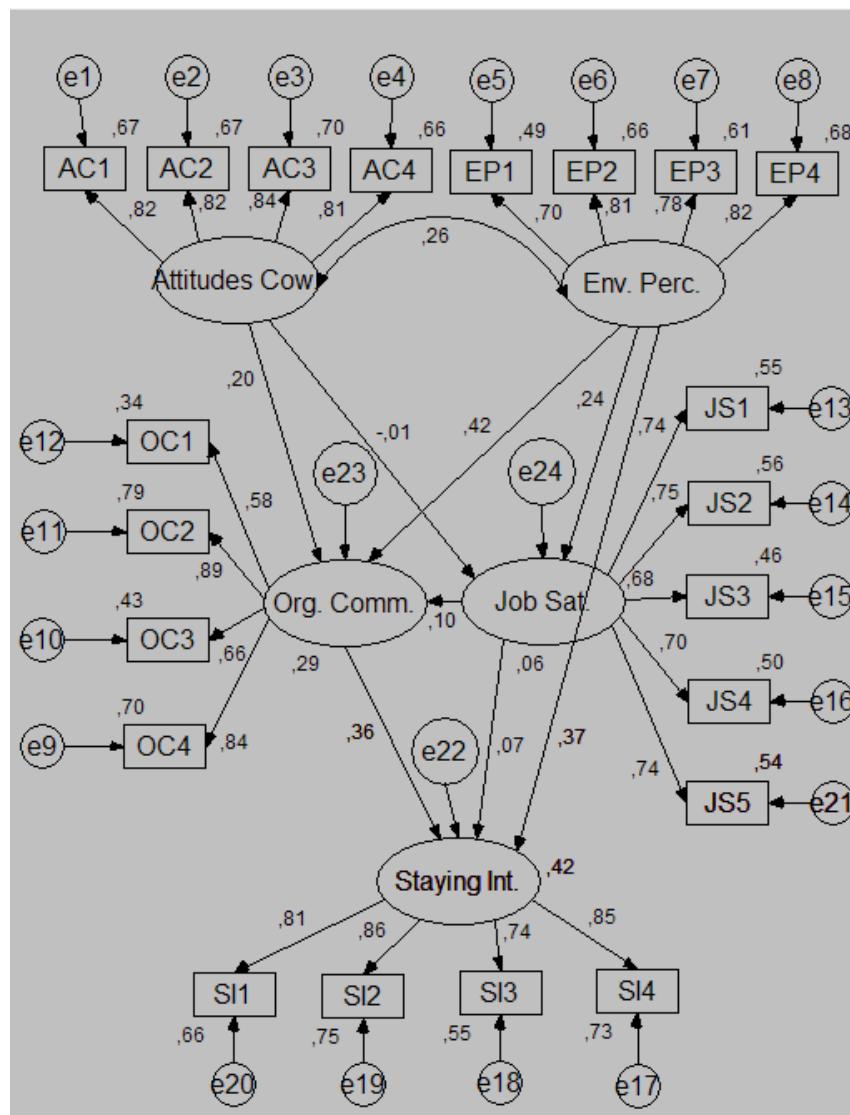


FIG. 21.34 – Le modèle modifié

Ici, nous décidons d'explorer un peu et de voir ce qui se passe si on ajoute le chemin entre E.P. et S.I. ; on obtient alors le modèle de la figure 21.34. On peut alors comparer les deux modèles. Le paramètre du nouveau chemin a une estimation de 0,37, ce qui est loin d'être négligeable. De plus le r^2 du concept S.I. est passé de 0,35 à 0,42. Et une fois de plus les charges factorielles n'ont pratiquement pas bougé, ce qui nous assure la stabilité de nos mesures.

CMIN

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	51	235,331	180	,004	1,307
Saturated model	231	,000	0		
Independence model	21	4439,239	210	,000	21,139

RMR, GFI

Model	RMR	GFI	AGFI	PGFI
Default model	,389	,947	,932	,738
Saturated model	,000	1,000		
Independence model	2,365	,341	,276	,310

Baseline Comparisons

Model	NFI	RFI	IFI	TLI	CFI
	Delta1	rho1	Delta2	rho2	
Default model	,947	,938	,987	,985	,987
Saturated model	1,000		1,000		1,000
Independence model	,000	,000	,000	,000	,000

Parsimony-Adjusted Measures

Model	PRATIO	PNFI	PCFI
Default model	,857	,812	,846
Saturated model	,000	,000	,000
Independence model	1,000	,000	,000

FIG. 21.35 – Les mesures d'ajustement

Pour ce qui est des mesures d'ajustement, on constate une amélioration de celles-ci. Même les indices de parcimonie sont meilleurs pour ce nouveau modèle. Il peut donc être intéressant de considérer celui-ci, surtout d'en comprendre l'interprétation (ce que ça nous apporte de plus dans le contexte de l'exemple).

NCP

Model	NCP	LO 90	HI 90
Default model	55,331	19,683	99,088
Saturated model	,000	,000	,000
Independence model	4229,239	4016,211	4449,537

FMIN

Model	FMIN	F0	LO 90	HI 90
Default model	,593	,139	,050	,250
Saturated model	,000	,000	,000	,000
Independence model	11,182	10,653	10,116	11,208

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	,028	,017	,037	1,000
Independence model	,225	,219	,231	,000

AIC

Model	AIC	BCC	BIC	CAIC
Default model	337,331	343,315	540,640	591,640
Saturated model	462,000	489,104	1382,870	1613,870
Independence model	4481,239	4483,703	4564,954	4585,954

FIG. 21.36 – Les mesures d'ajustement (suite)

ECVI

Model	ECVI	LO 90	HI 90	MECVI
Default model	,850	,760	,960	,865
Saturated model	1,164	1,164	1,164	1,232
Independence model	11,288	10,751	11,843	11,294

HOELTER

Model	HOELTER	HOELTER
	.05	.01
Default model	359	384
Independence model	22	24

FIG. 21.37 – Les mesures d'ajustement (fin)

Finalement, pour montrer combien il est profitable de considérer des concepts latents, la figure 21.38 présente le modèle structurel que nous avons étudié, mais maintenant chacun des concepts est considéré comme étant une variable observable (on a pris la moyenne des indicateurs). Si on compare les estimations de ce modèle à celles du modèle avec les variables latentes, on constate une nette détérioration de la force des liens. Par exemple, la variable *Staying Intentions* n'est plus expliquée qu'à 20 %, alors qu'elle l'était à 35 % dans l'autre modèle.

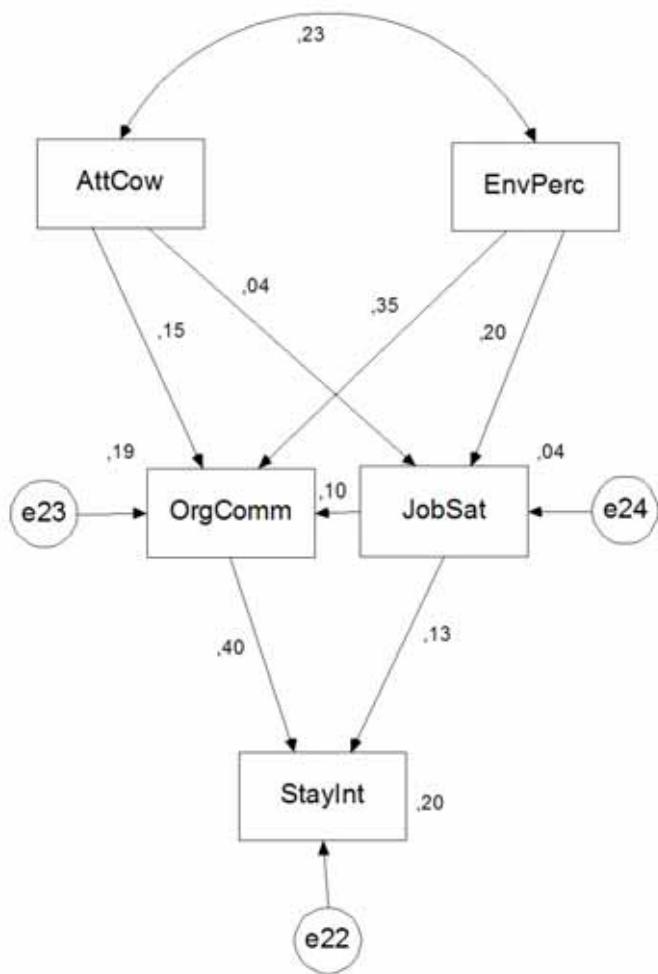


FIG. 21.38 – Le modèle sans variables latentes

21.8 Effets médiateurs et modérateurs

Deux types d'effet sont souvent étudiés en MES : les effets médiateurs et modérateurs. Cette section présente brièvement ces deux types d'effet, et propose une façon d'étudier les effets médiateurs. Nous ne présentons pas ici de méthodologie pour étudier les effets modérateurs ; nous vous référons à Roussel et Wacheux : « Management des Ressources Humaines : Méthodes de recherche en sciences humaines et sociales » (2005) et Hair *et al.* : « Multivariate Data Analysis » (2006).

21.8.1 Effet médiateur

Lors de la présentation de l'analyse des chemins à la section 21.2, les notions d'effet total, direct et indirect ont été introduites. Ces notions sont étroitement reliées à la notion d'effet médiateur. Pour illustrer ceci, prenons une variable X qui a une influence significative sur la variable Y , et supposons que la corrélation entre ces deux variables est a .

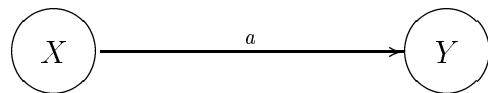


FIG. 21.39 – Influence de X sur Y

Supposons aussi qu'on a une troisième variable Z qui influence Y de façon significative, mais qui est elle aussi influencée significativement par Y ; on peut schématiser ceci comme dans la figure 21.40.

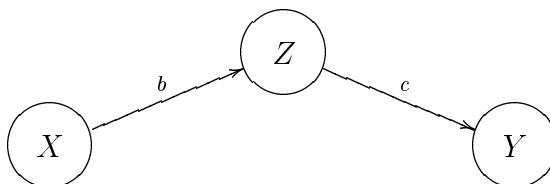
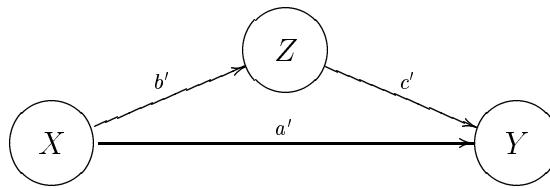


FIG. 21.40 – Ajout de la variable Z

FIG. 21.41 – Effet direct (a') et indirect ($b'c'$)

Alors l'effet total a qu'a Y sur X se décompose en effet direct (a') et indirect ($b'c'$), l'effet indirect transigeant par Z . Lorsque l'effet direct a' est nul (non-significatif), on dit alors que la médiation est **totale**. Sinon elle est dite **partielle**.

Hair *et al.* (2006) proposent de tester un effet médiateur de la façon suivante :

- Vérifier la présence de corrélations significatives entre les trois paires de variables (X avec Y , X avec Z et Z avec Y).
- Si le lien entre X et Y demeure significatif et inchangé lors de l'introduction de Z dans le modèle, alors l'hypothèse de médiation n'est pas supportée.
- Si le lien entre X et Y est réduit mais demeure significatif lors de l'introduction de Z dans le modèle, alors l'hypothèse de **médiation partielle** est supportée.
- Si le lien entre X et Y devient non significatif lors de l'introduction de Z dans le modèle, alors l'hypothèse de **médiation totale** est supportée.

Illustrons ces étapes avec un exemple.

Exemple 21.8.1 On s'intéresse ici à l'influence de la perception du soutien organisationnel (PSO) sur les comportements de citoyenneté organisationnelle (CCO). On pense que ce lien transige par l'engagement organisationnel affectif (EOA) ; on veut donc vérifier s'il y a une médiation par l'EOA, et si elle est partielle ou complète.

La figure 21.42 nous montre le lien entre la PSO et les CCO ; il est estimé à 0,25, et il est significatif (on le vérifie dans les sorties textes de Amos).

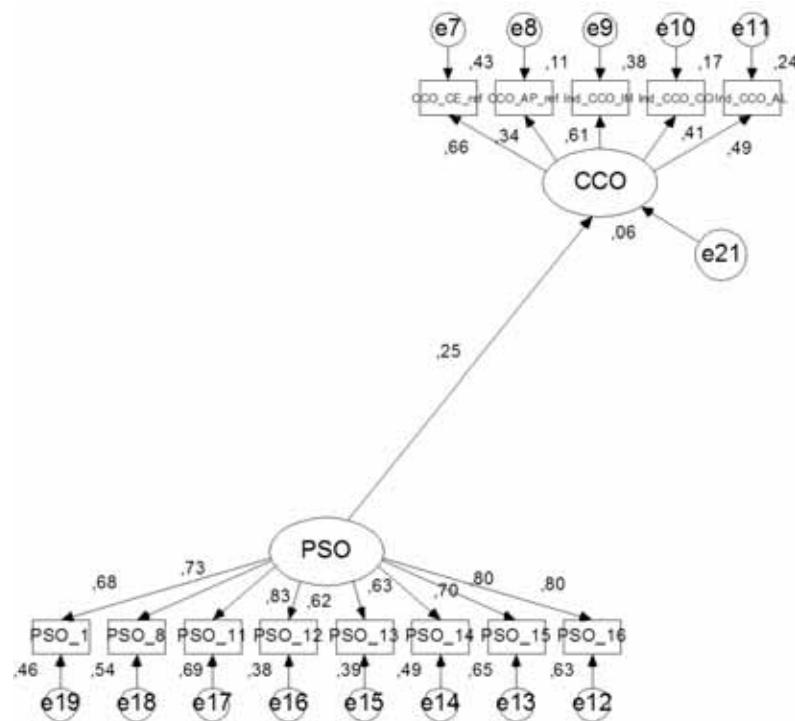


FIG. 21.42 – Lien significatif entre la PSO et les CCO

La figure 21.43 nous permet de constater qu'il y a un lien significatif de 0,72 entre la PSO et l'EOA, et de 0,42 entre l'EOA et les CCO (encore une fois on vérifie que ces liens sont significatifs dans les sorties textes de Amos).

On voit que le lien entre la PSO et les CCO est devenu non significatif à -0,04. Donc la médiation par l'EOA est complète.

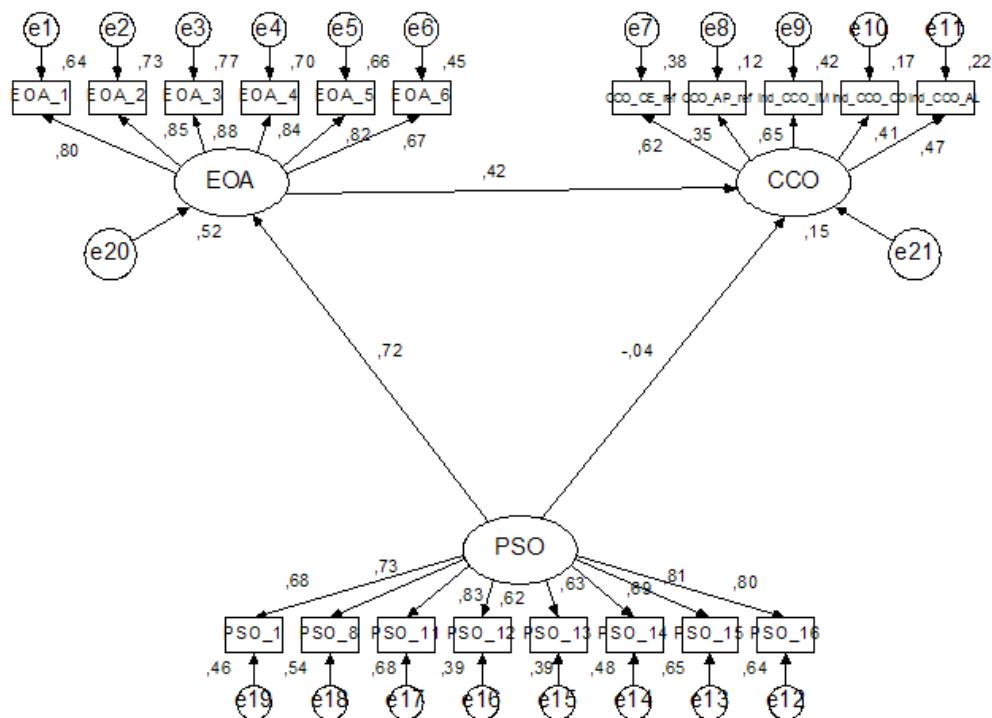


FIG. 21.43 – Introduction de l'EOA dans le modèle

21.8.2 Effet modérateur

Une **variable modératrice** est une variable qui vient influencer un lien entre deux variables. Elle peut modifier l'ampleur, le sens et la forme du lien de dépendance entre deux variables. Par exemple, un lien entre X et Y qui était modéré à l'origine pourrait être en fait très fort pour les hommes et faible pour les femmes ; la variable du sexe serait donc une variable modératrice dans ce cas. Les modèles multiplicatifs présentés au chapitre 10 permettent de tester si une variable discrète est modératrice. Malheureusement, il arrive souvent que des problèmes de multicolinéarité surviennent dans de tels modèles, rendant l'interprétation du modèle difficile.

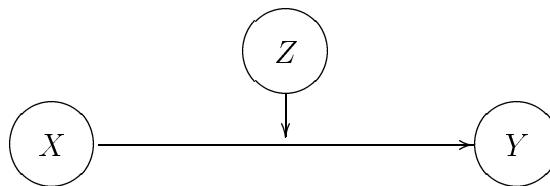


FIG. 21.44 – Effet modérateur de la variable Z sur le lien entre X et Y

Il existe diverses méthodes pour tester un effet modérateur avec les MES, mais ceci dépasse les objectifs de ce cours.

21.9 Introduction au logiciel AMOS

L'interface de Amos Graphics est très conviviale, il est donc possible d'être assez rapidement autonome pour monter un modèle. Pour vous familiariser avec cette interface, nous vous proposons de faire le modèle de la figure 21.45 étape par étape.

1. Commençons par créer un concept latent avec 4 indicateurs. Il faut d'abord créer la variable latente avec , puis ses 4 indicateurs en sélectionnant puis en cliquant 4 fois sur la variable latente.
2. Ensuite, on s'assure que notre variable latente et ses indicateurs sont sélectionnés (par exemple avec), puis à l'aide de , on duplique cette variable pour en

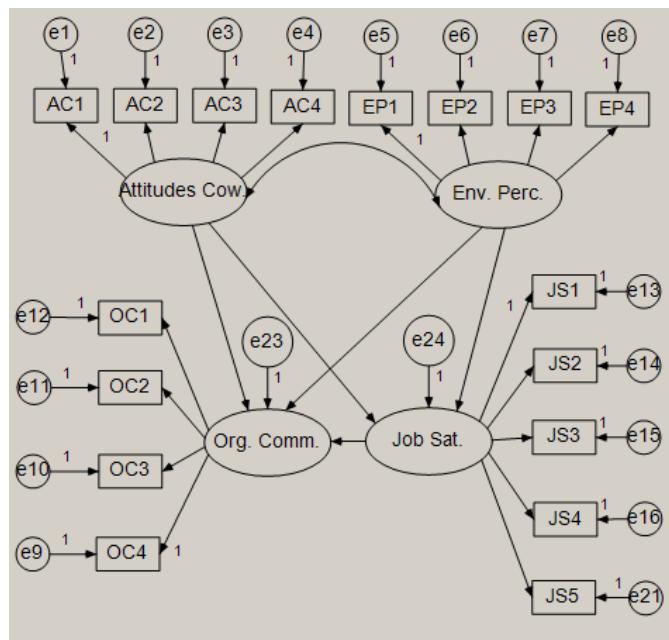


FIG. 21.45 – Modèle à reproduire

avoir 4 au total. Remarquez que c'est toujours le dernier objet dupliqué qui est sélectionné, on fait donc la copie suivante à partir de la copie la plus récente (figure 21.46).

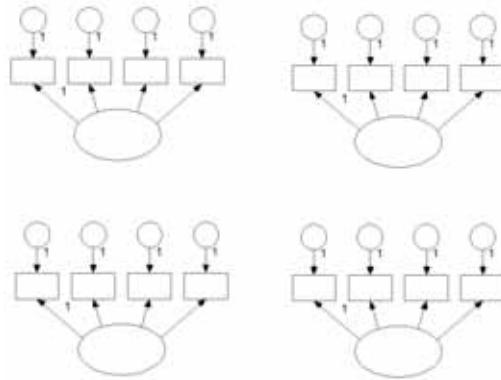


FIG. 21.46 –

3. On ajoute un indicateur à la variable latente en bas à droite à l'aide de  puisque le concept *Job Satisfaction* a 5 indicateurs.
4. On utilise l'outil  pour positionner les indicateurs des deux variables du bas comme dans la figure 21.47. On utilise aussi  si nécessaire pour repositionner un objet, en prenant soin de sélectionner la variable latente et toutes ses composantes que l'on veut repositionner avec  (un objet sélectionné devient bleu).

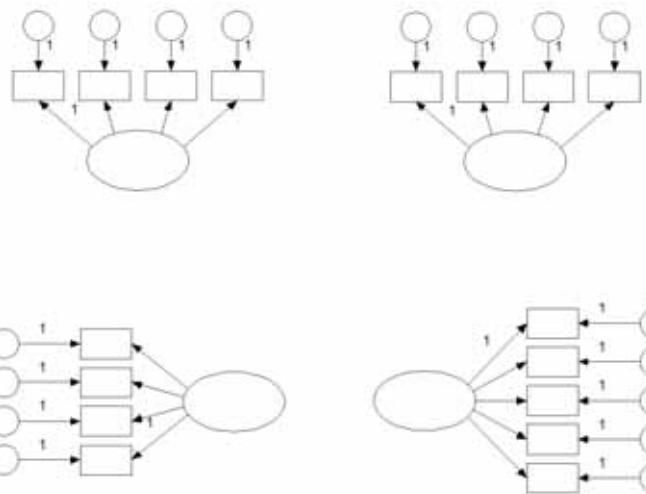


FIG. 21.47 –

5. On ajoute les termes d'erreur aux variables latentes du bas en utilisant , puis les liens de dépendance avec , et le lien d'interdépendance entre les variables exogènes à l'aide de  (figure 21.48). On peut s'aider des outils  et  pour améliorer le tout.
6. On lie maintenant notre modèle à la base de données `hbatsemmod.sav` à l'aide de . On utilise le bouton **File Name** pour trouver le fichier.
7. On utilise  pour faire afficher la liste des variables contenues dans la base de données sélectionnée à l'étape précédente. On associe les noms des variables à nos indicateurs avec des *drag-and-drop* (glisser-déposer). Pour que ce ne soit que

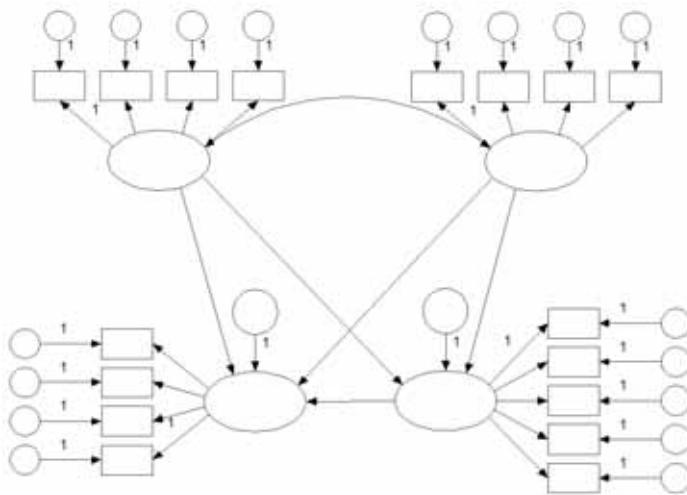


FIG. 21.48 –

les noms et non les *labels* des variables qui s'affichent, on va dans le menu **View** puis **Interface Properties...**, puis dans l'onglet **Misc** on désélectionne l'option **Display variable labels**. On a alors le diagramme de la figure 21.49.

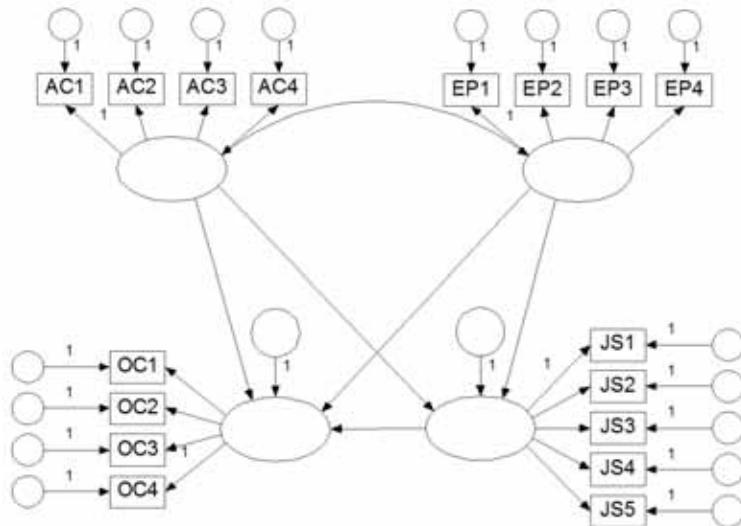


FIG. 21.49 –

8. Sur chaque concept latent, on clique avec le bouton droit, puis dans le menu on sélectionne **Object Properties...**. Ensuite on nomme les concepts latents en ta-

pant le nom voulu dans la fenêtre **Variable name** de l'onglet **Text**. Au besoin on ajuste la taille de la police (figure 21.50).

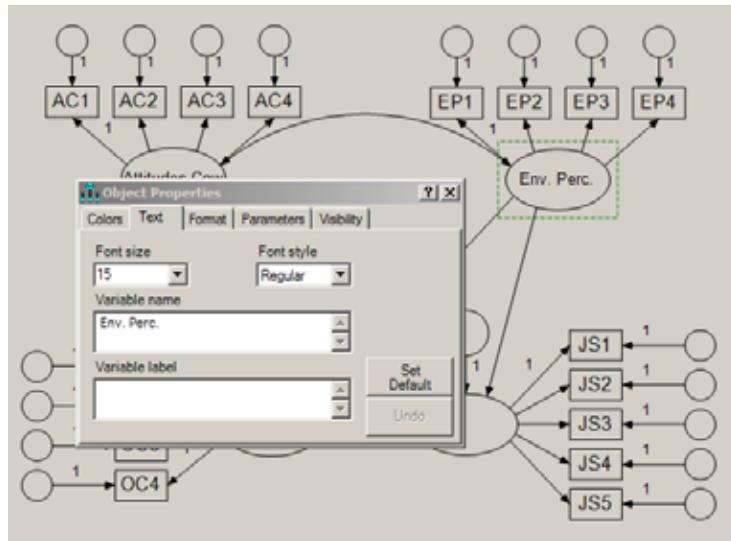


FIG. 21.50 –

9. Pour nommer les termes d'erreur, on utilise la macro **Name Unobserved Variables** du menu **Plugins**.
10. On est prêt à générer la solution pour ce modèle. On va d'abord dans pour les caractéristiques de l'analyse. Par défaut, dans l'onglet **Estimation**, c'est déjà **Maximum likelihood** qui est sélectionné. Dans l'onglet **Output**, on sélectionne tout ce que l'on désire avoir dans les sorties. Ensuite on fait calculer la solution avec . Pour faire apparaître les estimations sur le diagramme, on appuie sur . Par défaut ce sont les valeurs non standardisées qui apparaissent ; pour obtenir les estimations standardisées on sélectionne **Standardized estimates** comme dans le montre la figure 21.51. On obtient alors les estimés de la figure 21.52. Si certaines valeurs sont mal positionnées, on les repositionne avec .
11. Finalement, pour accéder aux différentes sorties, on utilise .

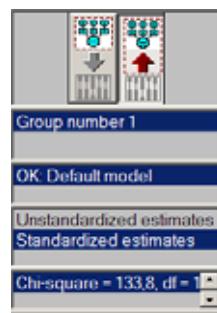


FIG. 21.51 –

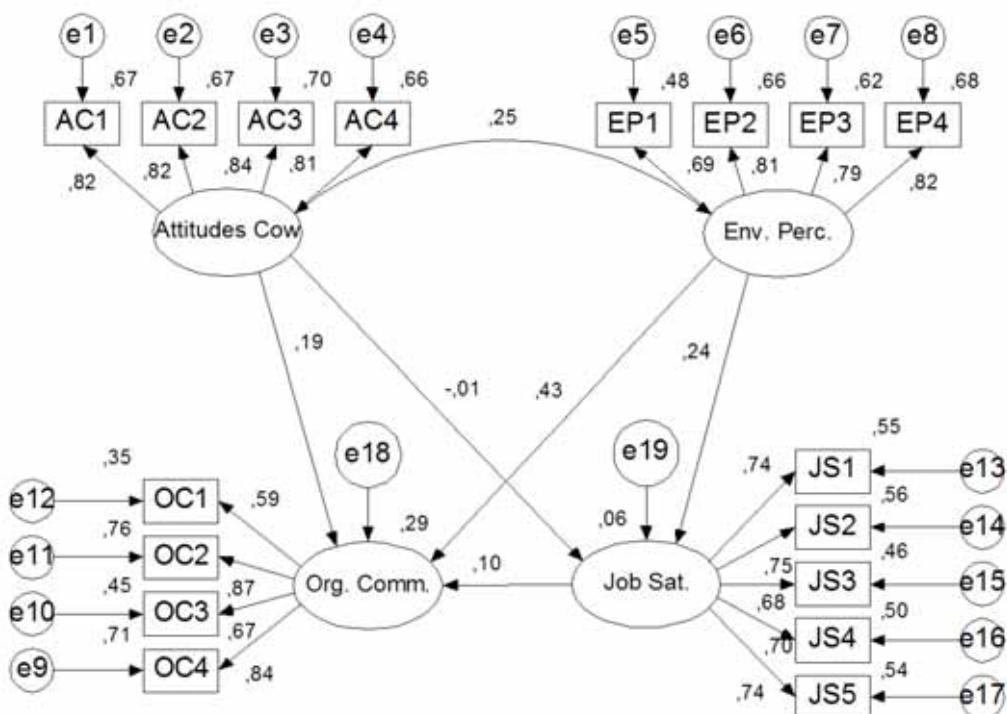


FIG. 21.52 – Le modèle avec les estimations standardisées

Dans les pages suivantes on retrouve de brèves descriptions des principaux outils de Amos Graphics.

-
- | | |
|---|---|
|  | Crée une variable observable (indicateur). |
|  | Crée une variable latente. |
|  | Crée, à l'aide d'un simple clic, un indicateur (attaché à un concept réflexif) et son erreur. |
|  | Crée un lien de dépendance. |
|  | Crée un lien d'interdépendance. |
|  | Ajoute d'un seul clic un terme d'erreur à une variable. |
|  | Permet d'ajouter un titre et des macros. |
|  | Affiche la liste des variables du modèle. |
|  | Affiche la liste des variables de la base de données. |
|  | Pour sélectionner un objet. |
|  | Pour sélectionner tous les objets. |
|  | Pour enlever toute sélection préalablement établie. |
-

	Duplique un objet.
	Déplace un objet.
	Supprime un objet.
	Change la forme d'un objet.
	Pour une rotation de 45° des indicateurs d'une variable latente.
	Applique une réflexion aux indicateurs d'une variable latente.
	Déplace la valeur des paramètres.
	Repositionne le diagramme sur la page.
	Permet d'ajuster d'un seul clic les flèches reliées à une variable (de façon à ce que ce soit le plus symétrique possible).
	Permet de sélectionner la base de données associée au modèle.
	Permet d'accéder aux propriétés de l'analyse.
	Pour calculer les estimés du modèle.



Permet de copier le diagramme, pour le coller dans un éditeur de texte par exemple.



Accès aux sorties textes des estimations du modèle.



Sauvegarde du diagramme.



Accès aux caractéristiques des objets.



Pour transférer une caractéristique d'un objet à un autre objet.



Préserve la symétrie.



Pour faire un zoom sur une partie que l'on sélectionne.



Zoom avant.



Zoom arrière.



Montre la page entière.



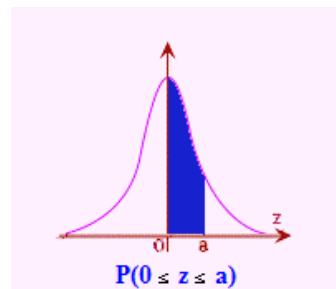
Redimensionne le diagramme pour qu'il occupe une page entière.



Loupe.

Annexe A

Tables de lois



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

FIG. A.1 – Table de la loi normale centrée réduite

dfp	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

FIG. A.2 – Table de la loi de Student

p d.d.l.	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,016	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,805	5,991	7,824	9,210	13,815
3	0,584	2,366	3,865	4,842	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,088	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,340	15,119	16,985	19,812	22,362	25,472	27,888	34,528
14	7,790	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	10,865	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	11,651	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	12,443	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	13,240	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	14,041	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	14,848	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	15,659	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	16,473	24,337	28,172	30,675	34,382	37,652	41,556	44,314	52,620
26	17,292	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052
27	18,114	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476
28	18,939	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,893
29	19,768	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302
30	20,599	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703

FIG. A.3 – Table de la loi du Chi-deux

Annexe B

Solutionnaires

B.1 Solutions de certains exercices du chapitre 3

Exercice 1

1. Le tableau de distribution des fréquences (figure B.1) nous indique qu'il y a 14 femmes (46,7 %) et 16 hommes (53,3 %).

sexe					
	Frequency	Percent	Valid Percent	Cumulative Percent	
Valid Féminin	14	46,7	46,7	46,7	
Masculin	16	53,3	53,3	100,0	
Total	30	100,0	100,0		

FIG. B.1 – Le tableau des fréquences

Dans le tableau des statistiques descriptives (figure B.2), on voit que l'intervalle de confiance de niveau 95 % pour la proportion des femmes qui occupent ce type de poste est [0,28, 0,66]. Ainsi, au niveau de la population, la proportion des femmes

occupant ce type de poste devrait se retrouver entre 28 % et 66 %, et ce 19 fois sur 20. Cet intervalle n'est pas très précis ; ceci est dû au fait que la taille de l'échantillon n'est pas très grande.

Descriptives			
		Statistic	Std. Error
sexé2	Mean	,47	,093
	95% Confidence Interval for Mean	,28 ,66	
	Lower Bound		
	Upper Bound		
	5% Trimmed Mean	,46	
	Median	,00	
	Variance	,257	
	Std. Deviation	,507	
	Minimum	0	
	Maximum	1	
	Range	1	
	Interquartile Range	1	
	Skewness	,141	,427
	Kurtosis	-2,127	,833

FIG. B.2 – L'intervalle de confiance de niveau 95 %

2. L'analyse de la sortie B.3 illustre que, ponctuellement, la moyenne salariale s'estime à 48 869,08 \$. La moyenne tronquée (ou 5 % trimmed mean) (qui a une valeur de 48 663,79 \$) et la médiane (qui a une valeur de 47 187,50 \$) sont près de la moyenne. Donc l'analyste est confiant que le salaire moyen μ_{salaire} tourne vraisemblablement autour de 48 869 \$.

Le coefficient de variation de la variable **salaire** est

$$CV = \frac{s}{\bar{x}} = \frac{10\ 803,71 \$}{48\ 869,08 \$} = 0,22.$$

Il faut donc faire attention à l'utilisation de cette moyenne, il semble il y avoir de la variation (peut-être d'une province à l'autre ?).

Les valeurs minimale et maximale (29 650,00 \$ et 73 400,00 \$) sont assez éloignées, ce qui confirme qu'il semble il y avoir une certaine variation dans les données.

La médiane nous indique que 50 % des cadres ont un salaire égal ou supérieur à 47 187,50 \$.

Descriptives			
salaire	Mean	48869,08	1972,478
	95% Confidence Interval for Mean	Lower Bound Upper Bound	44834,91 52903,25
	5% Trimmed Mean	48663,79	
	Median	47187,50	
	Variance	1,2E+08	
	Std. Deviation	10803,71	
	Minimum	29650,00	
	Maximum	73400,00	
	Range	43750,00	
	Interquartile Range	16345,00	
	Skewness	,308	,427
	Kurtosis	-,447	,833

FIG. B.3 – Les statistiques descriptives de la variable **salaire**

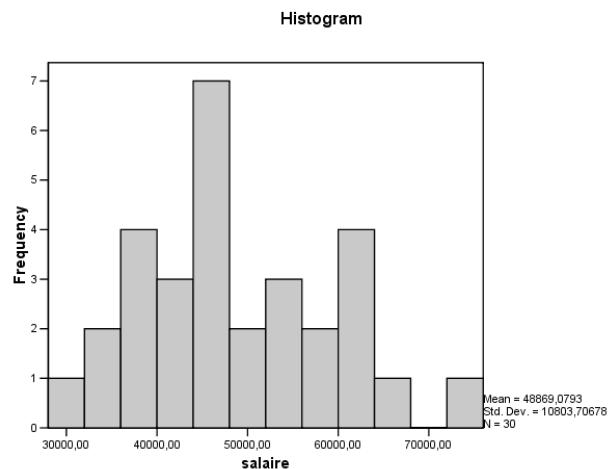
Le coefficient d'asymétrie (Skewness) de 0,308 nous permet de calculer un ratio de 0,721, ce qui en valeur absolue est < 2 et nous permet de croire que la distribution est symétrique (comme une distribution normale). Le coefficient de Kurtosis de -0,447 nous permet de calculer un ratio de -0,537, ce qui en valeur absolue est < 2 et signifie que la distribution est normalement aplatie. Donc la distribution de la variable **salaire** s'apparente à une distribution normale, comme le laisse présager l'histogramme (figure B.4).

3. Selon la figure B.3, le salaire moyen de la population des cadres ayant ce type d'emploi, μ_{salaire} , a une probabilité de 95 % de se retrouver entre 44 834,91 \$ et 52 903,25 \$.
4. Fixons le seuil de signification à $\alpha = 0,05$. Le test d'hypothèses à traiter est le suivant (c'est un test bilatéral) :

$$H_0 : \mu_{\text{salaire}} = 75\,000 \text{ \$}$$

$$H_1 : \mu_{\text{salaire}} \neq 75\,000 \text{ \$}$$

La figure B.5 nous permet de résoudre ce test.

FIG. B.4 – L'histogramme de la variable **salaire**

	One-Sample Test					
	Test Value = 75000				95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
salaire	-13,248	29	,000	-26130,92	-30165,1	-22096,8

FIG. B.5 – Test d'hypothèses

La règle de décision est la suivante : si la *p*-value est plus petite que $\alpha = 0,05$, on rejette H_0 et admettons H_1 comme étant vraisemblable. Sinon, on conserve H_0 . Ici, la *p*-value est égale à 0,000, donc on rejette H_0 . De plus, la cote *t* (qui a une valeur de -13,248 !) nous indique que ce rejet est fort (la valeur 75 000 \$ est à plus de 13 écarts-types de la moyenne échantillonnable qui vaut 48 869,08 \$!). On conclut donc, au risque de se tromper une fois sur 20, que dans la population, la moyenne salariale pour ce type d'emploi n'est pas de 75 000 \$. Ainsi nous n'appuyons pas l'affirmation du supérieur.

Exercice 2

1. Dans ce cas on remplace p par la valeur 0,5. On a $E = 0,035$ et $z_{0,05/2} = z_{0,025} = 1,96$. On obtient :

$$n = \frac{z_{0,025}^2 \cdot p(1-p)}{E^2} = \frac{1,96^2 \cdot 0,5 \cdot (1-0,5)}{0,035^2} = 784.$$

On devra donc planifier un échantillon de taille 784.

2. On a

$$E = z_{0,025} \cdot \sqrt{\frac{p(1-p)}{n}} = 1,96 \cdot \sqrt{\frac{0,25}{400}} = 0,049.$$

On pourra donc avoir une précision d'environ 5 % avec cette taille d'échantillon.

B.2 Solutions de certains exercices du chapitre 4

Exercice 2

Voici les recodifications qui ont été effectuées : la variable `anciennt` a été recodée en une nouvelle variable `ancien2` qui comporte les classes d'ancienneté à l'emploi suivantes :

- 5 ans et moins ;
- entre 5 et 9 ans ;
- entre 9 et 13 ans ;
- 13 ans et plus.

La variable `études` a été recodée en une nouvelle variable `etudes2` qui comporte les classes suivantes :

- `secondaire` (qui regroupe `secondaire` et `diplôme études professionnelles`) ;
- `collégial` (qui regroupe les deux niveaux collégiaux) ;
- `universitaire` (qui regroupe les trois niveaux universitaires).

La relation à étudier est `etudes2` \Rightarrow `ancien2` (il est aussi possible d'étudier la relation `ancien2` \Rightarrow `etudes2`).

On s'intéresse donc à tester les hypothèses

H_0 : Dans la population, l'ancienneté à l'emploi est indépendant du plus haut niveau d'études atteint.

H_1 : Dans la population, l'ancienneté à l'emploi est lié au plus haut niveau d'études atteint.

Fixons le seuil à $\alpha = 0,05$. Puisque les deux variables sont discrètes, on doit utiliser le test du chi-deux pour tester ces hypothèses.

On doit d'abord vérifier si la condition sur les fréquences théoriques est respectée. Or, si on jette un coup d'oeil aux fréquences théoriques (`Expected Count`) du tableau croisé (figure B.6) ou au commentaire dans le bas du tableau du test du chi-deux (figure B.7),

on remarque que l'on a 0 cellules qui ont une fréquence théorique inférieure à 5. On peut donc poursuivre l'analyse sans problème.

		etudes2 Crosstabulation			Total	
		Secondaire	Collégial	Universitaire		
ancien2	5 et moins	Count	43	67	51	161
		Expected Count	82,4	57,4	21,2	161,0
		% within etudes2	11,7%	26,1%	53,7%	22,3%
		Std. Residual	-4,3	1,3	6,5	
5 à 9		Count	72	40	18	130
		Expected Count	66,5	46,3	17,1	130,0
		% within etudes2	19,5%	15,6%	18,9%	18,0%
		Std. Residual	,7	-,9	,2	
9 à 13		Count	78	55	16	149
		Expected Count	76,3	53,1	19,6	149,0
		% within etudes2	21,1%	21,4%	16,8%	20,7%
		Std. Residual	,2	,3	-,8	
13 et +		Count	176	95	10	281
		Expected Count	143,8	100,2	37,0	281,0
		% within etudes2	47,7%	37,0%	10,5%	39,0%
		Std. Residual	2,7	-,5	-4,4	
Total		Count	369	257	95	721
		Expected Count	369,0	257,0	95,0	721,0
		% within etudes2	100,0%	100,0%	100,0%	100,0%

FIG. B.6 – Le tableau croisé

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	91,607 ^a	6	,000
Likelihood Ratio	91,366	6	,000
Linear-by-Linear Association	73,452	1	,000
N of Valid Cases	721		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 17,13.

FIG. B.7 – Le chi-deux

On peut donc faire le test du chi-deux qui permet de tester les hypothèses mentionnées auparavant. La p -value du test étant $0,000 < 0,05 = \alpha$ (figure B.7), on rejette H_0 . Ainsi, au risque de se tromper une fois sur 20, on admet que dans la population, l'ancienneté à l'emploi est lié au plus haut niveau d'études atteint.

Symmetric Measures					
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Nominal by Nominal	Phi	,356			,000
Nominal	Cramer's V	,252			,000
Ordinal by Ordinal	Gamma	-,384	,043	-8,225	,000
N of Valid Cases		721			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. B.8 – Cramer's V et Gamma

Le Cramer's V a une valeur de 0,252 (figure B.8), ce qui nous indique que la relation est faible. Les deux variables à l'étude étant ordinaires, on peut interpréter la statistique Gamma. Sa valeur est de -0,384. Ainsi on peut affirmer que nous sommes en présence d'une relation négative (plus le niveau d'études atteint est élevé, moins on est ancien), et la connaissance du niveau d'études d'un individu améliore nos prédictions sur son niveau d'ancienneté de 38,4 %.

On peut maintenant faire l'interprétation du tableau croisé (figure B.6). La majorité (47,7 %) de ceux qui ont un niveau d'études secondaire ont une ancienneté de 13 ans et plus ; 21,1 % ont entre 9 et 13 ans d'ancienneté, et 19,5 % ont entre 5 et 9 ans d'expérience. Donc seulement 11,7 % d'entre eux ont 5 ans ou moins d'expérience.

Pour ceux qui ont atteint le niveau collégial, la répartition dans les quatre classes d'ancienneté (de la classe de 5 ans ou moins en montant) est de 26,1 %, 15,6 %, 21,4 % et 37 % respectivement.

Finalement, ceux qui ont atteint le niveau universitaire ont en majorité (53,7 %) 5 ans ou moins d'expérience. La répartition dans les classes qui suivent va en diminuant (18,9 %, 16,8 % et 10,5 % respectivement).

On voit donc que ceux qui ont atteint le niveau universitaire ont de façon marquée 5 ans et moins d'ancienneté et de façon marquée ne sont pas les plus anciens.

Inversement, ceux dont le plus haut niveau d'études atteint est le secondaire se retrouvent de façon significative parmi les plus anciens et de façon marquée ne se retrouvent pas parmi les nouveaux.

Exercice 3

On étudie ici la relation **âge** \Rightarrow **risque** au seuil $\alpha = 0,05$. Ceci se fera à l'aide d'un test du χ^2 et d'un tableau croisé puisque les deux variables sont discrètes.

On s'intéresse à tester les hypothèses

H_0 : Dans la population, l'âge est indépendant du risque perçu.

H_1 : Dans la population, l'âge est lié au risque perçu.

Si on jette un coup d'oeil aux fréquences théoriques (**Expected Count**) du tableau croisé (sortie B.10) ou au commentaire dans le bas du tableau du test du chi-deux (première sortie de la figure B.9), on remarque que l'on a 0 cellule sur 12 qui ont une fréquence théorique inférieure à 5. On peut donc utiliser le test du chi-deux sans problème.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	91,242 ^a	6	,000
Likelihood Ratio	94,307	6	,000
Linear-by-Linear Association	61,215	1	,000
N of Valid Cases	155		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5,81.

Symmetric Measures

	Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Nominal by Nominal Phi	,767			,000
Nominal Cramer's V	,543			,000
Ordinal by Ordinal Gamma	,780	,062	10,237	,000
N of Valid Cases	155			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. B.9 – Le test du chi-deux, le Cramer's V et le Gamma.

La p -value du test étant $0,000 < 0,05 = \alpha$ (première sortie de la figure B.9), on rejette

H_0 . Ainsi, au risque de se tromper une fois sur 20, on admet que dans la population, l'âge est lié au risque perçu.

Ensuite, le Cramer's V (deuxième sortie de la figure B.9) nous indique que la relation est forte (il a une valeur de 0,543).

La statistique Gamma a une valeur de 0,78. Ainsi, la relation entre l'âge et le risque perçu est positive : plus un individu est âgé, plus le risque perçu est élevé. Aussi, dans cette relation, la connaissance de la valeur de la variable `age` chez un individu fait que nous avons une probabilité de 78 % de prédire la valeur de la variable `risque` pour ce même individu.

			risque * age Crosstabulation					
			age					
			Moins de 25 ans	Entre 25 et 35 ans	Entre 35 et 45 ans	Plus de 45 ans	Total	
risque	Peu risquée	Count	24	21	4	3	52	
		Expected Count	10,1	13,4	16,4	12,1	52,0	
		% within age	80,0%	52,5%	8,2%	8,3%	33,5%	
		Std. Residual	4,4	2,1	-3,1	-2,6		
	Moyennement risquée	Count	4	19	37	13	73	
		Expected Count	14,1	18,8	23,1	17,0	73,0	
		% within age	13,3%	47,5%	75,5%	36,1%	47,1%	
		Std. Residual	-2,7	,0	2,9	-1,0		
	Très risquée	Count	2	0	8	20	30	
		Expected Count	5,8	7,7	9,5	7,0	30,0	
		% within age	6,7%	,0%	16,3%	55,6%	19,4%	
		Std. Residual	-1,6	-2,8	-,5	4,9		
Total			30	40	49	36	155	
			Expected Count	30,0	40,0	49,0	36,0	
			% within age	100,0%	100,0%	100,0%	100,0%	
							100,0%	

FIG. B.10 – Le tableau croisé de la relation `textttage` ⇒ `risque`

On peut maintenant interpréter le tableau croisé. Chez les moins de 25 ans, 80 % d'entre eux perçoivent peu risquée l'utilisation des services bancaires en ligne, contre seulement 13,3 % et 6,7 % qui la perçoivent comme étant moyennement ou très risquée respectivement. Chez les 25-35 ans, 52,5 % considèrent que c'est peu risqué, et 47,5 % considèrent que c'est moyennement risqué ; 0 % des 25-35 ans considèrent l'utilisation des services bancaires en ligne comme étant très risquée ! Chez les 35-45 ans, la majorité considère que le risque est moyen (ils sont 75,5 %). Seulement 8,2 % d'entre eux consi-

dèrent que c'est peu risqué, et 16,3 % que c'est très risqué. Finalement, pour les plus de 45 ans, la majorité d'entre eux considère que l'utilisation est très risquée (55,6 %). Seulement que 8,6 % d'entre eux considèrent que c'est peu risqué, et 36,1 % considèrent que c'est moyennement risqué.

Ainsi, pour résumer les points forts de cette relation, on remarque d'abord que les moins de 25 ans considèrent de façon marquée que l'utilisation des services bancaires en ligne est peu risquée. Les 25-35 ans considèrent de façon significative que c'est peu risqué. Les 35-45 ans ne sont pas portés à considérer que c'est peu risqué, et ce de façon marquée. En fait ils indiquent de façon significative que c'est moyennement risqué. Finalement, les plus de 45 ans trouvent que l'utilisation des services bancaires en ligne est très risquée, et ce de façon marquée.

B.3 Solutions de certains exercices du chapitre 5

Il faut tout d'abord recoder la variable `bieres` tel que demandé ; on obtient la nouvelle variable `bieres2`

Étudions la relation `bieres2` \Rightarrow `satis`. Fixons le seuil de signification à $\alpha = 0,05$ pour tous les tests.

Faisons d'abord une analyse descriptive de la situation. Le tableau de la sortie B.11 permet de comparer simultanément les statistiques des trois groupes. La satisfaction moyenne du groupe des 0 à 2 bières est de 6,53 sur 10, celle du groupe des 3 à 5 bières est de 7,35 sur 10 et celle du groupe des 6 bières et plus est de 8,70 sur 10. Le CV du groupe des 0 à 2 bières est de 0,28, et celui du groupe des 3 à 5 bières est de 0,19. Donc il faut faire attention à l'utilisation des moyennes de ces groupes. Par contre le CV du groupe des 6 bières et plus est de 0,13, donc pour ce groupe la moyenne est représentative.

		Descriptives					
		Combien de bières avez-vous consommées?					
		0 à 2 bières		3 à 5 bières		6 bières et plus	
		Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Êtes-vous satisfait de votre soirée?	Mean	6,5261	,38460	7,3525	,21647	8,7000	,32310
	95% Confidence Interval for Mean	5,7285		6,9147		7,9889	
				7,7903		9,4111	
	Lower Bound						
	Upper Bound						
		7,3237					
	5% Trimmed Mean	6,6225		7,3889		8,7778	
	Median	6,3000		7,6000		8,9000	
	Variance	3,402		1,874		1,253	
	Std. Deviation	1,84446		1,36907		1,11925	
	Minimum	2,30		4,20		6,00	
	Maximum	8,90		9,80		10,00	
	Range	6,60		5,60		4,00	
	Interquartile Range	3,20		1,85		1,55	
	Skewness	-,471	,481	-,384	,374	-1,202	,637
	Kurtosis	-,224	,935	-,275	,733	2,056	1,232

FIG. B.11 – Statistiques descriptives

Le niveau moyen de satisfaction du groupe des 0 à 2 bières est compris entre 5,73 et 7,32, et ce, 19 fois sur 20 ; celui du groupe des 3 à 5 bières est compris entre 6,91 et 7,79, et ce, 19 fois sur 20, tandis que le niveau moyen de satisfaction du groupe des 6 bières et

plus est compris entre 7,99 et 9,41, et ce, 19 fois sur 20.

On remarque que les deux premiers intervalles de confiance se chevauchent, mais ne chevauchent pas le dernier, ce qui laisse présager une différence significative entre les moyennes.

Puisque la variable **bieres2** a trois modalités et que la variable **satis** est continue, l'analyse de la relation **bieres2** \Rightarrow **satis** se fera à l'aide de l'analyse de la variance. On doit donc d'abord vérifier si les conditions concernant la normalité et l'égalité des variances sont respectées.

Vérifions d'abord si les données des trois populations correspondant aux trois modalités de la variable **bieres2** suivent une loi normale. Pour ce faire, on doit résoudre le test d'hypothèses suivant pour chacune des populations :

H_0 : Les données de la population se répartissent selon une loi normale.

H_1 : Les données de la population ne se répartissent pas selon une loi normale.

On rejette H_0 si les p -values associées aux statistiques de Shapiro-Wilk et Kolmogorov-Smirnov sont toutes deux inférieures à α , sinon on conserve H_0 . On retrouve ces p -values dans la figure B.12.

Tests of Normality						
Combien de bières avez-vous consommées?	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Etes-vous satisfait de votre soirée?						
0 à 2 bières	,120	23	,200*	,935	23	,143
3 à 5 bières	,114	40	,200*	,976	40	,531
6 bières et +	,167	12	,200*	,899	12	,154

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. B.12 – Test de normalité

Tout d'abord, pour ceux qui ont consommé entre 0 et 2 bières, les p -values sont égales à 0,200 et 0,143 ; on ne rejette donc pas H_0 pour cette population.

Pour ceux qui ont consommé entre 3 et 5 bières, les p -values sont égales à 0,200 et 0,531 ; on ne rejette donc pas H_0 pour cette population.

Finalement, pour ceux qui ont consommé 6 bières et plus, les p -values sont égales à 0,200 et 0,154 ; donc encore une fois on ne rejette pas H_0 pour cette population.

Ainsi au seuil $\alpha = 0,05$ l'hypothèse de normalité est respectée pour chacune des trois populations.

Vérifions maintenant l'hypothèse d'égalité des variances. Le test d'hypothèses à résoudre est le suivant :

H_0 : Les variances des populations sont égales ($\sigma_{0 \text{ à } 2 \text{ bières}}^2 = \sigma_{3 \text{ à } 5 \text{ bières}}^2 = \sigma_{6 \text{ bières et +}}^2$).
 H_1 : Au moins une des variances est différente.

On rejette H_0 si la p -value est inférieure à α , sinon on conserve H_0 . On retrouve cette p -value dans la figure B.13.

Puisque cette p -value est égale à 0,085, on conserve H_0 . Ainsi au seuil $\alpha = 0,05$ on peut affirmer que l'hypothèse d'égalité des variances est respectée.

Test of Homogeneity of Variances			
Êtes-vous satisfait de votre soirée?			
Levene Statistic	df1	df2	Sig.
2,549	2	72	,085

FIG. B.13 – Test d'égalité des variances

Puisque les hypothèses de normalité et d'égalité des variances sont toutes deux respectées, on peut utiliser l'analyse de la variance pour résoudre le test d'hypothèses suivant :

H_0 : $\mu_{0 \text{ à } 2 \text{ bières}} = \mu_{3 \text{ à } 5 \text{ bières}} = \mu_{6 \text{ bières et +}}$

H_1 : Au moins une des moyennes est différente.

On rejette H_0 si la p -value de la table ANOVA est inférieure à α , sinon on conserve H_0 . On retrouve cette p -value dans la figure B.14.

Puisque cette p -value est égale à 0,001, on rejette H_0 . Ainsi, au risque de se tromper 1 fois sur 20, nous pouvons affirmer qu'au moins une des moyennes de niveaux de satisfaction est significativement différente des autres. Il reste à voir comment s'exprime cette différence.

ANOVA					
Êtes-vous satisfait de votre soirée?					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	37,390	2	18,695	8,323	,001
Within Groups	161,724	72	2,246		
Total	199,114	74			

FIG. B.14 – Table ANOVA

La figure B.14 nous permet de calculer le $\text{ETA}^2 = \frac{37,390}{199,114} = 0,188$. Ainsi 18,8 % de la variation de la satisfaction est expliquée par la consommation de bières. Aussi on a $\text{ETA} = 0,43$, et donc la relation est qualifiée de correcte.

Il reste à faire l'analyse Post Hoc pour voir comment s'exprime la différence entre les moyennes. Cette analyse se fera à partir de la sortie B.15.

Ici il y a $\frac{k(k-1)}{2} = \frac{3(3-1)}{2} = 3$ paires à considérer pour faire les comparaisons.

Multiple Comparisons

Dependent Variable: Êtes-vous satisfait de votre soirée?

Bonferroni

(I) Combien de bières avez-vous consommées?	(J) Combien de bières avez-vous consommées?	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
0 à 2 bières	3 à 5 bières	-,82641	,39219	,116	-1,7878	,1349
	6 bières et +	-2,17391*	,53370	,000	-3,4821	-,8657
3 à 5 bières	0 à 2 bières	,82641	,39219	,116	-,1349	1,7878
	6 bières et +	-1,34750*	,49329	,024	-2,5567	-,1383
6 bières et +	0 à 2 bières	2,17391*	,53370	,000	,8657	3,4821
	3 à 5 bières	1,34750*	,49329	,024	,1383	2,5567

*. The mean difference is significant at the .05 level.

FIG. B.15 – Analyse Post Hoc

1. μ_0 à 2 bières et μ_3 à 5 bières : la p -value associée à la différence des moyennes est de 0,116, donc au seuil 0,05 on conclut que la différence entre ces deux moyennes n'est pas significative. Donc μ_0 à 2 bières = μ_3 à 5 bières.
2. μ_0 à 2 bières et μ_6 bières et + : la p -value associée à la différence des moyennes est de 0,000, donc au seuil 0,05 on conclut que la différence entre ces deux moyennes est significative. La différence des moyennes s'estime ponctuellement à 2,17 en faveur du groupe des 6 bières et plus, donc μ_6 bières et + > μ_0 à 2 bières. Et il y a une probabilité de 95 % de retrouver la différence de ces moyennes entre 0,87 et 3,48 au niveau de la population.
3. μ_3 à 5 bières et μ_6 bières et + : la p -value associée à la différence des moyennes est de 0,024, donc au seuil 0,05 on conclut que la différence entre ces deux moyennes est significative. La différence des moyennes s'estime ponctuellement à 1,35 en faveur du groupe des 6 bières et plus, donc μ_6 bières et + > μ_3 à 5 bières. Et il y a une probabilité de 95 % de retrouver la différence de ces moyennes entre 0,1383 et 2,56 au niveau de la population.

On peut résumer la situation de la façon suivante :

$$(\mu_0 \text{ à 2 bières} = \mu_3 \text{ à 5 bières}) < \mu_6 \text{ bières et +}.$$

Il semble donc que la consommation de bière ait influencé positivement le niveau de satisfaction par rapport à la soirée...

Étudions maintenant la relation `celibat` \Rightarrow `satis`. Fixons le seuil de signification à $\alpha = 0,05$ pour tous les tests.

Faisons d'abord une analyse descriptive de la situation. Le tableau de la sortie B.16 permet de comparer simultanément les statistiques des deux groupes. La satisfaction moyenne du groupe des non célibataires est de 7,47 sur 10, et celle du groupe des célibataires est de 7,11 sur 10. Le CV du groupe des non célibataires est de 0,20, et celui du groupe des célibataires est de 0,26. Donc il faut faire attention à l'utilisation des moyennes.

		Descriptives			
		Êtes-vous célibataire?			
		non		oui	
		Statistic	Std. Error	Statistic	Std. Error
Êtes-vous satisfait de votre soirée?	Mean	7,4698	,22764	7,1063	,32238
	95% Confidence Interval for Mean	7,0104		6,4487	
		7,9292		7,7638	
	5% Trimmed Mean	7,5115		7,1979	
	Median	7,6000		7,2000	
	Variance	2,228		3,326	
	Std. Deviation	1,49277		1,82367	
	Minimum	4,20		2,30	
	Maximum	10,00		10,00	
	Range	5,80		7,70	
	Interquartile Range	2,50		2,60	
	Skewness	-,435	,361	-,698	,414
	Kurtosis	-,711	,709	,444	,809

FIG. B.16 – Statistiques descriptives

La différence entre les moyennes s'estime ponctuellement à 0,36 en faveur des non célibataires. Plus précisément, la satisfaction moyenne des non célibataires est comprise entre 7,01 et 7,93, et ce, 19 fois sur 20, tandis que la satisfaction moyenne des célibataires est comprise entre 6,45 et 7,76, et ce, 19 fois sur 20.

On remarque que les intervalles de confiance se chevauchent, ce qui laisse présager qu'il n'y a pas de différence significative entre les moyennes. Pour pouvoir confirmer ceci, on doit faire un Independant Samples T Test, qui exige que les deux groupes proviennent de populations normales. Nous devons donc résoudre le test suivant :

H_0 : Les données de la population se répartissent selon une loi normale.

H_1 : Les données de la population ne se répartissent pas selon une loi normale.

On rejette H_0 si les p -values associées aux statistiques de Shapiro-Wilk et Kolmogorov-Smirnov sont toutes deux inférieures à α , sinon on conserve H_0 . On retrouve ces p -values dans la figure B.17.

Tests of Normality						
ceibat	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
satis	.109	43	.200*	.960	43	.140
non	.107	32	.200*	.960	32	.272
oui						

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

FIG. B.17 – Test de normalité

Tout d'abord, pour les non célibataires, les p -values sont égales à 0,200 et 0,140, ce qui est plus grand que 0,05 ; on ne rejette donc pas H_0 pour cette population.

Pour les célibataires, les p -values sont égales à 0,200 et 0,272, ce qui est plus grand que 0,05 ; on ne rejette donc pas H_0 pour cette population.

Ainsi au seuil $\alpha = 0,05$ l'hypothèse de normalité est respectée pour chacune des deux populations.

On peut donc passer au Independant Samples T Test pour comparer les moyennes. Tout d'abord, pour savoir quelle ligne du tableau (figure B.18) utiliser, on doit tester l'égalité des variances :

H_0 : Les variances des deux populations sont égales ($\sigma_{\text{non célibataires}}^2 = \sigma_{\text{célibataires}}^2$).

H_0 : Les variances des deux populations ne sont pas égales ($\sigma_{\text{non célibataires}}^2 \neq \sigma_{\text{célibataires}}^2$).

On rejette l'hypothèse H_0 si la p -value du test de Levene est plus petite que le seuil $\alpha = 0,05$, sinon on ne rejette pas H_0 . Dans notre cas la p -value est égale à 0,304, ce qui

Independent Samples Test										
	Levene's Test for Equality of Variances		t-test for Equality of Means						95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference			
	1,071	,304	,949	73	,346	,36352	,38322	-,40024	1,12728	
Etes-vous satisfait de votre soirée?	Equal variances assumed Equal variances not assumed		,921	58,827	,361	,36352	,39465	-,42623	1,15327	

FIG. B.18 – Independant Samples T Test

est plus grand que 0,05. Ainsi on ne rejette pas H_0 , ce qui signifie que les variances sont égales. Ainsi, dans la suite, on utilisera les statistiques de la première ligne.

On peut maintenant résoudre le test

$$H_0 : \mu_{\text{non célibataires}} - \mu_{\text{célibataires}} = 0$$

$$H_1 : \mu_{\text{non célibataires}} - \mu_{\text{célibataires}} \neq 0$$

Pour ce faire, on utilise la p -value de la 5^e colonne du tableau B.18. On rejette l'hypothèse H_0 si cette p -value est plus petite que le seuil $\alpha = 0,05$, sinon on ne rejette pas H_0 . Dans notre cas, la p -value est égale à 0,346, et par conséquent on ne rejette pas H_0 . Ainsi, au seuil $\alpha = 0,05$, on affirme qu'il n'y a pas de différence significative entre le niveau moyen de satisfaction des non célibataires et le niveau moyen de satisfaction des célibataires ($\mu_{\text{non célibataires}} = \mu_{\text{célibataires}}$).

B.4 Solutions de certains exercices du chapitre 6

Exercice 1

- Le graphe de la relation est donné par la figure B.19. On peut affirmer que nous sommes en présence d'une relation linéaire (positive) puisque les points semblent se regrouper assez uniformément autour de la droite de régression.

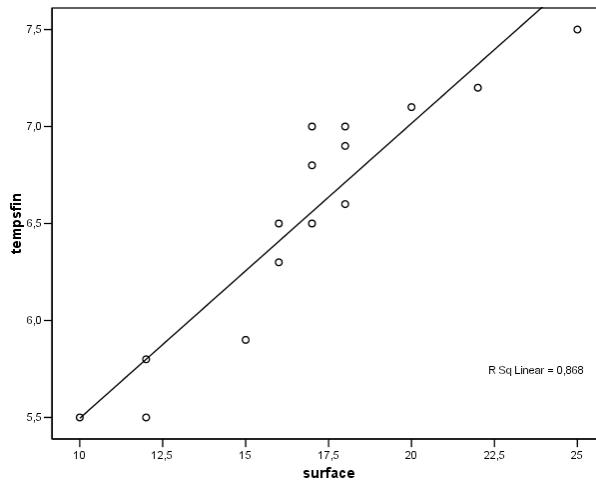


FIG. B.19 – Le graphe de la relation

- Ce % est donné par le coefficient de détermination r^2 , que l'on retrouve dans la première sortie de la figure B.20. On a $r^2 = 0,868$, donc il y a 86,8 % de la variation des temps de finition qui est expliqué par la surface des toiles.
- On retrouve les coefficients de la droite (b_0 et b_1) dans la dernière sortie de la figure B.20. Puisque $b_0 = 3,976$ et $b_1 = 0,152$, l'équation de la droite est

$$\hat{y}_{\text{tempsfin}} = 3,976 + 0,152x_{\text{surface}}.$$

Pour affirmer que la droite est significative, on doit résoudre le test suivant :

H_0 : La régression est non significative dans la population ($\beta_1 = 0$).

H_1 : La régression est significative dans la population ($\beta_1 \neq 0$).

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	,932 ^a	,868	,858	,2354	

a. Predictors: (Constant), surface

ANOVA ^b					
Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4,756	1	4,756	85,853
	Residual	,720	13	,055	
	Total	5,476	14		

a. Predictors: (Constant), surface

b. Dependent Variable: tempsfin

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	3,976	,283	14,030	,000
	surface	,152	,016	,9,266	,000

a. Dependent Variable: tempsfin

FIG. B.20 – Sorties de la régression

Pour ce faire on utilise la *p*-value de la table ANOVA (deuxième sortie de la figure B.20). Puisque la *p*-value est égale à $0,000 < 0,05$, on rejette H_0 . Ainsi, au risque de se tromper une fois sur 20, on peut affirmer que la régression est significative.

4. Il suffit de remplacer x_{surface} par 13 dans l'équation : $\hat{y}_{\text{tempsfin}} = 3,976 + 0,152x_{\text{surface}} = 3,976 + 0,152 \times 13 = 5,952$. Donc le temps moyen de finition pour des toiles dont la surface est de 13 m^2 est 5,952.
5. Calculons l'intervalle de confiance de niveau 95 % pour des toiles de 13 m^2 . SPSS nous donne l'intervalle suivant : [5,76, 6,14]. Ainsi le temps moyen de finition des toiles de 13 m^2 a une probabilité de 95 % de se retrouver entre 5,76 et 6,14.

Exercice 3 Tout d'abord, une brève analyse descriptive des variables `satis` et `musique` est de mise... je vous laisse la faire.

La sortie B.21 nous permet d'affirmer que nous sommes en présence d'une relation linéaire (positive) puisque les points semblent se regrouper assez uniformément autour de la droite de régression.

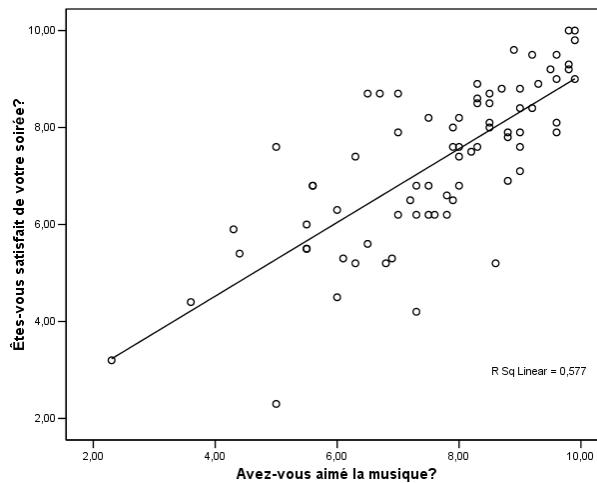


FIG. B.21 – Le graphe de la relation

La première sortie de la figure B.22 nous indique que le coefficient de corrélation r entre les variables `satis` et `musique` est de 0,759, ce qui indique une interrelation linéaire très forte.

Dans la même sortie, on retrouve aussi le coefficient de détermination $r^2 = 0,577$, qui nous indique que la satisfaction par rapport à la musique explique 57,7 % de la variation de la satisfaction générale.

La table ANOVA de la figure B.22 nous permet de tester les hypothèses suivantes :

H_0 : Dans la population, la régression n'est pas significative.

H_1 : Dans la population, la régression est significative.

Puisque la p -value = 0,000, on rejette H_0 au seuil $\alpha = 0,05$. Donc au risque de se tromper une fois sur 20, on peut affirmer que la régression est significative.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,759 ^a	,577	,571	1,07476

a. Predictors: (Constant), Avez-vous aimé la musique?

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	114,792	1	114,792	99,378	,000 ^a
	Residual	84,322	73	1,155		
	Total	199,114	74			

a. Predictors: (Constant), Avez-vous aimé la musique?

b. Dependent Variable: Êtes-vous satisfait de votre soirée?

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	1,483	,598	2,479	,015	,291	2,674
	Avez-vous aimé la musique?	,760	,076	,9,969	,000	,608	,912

a. Dependent Variable: Êtes-vous satisfait de votre soirée?

FIG. B.22 – Sorties de la régression

Finalement, la dernière sortie de la figure B.22 nous permet d'écrire la droite de régression. On trouve $b_0 = 1,483$ (la constante) et $b_1 = 0,76$ (coefficient associé à la variable **musique**). Donc la droite s'écrit $\hat{y}_{\text{satis}} = 1,483 + 0,76x_{\text{musique}}$.

On a l'interprétation suivante : le b_0 nous indique que théoriquement, une satisfaction nulle par rapport à la musique donnerait une satisfaction générale de 1,483 sur 10. Ceci n'a pas vraiment de sens dans ce contexte puisqu'il n'y a pas d'individu qui a indiqué une valeur nulle pour la variable **musique**. On voit aussi qu'il y a une probabilité de 95 % que le paramètre β_0 soit compris entre 0,291 et 2,674. Le b_1 , lui, nous indique que pour chaque point supplémentaire de satisfaction par rapport à la musique, la satisfaction générale augmente en moyenne de 0,76 point. Cette augmentation moyenne a une probabilité de

95 % de se retrouver entre 0,608 et 0,912 au niveau de la population.

Faisons quelques estimations. La satisfaction moyenne pour des individus dont la satisfaction par rapport à la musique est de 7 sur 10 est de 6,8 ($1,483 + 0,76 \times 7 = 6,8$). Cette valeur peut être obtenue dans SPSS (colonne PRE).

L'intervalle de confiance de niveau 95 % pour la satisfaction générale moyenne des individus dont la satisfaction par rapport à la musique est de 7 sur 10 est [6,53, 7,07]. Ainsi, au niveau de la population étudiée, la vraie moyenne de la satisfaction générale des individus dont la satisfaction par rapport à la musique est de 7 sur 10 a une probabilité de 95 % d'être comprise entre 6,53 et 7,07. Les bornes de cet intervalle sont obtenues de SPSS dans les colonnes LMCI (borne inférieure) et UMCI (borne supérieure).

B.5 Solution de l'exercice du chapitre 7

Tout d'abord, pour ce qui est des tableaux croisés, je ne ferai que donner quelques détails.

			sexe		Total	
			féminin	masculin		
salaried	250\$-450\$	Count	10	14	24	
		Expected Count	4,1	19,9	24,0	
		% within sexe	29,4%	8,4%	12,0%	
		Std. Residual	2,9	-1,3		
	450\$-650\$	Count	18	102	120	
		Expected Count	20,4	99,6	120,0	
		% within sexe	52,9%	61,4%	60,0%	
		Std. Residual	-.5	,2		
	650\$ et +	Count	6	50	56	
		Expected Count	9,5	46,5	56,0	
		% within sexe	17,6%	30,1%	28,0%	
		Std. Residual	-1,1	,5		
Total		Count	34	166	200	
		Expected Count	34,0	166,0	200,0	
		% within sexe	100,0%	100,0%	100,0%	

FIG. B.23 – Tableau croisé de la relation sexe \Rightarrow salairecl

D'après les sorties de la figure B.24, la relation sexe \Rightarrow salairecl existe et est qualifiée de faible. Il semble que les femmes gagnent significativement entre 250 \$ et 450 \$, ce qui est la classe la plus basse pour le salaire. On peut donc croire que les femmes gagnent moins que les hommes dans cette entreprise, mais peut-être y a-t-il une explication. (Formellement il faudrait bien sûr traiter ceci avec plus de détails.)

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	12,257 ^a	2	,002
Likelihood Ratio	10,167	2	,006
Linear-by-Linear Association	8,392	1	,004
N of Valid Cases	200		

a. 1 cells (16,7%) have expected count less than 5. The minimum expected count is 4,08.

Symmetric Measures		
	Value	Approx. Sig.
Nominal by Phi	,248	,002
Nominal Cramer's V	,248	,002
N of Valid Cases	200	

- a. Not assuming the null hypothesis.
- b. Using the asymptotic standard error assuming the null hypothesis.

FIG. B.24 – χ^2 et Cramer's V de la relation sexe \Rightarrow salairecl

La figure B.25 présente le tableau croisé de la relation sexe \Rightarrow salairecl avec la variable de contrôle fonction. L'analyse de ce tableau revient à faire l'analyse de trois tableaux croisés.

salairecl * sexe * fonction Crosstabulation				
fonction	salairecl	sexe		Total
		féminin	masculin	
administration	250\$-450\$	Count	7	14
		Expected Count	4,5	14,0
		% within sexe	70,0%	33,3%
		Std. Residual	1,2	,8
	450\$-650\$	Count	3	17
		Expected Count	5,5	17,0
		% within sexe	30,0%	66,7%
		Std. Residual	-1,1	,7
	Total		10	31
	Count	21	31,0	
	Expected Count	10,0	31,0	
	% within sexe	100,0%	100,0%	
production	250\$-450\$	Count	3	10
		Expected Count	1,6	10,0
		% within sexe	14,3%	6,5%
		Std. Residual	1,1	,5
	450\$-650\$	Count	13	91
		Expected Count	14,8	91,0
		% within sexe	61,9%	72,2%
		Std. Residual	,5	,2
	650\$ et +	Count	5	28
		Expected Count	4,6	28,0
		% within sexe	23,8%	21,3%
		Std. Residual	,2	-,1
	Total		21	129
	Count	108	129,0	
	Expected Count	21,0	129,0	
	% within sexe	100,0%	100,0%	
direction	450\$-650\$	Count	2	12
		Expected Count	,9	12,0
		% within sexe	66,7%	27,0%
		Std. Residual	1,2	-,3
	650\$ et +	Count	1	28
		Expected Count	2,1	28,0
		% within sexe	33,3%	73,0%
		Std. Residual	-,8	,2
	Total		3	40
	Count	37	40,0	
	Expected Count	3,0	40,0	
	% within sexe	100,0%	100,0%	

FIG. B.25 – Tableau croisé de la relation sexe \Rightarrow salairecl avec la variable de contrôle fonction

Chi-Square Tests					
		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
fonction	Pearson Chi-Square	3,677 ^a	1	,055	
	Continuity Correction ^b	2,346	1	,126	
	Likelihood Ratio	3,733	1	,053	
	Fisher's Exact Test				,121
	Linear-by-Linear Association	3,659	1	,059	
	N of Valid Cases	31			,063
production	Pearson Chi-Square	1,698 ^a	2	,428	
	Continuity Correction ^b	1,487	2	,476	
	Likelihood Ratio	,178	1	,874	
	Linear-by-Linear Association				
	N of Valid Cases	129			
direction	Pearson Chi-Square	2,076 ^a	1	,150	
	Continuity Correction ^b	,618	1	,432	
	Likelihood Ratio	1,869	1	,172	
	Fisher's Exact Test				,209
	Linear-by-Linear Association	2,024	1	,155	
	N of Valid Cases	40			,209

a. Computed only for a 2x2 table.

b. 1 cells (25,0%) have expected count less than 5. The minimum expected count is 4,52.

c. 2 cells (33,3%) have expected count less than 5. The minimum expected count is 1,63.

d. 2 cells (50,0%) have expected count less than 5. The minimum expected count is ,90.

Symmetric Measures

fonction			Value	Approx. Sig.
administration	Nominal by Nominal	Phi	,344	,055
		Cramer's V	,344	,055
	N of Valid Cases		31	
production	Nominal by Nominal	Phi	,115	,428
		Cramer's V	,115	,428
	N of Valid Cases		129	
direction	Nominal by Nominal	Phi	,228	,150
		Cramer's V	,228	,150
	N of Valid Cases		40	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

FIG. B.26 – χ^2 et Cramer's V de la relation sexe \Rightarrow salairecl avec la variable de contrôle fonction

Tout d'abord, la première sortie de la figure B.26 nous permet de tester les hypothèses suivantes :

H_0 : Dans la population, le sexe est indépendant du salaire hebdomadaire pour ceux exerçant une fonction administrative.

H_1 : Dans la population, le sexe est lié au salaire hebdomadaire pour ceux exerçant une fonction administrative.

H_0 : Dans la population, le sexe est indépendant du salaire hebdomadaire pour ceux exerçant une fonction de production.

H_1 : Dans la population, le sexe est lié au salaire hebdomadaire pour ceux exerçant une fonction de production.

H_0 : Dans la population, le sexe est indépendant du salaire hebdomadaire pour ceux exerçant une fonction de direction.

H_1 : Dans la population, le sexe est lié au salaire hebdomadaire pour ceux exerçant une fonction de direction.

En fait il y a un petit problème : le pré-requis au sujet des fréquences théoriques n'est respecté pour aucun des trois tableaux. Le premier a 25 % des cellules ayant une fréquence théorique en bas de 5, pour le deuxième tableau ce pourcentage monte à 33,3 %, et pour le troisième non seulement le pourcentage grimpe à 50 %, mais en plus le minimum des fréquences théoriques est de 0,9, ce qui est plus petit que le minimum exigé de 1.

Ainsi aucun des trois tests d'hypothèses ne peut vraiment être considéré comme étant valide lorsque traité avec le test du χ^2 . Cependant, nous traiterons le test du premier tableau puisque le 25 % est assez près de 20 %. Ensuite, nous ferons une analyse descriptive des tableaux pour quand même tenter de comprendre le phénomène.

Donc tout d'abord, traitons le test suivant au seuil $\alpha = 0,05$:

H_0 : Dans la population, le sexe est indépendant du salaire hebdomadaire pour ceux exerçant une fonction administrative.

H_1 : Dans la population, le sexe est lié au salaire hebdomadaire pour ceux exerçant une fonction administrative.

La p -value étant de $0,055 > 0,05$, nous ne rejetons pas H_0 au seuil $\alpha = 0,05$. Ainsi pour les employés ayant une fonction administrative, il n'y a pas de relation entre leur sexe et leur salaire.

Il semble qu'on pourrait tirer une conclusion semblable pour les fonctions de production et de direction car toutes les fréquences observées des tableaux sont semblables aux fréquences théoriques.

De façon descriptive, on voit pour la fonction administrative que 70 % des femmes gagnent entre 250 \$ et 450 \$, et 30 % gagnent entre 450 \$ et 650 \$. Pour les hommes, ces pourcentages sont respectivement de 33,3 % et 66,7 %. Et personne ne gagne plus de 650 \$. On voit qu'il y a quand même un certain phénomène (les hommes semblent avoir un meilleur salaire) ; on pourrait parler d'une tendance puisque le lien entre sexe et salaire aurait été admis au seuil de 10 % (et aurait été qualifié d'intéressant puisque le Cramer's V a une valeur de 0,344).

Pour la fonction de production, 14,3 % des femmes gagnent entre 250 \$ et 450 \$, 61,9 % gagnent entre 450 \$ et 650 \$, et 23,8 % gagnent plus de 650 \$. Pour les hommes, ces pourcentages sont 6,5 %, 72,2 % et 21,3 % respectivement, ce qui est assez semblable. Donc pour la fonction de production les hommes et les femmes semblent avoir un salaire semblable.

Finalement, pour la fonction de direction, 66,7 % des femmes gagnent entre 450 \$ et 650 \$, et 33,3 % plus de 650 \$. Pour les hommes les pourcentages sont 27 % et 73 % respectivement. Par contre il n'y a que 3 femmes occupant ce type de poste tandis qu'il y a 37 hommes, il n'est donc pas très approprié de comparer ces %.

Donc dans l'ensemble, on pourrait dire que la relation entre le sexe et le salaire détectée plus tôt (mais qui était quand même déjà faible) n'est qu'illusoire pour les fonctions de production et de direction, mais semble être quelque peu plausible pour la fonction d'administration. Cette analyse demeure sommaire, il faudrait d'autres informations (comme l'ancienneté et le niveau de scolarité par exemple) pour mieux comprendre.

Les autres sorties présentent les analyses qui peuvent aider à mieux comprendre pourquoi la relation entre sexe et salaire est plutôt illusoire. En effet, on voit que la variable fonction est en lien avec le sexe et le salaire, ce qui explique en partie la nature du lien entre sexe et salaire.

			fonction			Total	
			administratiion	production	direction		
salairecl	250\$-450\$	Count	14	10	0	24	
		Expected Count	3,7	15,5	4,8	24,0	
		% within fonction	45,2%	7,8%	,0%	12,0%	
		Std. Residual	5,3	-1,4	-2,2		
	450\$-650\$	Count	17	91	12	120	
		Expected Count	18,6	77,4	24,0	120,0	
		% within fonction	54,8%	70,5%	30,0%	60,0%	
		Std. Residual	-,4	1,5	-2,4		
	650\$ et +	Count	0	28	28	56	
		Expected Count	8,7	36,1	11,2	56,0	
		% within fonction	,0%	21,7%	70,0%	28,0%	
		Std. Residual	-2,9	-1,4	5,0		
Total		Count	31	129	40	200	
		Expected Count	31,0	129,0	40,0	200,0	
		% within fonction	100,0%	100,0%	100,0%	100,0%	

FIG. B.27 – Tableau croisé de la relation fonction \Rightarrow salairecl

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79,381 ^a	4	,000
Likelihood Ratio	75,189	4	,000
Linear-by-Linear Association	61,931	1	,000
N of Valid Cases	200		

a. 2 cells (22,2%) have expected count less than 5. The minimum expected count is 3,72.

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	,630	,000
Nominal	Cramer's V	,445	,000
N of Valid Cases		200	

- a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

FIG. B.28 – χ^2 et Cramer's V de la relation fonction \Rightarrow salairecl

			sexe		Total	
			féminin	masculin		
fonction	administration	Count	10	21	31	
		Expected Count	5,3	25,7	31,0	
		% within sexe	29,4%	12,7%	15,5%	
		Std. Residual	2,1	-,9		
	production	Count	21	108	129	
		Expected Count	21,9	107,1	129,0	
		% within sexe	61,8%	65,1%	64,5%	
		Std. Residual	-,2	,1		
	direction	Count	3	37	40	
		Expected Count	6,8	33,2	40,0	
		% within sexe	8,8%	22,3%	20,0%	
		Std. Residual	-1,5	,7		
Total		Count	34	166	200	
		Expected Count	34,0	166,0	200,0	
		% within sexe	100,0%	100,0%	100,0%	

FIG. B.29 – Tableau croisé de la relation sexe \Rightarrow fonction

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,721 ^a	2	,021
Likelihood Ratio	7,437	2	,024
Linear-by-Linear Association	7,268	1	,007
N of Valid Cases	200		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 5,27.

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	,196	,021
Nominal	Cramer's V	,196	,021
N of Valid Cases		200	

- a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

FIG. B.30 – χ^2 et Cramer's V de la relation sexe \Rightarrow fonction

B.6 Solution de l'exercice 3 du chapitre 8

On s'intéresse ici à la relation (**interface**, **couleurs**) \Rightarrow **appreciation**. Cette analyse se fera à l'aide d'une ANOVA à deux facteurs puisque les variables **interface** et **couleurs** sont discrètes tandis que la variable **appreciation** est continue.

Tout d'abord, une courte analyse descriptive est de mise pour prendre le pouls des données.

Dans la figure B.31 on voit d'abord les statistiques descriptives de la variable **appreciation** selon les groupements induits par la variable **interface**. Pour l'interface I, la moyenne de l'appréciation est de 5,5. Pour l'interface II, la moyenne est de 7,083, et pour l'interface III elle est de 6,5. Les CV pour ces groupes sont respectivement de 0,25, 0,225 et 0,454. Donc il faut faire attention à l'utilisation de la moyenne pour les interfaces I et II, et celle de l'interface III n'est pas représentative.

Ensuite, dans le deuxième tableau de la figure B.31, on voit les statistiques descriptives de la variable **appreciation** selon les groupements induits par la variable **couleurs**. Pour l'ensemble de couleurs I, la moyenne de l'appréciation est de 7,5. Pour l'ensemble de couleurs II, la moyenne est de 7,250 et pour l'ensemble de couleurs III on a une moyenne de 4,333. Les CV pour ces groupes sont respectivement de 0,189, 0,242 et 0,34. Donc il faut faire attention à l'utilisation de la moyenne pour les ensembles de couleurs I et II, tandis que la moyenne pour l'ensemble de couleurs III n'est pas représentative.

On perçoit donc certaines fluctuations pour l'appréciation moyenne selon l'interface et les couleurs. L'analyse avec l'ANOVA à deux facteurs nous permettra de voir si ces fluctuations sont présentes dans la population, et s'il y a une interaction entre l'interface et les couleurs. Fixons le seuil α à 0,05 pour tous les tests.

Descriptives

		interface					
		Interface I		Interface II		Interface III	
		Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
appreciation	Mean	5.500	.5627	7.083	.6509	6.500	1.2042
	95% Confidence Interval for Mean	Lower Bound	4.053	5.410		3.405	
		Upper Bound	6.947	8.756		9.595	
	5% Trimmed Mean		5.500	7.037		6.556	
	Median		5.250	6.500		7.500	
	Variance		1.900	2.542		8.700	
	Std. Deviation		1.3784	1.5943		2.9496	
	Minimum		3.5	5.5		2.0	
	Maximum		7.5	9.5		10.0	
	Range		4.0	4.0		8.0	
	Interquartile Range		2.1	2.9		5.0	
	Skewness		.086	.845	.773	.845	-.666
	Kurtosis		.173	1.741	-1.138	1.741	-.614
							1.741

Descriptives

		couleurs					
		Couleurs I		Couleurs II		Couleurs III	
		Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
appreciation	Mean	7.500	.5774	7.250	.7159	4.333	.6009
	95% Confidence Interval for Mean	Lower Bound	6.016	5.410		2.789	
		Upper Bound	8.984	9.090		5.878	
	5% Trimmed Mean		7.444	7.250		4.370	
	Median		7.250	7.500		4.500	
	Variance		2.000	3.075		2.167	
	Std. Deviation		1.4142	1.7536		1.4720	
	Minimum		6.0	5.0		2.0	
	Maximum		10.0	9.5		6.0	
	Range		4.0	4.5		4.0	
	Interquartile Range		2.1	3.4		2.5	
	Skewness		1.193	.845	-.167	.845	-.640
	Kurtosis		1.669	1.741	-1.557	1.741	-.300
							1.741

FIG. B.31 – Statistiques descriptives

En regardant les données, on voit que pour chaque traitement on a 2 essais, ce qui respecte le minimum requis qui est de 2 essais par traitement.

Tests of Between-Subjects Effects					
Dependent Variable: appreciation					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	66.028 ^a	8	8.253	10.072	.001
Intercept	728.347	1	728.347	888.831	.000
interface	7.694	2	3.847	4.695	.040
couleurs	37.194	2	18.597	22.695	.000
interface * couleurs	21.139	4	5.285	6.449	.010
Error	7.375	9	.819		
Total	801.750	18			
Corrected Total	73.403	17			

a. R Squared = .900 (Adjusted R Squared = .810)

FIG. B.32 – Table ANOVA à deux facteurs

On doit d'abord voir si au moins un des facteurs influence les ventes. Le test à résoudre est le suivant :

H_0 : Le modèle n'est pas significatif au niveau de la population.

H_1 : Le modèle est significatif au niveau de la population.

On résout ce test à l'aide de la table ANOVA (figure B.32). Puisque la p -value = $0,001 < 0,05$ (vis-à-vis **Corrected Model**), on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet que le modèle est significatif au niveau de la population.

On doit maintenant vérifier s'il y a une interaction entre les facteurs. Pour ce faire on résout le test suivant :

H_0 : Aucune interaction n'existe entre l'interface et les couleurs au niveau de la population (tous les $(\alpha\beta)_{ij} = 0$).

H_1 : L'interface et les couleurs interagissent au niveau de la population (au moins un des $(\alpha\beta)_{ij} \neq 0$)

On résout ce test à l'aide de la table ANOVA (figure B.32). Puisque la p -value =

$0,010 < 0,05$ (vis-à-vis `interface*couleurs`), on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet qu'il y a une interaction significative entre les deux facteurs.

Pour décrire cette interaction on peut s'appuyer sur le graphe de la figure B.33. (Il aurait aussi été possible de le faire en prenant le graphe de la figure B.34.)

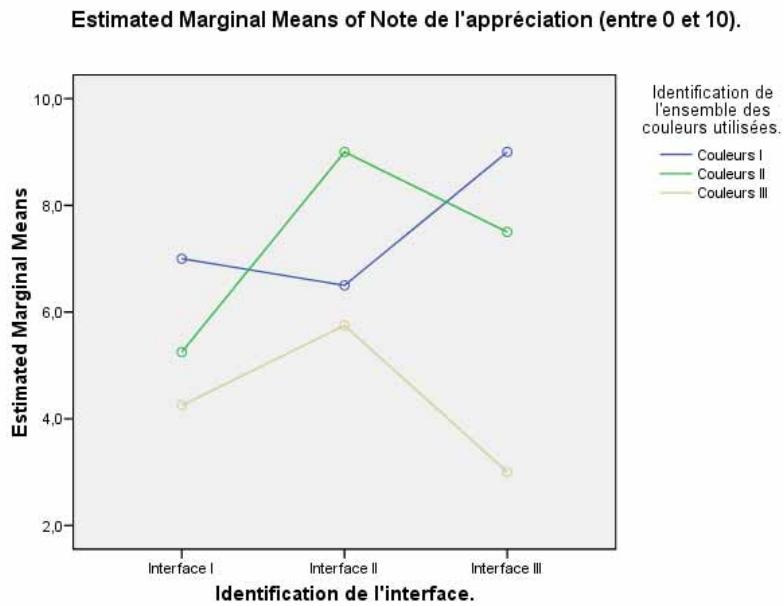


FIG. B.33 – Graphe de la relation

On remarque que ce sont les combinaisons InterfaceII-CouleursII et InterfaceIII-CouleursI qui obtiennent les meilleures appréciations. L'ensemble de couleurs III recueille les moins bonnes appréciations, sauf avec l'interface II où l'appréciation est légèrement meilleure qu'avec l'ensemble de couleurs II combiné avec l'interface I.

Si on regarde ce qui se passe pour chaque interface, on voit que la première obtient la meilleure appréciation avec les couleurs I, suivie de la combinaison avec les couleurs II.

Pour l'interface II, on sait déjà que la meilleure combinaison est avec les couleurs II, et elle est suivie de celle avec les couleurs I.

Pour l'interface III, tel que dit précédemment la meilleure combinaison est celle avec

les couleurs I, et c'est suivi de la combinaison avec les couleurs II.

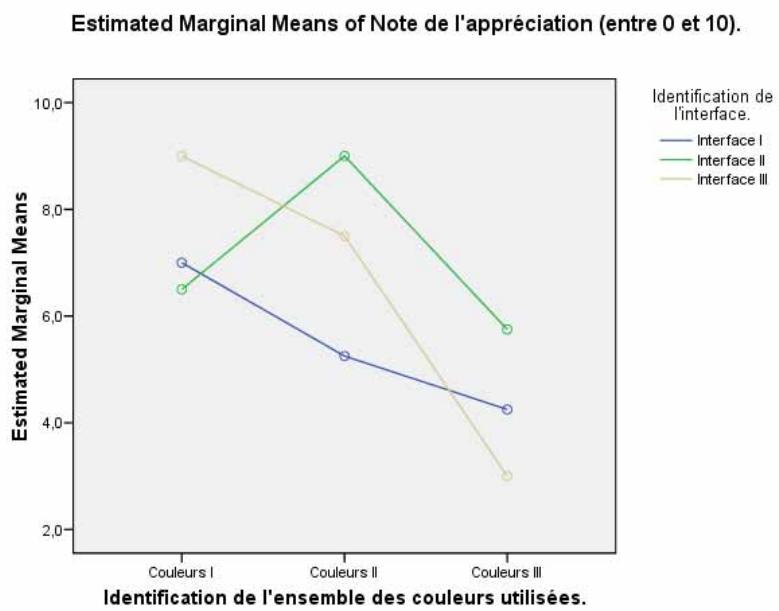


FIG. B.34 – Autre graphe possible

B.7 Solution de l'exercice 2 du chapitre 9

On s'intéresse ici à la relation (`revenu`, `nb_pers`, `scolarite`) \Rightarrow `soldé`. Cette analyse se fera à l'aide d'une régression linéaire multiple puisque les variables sont continues.

Tout d'abord, une courte analyse descriptive est de mise pour prendre le pouls des données. Je vous laisse la faire à l'aide du tableau de la figure B.35 (moyennes, CV, jeter un coup d'œil aux min-max pour voir si tout semble normal).

Descriptives								
	Total des soldes.		Revenu mensuel.		Nombre de personnes dans le ménage.		Niveau de scolarité du «chef» de famille en années.	
	Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Mean	5254,75	743,973	3458,30	286,314	3,25	,239	10,85	,483
95% Confidence Interval for Mean	Lower Bound	3697,60	2859,04		2,75		9,84	
	Upper Bound	6811,90	4057,56		3,75		11,86	
5% Trimmed Mean		5094,72	3396,89		3,28		10,72	
Median		4628,50	3158,00		3,50		10,00	
Variance		1,1E+07	1839518		1,145		4,661	
Std. Deviation		3327,148	1280,437		1,070		2,159	
Minimum		725	1800		1		8	
Maximum		12665	6222		5		16	
Range		11940	4422		4		8	
Interquartile Range		3721	1736		1		3	
Skewness		,596	,702	,512	-,842	,512	1,017	,512
Kurtosis		,018	,992	-,223	,992	,231	,992	,992

FIG. B.35 – Statistiques descriptives

Pour ce modèle on a trois variables explicatives, mais seulement 20 données. Or, si on regarde la table des coefficients de ce modèle (figure B.36), on voit que la variable `nb_pers` n'est pas significative, et l'est moins que la variable `scolarite` (la cote-*t* de `nb_pers` est de 0,626 alors que celle de `scolarite` est de -0,976). Il serait donc envisageable de retirer la variable `nb_pers` du modèle.

Pour voir si cette solution est la meilleure, comparons les $r^2_{\text{ajusté}}$ des modèles suivants :

$$Y_{\text{solde}} = \beta_0 + \beta_1 X_{\text{revenu}} + \beta_2 X_{\text{scolarite}} + \beta_3 X_{\text{nb_pers}} + \epsilon$$

$$Y_{\text{solde}} = \beta_0 + \beta_1 X_{\text{revenu}} + \beta_2 X_{\text{scolarite}} + \epsilon$$

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	39,427	3543,237		,011	,991	
	Niveau de scolarité du «chef» de famille en années.	-396,928	406,714	-,258	-,976	,344	,316 3,169
	Revenu mensuel.	2,433	,735	,936	3,310	,004	,275 3,640
	Nombre de personnes dans le ménage.	341,335	545,448	,110	,626	,540	,714 1,400

a. Dependent Variable: Total des soldes.

FIG. B.36 – Table des coefficients du modèle avec les 3 variables explicatives

$$Y_{\text{solde}} = \beta_0 + \beta_1 X_{\text{revenu}} + \beta_2 X_{\text{nb_pers}} + \epsilon$$

On trouve le $r^2_{\text{ajusté}}$ du premier modèle dans la figure B.37, il a une valeur de 0,582.

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,805 ^a	,648	,582	2149,900

a. Predictors: (Constant), Nombre de personnes dans le ménage., Niveau de scolarité du «chef» de famille en années., Revenu mensuel.

b. Dependent Variable: Total des soldes.

FIG. B.37 – $r^2_{\text{ajusté}}$ du modèle avec les trois variables explicatives

On trouve le $r^2_{\text{ajusté}}$ du deuxième modèle dans la figure B.38, il a une valeur de 0,597.

Finalement, le $r^2_{\text{ajusté}}$ du troisième modèle est dans la figure B.39, il a une valeur de 0,584. On voit donc que c'est le deuxième modèle qui est le plus performant, ce qui confirme que l'idée de retirer la variable nb_pers est bonne. De cette façon on a un meilleur modèle et on respecte la condition d'une dizaine de données par variable explicative.

Faisons donc l'analyse complète du modèle

$$Y_{\text{solde}} = \beta_0 + \beta_1 X_{\text{revenu}} + \beta_2 X_{\text{scolarite}} + \epsilon.$$

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,800 ^a	,640	,597	2111,079

a. Predictors: (Constant), Revenu mensuel., Niveau de scolarité du «chef» de famille en années.

b. Dependent Variable: Total des soldes.

FIG. B.38 – $r^2_{\text{ajusté}}$ du modèle sans la variable nb_pers

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,792 ^a	,627	,584	2146,891

a. Predictors: (Constant), Nombre de personnes dans le ménage., Revenu mensuel.

b. Dependent Variable: Total des soldes.

FIG. B.39 – $r^2_{\text{ajusté}}$ du modèle sans la variable scolarite

Fixons tous les seuils à $\alpha = 0,05$. On doit d'abord vérifier les hypothèses concernant les résidus. On doit résoudre le test suivant :

H_0 : Au niveau de la population, les résidus se distribuent selon une loi normale.

H_1 : Au niveau de la population, les résidus ne se distribuent pas selon une loi normale.

Les p -values de Kolmogorov-Smirnov et Shapiro-Wilk étant respectivement de 0,200 et 0,735 (figure B.40), on ne rejette pas H_0 au seuil $\alpha = 0,05$. Ainsi on admet que les résidus suivent une loi normale.

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	,092	20	,200*	,969	20	,735

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. B.40 – Test de normalité des résidus

En jetant un coup d'œil au graphe de la figure B.41, on voit que la répartition des résidus est assez uniforme, il ne semble pas il y avoir de problème de ce côté. Il n'y a pas non plus de résidus qui se détachent vraiment des autres (*outlier*). Ainsi les hypothèses concernant les résidus étant respectées, on peut poursuivre l'analyse.

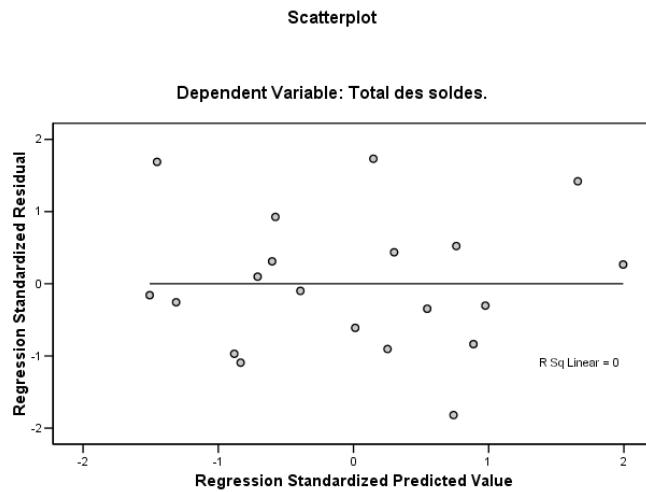


FIG. B.41 – Répartition des résidus

On peut maintenant qualifier le modèle dans son ensemble. On sait déjà que $r_{\text{ajusté}}^2 = 0,597$. Ainsi 59,7 % de la variation de la variable **soldé** est expliquée par le modèle, ce qui est bon.

Pour voir si le modèle est significatif dans son ensemble, on résout le test suivant :

H_0 : La régression est non significative dans la population (tous les $\beta_j = 0$).

H_1 : La régression est significative dans la population (au moins un des $\beta_j \neq 0$).

Puisque la p -value de la table ANOVA (figure B.42) est de $0,000 > 0,05$, on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet que la régression est significative.

On peut maintenant examiner chacun des paramètres du modèle. Les VIF étant inférieurs à 10 (ils ont une valeur de 2,604), on n'a pas de problème de multicolinéarité, on peut donc tester si les paramètres sont significatifs ou non. La p -value associée à la

ANOVA ^b					
Model		Sum of Squares	df	Mean Square	F
1	Regression	1,35E+08	2	67282612,39	15,097
	Residual	75763159	17	4456656,410	
	Total	2,10E+08	19		

a. Predictors: (Constant), Niveau de scolarité du «chef» de famille en années., Revenu mensuel.

b. Dependent Variable: Total des soldes.

FIG. B.42 – Table ANOVA de la régression

variable **revenu** étant de 0,000 ($< 0,05$) (figure B.43), on peut conclure que $\beta_1 \neq 0$. Par contre, la *p*-value associée à la variable **scolarité** étant de 0,181 ($> 0,05$), on conclut que $\beta_2 = 0$. Ainsi seul l'apport d'information de la variable **revenu** est jugé significatif (en présence de l'autre variable).

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1466,269	2663,183	,551	,589		
	Revenu mensuel.	2,678	,610	4,387	,000	,384	2,604
	Niveau de scolarité du «chef» de famille en années.	-504,390	362,027	-,327	-1,393	,181	,384

a. Dependent Variable: Total des soldes.

FIG. B.43 – Tableau des coefficients

L'équation de la régression s'écrit

$$\hat{y}_{\text{solde}} = 1466,269 + 2,678x_{\text{revenu}} - 504,39x_{\text{scolarité}}.$$

La valeur du b_0 nous donne l'estimation du solde lorsque le revenu et la scolarité sont nuls. Ainsi une personne sans revenu et sans scolarité aurait un solde de 1 466,27 \$. Ceci ne semble pas avoir beaucoup de sens, ce qui s'explique par le fait que les minimum des deux variables sont bien au-dessus de 0.

On a $b_1 = 2,678$, ce qui veut dire que pour chaque dollar additionnel de revenu et pour un niveau de scolarité fixe, le solde augmente en moyenne de 2,68 \$.

On a $b_2 = -504,39$, ce qui veut dire que pour chaque année additionnelle de scolarité et pour un revenu fixe, le solde diminue en moyenne de 504,39 \$.

Remarque Il semble étonnant de conclure que $\beta_2 = 0$ alors que $b_2 = -504,39$. Ceci est dû au fait que dans l'échantillon il y a peu de variation dans le nombre d'années de scolarité, ce qui fait qu'au bout du compte cette variable influence beaucoup moins le solde que ne le fait le revenu de la personne, variable pour laquelle il y a plus de variation.

B.8 Solutions de certains exercices du chapitre 10

Exercice 2

On s'intéresse à la relation (`spat`, `percept`, `visuo`, `sexe`) \Rightarrow `nombre`. Toutes les variables étant continues sauf celle du sexe, on utilisera l'analyse en régression linéaire multiple avec des variables explicatives continues et une discrète (dichotomique).

Une brève analyse descriptive est de mise. Je vous laisse la faire à l'aide des tableaux de la figure B.44 (moyennes, CV, jeter un coup d'œil aux min-max pour voir si tout semble normal, fréquences pour la variable discrète).

		Descriptives							
		nombre		spat		percept		visuo	
		Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Mean		150,93	1,777	104,36	2,456	94,49	2,467	96,91	2,511
95% Confidence Interval for Mean	Lower Bound	147,35		99,41		89,52		91,85	
	Upper Bound	154,52		109,30		99,46		101,97	
5% Trimmed Mean		150,78		104,21		94,17		96,70	
Median		149,00		102,00		94,00		97,00	
Variance		142,155		271,371		273,983		283,674	
Std. Deviation		11,923		16,473		16,552		16,843	
Minimum		129		76		65		68	
Maximum		175		135		129		130	
Range		46		59		64		62	
Interquartile Range		17		27		20		26	
Skewness		,192	,354	,123	,354	,261	,354	,238	,354
Kurtosis		-,597	,695	-,975	,695	-,398	,695	-,892	,695

sexe				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid masculin	21	46,7	46,7	46,7
féminin	24	53,3	53,3	100,0
Total	45	100,0	100,0	

FIG. B.44 – Statistiques descriptives

Le modèle qui nous intéresse est le suivant :

$$Y_{\text{nombre}} = \beta_0 + \beta_1 x_{\text{spat}} + \beta_2 x_{\text{percept}} + \beta_3 x_{\text{visuo}} + \beta_4 x_{\text{sexe}} + \epsilon.$$

Puisqu'on a 4 variables explicatives, l'échantillon devrait contenir au minimum 40 données, ce qui est le cas (il y en a 45).

On doit maintenant vérifier les hypothèses de validité relatives aux résidus. Fixons tous les seuils à $\alpha = 0,05$. On doit résoudre le test suivant :

H_0 : Au niveau de la population, les résidus se distribuent selon une loi normale.

H_1 : Au niveau de la population, les résidus ne se distribuent pas selon une loi normale.

Les p -values de Kolmogorov-Smirnov et Shapiro-Wilk étant respectivement de 0,200 et 0,128 (figure B.45), on ne rejette pas H_0 au seuil $\alpha = 0,05$. Ainsi on admet que les résidus suivent une loi normale.

	Tests of Normality			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ZRE_1	,096	45	,200*	,961	45	,128

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. B.45 – Test de normalité des résidus

En jetant un coup d'œil au graphe de la figure B.46, on voit que la répartition des résidus est assez uniforme, il ne semble pas il y avoir de problème de ce côté. Il n'y a pas non plus de résidus qui se détachent vraiment des autres (*outlier*). Ainsi les hypothèses concernant les résidus étant respectées, on peut poursuivre l'analyse.

On peut maintenant qualifier le modèle dans son ensemble. On voit que $r_{\text{ajusté}}^2 = 0,929$ (figure B.47). Ainsi 92,9 % de la variation de la variable **nombre** est expliquée par le modèle, ce qui plus qu'excellent !

Pour voir si le modèle est significatif dans son ensemble, on résout le test suivant :

H_0 : La régression est non significative dans la population (tous les $\beta_j = 0$).

H_1 : La régression est significative dans la population (au moins un des $\beta_j \neq 0$).

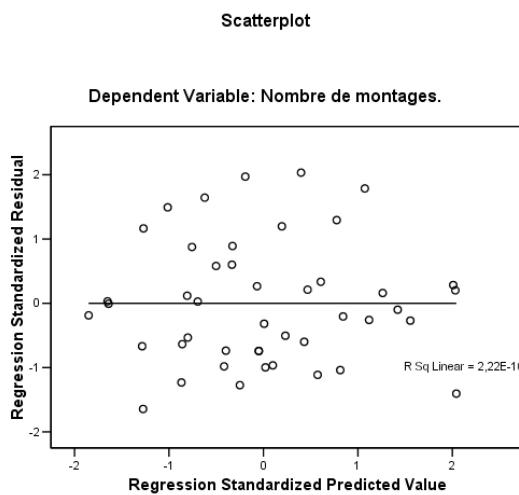


FIG. B.46 – Répartition des résidus

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,967 ^a	,935	,929	3,186

a. Predictors: (Constant), sexe, percept, visuo, spat
b. Dependent Variable: nombre

FIG. B.47 – $r^2_{\text{ajusté}}$ du modèle

Puisque la p -value de la table ANOVA (figure B.48) est de $0,000 < 0,05$, on rejette H_0 . Ainsi au risque de se tromper une fois sur 20 on admet que la régression est significative.

On peut maintenant examiner chacun des paramètres du modèle à l'aide de la figure B.49. Les VIF étant inférieurs à 10 (leurs valeurs étant de 2,566, 4,575, 2,562, 1,065), on n'a pas de problème de multicolinéarité, on peut donc tester si les paramètres sont significatifs ou non. Les p -values associées aux variables **spat**, **percept**, **visuo** étant de 0,000 ($< 0,05$), on peut conclure que $\beta_1 \neq 0$, $\beta_2 \neq 0$, $\beta_3 \neq 0$. Par contre, la p -value associée à la variable **sexe** étant de 0,506 ($> 0,05$), on conclut que $\beta_4 = 0$. Ainsi l'apport d'information de la variable **sexe** est jugé non significatif (en présence des autres variables).

ANOVA ^b					
Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	5848,692	4	1462,173	144,018	,000 ^a
Residual	406,108	40	10,153		
Total	6254,800	44			

a. Predictors: (Constant), sexe, percept, visuo, spat

b. Dependent Variable: nombre

FIG. B.48 – Table ANOVA de la régression

Parmi les variables jugées significatives, on voit que celle qui a le plus d'impact sur le nombre de montages est la variable **spat** (cote-*t* = 7,099). Vient ensuite la variable **visuo** avec une cote-*t* de 4,601, et finalement la variable **percept** avec une cote-*t* de 3,945.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	72,480	3,609		20,081	,000		
spat	,332	,047	,458	7,099	,000	,390	2,566
percept	,245	,062	,340	3,945	,000	,219	4,575
visuo	,210	,046	,297	4,601	,000	,390	2,562
sexe	,660	,983	,028	,672	,506	,939	1,065

a. Dependent Variable: nombre

FIG. B.49 – Tableau des coefficients

L'équation de la régression s'écrit

$$\hat{y}_{\text{nombre}} = 72,48 + 0,332x_{\text{spat}} + 0,245x_{\text{percept}} + 0,21x_{\text{visuo}} + 0,66x_{\text{sexe}}.$$

On peut tirer de cette équation une équation pour les hommes :

$$\hat{y}_{\text{nombre}} = 72,48 + 0,332x_{\text{spat}} + 0,245x_{\text{percept}} + 0,21x_{\text{visuo}}$$

et une autre pour les femmes :

$$\hat{y}_{\text{nombre}} = 73,14 + 0,332x_{\text{spat}} + 0,245x_{\text{percept}} + 0,21x_{\text{visuo}}.$$

On voit donc qu'effectivement il y a très peu de différence entre les hommes et les femmes, l'écart n'étant que de 0,66 montage... (Ceci est en fait l'interprétation du b_4 .)

Les b_0 nous donnent l'estimation du nombre de montages lorsque les résultats aux trois tests sont nuls (ce qui n'est pas nécessairement une bonne estimation puisqu'aucun individu n'a eu 0 à ces tests). Ainsi un tel individu ferait environ 72 montages s'il était un homme, et environ 73 montages s'il était une femme.

On a $b_1 = 0,332$, ce qui veut dire que pour chaque point de plus au test de spatialisation (lorsque les autres résultats sont fixes), le nombre de montages augmente en moyenne de 0,332 (et ce peu importe si l'individu est un homme ou une femme). L'interprétation de b_2 et b_3 est similaire.

On peut maintenant répondre à la question. Premièrement, on a pu voir dans le premier tableau de la figure B.44 que la moyenne du nombre de montages est de 150,93. Il suffit donc de faire des prédictions avec les données du tableau de l'énoncé et notre équation. On obtient les données de la figure B.50.

	id	nombre	spat	percept	visuo	sexe	PRE_1
43	43	132	78	70	75	féminin	131,68380
44	44	151	102	95	98	masculin	150,52160
45	45	175	131	127	127	féminin	173,98511
46	.	.	110	85	78	féminin	146,48003
47	.	.	104	102	98	masculin	152,93972
48	.	.	112	96	101	féminin	154,63770
AC							

FIG. B.50 – Prédictions

Ainsi les candidatures des individus B et C sont intéressantes puisque les nombres prévus de montages (152,9 et 154,6) sont supérieurs à la moyenne. La candidature de l'individu A ne l'est pas car le nombre prévu de montage est de 146,5, ce qui est en-dessous de la moyenne.

Exercice 3

Note : Les résultats de cet exercice ont été obtenus en filtrant l'individu no 43 (pour suivre ce qui avait été fait dans l'exemple des notes : cet individu n'a que 10 mois d'ancienneté et a le plus haut salaire).

On s'intéresse à la relation (`ancien`, `fonction`, `sexe`) \Rightarrow `salaire`. On utilisera l'analyse en régression linéaire multiple avec une variable explicative continue et des variables discrètes (une dichotomique et une polychotomique).

Une brève analyse descriptive est de mise. Je vous laisse la faire à l'aide des tableaux de la figure B.51 (moyennes, CV, jeter un coup d'œil aux min-max pour voir si tout semble normal, fréquences pour les variables discrètes).

Descriptives			
Mean	133,6583	5,40266	579,1658
95% Confidence Interval for Mean	Lower Bound	123,0042	563,3196
	Upper Bound	144,3124	595,0120
5% Trimmed Mean		132,3291	579,6393
Median		135,0000	586,0000
Variance		5808,559	12849,402
Std. Deviation		76,21391	113,35520
Minimum		2,00	271,00
Maximum		329,00	867,00
Range		327,00	596,00
Interquartile Range		123,00	166,00
Skewness		,141	-,068
Kurtosis		-,712	,172
		,343	-,059
			,343

sexe				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid masculin	165	82,9	82,9	82,9
féminin	34	17,1	17,1	100,0
Total	199	100,0	100,0	

fonction				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid administration	31	15,6	15,6	15,6
production	129	64,8	64,8	80,4
direction	39	19,6	19,6	100,0
Total	199	100,0	100,0	

FIG. B.51 – Statistiques descriptives

En regardant la table de la figure B.52, on voit que si l'on considère le modèle multiplicatif, les *p*-values associées aux variables multiplicatives *anciensex*, *ancienprod* et *ancienadmin* sont respectivement de 0,940, 0,152 et 0,535. Ainsi peu importe le seuil fixé, ces variables sont jugées non significatives, ce qui signifie qu'il n'y a pas d'interaction entre la variable *ancien* et les variables discrètes. Il n'est donc pas pertinent de considérer le modèle multiplicatif, et ainsi l'analyse se fera avec le modèle additif.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	603,063	30,372		,000		
	sexe	-24,263	27,527	-,081	,379	,279	3,585
	ancien	,475	,150	,319	,002	,230	4,350
	anciensexé	,015	,197	,007	,940	,270	3,707
	admin	-193,937	38,383	-,822	,000	,155	6,470
	production	-120,034	33,802	-,507	,000	,115	8,700
	ancienprod	,263	,183	,192	,152	,131	7,625
	ancienadmin	,172	,276	,058	,622	,272	3,679

a. Dependent Variable: salaire

FIG. B.52 – Tableau des coefficients du modèle multiplicatif

Le modèle qui nous intéresse est donc le suivant :

$$Y_{\text{nombre}} = \beta_0 + \beta_1 x_{\text{sexe}} + \beta_2 x_{\text{ancien}} + \beta_3 x_{\text{admin}} + \beta_4 x_{\text{production}} + \epsilon.$$

Puisqu'on a 4 variables explicatives, l'échantillon devrait contenir au minimum 40 données, ce qui est le cas (il y en a 199).

On doit maintenant vérifier les hypothèses de validité relatives aux résidus. Fixons tous les seuils à $\alpha = 0,05$. On doit résoudre le test suivant :

H_0 : Au niveau de la population, les résidus se distribuent selon une loi normale.

H_1 : Au niveau de la population, les résidus ne se distribuent pas selon une loi normale.

Les p -values de Kolmogorov-Smirnov et Shapiro-Wilk étant respectivement de 0,200 et 0,404 (figure B.53), on ne rejette pas H_0 au seuil $\alpha = 0,05$. Ainsi on admet que les résidus suivent une loi normale.

En jetant un coup d'œil au graphe de la figure B.54, on voit que la répartition des résidus est assez uniforme, il ne semble pas il y avoir de problème de ce côté. Il y a par contre un résidu très proche de -3 écarts-types. L'individu correspondant à ce résidu (no

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	,041	199	,200*	,993	199	,404

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

FIG. B.53 – Test de normalité des résidus

149) a un salaire peu élevé par rapport à son ancienneté et sa fonction (direction). Mais puisqu'il est quand même dans les ± 3 écarts-types, nous conservons cette donnée.

Ainsi les hypothèses concernant les résidus étant respectées, on peut poursuivre l'analyse.

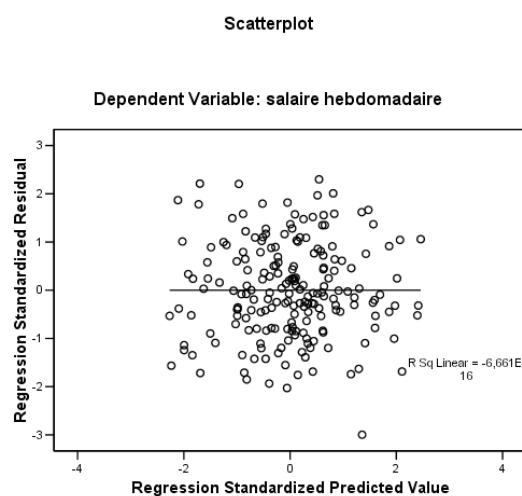


FIG. B.54 – Répartition des résidus

On peut maintenant qualifier le modèle dans son ensemble. On voit que $r_{\text{ajusté}}^2 = 0,538$ (figure B.55). Ainsi 53,8 % de la variation de la variable **salaire** est expliquée par le modèle, ce qui est mieux que le 39,7 % obtenu avec le modèle de l'exemple des notes (c'est-à-dire sans la variable de la fonction).

Pour voir si le modèle est significatif dans son ensemble, on résout le test suivant :

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,740 ^a	,548	,538	77,02725

a. Predictors: (Constant), production, sexe, ancien, admin
b. Dependent Variable: salaire

FIG. B.55 – $r^2_{\text{ajusté}}$ du modèle

H_0 : La régression est non significative dans la population (tous les $\beta_j = 0$).

H_1 : La régression est significative dans la population (au moins un des $\beta_j \neq 0$).

Puisque la p -value de la table ANOVA (figure B.56) est de $0,000 < 0,05$, on rejette H_0 .

Ainsi au risque de se tromper une fois sur 20 on admet que la régression est significative.

ANOVA ^b					
Model		Sum of Squares	df	Mean Square	F
1	Regression	1393141	4	348285,295	58,701
	Residual	1151040	194	5933,198	
	Total	2544182	198		

a. Predictors: (Constant), production, sexe, ancien, admin

b. Dependent Variable: salaire

FIG. B.56 – Table ANOVA de la régression

On peut maintenant examiner chacun des paramètres du modèle à l'aide de la figure B.57. Les VIF étant inférieurs à 10 (leurs valeurs étant de 1,041, 1,2, 1,859, 1,657), on n'a pas de problème de multicolinéarité, on peut donc tester si les paramètres sont significatifs ou non. La p -value associée à la variable `sexe` étant de 0,149 ($> 0,05$), on peut conclure que l'apport d'information de la variable `sexe` dans ce modèle n'est pas significatif. Par contre, les p -values associées aux variables `ancien`, `admin`, `production` étant de 0,000 ($< 0,05$), on conclut que $\beta_2 \neq 0$, $\beta_3 \neq 0$ et $\beta_4 \neq 0$. Ainsi l'apport d'information de la variable `ancien` et de la variable polychotomique de la fonction est jugé significatif.

Dans l'autre modèle (avec les variables `ancien` et `sexé`), la variable `sexé` était significative, et on avait comme conclusion que les femmes gagnent moins que les hommes. Le fait que la variable `sexé` n'est plus significative dans ce modèle s'explique peut-être par le fait qu'il y a plus de femmes pour les fonctions moins bien payées (et vice-versa).

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	569,517	19,109	29,803	,000		
	sexé	-21,471	14,805	-,071	-,1450	,149	,960
	ancien	,656	,079	,441	8,340	,000	,833
	admin	-161,720	20,528	-,519	-7,878	,000	,538
	production	-75,916	14,719	-,321	-5,158	,000	,604

a. Dependent Variable: salaire

FIG. B.57 – Tableau des coefficients

L'équation de la régression s'écrit

$$\hat{y}_{\text{salaire}} = 569,517 - 21,471x_{\text{sexé}} + 0,656x_{\text{ancien}} - 161,720x_{\text{admin}} - 75,916x_{\text{production}}.$$

On peut tirer de cette équation les équations suivantes :

homme, direction	$\hat{y}_{\text{salaire}} = 569,517 + 0,656x_{\text{ancien}}$
femme, direction	$\hat{y}_{\text{salaire}} = 548,046 + 0,656x_{\text{ancien}}$
homme, administration	$\hat{y}_{\text{salaire}} = 407,797 + 0,656x_{\text{ancien}}$
femme, administration	$\hat{y}_{\text{salaire}} = 386,326 + 0,656x_{\text{ancien}}$
homme, production	$\hat{y}_{\text{salaire}} = 493,601 + 0,656x_{\text{ancien}}$
femme, production	$\hat{y}_{\text{salaire}} = 472,13 + 0,656x_{\text{ancien}}$

Ici, les b_0 peuvent s'interpréter pour chacune des équations : ils représentent le salaire lorsque l'ancienneté est à 0. Par exemple, le salaire hebdomadaire d'une femme sans ancienneté à un poste administratif serait d'environ 386,33 \$.

Le b_0 de l'équation générale représente le salaire hebdomadaire d'un homme sans ancienneté à un poste de direction (569,52 \$).

On a $b_2 = 0,656$, ce qui indique que pour une année d'ancienneté de plus, on ajoute 0,66 \$ au salaire hebdomadaire (!) (lorsque les autres variables sont fixes).

On a $b_1 = -21,471$, ce qui signifie que pour une même ancienneté et même fonction, les femmes gagnent en moyenne 21,47 \$ de moins que les hommes.

On a $b_3 = -161,72$, ce qui signifie que pour une même ancienneté et même sexe, un individu qui a un poste administratif gagne en moyenne 161,72 \$ de moins qu'un individu qui a un poste de direction.

De même, puisque $b_4 = -75,916$, on a que pour une même ancienneté et même sexe, un individu qui a un poste de production gagne en moyenne 75,92 \$ de moins qu'un individu qui a un poste de direction.

B.9 Solutions des exercices du chapitre 15

Exercice 1

On fait ici une ACP à partir de 14 variables concernant les services utilisés par les clients d'un fournisseur de services de télécommunications. Il est à noter que certaines des variables sont dichotomiques (on peut les prendre dans l'ACP car elles sont correctement codées de façon binaire).

La première sortie de la figure B.58 nous donne le KMO de cette ACP : il est de 0,888, ce qui nous indique que l'ACP sera méritoire.

KMO and Bartlett's Test			Communalities	
			Initial	Extraction
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,888		
Bartlett's Test of Sphericity	Approx. Chi-Square	6230,901		
	df	91		
	Sig.	,000		
Extraction Method: Principal Component Analysis				

FIG. B.58 – Le KMO et les communalités

Cette prédiction semble confirmée par les *communalities* de la deuxième sortie de la figure B.58. En effet, celles-ci sont assez élevées, la plus basse étant à 0,527 pour `cartem` (ce qui signifie que cette variable est expliquée à 52,7 % par les facteurs).

Dans le tableau de la figure B.59, on voit qu'il y a trois valeurs propres plus grandes que 1, ce qui fait que l'on a 3 facteurs, qui expliquent 64,903 % de la variance totale des variables (colonne Cumulative %).

Component	Total Variance Explained								
	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,877	34,838	34,838	4,877	34,838	34,838	4,136	29,543	29,543
2	2,733	19,518	54,357	2,733	19,518	54,357	3,304	23,599	53,142
3	1,476	10,546	64,903	1,476	10,546	64,903	1,647	11,761	64,903
4	,697	4,979	69,883						
5	,612	4,368	74,251						
6	,508	3,628	77,879						
7	,470	3,358	81,237						
8	,461	3,289	84,526						
9	,423	3,021	87,548						
10	,418	2,988	90,536						
11	,377	2,695	93,231						
12	,350	2,503	95,733						
13	,339	2,424	98,158						
14	,258	1,842	100,000						

Extraction Method: Principal Component Analysis.

FIG. B.59 – Les valeurs propres

Après rotation, on voit que le premier facteur explique 29,543 % de la variance totale, le deuxième en explique 23,599 % et le dernier 11,761 %.

On peut maintenant examiner la matrice des facteurs (après rotation) de la figure B.60. On voit d'abord que le premier facteur est fortement corrélé avec les variables **affich**, **apellatt**, **transfert**, **conf** et **sfraism**. On pourrait donc dire que ce facteur est relié aux services qui sont pris en « extra ».

Le deuxième facteur est fortement corrélé avec les variables **equipm**, **internet**, **efact**, **sansfilm**, **pagette** et **boite**. On pourrait ici parler du groupe de services techniques.

Le troisième facteur est fortement corrélé aux variables **longdistm**, **lignmult** et **cartem**. Ici les variables **longdistm**, **cartem** se rapportent aux longues distances, et **lignmult** est un peu à part. On parlera donc du facteur des longues distances, même si ce n'est pas parfait.

	Rotated Component Matrix		
	1	2	3
Afficheur	,825	,056	,002
Appel en attente	,817	,031	,052
Transfert d'appel	,810	,060	,045
Appel conférence	,790	,045	,078
Numéro sans frais			
dernier mois	,768	,006	,210
Équipement dernier mois	,061	,853	,061
Internet	-,069	,793	-,048
Facturation électronique	-,133	,768	-,092
Sans fil dernier mois	,534	,649	,164
Service de pagette	,489	,609	,045
Boîte vocale	,478	,587	,061
Longue distance dernier mois	,052	-,145	,855
Lignes multiples	-,092	,435	,647
Carte d'appel dernier mois	,360	-,048	,629

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

FIG. B.60 – La matrice des facteurs

À partir de ceci, que peut-on dire de l'usage des services par les abonnés ? Quels sont les liens entre les facteurs, et que peut-on en tirer ?

D'abord, on voit que les variables `sansfilm`, `boite` et `pagette` sont quand même assez corrélées avec les facteurs 1 et 2, ce qui établit un lien entre ces deux facteurs (les extras et les techniques).

On voit aussi que la variable `lignmult` est assez corrélée avec les facteurs 2 et 3 (techniques et longues distances).

Ces liens suggèrent des idées pour les ventes. Par exemple, les consommateurs qui utilisent les services « extras » seraient sûrement mieux prédisposés à accepter une offre spéciale sur les services sans fil (`sansfilm`) plutôt qu'une offre sur les services internet.

Exercice 2

On fait ici l'ACP avec les 9 variables de la participation. On obtient tout d'abord le KMO (figure B.61). Il a une valeur de 0,742, ce qui indique que l'ACP sera moyenne.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy:		,742
Bartlett's Test of Sphericity	Approx. Chi-Square df Sig.	2657,086 36 ,000

FIG. B.61 – Le KMO

Pourtant les *communalities* (figure B.62) sont assez bonnes, la plus basse étant à 0,552 pour la variable `part_q3`. La plus élevée est à 0,822 ;, ainsi la variable `part_q1` sera expliquée à 82,2 % par les facteurs.

Communalities		
	Initial	Extraction
<code>part_q1</code>	1,000	,822
<code>part_q2</code>	1,000	,707
<code>part_q3</code>	1,000	,552
<code>part_q4</code>	1,000	,584
<code>part_q5</code>	1,000	,713
<code>part_q6</code>	1,000	,821
<code>part_q7</code>	1,000	,735
<code>part_q8</code>	1,000	,793
<code>part_q9</code>	1,000	,717

Extraction Method: Principal Component Analysis

FIG. B.62 – Les communalités

On voit dans la figure B.63 qu'il y a 3 valeurs propres qui ont une valeur supérieure à 1, et donc on aura 3 facteurs. Ces trois facteurs expliquent 71,595 % de la variance totale des variables.

Après rotation, on voit que le premier facteur explique 28,501 % de la variance totale, le deuxième en explique 22,303 % et le dernier 20,79 %.

On peut maintenant examiner la matrice des facteurs (après rotation) de la figure

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,571	39,682	39,682	3,571	39,682	39,682	2,565	28,501	28,501
2	1,841	20,459	60,141	1,841	20,459	60,141	2,007	22,303	50,805
3	1,031	11,454	71,595	1,031	11,454	71,595	1,871	20,790	71,595
4	,827	9,189	80,784						
5	,432	4,802	85,586						
6	,405	4,497	90,083						
7	,378	4,195	94,278						
8	,272	3,020	97,298						
9	,243	2,702	100,000						

Extraction Method: Principal Component Analysis.

FIG. B.63 – Les valeurs propres

B.64. On voit d'abord que le premier facteur est fortement corrélé avec les trois dernières variables.

	Rotated Component Matrix		
	1	2	3
part_q8	,877	,053	,147
part_q7	,852	,086	,047
part_q9	,835	,030	,134
part_q6	,080	,902	-,024
part_q5	-,001	,792	,291
part_q3	,394	,533	,335
part_q1	,207	-,023	,883
part_q2	,385	,304	,683
part_q4	-,142	,422	,621

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

FIG. B.64 – La matrice des facteurs

Le deuxième facteur est fortement corrélé avec la 3e, la 5e et la 6e variable, tandis que le troisième est fortement corrélé avec les deux premières et la quatrième variable.

On voit donc que les groupements obtenus ne sont pas ceux auxquels on se serait attendu intuitivement. De plus, les cohérences internes des regroupements intuitifs sont meilleures que celles des groupements suggérés par l'ACP.

B.10 Solutions de certains exercices du chapitre 17

Exercice 1

On veut ici classifier les employés selon leur satisfaction. Il est à noter que j'ai changé le nom des variables de la satisfaction afin qu'ils aient une certaine signification pour faciliter l'interprétation des tableaux. J'ai laissé le numéro de la question pour bien faire le lien avec les noms originaux.

On sélectionne d'abord 50 individus de façon aléatoire et on applique la méthode hiérarchique pour se donner une idée de combien de groupes (*clusters*) on veut créer. On obtient ainsi la figure B.65 qui contient la table d'agglomération.

On voit dans celle-ci que pour les coefficients de distance il y a un bon écart entre 5 et 4 clusters (on passe de 36,848 à 40,363). On retient donc que la solution à 5 clusters semble bien, et l'on va essayer avec 4, 5 et 6 clusters avec la méthode des nuées dynamiques.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	388	391	1,681	0	0	5
2	336	370	2,364	0	0	5
3	171	181	4,121	0	0	10
4	341	431	4,230	0	0	7
5	336	388	5,007	2	1	13
6	173	231	5,925	0	0	16
7	287	341	6,444	0	4	13
8	671	673	6,576	0	0	15
9	156	159	6,783	0	0	27
10	171	246	6,813	3	0	16
11	337	470	6,934	0	0	17
12	251	333	7,264	0	0	22
13	287	336	7,922	7	5	19
14	631	657	7,948	0	0	15
15	631	671	9,930	14	8	31
16	171	173	10,264	10	6	20
17	337	414	10,344	11	0	19
18	174	180	10,875	0	0	36
19	287	337	11,058	13	17	20
20	171	287	12,145	16	19	24
21	117	126	13,162	0	0	23
22	251	256	13,196	12	0	24
23	117	275	13,623	21	0	32
24	171	251	13,952	20	22	30
25	29	97	14,171	0	0	27
26	9	32	14,543	0	0	40
27	29	156	15,101	25	9	33
28	79	209	15,251	0	0	35
29	693	710	15,733	0	0	41
30	171	499	15,931	24	0	32
31	570	631	16,297	0	15	34
32	117	171	18,918	23	30	35
33	29	145	19,107	27	0	37
34	570	620	19,279	31	0	39
35	79	117	20,456	28	32	36
36	79	174	22,544	35	18	37
37	29	79	23,136	33	36	42
38	72	138	23,462	0	0	42
39	570	601	24,348	34	0	41
40	9	146	25,565	26	0	43
41	570	693	29,351	39	29	49
42	29	72	29,974	37	38	43
43	9	29	31,277	40	42	45
44	12	20	35,129	0	0	48
45	9	425	36,848	43	0	46
46	9	175	40,363	45	0	47
47	9	178	41,536	46	0	48
48	9	12	46,187	47	44	49
49	9	570	59,355	48	41	0

FIG. B.65 – La table d'agglomération

Pour les trois solutions les répartitions sont acceptables, les tables ANOVA aussi. Il faudra donc prendre la décision uniquement sur l'interprétation.

	Final Cluster Centers				
	1	2	3	4	5
Zdistouvr1	-,20289	-,53254	-,92426	,96811	,40135
Ztravint2	-,19122	-,59609	-1,11768	1,19047	,38917
Zorgtrav3	-,19166	-,62952	-1,02192	,98204	,47996
Ztravordre4	-,01365	-,64726	-,57765	,84288	,12476
Zchavan5	-,14835	-,70414	-,95721	1,14493	,33021
Zinfosup6	-,47927	,17999	-1,10815	1,18604	,34816
Zcomempl7	-,41507	,10067	-,88220	1,08315	,25023
Zcondphys8	-,34791	,12034	-,81084	,90581	,23369
Zresp9	-,32579	-,44267	-1,11219	1,26607	,42040
Zfactrav10	-,28792	-,27707	-1,20279	1,22089	,37731
Zestime11	-,32959	,02095	-1,12500	1,13002	,29672
Zdepadm12	-,47308	,03602	-,93038	1,27353	,25672
Zpaye13	-,34399	-,07598	-,71416	1,09723	,15084
Zsecemploi14	,01258	-,92676	-,65165	,86574	,25373
Zacctrav15	-,37390	,22377	-,81665	1,10458	,09595
Ztententesup16	-,50587	,39308	-1,08457	1,12264	,30373
Zcompsup17	-,60555	,37792	-,91586	1,05736	,37048
Ztravarie18	-,24304	-,54152	-1,00111	1,18173	,36300

FIG. B.66 – Les moyennes finales avec 5 clusters

Dans les solutions à 5 et 6 clusters, on voit certaines nuances d'un cluster à l'autre, mais les différences entre certains clusters ne me semblent pas beaucoup marquées (voir les sorties des figures B.66 et B.67). Donc interpréter ces solutions revient à donner du poids à de petites différences pour distinguer les clusters (mais ce serait quand même des solutions acceptables, et même préférables dans le cas où on veut vraiment aller cerner la moindre petite différence ; tout dépend du contexte). Voilà pourquoi je choisis la solution à 4 clusters, qui offre un portrait moins nuancé mais où les clusters se distinguent plus entre eux.

Final Cluster Centers

	Cluster					
	1	2	3	4	5	6
Zdistouvr1	1,00534	,42023	-,93822	-,18253	-,70264	-,25302
Ztravint2	1,23870	,46527	-1,21446	-,23013	-,47913	-,40117
Zorgtrav3	1,03515	,50426	-1,13535	-,36657	,01140	-,28812
Ztravordre4	,89044	,20146	-,67686	,06137	-,90041	-,42332
Zchavan5	1,20733	,38469	-,98729	-,12723	-,81568	-,47488
Zinfosup6	1,16252	,32483	-1,09602	-,35392	-,81940	,50053
Zcomempl7	1,07737	,19808	-,93318	-,48944	-,07109	,51934
Zcondphys8	,91755	,14328	-,93843	-,35797	-,19401	,61893
Zresp9	1,30951	,44060	-1,14146	-,38177	-,41807	-,16355
Zfactrav10	1,23302	,37803	-1,29988	-,32200	-,20959	-,03572
Zestime11	1,11894	,30102	-1,09716	-,30936	-,59473	,32407
Zdepadm12	1,26000	,26292	-,92581	-,40729	-,58629	,25807
Zpaye13	1,11631	,18000	-,66661	-,28018	-,75574	,11968
Zsecemploi14	,92722	,36049	-,73562	-,20527	,28139	-,94685
Zacctrav15	1,11549	-,03002	-,91873	-,47296	,30097	,64355
Ztententesup16	1,10674	,26445	-1,05360	-,35048	-,81820	,66083
Zcompsup17	1,04586	,32403	-,81532	-,39828	-1,00982	,59786
Ztravarie18	1,23454	,42097	-1,04474	-,26085	-,53562	-,37278

FIG. B.67 – Les moyennes finales avec 6 clusters

On peut donc procéder à l'analyse complète de la solution à 4 clusters. On voit premièrement à l'aide de la sortie B.68 que le processus a convergé complètement (les distances sont nulles) en 23 itérations. On a donc vraiment la solution optimale (pour 4 clusters).

Iteration	Change in Cluster Centers			
	1	2	3	4
1	5,257	6,513	6,541	6,470
2	,178	,437	,639	,316
3	,306	,205	,264	,399
4	,198	,137	,234	,464
5	,040	,126	,184	,317
6	,069	,148	,104	,234
7	,081	,123	,086	,145
8	,072	,150	,069	,147
9	,067	,125	,034	,127
10	,000	,103	,055	,106
11	,026	,065	,086	,091
12	,033	,071	,077	,089
13	,029	,026	,089	,052
14	,000	,036	,046	,050
15	,025	,048	,047	,050
16	,000	,021	,018	,021
17	,000	,034	,000	,027
18	,038	,022	,039	,022
19	,000	,000	,058	,033
20	,000	,000	,066	,038
21	,000	,000	,053	,031
22	,000	,013	,044	,033
23	,000	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 23. The minimum distance between initial centers is 9,688.

FIG. B.68 – Les itérations

On voit aussi d'après la table ANOVA (figure B.69) que cette solution semble bonne puisque toutes les *p*-values sont nulles. Il est toutefois bon de noter que c'était également le cas pour les deux autres solutions. Par contre, tous les *F* sauf un sont plus grands dans la solution à 4 clusters (comparativement aux *F* des solutions à 5 ou 6 clusters, mais je n'ai pas mis ici ces tables ANOVA...).

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zdistouvr1	88,635	3	,633	717	139,951	,000
Ztravint2	119,963	3	,502	717	238,852	,000
Zorgtrav3	93,823	3	,612	717	153,401	,000
Ztravordre4	45,251	3	,815	717	55,533	,000
Zchavan5	99,031	3	,590	717	167,897	,000
Zinfosup6	119,057	3	,506	717	235,272	,000
Zcomempl7	87,507	3	,638	717	137,149	,000
Zcondphys8	69,352	3	,714	717	97,131	,000
Zresp9	129,076	3	,464	717	278,111	,000
Zfactrav10	119,527	3	,504	717	237,122	,000
Zestime11	101,321	3	,580	717	174,617	,000
Zdepadm12	112,143	3	,535	717	209,626	,000
Zpaye13	77,028	3	,682	717	112,963	,000
Zsecemploi14	50,070	3	,795	717	63,007	,000
Zacctrav15	76,817	3	,683	717	112,506	,000
Zententesup16	113,384	3	,530	717	214,025	,000
Zcompsup17	104,510	3	,567	717	184,352	,000
Ztravarie18	109,733	3	,545	717	201,328	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

FIG. B.69 – Table ANOVA

L'interprétation des clusters se fait à l'aide de la sortie B.70. On voit que dans le premier groupe la moyenne pour chaque question est toujours plus élevée que les moyennes des autres groupes pour cette même question. Ainsi les individus de ce groupe sont les plus **satisfait**s. À l'intérieur du groupe on voit que la satisfaction à propos d'avoir des travailleurs sous ses ordres, les conditions de travail et la sécurité d'emploi sont un peu moins élevées. On voit selon la figure B.71 que ce cluster contient 119 individus, c'est le plus petit des quatre.

	Final Cluster Centers			
	1	2	3	4
Zdistouvr1	,96418	,35449	-,89707	-,22026
Ztravint2	1,19222	,35389	-1,00557	-,26443
Zorgtrav3	,98705	,42972	-,82930	-,33049
Ztravordre4	,84720	,06631	-,55586	-,13092
Zchavan5	1,14137	,28718	-,85276	-,27517
Zinfosup6	1,16984	,35618	-1,02519	-,24442
Zcomempl7	1,08068	,25362	-,80750	-,24581
Zcondphys8	,88966	,30675	-,74367	-,23543
Zresp9	1,26227	,38067	-,97907	-,33416
Zfactrav10	1,22039	,34024	-,96800	-,28840
Zestime11	1,12173	,29018	-,92114	-,22888
Zdepadm12	1,27097	,28860	-,79717	-,36934
Zpaye13	1,07829	,17912	-,69419	-,25031
Zsecemploi14	,86731	,16607	-,53085	-,23485
Zacctrav15	1,08273	,15103	-,71271	-,21917
Ztententesup16	1,10606	,36086	-1,03630	-,21176
Zcompsup17	1,04728	,40403	-,94930	-,26896
Ztravarie18	1,17752	,30993	-,93280	-,26424

FIG. B.70 – Les moyennes finales (4 clusters)

Dans le deuxième groupe, les individus se situent au-dessus de la moyenne, mais pas beaucoup. Ils ne démontrent pas d'insatisfaction flagrante en aucun point, on peut dire qu'ils sont **moyennement satisfait**s par rapport aux autres groupes. À l'intérieur du groupe il y a peu de variation ; on voit que pour la question 4 (avoir des travailleurs sous ses ordres) et les questions 13, 14 et 15 (paye, sécurité d'emploi et efforts pour éviter les accidents de travail) la satisfaction est évaluée un peu plus à la baisse. Ce cluster contient

203 individus.

Dans le troisième groupe se retrouvent de façon évidente les **insatisfaits** (par rapport aux autres groupes). Les points relatifs aux conditions de travail et aux travailleurs sous leurs ordres sont évalués un peu moins négativement que les autres. Ce cluster contient 146 individus.

Finalement, le quatrième groupe contient les **moyennement insatisfaits**; effectivement, ils se retrouvent en-dessous de la moyenne pour chaque question, mais toujours de façon moins négative que les individus du cluster 3. Il y a vraiment très peu de variation d'une question à l'autre à l'intérieur de ce groupe, qui est le plus gros (253 individus).

Number of Cases in each Cluster

Cluster	1	119,000
	2	203,000
	3	146,000
	4	253,000
Valid		721,000
Missing		,000

FIG. B.71 – Répartition

La répartition, dont on a déjà discuté et qui est tout à fait acceptable, se retrouve dans la figure B.71.

Il reste à interpréter les cartes perceptuelles (figures B.72 et B.73). Ici je me contente de parler des facteurs 1, 2 et 3 de la participation, que vous prendrez soin de bien nommer...;-)

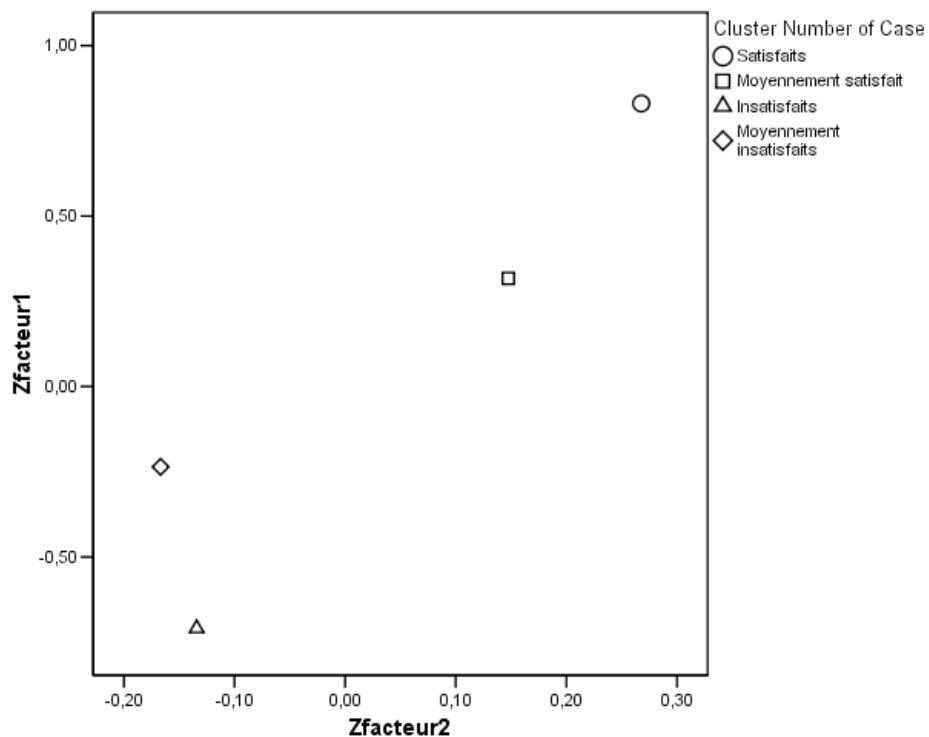


FIG. B.72 – Carte avec les facteurs 1 et 2

Cette carte nous permet de voir que par rapport aux facteurs 1 et 2, ce sont les plus satisfaits qui ont les meilleurs résultats. Suivent ensuite les moyennement satisfaits. Les insatisfaits ont la plus basse moyenne par rapport au facteur 1, mais ce sont les moyennement insatisfaits qui ont la plus basse moyenne par rapport au facteur 2.

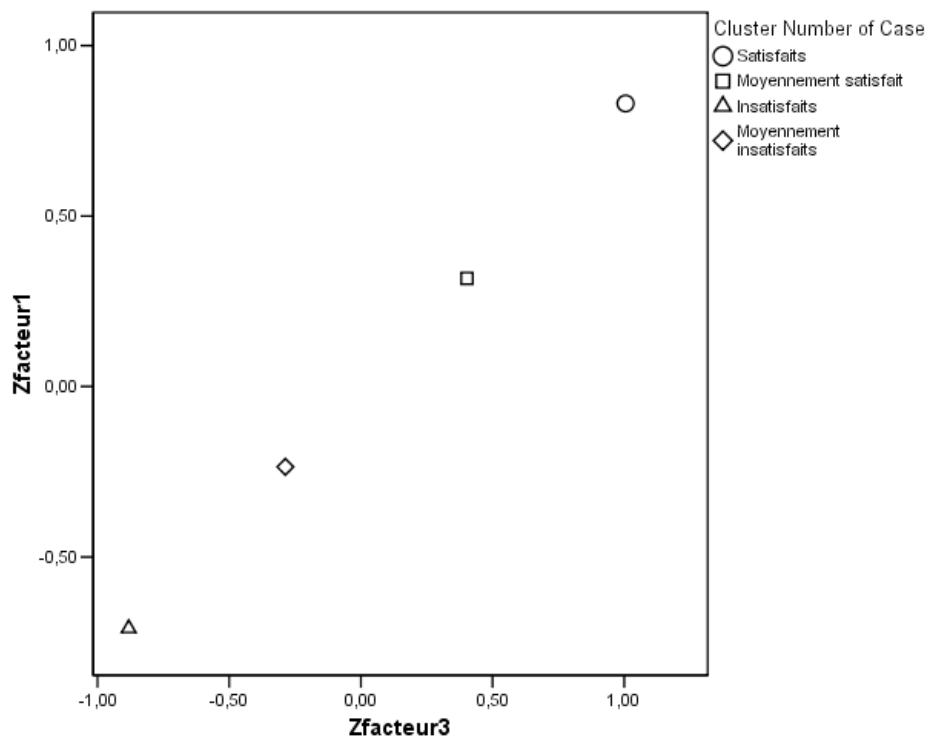


FIG. B.73 – Carte avec les facteurs 1 et 3

Cette deuxième carte nous montre que le classement des clusters par rapport aux facteurs 1 et 3 est le même : ce sont les satisfait qui ont donné les meilleurs résultats, suivis des moyennement satisfait, des moyennement insatisfait puis des insatisfait, un ordre tout à fait logique si l'on considère que la participation est reliée positivement à la satisfaction. Le fait que cet ordre ne soit pas respecté par rapport au facteur 2 (première carte) est intéressant et mèrriterait une investigation plus en profondeur...

Exercice 3

On doit d'abord faire une analyse de classification avec les 3 facteurs (scores factoriels, déjà standardisés) définis à l'exercice 1 du chapitre 15 (que j'ai nommés `extras`, `tech`, `longdist`).

Un premier essai avec la méthode hiérarchique avec un sous-échantillon de 50 individus suggère de prendre 4 ou 6 clusters (à vous de voir si vous arrivez à la même conclusion...).

Avec les nuées dynamiques, mon choix s'arrête sur la solution à 6 clusters. Pour les trois solutions les interprétations et les répartitions sont bonnes, mais pour l'analyse (tableau croisé) qui suivra, c'est avec la solution à 6 clusters que l'on arrive à mieux cerner la relation, d'où le choix.

Iteration	Change in Cluster Centers					
	1	2	3	4	5	6
1	1,011	1,201	1,265	1,059	,989	1,127
2	,140	,283	,314	,259	,099	,629
3	,060	,103	,232	,091	,060	,399
4	,057	,064	,240	,042	,041	,299
5	,050	,068	,183	,050	,031	,160
6	,071	,015	,158	,062	,055	,074
7	,050	,011	,168	,042	,067	,083
8	,026	,014	,230	,067	,145	,097
9	,031	,014	,174	,029	,154	,030
10	,025	,006	,151	,107	,195	,057
11	,054	,007	,051	,029	,073	,064
12	,018	,015	,063	,030	,077	,000
13	,019	,000	,036	,012	,046	,032
14	,000	,014	,025	,017	,032	,000
15	,006	,000	,008	,006	,014	,000
16	,006	,000	,000	,006	,007	,000
17	,000	,000	,000	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 17. The minimum distance between initial centers is 3,133.

FIG. B.74 – Les itérations

Le processus a convergé en 17 étapes, donc tout est correct de ce côté. Toutes les *p*-values de la table ANOVA sont nulles, donc la solution à 6 clusters semble donner de

bons résultats.

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
extras	155,126	5	,225	994	690,316	,000
tech	152,309	5	,239	994	637,575	,000
longdist	123,263	5	,385	994	320,169	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

FIG. B.75 – Table ANOVA

	Final Cluster Centers					
	Cluster					
	1	2	3	4	5	6
extras	-,86134	,93975	1,05411	-1,04281	-,56746	,03712
tech	-,39061	1,46752	-,74201	,93292	-,56641	-,66640
longdist	,71911	,02050	-,18941	-,21817	-,73928	2,68633

FIG. B.76 – Les moyennes finales

Pour l'interprétation, disons que le cluster 6 se démarque par l'utilisation prononcée des services de longues distances, et par une utilisation faible du groupe de services techniques.

Le cluster 5 est le seul à être dans le négatif partout, donc dans l'ensemble utilise peu les services comparativement aux autres groupes.

Le groupe 4 est celui qui utilise le moins les extras, et il est le deuxième pour l'utilisation du groupe technique.

Le groupe 3 est le groupe qui utilise le plus les extras, et utilise le moins le groupe technique.

Le groupe 2 est celui qui utilise le plus les services techniques, et est le deuxième pour les extras.

Number of Cases in each Cluster		
Cluster	1	136,000
	2	169,000
	3	245,000
	4	165,000
	5	229,000
	6	56,000
Valid		1000,000
Missing		,000

FIG. B.77 – Répartition

Finalement le groupe 1 est le deuxième pour les longues distances, et utilise peu les autres groupes.

On veut maintenant étudier la relation clusters \Rightarrow desabon. Je vous laisse faire correctement le test du chi carré et l'interprétation du Cramer's V (la relation est significative, et est qualifiée d'intéressante).

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	100,442 ^a	5	,000
Likelihood Ratio	107,930	5	,000
Linear-by-Linear Association	1,454	1	,228
N of Valid Cases	1000		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 15,34.

FIG. B.78 – Le chi carré

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	,317	,000
Nominal	Cramer's V	,317	,000
N of Valid Cases		1000	

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

FIG. B.79 – Le Cramer's V

			desabon * clusters Crosstabulation						Total	
			clusters					Longue distance 6	Total	
			2e longue distance 1	Tech, 2e pour extras 2	Extras, pas tech 3	Pas extras 4	Utilise peu 5			
desabon	Non	Count	123	99	201	86	163	54	726	
		Expected Count	98,7	122,7	177,9	119,8	166,3	40,7	726,0	
		% within clusters	90,4%	58,6%	82,0%	52,1%	71,2%	96,4%	72,6%	
		Std. Residual	2,4	-2,1	1,7	-3,1	-3	2,1		
	Oui	Count	13	70	44	79	66	2	274	
		Expected Count	37,3	46,3	67,1	45,2	62,7	15,3	274,0	
		% within clusters	9,6%	41,4%	18,0%	47,9%	28,8%	3,6%	27,4%	
		Std. Residual	-4,0	3,5	-2,8	5,0	,4	-3,4		
Total			136	169	245	165	229	56	1000	
			136,0	169,0	245,0	165,0	229,0	56,0	1000,0	
			100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	

FIG. B.80 – Le tableau croisé

Le tableau croisé, lui, nous révèle que ce sont les clusters 2 et 4 qui se désabonnent le plus, et les clusters 1 et 6 qui se désabonnent le moins. Connaissant quel profil de consommateur a le plus tendance à se désabonner, l'entreprise peut maintenant tenter d'ajuster son tir...