

Collection L'ESPRIT ÉCONOMIQUE
SÉRIE COURS PRINCIPAUX



Moad El kharrim

Introduction à l'Économétrie Appliquée

L'Harmattan

Collection « L'esprit économique »

fondée par Sophie Boutilier et Dimitri Uzunidis en 1996
dirigée par Sophie Boutilier, Blandine Laperche, Dimitri Uzunidis

Si l'apparence des choses se confondait avec leur réalité, toute réflexion, toute Science, toute recherche serait superflue. La collection « L'esprit économique » soulève le débat, textes et images à l'appui, sur la face cachée économique des faits sociaux : rapports de pouvoir, de production et d'échange, innovations organisationnelles, technologiques et financières, espaces globaux et microéconomiques de valorisation et de profit, pensées critiques et novatrices sur le monde en mouvement... Ces ouvrages s'adressent aux étudiants, aux enseignants, aux chercheurs en sciences économiques, politiques, sociales, juridiques et de gestion, ainsi qu'aux experts d'entreprise et d'administration des institutions.

La collection est divisée en six séries :

Dans la série *Economie et Innovation* sont publiés des ouvrages d'économie industrielle, financière et du travail et de sociologie économique qui mettent l'accent sur les transformations économiques et sociales suite à l'introduction de nouvelles techniques et méthodes de production. L'innovation se confond avec la nouveauté marchande et touche le cœur même des rapports sociaux et de leurs représentations institutionnelles.

La série *Economie formelle* a pour objectif de promouvoir l'analyse des faits économiques contemporains en s'appuyant sur les approches critiques de l'économie telle qu'elle est enseignée et normalisée mondialement. Elle comprend des livres qui s'interrogent sur les choix des acteurs économiques dans une perspective macroéconomique, historique et prospective.

Dans la série *Le Monde en Question* sont publiés des ouvrages d'économie politique traitant des problèmes internationaux. Les économies nationales, le développement, les espaces élargis, ainsi que l'étude des ressorts fondamentaux de l'économie mondiale sont les sujets de prédilection dans le choix des publications.

La série *Krisis* a été créée pour faciliter la lecture historique des problèmes économiques et sociaux d'aujourd'hui liés aux métamorphoses de l'organisation industrielle et du travail. Elle comprend la réédition d'ouvrages anciens, de compilations de textes autour des mêmes questions et des ouvrages d'histoire de la pensée et des faits économiques.

La série *Clichés* a été créée pour fixer les impressions du monde économique. Les ouvrages contiennent photos et texte pour faire ressortir les caractéristiques d'une situation donnée. Le premier thème directeur est : mémoire et actualité du travail et de l'industrie ; le second : histoire et impacts économiques et sociaux des innovations.

La série *Cours Principaux* comprend des ouvrages simples, fondamentaux et/ou spécialisés qui s'adressent aux étudiants en licence et en master en économie, sociologie, droit, et gestion. Son principe de base est l'application du vieil adage chinois : « le plus long voyage commence par le premier pas ».

Moad El kharrim

Introduction à l'Économétrie Appliquée

L'Harmattan

**© L'Harmattan, 2022
5-7, rue de l'École-Polytechnique ; 75005 Paris**

<http://www.editions-harmattan.fr/>

ISBN : 978-2-14-026703-1
EAN : 9782140267031

Préface

L'introduction à l'Économétrie Appliquée est conçue pour un premier cours d'économétrie de premier cycle. D'après notre expérience, pour que l'économétrie soit pertinente dans un cours d'introduction, des applications intéressantes doivent motiver la théorie qui doit elle aussi correspondre aux applications. Ce principe simple représente un écart important par rapport à l'ancienne génération des livres d'économétrie, dans laquelle les modèles théoriques et les hypothèses ne correspondent pas aux applications. Il n'est pas étonnant que certains étudiants remettent en question la pertinence de l'économétrie après avoir passé une grande partie de leur temps à apprendre les hypothèses qu'ils réalisent par la suite sont irréalistes, de sorte qu'ils doivent ensuite apprendre des «solutions» aux «problèmes» qui surviennent surtout lorsque les applications ne correspondent pas aux hypothèses. Nous pensons qu'il est préférable de motiver le besoin d'outils avec une application concrète, puis de fournir quelques hypothèses simples qui correspondent à l'application. Bien que les méthodes soient immédiatement pertinentes pour les applications, cette approche peut donner vie à l'économétrie.

Cet ouvrage fournit aux étudiants de premier cycle un aperçu des techniques économétriques utilisées par les économistes aujourd'hui. Le texte se concentre sur des techniques économétriques standards et aussi sur les développements récents dans le domaine.

Un aspect très utile de cet ouvrage est l'utilisation de données pour illustrer l'application des différentes techniques. Cette approche rend le texte vivant et très pertinent pour les étudiants et les chercheurs. Avec la grande disponibilité des progiciels économétriques tels que EViews, Stata, etc., les lecteurs peuvent acquérir une expérience pratique en les manipulant dans certains exercices fournis dans le texte. Il est cependant très important que les lecteurs comprennent les principes sous-jacents qui guident l'utilisation de la gamme de procédures des tests statistiques produites par ces programmes économétriques. De plus, l'utilisation et l'interprétation correctes des différentes techniques et procédures de test

sont essentielles à la bonne pratique économétrique. Le bon mélange de la théorie et de l'application fourni dans ce texte devrait être utile aux économistes appliqués et aux étudiants.

Ce livre a deux objectifs. Le premier, et le plus important des deux, est de préparer les étudiants à effectuer des travaux économétriques appliqués. C'est en grande partie pour cette raison que presque tous les concepts théoriques introduits sont illustrés à l'aide de données et des progiciels économétriques très populaires et faciles à utiliser tel que EViews et Stata.

Malgré l'accent mis sur l'application, l'ouvrage est solidement fondé sur la théorie économétrique. En effet, c'est le deuxième objectif de ce livre d'ancrer l'étudiant dans la théorie afin qu'il puisse procéder, en utilisant les éléments acquis ici comme base, pour faire un programme d'études supérieures plus avancé en théorie et pratique économétriques avec une certaine facilité.

Moad El kharrim

Professeur à L'Université AbdelMalek ESSAADI

Faculté des Sciences Économiques et Gestion, Tétouan, Maroc

melkharrim@uae.ac.ma

Table des matières

Préface	5
Table des matières	7
Chapitre 1 C'est quoi l'Économétrie ?	11
1.1 C'est quoi l'Économétrie?	11
1.2 Modèles Économiques et Modèles Économétriques	12
1.3 Portée et Méthodologie de l'Économétrie	14
1.4 Structure des Données Économiques	17
1.5 Causalité et Notion de « Ceteris Paribus » en Analyse Économétrique	20
1.6 Qu'est-ce qui constitue un Test pour une Théorie Économique?	21
1.7 Résumé	22
Chapitre 2 Modèle de Régression Linéaire Simple	23
2.1 Relations entre Variables	23
2.1.1 Relation Fonctionnelle entre Deux Variables	24
2.1.2 Relation Statistique entre Deux Variables	25
2.2 Modèles de Régression et leurs Utilisations	27
2.2.1 Origines Historiques	27
2.2.2 Concepts de Base	27
2.2.3 Construction des Modèles de Régression	29
2.2.4 Fins d'Analyse de Régression	31
2.2.5 Régression et Causalité	32
2.2.6 Utilisation des Ordinateurs	32
2.3 Modèle de Régression Linéaire Simple avec Distribution des Termes d'Erreur Non Spécifiés	33
2.3.1 Présentation Formelle du Modèle	33
2.3.2 Estimation des Moindres Carrés et les Hypothèses Classiques	35
2.3.3 Propriétés Statistiques des Estimateurs des Moindres Carrés	42

TABLE DES MATIÈRES

2.3.4	Estimation de la Variance des Erreurs σ_ε^2	46
2.3.5	Conséquences de la Normalité des Erreurs	47
2.3.6	Estimation par Méthode de Maximum de Vraisemblance	50
2.4	Inférences dans la Régression et Analyse de Corrélation .	51
2.4.1	Intervalles de Confiance	53
2.4.2	Les Tests de Signification des Paramètres de Régression	54
2.4.3	Inférence sur le Coefficient de Corrélation	56
2.4.4	Analyse de la Variance ANOVA et Mesure de la Qualité d'Ajustement	58
2.4.5	Exemple Numérique : Impact de l'Education sur les Salaires	63
2.4.6	Prévision dans le Modèle de Régression Simple .	66
2.4.7	Analyse Résiduelle	69
2.4.8	La Régression à Travers l'Origine	70
2.4.9	Quelques Considérations sur les Inférences sur les Paramètres β_0 et β_1 de la Régression Simple	73
2.4.10	Exemple Numérique : Impact du Revenu sur la Consommation	79
2.5	Annexe	82
2.6	Exercices	86
Chapitre 3	Modèle de Régression Linéaire Multiple	113
3.1	Présentation Formelle du Modèle	114
3.2	Modèle de Régression Linéaire Général	114
3.2.1	Présentation Matricielle du Modèle de Régression Linéaire Général	115
3.3	Estimation des Coefficients de Régression	117
3.3.1	Matrice Hat	119
3.4	Hypothèses de Régression	120
3.4.1	Hypothèses Stochastiques	120
3.4.2	Hypothèses Structurelles	120
3.5	Estimation du Maximum de Vraisemblance (EMV)	121
3.6	Analyse de la Variance ANOVA et Coefficient de Détermination Multiple	123
3.7	Les Tests de Signification des Paramètres de Régression .	129
3.7.1	Intervalle de Confiance de β_k	130
3.7.2	Intervalle de Confiance de la Variance de l'Erreur .	130
3.7.3	Test de Signification d'un Seul Coefficient $\beta_k = 0$.	131
3.7.4	Test de Signification de plusieurs Coefficients β_k .	132
3.7.5	Autres Tests	133

TABLE DES MATIÈRES

3.7.6	Coefficients de Détermination Partielle	135
3.7.7	Coefficients de Corrélation Partielle	137
3.7.8	Interprétation des Coefficients de Corrélation Simples et Partiels	138
3.8	Tests de Stabilité	140
3.8.1	Tests d'Analyse de Variance	140
3.8.2	Tests Prédictifs de Stabilité	141
3.8.3	Test de Spécification de Ramsey (<i>Ramsey's RESET Test</i>)	146
3.9	Estimation de la Réponse Moyenne et Prévision de Nouvelles Observations	149
3.9.1	Estimation d'Intervalle de $E(Y_h)$	149
3.9.2	Prévision de Nouvelle Observation $Y_{h(nouvelle)}$	150
3.9.3	Prédiction de la Moyenne de p Nouvelles Observations à X_h	150
3.9.4	Précautions Concernant des Extrapolations Cachees	151
3.10	Variables Indicatrices (Dummy Variables)	151
3.11	Exemple Empirique	157
3.12	Modèle de Régression Multiple Standardisé	165
3.13	Formes Fonctionnelles des Modèles de Régression	171
3.13.1	Comment Mesurer l'Elasticité : le Modèle Log-Linéaire	172
3.13.2	Modèles Semilog : Modèles Log-Lin et Lin-Log	176
3.13.3	Modèles Réciproques	179
3.14	Choix de la Forme Fonctionnelle	181
3.15	Annexe	183
3.16	Exercices	185
Chapitre 4 Violations des Hypothèses Classiques		223
4.1	La Multicolinéarité et Ses Effets	223
4.1.1	Variables Prédictives Non Corrélées	224
4.1.2	Nature du Problème lorsque les Variables Prédictives sont Parfaitement Corrélées	227
4.1.3	Effets de la Multicolinéarité	229
4.1.4	Diagnostic de Multicolinéarité - Facteur d'Inflation de la Variance	241
4.1.5	Mesures Correctives	246
4.2	Autocorrélation des Erreurs	248
4.2.1	Qu'est-ce qui Cause l'Autocorrélation ?	249
4.2.2	Conséquences pour les Moindres Carrés Ordinaires	252
4.2.3	Détection de l'Autocorrélation	254

TABLE DES MATIÈRES

4.2.4	Correction de l'Autocorrélation	265
4.3	Hétéroscédasticité : Que se passe-t-il si la variance d'erreur n'est pas constante ?	273
4.3.1	La Nature de l'Hétéroscédasticité	273
4.3.2	Estimation des MCO en Présence d'Hétéroscédasticité	275
4.3.3	La Méthode des Moindres Carrés Généralisés (MCG)	276
4.3.4	Conséquences de l'Hétéroscédasticité pour les Estimateurs des MCO	280
4.3.5	Détection de l'Hétéroscédasticité	283
4.3.6	Correction de l'Hétéroscédasticité	299
4.4	Questions et Exercices	303
4.5	Annexe	317
Chapitre 5	Modèles à Équations Simultanées	321
5.1	Le Biais Simultané	322
5.2	Le Problème d'Identification	326
5.3	Estimation des Modèles à Équations Simultanées	330
5.3.1	Estimation d'une Équation Identifiée avec Précision : La Méthode MCI	330
5.3.2	Estimation d'une Équation Sur-Identifiée : La Méthode DMC	331
Annexes		337
A1.	La Loi Normale Centrée Réduite	338
A2.	Table de la Loi de Fisher-Snedecor	339
A3.	Table de La loi du Chi-Deux	341
A4.	Table de La loi de Student	342
A5.	Table de Durbin-Watson	343
Bibliographie		347
Index		349

Chapitre 1

C'est quoi l'Économétrie ?

1.1 C'est quoi l'Économétrie ?

Littéralement parlant, le mot "économétrie" signifie "mesure en économie". Cette définition est trop large pour être utile, car la plupart des considérations économiques concernent la mesure. Nous mesurons notre produit national brut, l'emploi, la masse monétaire, les exportations, les importations, les indices de prix, etc. Ce que nous entendons par économétrie est :

L'application des méthodes statistiques et mathématiques à l'analyse des données économiques, dans le but de donner un contenu empirique aux théories économiques et de les vérifier ou de les réfuter.

À cet égard, on distingue l'économétrie de l'économie mathématique, qui consiste uniquement en une application des mathématiques, et les théories dérivées n'ont pas nécessairement de contenu empirique.

L'application d'outils statistiques aux données économiques a une très longue histoire. Stigler note que le premier tableau de la demande "empirique" a été publié en 1699 par Charles Davenant et que les premières études de la demande statistiques modernes ont été réalisées par le statisticien italien Rodulfo Enini en 1907. L'impulsion principale du développement de l'économétrie avec la création de la Société d'économétrie en 1930 et la publication du journal *Econometrica* en janvier 1933. Avant toute analyse statistique avec des données économiques, il faut une formulation mathématique claire de la théorie économique pertinente. Pour prendre un exemple très simple, dire que la courbe de la demande est

1 C'est quoi l'Économétrie ?

en pente descendante ne suffit pas. Nous devons écrire la déclaration sous forme mathématique. Cela peut être fait de plusieurs manières. Par exemple, en définissant q comme quantité demandée et p comme prix, on peut écrire

$$q = \beta_0 + \beta_1 p \quad \beta_1 < 0$$

ou

$$q = Ap^{\beta_1} \quad \beta_1 < 0$$

1.2 Modèles Économiques et Modèles Économétriques

La première tâche d'un économètre est celle de formuler un modèle économétrique. Qu'est ce qu'un modèle ?

Un modèle est une représentation simplifiée d'un processus du monde réel. Par exemple, dire que la quantité demandée d'oranges dépend du prix d'oranges est une représentation simplifiée, car il existe une foule d'autres variables auxquelles on peut penser qui déterminent la demande d'oranges. Par exemple, le revenu des consommateurs, une augmentation de la conscience de l'alimentation (boire de l'alcool provoque le cancer, il est donc préférable de passer au jus d'orange, etc.), une augmentation ou une diminution du prix des pommes, etc. Cependant, ce flux d'autres variables n'a pas de fin.

De nombreux scientifiques ont plaidé en faveur de la simplicité, car les modèles simples sont plus faciles à comprendre, à communiquer et à tester de manière empirique avec des données. C'est la position de Karl Popper et Milton Friedman. Le choix d'un modèle simple pour expliquer des phénomènes complexes du monde réel conduit à deux critiques :

- (1) Le modèle est trop simplifié.
- (2) Les hypothèses sont irréalistes.

Par exemple, dans notre exemple de demande d'oranges, affirmer que cela dépend uniquement du prix des oranges est une simplification excessive et une hypothèse irréaliste. À la critique de la simplification excessive, on peut dire qu'il est préférable de commencer avec un modèle simplifié et de construire progressivement des modèles plus compliqués. C'est l'idée exprimée par Koopmans. D'autre part, certains préconisent de commencer par un modèle très général et de le simplifier progressivement en fonction des données disponibles. Le célèbre statisticien L. J. (Jimmy) Savage avait l'habitude de dire qu'un *modèle devrait être aussi gros qu'un éléphant*. Quels que soient les mérites relatifs de cette

1 C'est quoi l'Économétrie ?

approche alternative, nous commencerons par des modèles simples pour construire progressivement des modèles plus complexes.

L'autre critique que nous avons mentionnée est celle des "hypothèses irréalistes". Friedman soutenait que les hypothèses d'une théorie ne sont jamais réalistes du point de vue descriptif. Il dit :

La question pertinente à se poser sur les "hypothèses" d'une théorie n'est pas de savoir si elles sont descriptives "réalistes" car elles ne le sont jamais, mais si elles constituent des approximations suffisamment bonnes pour le but recherché. Et on peut répondre à cette question en ne regardant que si la théorie fonctionne, ce qui signifie si elle fournit des prédictions suffisamment précises.

Pour revenir à notre exemple de demande d'oranges, affirmer que cela ne dépend que du prix des oranges est une hypothèse peu réaliste. Cependant, l'inclusion d'autres variables, telles que le revenu et le prix des pommes dans le modèle, ne rend pas le modèle plus réaliste sur le plan descriptif. Même ce modèle peut être considéré comme basé sur des hypothèses irréalistes car il laisse de nombreuses autres variables (comme la conscience de la santé, etc.). Mais le problème est de savoir quel modèle est le plus utile pour prévoir la demande d'oranges. Cette question ne peut être décidée que sur la base des données dont nous disposons et des données que nous pouvons obtenir.

En pratique, nous incluons dans notre modèle toutes les variables que nous pensons être pertinentes pour notre objectif et nous plaçons le reste des variables dans un panier appelé "perturbation". Cela nous amène à la distinction entre un modèle économique et un modèle économétrique.

Un modèle économique est un ensemble d'hypothèses décrivant approximativement le comportement d'une économie (ou d'un secteur d'une économie). Un modèle économétrique comprend les éléments suivants :

- (1) Un ensemble d'équations comportementales dérivées du modèle économique. Ces équations impliquent certaines variables observées et certaines "perturbations" (qui sont un fourre-tout pour toutes les variables considérées comme non pertinentes aux fins de ce modèle ainsi que pour tous les événements imprévus).
- (2) Une déclaration indiquant s'il y a des erreurs d'observation dans les variables observées.
- (3) Une spécification de la distribution de probabilité des "perturbations" (et des erreurs de mesure).

1 C'est quoi l'Économétrie ?

Avec ces spécifications, nous pouvons tester la validité empirique du modèle économique et l'utiliser pour établir des prévisions ou l'utiliser dans l'analyse des politiques.

Prenant l'exemple le plus simple d'un modèle de demande, le modèle économétrique consiste généralement à :

- (1) Une équation comportementale

$$q = \beta_0 + \beta_1 p + \varepsilon$$

où q est la quantité demandée et p le prix. Ici, p et q sont les variables observées et ε est le terme d'erreur.

- (2) Une spécification de la distribution de probabilité de ε qui dit que $E(\varepsilon/p) = 0$ et que les valeurs de ε pour les différentes observations sont indépendamment et normalement distribuées avec une moyenne nulle et une variance σ_ε^2 .

Avec ces spécifications, on procède à un test empirique de la loi de la demande ou de l'hypothèse $\beta_1 < 0$. On peut également utiliser la fonction de la demande estimée à des fins de prévision et de politique.

1.3 Portée et Méthodologie de l'Économétrie

Les objectifs de l'économétrie sont les suivants :

- (1) Formulation d'un modèle économétrique, c'est-à-dire formulation d'un modèle économique sous une forme testable empiriquement. Généralement, il existe plusieurs façons de formuler le modèle économétrique à partir d'un modèle économique, car nous devons choisir la forme fonctionnelle, la spécification de la structure stochastique des variables, etc. Cette partie constitue l'aspect spécification du travail économétrique.
- (2) Estimation et test de ces modèles avec les données observées. Cette partie constitue l'aspect inférence du travail économétrique.
- (3) Utilisation de ces modèles à des fins de prévision et de politique.

Au cours des années 1950 et 1960, l'inférence a suscité beaucoup d'attention et l'aspect spécification a été très peu pris en compte. La principale préoccupation des économétriciens était l'estimation statistique de modèles économétriques correctement spécifiés. À la fin des années 1940, la Foundation Cowles constitua une avancée majeure à cet égard, mais l'analyse statistique posait d'énormes problèmes de calcul. Ainsi, les années 1950 et 1960 ont été principalement consacrées à la conception de

1 C'est quoi l'Économétrie ?

méthodes d'estimation alternatives et d'algorithmes informatiques alternatifs. Peu d'attention a été portée aux erreurs de spécification ou aux erreurs d'observations. Avec l'avènement des ordinateurs à grande vitesse, tout cela a toutefois changé. Les problèmes d'estimation ne sont plus redoutables et les économétriciens ont porté leur attention sur d'autres aspects de l'analyse économique.

Nous pouvons décrire schématiquement les différentes étapes d'une analyse économique, comme cela a été fait avant l'accent mis sur l'analyse de spécification. Ceci est illustré à la figure (1.1). Comme les entrées dans les cases sont explicites, nous ne les détaillerons pas. La seule case qui nécessite une explication est la case 4, "information préalable". Cela fait référence à toute information que nous pourrions avoir sur les paramètres inconnus du modèle. Ces informations peuvent provenir de la théorie économique ou d'études empiriques antérieures.

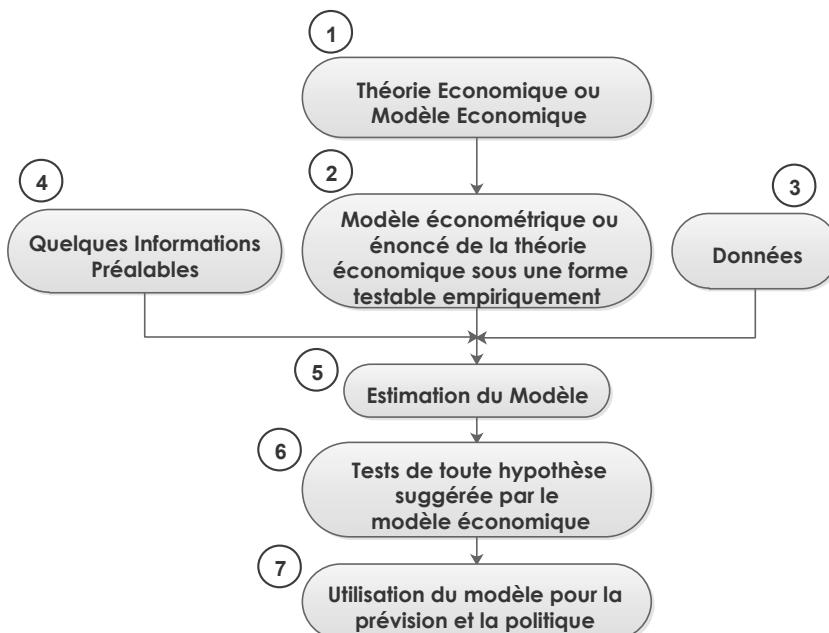


FIGURE 1.1 – Description schématique des étapes d'une analyse économique des modèles économiques

Le schéma présenté à la figure (1.1) a toutefois suscité un mécontentement considérable. Bien que l'on puisse trouver des exemples d'insatisfaction plus tôt, c'est principalement dans les années 1970 que des arguments ont été invoqués contre le trafic à sens unique illustré à la figure (1.1). Nous discuterons de trois de ces arguments.

1 C'est quoi l'Économétrie ?

- (1) Dans la figure (1.1), les tests économétriques des théories économiques ne donnent pas d'informations en retour sur la formulation de théories économiques (c'est-à-dire de la case 6 à la case 1). Il a été avancé que les économétriciens ne sont pas simplement des filles des théoriciens de l'économie. Ce n'est pas vrai qu'ils prennent juste les théories qui leur sont données et les testent sans rien apprendre des tests. Nous avons donc besoin d'une flèche allant de la case 6 à la case 1.
- (2) Il en va de même pour les agences de collecte de données. Il n'est pas vrai qu'ils rassemblent toutes les données possibles et les économétriciens utilisent les données qui leur sont fournies. (Le mot data provient du mot latin datum, ce qui signifie donné.) Il devrait y avoir des feedback des cases 2 et 5 vers la case 3.
- (3) En ce qui concerne la case 6 dans la figure (1.1), il a été avancé que le test d'hypothèse ne se réfère qu'aux hypothèses suggérées par le modèle économique initial. Cela dépend de l'hypothèse que la spécification adoptée dans la case 2 est correcte. Cependant, ce que nous devrions faire, c'est également tester l'adéquation de la spécification d'origine. Nous avons donc besoin d'une boîte supplémentaire de tests de spécifications et de vérification de diagnostic. Il y aura également un retour d'information de cette case à la case 2, c'est-à-dire que les tests de spécification aboutiront à une nouvelle spécification pour les modèles économétriques.

Les nouveaux développements suggérés sont montrés dans la figure (1.2). Certaines des cases d'origine ont été supprimés ou combinés. La description schématique de la figure (1.2) n'est qu'illustrative et ne doit pas être interprétée littéralement. Les points importants à noter sont les feedbacks :

- (1) Des résultats économétriques à la théorie économique ;
- (2) Des tests de spécification et vérification diagnostique à la spécification révisée du modèle économique ;,
- (3) Du modèle économique aux données.

Dans le schéma précédent, nous n'avons parlé que d'une théorie, mais il existe souvent de nombreuses théories concurrentes, et l'un des principaux objectifs de l'économétrie est d'aider à choisir entre des théories concurrentes, celà remet aussi en question le problème de sélection du modèle.

1 C'est quoi l'Économétrie ?

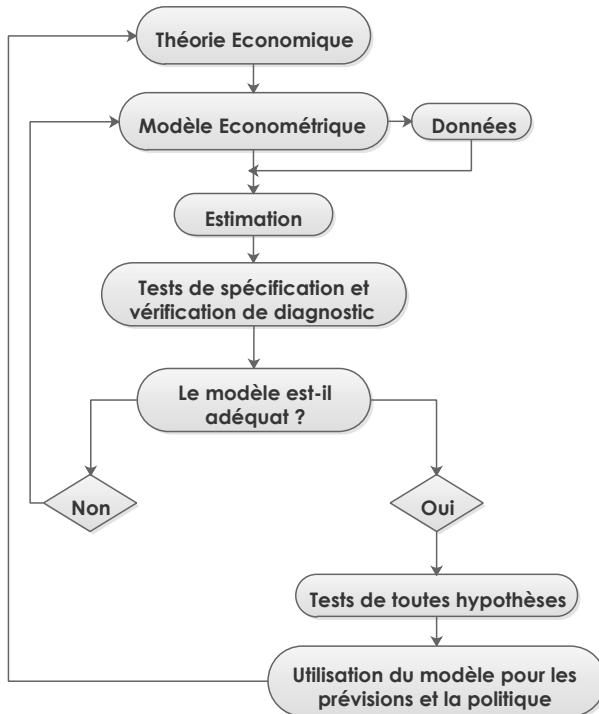


FIGURE 1.2 – Description schématique révisée des étapes d'une analyse économétrique des modèles économiques

1.4 Structure des Données Économiques

Il existe différents types de données économiques. Bien que certaines méthodes économétriques peuvent être appliquées directement à de nombreux types d'ensemble de données, il est essentiel d'examiner les particularités de certains ensembles. Nous décrivons ensuite les structures de données les plus importantes rencontrées en économétrie appliquée.

Données transversales

Un ensemble de données transversales comprend un échantillon de personnes, de ménages, d'entreprises, de villes, d'États, de pays ou de diverses autres unités prises à un moment donné. Parfois, les données sur toutes les unités ne correspondent pas exactement à la même période. Par exemple, plusieurs familles peuvent être enquêtées au cours de différentes semaines de l'année. Dans une analyse transversale pure, nous ignorions les différences de synchronisation mineures dans la collecte des données. Si un ensemble de familles a été interrogé au cours de

1 C'est quoi l'Économétrie ?

différentes semaines de la même année, nous considérons tout de même qu'il s'agit d'un ensemble de données transversal.

Parfois, l'échantillonnage aléatoire n'est pas une hypothèse appropriée pour l'analyse des données transversales. Par exemple, supposons que nous soyons intéressés à étudier les facteurs qui influencent l'accumulation de la richesse familiale. Nous pourrions interroger un échantillon aléatoire de familles, mais certaines familles pourraient refuser de déclarer leur richesse. Si, par exemple, les familles les plus riches sont moins susceptibles de révéler leur patrimoine, l'échantillon de patrimoine obtenu ne constitue pas un échantillon aléatoire de la population de toutes les familles.

Une autre violation de l'échantillonnage aléatoire se produit lorsque nous échantillonnons dans des unités de grande taille par rapport à la population, en particulier des unités géographiques. Le problème potentiel dans de tels cas est que la population n'est pas assez nombreuse pour supposer raisonnablement que les observations sont indépendantes.

Données en séries chronologiques

Un ensemble de données chronologiques comprend des observations sur une ou plusieurs variables dans le temps. Les exemples de données de séries chronologiques incluent les prix des actions, la masse monétaire, l'indice des prix à la consommation, le produit intérieur brut, les taux d'homicides annuels et les chiffres des ventes d'automobiles. Comme les événements passés peuvent influer sur les événements futurs et que les retards de comportement prévalent dans les sciences sociales, le temps est une dimension importante dans un ensemble de données de séries chronologiques. Contrairement à la disposition des données transversales, le classement chronologique des observations dans une série chronologique véhicule des informations potentiellement importantes.

Une caractéristique essentielle des données chronologiques qui rend l'analyse plus difficile que les données transversales est le fait que les observations économiques peuvent rarement, voire jamais, être considérées comme indépendantes dans le temps. La plupart des séries économiques et autres sont liées, souvent étroitement liées, à leurs histoires récentes. Par exemple, connaître le produit intérieur brut du dernier trimestre nous en dit long sur la fourchette probable du PIB au cours de ce trimestre, car le PIB a tendance à rester assez stable d'un trimestre à l'autre. Bien que la plupart des procédures économétriques puissent être utilisées avec des données à la fois transversales et chronologiques, il reste encore beaucoup à faire pour spécifier des modèles économétriques

1 C'est quoi l'Économétrie ?

pour les données chronologiques avant de pouvoir justifier des méthodes économétriques standard. En outre, des modifications et des améliorations aux techniques économétriques standard ont été développées pour prendre en compte et exploiter la nature dépendante des séries chronologiques économiques et pour traiter d'autres problèmes, tels que le fait que certaines variables économiques ont tendance à afficher des tendances claires dans le temps.

Une autre caractéristique des données chronologiques pouvant nécessiter une attention particulière est la fréquence à laquelle les données sont collectées. En économie, les fréquences les plus courantes sont quotidiennes, hebdomadaires, mensuelles, trimestrielles et annuelles. Les cours des actions sont enregistrés tous les jours (sauf samedi et dimanche). De nombreuses séries macroéconomiques sont mises en tableaux tous les mois, y compris les taux d'inflation et d'emploi. Les autres séries macro sont enregistrées moins fréquemment, par exemple tous les trois mois (tous les trimestres). Le produit intérieur brut est un exemple important de série trimestrielle.

Données en Coupes Transversales

Certains ensembles de données ont à la fois des caractéristiques transversales et chronologiques. Par exemple, supposons que deux enquêtes transversales sur les ménages soient menées aux États-Unis, une en 1985 et une en 1990. En 1985, un échantillon aléatoire de ménages est interrogé pour des variables telles que le revenu, l'épargne, la taille de la famille, etc. . En 1990, un nouvel échantillon aléatoire de ménages est constitué à l'aide des mêmes questions de l'enquête. Afin d'augmenter la taille de notre échantillon, nous pouvons former une section transversale groupée en combinant les deux années. Étant donné que des échantillons aléatoires sont prélevés chaque année, ce serait un hasard si le même ménage faisait partie de l'échantillon au cours des deux années. (La taille de l'échantillon est généralement très petite comparée au nombre de ménages aux États-Unis.) Ce facteur important distingue un échantillon en coupe groupé d'un ensemble de données de panel.

La collecte des données en coupes transversales de différentes années est souvent un moyen efficace d'analyser les effets d'une nouvelle politique gouvernementale. L'idée est de collecter des données des années avant et après un changement d'une politique.

1 C'est quoi l'Économétrie ?

Données de Panel ou Longitudinales

Un ensemble de données de panel (ou données longitudinales) consiste en une série chronologique pour chaque élément en coupe de l'ensemble de données. A titre d'exemple, supposons que nous ayons des antécédents de salaire, d'éducation et d'emploi pour un ensemble d'individus suivis sur une période de dix ans. Nous pourrions également collecter des informations, telles que des données d'investissement et financières, sur le même ensemble d'entreprises sur une période de cinq ans. Les données de panel peuvent également être collectées sur des unités géographiques. Par exemple, nous pouvons collecter des données pour le même ensemble de pays sur les flux d'immigration, les taux d'imposition, les taux de rémunération, les dépenses de l'État, etc., pour les années 1980, 1985 et 1990.

La caractéristique principale des données de panel qui la distingue d'une coupe transversale est le fait que les mêmes coupes transversales (individus, entreprises ou départements) sont suivies sur une période donnée.

Étant donné que les données de panel nécessitent la réPLICATION des mêmes unités dans le temps, il est plus difficile d'obtenir des coupe de données de panel, en particulier ceux concernant les individus, les ménages et les entreprises, que les coupes transversales. L'observation des mêmes unités dans le temps présente plusieurs avantages par rapport aux données transversales. L'avantage sur lequel nous allons nous concentrer dans ce texte est que le fait de disposer de plusieurs observations sur les mêmes unités nous permet de contrôler certaines caractéristiques non observées d'individus, d'entreprises, etc. Ainsi, l'utilisation de plusieurs observations peut faciliter l'inférence causale dans des situations où il serait très difficile d'inférer un lien de causalité si une seule section était disponible. Un deuxième avantage des données de panel est qu'elles nous permettent souvent d'étudier l'importance des décalages dans le comportement ou du résultat de la prise de décision. Cette information peut être importante, car de nombreuses politiques économiques ne devraient avoir un impact qu'après un certain temps.

1.5 Causalité et Notion de « Ceteris Paribus » en Analyse Économétrique

Dans la plupart des tests de la théorie économique, et certainement pour l'évaluation des politiques publiques, l'économiste a pour objectif de déduire qu'une variable a un **effet causal** sur une autre (comme le

1 C'est quoi l'Économétrie ?

taux de criminalité ou la productivité des travailleurs). Le simple fait de trouver une association entre deux variables ou plus pourrait être suggestif, mais à moins que la causalité ne puisse être établie, elle est rarement convaincante. La notion de **ceteris paribus** - qui signifie « les autres facteurs (pertinents) étant égaux par ailleurs » - joue un rôle important dans l'analyse causale.

Vous vous souvenez probablement à partir de l'introduction à la théorie économique que la plupart des questions économiques sont ceteris paribus par nature. Par exemple, en analysant la demande des consommateurs, nous souhaitons connaître l'effet de la modification du prix d'un bien sur la quantité demandée, tout en maintenant fixes tous les autres facteurs, tels que le revenu, les prix des autres biens et les goûts individuels. Si d'autres facteurs ne sont pas conservés, nous ne pouvons pas connaître l'effet causal d'un changement de prix sur la quantité demandée.

La question clé dans la plupart des études empiriques est la suivante : suffisamment de facteurs ont-ils été résolus pour justifier la causalité ? Une étude économétrique est rarement évaluée sans soulever cette question.

1.6 Qu'est-ce qui constitue un Test pour une Théorie Économique ?

Nous avons indiqué précédemment que l'un des objectifs de l'économétrie était de tester les théories économiques. Une question importante qui se pose est la suivante : en quoi consiste un test ? Comme preuve de la réussite du test de la théorie économique, il est habituel de signaler que les signes des coefficients estimés dans un modèle économétrique sont corrects. Cette approche peut être qualifiée d'approche de confirmation des théories économiques. Le problème de cette approche est que, comme le souligne Mark Blaug :

Dans de nombreux domaines de l'économie, différentes études économétriques aboutissent à des conclusions contradictoires et, compte tenu des données disponibles, il n'existe fréquemment aucune méthode efficace pour décider quelle conclusion est correcte. En conséquence, des hypothèses contradictoires continuent de coexister, parfois pendant des décennies ou plus.

Un test plus valable d'une théorie économique consiste à déterminer

1 C'est quoi l'Économétrie ?

si elle peut donner des prévisions meilleures que celles des théories alternatives suggérées plus tôt. Il faut donc comparer un modèle donné avec les modèles précédents. Cette approche consistant à évaluer des théories alternatives a reçu une attention accrue ces dernières années.

1.7 Résumé

Dans ce chapitre introductif, nous avons discuté de l'objectif et de la portée de l'analyse économétrique. L'économétrie est utilisée dans tous les domaines économiques appliqués pour tester les théories économiques, informer les décideurs publics et privés et pour prévoir les séries chronologiques économiques. Parfois, un modèle économétrique est dérivé d'un modèle économique formel, mais dans d'autres cas, les modèles économétriques sont basés sur un raisonnement économique et une intuition informels. Toute analyse économétrique a pour objectif d'estimer les paramètres du modèle et de tester des hypothèses sur ces paramètres ; les valeurs et les signes des paramètres déterminent la validité d'une théorie économique et les effets de certaines politiques.

Les données transversales, les séries chronologiques, les coupes transversales et les données de panel sont les types les plus courants de structures de données utilisées dans l'économétrie appliquée. Les ensembles de données impliquant une dimension temporelle, tels que les séries chronologiques et les données de panel, nécessitent un traitement spécial en raison de la corrélation dans le temps des séries chronologiques les plus économiques. D'autres problèmes, tels que les tendances et la saisonnalité, se posent lors de l'analyse des données de séries chronologiques, mais pas des données transversales.

Nous avons discuté des notions de *ceteris paribus* et d'*inférence causale*. Dans la plupart des cas, les hypothèses en sciences sociales sont, par nature, *ceteris paribus* : tous les autres facteurs pertinents doivent être fixés lors de l'étude de la relation entre deux variables. En raison de la nature non expérimentale de la plupart des données recueillies en sciences sociales, il est très difficile de découvrir des relations de cause à effet.

Chapitre 2

Modèle de Régression

Linéaire Simple

L'analyse de régression est une méthodologie statistique qui utilise la relation entre deux variables quantitatives ou plus, de sorte qu'une variable puisse être prédite à partir de l'autre ou d'autres. Cette méthodologie est largement utilisée dans divers domaines comme l'économie, les sciences sociales, la biologie et dans nombreuses autres disciplines. Quelques exemples d'applications sont :

- Les ventes d'un produit peuvent être prédites en utilisant la relation entre les ventes et le montant des dépenses publicitaires.
- La performance d'un employé au travail peut être prédite en utilisant la relation entre la performance et les tests d'aptitude.
- La durée de l'hospitalisation d'un patient opéré peut être prédite en utilisant la relation entre le temps passé à l'hôpital et la gravité de l'opération.

Dans ce chapitre, nous examinons tout d'abord les idées de base de l'analyse de régression, puis nous discutons l'estimation des paramètres de modèle de régression linéaire contenant une seule variable prédictive.

2.1 Relations entre Variables

Le concept de relation entre deux variables, telles qu'entre le revenu familial et les dépenses familiales pour le logement, est un concept familier. Nous distinguons une relation fonctionnelle et une relation statistique et considérons chacune d'elles tour à tour.

2 Modèle de Régression Linéaire Simple

2.1.1 Relation Fonctionnelle entre Deux Variables

Une relation fonctionnelle entre deux variables est exprimée par une formule mathématique. Si X désigne la variable indépendante et Y la variable dépendante, une relation fonctionnelle est de la forme :

$$Y = f(X)$$

Étant donné une valeur particulière de X , la fonction f indique la valeur correspondante de Y .

Exemple : Examinons la relation entre les ventes en dirham (Y) d'un produit vendu à prix fixe et le nombre d'unités vendues (X). Si le prix de vente est de 2 dh l'unité, la relation est exprimée par l'équation :

$$Y = 2X$$

Cette relation fonctionnelle est illustrée à la figure (2.1). Le nombre d'unités vendues et les ventes en dirham au cours de trois périodes récentes (alors que le prix unitaire est resté constant à 2 dh) sont les suivants :

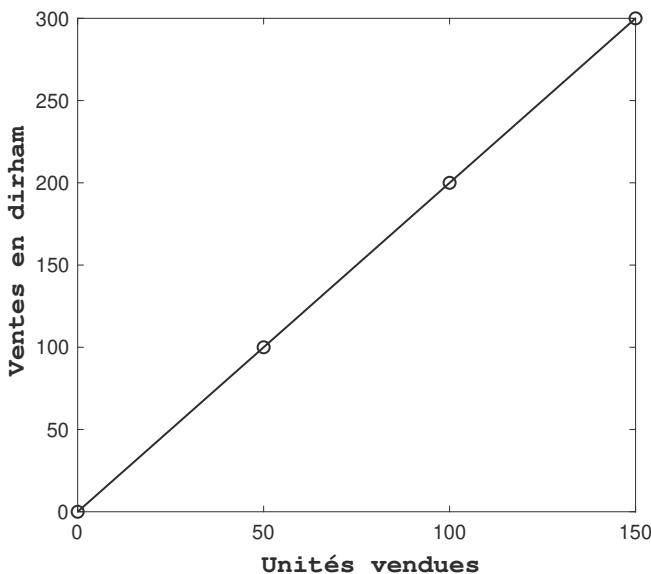


FIGURE 2.1 – Exemple d'une Relation Fonctionnelle

2 Modèle de Régression Linéaire Simple

2.1.2 Relation Statistique entre Deux Variables

Une relation statistique, contrairement à une relation fonctionnelle, n'est pas parfaite. En général, les observations pour une relation statistique ne tombent pas directement sur la courbe de relation.

Exemple 1 : Des évaluations de performance pour 10 employés ont été obtenues à la mi-année et à la fin de l'année. Ces données sont représentées à la figure (2.2a). Les évaluations de fin d'année sont considérées comme la variable *dépendante* ou de *réponse Y* et les évaluations de mi-année comme la variable *indépendante, explicative* ou *prédictive X*. Le tracé est effectué comme auparavant. Par exemple, les évaluations du rendement du premier employé à mi-année et en fin d'année sont tracées en $X = 90$, $Y = 94$.

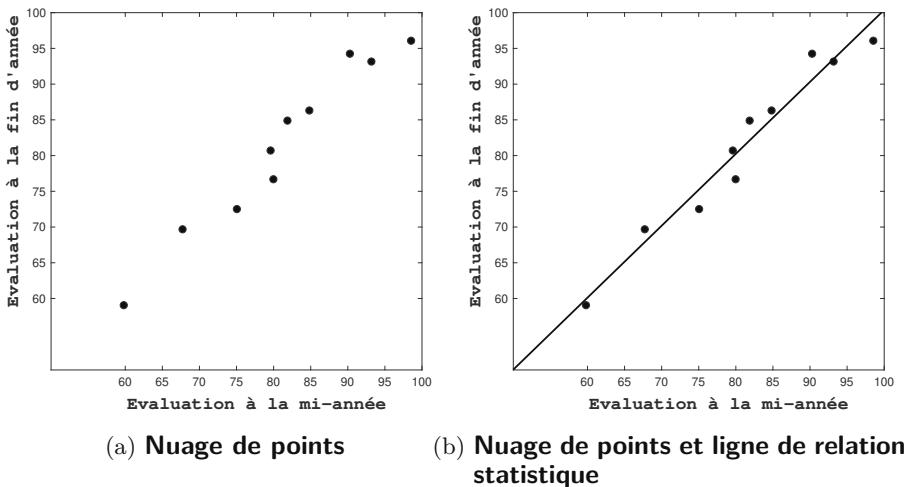


FIGURE 2.2 – Relation statistique entre l'évaluation du rendement à mi-année et l'évaluation de fin d'année

La figure (2.2a) montre clairement qu'il existe une relation entre les évaluations de mi-année et de fin d'année; plus élevée est l'évaluation de mi-année, le plus élevée est l'évaluation de fin d'année. Cependant, la relation n'est pas parfaite. Il y a une dispersion des points, ce qui suggère qu'une partie de la variation dans les évaluations de fin d'année n'est pas prise en compte dans les évaluations de performance de mi-année. Par exemple, deux employés ont eu des évaluations de $X = 80$ en milieu d'année, mais ils ont reçu des évaluations de fin d'année un peu différentes. La dispersion des points dans une relation statistique, la figure (2.2a) est appelée *diagramme de dispersion* ou *graphique de*

2 Modèle de Régression Linéaire Simple

dispersion. En terminologie statistique, chaque point du diagramme de dispersion représente un essai ou un cas.

Dans la figure (2.2b), nous avons tracé une relation qui décrit la relation statistique entre les évaluations de mi-année et de fin d'année. Le graphique indique la tendance générale selon laquelle les évaluations de fin d'année varient avec le niveau d'évaluation des performances en milieu d'année. Notez que la plupart des points ne tombent pas directement dans la relation statistique. Cette dispersion des points autour de la ligne représente une variation dans les évaluations de fin d'année qui n'est pas associée à une évaluation de performance en milieu d'année et qui est généralement considérée comme étant de nature aléatoire. Les relations statistiques peuvent être très utiles, même si elles ne possèdent pas l'exactitude d'une relation fonctionnelle.

Exemple 2 : La figure (2.3) présente des données sur l'âge et le niveau plasmatique d'un stéroïde chez 27 femmes en bonne santé âgées de 8 à 25 ans. Les données suggèrent fortement que la relation statistique est *curviligne* (non linéaire). La courbe de relation a également été dessinée à la figure (2.3). Cela implique que, à mesure que l'âge augmente, le niveau de stéroïdes augmente jusqu'à un point puis il commence à se stabiliser. Notez à nouveau la dispersion des points autour de la courbe de la relation statistique parrait typique comme dans toutes les relations statistiques.

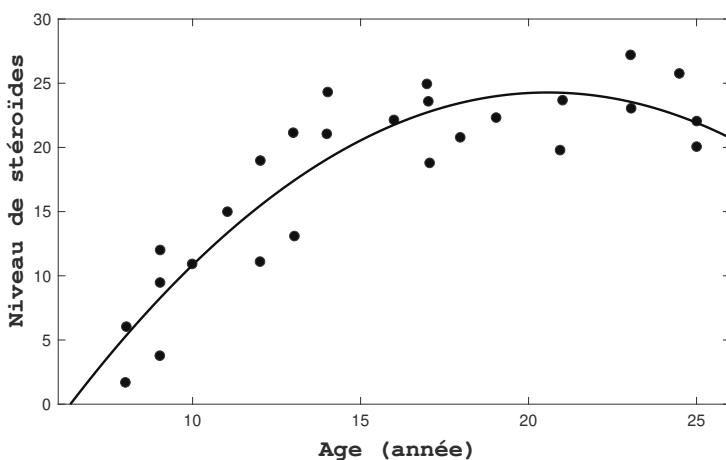


FIGURE 2.3 – Relation statistique curviligne entre l'âge et le niveau de stéroïdes chez des femmes en bonne santé âgées de 8 à 25 ans

2.2 Modèles de Régression et leurs Utilisations

2.2.1 Origines Historiques

L'analyse de régression a été développée par Francis Galton vers la fin du XIXe siècle. Galton avait étudié la relation entre les hauteurs des parents et des enfants et avait noté que les hauteurs des enfants des parents grands et petits semblaient "revenir" ou "régresser" à la moyenne du groupe. Il a considéré cette tendance comme une régression vers la "médiocrité". Galton a développé une description mathématique de cette tendance à la régression, précurseur des modèles de régression actuels.

Le terme *régression* persiste encore à ce jour pour décrire les relations statistiques entre variables.

2.2.2 Concepts de Base

Un modèle de régression est un moyen formel d'exprimer les deux ingrédients essentiels d'une relation statistique :

- Une tendance de la variable dépendante Y à varier avec la variable prédictive X de manière systématique.
- Une dispersion de points autour de la courbe de relation statistique.

Ces deux caractéristiques sont incorporées dans un modèle de régression en postulant que :

- Il existe une distribution de probabilité de Y pour chaque niveau de X .
- Les moyennes de ces distributions de probabilité varient de façon systématique avec X .

Exemple : Reprenons l'exemple d'évaluation de performance présenté à la figure (2.2). L'évaluation de fin d'année Y est traitée dans un modèle de régression comme une variable aléatoire. Une distribution de probabilité de Y est postulée pour chaque niveau d'évaluation de performance en milieu d'année. La figure (2.4) illustre cette distribution de probabilité pour $X = 90$, et qui correspond à l'évaluation en milieu d'année du premier employé.

L'évaluation réelle de fin d'année de cet employé, $Y = 94$, est ensuite considérée comme une sélection aléatoire de cette distribution de probabilité. La figure (2.4) montre également les distributions de probabilité de Y pour les niveaux d'évaluation de mi-année $X = 50$ et $X = 70$.

2 Modèle de Régression Linéaire Simple

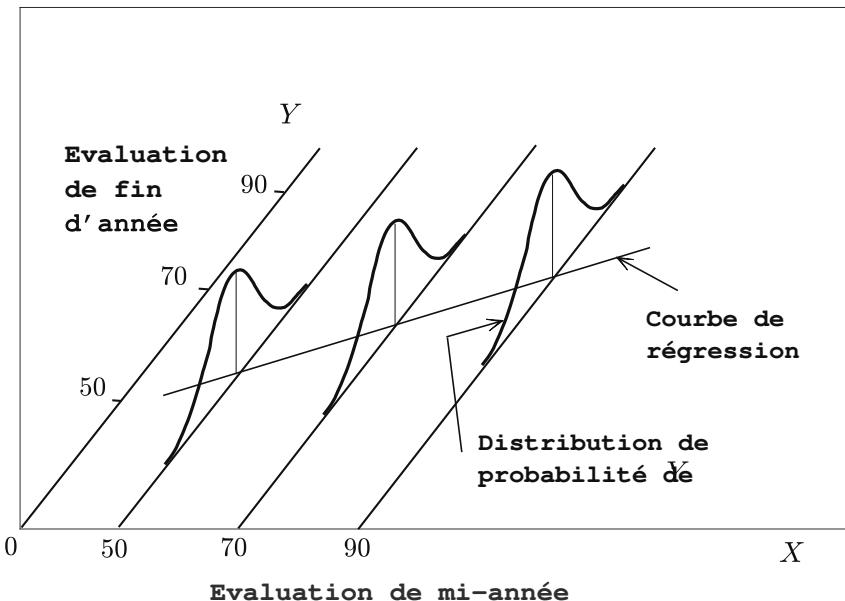


FIGURE 2.4 – Représentation graphique du modèle de régression

Notez que les moyennes des distributions de probabilité ont une relation systématique avec le niveau de X . Cette relation systématique est appelée la fonction de régression de Y sur X .

Le graphique de la fonction de régression s'appelle la courbe de régression. Notez que sur la figure (2.4), la fonction de régression est légèrement curviligne. Cela impliquerait, pour notre exemple, que l'augmentation de l'évaluation attendue (moyenne) en fin d'année avec une augmentation de l'évaluation de la performance en milieu d'année est retardée à des niveaux plus élevés de performance en milieu d'année.

Les modèles de régression peuvent différer dans la forme fonctionnelle de la relation de régression (linéaire, curviligne), dans la forme des distributions de probabilité de Y (symétrique, asymétrique) et aussi dans divers autres formes. Quelle que soit la variation, le concept de distribution de probabilité de Y pour un X donné est la contrepartie formelle de la dispersion empirique dans une relation statistique. De même, la courbe de régression, qui décrit la relation entre les moyennes des distributions de probabilité de Y et le niveau de X , est la contrepartie de la tendance générale de Y à varier avec X systématiquement dans une relation statistique.

Modèles de régression avec plusieurs variables prédictives : Les

2 Modèle de Régression Linéaire Simple

modèles de régression peuvent contenir plus d'une variable prédictive. Trois exemples peuvent être cités :

- (1) Dans une étude sur l'efficacité de 67 succursales d'une chaîne de crédit à la consommation, la variable dépendante était le coût d'exploitation direct pour l'année écoulée. Il y avait quatre variables prédictives : la taille moyenne des prêts non remboursés au cours de l'année, le nombre moyen de prêts non remboursés, le nombre total de nouvelles demandes de prêt traitées et un indice des salaires.
- (2) Dans une étude sur les achats de tracteurs, la variable dépendante était le volume des achats de tracteurs sur le territoire de vente d'une entreprise de matériel agricole. Il y avait neuf variables prédictives, y compris l'âge moyen des tracteurs dans les exploitations du territoire, le nombre d'exploitations sur le territoire et un indice de quantité de la production végétale sur le territoire.
- (3) Dans une étude médicale sur des enfants de petite taille, la variable dépendante était le niveau maximal d'hormone de croissance plasmatique. Il y avait 14 variables prédictives, notamment l'âge, le sexe, la taille, le poids et 10 mesures du pli cutané.

Les caractéristiques de modèle représentées à la figure (2.4) doivent être étendues à d'autres dimensions lorsqu'il existe plusieurs variables prédictives. Avec deux variables prédictives X_1 et X_2 , par exemple, le modèle de régression suppose une distribution de probabilité de Y pour chaque combinaison (X_1, X_2) . La relation systématique entre les moyennes de ces distributions de probabilité et les variables prédictives X_1 et X_2 est donnée par une surface de régression.

2.2.3 Construction des Modèles de Régression

Sélection des variables prédicteurs : Comme la réalité doit être réduite à des proportions raisonnables chaque fois que nous construisons des modèles, seul un nombre limité de variables explicatives ou prédictives peut - ou doit - être inclus dans un modèle de régression pour toute situation présentant un intérêt. Le problème central dans de nombreuses études exploratoires est donc celui de choisir, pour un modèle de régression, un ensemble de variables prédictives qui soit "bon" aux fins de l'analyse. Ce choix dépend en grande partie de la mesure dans laquelle une variable choisie contribue à la réduction de la variation restante de Y et après la prise en compte des contributions d'autres variables prédictives qui ont été provisoirement incluses dans le modèle de

2 Modèle de Régression Linéaire Simple

régression. D'autres considérations incluent l'importance de la variable en tant qu'agent causal dans le processus analysé ; la mesure dans laquelle les observations sur la variable peuvent être obtenues de manière plus précise, rapide ou économique que sur des variables concurrentes ; et le degré de contrôle de la variable.

Forme fonctionnelle de la relation de régression : Le choix de la forme fonctionnelle de la relation de régression est lié au choix des variables prédictives. Parfois, une théorie pertinente peut indiquer la forme fonctionnelle appropriée. La théorie de l'apprentissage, par exemple, peut indiquer que la fonction de régression liant le coût de production unitaire au nombre de fois où l'article a été produit devrait avoir une forme spécifiée avec des propriétés asymptotiques particulières. Le plus souvent, cependant, la forme fonctionnelle de la relation de régression n'est pas connue à l'avance et doit être décidée de manière empirique une fois les données collectées. Les fonctions de régression linéaire ou quadratique sont souvent utilisées comme première approximation satisfaisante de fonctions de régression de nature inconnue. En effet, ces types simples de fonctions de régression peuvent être utilisés même lorsque la théorie fournit la forme fonctionnelle pertinente, notamment lorsque la forme connue est très complexe mais peut être raisonnablement approchée par une fonction de régression linéaire ou quadratique. La figure (2.5a) illustre un cas où la fonction de régression complexe peut être raisonnablement estimée par une fonction de régression linéaire. La figure (2.5b) donne un exemple où deux fonctions de régression linéaire peuvent être utilisées "par morceaux" pour approcher une fonction de régression complexe.

Portée du modèle : Lors de la formulation d'un modèle de régression, il est généralement nécessaire de limiter la couverture du modèle à un intervalle ou une région de valeurs de la ou des variables prédictives. L'étendue est déterminée soit par la conception de l'enquête, soit par l'étendue des données disponibles. Par exemple, une entreprise qui a étudié l'effet des prix sur le volume des ventes a examiné six niveaux de prix, allant de 4.95 dh à 6.95 dh. Ici, la portée du modèle est limitée à des niveaux de prix allant de près de 5 dhs à près de 7 dh. La forme de la fonction de régression située sensiblement en dehors de cet intervalle ferait l'objet d'un sérieux doute, car l'enquête n'a fourni aucune preuve quant à la nature de la relation statistique inférieure à 4.95 dh ou supérieure à 6.95 dh.

2 Modèle de Régression Linéaire Simple

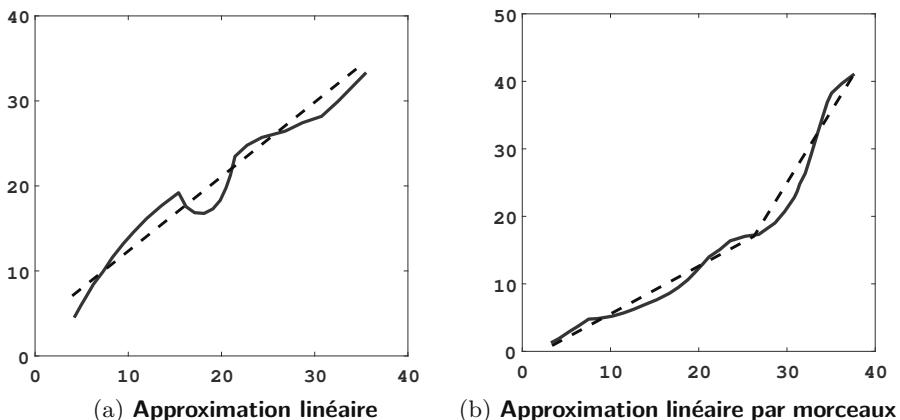


FIGURE 2.5 – Utilisations des fonctions de régression linéaire pour approximer les fonctions de régression complexe - La ligne en gras est la fonction de régression réelle et la ligne en pointillé est l'approximation de la régression

2.2.4 Fins d'Analyse de Régression

L'analyse de régression sert trois objectifs principaux : (1) la description, (2) le contrôle et (3) la prédiction. Ces objectifs sont illustrés par les trois exemples cités précédemment. L'étude sur l'achat de tracteurs avait un but descriptif. Dans l'étude des coûts de fonctionnement des succursales, l'objectif principal était le contrôle administratif; en développant une relation statistique utilisable entre le coût et les variables prédictives, la direction a pu établir des normes de coûts pour chaque succursale de la chaîne. Dans l'étude médicale des enfants de petite taille, l'objectif était la prédiction. Les cliniciens ont pu utiliser la relation statistique pour prédire les déficiences en hormone de croissance chez les enfants de petite taille en utilisant des mesures simples.

Les objectifs de l'analyse de régression se chevauchent souvent dans la pratique. L'exemple de la succursale est un exemple typique. La connaissance de la relation entre les coûts d'exploitation et les caractéristiques de la succursale a non seulement permis à la direction d'établir des normes de coûts pour chaque bureau, mais elle pouvait aussi prévoir les coûts et, à la fin de l'exercice, elle pouvait comparer les coûts réels de la succursale aux coûts prévus.

2.2.5 Régression et Causalité

L'existence d'une relation statistique entre la variable dépendante Y et la variable explicative ou prédictive X n'implique en aucune manière que Y dépende de X de manière causale. Quelle que soit la force de la relation statistique entre X et Y , aucun effet causal n'est nécessairement impliqué par le modèle de régression. Par exemple, les données sur la taille du vocabulaire (X) et la vitesse d'écriture (Y) pour un échantillon de jeunes enfants âgés de 5 à 10 ans montrent une relation de régression positive.

Cette relation n'implique cependant pas qu'une augmentation du vocabulaire entraîne une vitesse d'écriture plus rapide. Ici, d'autres variables explicatives, telles que l'âge de l'enfant et le niveau d'éducation, affectent à la fois le vocabulaire (X) et la vitesse d'écriture (Y). Les enfants plus âgés ont un plus grand vocabulaire et une vitesse d'écriture plus rapide. Même lorsqu'une relation statistique forte reflétant des conditions de causalité, celles-ci peuvent agir dans la direction opposée, de Y à X . Considérons, par exemple, l'étalonnage d'un thermomètre. Ici, les lectures du thermomètre sont prises à différentes températures connues et la relation de régression est étudiée de manière à pouvoir évaluer la précision des prédictions établies à l'aide des lectures du thermomètre. À cette fin, le thermomètre indique la variable de prédiction X et la température réelle est la variable dépendante Y à prédire.

Cependant, le modèle de causalité ne va pas ici de X à Y , mais dans le sens opposé : la température réelle (Y) affecte la lecture du thermomètre (X).

Ces exemples démontrent la nécessité de prendre des précautions pour tirer des conclusions sur les relations de causalité à partir d'une analyse de régression. L'analyse de régression en elle-même ne fournit aucune information sur les modèles de causalité et doit être complétée par des analyses supplémentaires pour obtenir des informations sur les relations de causalité.

2.2.6 Utilisation des Ordinateurs

Comme l'analyse de régression nécessite souvent des calculs longs et fastidieux, les ordinateurs sont généralement utilisés pour effectuer les calculs nécessaires. Presque tous les logiciels de statistiques pour ordinateurs contiennent un composant de régression. Bien que les packages diffèrent dans de nombreux détails, leur la régression de base a tendance à être assez similaire.

2 Modèle de Régression Linéaire Simple

Après une première explication des calculs de régression requis, voici quelque logiciels informatiques qui peuvent être utilisés pour effectuer les calculs nécessaires : SAS, R, SPSS, MATLAB, Eviews etc.

2.3 Modèle de Régression Linéaire Simple avec Distribution des Termes d'Erreur Non Spécifiés

2.3.1 Présentation Formelle du Modèle

Nous considérons un modèle de régression de base où il n'existe qu'une seule variable prédictive et où la fonction de régression est linéaire. Le modèle peut être défini comme suit :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n \quad (2.1)$$

où Y_i désigne la i -ième observation de la variable dépendante Y qui pourrait être la consommation, l'investissement ou la production, et X_i désigne la i -ième observation de la variable indépendante X qui pourrait être le revenu disponible, le taux d'intérêt ou un input. Ces observations pourraient être recueillies à partir des entreprises ou des ménages à un moment donné, auquel cas nous appelons ces données "*une coupe instantanée*". Alternativement, ces observations peuvent être collectées au fil du temps pour une industrie ou un pays spécifique, auquel cas nous appelons ces derniers données "*une série chronologique*". n est le nombre d'observations, qui pourrait être le nombre d'entreprises ou de ménages en coupe instantanée ou le nombre d'années si les observations sont collectées chaque année. β_0 et β_1 sont **la constante** et **la pente** de cette relation linéaire simple entre Y et X . On suppose qu'ils sont des paramètres inconnus, et donc à estimer à partir des données.

La représentation graphique des données, c'est-à-dire Y en fonction de X serait très illustrative pour montrer quel type de relation existe *empiriquement* entre ces deux variables. Par exemple, si Y est la consommation et X est le revenu disponible, nous nous attendons à une relation positive entre ces variables et les données peuvent ressembler à la figure (2.6a) lorsqu'elles sont tracées pour un échantillon aléatoire de ménages. Si β_0 et β_1 étaient connus, on pourrait tracer la ligne droite $(\beta_0 + \beta_1 X)$ comme le montre la figure (2.6a). Il est clair que toutes les observations (X_i, Y_i) ne sont pas sur la ligne droite $(\beta_0 + \beta_1 X)$. En fait, l'équation (2.1) indique que la différence entre chaque Y_i et le point correspondant $(\beta_0 + \beta_1 X_i)$ est due à une erreur aléatoire ε_i . Cette erreur peut être due à :

- (i) l'omission de facteurs pertinents qui pourraient influencer la consom-

2 Modèle de Régression Linéaire Simple

mation, autres que le revenu disponible, comme de la richesse réelle ou des goûts variés, ou des événements imprévus qui incitent les ménages à consommer plus ou moins,

- (ii) une erreur de mesure qui pourrait résulter des ménages qui ne déclarent pas leur consommation ou leur revenu avec précision,
- (iii) un mauvais choix de la relation linéaire entre la consommation et le revenu, lorsque la vraie relation peut être non linéaire.

Ces différentes causes du terme d'erreur auront des effets différents sur la distribution de cette erreur. Dans ce qui suit, nous ne considérons que des erreurs qui satisfont certaines hypothèses restrictives. **Dans les chapitres ultérieurs, nous détendons ces hypothèses pour tenir compte des types plus généraux des termes d'erreur.**

Dans la vie réelle, β_0 et β_1 ne sont pas connus et doivent être estimés à partir des données observées (X_i, Y_i) pour $i = 1, \dots, n$. Cela signifie également que la vraie ligne $(\beta_0 + \beta_1 X)$ ainsi que les erreurs réelles (les ε_i 's) ne sont pas observables. Dans ce cas, β_0 et β_1 pourraient être estimés par la meilleure ligne d'ajustement à travers les données. Différents chercheurs peuvent tracer différentes lignes à travers les mêmes données. Qu'est-ce qui rend une ligne meilleure qu'une autre ? Une mesure d'une aberration est la quantité d'erreur du Y_i observée à la ligne estimée, notons cette dernière par $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, où le chapeau ($\hat{\cdot}$) désigne une estimation du paramètre ou de la variable appropriée. Chaque observation (X_i, Y_i) aura une erreur observable correspondante qui lui sera attachée, que nous appellerons $e_i = Y_i - \hat{Y}_i$, voir la figure (2.6b). En d'autres termes, nous obtenons \hat{Y}_i estimée (c-à-d \hat{Y}_i) correspondant à chaque X_i à partir de la droite estimée, $\hat{\beta}_0 + \hat{\beta}_1 X_i$.

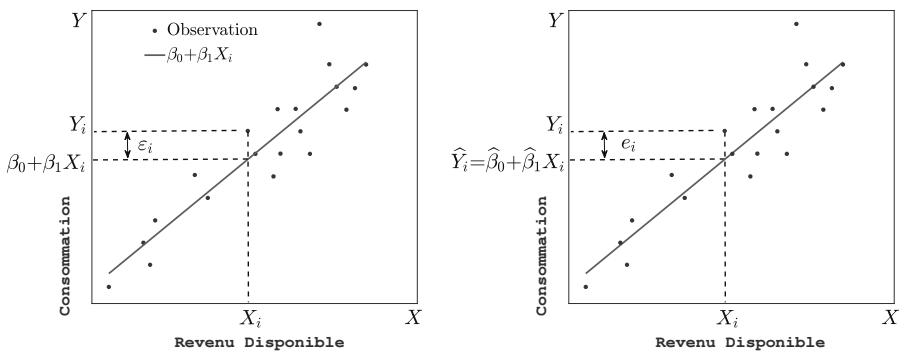


FIGURE 2.6 – Fonction de la consommation

2 Modèle de Régression Linéaire Simple

Ensuite, nous trouvons notre erreur d'estimation de Y_i , en soustrayant Y_i réel du \hat{Y}_i estimé. La seule différence entre la figure (2.6a) et la figure (2.6b) est que la figure (2.6a) trace la vraie ligne de consommation qui est inconnue au chercheur, alors que la figure (2.6b) est une ligne de consommation supposée tirée à travers les données. Par conséquent, alors que les ε_i 's sont non observables, les e_i 's sont observables. Notez qu'il y aura n erreurs pour chaque ligne, une erreur correspondant à chaque observation.

De même, il y aura un autre ensemble de n erreurs pour une autre ligne devinée tirée à travers les données. Pour chaque ligne devinée, on peut résumer ses erreurs correspondantes par un seul nombre qui est *la somme des carrés de ces erreurs*, ce qui semble être un critère naturel pour pénaliser une mauvaise estimation. Notez qu'une simple somme de ces erreurs n'est pas un bon choix pour une mesure d'écart puisque les erreurs positives finissent par annuler les erreurs négatives quand les deux doivent être comptées dans notre mesure. Cependant, cela ne signifie pas que la somme des erreurs au carré est la seule mesure de l'écart. D'autres mesures comprennent la somme des erreurs absolues, mais cette dernière mesure est mathématiquement plus difficile à gérer. Une fois la mesure de l'écart choisi, on peut alors estimer β_0 et β_1 en minimisant cette mesure. En fait, c'est l'idée derrière « l'estimation des moindres carrés ».

2.3.2 Estimation des Moindres Carrés et les Hypothèses Classiques

La méthode des moindres carrés minimise **la somme des carrés des écarts résiduels**, où les résidus sont donnés par

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \quad i = 1, 2, \dots, n$$

et $\hat{\beta}_0$ et $\hat{\beta}_1$ désignent des estimations sur les paramètres de régression β_0 et β_1 , respectivement. La somme des carrés des résidus notée par

$$SCR = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

est minimisée par les deux conditions de premier ordre :

$$\frac{\partial \left(\sum_{i=1}^n e_i^2 \right)}{\partial \beta_0} = -2 \sum_{i=1}^n e_i = 0 \text{ ou } \sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n X_i = 0 \quad (2.2)$$

2 Modèle de Régression Linéaire Simple

$$\frac{\partial \left(\sum_{i=1}^n e_i^2 \right)}{\partial \beta_1} = -2 \sum_{i=1}^n e_i X_i = 0 \text{ ou } \sum_{i=1}^n Y_i X_i - \hat{\beta}_0 \sum_{i=1}^n X_i - \hat{\beta}_1 \sum_{i=1}^n X_i^2 = 0 \quad (2.3)$$

Résoudre les équations normales des moindres carrés données en (2.2) et (2.3) pour obtenir $\hat{\beta}_0$ et $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad \text{et} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (2.4)$$

où

$$\bar{Y} = \sum_{i=1}^n \frac{Y_i}{n}, \quad \bar{X} = \sum_{i=1}^n \frac{X_i}{n}, \quad \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$$

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}$$

Ces estimateurs sont des estimateurs des moindres carrés ordinaires (MCO). Les résidus $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ satisfont automatiquement les deux relations (2.2) et (2.3). La première relation stipule que (i) $\sum_{i=1}^n e_i = 0$ la somme des résidus est nulle. Cela est vrai tant qu'il y a une constante dans la régression. Cette propriété numérique des résidus des moindres carrés implique également que la ligne de régression estimée passe à travers le point moyen (barycentre) de l'échantillon (\bar{X}, \bar{Y}) , et on obtient à partir de (2.2) $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$. La deuxième relation indique que : (ii) $\sum_{i=1}^n e_i X_i = 0$, les résidus et la variable explicative ne sont pas corrélés. Les autres *propriétés numériques* que les estimateurs MCO satisfont sont les suivantes : (iii) $\sum_{i=1}^n \hat{Y}_i = \sum_{i=1}^n Y_i$ et (iv) $\sum_{i=1}^n e_i \hat{Y}_i = 0$. La propriété (iii) indique que la somme des \hat{Y}_i 's estimés ou les \hat{Y}_i 's prévus de l'échantillon est égale à la somme des Y_i 's réels. La propriété (iv) indique que les résidus des MCO et les \hat{Y}_i 's prédicts ne sont pas corrélés.

Afin d'étudier les propriétés statistiques des estimateurs MCO β_0 et β_1 , nous devons imposer des hypothèses statistiques sur le modèle générant les données.

Hypothèse 1 : $E(\varepsilon_i) = 0$ pour tout $i = 1, \dots, n$, c'est-à-dire, les erreurs ont une moyenne nulle. Cette hypothèse est nécessaire pour s'assurer qu'en moyenne nous sommes sur la droite réelle.

2 Modèle de Régression Linéaire Simple

Pour voir ce qui se passe si $E(\varepsilon_i) \neq 0$, considérons le cas où les ménages sous-déclarent constamment leur consommation d'un montant constant de δ Dirhams, alors que leur revenu est mesuré avec précision, par exemple via leurs bulletins de paies. Dans ce cas,

$$(\text{Consommation observée}) = (\text{Consommation réelle}) - \delta$$

et notre équation de régression est vraiment

$$(\text{Consommation réelle})_i = \beta_0 + \beta_1 (\text{Revenu})_i + \varepsilon_i$$

Mais nous observons,

$$(\text{Consommation observée})_i = \beta_0 + \beta_1 (\text{Revenu})_i + \varepsilon_i - \delta$$

Cela peut être considéré comme l'ancienne équation de régression avec un nouveau terme d'erreur $\varepsilon_i^* = \varepsilon_i - \delta$. En utilisant le fait que $\delta > 0$ et $E(\varepsilon_i) = 0$, on obtient $E(\varepsilon_i^*) = -\delta < 0$. Cela signifie que pour tous les ménages ayant le même revenu, disons 10000 Dhs, leur consommation observée sera en moyenne inférieure à celle prédictive à partir de la vraie ligne $[\beta_0 + \beta_1 (10000 \text{Dhs})]$ d'un montant δ . Heureusement, on peut traiter ce problème de la constante δ mais pas celui de la moyenne nulle des erreurs et ce en reparamétrisant le modèle comme suit

$$(\text{Consommation observée})_i = \beta_0^* + \beta_1 (\text{Revenu})_i + \varepsilon_i$$

où $\beta_0^* = \beta_0 - \delta$. Dans ce cas, $E(\varepsilon_i) = 0$, ainsi β_0^* et β_1 peuvent être estimés à partir de la régression. Notez que si β_0^* est estimable, alors β_0 et δ sont non estimables. Notez également que pour tous les ménages à revenu de 10000 Dhs, leur consommation moyenne est de $[(\beta_0 - \delta) + \beta_1 (10000 \text{Dhs})]$.

Hypothèse 2 : Homoscédasticité ; c'est-à-dire $\text{var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$, pour tout $i = 1, \dots, n$, c'est-à-dire, les erreurs ont une variance constante. Cela assure que chaque observation est également fiable.

Pour voir ce que cette supposition signifie, considérons le cas où $\text{var}(\varepsilon_i) = \sigma_i^2$ pour tout $i = 1, \dots, n$. Dans ce cas, chaque observation a une variance

2 Modèle de Régression Linéaire Simple

différente. Une observation avec une variance importante est moins fiable qu'une observation avec une variance plus faible. Mais, comment cette différence peut-elle se produire ? Dans le cas de la consommation, les ménages ayant un revenu disponible important (un X_i grand, disons 20000 Dhs) peuvent économiser davantage (ou emprunter plus pour dépenser plus) que les ménages à revenu plus faible (un petit X_i , disons 10000 Dhs). Dans ce cas, la variation de la consommation pour le ménage à revenu de 20000 Dhs sera beaucoup plus grande que pour le ménage à revenu de 10000 Dhs. Par conséquent, la variance correspondante pour l'observation de 20000 Dhs sera plus grande que pour l'observation de 10000 Dhs.

Hypothèse 3 : $E(\varepsilon_i \varepsilon_j) = 0$ pour $i \neq j$ et $i, j = 1, \dots, n$, c'est-à-dire, les erreurs ne sont pas corrélées. La connaissance de la i-ème erreur ne nous dit rien sur la j-ème erreur car $i \neq j$.

Pour l'exemple de la consommation, l'erreur imprévue qui a mené le i-ème ménage (comme la visite des parents) à consommer plus n'a rien à voir avec les erreurs imprévues de tout autre ménage. Ceci est susceptible de s'appliquer à un échantillon aléatoire de ménages. Cependant, il est moins probable qu'il soit retenu pour une étude chronologique de la consommation pour l'économie globale, où une erreur en 1945, soit une année de guerre, est susceptible d'affecter la consommation pendant plusieurs années après cela. Dans ce cas, nous disons que l'erreur de 1945 est liée aux erreurs de 1946, 1947, etc.

Hypothèse 4 : La variable explicative X est non-stochastique, c'est-à-dire fixée dans des échantillons répétés et, par conséquent, non corrélée

avec les erreurs. De plus, $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \neq 0$ a une limite finie lorsque n tend vers l'infini.

Cette hypothèse indique que nous avons au moins deux valeurs distinctes pour X . Cela est logique, car nous avons besoin d'au moins deux points distincts pour tracer une ligne droite. Sinon $\bar{X} = X$, la valeur

commune, et $X - \bar{X} = 0$, qui viole $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \neq 0$. En pratique, on a toujours plusieurs valeurs distinctes de X . Plus important encore, cette hypothèse implique que X n'est pas une variable aléatoire et n'est donc pas corrélée avec les erreurs.

Cependant, et pour être réaliste, nous devrions assouplir l'hypothèse d'un X non-stochastique. Fondamentalement, X devient une variable aléatoire et nos hypothèses doivent être reformulées *conditionnellement* à l'ensemble des X observés. C'est le cas le plus réaliste avec des données

2 Modèle de Régression Linéaire Simple

économiques. L'hypothèse de la moyenne nulle devient $E(\varepsilon_i/X) = 0$, l'hypothèse de variance constante devient $\text{var}(\varepsilon_i/X) = \sigma^2$, l'hypothèse de non-corrélation des erreurs devient $E(\varepsilon_i\varepsilon_j/X) = 0$ pour $i \neq j$. L'espérance conditionnelle ici est par rapport à toute observation sur X_i pour $i = 1, \dots, n$. Bien sûr, on peut montrer que si $E(\varepsilon_i/X) = 0$ pour tout i , alors X_i et ε_i ne sont pas corrélés. L'inverse n'est pas nécessairement vrai. Cependant, deux variables aléatoires, disons ε_i et X_i , pourraient ne pas être corrélées, c'est-à-dire non liées linéairement lorsqu'elles sont non linéairement reliées, c'est-à-dire on a $\varepsilon_i = X_i^2$. Par conséquent, $E(\varepsilon_i/X_i) = 0$ est une hypothèse plus forte que le fait que ε_i et X_i ne sont pas corrélés. Selon la loi des espérances itératives, $E(\varepsilon_i/X) = 0$ implique que $E(\varepsilon_i) = 0$. Cela implique aussi que ε_i n'est corrélé avec aucune fonction de X_i . C'est une supposition plus forte que le fait que ε_i n'est pas corrélée avec X_i . Donc, conditionnellement à X_i , la moyenne des erreurs est nulle et ne dépend pas de X_i . Dans ce cas, $E(Y_i/X_i) = \beta_0 + \beta_1 X_i$ est linéaire en β_0 et β_1 et est supposée être *la vraie espérance conditionnelle* de Y sachant X .

Pour voir ce que signifie une violation de l'hypothèse 4, supposons que X est une variable aléatoire et que X et ε sont positivement corrélés, alors dans l'exemple de consommation, les ménages dont le revenu est supérieur au revenu moyen seront associés à des erreurs supérieures à zéro, et donc des erreurs positives. De même, les ménages ayant un revenu inférieur au revenu moyen seront associés à des erreurs inférieures à leur moyenne nulle, et donc à des erreurs négatives. Cela signifie que les erreurs sont systématiquement affectées par les valeurs de la variable explicative et que la dispersion des données ressemblera à la figure (2.7). Notez que si nous effaçons maintenant la vraie ligne ($\beta_0 + \beta_1 X$) et estimons cette ligne à partir des données, la ligne des moindres carrés tracée à travers les données va avoir une plus petite intersection et une pente plus grande que celles de la vraie ligne. La dispersion devrait ressembler à la figure (2.8) où les erreurs sont des variables aléatoires, non corrélées avec les X_i 's, tirées d'une distribution avec une moyenne nulle et une variance constante. Les hypothèses 1 et 4 assurent que $E(Y_i/X_i) = \beta_0 + \beta_1 X_i$, c'est-à-dire, en moyenne, nous sommes sur la vraie ligne.

Nous générerons maintenant un ensemble de données qui satisfait aux quatre hypothèses classiques. Supposons que β_0 et β_1 prennent les valeurs arbitraires, disons 10 et 0,5 respectivement, et considérons un ensemble de 20 X 's fixes, disons les classes de revenu de 10 à 105 (en milliers de dirhams), par tranches de 5, 10, 15, 20, 25,..., 105. Notre variable de consommation Y_i est construite comme $(10 + 0.5X_i + \varepsilon_i)$ où ε_i

2 Modèle de Régression Linéaire Simple

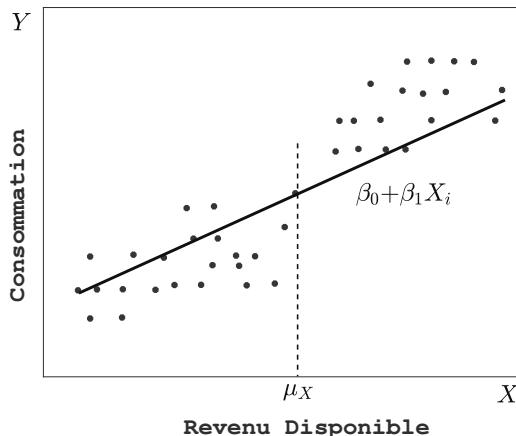


FIGURE 2.7 – **Fonction de consommation avec $\text{cov}(X, \varepsilon) > 0$**

est une erreur tirée aléatoirement d'une distribution avec une moyenne nulle et une variance constante, soit $\sigma^2 = 9$. Les ordinateurs génèrent des nombres aléatoires avec diverses distributions.

Dans ce cas, la figure (2.8) décrirait nos données, avec la vraie ligne étant $10 + 0,5X$ et les ε_i aléatoires qui sont par construction indépendants et identiquement distribués avec une moyenne nulle et une variance égale à 9. Pour chaque ensemble de 20 ε_i 's est généré aléatoirement, étant donné les X_i 's fixés, nous obtenons un ensemble correspondant de 20 Y_i 's à partir de notre modèle de régression linéaire. C'est ce que nous voulons dire dans l'hypothèse 4 quand nous disons que les X 's sont fixés dans des échantillons répétés. Les expériences de Monte Carlo génèrent un grand nombre d'échantillons, disons 1000, de la manière décrite ci-dessus. Pour chaque ensemble de données généré, les moindres carrés peuvent être réalisés et les propriétés des estimateurs résultants qui sont déduits analytiquement dans le reste de ce chapitre, peuvent être vérifiées. Par exemple, la moyenne des 1000 estimations de β_0 et β_1 peut être comparée à leurs vraies valeurs pour voir si ces estimations des moindres carrés sont sans biais. Notez ce qui va arriver à la figure 2.8 si $E(\varepsilon_i) = -\delta$ où $\delta > 0$, ou $\text{var}(\varepsilon_i) = \sigma_i^2$ pour $i = 1, \dots, n$. Dans le premier cas, la moyenne de $f(\varepsilon)$, la fonction de densité de probabilité de ε , décale la vraie ligne ($10 + 0,5X$) de $-\delta$. En d'autres termes, nous pouvons penser que les distributions des ε_i 's, illustrées à la figure (2.8), sont centrées sur une nouvelle ligne imaginaire parallèle à la vraie ligne mais inférieure d'une distance δ . Cela signifie que l'on est plus susceptible de tirer des perturbations négatives que les perturbations positives,

2 Modèle de Régression Linéaire Simple

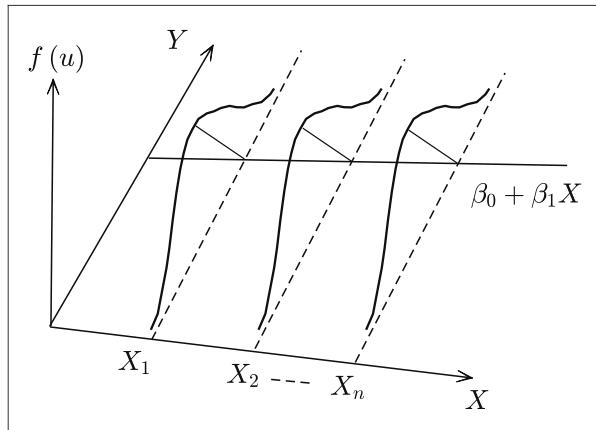


FIGURE 2.8 – Erreurs aléatoires autour de la droite de régression

et les Y_i 's observés sont plus susceptibles d'être en dessous de la vraie ligne qu'au-dessus. Dans le second cas, chaque $f(\varepsilon_i)$ aura une variance différente, donc la dispersion de cette fonction de densité de probabilité variera avec chaque observation. Dans ce cas, la figure (2.8) aura une distribution pour les ε_i 's qui a une dispersion différente pour chaque observation. En d'autres termes, si les ε_i 's sont normalement distribués, alors ε_1 est tiré d'une distribution $N(0, \sigma_1^2)$, alors que ε_2 est tiré d'une distribution $N(0, \sigma_2^2)$, et ainsi de suite.

Hypothèse 5 : $\text{cov}(X_i, \varepsilon_i) = 0$ pour tout $i = 1, \dots, n$; toutes les variables explicatives ne sont pas corrélées avec le terme d'erreur. On suppose que les valeurs observées des variables explicatives sont indépendantes des valeurs du terme d'erreur.

Si une variable explicative et le terme d'erreur étaient plutôt corrélés l'un à l'autre, les estimations des MCO seraient susceptibles d'attribuer au X une partie de la variation de Y qui provenait en fait du terme d'erreur. Si le terme d'erreur et X étaient positivement corrélés par exemple, alors le coefficient estimé serait probablement plus élevé qu'il ne l'aurait été autrement (biaisé à la hausse), car les MCO attribuerait à tort la variation de Y causée par ε à X . Par conséquent, il est important de s'assurer que les variables explicatives ne sont pas corrélées avec le terme d'erreur.

L'hypothèse 5 est violée souvent lorsqu'un chercheur omet une variable indépendante importante du modèle. L'une des principales composantes du terme d'erreur stochastique est les variables omises, donc si une variable a été omise, le terme d'erreur changera lorsque la variable omise

2 Modèle de Régression Linéaire Simple

change. Si cette variable omise est corrélée avec une variable indépendante incluse (comme cela arrive souvent en économie), le terme d'erreur est également corrélé avec cette variable indépendante. Lorsqu'on viole l'hypothèse 5, MCO attribuera l'impact de la variable omise à la variable incluse, dans la mesure où les deux variables sont corrélées.

2.3.3 Propriétés Statistiques des Estimateurs des Moindres Carrés Estimateurs Sans Biais

Compte tenu des hypothèses 1 et 4, il est facile de montrer que $\hat{\beta}_1$ est un estimateur non biaisé de β_1 .

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\varepsilon}$$

$$Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (\varepsilon_i - \bar{\varepsilon})$$

En remplaçant $(Y_i - \bar{Y})$ dans l'équation (2.4) on peut écrire

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X}) \varepsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &\text{car } \bar{\varepsilon} \sum_{i=1}^n (X_i - \bar{X}) = \bar{\varepsilon} \sum_{i=1}^n X_i - \bar{\varepsilon} n \bar{X} = \bar{\varepsilon} n \bar{X} - \bar{\varepsilon} n \bar{X} = 0. \end{aligned} \quad (2.5)$$

En prenant les espérances des deux côtés de (2.5) et en utilisant les hypothèses 1 et 4, on peut montrer que $E(\hat{\beta}_1) = \beta_1$.

Une démonstration similaire peut être utilisée pour montrer que $E(\hat{\beta}_0) = \beta_0$.

En fait,

$$\left. \begin{aligned} \bar{Y} &= \beta_0 + \beta_1 \bar{X} + \bar{\varepsilon} \\ \bar{Y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \end{aligned} \right\} \implies \hat{\beta}_0 = \beta_0 + \bar{\varepsilon} - (\hat{\beta}_1 - \beta_0) \bar{X}$$

$$E(\hat{\beta}_0) = \beta_0 + E(\bar{\varepsilon}) - E((\hat{\beta}_1 - \beta_0) \bar{X}) = \beta_0$$

car $E(\bar{\varepsilon}) = 0$ et $E(\hat{\beta}_1 - \beta_1) = 0$

2 Modèle de Régression Linéaire Simple

où la dernière égalité utilise les hypothèses 2 et 3, c'est-à-dire que les ε_i 's ne sont pas corrélés entre eux et que leur variance est constante. Notons que la variance de l'estimateur des MCO de β_1 dépend de σ^2 , la variance des erreurs dans le modèle réel, et de la variation de X . Plus la variation de X est grande, plus $\sum_{i=1}^n (X_i - \bar{X})^2$ est grande et plus petite est la variance de $\hat{\beta}_1$.

Consistance (ou Convergence)

Ensuite, nous montrons que $\hat{\beta}_1$ est cohérent pour β_1 . Une condition suffisante de consistance est que $\hat{\beta}_1$ est non biaisé et que sa variance tend vers zéro lorsque n tend vers l'infini. Nous avons déjà montré que $\hat{\beta}_1$ est non biaisé, il reste à montrer que sa variance tend vers zéro alors que n tend vers l'infini.

On calcule la variance de $\hat{\beta}_1$ à partir de (2.5),

$$\text{var}(\hat{\beta}_1) = E(\hat{\beta}_1 - \beta_1)^2 = E\left(\frac{\sum_{i=1}^n (X_i - \bar{X}) \varepsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)^2$$

On supposant un changement de variable $Z_i = \frac{\sum_{i=1}^n (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$, on obtient :

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (2.6)$$

Alors

$$\lim_{n \rightarrow \infty} \text{var}(\hat{\beta}_1) = \lim_{n \rightarrow \infty} \left[\frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] = \lim_{n \rightarrow \infty} \left[\frac{\sigma^2/n}{\sum_{i=1}^n (X_i - \bar{X})^2/n} \right] = 0$$

où la deuxième égalité découle du fait que $(\sigma^2/n) \rightarrow 0$ et $\left(\sum_{i=1}^n (X_i - \bar{X})^2/n \right) \neq 0$ qui a une limite finie, voir l'hypothèse 4. Par conséquent, $\hat{\beta}_1$ est un estimateur consistant de β_1 .

De même, on peut montrer que $\hat{\beta}_0$ est non biaisé et consistant avec

2 Modèle de Régression Linéaire Simple

une variance égale à

$$var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] \quad (2.7)$$

$$\lim_{n \rightarrow \infty} var(\hat{\beta}_0) = 0$$

On remarque que

$$var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{X}^2 var(\hat{\beta}_1) \quad \text{et} \quad cov(\hat{\beta}_0, \hat{\beta}_1) = -\bar{X} var(\hat{\beta}_1)$$

Meilleur Estimateur Linéaire Sans Biais (BLUE)

En utilisant (2.5) on peut écrire $\hat{\beta}_1$ comme

$$\sum_{i=1}^n w_i Y_i \text{ où } w_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Ceci prouve que $\hat{\beta}_1$ est une combinaison linéaire des Y_i 's, avec des poids satisfaisant les propriétés suivantes :

$$\sum_{i=1}^n w_i = 0; \quad \sum_{i=1}^n w_i X_i = 1; \quad \sum_{i=1}^n w_i^2 = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (2.8)$$

Le théorème suivant montre que parmi tous les estimateurs linéaires sans biais de β_1 , c'est $\hat{\beta}_1$ qui a la plus petite variance. Ceci est connu comme le théorème de Gauss-Markov.

Théorème 1 : Considérons tout estimateur linéaire arbitraire $\tilde{\beta}_1 = \sum_{i=1}^n a_i Y_i$ pour β_1 , où les a_i 's désignent des constantes arbitraires. Si $\tilde{\beta}_1$ est un estimateur non biaisé de β_1 , et les hypothèses 1 à 4 sont satisfaites, alors $var(\tilde{\beta}_1) \geq var(\hat{\beta}_1)$.

Preuve :

En substituant Y_i de (2.1) en $\tilde{\beta}_1$, on obtient $\tilde{\beta}_1 = \alpha \sum_{i=1}^n a_i + \beta \sum_{i=1}^n a_i X_i + \sum_{i=1}^n a_i \varepsilon_i$.

Pour que $\tilde{\beta}_1$ soit sans biais pour β_1 , il faut que $E(\tilde{\beta}_1) = \alpha \sum_{i=1}^n a_i + \beta \sum_{i=1}^n a_i X_i$.

2 Modèle de Régression Linéaire Simple

$\beta_1 \sum_{i=1}^n a_i X_i = \beta_1$ pour toutes les observations $i = 1, \dots, n$. Cela signifie que $\sum_{i=1}^n a_i = 0$ et $\sum_{i=1}^n a_i X_i = 1$ pour tout $i = 1, \dots, n$.

Donc, $\tilde{\beta}_1 = \beta_1 + \sum_{i=1}^n a_i \varepsilon_i$ avec $\text{var}(\tilde{\beta}_1) = \text{var}\left(\sum_{i=1}^n a_i \varepsilon_i\right) = \sigma^2 \sum_{i=1}^n a_i^2$ où la dernière égalité découle des hypothèses 2 et 3. Mais les a_i 's sont des constantes qui diffèrent des w_i 's ; les poids de l'estimateur des MCO, par certains d'autres constantes, disons d_i 's, c'est-à-dire $a_i = w_i + d_i$ pour $i = 1, \dots, n$.

En utilisant les propriétés des a_i 's et w_i , on peut en déduire des propriétés similaires sur les d_i 's, c'est-à-dire $\sum_{i=1}^n d_i = 0$ et $\sum_{i=1}^n d_i X_i = 0$. En fait

$$\sum_{i=1}^n a_i^2 = \sum_{i=1}^n d_i^2 + \sum_{i=1}^n w_i^2 + 2 \sum_{i=1}^n w_i d_i$$

$$\text{où } \sum_{i=1}^n w_i d_i = \frac{\sum_{i=1}^n x_i d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0. \text{ Cela découle de la définition de } w_i$$

et le fait que $\sum_{i=1}^n d_i = \sum_{i=1}^n d_i X_i = 0$. Par conséquent

$$\text{var}(\tilde{\beta}_1) = \sigma^2 \sum_{i=1}^n a_i^2 = \sigma^2 \sum_{i=1}^n d_i^2 + \sigma^2 \sum_{i=1}^n w_i^2 = \text{var}(\hat{\beta}_1) + \sigma^2 \sum_{i=1}^n d_i^2$$

Puisque $\sigma^2 \sum_{i=1}^n d_i^2$ est non négatif, cela prouve que $\text{var}(\tilde{\beta}_1) \geq \text{var}(\hat{\beta}_1)$, avec l'égalité ne tenant que si $d_i = 0$ pour tout $i = 1, \dots, n$, c'est-à-dire seulement si $a_i = w_i$, auquel cas $\tilde{\beta}_1$ se réduit à $\hat{\beta}_1$. Par conséquent, tout estimateur linéaire de β_1 , comme $\tilde{\beta}_1$ qui est sans biais pour β_1 a une variance au moins aussi grande que $\text{var}(\hat{\beta}_1)$. Cela prouve que $\hat{\beta}_1$ est BLUE (Best Linear Unbiased Estimator), meilleur parmi tous les estimateurs linéaires sans biais de β_1 .

De même, on peut montrer que $\hat{\beta}_0$ est linéaire en Y_i et a la plus petite variance parmi tous les estimateurs linéaires sans biais de β_0 , si les hypothèses 1 à 4 sont satisfaites. Ce résultat implique que l'estimateur des MCO de β_0 est également BLUE.

2 Modèle de Régression Linéaire Simple

2.3.4 Estimation de la Variance des Erreurs σ_ε^2

La variance des erreurs de régression σ^2 est inconnue et doit être estimée. En fait, la variance de $\hat{\beta}_1$ et celle de $\hat{\beta}_0$ dépendent de σ_ε^2 , voir

(2.6). Un estimateur sans biais pour σ^2 est $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$. Pour prouver cela, nous avons besoin du fait que

$$\begin{aligned} e_i &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X}) \\ &= (\beta_1 - \hat{\beta}_1) (X_i - \bar{X}) + (\varepsilon_i - \bar{\varepsilon}) \end{aligned}$$

où $\bar{\varepsilon} = \frac{\sum_{i=1}^n \varepsilon_i}{n}$. La deuxième égalité substitue $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ et la troisième égalité remplace $(Y_i - \bar{Y}) = \beta_1 (X_i - \bar{X}) + (\varepsilon_i - \bar{\varepsilon})$. Par conséquent

$$\begin{aligned} \sum_{i=1}^n e_i^2 &= (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\ &\quad + \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 - 2(\hat{\beta}_1 - \beta_1) \sum_{i=1}^n (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon}) \end{aligned}$$

et

$$\begin{aligned} E\left(\sum_{i=1}^n e_i^2\right) &= \sum_{i=1}^n (X_i - \bar{X})^2 var(\hat{\beta}_1) + (1-n)\sigma^2 - 2 \frac{E\left(\sum_{i=1}^n (X_i - \bar{X}) \varepsilon_i\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \sigma^2 + (n-1)\sigma^2 - 2\sigma^2 = (n-2)\sigma^2 \end{aligned}$$

où la première égalité utilise le fait que $E\left(\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2\right) = (n-1)\sigma^2$
 et $\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}) \varepsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$. La deuxième égalité utilise le fait que

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad \text{et} \quad E\left(\sum_{i=1}^n (X_i - \bar{X}) \varepsilon_i\right)^2 = \sigma^2 \sum_{i=1}^n (X_i - \bar{X})^2$$

2 Modèle de Régression Linéaire Simple

Donc,

$$E(\hat{\sigma}^2) = E\left(\frac{\sum_{i=1}^n e_i^2}{n-2}\right) = \sigma^2$$

Intuitivement, l'estimateur de σ^2 pourrait être obtenu à partir de $\frac{\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2}{n-1}$ si les erreurs réelles étaient connues. Puisque les ε_i 's ne sont pas connus, des estimations cohérentes sont utilisées. Ce sont les e_i 's. Puisque $\sum_{i=1}^n e_i = 0$, notre estimateur de σ^2 devient $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$. En prenant des espérances, nous trouvons que le diviseur correct devrait être $(n-2)$ et non $(n-1)$ pour que cet estimateur soit sans biais pour σ^2 . Ceci est plausible, puisque nous avons estimé deux paramètres β_0 et β_1 pour obtenir les e_i 's, et il ne reste que $n-2$ éléments d'information indépendants dans les données. Pour prouver ce fait, considérons les équations normales des MCO données en (2.2) et (2.3). Ces équations représentent deux relations impliquant les e_i 's. Donc, connaissant $(n-2)$ des e_i 's, nous pouvons déduire les deux e_i 's restants à partir de (2.2) et (2.3).

2.3.5 Conséquences de la Normalité des Erreurs

Quelle que soit la forme de la distribution des termes d'erreur ε_i ; (et donc de Y_i), la méthode des moindres carrés fournit des estimateurs ponctuels non biaisés de β_0 et β_1 , qui présentent une variance minimale entre tous les estimateurs linéaires non biaisés. Pour établir des estimations d'intervalle et effectuer des tests, nous devons toutefois émettre une hypothèse sur la forme de la distribution de ε_i . L'hypothèse standard est que les termes d'erreur ε_i ; sont normalement distribués, et nous allons l'adopter ici. Un terme d'erreur normal simplifie grandement la théorie de l'analyse de régression et, comme nous l'expliquerons tout à l'heure, est justifiable dans de nombreuses situations réelles où l'analyse de régression est appliquée.

Le modèle de régression à erreur normale est le suivant :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n \quad (2.9)$$

où

ε_i , sont indépendants $N(0, \sigma^2)$

2 Modèle de Régression Linéaire Simple

- (1) Le symbole $N(0, \sigma^2)$ représente la distribution normale, avec une moyenne égale à 0 et une variance σ^2 .
- (2) Le modèle à erreur normale (2.9) est identique au modèle de régression (2.1) avec une distribution d'erreur non spécifiée, sauf que le modèle (2.9) suppose que les erreurs ε_i ; sont normalement distribués.
- (3) Étant donné que le modèle de régression (2.9) suppose que les erreurs sont normalement distribuées, l'hypothèse de non corrélation de ε_i ; dans le modèle de régression (2.1) devient une indépendance dans le modèle d'erreur normale. Par conséquent, l'issue d'un procès n'a aucun effet sur l'erreur générée pour un autre procès, qu'il soit positif ou négatif, petit ou grand.
- (4) Le modèle de régression (2.9) implique que les Y_i sont des variables aléatoires normales indépendantes, avec une moyenne de $E(Y_i) = \beta_0 + \beta_1 X_i$ et une variance σ^2 . La figure 2.9 illustre ce modèle à erreur normale. Chacune des distributions de probabilité de Y dans la figure 2.9 est normalement distribuée, avec une variabilité constante, et la fonction de régression est linéaire.
- (5) L'hypothèse de normalité pour les termes d'erreur est justifiable dans de nombreuses situations car les termes d'erreur représentent souvent les effets de facteurs omis du modèle qui affectent la variable dépendante dans une certaine mesure et qui varient de manière aléatoire sans référence à la variable X .

Exemple : La société Astorias fabrique des équipements de réfrigération ainsi que de nombreuses pièces de rechange. Dans le passé, l'une des pièces de rechange était produite périodiquement en lots de différentes tailles. Lorsqu'un programme d'amélioration des coûts était entrepris, les responsables de la société souhaitaient déterminer la taille optimale du lot pour la production de cette pièce. La production de cette pièce implique la mise en place du processus de fabrication (quelle que soit la taille du lot) et des opérations d'usinage et d'assemblage. La relation entre « la taille du lot » et les « heures de main-d'œuvre » nécessaires à la production du lot a été l'un des éléments clés permettant de déterminer la taille optimale du lot. Pour déterminer cette relation, nous avons utilisé des données sur la taille du lot et les heures de travail pour 25 séries de production récentes. Les conditions de production sont restées stables au cours de la période de six mois au cours de laquelle les 25 séries ont été effectuées et devraient rester les mêmes au cours des trois prochaines années, période de planification pour

2 Modèle de Régression Linéaire Simple

laquelle le programme d'amélioration des coûts est en cours.

Pour la société Astorias, les effets de facteurs tels que le temps écoulé depuis la dernière production, les machines utilisées, la saison de l'année et le personnel employé pourraient varier plus ou moins au hasard d'une campagne à l'autre, indépendamment de la taille du lot. En outre, il pourrait y avoir des erreurs de mesure aléatoires dans l'enregistrement de Y , les heures requises. Dans la mesure où ces effets aléatoires ont un degré d'indépendance mutuelle, le terme d'erreur composite ε_i représentant tous ces facteurs tendrait à respecter le théorème de la limite centrale et la distribution du terme d'erreur se rapprocherait de la normalité à mesure que le nombre d'effets de facteur deviendrait grand.

Une deuxième raison pour laquelle l'hypothèse de normalité des termes d'erreur est souvent justifiable est que les procédures d'estimation et de test à traiter sont basées sur la distribution de Student et ne sont généralement sensibles qu'aux écarts importants par rapport à la normalité. Ainsi, à moins que les écarts par rapport à la normalité ne soient sérieux, en particulier en ce qui concerne l'asymétrie, les coefficients de confiance et les risques d'erreur réels seront proches des niveaux de normalité exacte.

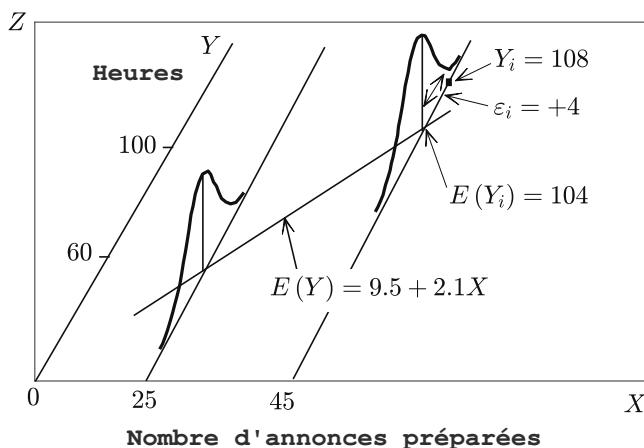


FIGURE 2.9 – Représentation graphique du modèle de régression

2 Modèle de Régression Linéaire Simple

2.3.6 Estimation par Méthode de Maximum de Vraisemblance

Lorsque la forme fonctionnelle de la distribution de probabilité des termes d'erreur est spécifiée, des estimateurs des paramètres β_0 , β_1 et σ^2 peuvent être obtenus par la méthode du maximum de vraisemblance. Essentiellement, la méthode du maximum de vraisemblance choisit comme estimations les valeurs des paramètres les plus cohérents avec les données de l'échantillon.

Hypothèse 5 : Les ε_i 's sont indépendants et identiquement distribués $N(0, \sigma^2)$.

Cette hypothèse nous permet d'obtenir des distributions d'estimateurs et d'autres statistiques de test. En fait, en utilisant (2.5) on peut facilement constater que $\hat{\beta}_1$ est une combinaison linéaire des ε_i 's. Mais, une combinaison linéaire de variables aléatoires normales est aussi une variable aléatoire normale.

$$\text{Donc } \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

$$\text{De même, } \hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n X_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right), \text{ et } Y_i \text{ est } N(\beta_0 + \beta_1 X_i, \sigma^2).$$

De plus, nous pouvons écrire la fonction de densité de probabilité conjointe des ε_i 's comme

$$f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n; \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^n \varepsilon_i^2}{2\sigma^2}\right)$$

Pour obtenir la fonction de vraisemblance, nous faisons la transformation $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$ et notons que le jacobien de la transformation est 1. Par conséquent,

$$f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n; \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right) \quad (2.10)$$

2 Modèle de Régression Linéaire Simple

Prenant le logarithme de cette vraisemblance, nous obtenons

$$\log L(\beta_0, \beta_1, \sigma^2) = -\left(\frac{n}{2}\right) \log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \quad (2.11)$$

En maximisant cette vraisemblance par rapport à β_0 , β_1 et σ^2 on obtient **les estimateurs de maximum de vraisemblance (EMV)**. Cependant, seul le second terme de la log-vraisemblance contient β_0 et β_1 et ce terme (sans le signe négatif) a été déjà minimisé par rapport à β_0 et β_1 dans (2.2) et (2.3) ce qui nous donne les estimateurs des MCO. Donc, $\hat{\beta}_0 = \hat{\beta}_{0(EMV)}$ et $\hat{\beta}_1 = \hat{\beta}_{1(EMV)}$. De même, en dérivant $\log L$ par rapport à σ^2 et en mettant cette dérivée à zéro, on obtient

$$\hat{\sigma}_{(EMV)}^2 = \frac{\sum_{i=1}^n e_i^2}{n}. \text{ Notez que cela ne diffère de } \hat{\sigma}^2 \text{ que dans le numérateur. En fait, } E(\hat{\sigma}_{(EMV)}^2) = \frac{(n-2)\sigma^2}{n} \neq \sigma^2. \text{ Par conséquent, } \hat{\sigma}_{(EMV)}^2 \text{ est biaisé mais notez qu'il est toujours asymptotiquement non biaisé.}$$

Jusqu'à présent, les gains découlant de l'imposition de l'hypothèse 5 sont les suivants : La vraisemblance peut être établie, des estimateurs du maximum de vraisemblance peuvent être dérivés et des distributions peuvent être obtenues pour ces estimateurs. En fait, on peut montrer, en suivant la théorie des statistiques complètes, que $\hat{\beta}_0$ et $\hat{\beta}_1$ et $\hat{\sigma}^2$ sont des estimateurs sans biais de variance minimale de β_0 , β_1 et σ^2 . C'est un résultat plus fort (pour $\hat{\beta}_0$ et $\hat{\beta}_1$) que celui obtenu en utilisant le théorème de Gauss-Markov. Il indique que parmi tous les estimateurs sans biais de β_0 et β_1 , les estimateurs des MCO sont les meilleurs. En d'autres termes, notre ensemble d'estimateurs inclut maintenant *tous* les estimateurs non biaisés et pas seulement les estimateurs *linéaires* sans biais. Ce résultat fort est obtenu au détriment d'une hypothèse distributionnelle plus forte, c'est-à-dire, la normalité. Si la distribution des erreurs n'est pas normale, alors MCO n'est plus EMV. Dans ce cas, EMV sera plus efficace que MCO tant que la distribution des erreurs est correctement spécifiée.

2.4 Inférences dans la Régression et Analyse de Corrélation

Dans cette section, nous abordons d'abord les inférences concernant les paramètres de régression β_0 et β_1 , en considérant à la fois l'estimation par intervalles de ces paramètres et les tests les concernant. Nous discutons ensuite les intervalles de prédiction pour une nouvelle observation

2 Modèle de Régression Linéaire Simple

Y , l'analyse de variance de la régression et l'approche générale des tests linéaires. Enfin, nous examinons le coefficient de corrélation, une mesure d'association entre X et Y lorsque X et Y sont des variables aléatoires.

Tout au long de cette section et, sauf indication contraire, nous supposons que le modèle de régression (2.1) dont les erreurs sont normalement distribuées est applicable. Ce modèle est le suivant :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n \quad (2.12)$$

où

β_0 et β_1 sont des paramètres,

X_i sont des constantes connues,

ε_i sont indépendants et identiquement distribués $N(0, \sigma^2)$.

Nous avons trouvé les distributions de $\hat{\beta}_0$ et $\hat{\beta}_1$, maintenant nous donnons celle de $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - 2} = \frac{SCR}{n - 2}$$

Pour le modèle de régression (2.12), $\frac{SCR}{\sigma^2}$ est distribué selon χ_{n-2}^2 (chi-carré) avec du $(n - 2)$ degrés de liberté et est indépendante de $\hat{\beta}_0$ et $\hat{\beta}_1$.

De même, $\hat{\sigma}^2$ est indépendante de $\hat{\beta}_0$ et $\hat{\beta}_1$. Ceci est utile pour tester des hypothèses. En fait, le gain majeur de l'hypothèse 5 de normalité des erreurs est que nous pouvons effectuer un *test d'hypothèses*.

En normalisant la variable aléatoire normale $\hat{\beta}_1$, on obtient

$$z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}}} \sim N(0, 1)$$

De plus, $\frac{(n - 2) \hat{\sigma}^2}{\sigma^2}$ est distribué selon χ_{n-2}^2 . Par conséquent, on peut diviser z qui est une variable aléatoire $N(0, 1)$, par la racine carrée de $\frac{(n - 2) \hat{\sigma}^2}{\sigma^2}$ divisée rapportée par ses degrés de liberté $(n - 2)$ pour obtenir une statistique t de Student avec $(n - 2)$ degrés de liberté. La statistique

2 Modèle de Régression Linéaire Simple

résultante est

$$t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}$$

où

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Cette statistique peut être utilisée pour appliquer les tests de signification de β_1 comme on va discuter par la suite dans une autre sous-section.

2.4.1 Intervalles de Confiance

On peut obtenir un intervalle de confiance pour β_1 en utilisant le fait que

$$\Pr \left[-t_{n-2}^{\alpha/2} < t_{\hat{\beta}_1}^* < t_{n-2}^{\alpha/2} \right] = 1 - \alpha$$

et en substituant $t_{\hat{\beta}_1}^*$ par sa valeur $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}}$. Puisque les valeurs critiques

sont connues, $\hat{\beta}_1$ et $\hat{\sigma}_{\hat{\beta}_1}$ peuvent être calculés à partir des données, et l'intervalle de confiance de β_1 avec un niveau de confiance $(1 - \alpha)\%$ se présente comme

$$IC_{\beta_1} = \left[\hat{\beta}_1 \pm t_{n-2}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_1} \right]$$

L'intervalles de confiance de l'estimateur β_0 et σ^2 peuvent être construits de manière similaire en utilisant la distribution normale de $\hat{\beta}_0$ et la distribution χ_{n-2}^2 de $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$.

$$IC_{\beta_0} = \left[\hat{\beta}_0 \pm t_{n-2}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_0} \right]$$

où

$$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$$

2.4.2 Les Tests de Signification des Paramètres de Régression

Une approche *alternative mais complémentaire à la méthode des intervalles de confiance* pour tester des hypothèses statistiques est l'approche des tests de signification développée indépendamment par R. A. Fisher et conjointement par Neyman et Pearson. De manière générale, un test de signification est une procédure par laquelle des résultats d'échantillons sont utilisés pour vérifier la véracité ou la fausseté d'une hypothèse nulle. L'idée clé derrière les tests de signification est celle d'une « statistique de test » (estimateur) et de la distribution d'échantillonnage d'une telle statistique sous l'hypothèse nulle. La décision d'accepter ou de rejeter H_0 est prise sur la base de la valeur de la statistique de test obtenue à partir des données disponibles.

Puisque $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}}$ est distribué selon une loi de Student à $(n - 2)$ degrés de liberté, les tests concernant β_1 et β_0 peuvent être réalisés de manière ordinaire à l'aide de la même loi de distribution.

Test Bilatéral

Un analyste des coûts de la société Astorias souhaite tester, à l'aide du modèle de régression (2.12), s'il existe ou non une association linéaire entre les heures de travail et la taille du lot, c'est-à-dire si c'est oui ou non $\beta_1 = 0$. Les deux alternatives possibles sont :

$$\begin{aligned} H_0: \beta_1 &= 0 \\ H_1: \beta_1 &\neq 0 \end{aligned} \tag{2.13}$$

Le test bilatéral consiste à tester l'hypothèse H_0 contre l'hypothèse H_1 , soit à comparer le ratio empirique de Student $|t_{\hat{\beta}_1}^*|$ (dans ce cas $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$) à la valeur de t de Student lue dans la table de la loi de Student à $(n - 2)$ degrés de liberté pour un seuil de probabilité égal à 5%.

La règle de décision avec ce test pour contrôler le niveau de signification est la suivante :

$$\begin{aligned} \text{Si } |t_{\hat{\beta}_1}^*| &\leq t_{n-2}^{\alpha/2} \quad \text{on accepte } H_0 \\ \text{Si } |t_{\hat{\beta}_1}^*| &> t_{n-2}^{\alpha/2} \quad \text{on accepte } H_1 \end{aligned} \tag{2.14}$$

2 Modèle de Régression Linéaire Simple

Si $n - 2 > 30$ alors $t_{\infty}^{\alpha/2} = 1,96$ ¹

Exemple

Reprendons notre exemple de l'impact de l'éducation sur le salaire. Nous savons que $\hat{\beta}_1 = 0.7240$, $\hat{\sigma}_{\hat{\beta}_1} = 0.0700$ et $df = 11$. Si nous supposons que $\alpha = 5\%$, $t_{11}^{\alpha/2} = 2.201$. Si nous supposons que $H_0: \beta_1 = 0.5$ et $H_1: \beta_1 \neq 0.5$, on a :

$$t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0.7240 - 0.5}{0.0700} = 3.2$$

Puisque $|t_{\hat{\beta}_1}^*| > t_{n-2}^{\alpha/2} = t_{11}^{0,025} = 2.201$, la pente β_1 est significativement différente de 0.5.

Test Unilatéral

Supposons que l'analyste ait voulu vérifier si β_1 est positif ou non, pour contrôler le niveau de signification à un seuil de probabilité égal à 5%. Les alternatives seraient alors :

$$\begin{aligned} H_0: \beta_1 &\leq 0 \\ H_1: \beta_1 &> 0 \end{aligned} \tag{2.15}$$

Le test unilatéral consiste donc à tester l'hypothèse H_0 contre l'hypothèse H_1 , soit à comparer le ratio empirique de Student $|t_{\hat{\beta}_1}^*|$ à la valeur de t de Student lue dans la table de la loi de Student à $(n - 2)$ degrés de liberté pour un seuil de probabilité égal à 5%.

La règle de décision avec ce test pour contrôler le niveau de signification est la suivante :

$$\begin{aligned} \text{Si } |t_{\hat{\beta}_1}^*| &\leq t_{n-2}^{\alpha} & \text{on accepte } H_0 \\ \text{Si } |t_{\hat{\beta}_1}^*| &> t_{n-2}^{\alpha} & \text{on accepte } H_1 \end{aligned} \tag{2.16}$$

Il n'y a que de rares occasions où nous souhaitons faire des déductions concernant le paramètre β_0 , l'interception de la droite de régression. Celles-ci se produisent lorsque la portée du modèle inclut $X = 0$. Les tests de signification pour le paramètre β_0 peuvent être réalisés d'une

1. Si le degré de liberté est supérieur à 30, la loi de Student peut être approximée par une loi normale.

2 Modèle de Régression Linéaire Simple

manière similaire aux tests réalisés sur le paramètre β_1 et ce après le

calcul de $\hat{\sigma}_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$ et la statistique empirique de Student $t_{\hat{\beta}_1}^*$ comme signalé avant pour l'intervalle de confiance.

Remarques :

- (1) Parfois, il est souhaitable de tester si β_1 est égal ou non à une valeur spécifiée et non nulle notée ρ , qui peut être une norme historique, une valeur d'un processus comparable ou une spécification technique. Les alternatives sont maintenant :

$$\begin{aligned} H_0; \beta_1 &= \rho \\ H_1; \beta_1 &\neq \rho \end{aligned} \tag{2.17}$$

et la statistique de test appropriée est : $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - \rho}{\hat{\sigma}_{\hat{\beta}_1}}$, et la règle de décision avec ce test est celle du (2.14).

- (2) Si nous rejetons l'hypothèse H_0 pour un test bilatéral, alors nous rejetons forcément (pour un même seuil de probabilité) l'hypothèse H_0 pour un test unilatéral.
- (3) La probabilité critique - risque de rejeter à tort l'hypothèse H_0 - ou encore *risque de premier espèce* est donnée par la valeur de la probabilité α^c telle que : $t_{n-2}^{\alpha^c} = t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$, par la lecture sur la table de Student à $(n - 2)$ degrés de liberté.
- (4) La probabilité critique d'un test unilatéral peut directement se déduire de la probabilité critique d'un test bilatéral par une simple division par 2 $\rightarrow \alpha_{unilatéral}^c = \alpha_{bilatéral}^c / 2$ (opération inverse de celle de la lecture de la table de Student).

Exemple : Pour $t_{\hat{\beta}_1}^* = 2,53$ et $n - 2 = 20$, soit à déterminer α^c tel que $t_{20}^{\alpha^c} = 2,53$. Après la lecture de la table de Student à 20 degrés de liberté, nous trouvons $\alpha^c = 2\%$ pour un test bilatéral et donc 1% pour un test unilatéral.

2.4.3 Inférence sur le Coefficient de Corrélation

L'utilisation principale du modèle de corrélation bivariée est d'étudier la relation entre deux variables. Dans un modèle normal bivarié, le pa-

2 Modèle de Régression Linéaire Simple

ramètre $r_{X,Y}$ fournit des informations sur le degré de relation linéaire entre les deux variables X et Y .

L'estimateur du maximum de vraisemblance du coefficient de corrélation linéaire $r_{X,Y}$, noté $\rho_{X,Y}$, est donné par :

$$\rho_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Cet estimateur est souvent appelé coefficient de corrélation de Pearson. Il s'agit d'un estimateur biaisé de $r_{X,Y}$, (sauf si $\rho_{X,Y} = 0$ ou 1), mais le biais est faible lorsque n est grand. On peut montrer que la plage de $\rho_{X,Y}$ est :

$$-1 \leq \rho_{X,Y} \leq 1$$

En règle générale, les valeurs de $\rho_{X,Y}$ près de 1 indiquent une forte association linéaire positive (directe) entre X et Y , tandis que les valeurs de $\rho_{X,Y}$ près de -1 indiquent une forte association linéaire négative (indirecte). Les valeurs de $\rho_{X,Y}$ proches de 0 indiquent une association linéaire faible ou nulle entre X et Y .

Remarques

- (1) Le coefficient de corrélation r est indépendant de l'origine et de l'échelle ; c'est-à-dire que si nous définissons $X_i^* = aX_i + c$ et $Y_i^* = bY_i + d$, où $a > 0$, $b > 0$ et c et d sont des constantes, alors le r entre X^* et Y^* est identique à celui entre les variables d'origine X et Y .
- (2) Si X et Y sont statistiquement indépendants, le coefficient de corrélation entre eux est égal à zéro. mais si $r = 0$, cela ne signifie pas que deux variables sont indépendantes. En d'autres termes, la corrélation nulle n'implique pas nécessairement l'indépendance.
- (3) C'est une mesure uniquement d'*association linéaire ou de dépendance linéaire* ; cela n'a aucun sens pour décrire les relations non linéaires.
- (4) Bien qu'il s'agisse d'une mesure d'association linéaire entre deux variables, cela n'implique pas nécessairement une relation de cause à effet.
- (5) Dans le contexte de la régression, r^2 est une mesure plus significative que r ,² car le premier nous indique la proportion de variation

2. Dans la modélisation par régression, la théorie sous-jacente indiquera la direction de la causalité entre Y et X , qui, dans le contexte des modèles à équation simple, va généralement de X à Y .

2 Modèle de Régression Linéaire Simple

de la variable dépendante expliquée par la ou les variables explicatives et fournit donc une mesure globale de la mesure dans laquelle la variation d'une variable détermine la variation dans l'autre (r^2 est appelé aussi coefficient de détermination, noté R^2).

Test de Signification du coefficient de corrélation

Il est fréquemment souhaitable de tester si le coefficient de corrélation est nul. Les deux hypothèses pour ce test sont :

$$H_0; r_{X,Y} = 0$$

$$H_1; r_{X,Y} \neq 0$$

Le ratio empirique de Student s'écrit

$$t_{r_{X,Y}}^* = \frac{\rho_{X,Y}}{\sqrt{\frac{1 - \rho_{X,Y}^2}{n - 2}}}$$

La règle de décision avec ce test pour contrôler le niveau de signification est la suivante :

$$\begin{aligned} \text{Si } |t_{r_{X,Y}}^*| &\leq t_{n-2}^{\alpha/2} && \text{on accepte } H_0 \\ \text{Si } |t_{r_{X,Y}}^*| &> t_{n-2}^{\alpha/2} && \text{on accepte } H_1 \end{aligned}$$

Le test consiste à tester l'hypothèse H_0 contre l'hypothèse H_1 , soit à comparer le ratio empirique de Student $|t_{r_{X,Y}}^*|$ à la valeur de t de Student lue dans la table de la loi de Student à $(n - 2)$ degrés de liberté pour un seuil de probabilité égal à 5%.

2.4.4 Analyse de la Variance ANOVA et Mesure de la Qualité d'Ajustement

Nous avons obtenu les estimations des moindres carrés de β_0 , β_1 et σ^2 et trouvé leurs distributions sous l'hypothèse de la normalité des erreurs. Nous avons également appris à tester des hypothèses concernant ces paramètres. Nous passons maintenant à une mesure d'ajustement pour cette droite de régression estimée.

Rappelons que $e_i = Y_i - \hat{Y}_i$ où \hat{Y}_i désigne le Y_i prédit à partir de la ligne de régression par les moindres carrés à la valeur X_i , c'est-à-dire, $\hat{\beta}_0 + \hat{\beta}_1 X_i$. En utilisant le fait que $\sum_{i=1}^n e_i = 0$, on déduit que $\sum_{i=1}^n Y_i = \sum_{i=1}^n$

2 Modèle de Régression Linéaire Simple

\hat{Y}_i , et donc $\bar{Y} = \bar{\hat{Y}}$. Les valeurs réelles et prédictes de Y ont la même moyenne d'échantillon, voir les propriétés numériques des estimateurs des MCO discutés dans la sous-section 2.3.3. Cela est vrai tant qu'il y a une constante dans la régression. En ajoutant et en soustrayant \bar{Y} de e_i , on obtient

$$e_i + \bar{Y} - \bar{Y} = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y}) \quad \text{ou encore} \quad Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + e_i$$

Mettre au carré et la sommation des deux termes de l'équation donnera :

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2 + 2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y}) e_i \quad (2.18)$$

où la dernière égalité découle du fait que

$$(\hat{Y}_i - \bar{Y}) = \beta_1 (X_i - \bar{X}) \quad \text{et} \quad \sum_{i=1}^n (X_i - \bar{X}) e_i$$

$$\text{En fait, } \sum_{i=1}^n (\hat{Y}_i - \bar{Y}) e_i = \sum_{i=1}^n e_i \hat{Y}_i = 0$$

Celà signifie que les résidus des MCO ne sont pas corrélés avec les valeurs prédictes de la régression, voir les propriétés numériques (ii) et (iv) des estimations des MCO discutées dans la sous-section 2.3.2. En d'autres termes, (2.18) affirme que *la somme des carrés totale* (SCT) de Y_i , autour de sa moyenne \bar{Y} , notée $SCT = \sum_{i=1}^n (Y_i - \bar{Y})^2$, peut être décomposée en deux parties : la première est *la somme des carrés expliquée* (SCE), $SCE = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}^2 \sum_{i=1}^n (X_i - \bar{X})^2$ et la seconde est *la somme des carrés des résidus* (SCR), $SCR = \sum_{i=1}^n e_i^2$.

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2 \\ SCT &= SCE + SCR \end{aligned} \quad (2.19)$$

En utilisant cette décomposition, on peut définir le pouvoir explicatif de la régression comme le rapport de la somme des carrés expliquée *SCE* sur la somme des carrés totale *SCT*. En d'autres termes, on définit le

2 Modèle de Régression Linéaire Simple

coefficient R^2 :

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (\text{Coefficient de détermination})$$

et cette valeur est clairement comprise entre 0 et 1. En effet, diviser (2.19) par $\sum_{i=1}^n (Y_i - \bar{Y})^2$, on obtient

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{SCR}{SCT}$$

$SCR = \sum_{i=1}^n e_i^2$ est une mesure d'écart minimisée par les moindres carrés.

Si $\sum_{i=1}^n e_i^2$ est grande, cela signifie que la régression n'explique pas une grande partie de la variation de Y et, par conséquent, la valeur de R^2 serait faible. Alternativement, si $\sum_{i=1}^n e_i^2$ est petite, alors l'ajustement est bon et la valeur de R^2 est grande.

En effet, pour un ajustement parfait, où toutes les observations reposent sur la droite ajustée, $Y_i = \hat{Y}_i$ et $e_i = 0$, ce qui signifie que $\sum_{i=1}^n e_i^2 = 0$ et

$R^2 = 1$. L'autre cas extrême est celui où la somme des carrés de expliquée est nulle $SCE = 0$, en d'autres termes, la régression linéaire n'explique rien de la variation de Y_i . Dans ce cas, $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n e_i^2$ et $R^2 = 0$.

Notez que puisque $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = 0$ implique que $(\hat{Y}_i - \bar{Y}) = 0$ pour tout i , ce qui signifie que $\hat{Y}_i = \bar{Y}$ pour chaque i . La droite de régression ajustée est une ligne horizontale tracée à $Y = \bar{Y}$, et la variable indépendante X n'a aucun pouvoir explicatif dans une relation linéaire avec Y .

Notez que R^2 a deux significations alternatives :

(i) C'est le coefficient de corrélation simple au carré entre Y_i et \hat{Y}_i .

Aussi, pour le cas de régression simple,

(ii) C'est la corrélation simple au carré entre X et Y .

Cela signifie qu'avant d'exécuter la régression de Y sur X , on peut calculer $r_{x,y}^2$ qui à son tour nous indique la proportion de la variation de Y qui sera expliquée par X . Si ce nombre est assez faible, nous avons

2 Modèle de Régression Linéaire Simple

une faible relation linéaire entre Y et X et nous savons qu'un mauvais ajustement résultera si Y est régressé sur X . Il convient de souligner que R^2 est une mesure de l'association *linéaire* entre Y et X . Il pourrait exister, par exemple, une relation quadratique parfaite entre X et Y , mais la ligne des moindres carrés estimée à travers les données est une ligne plate impliquant que $R^2 = 0$. On devrait également se méfier des régressions des moindres carrés avec R^2 très proche de 1. Dans certains cas, nous pouvons ne pas vouloir inclure une constante dans la régression. Dans de tels cas, il convient d'utiliser un R^2 *non centré* comme mesure de qualité d'ajustement. L'annexe de ce chapitre définit R^2 à la fois *centré* et *non centré* et explique la différence entre eux.

Tableau d'Analyse de la Variance

Le tableau 2.1 présente l'analyse de la variance pour un modèle de régression simple.

Source de variation	Somme des Carrés	Degrés de liberté ³	Carrés Moyens
(ddl)			
X	$SCE = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	1	$MCE = \frac{SCE}{1}$
Résidu	$SCR = \sum_{i=1}^n e_i^2$	$n - 2$	$MCR = \frac{SCR}{n - 2}$
Total	$SCT = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$	

TABLE 2.1 – Analyse de la Variance par une Régression Linéaire Simple

Le test d'hypothèse $H_0; \beta_1 = 0$ est équivalent au test d'hypothèse⁴ $H_0; SCE = 0$ (la variable explicative X_i ne contribue pas à l'explication du modèle).

3. Les degrés de liberté correspondent au nombre de valeurs que nous pouvons choisir arbitrairement (par exemple, pour la variabilité totale, connaissant $(n - 1)$ valeurs, nous pourrons en déduire la $n - ième$, puisque nous connaissons la moyenne \bar{Y}).

4. Cela n'est valable que dans le cas du modèle de régression linéaire simple.

2 Modèle de Régression Linéaire Simple

Soit le test d'hypothèses $H_0; SCE = 0$ contre l'hypothèse $H_1; SCE \neq 0$. La statistique de ce test est donnée par⁵ :

$$F^* = \frac{SCE}{\frac{ddl_{SCE}}{SCR}} = \frac{MCE}{MCR} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\frac{1}{\sum_{i=1}^n e_i^2} n - 2}$$

ou encore

$$F^* = \frac{\frac{SCE}{SCR}}{\frac{ddl_{SCE}}{ddl_{SCR}}} = \frac{MCE}{MCR} = \frac{\frac{\sum_{i=1}^n (\bar{Y}_i - \bar{Y})^2}{\sum_{i=1}^n e_i^2}}{\frac{1}{n-2}} = \frac{\frac{R^2}{1-R^2}}{\frac{n-2}{n-2}}$$

La statistique F^* est le rapport de la somme des carrés expliqués par X_i sur la somme des carrés des résidus, chacune de ces sommes étant divisée par son degré de liberté respectif. Ainsi, si la variance expliquée est significativement supérieure à la variance résiduelle, la variable X_i est considérée comme étant une variable réellement explicative.

F^* suit une statistique de Fisher à 1 et $n - 2$ degrés de liberté. Si $F^* > F_{1;n-2}^\alpha$ nous rejetons au seuil α l'hypothèse H_0 d'égalité des variances, la variable X_i est significative ; dans le cas contraire, nous acceptons l'hypothèse d'égalité des variances, la variable X_i n'est pas explicative de la variable Y_i .

Preuve :

Preuve : $\frac{\left(\hat{\beta}_1 - \beta_1\right)^2}{\sigma^2} \sim \chi_1^2$ (carré d'une variable aléatoire normale centrée réduite)

$$\frac{\sum_{i=1}^n e_i^2}{\sigma^2} = \frac{(n-2) \hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2 \text{ (Somme au carré de } n-2 \text{ variables aléatoires indépendantes normales centrées réduites)}$$

En faisant le rapport des deux chi-deux on obtient :

5. Nous comparons la somme des carrés expliqués *SCE* à la somme des carrés des résidus *SCR* qui est représentative de la somme des carrés théoriquement la plus faible.

2 Modèle de Régression Linéaire Simple

$$F^* = \frac{\left(\hat{\beta}_1 - \beta_1\right)^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n e_i^2}$$

soit sous l'hypothèse $H_0: \beta_1 = 0$, $F^* = \frac{\hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n e_i^2} = \frac{\frac{SCE}{SCR}}{\frac{n-2}{n-2}}$ suit

une loi de Fisher à 1 et $n-2$ degrés de libertés (rapport de deux chi-deux divisés par leurs degrés de liberté).

En effet, on a $SCE = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$

car $\hat{Y}_i - \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = \hat{\beta}_1 (X_i - \bar{X})$

Remarque :

$$\begin{aligned} F^* &= \left(t_{\hat{\beta}_1}^*\right)^2 = \left(\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}\right)^2 = \frac{\frac{\hat{\beta}_1^2}{\hat{\sigma}^2}}{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-2}} \\ &= \frac{\frac{\hat{\beta}_1^2}{\sum_{i=1}^n e_i^2} \sum_{i=1}^n (X_i - \bar{X})^2}{\frac{\sum_{i=1}^n e_i^2}{n-2}} = \frac{\frac{R^2}{1}}{\frac{1-R^2}{n-2}} \end{aligned}$$

2.4.5 Exemple Numérique : Impact de l'Education sur les Salaires

Nous illustrons la théorie économétrique développée jusqu'à présent en prenant en compte les données du tableau (2.2), qui établit une relation entre le salaire horaire moyen (Y) et le nombre d'années d'étude (X). La théorie fondamentale de l'économie du travail nous dit que, parmi de nombreuses variables, l'éducation est un déterminant important des salaires.

Les tableaux (2.2) et (2.3) présentent les données brutes nécessaires pour estimer l'impact quantitatif de l'éducation sur les salaires.

<i>Obs</i>	Y_i	X_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$	X_i^2
1	4.4567	6	-6	-4.218	36	25.308	36
2	5.77	7	-5	-2.9047	25	14.5235	49
3	5.9787	8	-4	-2.696	16	10.784	64
4	7.3317	9	-3	-1.343	9	4.029	81
5	7.3182	10	-2	-1.3565	4	2.713	100
6	6.5844	11	-1	-2.0903	1	2.0903	121
7	7.8182	12	0	-0.8565	0	0	144
8	7.8351	13	1	-0.8396	1	-0.8396	169
9	11.0223	14	2	2.3476	4	4.6952	196
10	10.6738	15	3	1.9991	9	5.9973	225
11	10.8361	16	4	2.1614	16	8.6456	256
12	13.615	17	5	4.9403	25	24.7015	289
13	13.531	18	6	4.8563	36	29.1378	324
\sum	112.7712	156	0	0	182	131.785	2054
<i>moy</i>	8.6747	12					

TABLE 2.2 – Exemple Numérique : Tableau des Données

Estimation des Paramètres de Régression :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{131.7856}{182.0} = 0.7240967$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 8.674708 - 0.7240967 \times 12 = -0.01445$$

Estimation des Variances des Paramètres :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - 2} = \frac{9.83017}{11} = 0.893652 \quad s = 0.945332$$

2 Modèle de Régression Linéaire Simple

Obs	Y_i^2	\hat{Y}_i	$e_i = Y_i - \hat{Y}_i$	e_i^2
1	19.86217	4.165294	0.291406	0.084917
2	33.2929	4.916863	0.853137	0.727843
3	35.74485	5.668432	0.310268	0.096266
4	53.75382	6.420001	0.911699	0.831195
5	53.55605	7.17157	0.14663	0.0215
6	43.35432	7.923139	-1.33874	1.792222
7	61.12425	8.674708	-0.85651	0.733606
8	61.38879	9.426277	-1.59118	2.531844
9	121.4911	10.17785	0.844454	0.713103
10	113.93	10.92941	-0.25562	0.065339
11	117.4211	11.68098	-0.84488	0.713829
12	185.3682	12.43255	1.182447	1.398181
13	183.088	13.18412	0.346878	0.120324
\sum	1083.376	112.7712	0	9.83017

TABLE 2.3 – **Exemple Numérique : Tableau des Données (Suite)**

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{0.893652}{182.0} = 0.004910$$

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{0.004910} = 0.070072$$

$$\hat{\sigma}_{\hat{\beta}_0}^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] = 0.893652 \left[\frac{1}{13} + \frac{(12)^2}{182.0} \right] = 0.775808$$

$$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{0.775808} = 0.88080$$

Coefficient de Détermination :

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{9.83017}{105.1188} = 0.906485$$

À partir des données de l'exemple numérique, nous obtenons la droite

2 Modèle de Régression Linéaire Simple

de régression estimée comme suit :

$$\hat{Y}_i = -0.01445 + 0.7240X_i$$

La valeur de $\hat{\beta}_0 = -0.01445$, qui est l'interception de la ligne, et qui indique le niveau de salaire moyen lorsque le niveau d'éducation est nul. Une telle interprétation littérale de l'interception dans cet exemple n'a aucun sens. Comment pourrait-il y avoir des salaires négatifs ? Comme nous le verrons tout au long de cet ouvrage, très souvent le terme d'interception n'a aucune signification pratique viable. En outre, le niveau d'éducation zéro n'est pas dans le niveau d'éducation observé dans notre échantillon.

La valeur de R^2 est 0.906485, celà suggère que l'éducation explique environ 90.64% de la variation du salaire horaire. Étant donné que R^2 peut être au plus égal à 1, notre droite de régression correspond très bien aux données. Le coefficient de corrélation, $r_{X,Y} = 0.9521$, montre que les salaires et l'éducation sont très fortement corrélés.

Notez que notre modèle est extrêmement simple. La théorie de l'économie du travail nous dit que, à côté de l'éducation, des variables telles que le sexe, la race, le lieu de travail, les syndicats et la langue sont également des facteurs importants dans la détermination du salaire horaire.

2.4.6 Prévision dans le Modèle de Régression Simple

Losque les coefficients du modèle de régression sont estimés, il est serait aisément de calculer une prévision futur à un horizon h .

Soit le modèle de régression estimé sur une période $i = 1, \dots, n$

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \varepsilon_i \quad i = 1, \dots, n$$

Si la valeur de la variable explicative X_i est connue en $n+1$ soit X_{n+1} , la prévision est donnée par : $Y_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1}$, on parle d'une *prévision ponctuelle sans biais*.

Preuve :

L'erreur de prévision est égale à $e_{n+1} = y_{n+1} - \hat{y}_{n+1}$
que l'on peut exprimer comme :

$$e_{n+1} = (\beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1}) - (\hat{\beta}_0 + \hat{\beta}_1 X_{n+1})$$

ou encore

$$e_{n+1} = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) X_{n+1} + \varepsilon_{n+1} \quad (2.20)$$

2 Modèle de Régression Linéaire Simple

En se référant au hypothèses du modèle, on a : $E(e_{n+1}) = 0$

Une démonstration analogue permet d'obtenir $E(e_{n+h}) = 0$

La prévision ponctuelle sans biais est donc obtenue par l'application directe du modèle de régression estimé. Cependant, dans la pratique, il n'est que de peu d'utilité de connaître la prévision si nous ne savons pas quel degré de confiance nous pouvons lui accorder. Nous allons donc calculer la variance de l'erreur de prévision qui nous permet de déterminer un intervalle de confiance bornant la prévision (on parle d'un *intervalle de prédiction*).

A partir de (2.20), la variance de l'erreur de prévision est donnée par :

$$var(e_{n+1}) = var\left(\left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right) X_{n+1} + \varepsilon_{n+1}\right)$$

Alors que la variable X_{n+1} est certaine et l'erreur ε_{n+1} est non autocorrélée avec les ε_i , la variance peut s'écrire comme :

$$var(e_{n+1}) = var(\hat{\beta}_0) + X_{n+1}^2 var(\hat{\beta}_1) + 2X_{n+1} cov(\hat{\beta}_0, \hat{\beta}_1) + var(\varepsilon_{n+1})$$

En remplaçant les variances et la covariance des coefficients par leurs expressions et connaissant $var(\varepsilon_{n+1}) = \sigma^2$, on obtient :

$$var(e_{n+1}) = \frac{\sigma^2}{n} + \bar{X}^2 var(\hat{\beta}_1) + X_{n+1}^2 var(\hat{\beta}_1) - 2X_{n+1}\bar{X} var(\hat{\beta}_1) + \sigma^2$$

ou encore :

$$var(e_{n+1}) = var(Y_{n+1} - \hat{Y}_{n+1}) = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] \quad (2.21)$$

Nous constate que, dans cette formule, la variance de l'erreur de prévision est fonction de l'écart quadratique entre la variable exogène prévue et la moyenne de cette même variable : plus la valeur prévue s'éloigne de cette moyenne, plus le risque d'erreur est important. De même, nous remarquons que la variance de l'erreur de prévision est une fonction inverse de la variabilité de la série explicative.

L'hypothèse de normalité de ε_i permet alors de déterminer un intervalle à $(1 - \alpha)\%$ pour la prévision :

$$e_{n+1} = Y_{n+1} - \hat{Y}_{n+1} \sim N\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]\right)$$

2 Modèle de Régression Linéaire Simple

Soit $\frac{\hat{\beta}_0 + \hat{\beta}_1 X_{n+1} - Y_{n+1}}{\hat{\sigma} \sqrt{\left[1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}} \sim t_{n-2}$

$$Y_{n+1} = \hat{Y}_{n+1} \pm t_{n-2}^{\alpha/2} \hat{\sigma} \sqrt{\left[1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$$

Remarque :

Lorsque nous utilisons le modèle de régression simple pour calculer une droite de tendance (moindres carrés sur le temps), le modèle est spécifié ainsi :

$$T_i = \hat{\beta}_0 + \hat{\beta}_1 t + e_i \quad i = 1, \dots, n$$

Pour calculer la prévision à l'horizon h , nous employons la formule d'extrapolation :

$$\hat{T}_{n+h} = \hat{\beta}_0 + \hat{\beta}_1 (n + h)$$

et l'intervalle de prédiction se trouve alors sur deux branches d'hyperbole $(n + h - \bar{t})$ comme illustré dans la figure

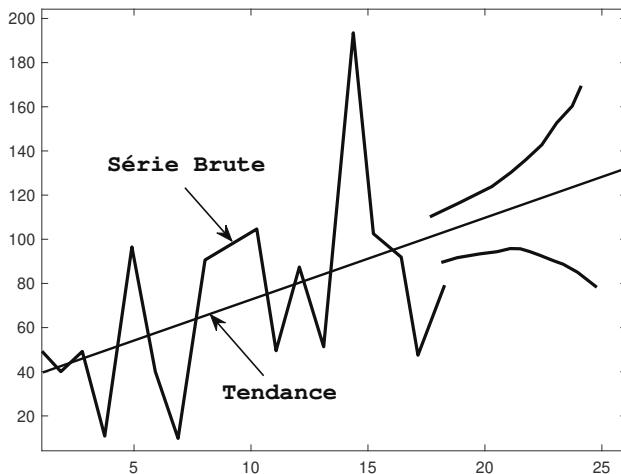


FIGURE 2.10 – Intervalle de la prévision par extrapolation de tendance

2.4.7 Analyse Résiduelle

Un graphique des résidus de la régression est très important. Les résidus sont des estimations cohérentes des erreurs réelles. Mais contrairement aux ε_i 's, ces e_i 's ne sont pas indépendants. En fait, les équations normales des MCO (2.2) et (2.3) nous donnent deux relations entre ces résidus. Par conséquent, connaissant $(n - 2)$ de ces résidus, les deux résidus restants peuvent être déduits. Si nous avions les vrais ε_i 's, et nous les avons tracés, ils devraient ressembler à une dispersion aléatoire autour de l'axe horizontal sans motif spécifique. Un graphique des e_i 's qui montre un motif comme un ensemble de résidus positifs suivi d'un ensemble de résidus négatifs comme indiqué dans la figure (2.11a) et peut être révélateur d'une violation de l'une des 5 hypothèses imposées au modèle, ou simplement en indiquant une mauvaise forme fonctionnelle. Par exemple, si l'hypothèse 3 est violée, de sorte que les ε_i 's soient positivement corrélés, alors il est probable qu'il y ait un résidu positif suivi d'un autre positif, et un résidu négatif suivi d'un autre négatif, comme le montre la figure (2.11b). Alternativement, si nous ajustons une droite de régression linéaire à une vraie relation quadratique entre Y et X , alors une dispersion de résidus comme celle de la figure (2.11c) sera générée.

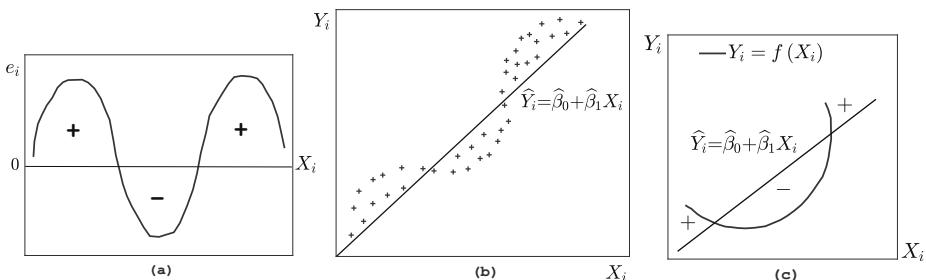


FIGURE 2.11 – Résidus positivement corrélos

Les grands résidus sont des indicateurs de mauvaises prédictions dans l'échantillon. Un résidu important pourrait être une faute de frappe, où le chercheur a mal compris cette observation. Alternativement, il peut s'agir d'une observation influente, ou d'une valeur aberrante qui se comporte différemment des autres observations de l'échantillon et, par conséquent, est plus éloignée de la droite de régression estimée que les autres observations. Le fait que les MCO minimise la somme des carrés de ces résidus signifie qu'un poids important est mis sur cette observation et donc elle est influente. En d'autres termes, la suppression de cette observation de l'échantillon peut modifier de manière significative les

2 Modèle de Régression Linéaire Simple

estimations et la droite de régression. Pour en savoir plus sur l'étude des observations influentes, voir Belsely, Kuh et Welsch (1980).

On peut aussi tracer les résidus par rapport aux X_i 's. Si un modèle comme celui de la figure (2.12) émerge, cela pourrait indiquer une violation de l'hypothèse 2 parce que la variation des résidus augmente avec X_i alors qu'elle devrait être constante pour toutes les observations. Si non, cela pourrait impliquer une relation entre les X_i 's et les erreurs réelles, ce qui constitue une violation de l'hypothèse 4. En résumé, il faut toujours tracer les résidus pour vérifier les données, identifier les observations influentes et vérifier les violations des 5 hypothèses sous-jacentes au modèle de régression.

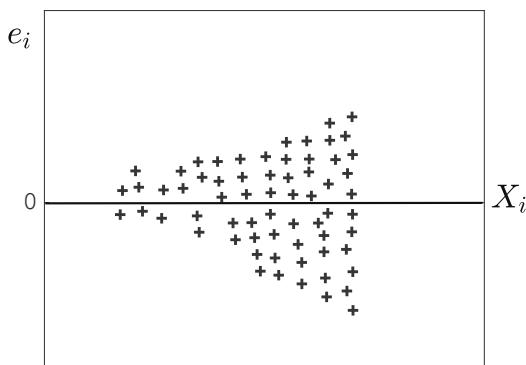


FIGURE 2.12 – Variation résiduelle croissante avec X

2.4.8 La Régression à Travers l'Origine

Il existe des cas où la fonction de régression à une variable explicative prend la forme suivante :

$$Y_i = \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n$$

Dans ce modèle, le terme d'interception β_0 est absent ou nul, d'où le nom de régression par l'origine.

À titre d'illustration, considérons le *Modèle d'Evaluation des Actifs Financiers (MEDAF)* de la théorie moderne du portefeuille, et dont la forme de la prime de risque, peut être exprimée comme :

$$(E(R_i) - r_f) = \beta_i (E(R_m) - r_f)$$

où

2 Modèle de Régression Linéaire Simple

$E(R_i)$ = Taux de rendement espéré d'un actif i ;

$E(R_m)$ = Taux de rendement espéré du portefeuille de marché tel que représenté par exemple, par un indice boursier ;

r_f = Taux de rendement d'un actif sans risque, par exemple le rendement des bons du Trésor à 90 jours.

β_i = Le coefficient bêta, mesure du risque systématique, c'est-à-dire un risque qui ne peut être éliminé par la diversification. En outre, une mesure de l'ampleur avec laquelle le taux de rendement du i -ème actif évolue avec le marché. Un $\beta_i > 1$ implique un actif volatil ou agressive, alors qu'un $\beta_i < 1$ suggère qu'il s'agit un actif défensif.

Remarque : Ne confondez pas ce β_i avec le coefficient de pente de la régression à une variable explicative, β_1 .

Si les marchés des capitaux fonctionnent efficacement, le MEDAF postule que la prime de risque attendue de l'actif i ($= E(R_i) - r_f$) est égale au coefficient β de cet actif multiplié par la prime de risque attendue du marché ($= E(R_m) - r_f$). Si le MEDAF tient, nous avons la situation décrite à la figure (2.13). La ligne illustrée dans la figure est connue sous le nom de ligne du marché des actifs (Security Market Line (SML)). Pour des fins empiriques, l'équation précédente est souvent exprimée comme :

$$R_i - r_f = \beta_i (R_m - r_f) + \varepsilon_i$$

ou encore

$$R_i - r_f = \alpha_i + \beta_i (R_m - r_f) + \varepsilon_i$$

Ce dernier modèle est connu sous le nom de **Modèle de Marché**. Si le MEDAF est valide, α_i devrait être égal à zéro. Voir la figure (2.14)

Notons que dans l'équation précédente, la variable dépendante, Y , est $(R_i - r_f)$ et la variable explicative, X , est β_i , le coefficient de volatilité, et non $(R_m - r_f)$. Par conséquent, pour exécuter la régression, il faut d'abord estimer le paramètre β_i .

Comme le montre cet exemple, la théorie sous-jacente impose parfois que le terme d'interception soit absent du modèle. L'hypothèse de revenu permanent de Milton Friedman, qui énonce que la consommation permanente est proportionnelle au revenu permanent, est un autre exemple d'un modèle approprié. La théorie de l'analyse des coûts, où il est postulé que le coût de production variable est proportionnel à la production ; et certaines versions de la théorie monétariste selon lesquelles le taux de variation des prix (c'est-à-dire le taux d'inflation) est proportionnel au taux de variation de la masse monétaire.

Pour estimer le modèle précédent nous écrivons d'abord la fonction de

2 Modèle de Régression Linéaire Simple

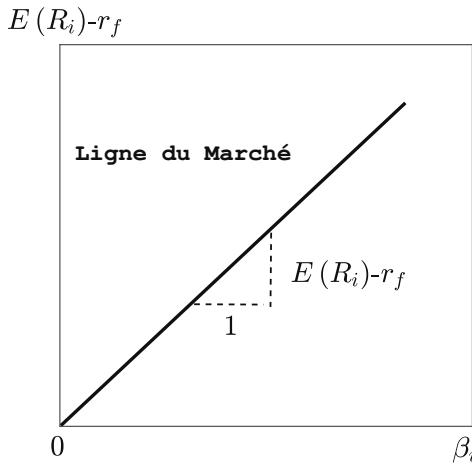


FIGURE 2.13 – Risque Systématique

régression, à savoir :

$$Y_i = \beta_1 X_i + \varepsilon_i$$

Appliquons maintenant la méthode des moindres carrés ordinaires (MCO) à l'équation précédente (la même méthode appliquée dans l'équation (2.3)). Nous obtenons les formules suivantes pour $\hat{\beta}_1$ et sa variance :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n X_i^2}$$

où σ^2 est estimé par :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - 1}$$

2 Modèle de Régression Linéaire Simple

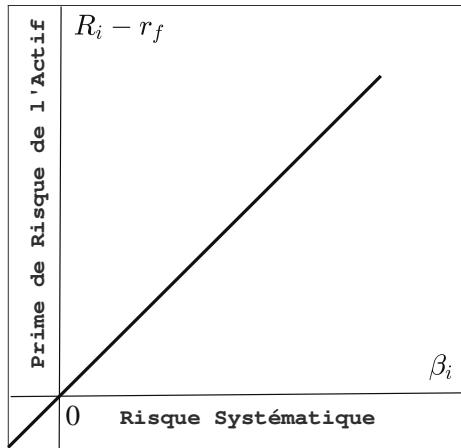


FIGURE 2.14 – **Le modèle de marché de la théorie du portefeuille (en supposant que $\alpha_i = 0$)**

2.4.9 Quelques Considérations sur les Inférences sur les Paramètres β_0 et β_1 de la Régression Simple

Effets des Déviations de Normalité

Si les distributions de probabilité de Y ne sont pas exactement normales mais ne dévient pas sérieusement de la normale, alors les distributions d'échantillonnage de $\hat{\beta}_0$ et $\hat{\beta}_1$ seront approximativement normales et l'utilisation de la distribution de Student fournira approximativement le coefficient de confiance spécifié ou le niveau de signification. Même si les distributions de Y sont loin de la normale, les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ ont généralement la propriété de *normalité asymptotique* : leurs distributions approchent de la normalité dans des conditions très générales lorsque la taille de l'échantillon augmente. Ainsi, avec des échantillons suffisamment grands, les intervalles de confiance et les règles de décision donnés précédemment s'appliquent toujours, même si les distributions de probabilité de Y s'éloignent de la normalité. Pour les échantillons de grande taille, la t de Student est bien sûr remplacée par la z de la distribution normale.

2 Modèle de Régression Linéaire Simple

Interprétation du Seuil Coefficient de Confiance et des Risques des Erreurs

Comme le modèle de régression (2.12) suppose que les X_i sont des constantes connues, le seuil de confiance et les risques d'erreur sont interprétés en ce qui concerne la prise d'échantillons répétés dans lesquels les X 's observations sont maintenues aux mêmes niveaux que dans l'échantillon observé. Par exemple, nous avons construit un intervalle de confiance pour β_1 , avec un seuil de confiance de 0,95. Ce coefficient est interprété comme signifiant que si de nombreux échantillons indépendants sont prélevés où les niveaux de X sont les mêmes que dans l'ensemble de données et qu'un intervalle de confiance de 95% est construit pour chaque échantillon, 95% des intervalles contiendront la vraie valeur de β_1 .

Espacement des Données Observées X

L'inspection des formules (2.6) et (2.7) pour les variances de $\hat{\beta}_0$ et $\hat{\beta}_1$, respectivement, indique que, pour n et σ^2 donnés, ces variances sont affectées par l'espacement des niveaux de X dans les données observées. Par exemple, plus l'écart entre les niveaux X est grand, plus la quantité $\sum_{i=1}^n (X_i - \bar{X})^2$ est grande et plus la variance de $\hat{\beta}_1$ est faible.

Puissance des Tests

Considérons, par exemple, le test concernant β_1 :

$$\begin{aligned} H_0: \beta_1 &= \beta_{10} \\ H_1: \beta_1 &\neq \beta_{10} \end{aligned}$$

pour laquelle la statistique de test utilisée :

$$t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}_{\hat{\beta}_1}}$$

et la règle de décision au niveau de signification α est :

$$\begin{aligned} \text{Si } |t_{\hat{\beta}_1}^*| &\leq t_{n-2}^{\alpha/2} \quad \text{on accepte } H_0 \\ \text{Si } |t_{\hat{\beta}_1}^*| &> t_{n-2}^{\alpha/2} \quad \text{on accepte } H_1 \end{aligned}$$

La puissance de ce test est la probabilité que la règle de décision conduise

2 Modèle de Régression Linéaire Simple

à accepter H_1 . Plus précisément, la puissance est donnée par :

$$Puissance = P \left\{ |t_{\hat{\beta}_1}^*| > t_{n-2}^{\alpha/2} |\delta| \right\}$$

où δ est la *mesure de non-centralité*, c'est-à-dire une mesure de la distance entre la valeur réelle de β_1 et β_{10} :

$$\delta = \frac{|\hat{\beta}_1 - \beta_{10}|}{\hat{\sigma}_{\beta_1}}$$

Former les Hypothèses Nulle et Alternative

Compte tenu des hypothèses nulles et alternatives, leur test de signification statistique ne devrait plus être un mystère. Mais comment formuler ces hypothèses ? Il n'y a pas de règles strictes. Très souvent, le phénomène à l'étude suggérera la nature des hypothèses nulle et alternative. Par exemple, considérons la ligne du marché des capitaux (LMC) de la théorie du portefeuille, qui postule que $E_i = \beta_0 + \beta_1 \sigma_i$, où E = *rendement espéré du portefeuille* et σ = *écart-type du rendement, mesure du risque*. Etant donné que le rendement et le risque devraient être liés positivement (*plus le risque est élevé, plus le rendement est élevé*), l'hypothèse alternative naturelle à l'hypothèse nulle selon laquelle $\beta_1 = 0$ serait $\beta_1 > 0$. Autrement dit, on ne choisirait pas de considérer des valeurs de β_1 inférieures à zéro.

Mais considérons le cas de la demande monétaire. L'un des déterminants importants de la demande de monnaie est le revenu. Des études antérieures sur les fonctions de demande de monnaie ont montré que l'élasticité-revenu de la demande de monnaie (la variation en pourcentage de la demande de monnaie pour une variation de 1% du revenu) était généralement comprise entre 0,7 et 1,3. Par conséquent, dans une nouvelle étude de la demande de monnaie, si l'on postule que le coefficient d'élasticité du revenu β_1 est égal à 1, l'hypothèse alternative pourrait être que $\beta_1 \neq 1$, c'est à dire une hypothèse alternative à deux cotés.

Ainsi, on peut s'appuyer sur des attentes théoriques ou sur des travaux empiriques antérieurs, ou sur les deux, pour formuler des hypothèses. Mais quelle que soit la façon dont les hypothèses sont formées, *il est extrêmement important que le chercheur établisse ces hypothèses avant de mener l'enquête empirique*. Sinon, il sera coupable d'un raisonnement circulaire ou des intuitions auto-réalisatrices. En d'autres termes, si l'on formule des hypothèses après avoir examiné les résultats empiriques, on

2 Modèle de Régression Linéaire Simple

peut tenté de formuler des hypothèses qui justifient ses résultats. Une telle pratique devrait être évitée à tout prix, au moins dans un souci d'objectivité scientifique.

Le Choix du le Seuil de Signification α

Il ressort clairement de la discussion jusqu'à présent que le rejet ou non de l'hypothèse nulle dépend de manière critique de α ; le niveau de signification ou de *la probabilité de commettre une Erreur de Type I*, c'est-à-dire de la probabilité de rejeter l'hypothèse vraie. Dans l'annexe, nous discutons pleinement de la nature d'une Erreur de Type I, de sa relation avec *une Erreur de Type II* (la probabilité d'accepter la fausse hypothèse) et des raisons pour lesquelles les statistiques classiques se concentrent généralement sur une Erreur de Type I. Mais même dans ce cas, pourquoi α est-il couramment fixé aux niveaux de 1, 5 ou au plus 10% ? En réalité, ces valeurs n'ont rien de sacro-saint ; toute autre valeur fera aussi bien l'affaire.

Dans un ouvrage d'introduction comme celui-ci, il n'est pas possible de discuter en profondeur des raisons pour lesquelles on choisit les niveaux de signification de 1, 5 ou 10%, car cela nous mènera dans le domaine de la prise de décision statistique, une discipline en soi. Un bref résumé peut toutefois être proposé. Comme nous le verrons à l'annexe, pour une taille d'échantillon donnée, si nous essayons de réduire une *Erreur de Type I*, une *Erreur de Type II* augmente, et inversement. C'est-à-dire que, compte tenu de la taille de l'échantillon, si nous essayons de réduire la probabilité de rejeter la vraie hypothèse, nous augmentons en même temps la probabilité d'accepter la fausse hypothèse. Il existe donc un compromis entre ces deux types d'erreurs, compte tenu de la taille de l'échantillon. Alors,⁶

Si l'erreur de rejet de l'hypothèse nulle qui est en fait vraie (Erreur de Type I) est coûteuse par rapport à l'erreur de ne pas rejeter l'hypothèse nulle qui est en fait fausse (Erreur de Type II), il sera rationnel de définir la probabilité du premier type d'erreur faible. Si, en revanche, le coût de l'Erreur de Type I est faible par rapport à celui de l'Erreur de Type II, il sera alors plus rentable de rendre la probabilité du premier type d'erreur élevée (rendant ainsi la probabilité du second type d'erreur faible).

6. Jan Kmenta, *Elements of Econometrics*, Macmillan, New York, 1971, pp. 126–127.

2 Modèle de Régression Linéaire Simple

Bien sûr, l'inconvénient, c'est que nous connaissons rarement le coût de ces deux types d'erreurs. Ainsi, les économétriciens pratiquants suivent généralement la pratique consistant à fixer la valeur de α à un niveau égal à 1, 5 ou au plus égal à 10% et choisissent une statistique de test permettant de réduire autant que possible la probabilité de commettre une Erreur de Type II. Etant donné que $(1 - \text{la probabilité de commettre une Erreur de Type II})$ est appelé la **puissance du test**, cette procédure revient à maximiser la puissance du test.

Le Niveau Exact de Signification : la Valeur p

Comme on vient de le noter, le talon d'Achille de l'approche classique du test d'hypothèse est son caractère arbitraire dans la sélection de α . Une fois qu'une statistique de test (par exemple, la statistique empirique t de Student) est obtenue dans un exemple donné, pourquoi ne pas simplement consulter le tableau statistique approprié et rechercher la probabilité réelle d'obtenir une valeur de la statistique de test égale ou supérieure à celle obtenue en l'exemple ? Cette probabilité est appelée la **valeur p** (c'est-à-dire la **valeur de probabilité**), également appelée **niveau de signification observé ou exact**, ou la **probabilité exacte de commettre une Erreur de Type I**. Plus techniquement, la valeur p est définie comme le **niveau de signification le plus bas auquel une hypothèse nulle peut être rejetée**.

Pour illustrer notre propos, revenons à notre exemple de l'impact de l'éducation sur le salaire. Étant donné l'hypothèse nulle que le coefficient réel d'éducation est de 0.5, nous avons obtenu une valeur t de 3.2 dans l'équation. Quelle est la valeur p pour obtenir une valeur t égale ou supérieure à 3.2 ? En regardant la table de Student, nous observons que pour 11 ddl, la probabilité d'obtenir une telle valeur t doit être inférieure à 0.005 (une queue) ou à 0.010 (deux queues).

Si vous utilisez des progiciels statistiques comme EViews par exemple, vous constaterez que la valeur p pour obtenir une valeur t de 3.2 ou plus est d'environ 0.00001, c'est-à-dire extrêmement petite. C'est la valeur p de la statistique t observée. Ce niveau de signification exact de la statistique t est beaucoup plus petit que le niveau de signification conventionnel celui arbitrairement fixé en 1, 5 ou 10%. En fait, si nous utilisions la valeur p qui vient d'être calculée et rejetions l'hypothèse nulle selon laquelle le vrai coefficient d'éducation est égal à 0.5, la probabilité que nous commettions une Erreur de Type I ne serait que d'environ 1 sur 100000.

Comme nous l'avons noté précédemment, si les données ne supportent

2 Modèle de Régression Linéaire Simple

pas l'hypothèse nulle, $|t|$ obtenue sous l'hypothèse nulle sera “grande” et donc la valeur p d'obtention d'un tel $|t|$ sera “petite”. En d'autres termes, pour une taille d'échantillon donnée, comme $|t|$ augmente, la valeur p diminue et on peut donc rejeter l'hypothèse nulle avec une confiance croissante.

Quelle est la relation entre la valeur p et le niveau de signification α ? Si nous faisons en sorte de fixer α comme égal à la valeur p d'une statistique de test (par exemple, la statistique t de Student), il n'y a pas de conflit entre les deux valeurs. En d'autres termes, il vaut mieux renoncer à fixer α de façon arbitraire à un certain niveau et simplement choisir la valeur α de la statistique de test. Il est préférable de laisser au lecteur le choix de rejeter ou non l'hypothèse nulle selon la valeur p donnée. Si, dans une application, la valeur p d'une statistique de test est par exemple, 0.145 ou 14.5%, et si le lecteur souhaite rejeter l'hypothèse nulle à ce niveau (exactement) de signification, qu'il en soit ainsi. De même, comme dans notre exemple sur l'éducation et les salaires, il n'y a rien de mal à ce que le chercheur veuille choisir une valeur p d'environ 0.02% et ne pas courir le risque de se tromper plus de 2 fois sur 10000. Après tout, certains enquêteurs sont peut-être des amoureux des risques.

La Signification Statistique versus La Signification Pratique

Lorsqu'on suppose une telle hypothèse nulle, disons, $H_0: \beta_1 = 1$, l'intention probable dans ce cas est que β_1 est *proche* de 1, si proche qu'il peut être traité à toutes fins pratiques *comme s'il était* 1. Mais si 1.1 « est pratiquement égal » à 1.0, c'est une question d'économie, pas de statistique. On ne peut résoudre le problème en s'appuyant sur un test d'hypothèse, car la statistique de test $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}_{\hat{\beta}_1}}$ mesure le coefficient estimé en unités d'erreur, qui ne sont pas des unités significatives dans lesquelles mesurer le paramètre économique $(\hat{\beta}_1 - 1)$. Il peut être judicieux de réserver le terme «signification» au concept statistique, en adoptant le terme «substantiel» pour le concept économique.

Le but de toute cette discussion est qu'il ne faut pas confondre la signification statistique avec la signification pratique ou économique.

Ce point important a été soulevé par Goldberger. À mesure que la taille de l'échantillon devient très grande, les questions d'importance statistique deviennent beaucoup moins importantes, mais les questions d'importance économique deviennent critiques. En effet, comme avec de très grands échantillons presque toutes les hypothèses nulles seront rejetées, il peut y avoir des études dans lesquelles la magnitude des estimations

2 Modèle de Régression Linéaire Simple

ponctuelles peut être le seul problème.

2.4.10 Exemple Numérique : Impact du Revenu sur la Consommation

Le tableau (2.4) donne la consommation annuelle de 10 ménages choisis au hasard parmi un groupe de ménages ayant un revenu personnel disponible fixe. Le revenu et la consommation sont évalués à 10 000dhs, de sorte que le premier ménage gagne 50 000dhs et consomme 46 000dhs par année. Il vaut la peine de faire les calculs nécessaires pour obtenir les estimations de la régression des moindres carrés de la consommation sur le revenu dans ce cas simple et de les comparer avec celles obtenues à partir d'un ensemble de régressions. Pour ce faire, nous calculons d'abord $\bar{Y} = 6,5$ et $\bar{X} = 7,5$ et formons deux nouvelles colonnes de données composées de $Y_i - \bar{Y}$ et $X_i - \bar{X}$.

Obs	Consommation Y_i	Revenu X_i	$Y_i - \bar{Y}$	$X_i - \bar{X}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	\hat{Y}_i	e_i
1	4,6	5	-1,9	-2,5	4,75	6,25	4,476190	0,123810
2	3,6	4	-2,9	-3,5	10,15	12,25	3,666667	-0,066667
3	4,6	6	-1,9	-1,5	2,85	2,25	5,285714	-0,685714
4	6,6	8	0,1	0,5	0,05	0,25	6,904762	-0,304762
5	7,6	8	1,1	0,5	0,55	0,25	6,904762	0,695238
6	5,6	7	-0,9	-0,5	0,45	0,25	6,095238	-0,495238
7	5,6	6	-0,9	-1,5	1,35	2,25	5,285714	0,314286
8	8,6	9	2,1	1,5	3,15	2,25	7,714286	0,885714
9	8,6	10	2,1	2,5	5,25	6,25	8,523810	0,076190
10	9,6	12	3,1	4,5	13,95	20,25	10,142857	-0,542857
Somme	65	75	0	0	42,5	52,5	65	0
Moyenne	6,5	7,5					6,5	

TABLE 2.4 – Calculs d'une régression simple

2 Modèle de Régression Linéaire Simple

Pour obtenir $\hat{\beta}_1$, on a $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{42,5}{52,5} = 0,8095$ qui

est la propension marginale à consommer estimée. C'est la consommation supplémentaire provoquée par un dirham supplémentaire de revenu disponible.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 6,5 - (0,8095)(7,5) = 0,4286$$

C'est la consommation estimée à zéro revenu disponible (*consommation incompressible*). Les valeurs ajustées ou prédictes de cette régression sont calculées à partir de $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 0,4286 + 0,8095 X_i$ et sont données dans le tableau (2.4). Notez que la moyenne de \hat{Y}_i est égale à la moyenne de Y_i confirmant l'une des propriétés numériques des moindres carrés. Les résidus sont calculés à partir de $e_i = Y_i - \hat{Y}_i$ et ils satisfont $\sum_{i=1}^n e_i = 0$.

Il est laissé au lecteur de vérifier que $\sum_{i=1}^n e_i X_i = 0$. La somme des carrés des résidus est obtenue en mettant au carré la colonne des résidus et en la sommant. Cela nous donne $\sum_{i=1}^n e_i^2 = 2,495238$. Cela signifie que

$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = 0,311905$. Sa racine carrée est donnée par $\hat{\sigma} = 0,558$. Ceci est connu comme l'erreur standard de la régression. Dans ce cas, la valeur estimée $\hat{\sigma}_{\hat{\beta}_1}^2$ est $\frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0,005941$ et la valeur estimée

$$\hat{\sigma}_{\hat{\beta}_0}^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] = 0,311905 \left[\frac{1}{10} + \frac{(7,5)^2}{52,5} \right] = 0,365374$$

En prenant la racine carrée de ces variances estimées, nous obtenons les erreurs-types estimées de $\hat{\beta}_0$ et $\hat{\beta}_1$ données par $\hat{\sigma}_{\hat{\beta}_0} = 0,60446$ et $\hat{\sigma}_{\hat{\beta}_1} = 0,077078$. Puisque les erreurs sont normales, les estimateurs des MCO sont aussi les estimateurs du maximum de vraisemblance et sont normalement distribués eux-mêmes. Pour l'hypothèse nulle $H_0: \beta_1 = 0$; la statistique t de Student observée est

$$t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0,809524}{0,077078} = 10,50$$

2 Modèle de Régression Linéaire Simple

on a $t_{\hat{\beta}_1}^* > t_{n-2}^{\alpha/2} = t_8^{0,025} = 2,306$; ceci est hautement significatif, puisque $\Pr \left[\left| t_{\hat{\beta}_1}^* \right| > 10,5 \right] < 0,0001$. Cette probabilité peut être obtenue en utilisant plusieurs packages de régression. Elle est également connu comme la *p*-value ou la valeur de probabilité. Elle montre que cette valeur *t* est hautement improbable et nous rejetons $H_0: \beta_1 = 0$. De même, l'hypothèse nulle $H_0: \beta_0 = 0$, donne une statistique *t* observée de $t_{\hat{\beta}_0}^* = \frac{0,428571}{0,604462} = 0,709 < t_8^{0,025} = 2,306$, ce qui n'est pas significatif, puisque sa valeur *p* est $\Pr \left[\left| t_{\hat{\beta}_0}^* \right| > 0,709 \right] < 0,709$. Par conséquent, nous ne rejetons pas l'hypothèse nulle H_0 que $\beta_0 = 0$.

La somme totale des carrés est $\sum_{i=1}^n (Y_i - \bar{Y})^2$, ce qui peut être obtenu en mettant au carré la colonne $Y_i - \bar{Y}$ du tableau (2.4) et en sommant. Cela donne $n = 36,9$. En outre, la somme des carrés expliquée $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ peut être obtenue en soustrayant $\bar{Y} = \hat{Y} = 6,5$ de la colonne \hat{Y}_i , et en mettant au carré cette colonne et on sommant. Cela donne 34,404762. Cela aurait également pu être obtenu à partir de

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 = (0,809524)^2 (52,5) = 34,404762$$

Une vérification finale est que

$$\begin{array}{rcl} \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 & = & \sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n e_i^2 \\ SCE & = & SCT - SCR \\ 34,404762 & = & 36,9 - 2,495238 \end{array}$$

Rappelons que

$$R^2 = r_{x,y}^2 = \frac{\left[\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) \right]^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2 \right) \left(\sum_{i=1}^n (Y_i - \bar{Y})^2 \right)} = \frac{(42,5)^2}{(52,5)(36,9)} = 0,9324$$

Cela pourrait aussi avoir été obtenu comme

$$R^2 = 1 - \left(\frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \right) = 0,9324$$

2 Modèle de Régression Linéaire Simple

ou comme

$$R^2 = r_{y,\hat{y}}^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{34,404762}{36,9} = 0,9324$$

Cela signifie que le revenu personnel disponible explique 93,24% de la variation de la consommation. Un graphique des valeurs réelles, prédictes et résiduelles en fonction du temps est donné à la figure (2.15). Cela a été fait en utilisant EViews.

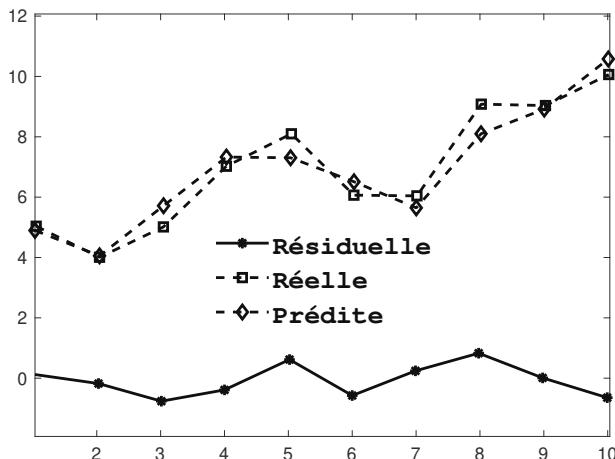


FIGURE 2.15 – Tracés des résidus

2.5 Annexe

R^2 centré et non centré

De la régression des MCO en (2.1), nous obtenons

$$Y_i = \hat{Y}_i + e_i \quad i = 1, 2, \dots, n \quad (2.22)$$

où $\hat{Y}_i = \hat{\beta}_0 + X_i \hat{\beta}_1$. En mettant le tout au carré et en sommant l'équation (2.22) pour toutes les observations, on obtient

$$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n \hat{Y}_i^2 + \sum_{i=1}^n e_i^2 \quad (2.23)$$

2 Modèle de Régression Linéaire Simple

tant que $\sum_{i=1}^n \hat{Y}_i e_i = 0$. Le R^2 non centré est donné par

$$R^2 \text{ non centré} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n Y_i^2} = \frac{\sum_{i=1}^n \hat{Y}_i^2}{\sum_{i=1}^n Y_i^2} \quad (2.24)$$

Notez que la somme totale des carrés pour Y_i n'est pas exprimée en écart par rapport à la moyenne \bar{Y} de l'échantillon. En d'autres termes, le R^2 non centré est la proportion de variation de $\sum_{i=1}^n Y_i^2$ qui est expliquée par la régression sur X . Les packages de régression rapportent habituellement

le R^2 centré qui a été défini comme $1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$. Cette dernière mesure vise à expliquer la variation de Y_i après ajustement de la constante.

À partir de la sous-section 2.4.4, nous avons vu qu'un modèle simple avec seulement une constante donne \bar{Y} comme l'estimation de la constante. La variation en Y_i qui n'est pas expliquée par ce modèle simple est $\sum_{i=1}^n (Y_i - \bar{Y})^2$. En soustrayant $n\bar{Y}^2$ des deux côtés de (2.23), nous obtenons

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n \hat{Y}_i^2 - n\bar{Y}^2 + \sum_{i=1}^n e_i^2$$

et le R^2 centré est

$$R^2 \text{ centré} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n \hat{Y}_i^2 - n\bar{Y}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (2.25)$$

S'il y a une constante dans le modèle $\bar{Y} = \hat{Y}$, voir la sous-section 2.4.4, et $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n \hat{Y}_i^2 - n\bar{Y}^2$. Par conséquent, le R^2 centré = $\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$ qui est le R^2 rapporté par les packages de régression.

S'il n'y a pas de constante dans le modèle, certains packages de régres-

2 Modèle de Régression Linéaire Simple

sion vous donnent l'option de (pas de constante) et le R^2 rapporté est généralement le R^2 non centré.

Erreur de Type I et Erreur de Type II

La méthode façon de procéder est de donner un exemple.

Exemple

On suppose que la distribution de la taille des hommes dans une population est normalement distribuée avec une moyenne $= \mu$ en pouces et $\sigma = 2.5$ pouces. Un échantillon de 100 hommes choisis au hasard dans cette population avait une taille moyenne de 67 pouces. Établissez un intervalle de confiance de 95% pour la taille moyenne ($= \mu$) de la population dans son ensemble.

Comme indiqué, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, qui dans ce cas devient $\bar{X} \sim N\left(\mu, \frac{2.5^2}{100}\right)$.

A partir du tableau A1 dans les annexes, on peut voir que

$$\bar{X} - 1.96 \left(\frac{\sigma}{\sqrt{n}} \right) \leq \mu \leq \bar{X} + 1.96 \left(\frac{\sigma}{\sqrt{n}} \right)$$

couvre 95% de l'aire sous la courbe normale. Par conséquent, l'intervalle précédent fournit un intervalle de confiance de 95% pour μ . En connectant les valeurs données de \bar{X} , σ et n , nous obtenons l'intervalle de confiance à 95% comme suit :

$$66.51 \leq \mu \leq 67.49$$

Dans des mesures répétées, les intervalles ainsi établis incluront le μ réel avec une confiance de 95%. Un point technique peut être noté ici. Bien que nous puissions dire que la probabilité que l'intervalle aléatoire $[\bar{X} \pm 1.96 (\sigma/\sqrt{n})]$ inclue μ soit de 95%, nous ne pouvons pas dire que la probabilité que l'intervalle particulier (66.51; 67.49) inclue μ . Une fois que cet intervalle est fixé, la probabilité qu'il inclue μ est égal à 0 ou 1. Ce que nous pouvons dire, c'est que si nous construisons 100 intervalles de ce type, 95 sur 100 intervalles incluront le vrai μ ; nous ne pouvons pas garantir qu'un intervalle particulier inclura nécessairement μ .

Dans le langage utilisé pour le test d'hypothèse, l'intervalle de confiance que nous avons défini est appelé région d'acceptation et la (ou les) zones situées en dehors de la région d'acceptation sont appelées région(s) critique(s) ou région(s) de rejet de l'hypothèse nulle. Les limites inférieure et supérieure de la région d'acceptation (qui la démarque des régions de rejet) sont appelées valeurs critiques. Dans ce langage de test d'hypo-

2 Modèle de Régression Linéaire Simple

thèses, si la valeur supposée se situe dans la zone d'acceptation, on ne peut pas rejeter l'hypothèse nulle; sinon on peut la rejeter.
Dans l'exemple on avait comme données

$$X_i \sim N(\mu, \sigma^2) = N(\mu, 2.5^2)$$

$$\bar{X} = 67 \quad n = 100$$

On suppose

$$H_0: \mu = \mu^* = 69$$

$$H_1: \mu \neq 0$$

Il est important de noter qu'en décidant de rejeter ou de ne pas rejeter H_0 , nous sommes susceptibles de commettre deux types d'erreurs : (1) nous pouvons rejeter H_0 lorsqu'elle est en fait vraie; cela s'appelle une *Erreur de Type I* (ainsi, dans l'exemple précédent, $\bar{X} = 67$ pourrait provenir de la population avec une valeur moyenne de 69), ou (2) nous ne pouvons pas rejeter H_0 lorsqu'elle est en fait fausse; c'est ce qu'on appelle une *Erreur de Type II*. Par conséquent, un test d'hypothèse ne permet pas d'établir la valeur de μ réel. Elle fournit simplement un moyen de décider si nous pouvons agir comme si $\mu = \mu^*$.

Erreur de Type I et Erreur de Type II

Schématiquement, nous avons

État de Nature		
Décision	H_0 est Vraie	H_0 est Fausse
Rejeter	Erreur de Type I	Pas d'Erreur
Ne pas Rejeter	Pas d'Erreur	Erreur de Type II

Idéalement, nous aimerais minimiser les erreurs de Type I et de Type II. Mais malheureusement, quelle que soit la taille de l'échantillon, il n'est pas possible de minimiser les deux erreurs simultanément. L'approche classique de ce problème, concrétisée par les travaux de Neyman et Pearson, est de supposer qu'une Erreur de Type I est probablement plus grave en pratique qu'une Erreur de Type II. Par conséquent, il faut essayer de maintenir la probabilité de commettre une Erreur de Type I à un niveau assez bas, tel que 0.01 ou 0.05, puis de minimiser autant que possible la probabilité d'avoir une Erreur de Type II.

Dans la littérature, la probabilité d'une Erreur de Type I est désignée par α et est appelée « niveau de signification » et la probabilité d'une Erreur de Type II est désignée par β . La probabilité de ne pas commettre

2 Modèle de Régression Linéaire Simple

d'Erreur de Type II s'appelle la puissance du test. En d'autres termes, la puissance d'un test comme on a discuté dans la sous-section 2.4.8 réside dans sa capacité à rejeter une hypothèse nulle erronée. L'approche classique du test d'hypothèse consiste à fixer α à des niveaux tels que 0.01 (ou 1%) ou 0.05 (5%), puis essayer de maximiser la puissance du test ; c'est-à-dire de minimiser θ .

2.6 Exercices

Exercice 1

Un producteur s'intéresse à la liaison pouvant exister entre le rendement d'un produit (X_i) la quantité de matière première utilisée (Y_i). Il relève 9 couples de données consignés dans le tableau ci-dessous :

Y_i	4	8	10	15	19	30	39	45	50
X_i	2	4	6	9	14	24	30	36	45

- (1) Estimer les deux paramètres de la droite de régression simple β_0 et β_1 et écrire la fonction de la droite.
- (2) Tester la significativité de la pente.
- (3) Construire l'intervalle de confiance au niveau de confiance de 95% pour le paramètre β_1 .
- (4) Calculer le coefficient de détermination R^2 et effectuer le test de Fisher permettant de déterminer si la régression est significative dans son ensemble.
- (5) Le producteur prévoit respectivement 57 et 69 pour le rendement de produit. Déterminer les valeurs prévues de la quantité de matière première pour ces deux prévisions, ainsi que l'intervalle de prévision au niveau de confiance de 95%.

Solution

2 Modèle de Régression Linéaire Simple

Y_i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	\hat{Y}_i	$(\hat{Y}_i - \bar{Y})^2$
4	2	-16,89	285,23	-20,44	345,28	417,98	5,91	343,39
8	4	-14,89	221,68	-16,44	244,84	270,42	8,11	266,88
10	6	-12,89	166,12	-14,44	186,17	208,64	10,30	199,99
15	9	-9,89	97,79	-9,44	93,40	89,20	13,59	117,73
19	14	-4,89	23,90	-5,44	26,62	29,64	19,08	28,77
30	24	5,11	26,12	5,56	28,40	30,86	30,05	31,45
39	30	11,11	123,46	14,56	161,73	211,86	36,64	148,63
45	36	17,11	292,79	20,56	351,73	422,53	43,22	352,49
50	45	26,11	681,79	25,56	667,28	653,09	53,09	820,80
\sum	220	170	1918,89		2105,44	2334,22		2310,14
Moy	24,44	18,89						

(1) Estimation des paramètres de régression linéaire simple β_0 et β_1 :

$$\text{On a } \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{2105,44}{1918,89} = 1,0972$$

$$\text{et } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 24,44 - (1,0972 \times 18,89) = 3,7191$$

l'équation de la droite est donc :

$$Y_i = 3,7191 + 1,0972 X_i + e_i$$

(2) Il s'agit d'un Test Bilatéral de la pente $\hat{\beta}_1$:

$$H_0; \beta_1 = 0$$

$$H_1; \beta_1 \neq 0$$

Calcul de ratio empirique de Student $t_{\hat{\beta}_1}^*$ (dans ce cas $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$)

$$\text{on a } \hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \text{ avec } \hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{SCR}{n-2}$$

2 Modèle de Régression Linéaire Simple

d'après le tableau on calcul SCR :

$$SCR = SCT - SCE = 2334,22 - 2310,14 = 24,08$$

alors $\hat{\sigma}^2 = \frac{SCR}{n-2} = \frac{24,08}{7} = 3,44$

ainsi $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{1,0972}{\sqrt{\frac{3,44}{1918,89}}} = 25,91$

Enfin $|t_{\hat{\beta}_1}^*| > t_{n-2}^{\alpha/2} = t_7^{0,025} = 2,365$ (voir la valeur de $t_7^{0,025}$ dans la table de la loi de Student)

la pente β_1 est significativement différente de zéro.

- (3) Intervalle de confiance au niveau de confiance de 95% pour le paramètre β_1**

on a

$$\begin{aligned} IC_{\beta_1} &= \left[\hat{\beta}_1 \pm t_{n-2}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_1} \right] = \left[1,0972 \pm \left(2,365 \times \sqrt{\frac{3,44}{1918,89}} \right) \right] \\ &= [0,9970 ; 1,1973] \end{aligned}$$

- (4) Coefficient de détermination R^2**

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{2310,14}{2334,22} = 0,9896. \text{ Il s'agit d'un}$$

très bon ajustement linéaire (98,96% de la variabilité de Y autour de sa moyenne est expliquée par la variabilité de X)

Test de Fisher

Calcul de la statistique empirique de Fisher

$$F^* = \frac{\frac{SCE}{ddl}}{\frac{SCR}{ddl_{SCR}}} = \frac{\frac{2310,14}{1}}{\frac{24,08}{7}} = 671,55$$

$F^* > F_{1;n-2}^\alpha = F_{1;7}^{0,05} = 5,59$ (voir la valeur de $F_{1;7}^{0,05}$ dans la table de la loi de Fisher)

Le modèle est globalement significatif

- (5) Prévision**

2 Modèle de Régression Linéaire Simple

Prévisions ponctuelles respectivement des observations $X_{10} = 57$ et $X_{11} = 65$

$$\hat{Y}_{10} = \hat{\beta}_0 + \hat{\beta}_1 X_{10} = 3,7191 + (1,0972 \times 57) = 66,259$$

$$\hat{Y}_{11} = \hat{\beta}_0 + \hat{\beta}_1 X_{11} = 3,7191 + (1,0972 \times 69) = 75,037$$

Intervalles de prédiction

$$Y_{10} = \hat{Y}_{10} \pm t_{n-2}^{\alpha/2} \hat{\sigma} \sqrt{\left[1 + \frac{1}{n} + \frac{(X_{10} - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$$

$$Y_{10} = 66,259 \pm 2,365 \sqrt{3,44} \sqrt{\left[1 + \frac{1}{9} + \frac{(57 - 18,89)^2}{1918,89} \right]}$$

$$Y_{10} = [60,263; 72,254]$$

Après application de la même formule pour l'observation $X_{11} = 65$, on trouve

$$Y_{11} = [68,502; 81,571]$$

Exercice 2

Un statisticien s'intéresse à la liaison pouvant exister entre la croissance économique d'un pays (Y_i) et est le niveau d'inflation (X_i). Il relève 8 couples de données consignés dans le tableau ci-dessous :

Y_i	38	35	30	25	20	17	15	13
X_i	7	10	13	15	19	21	23	25

- (1) Déterminer l'équation de la droite de régression de Y sur X .
- (2) Le coefficient de la variable X est-il significativement inférieur à (-1) ?
- (3) Juger la qualité de cet ajustement et effectuer le test de Fisher permettant de déterminer si la régression est significative dans son ensemble. Commenter.
- (4) Le statisticien prévoit respectivement 26 et 30 pour le niveau d'inflation. Déterminer les valeurs prévues pour la variable croissance, ainsi que les intervalles de prévision au niveau de confiance de 95%.

2 Modèle de Régression Linéaire Simple

Solution

Y_i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	\hat{Y}_i	$(\hat{Y}_i - \bar{Y})^2$
38	7	-9,63	92,64	13,88	-133,55	192,52	38,19	197,78
35	10	-6,63	43,89	10,88	-72,05	118,27	33,81	93,70
30	13	-3,63	13,14	5,88	-21,30	34,52	29,42	28,05
25	15	-1,63	2,64	0,88	-1,42	0,77	26,50	5,64
20	19	2,38	5,64	-4,13	-9,80	17,02	20,65	12,04
17	21	4,38	19,14	-7,13	-31,17	50,77	17,73	40,86
15	23	6,38	40,64	-9,13	-58,17	83,27	14,81	86,76
13	25	8,38	70,14	-11,13	-93,17	123,77	11,89	149,74
\sum	193	133	287,88		-420,63	620,88		614,59
<i>Moy</i>	24,13	16,63						

(1) Estimation des paramètres de régression linéaire simple β_0 et β_1 :

$$\text{On a } \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{-420,63}{287,88} = -1,461$$

$$\text{et } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 24,13 - (-1,461 \times 16,63) = 48,416$$

l'équation de la droite est donc :

$$Y_i = 48,416 - 1,461X_i + e_i$$

(2) Il s'agit d'un **Test Unilatéral** de la pente $\hat{\beta}_1$:

$$H_0; \beta_1 = -1$$

$$H_1; \beta_1 < -1$$

Calcul de ratio empirique de Student $t_{\hat{\beta}_1}^*$ (dans ce cas $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 + 1}{\hat{\sigma}_{\hat{\beta}_1}}$)

2 Modèle de Régression Linéaire Simple

on a $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$ avec $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{SCR}{n-2}$

d'après le tableau on calcul SCR :

$$SCR = SCT - SCE = 620,88 - 614,59 = 6,28$$

alors $\hat{\sigma}^2 = \frac{SCR}{n-2} = \frac{6,28}{6} = 1,046$

ainsi $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 + 1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-0,461}{\sqrt{\frac{1,046}{287,88}}} = -7,647$

Enfin $|t_{\hat{\beta}_1}^*| > t_{n-2}^\alpha = t_6^{0,05} = 1,943$ la pente β_1 est significativement inférieur à (-1).

(3) Coefficient de détermination R^2

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{614,59}{620,88} = 0,9898.$$

Il s'agit d'un

très bon ajustement linéaire (98,98% de la variabilité de Y autour de sa moyenne est expliquée par la variabilité de X)

Test de Fisher

Calcul de la statistique empirique de Fisher

$$F^* = \frac{\frac{SCE}{ddl_{SCE}}}{\frac{SCR}{ddl_{SCR}}} = \frac{\frac{MCE}{MCR}}{\frac{1}{6,28}} = \frac{\frac{614,59}{1}}{6} = 587,18$$

$F^* > F_{1;n-2}^\alpha = F_{1;6}^{0,05} = 5,99$. Le modèle est globalement significatif

(4) Prévision

Prévisions ponctuelles respectivement des observations $X_9 = 26$ et $X_{10} = 30$

$$\hat{Y}_9 = \hat{\beta}_0 + \hat{\beta}_1 X_9 = 48,416 + (-1,461 \times 26) = 10,43$$

$$\hat{Y}_{10} = \hat{\beta}_0 + \hat{\beta}_1 X_{10} = 48,416 + (-1,461 \times 30) = 4,586$$

2 Modèle de Régression Linéaire Simple

Intervalles de prédition

$$Y_9 = \hat{Y}_9 \pm t_{n-2}^{\alpha/2} \hat{\sigma} \sqrt{\left[1 + \frac{1}{n} + \frac{(X_9 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$$

$$Y_9 = 10,43 \pm 2,447 \sqrt{1,046} \sqrt{\left[1 + \frac{1}{8} + \frac{(26 - 16,63)^2}{287,88} \right]}$$

$$Y_9 = [7,437; 13,442]$$

Après application de la même formule pour l'observation $X_{10} = 30$, on trouve

$$Y_{10} = [1,279; 7,892]$$

Exercice 3

Soit un modèle linéaire simple : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n$
on donne les informations suivantes :

$$\sum_{i=1}^n X_i Y_i = 184500 \quad \sum_{i=1}^n Y_i^2 = 26350 \quad \sum_{i=1}^n X_i^2 = 1400000$$

$$\bar{Y} = 60 \quad \bar{X} = 400 \quad n = 7$$

- (1) Estimer les coefficients du modèle
- (2) Evaluer la qualité de cet ajustement
- (3) Tester la significativité globale du modèle
- (4) Calculer l'intervalle de confiance de l'estimateur β_1 .

Solution

$$(1) \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2} = \frac{184500 - 7 \times 400 \times 60}{1400000 - 7 \times (400)^2} = 0,0589$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 60 - (0,0589 \times 400) = 36,44$$

$$(2) \quad R^2 = r_{x,y}^2 = \frac{\left[\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} \right]^2}{\left(\sum_{i=1}^n X_i^2 - n \bar{X}^2 \right) \left(\sum_{i=1}^n Y_i^2 - n \bar{Y}^2 \right)} = 0,8454$$

2 Modèle de Régression Linéaire Simple

Il s'agit d'un très bon ajustement linéaire (84,54% de la variabilité de Y autour de sa moyenne est expliquée par la variabilité de X)

(3) Test de Fisher

Calcul de la statistique empirique de Fisher

$$F^* = \frac{\frac{SCE}{ddl_{SCE}}}{\frac{SCR}{ddl_{SCR}}} = \frac{MCE}{MCR} = \frac{\frac{R^2}{1 - R^2}}{\frac{1}{n - 2}} = \frac{\frac{0,8454}{1 - 0,8454}}{5} = 27,341$$

$F^* > F_{1;n-2}^\alpha = F_{1;5}^{0,05} = 6,61$. Le modèle est globalement significatif

(4) Intervalle de confiance au niveau de confiance de 95% pour le paramètre β_1

on a

$$IC_{\beta_1} = \left[\hat{\beta}_1 \pm t_{n-2}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_1} \right]$$

Calcul de $\hat{\sigma}_{\hat{\beta}_1}$

On connaît à partir des propriétés de la régression linéaire simple

$$\text{que : } F^* = \left(t_{\hat{\beta}_1}^* \right)^2 = \left(\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right)^2$$

$$\text{alors } \hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\left(\hat{\beta}_1 \right)^2}{F^*}} = \sqrt{\frac{(0,0589)^2}{27,341}} = 0,0112644$$

Enfin

$$\begin{aligned} IC_{\beta_1} &= \left[\hat{\beta}_1 \pm t_{n-2}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_1} \right] = [0,0589 \pm (2,571 \times 0,0112644)] \\ &= [0,0299 ; 0,0878] \end{aligned}$$

Exercice 4

On considère les résultats suivants d'une étude économétrique réalisée sur un échantillon de 20 individus.

$$\begin{aligned} Y_i &= 1,651X_i - 26,27 + e_i \\ R^2 &= 0,26 \\ n &= 20 \\ \hat{\sigma} &= 10,50 \end{aligned}$$

(1) A partir les résultats de l'estimation, on demande de retrouver les statistiques suivantes :

(a) la somme des carrés des résidus SCR ,

2 Modèle de Régression Linéaire Simple

- (b) la somme des carrés totaux SCT ,
 (c) la somme des carrés expliqués SCE ,
 (d) la valeur de la statistique empirique de Fisher F^*
 (e) et l'écart type du coefficient $\hat{\beta}_1$, $(\hat{\sigma}_{\hat{\beta}_1})$
- (2) Le coefficient de la variable X est-il significativement supérieur à 1 ?

Solution

- (1)(a) Somme des carrés des résidus SCR

$$\text{on a } \hat{\sigma}^2 = \frac{SCR}{n - 2}$$

$$\text{alors } SCR = \hat{\sigma}^2 (n - 2) = (10,50)^2 \times 18 = 1984,5$$

- (b) Somme des carrés totaux SCT

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

$$SCT = \frac{SCR}{1 - R^2} = \frac{1984,5}{1 - 0,26} = 2681,75$$

- (c) Somme des carrés expliqués SCE

$$SCE = SCT - SCR = 2681,75 - 1984,5 = 697,25$$

- (d) Statistique empirique de Fisher F^*

$$\text{Méthode 1 : } F^* = \frac{\frac{SCE}{ddl_{SCE}}}{\frac{SCR}{ddl_{SCR}}} = \frac{\frac{697,25}{1}}{\frac{1984,5}{18}} = 6,324$$

$$\text{Méthode 2 : } F^* = \frac{\frac{R^2}{1 - R^2}}{\frac{0,26}{1 - 0,26}} = \frac{\frac{0,26}{18}}{\frac{0,26}{18}} = 6,324$$

- (e) Calcul de $\hat{\sigma}_{\hat{\beta}_1}$

On connaît à partir des propriétés de la régression linéaire simple que : $F^* = \left(t_{\hat{\beta}_1}^* \right)^2 = \left(\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right)^2$

$$\text{alors } \hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\left(\hat{\beta}_1 \right)^2}{F^*}} = \sqrt{\frac{(1,651)^2}{6,324}} = 0,65652$$

2 Modèle de Régression Linéaire Simple

(2) Il s'agit d'un **Test Unilatéral** de la pente $\hat{\beta}_1$:

$$\begin{aligned} H_0; \beta_1 &= 1 \\ H_1; \beta_1 &> 1 \end{aligned}$$

Calcul de ratio empirique de Student $t_{\hat{\beta}_1}^*$ (dans ce cas $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}_{\hat{\beta}_1}}$)

$$\text{ainsi } t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0,651}{0,65652} = 0,9915$$

Enfin $|t_{\hat{\beta}_1}^*| < t_{n-2}^\alpha = t_{18}^{0,05} = 1,734$ la pente β_1 est non significativement supérieur à 1.

Exercice 5

Soit un modèle linéaire simple suivant : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n$

Y_i : salaire moyen horaire par jour (en Dh)

X_i : nombre d'années d'études

On donne par ailleurs les informations suivantes :

$$r_{X,Y} = 0,951916 \quad \sigma_X = 3,894440 \quad \sigma_Y = 2,945636$$

Après estimation, sur base d'un échantillon de 13 observations, un étudiant présente les résultats incomplets ci-après :

$$\hat{Y}_i = 0,030769 + \dots \dots \dots X_i$$

- (1) Compléter les pointillés
- (2) Tester la significativité du $r_{X,Y}$
- (3) Calculer le coefficient de détermination R^2 . Interpréter ces résultats. Semblent-ils logiques ?
- (4) Tester la significativité de la pente (β_1) et la significativité d'ensemble du modèle.

Solution

- (1) **Calcul du paramètre $\hat{\beta}_1$:**

2 Modèle de Régression Linéaire Simple

on a $r_{X,Y} = \frac{\hat{\beta}_1 \times \sigma_x}{\sigma_Y}$ alors $\hat{\beta}_1 = \frac{r_{X,Y} \times \sigma_Y}{\sigma_X}$

$$\hat{\beta}_1 = \frac{0,951916 \times 2,945636}{3,894440} = 0,72$$

(2) Test de la significativité du $r_{X,Y}$

$$H_0; r_{X,Y} = 0$$

$$H_1; r_{X,Y} \neq 0$$

$$t_{r_{X,Y}}^* = \frac{r_{X,Y}}{\sqrt{\frac{1 - r_{X,Y}^2}{n - 2}}} = \frac{0,951916}{\sqrt{\frac{1 - (0,951916)^2}{11}}} = 10,30$$

$|t_{r_{X,Y}}^*| > t_{n-2}^{\alpha/2} = t_{11}^{0,025} = 2,201$. Le coefficient de corrélation $r_{X,Y}$ est significativement différent de zéro.

(3) Coefficient de détermination R^2

$$R^2 = r_{X,Y}^2 = (0,951916)^2 = 0,9061$$

Il s'agit d'un bon ajustement linéaire (90,61% de la variabilité de Y autour de sa moyenne est expliquée par la variabilité de X)

Il y a lien fort et positif entre le salaire moyen horaire par jour et le nombre d'années d'études. En effet, ces résultats semblent logiques pour cet échantillon car il est tout à fait normal que ceux qui beaucoup étudié gagnent un peu plus que ceux qui ont étudié un peu moins.

(4) Il s'agit d'un Test Bilatéral de la pente $\hat{\beta}_1$:

$$H_0; \beta_1 = 0$$

$$H_1; \beta_1 \neq 0$$

Calcul de ratio empirique de Student $t_{\hat{\beta}_1}^*$ (dans ce cas $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$)

On connaît à partir des propriétés de la régression linéaire simple que :

$$F^* = \left(t_{\hat{\beta}_1}^* \right)^2 = \left(\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right)^2 = \frac{R^2}{\frac{1 - R^2}{n - 2}} = \frac{r_{X,Y}^2}{\frac{1 - r_{X,Y}^2}{n - 2}}$$

2 Modèle de Régression Linéaire Simple

alors

$$t_{\beta_1}^* = \sqrt{\frac{r_{X,Y}^2}{\frac{1-r_{X,Y}^2}{n-2}}} = \sqrt{\frac{(0,951916)^2}{\frac{1-(0,951916)^2}{11}}} = 10,30$$

Enfin $|t_{\beta_1}^*| > t_{n-2}^{\alpha/2} = t_{11}^{0,025} = 2,201$

la pente β_1 est significativement différente de zéro.

Test de Fisher

Calcul de la statistique empirique de Fisher

$$F^* = \frac{R^2}{\frac{1-R^2}{n-2}} = \frac{0,9061}{\frac{1-0,9061}{11}} = 106,145$$

$F^* > F_{1;n-2}^{\alpha} = F_{1;11}^{0,05} = 4,84$. Le modèle est globalement significatif

Remarque :

On remarque d'après les question (3) et (5) que $t_{\beta_1}^* = t_{r_{X,Y}}^*$ ce qui confirme encore le fait que le test d'hypothèse $H_0; \beta_1 = 0$ est équivalent au test d'hypothèse $H_0; SCE = 0$.

Exercice 6

Un producteur cherche à estimer la relation liant la production d'un article Y_i au taux d'un produit chimique X_i existant dans la matière première en formalisant la relation suivante :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n$$

A partir d'une étude statistique portant sur 85 tonnes de matière première, un économètre lui fournit les résultats suivants :

$$Y_i = 132,80 - 1,1 X_i + e_i \\ (4,3) \quad (10,2)$$

(.) = Ratio empirique de Student

$$\sum_{i=1}^{85} e_i^2 = 6234,32$$

- (1) Montrer que tester l'hypothèse $H_0; \beta_1 = 0$ revient à tester l'hypothèse $H_0; r_{X,Y} = 0$, où $r_{X,Y}$ est le coefficient de corrélation linéaire simple entre Y_i et X_i , le calculer.

2 Modèle de Régression Linéaire Simple

- (2) Construire l'équation d'analyse de la variance et vérifier les résultats obtenus dans la question (1) à partir du test de Fisher.
 (3) Le coefficient β_1 est-il significativement inférieur à (-1) .

Solution

(1) on a

$$r_{x,y}^2 = \frac{\left[\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) \right]^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2 \right) \left(\sum_{i=1}^n (Y_i - \bar{Y})^2 \right)}$$

et $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$ alors

$$r_{x,y}^2 = \frac{\hat{\beta}_1 \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{SCE}{SCT} = R^2$$

car :

$$\begin{aligned} \hat{\beta}_1 \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) &= \hat{\beta}_1 \times \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^n (\hat{\beta}_1 X_i - \hat{\beta}_1 \bar{X})^2 &= \sum_{i=1}^n (\hat{Y}_i - \hat{\beta}_0 - \bar{Y} + \hat{\beta}_0)^2 \\ &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = SCE \end{aligned}$$

et puisqu'on a une égalité entre R^2 et $r_{X,Y}^2$ dans un modèle de régression linéaire simple, alors :

$$F^* = \frac{R^2}{(1 - R^2) / (n - 2)} = \frac{r_{X,Y}^2}{(1 - r_{X,Y}^2) / (n - 2)}$$

2 Modèle de Régression Linéaire Simple

ainsi

$$t_{\hat{\beta}_1}^* = \frac{r_{X,Y} \sqrt{(n-2)}}{\sqrt{(1 - r_{X,Y}^2)}} \sim t_{n-2}$$

Ceci permet de tester si la relation entre Y_i et X_i est significative, ou encore si le coefficient $r_{X,Y}$ est significativement différent de 0.

- (2) On a $t_{\hat{\beta}_1}^* = 10,2$ et $SCR = \sum_{i=1}^{85} e_i^2 = 6234,32$

$$F^* = \frac{\frac{SCE}{ddl_{SCE}}}{\frac{SCR}{ddl_{SCR}}} = \left(t_{\hat{\beta}_1}^* \right)^2 = \frac{\frac{SCE}{1}}{\frac{6234,32}{83}} = (10,2)^2 = 104,04$$

$$\text{alors } SCE = \frac{6234,32 \times (10,2)^2}{83} = 7814,682 \text{ et } SCT = SCE + SCR = 7814,682 + 6234,32 = 14049$$

Test de Fisher

Calcul de la statistique empirique de Fisher

$$F^* = \frac{\frac{SCE}{ddl_{SCE}}}{\frac{SCR}{ddl_{SCR}}} = \frac{\frac{R^2}{1}}{\frac{1 - R^2}{n - 2}} = \frac{\frac{7814,682}{6234,32}}{\frac{83}{83}} = 104,03$$

$F^* > F_{1;n-2}^\alpha = F_{1;83}^{0,05} = 3,92$. Le modèle est globalement significatif

- (3) Il s'agit d'un **Test Unilatéral** de la pente $\hat{\beta}_1$:

$$\begin{aligned} H_0; \beta_1 &= -1 \\ H_1; \beta_1 &< -1 \end{aligned}$$

Calcul de ratio empirique de Student $t_{\hat{\beta}_1}^*$ (dans ce cas $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 + 1}{\hat{\sigma}_{\hat{\beta}_1}}$)

calcul de $\hat{\sigma}_{\hat{\beta}_1}$:

d'après le ratio empirique bilatéral de Student donné dans l'exercice on peut déduire la valeur de $\hat{\sigma}_{\hat{\beta}_1}$, c'est-à-dire

$$t_{\hat{\beta}_1}^* (\text{bilatéral}) = \left| \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right| \Rightarrow \hat{\sigma}_{\hat{\beta}_1} = \left| \frac{\hat{\beta}_1}{t_{\hat{\beta}_1}^* (\text{bilatéral})} \right| = \frac{1,1}{10,2} = 0,107843$$

2 Modèle de Régression Linéaire Simple

$$\text{ainsi } t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 + 1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-0,1}{0,107843} = -0,9272739$$

Enfin $|t_{\hat{\beta}_1}^*| < t_{n-2}^\alpha = t_{83}^{0,05} = 1,664$ la pente β_1 est non significativement inférieur à (-1) .

Exercice 7

Soit les résultats d'une estimation économétrique :

$$\begin{aligned} Y_i &= 1,651X_i - 29,27 + e_i \\ &\quad (3,60) \\ (\cdot) &= \text{Ratio empirique de Student} \\ n &= 20 \\ \hat{\sigma} &= 10,50 \end{aligned}$$

- (1) A partir des informations connues, on demande de retrouver les statistiques suivantes : la somme des carrés des résidus (SCR), la somme des carrés expliqués (SCE) et la somme des carrés totaux (SCT).
- (2) Le coefficient de la variable X est-il significativement supérieur à 1?
- (3) On considère $r_{X,Y}$ le coefficient de corrélation linéaire de X et Y . Tester la significativité de $r_{X,Y}$.

Solution

- (1) Somme des carrés des résidus SCR

$$\text{on a } \hat{\sigma}^2 = \frac{SCR}{n-2}$$

$$\text{alors } SCR = \hat{\sigma}^2(n-2) = (10,50)^2 \times 18 = 1984,5$$

Somme des carrés expliqués SCE

on a

$$F^* = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} = \left(t_{\hat{\beta}_1}^* \right)^2 = \frac{\frac{SCE}{1}}{\frac{1984,5}{18}} = (3,60)^2 = 12,96$$

$$SCE = \frac{1984,5 \times 12,96}{18} = 1428,84$$

2 Modèle de Régression Linéaire Simple

Somme des carrés totaux SCT

$$SCT = SCE + SCR = 1428,84 + 1984,5 = 3413,34$$

(2) Il s'agit d'un **Test Unilatéral** de la pente $\hat{\beta}_1$:

$$H_0; \beta_1 = 1$$

$$H_1; \beta_1 > 1$$

Calcul de ratio empirique de Student $t_{\hat{\beta}_1}^*$ (dans ce cas $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}_{\hat{\beta}_1}}$)

Calcul de $\hat{\sigma}_{\hat{\beta}_1}$

on a

$$t_{\hat{\beta}_1}^*(bilateral) = \left| \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right| \Rightarrow \hat{\sigma}_{\hat{\beta}_1} = \left| \frac{\hat{\beta}_1}{t_{\hat{\beta}_1}^*(bilateral)} \right| = \frac{1,651}{3,60} = 0,4586$$

$$\text{ainsi } t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0,651}{0,4586} = 1,41953$$

Enfin $|t_{\hat{\beta}_1}^*| < t_{n-2}^{\alpha} = t_{18}^{0,05} = 1,734$ la pente β_1 n'est pas significativement supérieur à 1.

(3) Calcul de $r_{X,Y}$ coefficient de corrélation linéaire de X et Y :

on a

$$R^2 = \frac{SCE}{SCT} = \frac{1428,84}{3413,34} = 0,4186 \Rightarrow r_{X,Y} = \sqrt{R^2} = 0,6470$$

Test de Student sur $r_{X,Y}$

On a

$$H_0; r_{X,Y} = 0$$

$$H_1; r_{X,Y} \neq 0$$

ainsi $t_{r_{X,Y}}^* = \frac{|r_{X,Y}|}{\sqrt{\frac{1 - r_{X,Y}^2}{n - 2}}} = 3,6$. Enfin $t_{r_{X,Y}}^* > t_{n-2}^{\alpha/2} = t_{18}^{0,025} =$

2 Modèle de Régression Linéaire Simple

2, 101, le coefficient de corrélation $r_{X,Y}$ est significativement différent de 0.

Exercice 8

On considère les résultats suivants d'une étude économétrique réalisée sur un échantillon de 35 individus.

$$Y_i = 12,27 + 0,8X_i + e_i \quad (2,7)$$

(.) = Ratio empirique de Student

$$R^2 = 0,50$$

$$Cov(X, Y) = 27$$

$$\sum_{i=1}^{35} X_i = 210$$

Construire l'intervalle de confiance au niveau de confiance de 95% pour le paramètre β_0 .

Solution

L'intervalle de confiance pour l'estimateur β_0 , on a

$$IC_{\beta_0} = \left[\hat{\beta}_0 \pm t_{n-2}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_0} \right]$$

Calcul de $\hat{\sigma}_{\hat{\beta}_0}$

$$\hat{\sigma}_{\hat{\beta}_0}^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]$$

$$\hat{\sigma}_{\hat{\beta}_0}^2 = \frac{\hat{\sigma}^2}{n} + \bar{X}^2 \hat{\sigma}_{\hat{\beta}_1}^2$$

Calcul $\hat{\sigma}_{\hat{\beta}_1}^2$

$$t_{\hat{\beta}_1}^*(bilatéral) = \left| \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right| \Rightarrow \hat{\sigma}_{\hat{\beta}_1} = \left| \frac{\hat{\beta}_1}{t_{\hat{\beta}_1}^*(bilatéral)} \right| = \frac{0,8}{2,7} = 0,296$$

$$\hat{\sigma}_{\hat{\beta}_1}^2 = 0,08761$$

2 Modèle de Régression Linéaire Simple

Calcul \bar{X}^2

$$\bar{X}^2 = \left(\frac{\sum_{i=1}^n X_i}{n} \right)^2 = \left(\frac{210}{35} \right)^2 = 36$$

Calcul $\hat{\sigma}^2$

On a

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{Cov(X, Y)}{\sigma_X^2} = \frac{Cov(X, Y)}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n \times Cov(X, Y)}{\hat{\beta}_1} = \frac{35 \times 27}{0,8} = 1181.25$$

on déduit la valeur de $\hat{\sigma}^2$ à partir de $\hat{\sigma}_{\hat{\beta}_1}^2$

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \implies \hat{\sigma}^2 = 1181,25 \times 0,08761 = 103.489$$

ainsi

$$\hat{\sigma}_{\hat{\beta}_0}^2 = \frac{\hat{\sigma}^2}{n} + \bar{X}^2 \hat{\sigma}_{\hat{\beta}_1}^2 = \frac{103.489}{35} + 36 \times 0,08761 = 6,1107$$

Enfin

$$\begin{aligned} IC_{\beta_0} &= [\hat{\beta}_0 \pm t_{n-2}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_0}] = [12,27 \pm 2,021 \times \sqrt{6,1107}] \\ &= [7,2741; 17,2658] \end{aligned}$$

Exercice 9

Le coefficient de corrélation linéaire entre deux variables X et Y est $r_{X,Y} = 0,6$. Si les écarts type de X et Y sont respectivement 1,50 et 2 ; et leurs moyennes, respectivement, 10 et 20. Trouvez les équations de régression de Y en X et de X en Y .

Solution

Equation de régression de Y en X : $(Y_i = \beta_0 + \beta_1 X_i)$

2 Modèle de Régression Linéaire Simple

on a

$$\hat{\beta}_1 = \frac{Cov(X, Y)}{\sigma_X^2} = \frac{r_{X,Y} \times \sigma_Y}{\sigma_X} = \frac{0,6 \times 2}{1,50} = 0,8$$

et $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 20 - 0,8 \times 10 = 12$

Enfin $Y_i = 12 + 0,8X_i$

Equation de régression de X en Y : $(X_i = \beta'_0 + \beta'_1 Y_i)$

on a

$$\hat{\beta}'_1 = \frac{Cov(X, Y)}{\sigma_Y^2} = \frac{r_{X,Y} \times \sigma_X}{\sigma_Y} = \frac{0,6 \times 1,50}{2} = 0,45$$

et $\hat{\beta}'_0 = \bar{X} - \hat{\beta}'_1 \bar{Y} = 10 - 0,45 \times 20 = 1$

Enfin $X_i = 1 + 0,45Y_i$

Exercice 10

On considère les résultats suivants d'une étude économétrique réalisée sur un échantillon de 27 individus.

$$Y_i = 10,5 + 1,3X_i + e_i$$

$$(5,10)$$

(.) = Ratio empirique de Student

$$SCT = 227$$

$$n = 27$$

$$\sum_{i=1}^{27} Y_i = 880,2$$

(1) Le coefficient $\hat{\beta}_1$ de la variable X est-il significativement supérieur à 1 ?

(2) On prévoit les valeurs 16 et 27 respectivement pour les observations X_{28} et X_{29} .

Déterminer les valeurs prévues ainsi que les intervalles de prévision pour ces deux observations.

Solution

(1) Il s'agit d'un **Test Unilatéral** de la pente $\hat{\beta}_1$:

$$H_0; \beta_1 = 1$$

$$H_1; \beta_1 > 1$$

Calcul de ratio empirique de Student $t_{\hat{\beta}_1}^*$ (dans ce cas $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} =$

2 Modèle de Régression Linéaire Simple

$$\frac{\hat{\beta}_1 - 1}{\hat{\sigma}_{\hat{\beta}_1}})$$

Calcul de $\hat{\sigma}_{\hat{\beta}_1}$

on a

$$t_{\hat{\beta}_1}^* (\text{bilatéral}) = \left| \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right| \Rightarrow \hat{\sigma}_{\hat{\beta}_1} = \left| \frac{\hat{\beta}_1}{t_{\hat{\beta}_1}^* (\text{bilatéral})} \right| = \frac{1,3}{5,10} = 0,2549$$

$$\text{ainsi } t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0,3}{0,2549} = 1,1769$$

Enfin $|t_{\hat{\beta}_1}^*| < t_{n-2}^\alpha = t_{25}^{0,05} = 1,708$. La pente β_1 est non significativement supérieur à 1.

(2) Prévision

Prévisions ponctuelles respectivement des observations $X_{28} = 16$ et $X_{29} = 27$

$$\hat{Y}_{28} = \hat{\beta}_0 + \hat{\beta}_1 X_{28} = 10,5 + (1,3 \times 16) = 31,3$$

$$\hat{Y}_{29} = \hat{\beta}_0 + \hat{\beta}_1 X_{29} = 10,5 + (1,3 \times 27) = 45,6$$

Intervalles de prédiction

$$Y_{28} = \hat{Y}_{28} \pm t_{n-2}^{\alpha/2} \hat{\sigma} \sqrt{\left[1 + \frac{1}{n} + \frac{(X_{28} - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$$

$$\text{on a } F^* = \left(t_{\hat{\beta}_1}^* \right)^2 = \frac{R^2}{1 - R^2} = (5,10)^2 = 26,01$$

$$R^2 = \frac{26,01}{26,01 + n - 2} = 0,5099$$

Calcul de $\hat{\sigma}^2$

$$\text{on a } R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \implies SCR = SCT(1 - R^2)$$

$$SCR = 227 \times (1 - 0,5099) = 111,25$$

$$\hat{\sigma}^2 = \frac{SCR}{n - 2} = \frac{111,25}{25} = 4,45$$

2 Modèle de Régression Linéaire Simple

Calcul de $\sum_{i=1}^n (X_i - \bar{X})^2$

On peut la déduire à partir de la formule de $\hat{\sigma}_{\hat{\beta}_1}^2$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{\hat{\sigma}^2}{\hat{\sigma}_{\hat{\beta}_1}^2} = \frac{4,45}{(0,2549)^2} = 68,4889$$

Calcul de \bar{X}

$$\bar{X} = \frac{\bar{Y} - \hat{\beta}_0}{\hat{\beta}_1} \quad \text{et} \quad \bar{Y} = \frac{\sum_{i=1}^{27} Y_i}{27}$$

$$\bar{X} = \frac{(880,2/27) - 10,5}{1,3} = 17$$

Enfin

$$Y_{28} = 31,3 \pm 2,060\sqrt{4,45} \sqrt{\left[1 + \frac{1}{27} + \frac{(16-17)^2}{68,4889}\right]}$$

$$Y_{28} = [26,843; 35,756]$$

Après application de la même formule pour l'observation $X_{29} = 27$, on trouve

$$Y_{29} = [38,732; 52,467]$$

Exercice 11

Un chercheur s'intéresse à la relation liant le salaire et la durée de formation. Pour ce faire, il dispose d'un échantillon de 35 hommes et 27 femmes ayant le même âge, dont il relève la rémunération annuelle (Y_i), exprimée en milliers de dirhams, et le nombre d'années de formation (X_i). L'estimation a conduit aux résultats suivants :

Pour les hommes

$$Y_i = 16,20 + 1,5X_i + e_i \quad (8) \quad (3,8)$$

(.) = Ratio empirique de Student $R^2 = 0,57$ $(.) =$ Ratio empirique de Student $R^2 = 0,40$

Pour les femmes

$$Y_i = 12,50 + 1,62X_i + e_i \quad (12) \quad (3,12)$$

(1) L'estimateur $\hat{\beta}_1$ obtenu pour le cas des hommes est-il significatif ?

2 Modèle de Régression Linéaire Simple

- vement supérieur à 0,5 ?
- (2) L'estimateur $\hat{\beta}'_1$ obtenu pour le cas des femmes est-il significativement supérieur à 0,5 ?
- (3) Existe-t-il une différence significative de l'impact de la durée de formation sur le salaire des hommes et des femmes ?

Solution

(1) Pour le cas des hommes

Il s'agit d'un **Test Unilatéral** de la pente $\hat{\beta}_1$:

$$H_0; \beta_1 = 0,5$$

$$H_1; \beta_1 > 0,5$$

Calcul de ratio empirique de Student $t_{\hat{\beta}_1}^*$ (dans ce cas $t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0,5}{\hat{\sigma}_{\hat{\beta}_1}}$)

Calcul de $\hat{\sigma}_{\hat{\beta}_1}$

on a

$$t_{\hat{\beta}_1}^* (\text{bilatéral}) = \left| \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right| \Rightarrow \hat{\sigma}_{\hat{\beta}_1} = \left| \frac{\hat{\beta}_1}{t_{\hat{\beta}_1}^* (\text{bilatéral})} \right| = \frac{1,5}{3,8} = 0,39473$$

$$\text{ainsi } t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - 0,5}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{1}{0,39473} = 2,533$$

Enfin $|t_{\hat{\beta}_1}^*| > t_{n-2}^\alpha = t_{33}^{0,05} = 1,684$. La pente β_1 est significativement supérieur à 0,5.

(2) Pour le cas des femmes

Il s'agit d'un **Test Unilatéral** de la pente $\hat{\beta}'_1$:

$$H_0; \beta'_1 = 0,5$$

$$H_1; \beta'_1 > 0,5$$

Calcul de ratio empirique de Student $t_{\hat{\beta}'_1}^*$ (dans ce cas $t_{\hat{\beta}'_1}^* = \frac{\hat{\beta}'_1 - \beta'_1}{\hat{\sigma}_{\hat{\beta}'_1}} = \frac{\hat{\beta}'_1 - 0,5}{\hat{\sigma}_{\hat{\beta}'_1}}$)

2 Modèle de Régression Linéaire Simple

Calcul de $\hat{\sigma}_{\hat{\beta}'_1}$

on a

$$t_{\hat{\beta}'_1}^*(\text{bilatéral}) = \left| \frac{\hat{\beta}'_1}{\hat{\sigma}_{\hat{\beta}'_1}} \right| \Rightarrow \hat{\sigma}_{\hat{\beta}'_1} = \left| \frac{\hat{\beta}'_1}{t_{\hat{\beta}'_1}^*(\text{bilatéral})} \right| = \frac{1,62}{3,12} = 0,5192$$

$$\text{ainsi } t_{\hat{\beta}'_1}^* = \frac{\hat{\beta}'_1 - 0,5}{\hat{\sigma}_{\hat{\beta}'_1}} = \frac{1,12}{0,5192} = 2,1571$$

Enfin $|t_{\hat{\beta}'_1}^*| > t_{n-2}^\alpha = t_{25}^{0,05} = 1,708$. La pente β_1 est significativement supérieur à 0,5.

- (3) Le problème dans cette question se ramène à un **test bilatéral** de différence de moyennes de variables aléatoires normales indépendantes et de variances inégales.

Soit d la différence entre les deux estimateurs β_1 et $\hat{\beta}'_1$. Les hypothèses pour ce test se présente comme suit :

$$H_0; \beta_1 = \hat{\beta}'_1 \iff H_0; d = \beta_1 - \hat{\beta}'_1 = 0$$

$$H_1; \beta_1 \neq \hat{\beta}'_1 \iff H_1; d = \beta_1 - \hat{\beta}'_1 \neq 0$$

On a

$$\frac{\hat{d}}{\hat{\sigma}_{\hat{d}}} = \frac{(\hat{\beta}_1 - \hat{\beta}'_1) - (\beta_1 - \hat{\beta}'_1)}{\hat{\sigma}_{\hat{\beta}_1 - \hat{\beta}'_1}} \sim t_{n_1+n_2-4}^{0,025}$$

Avec $\hat{\sigma}_{\hat{d}}^2 = \hat{\sigma}_{\hat{\beta}_1}^2 + \hat{\sigma}_{\hat{\beta}'_1}^2$ et $\hat{d} = \hat{\beta}_1 + \hat{\beta}'_1$

Sous l'hypothèse H_0 , le ratio empirique de Student se présente comme :

$$t_{\hat{d}}^* = \frac{\hat{d}}{\hat{\sigma}_{\hat{d}}} = \frac{(1,5 - 1,62)}{\sqrt{(0,39473)^2 + (0,5192)^2}} = -0,183989$$

Enfin $|t_{\hat{d}}^*| < t_{n_1+n_2-4}^{0,025} = t_{58}^{0,025} = 2,000$. La différence d est non significativement différente de 0, c'est-à-dire il n'y a pas une différence significative de l'impact de la durée de formation sur le salaire des hommes et des femmes.

Remarque :

Dans la question (3) on a utilisé la propriété suivante :

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{Cov}(X, Y)$$

2 Modèle de Régression Linéaire Simple

on a $\text{cov}(\hat{\beta}_1, \hat{\beta}'_1) = 0$, car les deux régressions sont **indépendantes**

Exercice 12

On considère un modèle de régression linéaire simple sans constante

$$Y_i = \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n \text{ avec } \varepsilon_i \sim N(0, \sigma_\varepsilon^2).$$

- (1) Dérivez l'estimateur MCO de β_1 et trouvez sa variance.
- (2) Quelles propriétés numériques des estimateurs des MCO décrites dans le cas d'un modèle avec constante sont encore valables pour ce modèle ?

Solution

- (1) La minisation de la somme des carrés des résidus par rapport à $\hat{\beta}_1$:

$$\begin{aligned} \sum_{i=1}^n e_i^2 &= \sum_{i=1}^n (Y_i - \hat{\beta}_1 X_i)^2 \\ \frac{\partial \left(\sum_{i=1}^n e_i^2 \right)}{\partial \hat{\beta}_1} &= -2 \sum_{i=1}^n (Y_i - \hat{\beta}_1 X_i) X_i = 0 \end{aligned}$$

$$\text{en simplifiant l'équation, on trouve } \hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

on remplaçant $Y_i = \beta_1 X_i + \varepsilon_i$ dans la formule de $\hat{\beta}_1$ on trouve :

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n X_i \varepsilon_i}{\sum_{i=1}^n X_i^2} \text{ avec } E(\hat{\beta}_1) = \beta_1 \text{ puisque } X_i \text{ est non stochastique et } E(\varepsilon_i) = 0.$$

ainsi

$$\text{var}(\hat{\beta}_1) = \text{var}(\hat{\beta}_1 - \beta_1)^2 = E \left(\frac{\sum_{i=1}^n X_i \varepsilon_i}{\sum_{i=1}^n X_i^2} \right)^2 = \frac{\sigma^2 \sum_{i=1}^n X_i^2}{\left(\sum_{i=1}^n X_i^2 \right)^2} = \frac{\sigma^2}{\sum_{i=1}^n X_i^2}$$

2 Modèle de Régression Linéaire Simple

- (2) A partir de la condition du premier ordre dans la question (1), on obtient $\sum_{i=1}^n e_i X_i = 0$, où $e_i = Y_i - \hat{\beta}_1 X_i$. Par conséquent, $\sum_{i=1}^n e_i$ n'est pas nécessairement nul. Toutefois $\sum_{i=1}^n Y_i$ n'est pas nécessairement égale à $\sum_{i=1}^n \hat{Y}_i$. De plus, $\sum_{i=1}^n e_i \hat{Y}_i = \hat{\beta}_1 \sum_{i=1}^n e_i X_i = 0$, car $\hat{Y}_i = \hat{\beta}_1 X_i$ et $\sum_{i=1}^n e_i X_i = 0$. Par conséquent, seules deux des propriétés numériques considérées sont valables pour cette régression sans constante.

Exercice 13

En utilisant le fait que $(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + e_i$

Montrer que $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y}) = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$, et déduire par conséquent que

$$r_{Y,\hat{Y}}^2 = \frac{\left[\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) \right]^2}{\left(\sum_{i=1}^n (Y_i - \bar{Y})^2 \right) \left(\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \right)} = r_{X,Y}^2$$

Solution

En multipliant les deux termes de $(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + e_i$ par $(\hat{Y}_i - \bar{Y})$ en sommant on obtient :

$$\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

car $\sum_{i=1}^n (\hat{Y}_i - \bar{Y}) e_i = 0$ d'après les propriétés de la régression linéaire simple.

Ainsi

$$r_{Y,\hat{Y}}^2 = \frac{\left[\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) \right]^2}{\left(\sum_{i=1}^n (Y_i - \bar{Y})^2 \right) \left(\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \right)}$$

2 Modèle de Régression Linéaire Simple

$$r_{Y,\widehat{Y}}^2 = \frac{\left[\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2 \right]^2}{\left(\sum_{i=1}^n (Y_i - \bar{Y})^2 \right) \left(\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2 \right)}$$

$$= \frac{\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{SCE}{SCT} = R^2 = r_{X,Y}^2$$

Exercice 14

Montrer que :

$$\text{cov}(\widehat{\beta}_0; \widehat{\beta}_1) = -\bar{X} \text{var}(\widehat{\beta}_1) = -\sigma^2 \frac{\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Cela signifie que le signe de la covariance est déterminé par le signe de \bar{X} . Si \bar{X} est positif, cette covariance sera négative. Cela signifie également que si $\widehat{\beta}_0$ est surestimé, $\widehat{\beta}_1$ sera sous-estimé.

Solution

on a

$$\begin{aligned} \text{cov}(\widehat{\beta}_0; \widehat{\beta}_1) &= \text{cov}(\bar{Y} - \widehat{\beta}_1 \bar{X}; \widehat{\beta}_1) = E[((-\widehat{\beta}_1 \bar{X} + \bar{X}\beta_1)(\widehat{\beta}_1 - \beta_1))] \\ &= -\bar{X} E(\widehat{\beta}_1 - \beta_1)^2 = -\bar{X} \text{var}(\widehat{\beta}_1) = -\sigma^2 \frac{\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}$$

Chapitre 3

Modèle de Régression Linéaire Multiple

Lorsque nous avons introduit l'analyse de régression dans le chapitre précédent, nous avons parlé de modèles de régression contenant un certain nombre de variables prédictives.

Dans tous ces exemples, une seule variable prédictive du modèle aurait fourni une description inadéquate puisqu'un certain nombre de variables clés affectent la variable dépendante de manière importante et distincte. De plus, dans des situations de ce type, nous trouvons fréquemment que les prédictions de la variable dépendante basées sur un modèle ne contenant qu'une seule variable prédictive sont trop imprécises. Un modèle plus complexe, contenant des variables de prédiction supplémentaires est généralement plus utile pour fournir des prédictions suffisamment précises de la variable dépendante.

L'analyse de régression multiple est également très utile dans des situations expérimentales où l'expérimentateur peut contrôler les variables de prédiction. Un expérimentateur souhaitera généralement étudier simultanément un certain nombre de variables prédictives car presque toujours plus qu'une variable prédictive clé influencent la variable dépendante. Par exemple, dans une étude sur la productivité des équipes de travail, l'expérimentateur peut souhaiter contrôler à la fois la taille de l'équipe et le niveau de la prime.

Les modèles de régression multiple que nous décrivons maintenant peuvent être utilisés soit pour des données d'observation, soit pour des données expérimentales à partir d'un plan complètement aléatoire.

3 Modèle de Régression Linéaire Multiple

3.1 Présentation Formelle du Modèle

Nous considérons maintenant le cas où il y a m variables explicatives X_1, \dots, X_m . Le modèle de régression s'écrit :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \varepsilon_i \quad k = 1, \dots, m \quad (3.1)$$

est appelé un modèle de premier ordre avec m variables prédictives. On peut aussi écrire :

$$Y_i = \beta_0 + \sum_{k=1}^m \beta_k X_{ik} + \varepsilon_i \quad (3.2)$$

ou, si nous laissons $X_{i0} \equiv 1$, alors (3.2) peut être écrite ainsi :

$$Y_i = \sum_{k=0}^m \beta_k X_{ik} + \varepsilon_i \quad \text{où} \quad X_{i0} \equiv 1, \quad k = 0, \dots, m \quad (3.3)$$

ainsi l'équation (3.1) sera :

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \varepsilon_i \quad k = 1, \dots, m \quad (3.4)$$

En supposant que $E(\varepsilon_i) = 0$, la fonction du modèle de régression (3.1) est la suivante :

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m \quad (3.5)$$

Cette fonction de régression est un *hyperplan*, qui est un plan à plus de deux dimensions. Il n'est plus possible de visualiser la surface de cette fonction, comme nous pouvons le faire pour le cas de deux variables prédictives. Néanmoins, la signification des paramètres est analogue au cas de deux variables prédictives. Le paramètre β_m indique l'évolution de $E(Y)$ avec une augmentation unitaire de la variable prédictive X_m lorsque toutes les autres variables de prédiction du modèle de régression sont maintenues constantes.

3.2 Modèle de Régression Linéaire Général

Nous définissons le modèle général de régression linéaire, avec des termes d'erreur normaux en termes de X variables :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \varepsilon_i \quad k = 1, \dots, m \quad (3.6)$$

où

$\beta_0, \beta_1, \dots, \beta_m$ sont des paramètres

X_{i1}, \dots, X_{im} sont des constantes connues

3 Modèle de Régression Linéaire Multiple

ε_i sont indépendants et normalement distribués $N(0, \sigma^2)$

Remarque :

Le modèle de régression généralisée (3.6) englobe non seulement des variables prédictives quantitatives, mais également des *variables qualitatives*, telles que le sexe (homme, femme) ou le statut d'invalidité (non handicapé, partiellement ou totalement handicapé). Nous utilisons des variables indicatrices qui prennent les valeurs binaires 0 et 1 pour identifier les classes d'une variable qualitative.

Considérons une analyse de régression pour prédire la durée du séjour à l'hôpital (Y) en fonction de l'âge (X_1) et du sexe (X_2) du patient. Nous définissons (X_2) comme suit :

$$X_2 = \begin{cases} 1 & \text{si le patient est une femme} \\ 0 & \text{si le patient est un homme} \end{cases}$$

3.2.1 Présentation Matricielle du Modèle de Régression Linéaire Général

Nous présentons maintenant les principaux résultats de la régression linéaire générale (3.6) en termes de matrice.

C'est une propriété remarquable de l'algèbre matricielle que les résultats du modèle de régression linéaire général (3.6) en notation matricielle apparaissent exactement comme ceux du modèle de régression linéaire simple. Seuls les degrés de liberté et autres constantes liés au nombre de variables X et aux dimensions de certaines matrices qui sont différents. Nous sommes donc en mesure de présenter les résultats de manière très concise.

Certes, la notation matricielle peut masquer d'énormes complexités de calcul. Pour trouver l'inverse d'une matrice A d'ordre 10×10 , celà nécessite une énorme quantité de calcul, mais elle est simplement représentée par A^{-1} . Notre raison de mettre l'accent sur l'algèbre matricielle est qu'elle indique les étapes conceptuelles essentielles de la solution. Les calculs réels, dans tous les cas sauf les plus simples, seront effectués par ordinateur. peu importe pour nous que $(X'X)^{-1}$ représente l'inverse d'une matrice d'ordre 2×2 ou 10×10 . Le point important ici est de savoir ce que représente l'inverse de la matrice.

Pour exprimer un modèle de régression linéaire général (3.6) sous forme matricielle, nous devons définir les matrices suivantes :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \varepsilon_i \quad k = 1, \dots, m$$

3 Modèle de Régression Linéaire Multiple

$$\begin{aligned}
 Y_{n \times 1} &= \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} & X_{n \times (m+1)} &= \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,m} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,m} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,m} \end{pmatrix} \\
 \varepsilon_{n \times 1} &= \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} & \beta_{(m+1) \times 1} &= \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}
 \end{aligned} \tag{3.7}$$

En termes matriciels, le modèle de régression linéaire général (3.6) est le suivant :

$$Y_{n \times 1} = X_{n \times (m+1)} \times \beta_{(m+1) \times 1} + \varepsilon_{n \times 1} \tag{3.8}$$

- Y est le vecteur de la variable dépendante
- β est le vecteur des paramètres
- X est la matrice des constantes
- ε est le vecteur des variables indépendantes et normalement distribuées avec une moyenne $E(\varepsilon) = 0$ est une matrice variance-covariance égale à :

$$E(\varepsilon \varepsilon') = \begin{pmatrix} E(\varepsilon_1^2) & E(\varepsilon_1 \varepsilon_2) & E(\varepsilon_1 \varepsilon_3) & \cdots & E(\varepsilon_1 \varepsilon_n) \\ E(\varepsilon_2 \varepsilon_1) & E(\varepsilon_2^2) & E(\varepsilon_2 \varepsilon_3) & \cdots & E(\varepsilon_2 \varepsilon_n) \\ E(\varepsilon_3 \varepsilon_1) & E(\varepsilon_3 \varepsilon_2) & E(\varepsilon_3^2) & \cdots & E(\varepsilon_3 \varepsilon_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E(\varepsilon_n \varepsilon_1) & E(\varepsilon_n \varepsilon_2) & E(\varepsilon_n \varepsilon_3) & \cdots & E(\varepsilon_n^2) \end{pmatrix}$$

Puisque chaque terme d'erreur, ε_i a une moyenne nulle, les éléments diagonaux de cette matrice représenteront la variance des erreurs et les termes non diagonaux seront les covariances entre les différentes erreurs. Par conséquent, en utilisant les hypothèses 2 ($\text{var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$) et 3 ($\text{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0$), la matrice variance-covariance sera sous la forme :

$$\sigma^2(\varepsilon) = E(\varepsilon \varepsilon') = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 I$$

Par conséquent, le vecteur aléatoire Y a une moyenne :

$$E(Y) = X\beta_{n \times 1} \tag{3.9}$$

3 Modèle de Régression Linéaire Multiple

et la matrice variance-covariance de Y est la même que celle de ε :

$$\sigma^2_{n \times n}(Y) = \sigma^2 I \quad (3.10)$$

3.3 Estimation des Coefficients de Régression

Le critère des moindres carrés est généralisé comme suit pour le modèle général de régression linéaire (3.6) :

$$S = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_m X_{im})^2 \quad (3.11)$$

Pour dériver les équations normales par la méthode des moindres carrés, nous minimisons la quantité :

$$S = (Y - X\beta)' (Y - X\beta) \quad (3.12)$$

En développant (3.12), on obtient :

$$S = Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta$$

puisque $(X\beta)' = \beta'X'$. Notez maintenant que $Y'X\beta$ est une matrice d'ordre 1×1 , donc est égal à son transposé, $\beta'X'Y = (Y'X\beta)'$. Ainsi, on trouve :

$$S = Y'Y - 2\beta'X'Y + \beta'X'X\beta \quad (3.13)$$

Pour trouver la valeur de β qui minimise S , nous dérivons S par rapport à β . soit :

$$\frac{\partial S}{\partial \beta} = \begin{pmatrix} \frac{\partial S}{\partial \beta_0} \\ \frac{\partial S}{\partial \beta_1} \\ \vdots \\ \frac{\partial S}{\partial \beta_m} \end{pmatrix} \quad (3.14)$$

Ensuite, il s'ensuit que :

$$\frac{\partial S}{\partial \beta} = -2X'Y + 2X'X\beta \quad (3.15)$$

$$\hat{\beta} = (X'X)^{-1} X'Y \quad (3.16)$$

3 Modèle de Régression Linéaire Multiple

La solution obtenue est réalisable si la matrice carrée $(X'X)$ d'ordre $((m+1) \times (m+1))$ est inversible. En cas de colinéarité parfaite entre deux variables explicatives, la matrice $(X'X)$ est singulière et la méthode des MCO dans ce cas serait défaillante.

Soit \hat{Y} le vecteur des valeurs estimées \hat{Y}_i :

$$\hat{Y}_{n \times 1} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix}$$

En notation matricielle, on a alors :

$$\hat{Y}_{n \times 1} = X_{n \times (m+1)} \times \hat{\beta}_{(m+1) \times 1} \quad (3.17)$$

Puisque $(X'X)^{-1}X'$ est une matrice des constantes, les éléments de $\hat{\beta}$ sont des fonctions linéaires des Y . $\hat{\beta}$ est donc un *estimateur linéaire*. De plus, en remplaçant (3.8) en (3.16), nous obtenons

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$$

Puisque $E(\varepsilon) = 0$, nous avons $E(\hat{\beta}) = \beta$. Donc β est un *estimateur sans biais*. On a aussi

$$\begin{aligned} var(\hat{\beta}) &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1} \\ &= (X'X)^{-1}\sigma^2 \end{aligned}$$

Remarque :

Dans le cas d'estimation avec des données centrées¹, on peut déduire l'estimateur de β à partir des matrices des variances et covariances empiriques :

1. Données centrées sur la moyenne : soit X_i une variable observée sur n observations et \bar{X} sa moyenne, nous pouvons calculer une nouvelle variable $x_i = X_i - \bar{X}$ dont la somme est nulle par construction : $\sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n x_i = 0$. Par conséquent $\bar{x} = 0$.

3 Modèle de Régression Linéaire Multiple

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_m \end{pmatrix} = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & cov(X_1, X_3) & \cdots & cov(X_1, X_m) \\ cov(X_2, X_1) & var(X_2) & cov(X_2, X_3) & \cdots & cov(X_2, X_m) \\ cov(X_3, X_1) & cov(X_3, X_2) & var(X_3) & \cdots & cov(X_3, X_m) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(X_m, X_1) & cov(X_m, X_2) & cov(X_m, X_3) & \cdots & var(X_m) \end{pmatrix}^{-1} \times \begin{pmatrix} cov(X_1, Y) \\ cov(X_2, Y) \\ cov(X_3, Y) \\ \vdots \\ cov(X_m, Y) \end{pmatrix}$$

avec $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 - \dots - \hat{\beta}_m \bar{X}_m$.

3.3.1 Matrice Hat

Nous pouvons exprimer le résultat de la matrice pour \hat{Y} dans (3.17) comme suit en utilisant l'expression pour $\hat{\beta}$ dans (3.16) :

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$$

ou équivalent :

$$\hat{Y}_{n \times 1} = H_{n \times n} \times Y_{n \times 1} \quad (3.18)$$

où

$$H_{n \times n} = X(X'X)^{-1}X' \quad (3.19)$$

La matrice H ne concerne que les observations sur la variable prédictive X , comme il ressort de (3.19).

La matrice carrée H est appelée *matrice Hat*. Elle joue un rôle important dans le diagnostic pour l'analyse de régression, comme nous le verrons surtout lorsque nous examinerons si les résultats de la régression sont indûment influencés par une ou plusieurs observations. La matrice H est symétrique et possède la propriété spéciale (appelée *idempotence*) :

$$HH = H \quad (3.20)$$

En général, une matrice M est dite idempotente si $MM = M$.

3.4 Hypothèses de Régression

Le modèle de régression est linéaire en X (ou sur ces coefficients) et nous distinguons les hypothèses stochastiques (liées à l'erreur ε) des hypothèses structurelles.

3.4.1 Hypothèses Stochastiques

- $H_1 : E(\varepsilon_i) = 0$ pour tout $i = 1, \dots, n$, c'est-à-dire, les erreurs ont une moyenne nulle.
- $H_2 : \text{var}(\varepsilon_i) = \sigma^2$, pour tout $i = 1, \dots, n$, c'est-à-dire, les erreurs ont une variance constante (homoscédasticité).
- $H_3 : E(\varepsilon_i \varepsilon_j) = 0$ pour $i \neq j$ et $i, j = 1, \dots, n$, c'est-à-dire, les erreurs ne sont pas corrélées.
- H_4 : La variable explicative X est non-stochastique, c'est-à-dire fixée dans des échantillons répétés et, par conséquent, non corrélée avec les erreurs.
- $H_5 : \text{cov}(X_i, \varepsilon_i) = 0$ pour tout $i = 1, \dots, n$; Toutes les variables explicatives ne sont pas corrélées avec le terme d'erreur.

3.4.2 Hypothèses Structurelles

- H_6 : Absence de colinéarité entre les variables explicatives, cela implique que la matrice $(X'X)$ est régulière et que la matrice inverse $(X'X)^{-1}$ existe.
- $H_7 : (X'X)/n$ tend vers une matrice finie non singulière.
- H_8 : Le nombre d'observations est supérieur au nombre des séries explicatives.

Résidus

Soit le vecteur des résidus $e_i = Y_i - \hat{Y}_i$ noté par :

$$e_{n \times 1} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \quad (3.21)$$

En notation matricielle, on a alors :

$$e_{n \times 1} = Y_{n \times 1} - \hat{Y}_{n \times 1} = Y_{n \times 1} - X\hat{\beta}_{n \times 1} \quad (3.22)$$

3 Modèle de Régression Linéaire Multiple

Matrice de Variance-Covariance des Résidus

Les résidus e_i , comme les valeurs estimées \hat{Y}_i , peuvent être exprimés sous forme de combinaisons linéaires des observations de la variable dépendante Y_i , en utilisant le résultat obtenu en (3.18) pour \hat{Y} :

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$

Nous avons donc le résultat important suivant :

$$\underset{n \times 1}{e} = \left(\underset{n \times n}{I} - \underset{n \times n}{H} \right) \underset{n \times 1}{Y} \quad (3.23)$$

où H est la matrice Hat définie dans (3.19). La matrice $(I - H)$, comme la matrice H , est symétrique et idempotente.

La matrice variance-covariance du vecteur des résidus e fait intervenir la matrice $(I - H)$:

$$\underset{n \times n}{\sigma^2(e)} = \sigma^2(I - H) \quad (3.24)$$

et estimée par :

$$\hat{\sigma}^2(e) = MCR(I - H) \quad (3.25)$$

Remarques

La matrice variance-covariance de e dans (3.24) peut être présentée comme en (3.26). Puisque $e = (I - H)Y$, on obtient :

$$\sigma^2(e) = [I - H] \sigma^2(Y) [I - H]' \quad (3.26)$$

Maintenant, $\sigma^2(Y) = \sigma^2(\varepsilon) = \sigma^2 I$ pour le modèle d'erreur normale. Aussi, $(I - H)' = (I - H)$ à cause de la symétrie de la matrice. D'où :

$$\begin{aligned} \sigma^2(e) &= \sigma^2(I - H) I (I - H) \\ &= \sigma^2(I - H) (I - H) \end{aligned} \quad (3.27)$$

Etant donné que la matrice $(I - H)$ est idempotente, nous connaissons que

$$(I - H)(I - H) = I - H$$

et nous obtenons la formule (3.24).

3.5 Estimation du Maximum de Vraisemblance (EMV)

Au chapitre 2, nous avons présenté la méthode d'estimation du maximum de vraisemblance, qui consiste à spécifier la distribution à partir de

3 Modèle de Régression Linéaire Multiple

laquelle nous effectuons l'échantillonnage et à écrire la densité de joint de notre échantillon. Cette densité conjointe est alors appelée fonction de vraisemblance car elle donne, pour un ensemble de paramètres spécifiant la distribution, la probabilité d'obtenir l'échantillon observé. Pour l'équation de régression, spécifier la distribution des perturbations spécifie à son tour la fonction de vraisemblance. Ces perturbations peuvent être de type Poisson, Exponentielle, Normale, etc. Une fois cette distribution choisie, la fonction de vraisemblance est maximisée et l'EMV des paramètres de régression est obtenue. Les estimateurs de maximum de vraisemblance sont souhaitables car ils sont (1) *cohérents dans des conditions relativement générales*, (2) *asymptotiquement normaux*, (3) *asymptotiquement efficaces* et (4) *invariants aux reparamétrisations du modèle*. Certaines des propriétés indésirables de l'EMV sont les suivantes :

- (1) elle nécessite des hypothèses de distribution explicites sur les perturbations,
- (2) pour les EMV, les propriétés d'échantillon fini peuvent être très différentes de leurs propriétés asymptotiques.

Par exemple, les EMV peuvent être biaisés même s'ils sont cohérents et leurs estimations de covariance peuvent être trompeuses pour les petits échantillons. Dans cette section, nous dérivons l'EMV sous la normalité des perturbations.

L'Hypothèse de Normalité $\varepsilon \sim N(0, \sigma^2 I)$

Cette hypothèse supplémentaire nous permet de déduire des distributions d'estimateurs et d'autres variables aléatoires. Ceci est important pour la construction d'intervalles de confiance et de tests d'hypothèses. En fait, en utilisant le fait que $\hat{\beta} = \beta + (X' X)^{-1} X' \varepsilon$, on peut facilement voir que $\hat{\beta}$ est une combinaison linéaire des ε 's. Cependant, une combinaison linéaire de variables aléatoires normales est elle-même une variable aléatoire normale. Donc, $\hat{\beta}$ est $N(\beta, \sigma^2 (X' X)^{-1})$. De même Y est $N(X\beta, \sigma^2 I)$ et ε est $N(0, \sigma^2 (I - H))$. De plus, nous pouvons écrire la fonction de densité de probabilité conjointe des ε 's comme :

$$f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n; \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left(-\frac{\varepsilon' \varepsilon}{2\sigma^2} \right)$$

Pour obtenir la fonction de vraisemblance, nous effectuons la transformation $\varepsilon = Y - X\beta$ et notons que le jacobien de la transformation est

3 Modèle de Régression Linéaire Multiple

1.

$$f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n; \beta\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^2}\right)$$

En prenant le log de cette vraisemblance, nous obtenons

$$\log L(\beta, \sigma^2) = -\left(\frac{n}{2}\right) \log(2\pi\sigma^2) - \frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^2}$$

En maximisant cette vraisemblance par rapport à β et σ^2 , on obtient les estimateurs maximum de vraisemblance (EMV)

$$\begin{aligned} \frac{\partial \log L(\beta, \sigma^2)}{\partial \beta} &= \frac{2X'Y - 2X'X\beta}{2\sigma^2} \\ \frac{\partial \log L(\beta, \sigma^2)}{\partial \sigma^2} &= \frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^4} - \frac{n}{2\sigma^2} \end{aligned}$$

En fixant ces conditions de premier ordre à zéro, on obtient

$$\hat{\beta}_{EMV} = \hat{\beta}_{MCO} \quad \text{et} \quad \hat{\sigma}_{EMV}^2 = \frac{(Y - X\beta)'(Y - X\beta)}{n} = \frac{SCR}{n} = \frac{e'e}{n}$$

Intuitivement, seul le second terme du log vraisemblance contient β et ce terme (sans le signe négatif) a déjà été minimisé par rapport à β , ce qui nous donne l'estimateur des MCO. Notez que $\hat{\sigma}_{EMV}^2$ ne diffère de $\hat{\sigma}_{MCO}^2$ que par les degrés de liberté. Il est clair que $\hat{\beta}_{EMV}$ est non biaisé pour β alors que $\hat{\sigma}_{EMV}^2$ ne l'est pas pour σ^2 .

3.6 Analyse de la Variance ANOVA et Coefficient de Détermination Multiple

Sommes des Carrés

Pour voir comment les sommes des carrés sont exprimées en notation matricielle, commençons par la **somme des carrés totale** :

$$SCT = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \quad (3.28)$$

3 Modèle de Régression Linéaire Multiple

Nous savons que :

$$Y'Y = \sum_{i=1}^n Y_i^2$$

$$\left(\sum_{i=1}^n Y_i \right)^2$$

Le terme de soustraction $\frac{n}{n}$ sous forme matricielle utilise la matrice J , qui est une matrice dont tout les termes sont égales à 1 est définie comme suit :

$$J_{n \times n} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \quad (3.29)$$

alors

$$\frac{\left(\sum_{i=1}^n Y_i \right)^2}{n} = \left(\frac{1}{n} \right) Y' J Y \quad (3.30)$$

Par conséquent, il s'ensuit que :

$$SCT = Y'Y - \left(\frac{1}{n} \right) Y' J Y \quad (3.31)$$

De la même manière, la *somme des carrés des résidus* peut être exprimée matriciellement comme dans (2.13) :

$$SCR = e'e = (Y - X\hat{\beta})' (Y - X\hat{\beta}) \quad (3.32)$$

qui est égale à :

$$SCR = Y'Y - \hat{\beta}' X' Y \quad (3.33)$$

Enfin on peu déduire la valeur de la *somme des carrés expliquée SCE*

$$SCE = \hat{\beta}' X' Y - \left(\frac{1}{n} \right) Y' J Y \quad (3.34)$$

Somme des Carrés en tant que Formes Quadratiques

Dans l'analyse de la variance ANOVA, les statistiques *SCE*, *SCR* et *SCT* ont toutes des formes quadratiques², comme on peut le voir en reformulant $\hat{\beta}' X'$.

2. **Rappel** : En général, une forme quadratique est définie comme :

$$Y' A Y = \sum_{1 \times 1}^n \sum_{i=1}^n a_{ij} Y_i Y_j$$
 où $a_{ij} = a_{ji}$ (A est une matrice symétrique d'ordre $n \times n$)

3 Modèle de Régression Linéaire Multiple

D'après (3.17), nous savons que :

$$\widehat{Y}' = \left(X\widehat{\beta} \right)' = \widehat{\beta}' X'$$

Nous utilisons maintenant le résultat en (3.18) pour obtenir :

$$\widehat{\beta}' X' = (HY)'$$

Puisque H est une matrice symétrique et que $H' = H$, on obtient finalement :

$$\widehat{\beta}' X' = Y' H \quad (3.35)$$

Ce résultat nous permet d'exprimer les sommes de carrés comme suit :

$$SCT = Y' Y - \left(\frac{1}{n} \right) Y' J Y = Y' \left[I - \left(\frac{1}{n} \right) J \right] Y \quad (3.36)$$

$$SCR = e' e = \left(Y - X\widehat{\beta} \right)' \left(Y - X\widehat{\beta} \right) = Y' Y - \widehat{\beta}' X' Y = Y' (I - H) Y \quad (3.37)$$

$$SCE = \widehat{\beta}' X' Y - \left(\frac{1}{n} \right) Y' J Y = Y' \left[H - \left(\frac{1}{n} \right) J \right] Y \quad (3.38)$$

Le tableau (3.1) présente l'analyse de la variance pour un modèle de régression multiple, ainsi que les carrés moyens MCE et MCR :

$$MCE = \frac{SCE}{m}$$

$$MCR = \frac{SCR}{n - m - 1} = \frac{e' e}{n - m - 1} = \hat{\sigma}^2$$

L'espérance de MCR est σ^2 , comme pour la régression linéaire simple. L'espérance de MCE est σ^2 plus une quantité non négative. Par exemple, lorsque $m = 2$, nous avons :

$$\begin{aligned} E(MCE) &= \sigma^2 + \frac{1}{2} \left[\beta_1^2 \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 + \beta_2^2 \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2 \right. \\ &\quad \left. + 2\beta_1\beta_2 \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) \right] \end{aligned}$$

Notez que si β_1 et β_2 sont égaux à zéro, $E(MCE) = \sigma^2$, sinon $E(MCE) > \sigma^2$.

Source de variation	Somme des Carrés	Degrés de liberté (ddl)	Carrés Moyens
X	$\widehat{\beta}' X' Y - \left(\frac{1}{n}\right) Y' JY$	m	$MCE = \frac{SCE}{m}$
Résidu	$Y' Y - \widehat{\beta}' X' Y$	$n - m - 1$	$MCR = \frac{SCR}{n - m - 1}$
Total	$Y' Y - \left(\frac{1}{n}\right) Y' JY$	$n - 1$	

TABLE 3.1 – Analyse de la Variance par une Régression Linéaire Multiple

Coefficient de Détermination Multiple

Le coefficient de détermination multiple, noté R^2 , est défini comme suit :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \quad (3.39)$$

Il mesure la réduction proportionnelle de la variation totale de Y associée à l'utilisation des variables X_1, \dots, X_m . Le coefficient de détermination multiple R^2 se réduit au coefficient de détermination simple d'une régression linéaire simple lorsque $m = 1$, c'est-à-dire lorsqu'une seule variable X figure dans le modèle de régression. Juste comme avant, nous avons :

$$0 \leq R^2 \leq 1 \quad (3.40)$$

où R^2 prend la valeur 0 lorsque tout les $\beta_m = 0$ ($k = 1, \dots, m$) et la valeur 1 lorsque toutes les observations Y tombent exactement sur la surface de régression estimée, c.-à-d., lorsque $Y_i = \widehat{Y}_i$ pour tout i .

Ajouter plus de variables prédictives X au modèle de régression ne peut qu'accroître R^2 et ne jamais le réduire, car SCR ne peut jamais devenir plus grande avec plus de variables X et SCT est toujours identique pour la variable dépendante. Etant donné que R^2 peut généralement être agrandi en incluant un plus grand nombre de variables prédictives, il est parfois suggéré d'utiliser une mesure modifiée qui ajuste le nombre de variables X dans le modèle.

Le coefficient de détermination multiple ajusté, noté \bar{R}^2 , ajuste R^2 en

3 Modèle de Régression Linéaire Multiple

divisant chaque somme de carrés par ses degrés de liberté associés :

$$\bar{R}^2 = 1 - \frac{\frac{SCR}{SCT}}{n-1} = 1 - \left(\frac{n-1}{n-m-1} \right) \frac{SCR}{SCT} \quad (3.41)$$

Ce coefficient ajusté de détermination multiple peut en fait devenir plus petit quand une autre variable X est introduite dans le modèle, car toute diminution de SCR peut être plus que compensée par la perte d'un degré de liberté du dénominateur ($n - m - 1$).

Remarques :

- (1) Pour distinguer les coefficients de détermination de la régression simple des régressions multiples, nous désignerons désormais les premiers par le coefficient de détermination simple.
- (2) On peut montrer que le coefficient de détermination multiple R^2 peut être considéré comme un coefficient de détermination simple entre les variables dépendantes Y_i ; et les valeurs ajustées \hat{Y}_i .
- (3) Une valeur élevée de R^2 n'implique pas nécessairement que le modèle ajusté est utile. Par exemple, des observations des variables prédictives qui ont été collectées à quelques niveaux non suffisantes. Malgré un R^2 élevé dans ce cas, le modèle ajusté peut ne pas être utile si la plupart des prévisions nécessitent des extrapolations en dehors de la région des observations. Encore une fois, même si R^2 est grand, SCE peut être encore trop grand pour que les inférences soient utiles lorsqu'une précision élevée est requise.

Coefficient de Corrélation Multiple

Le coefficient de corrélation multiple R est la racine carrée positive de R^2 :

$$R = \sqrt{R^2} \quad (3.42)$$

Lorsqu'il y a une variable X dans le modèle de régression (3.8), c'est-à-dire lorsque $m = 1$, le coefficient de corrélation multiple r est égal à la valeur absolue du coefficient de corrélation simple $r_{X,Y}$.

Test de Fisher pour une Régression Multiple

Pour vérifier s'il existe une relation de régression entre la variable dépendante Y et l'ensemble des variables X_1, \dots, X_m , c'est-à-dire choisir

3 Modèle de Régression Linéaire Multiple

entre les hypothèses suivantes

$$\begin{aligned} H_0: \quad & \beta_1 = \beta_2 = \dots = \beta_m = 0 \\ H_1: \quad & \text{Il existe au moins un de ces coefficients } \beta_k \ (k = 1, \dots, m) \text{ non nul} \end{aligned} \quad (3.43)$$

calcul de la statistique empirique de Fisher F^* :

$$F^* = \frac{\frac{SCE}{m}}{\frac{SCR}{n-m-1}} = \frac{SCE(X_1, \dots, X_m)}{\frac{m}{SCR(X_1, \dots, X_m)}} \quad (3.44)$$

La règle de décision avec ce test pour un niveau de signification α est la suivante :

$$\begin{aligned} \text{Si } F^* \leq F_{m;n-m-1}^\alpha & \text{ on accepte } H_0 \\ \text{Si } F^* > F_{m;n-m-1}^\alpha & \text{ on accepte } H_1 \end{aligned} \quad (3.45)$$

L'existence d'une relation de régression en elle-même ne garantit évidemment pas que des prédictions utiles puissent être faites.

Notez que lorsque $m = 1$, ce test se réduit au test de Fisher pour une régression linéaire simple, que ce soit $\beta_1 = 0$ ou non.

R^2 versus \bar{R}^2

Puisque MCO minimise la somme des carrés des résidus, l'ajout d'une ou plusieurs variables à la régression ne peut pas augmenter cette somme résiduelle. Après tout, nous minimisons un ensemble de paramètres de dimension plus grande et le minimum y est inférieur ou égal à celui d'un sous-ensemble de l'espace de paramètre. Par conséquent, pour la même variable dépendante Y , l'ajout de variables fait que $\sum_{i=1}^n e_i^2$ non croissant

et R^2 non décroissant, puisque $R^2 = 1 - \left(\sum_{i=1}^n e_i^2 \right) / \left(\sum_{i=1}^n (Y_i - \bar{Y})^2 \right)$.

Par conséquent, un critère de sélection d'une régression qui "maximise R^2 " n'a pas de sens, car nous pouvons ajouter plus de variables à cette régression et améliorer ce R^2 (ou au pire le laisser le même). Afin de pénaliser le chercheur pour avoir ajouté une variable supplémentaire, on calcule

$$\bar{R}^2 = 1 - \frac{\left[\left(\sum_{i=1}^n e_i^2 \right) / (n - (m + 1)) \right]}{\left[\left(\sum_{i=1}^n (Y_i - \bar{Y})^2 \right) / (n - 1) \right]}$$

où $\sum_{i=1}^n e_i^2$ et $\sum_{i=1}^n (Y_i - \bar{Y})^2$ ont été ajustés en fonction de leur degré de

3 Modèle de Régression Linéaire Multiple

liberté. Notez que le numérateur est le $\hat{\sigma}^2 = \left(\sum_{i=1}^n e_i^2 \right) / (n - (m + 1))$. $\sum_{i=1}^n e_i^2$ n'augmente pas à mesure que nous ajoutons plus de variables, mais les degrés de liberté diminuent de 1 à chaque variable ajoutée. Par conséquent, $\hat{\sigma}^2$ ne diminuera que si l'effet de la diminution de $\sum_{i=1}^n e_i^2$ l'emporte sur l'effet de la perte d'un degré de liberté sur $\hat{\sigma}^2$. C'est exactement l'idée qui sous-tend \bar{R}^2 , c'est-à-dire pénaliser chaque variable ajoutée en diminuant les degrés de liberté par 1. Par conséquent, cette variable augmentera \bar{R}^2 uniquement si la réduction de $\sum_{i=1}^n e_i^2$ l'emporte sur cette perte, c'est-à-dire que si $\hat{\sigma}^2$ est diminué. En utilisant la définition de \bar{R}^2 , on peut la relier à R^2 comme suit :

$$\bar{R}^2 = 1 - \left(1 - R^2 \right) \left[\frac{n - 1}{n - (m + 1)} \right]$$

3.7 Les Tests de Signification des Paramètres de Régression

Les estimateurs des moindres carrés et du maximum de vraisemblance de $\hat{\beta}$ sont non biaisés :

$$E(\hat{\beta}) = \beta \quad (3.46)$$

La matrice de variance-covariance $\sigma^2(\hat{\beta})$:

$$\begin{aligned} \sigma^2(\hat{\beta})_{(m+1) \times (m+1)} &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\ &= \begin{pmatrix} \sigma^2(\hat{\beta}_0) & \sigma(\hat{\beta}_0, \hat{\beta}_1) & \dots & \sigma(\hat{\beta}_0, \hat{\beta}_m) \\ \sigma(\hat{\beta}_1, \hat{\beta}_0) & \sigma^2(\hat{\beta}_1) & \dots & \sigma(\hat{\beta}_1, \hat{\beta}_m) \\ \vdots & \vdots & & \vdots \\ \sigma(\hat{\beta}_m, \hat{\beta}_0) & \sigma(\hat{\beta}_m, \hat{\beta}_1) & \dots & \sigma^2(\hat{\beta}_m) \end{pmatrix} \end{aligned} \quad (3.47)$$

est donnée par :

$$\sigma^2(\hat{\beta})_{(m+1) \times (m+1)} = \sigma^2 \times (X'X)^{-1} \quad (3.48)$$

3 Modèle de Régression Linéaire Multiple

La matrice de variance-covariance estimée $\hat{\sigma}^2(\hat{\beta})$:

$$\hat{\sigma}^2(\hat{\beta})_{(m+1) \times (m+1)} = \begin{pmatrix} \hat{\sigma}^2(\hat{\beta}_0) & \hat{\sigma}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \hat{\sigma}(\hat{\beta}_0, \hat{\beta}_m) \\ \hat{\sigma}(\hat{\beta}_1, \hat{\beta}_0) & \hat{\sigma}^2(\hat{\beta}_1) & \cdots & \hat{\sigma}(\hat{\beta}_1, \hat{\beta}_m) \\ \vdots & \vdots & & \vdots \\ \hat{\sigma}(\hat{\beta}_m, \hat{\beta}_0) & \hat{\sigma}(\hat{\beta}_m, \hat{\beta}_1) & \cdots & \hat{\sigma}^2(\hat{\beta}_m) \end{pmatrix} \quad (3.49)$$

est donnée par :

$$\hat{\sigma}^2(\hat{\beta})_{(m+1) \times (m+1)} = \hat{\sigma}^2 \times (X' X)^{-1} = MCR \times (X' X)^{-1} \quad (3.50)$$

À partir de $\hat{\sigma}^2(\hat{\beta})$, on peut obtenir $\hat{\sigma}^2(\hat{\beta}_0)$, $\hat{\sigma}^2(\hat{\beta}_1)$ ou toute autre variance nécessaire, ou toute covariance nécessaire.

3.7.1 Intervalle de Confiance de β_k

Pour le modèle de régression à erreur normale (3.8), nous avons :

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}(\hat{\beta}_k)} \sim t_{n-m-1} \quad k = 0, 1, \dots, m \quad (3.51)$$

Par conséquent, l'intervalle de confiance pour β_k avec un niveau de confiance $(1 - \alpha)\%$ se présente comme :

$$IC_{\beta_k} = \left[\hat{\beta}_k \pm t_{n-2}^{\alpha/2} \hat{\sigma}(\hat{\beta}_k) \right] \quad (3.52)$$

3.7.2 Intervalle de Confiance de la Variance de l'Erreur

L'intervalle de confiance de la variance de l'erreur permet de déterminer une fourchette de variation de l'amplitude de l'erreur. Pour un intervalle à $(1 - \alpha)\%$, il est donné par

$$IC = \left[\frac{(n - m - 1) \hat{\sigma}^2}{\chi_1^2}; \frac{(n - m - 1) \hat{\sigma}^2}{\chi_2^2} \right]$$

Avec χ_1^2 à $n - m - 1$ degrés de liberté et $\alpha/2$ de probabilité³ d'être dépassée et χ_2^2 à $n - m - 1$ degrés de liberté et $(1 - \alpha/2)$ de probabilité

3. **Attention**, la loi du chi-deux n'est pas symétrique, il convient donc de lire sur la table les deux probabilités $(1 - \alpha/2)$ et $\alpha/2$.

3 Modèle de Régression Linéaire Multiple

d'être dépassée.

3.7.3 Test de Signification d'un Seul Coefficient $\beta_k = 0$

Il s'agit d'un test de Fisher partiel permettant de déterminer si un coefficient de régression particulier β_k est égal à zéro. Les alternatives sont :

$$\begin{aligned} H_0: \beta_k &= 0 \\ H_1: \beta_k &\neq 0 \end{aligned} \quad (3.53)$$

et la statistique empirique F de Fisher est égales à :

$$F^* = \frac{\frac{SCE(X_k | X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_m)}{1}}{\frac{SCR(X_1, \dots, X_m)}{n - m - 1}} \quad (3.54)$$

$$= \frac{\frac{SCR(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_m) - SCR(X_1, \dots, X_m)}{1}}{\frac{SCR(X_1, \dots, X_m)}{n - m - 1}} \quad (3.55)$$

$$= \frac{\frac{SCR(R) - SCR(C)}{dll(R) - dll(C)}}{\frac{SCR(C)}{dll(C)}}$$

où $SCR(R)$ concerne SCR du modèle de régression *réduit*, c'est-à-dire sans inclusion de la seule variable prédictive X_k sujet de ce test de signification, et $SCR(C)$ concerne SCR du modèle de régression *complet* dont toutes les variables prédictives sont incluses.

$$(C): Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{k-1} X_{i,k-1} + \beta_{k+1} X_{i,k+1} + \dots + \beta_m X_{i,m} + \varepsilon_i$$

$$(R): Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{i,m} + \varepsilon_i \quad (3.56)$$

Si H_0 tient, $F^* \sim F_{1;n-m-1}$. Les grandes valeurs de F^* mènent à l'acceptation de H_1 . Les statistiques qui fournissent des sommes de carrés supplémentaires permettent d'utiliser ce test sans avoir à s'adapter au modèle réduit.

Une statistique de test équivalente est celle de Student dans le cas d'une

3 Modèle de Régression Linéaire Multiple

régression linéaire simple c'est-à-dire de $t_{\hat{\beta}_k}^*$:

$$t_{\hat{\beta}_k}^* = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \quad (3.57)$$

Si H_0 tient, $t_{\hat{\beta}_k}^* \sim t_{n-m-1}$. Les grandes valeurs de $t_{\hat{\beta}_k}^*$ mènent à l'acceptation de H_1 . La puissance du test de Student peut être obtenue avec les degrés de liberté modifiés à $(n - m - 1)$.

Comme les deux tests sont équivalents, le choix est généralement fait en termes d'informations disponibles fournies par l'output de la régression. Nous pouvons appliquer le test de Student pour les hypothèses :

$$\begin{aligned} H_0: \beta_k &= c \\ H_1: \beta_k &\neq c \end{aligned}$$

avec c une constante fixe. Dans ce cas la statistique empirique de Student sous H_0 est :

$$t_{\hat{\beta}_k}^* = \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} = \frac{\hat{\beta}_k - c}{\hat{\sigma}_{\hat{\beta}_k}}$$

3.7.4 Test de Signification de plusieurs Coefficients β_k

Ceci est un autre test partiel de Fisher. Ici, les hypothèses sont :

$$\begin{aligned} H_0: \beta_q &= \beta_{q+1} = \dots = \beta_m = 0 \\ H_1: \text{Il existe au moins un de ces coefficients } \beta_k &(k = 1, \dots, m) \text{ non nul} \end{aligned} \quad (3.58)$$

où, par commodité, nous arrangeons le modèle de sorte que les derniers coefficients $(m - q + 1)$ soient ceux à tester. La statistique empirique de test est :

$$F^* = \frac{\frac{SCE(X_q, \dots, X_m | X_1, \dots, X_{q-1})}{m - q + 1}}{\frac{SCR(X_1, \dots, X_m)}{n - m - 1}} \quad (3.59)$$

ou bien

$$F^* = \frac{\frac{SCR(X_1, \dots, X_{q-1}) - SCR(X_1, \dots, X_m)}{m - q + 1}}{\frac{SCR(X_1, \dots, X_m)}{n - m - 1}}$$

Si H_0 tient, $F^* \sim F_{m-q+1; n-m-1}$. Les grandes valeurs de F^* mènent à l'acceptation de H_1 .

3 Modèle de Régression Linéaire Multiple

Notez que la statistique de test (3.59) englobe en réalité les deux cas précédents. Si $q = 1$, le test consiste à déterminer si tous les coefficients de régression sont égaux à zéro. Si $q = p - 1$, le test consiste à déterminer si un coefficient de régression unique est égal à zéro. Notez également que la statistique de test (3.59) peut être calculée sans devoir s'adapter au modèle réduit si le package de régression fournit les sommes en carrées supplémentaires nécessaires :

$$\begin{aligned} SCE(X_q, \dots, X_m | X_1, \dots, X_{q-1}) &= SCE(X_q | X_1, \dots, X_{q-1}) + \dots \\ &\quad + SCE(X_m | X_1, \dots, X_{m-1}) \end{aligned} \quad (3.60)$$

La statistique de test (3.59) peut être énoncée de manière équivalente en termes de coefficients de détermination multiple pour les modèles complet et réduit lorsque ces modèles contiennent le terme d'interception β_0 , comme suit :

$$F^* = \frac{\frac{R_{Y|1\dots m}^2 - R_{Y|1\dots q-1}^2}{m - q + 1}}{\frac{1 - R_{Y|1\dots m}^2}{n - m - 1}} \quad (3.61)$$

où $R_{Y|1\dots m}^2$ désigne le coefficient de détermination multiple lorsque Y régresse sur toutes les variables X , et $R_{Y|1\dots q-1}^2$ désigne le coefficient lorsque Y régresse uniquement sur X_1, \dots, X_{q-1} .

3.7.5 Autres Tests

Lorsque vous souhaitez effectuer des tests sur les coefficients de régression sans indiquer si un ou plusieurs β_k sont égaux à zéro, vous ne pouvez pas utiliser de sommes supplémentaires de carrés et la méthode des tests linéaires généraux requiert des ajustements séparés des modèles complet et réduit.

Exemple 1 : Pour le modèle complet contenant trois variables prédictives :

$$(C) : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \quad \text{Modèle complet} \quad (3.62)$$

on veut par exemple tester :

$$\begin{aligned} H_0; \beta_1 &= \beta_2 \\ H_1; \beta_1 &\neq \beta_2 \end{aligned} \quad (3.63)$$

La procédure serait d'ajuster le modèle complet (3.62), puis le modèle

3 Modèle de Régression Linéaire Multiple

réduit :

$$(R) : Y_i = \beta_0 + \beta_c (X_{i1} + X_{i2}) + \beta_3 X_{i3} + \varepsilon_i \quad \text{Modèle réduit} \quad (3.64)$$

où β_c désigne le coefficient commun pour β_1 et β_2 sous H_0 et $(X_{i1} + X_{i2})$ est la nouvelle variable X correspondante. Nous utilisons ensuite la statistique de test linéaire générale F^* de Fisher avec 1 et $n - 4$ degrés de liberté.

Notez que H_0 peut être réécrite en tant que $H_0; \beta_1 - \beta_2 = 0$. Ceci peut être testé en utilisant une statistique empirique de Student qui teste si $d = \beta_1 - \beta_2$ est égal à zéro. A partir de la régression sans restriction (modèle complet), nous pouvons obtenir $\hat{d} = \hat{\beta}_1 - \hat{\beta}_2$ avec $\text{var}(\hat{d}) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) - 2\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$. La matrice de variance-covariance des coefficients de régression peut être donnée avec n'importe quel logiciel de régression. Cela signifie que $\hat{\sigma}(\hat{d}) = \sqrt{\text{var}(\hat{d})}$ et que la statistique empirique de Student est $t_d^* = \frac{\hat{d}}{\hat{\sigma}(\hat{d})}$ qui est distribuée selon

une loi de Student t_{n-4} sous H_0 . Alternativement, on peut exécuter un test de Fisher avec la somme des carrés des résidus SCR du modèle restreint (réduit) obtenu en exécutant la régression (3.64).

Exemple 2 : Ci-dessous un autre exemple où des sommes supplémentaires des carrés ne peuvent pas être utilisées dans le test pour le modèle de régression (3.62) :

$$\begin{aligned} H_0; \quad &\beta_1 = 3, \quad \beta_3 = 5 \\ H_1; \quad &\beta_1 \neq 3, \quad \beta_3 \neq 5 \end{aligned} \quad (3.65)$$

Ici, le modèle réduit serait :

$$(R) : Y_i - 3X_{i1} - 5X_{i3} = \beta_0 + \beta_2 X_{i2} + \varepsilon_i \quad \text{Modèle réduit} \quad (3.66)$$

Notez la nouvelle variable dépendante $Y_i - 3X_1 - 5X_3$ dans le modèle réduit, puisque $\beta_1 X_1$ et $\beta_3 X_3$ sont des constantes connues sous H_0 . Nous utilisons ensuite la statistique de test linéaire générale F^* de Fisher présenté dans l'équation (3.55) avec 2 et $n - 4$ degrés de liberté.

Exemple 3 : Tester l'hypothèse commune $H_0; \beta_3 = 1$ et $\beta_2 - 2\beta_4 = 0$. Ces deux restrictions sont généralement obtenues à partir d'informations préalables ou imposées par la théorie. La première restriction est $\beta_3 = 1$. La valeur 1 aurait pu être une autre constante. La deuxième restriction montre qu'une combinaison linéaire de β_2 et β_4 est égale à zéro. En

3 Modèle de Régression Linéaire Multiple

substituant ces restrictions en (3.6) nous obtenons :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + X_{i3} + \frac{1}{2}\beta_2 X_{i4} + \beta_5 X_{i5} + \dots + \beta_m X_{i,m} + \varepsilon_i$$

qui peut être écrit comme

$$Y_i - X_{i3} = \beta_0 + \beta_1 X_{i1} + \beta_2 \left(X_{i2} + \frac{1}{2}X_{i4} \right) + \beta_5 X_{i5} + \dots + \beta_m X_{i,m} + \varepsilon_i$$

Par conséquent, SCR du modèle réduit peut être obtenu en régressant $(Y - X_3)$ sur $\left(X_2 + \frac{1}{2}X_4 \right), X_5, \dots, X_m$. Cette régression a $n - (m - 2)$ degrés de liberté. SCR du modèle non restreint (complet) est obtenue à partir de la régression avec tous les X inclus. La statistique empirique de Fisher F^* résultante a 2 et $(n - m - 1)$ degrés de liberté.

3.7.6 Coefficients de Détermination Partielle

Les sommes supplémentaires de carrés sont non seulement utiles pour les tests sur les coefficients d'un modèle de régression multiple, mais elles se rencontrent également dans les mesures descriptives de la relation appelées *coefficients de détermination partielle*. Rappelons que le coefficient de détermination multiple, R^2 , mesure la réduction proportionnelle de la variation de Y obtenue par l'introduction de l'ensemble des variables prédictives X considérées dans le modèle. En revanche, un coefficient de détermination partielle mesure la *contribution marginale d'une variable X* lorsque toutes les autres sont déjà incluses dans le modèle.

Cas de deux variables prédictives

Nous considérons d'abord un modèle de régression multiple de premier ordre avec deux variables prédictives :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

$SCR(X_2)$ mesure la variation de Y lorsque X_2 est inclus dans le modèle. $SCR(X_1, X_2)$ mesure la variation de Y lorsque X_1 et X_2 sont inclus dans le modèle. Par conséquent, la réduction marginale relative de la variation de Y associée à X_1 lorsque X_2 est déjà dans le modèle est :

$$\frac{SCR(X_2) - SCR(X_1, X_2)}{SCR(X_2)} = \frac{SCE(X_1 | X_2)}{SCR(X_2)} \quad (3.67)$$

Cette mesure est le coefficient de détermination partielle entre Y et X_1 ,

3 Modèle de Régression Linéaire Multiple

étant donné que X_2 est dans le modèle, notons cette mesure par $R_{Y1|2}^2$:

$$R_{Y1|2}^2 = \frac{SCR(X_2) - SCR(X_1, X_2)}{SCR(X_2)} = \frac{SCE(X_1 | X_2)}{SCR(X_2)} \quad (3.68)$$

Ainsi, $R_{Y1|2}^2$ mesure la réduction proportionnelle de la variation de Y restant après X_2 est incluse dans le modèle obtenu en incluant également X_1 dans le modèle.

Le coefficient de détermination partielle entre Y et X_2 , étant donné que X_1 est dans le modèle, est défini en conséquence par :

$$R_{Y2|1}^2 = \frac{SCE(X_2 | X_1)}{SCR(X_1)} \quad (3.69)$$

Cas Général

La généralisation des coefficients de détermination partielle à trois variables X ou plus dans le modèle est immédiate. Par exemple :

$$R_{Y1|23}^2 = \frac{SCE(X_1 | X_2, X_3)}{SCR(X_2, X_3)} \quad (3.70)$$

$$R_{Y2|13}^2 = \frac{SCE(X_2 | X_1, X_3)}{SCR(X_1, X_3)} \quad (3.71)$$

$$R_{Y3|12}^2 = \frac{SCE(X_3 | X_1, X_2)}{SCR(X_1, X_2)} \quad (3.72)$$

$$R_{Y4|123}^2 = \frac{SCE(X_4 | X_1, X_2, X_3)}{SCR(X_1, X_2, X_3)} \quad (3.73)$$

Notez que dans les indices de R^2 , les entrées à gauche de la barre verticale montrent tour à tour la variable prise comme dépendante et la variable X est ajoutée. Les entrées à droite de la barre verticale indiquent les variables X déjà présentes dans le modèle.

Remarques

- (1) Les coefficients de détermination palliale peuvent prendre des valeurs comprises entre 0 et 1.
- (2) Un coefficient de détermination partielle peut être interprété comme un coefficient de détermination simple. Considérons un modèle de régression multiple avec deux variables prédictives. Supposons que nous régressions Y sur X_2 et obtenions les résidus :

$$e_i(Y | X_2) = Y_i - \hat{Y}_i(X_2)$$

3 Modèle de Régression Linéaire Multiple

où $\widehat{Y}_i(X_2)$ désigne les valeurs ajustées de Y lorsque X_2 est dans le modèle. Supposons que nous régressions davantage X_1 sur X_2 et obtenions les résidus :

$$e_i(X_1|X_2) = X_{i1} - \widehat{X}_{i1}(X_2)$$

où $\widehat{X}_{i1}(X_2)$ désigne les valeurs ajustées de X_1 dans la régression de X_1 sur X_2 . Le coefficient de détermination simple R^2 entre ces deux ensembles de résidus est égale au coefficient de détermination partielle $R^2_{Y1|2}$. Ainsi, ce coefficient mesure la relation entre Y et X_1 lorsque ces deux variables ont été ajustées pour leurs relations linéaires à X_2 .

3.7.7 Coefficients de Corrélation Partielle

La racine carrée d'un coefficient de détermination partielle s'appelle un coefficient de corrélation partielle. Il a le même signe que celui du coefficient de régression correspondant à la fonction de régression ajustée. Les coefficients de corrélation partielle sont fréquemment utilisés dans la pratique, bien qu'ils n'aient pas une signification aussi claire que les coefficients de détermination partielle. Les coefficients de corrélation partielle sont notamment utilisés dans les routines informatiques pour trouver la meilleure variable prédictive à sélectionner ensuite pour l'inclure dans le modèle de régression.

Remarque

Les coefficients de détermination partielle peuvent être exprimés en termes de coefficients de corrélation simples ou autres. Par exemple :

$$R^2_{Y2|1} = (r_{Y2|1})^2 = \frac{(r_{Y2} - r_{12}r_{Y1})^2}{(1 - r_{12}^2)(1 - r_{Y2}^2)} \quad (3.74)$$

$$R^2_{Y2|13} = (r_{Y2|13})^2 = \frac{(r_{Y2|3} - r_{12|3}r_{Y1|3})^2}{(1 - r_{12|3}^2)(1 - r_{Y1|3}^2)} \quad (3.75)$$

où r_{Y1} désigne le coefficient de corrélation simple entre Y et X_1 , r_{12} désigne le coefficient de corrélation simple entre X_1 et X_2 , et ainsi de suite. $r_{Y2|13}$ désigne le coefficient de corrélation entre Y et X_2 en maintenant constants X_1 et X_3 , et ainsi de suite.

3.7.8 Interprétation des Coefficients de Corrélation Simples et Partiels

Dans le cas d'une seule variable explicative, le coefficient de corrélation simple r avait un sens simple : il mesurait le degré d'association (linéaire) (et non de causalité) entre la variable dépendante Y et la variable explicative unique X . Mais une fois, nous allons au-delà de la variable à deux Dans ce cas, nous devons faire très attention à l'interprétation du coefficient de corrélation simple. À partir des équations (3.74) et (3.75), par exemple, on déduit que :

- (1) Même si $r_{Y1} = 0$, $r_{Y1|2}$ ne sera pas nul sauf si r_{Y2} ou r_{12} ou les deux sont égaux à zéro

$$\text{Rappel : } r_{Y1|2} = \frac{r_{Y1} - r_{21}r_{Y2}}{\sqrt{(1 - r_{21}^2)}\sqrt{(1 - r_{Y1}^2)}}$$

- (2) Si $r_{Y1} = 0$ et que r_{Y2} et r_{12} sont non nuls et ont le même signe, $r_{Y1|2}$ sera négatif, alors que s'ils sont de signes opposés, $r_{Y1|2}$ sera positif. Un exemple clarifiera ce point. Soit Y = rendement agricole, X_1 = précipitations et X_2 = température. Supposons que $r_{Y1} = 0$, c'est-à-dire qu'il n'y a aucune association entre le rendement agricole et les précipitations. Supposons en outre que r_{Y2} est positif et que r_{12} est négatif. Ensuite, $r_{Y1|2}$ sera positif ; c'est-à-dire, en maintenant la température constante, il existe une association positive entre le rendement et les précipitations. Ce résultat apparemment paradoxal n'est cependant pas surprenant. Étant donné que la température X_2 affecte à la fois le rendement Y et les précipitations X_1 , afin de déterminer le rapport net entre le rendement des cultures et les précipitations, nous devons éliminer l'influence de la «nuisance» variable de température. Cet exemple montre comment on peut être induit en erreur par le simple coefficient de corrélation.

- (3) Les termes $r_{Y1|2}$ et r_{Y1} (et des comparaisons similaires) ne doivent pas nécessairement avoir le même signe.
(4) Dans le cas d'une seule variable explicative, nous avons vu que r^2 entre X et Y est compris entre 0 et 1. La même propriété est vraie pour les coefficients de corrélation partielle au carré. En utilisant ce fait, le lecteur devrait vérifier que l'on peut obtenir l'expression suivante :

$$0 \leq r_{Y1}^2 + r_{Y2}^2 + r_{12}^2 - 2r_{Y1}r_{Y2}r_{12} \leq 1$$

qui donne les interrelations entre les trois coefficients de corrélation

3 Modèle de Régression Linéaire Multiple

d'ordre zéro. Des expressions similaires peuvent être dérivées.

- (5) Supposons que $r_{Y2} = r_{12} = 0$. Est-ce que cela signifie que r_{Y1} est également égal à zéro ?

La réponse est évidente à partir de l'équation en (4). Le fait que Y, X_2, X_1 et X_2 ne soient pas corrélés ne signifie pas que Y et X_1 ne sont pas corrélés.

L'expression $r_{Y1|2}^2$ peut être appelée *coefficient de détermination partielle* et peut être interprétée comme la proportion de la variation de Y non expliquée par la variable X_2 expliquée par l'inclusion de X_1 dans le modèle. Conceptuellement, il est similaire à R^2 .

Avant de poursuivre, notez les relations suivantes entre R^2 , le coefficient de corrélation simple et les coefficients de corrélation partielle :

$$R^2 = \frac{r_{Y1}^2 + r_{Y2}^2 - 2r_{Y1}r_{Y2}r_{12}}{1 - r_{12}^2}$$

$$R^2 = r_{Y1}^2 + (1 - r_{Y1}^2)r_{Y2|1}^2$$

$$R^2 = r_{Y2}^2 + (1 - r_{Y2}^2)r_{Y1|2}^2 \quad \text{ou} \quad (1 - R^2) = (1 - r_{Y2|1}^2)(1 - r_{Y1}^2)$$

Pour conclure cette section, considérons ce qui suit : Il a été dit précédemment que R^2 ne diminuerait pas si une variable explicative supplémentaire était introduite dans le modèle, ce qui peut être clairement vu à partir de la deuxième équation précédente de R^2 . Cette équation indique que la proportion de la variation de Y expliquée conjointement par X_1 et X_2 est la somme de deux parties : la partie expliquée par X_1 seule ($= r_{Y1}^2$) et la partie non expliquée par X_1 ($= 1 - r_{Y1}^2$) multipliée par la proportion expliquée par X_2 après avoir maintenu l'influence de X_1 constante. Maintenant $R^2 > r_{Y1}^2$ tant que $r_{Y2|1}^2 > 0$. Dans le pire des cas, $r_{Y2|1}^2$ sera égal à zéro, auquel cas $R^2 = r_{Y1}^2$.

L'équation précédente peut être généralisée pour un modèle à trois variables explicatives comme :

$$(1 - R^2) = (1 - r_{Y1}^2)(1 - r_{Y2|1}^2)(1 - r_{Y3|12}^2)$$

Pour un modèle à quatre variables explicatives on aura⁴ :

$$(1 - R^2) = (1 - r_{Y1}^2)(1 - r_{Y2|1}^2)(1 - r_{Y3|12}^2)(1 - r_{Y4|123}^2)$$

4. On note ici que les indices peuvent permutoer, pour un ordre des variables 4, 2, 1, 3, on aura $(1 - R^2) = (1 - r_{Y4}^2)(1 - r_{Y2|4}^2)(1 - r_{Y1|42}^2)(1 - r_{Y3|421}^2)$

3 Modèle de Régression Linéaire Multiple

Remarque

Dans un modèle à m variables explicatives, il existe une relation entre le coefficient de corrélation partielle et le t empirique de Student :

$$r_{Y_i|\text{autres variables}}^2 = \frac{t_i^2}{t_i^2 + (n - m - 1)}$$

Il est à noter que cette relation n'est vérifiée que pour un coefficient de corrélation partielle d'ordre $m - 1$.

3.8 Tests de Stabilité

Lorsque nous estimons une équation de régression multiple et l'utilisons pour des prédictions à des moments futurs, nous supposons que les paramètres sont constants sur toute la période d'estimation et de prédiction. Pour tester cette hypothèse de constance (ou de stabilité) des paramètres, des tests ont été proposés. Ces tests peuvent être décrits comme suit :

- (1) Tests d'Analyse de Variance
- (2) Tests Prédicteurs de Stabilité

3.8.1 Tests d'Analyse de Variance

Supposons que nous ayons deux ensembles de données indépendants avec des tailles d'échantillon (ou périodes si on parlait d'un modèle en série chronologique) n_1 et n_2 , respectivement. L'équation de régression est :

$$Y_1 = \beta_{01} + \beta_{11}X_{11} + \beta_{12}X_{12} + \dots + \beta_{1m}X_{1,m} + \varepsilon_1 \quad (\text{le 1er ensemble})$$

$$Y_2 = \beta_{02} + \beta_{21}X_{21} + \beta_{22}X_{22} + \dots + \beta_{2m}X_{2,m} + \varepsilon_2 \quad (\text{le 2ème ensemble})$$

Pour les β 's, le premier indice désigne l'ensemble de données et le second indice désigne l'ordre de la variable explicative. Un test de stabilité des paramètres entre les populations ayant généré les deux ensembles de données est un test de l'hypothèse :

$$H_0 : \beta_{11} = \beta_{21}, \beta_{12} = \beta_{22}, \dots, \beta_{1m} = \beta_{2m}, \beta_{01} = \beta_{02}$$

Si l'hypothèse H_0 est vraie, nous pouvons estimer une seule équation pour l'ensemble de données obtenu en regroupant les deux ensembles de données, c.-à-d. :

$$Y_1 = \beta_{01} + \beta_{11}X_{11} + \beta_{12}X_{12} + \dots + \beta_{1m}X_{1,m} + \varepsilon_1$$

3 Modèle de Régression Linéaire Multiple

Le test de Fisher que nous utilisons est le test Fisher décrit à la section des test de signification basé sur la somme des carrés des résidus. Pour obtenir somme des carrés des résidus non restreinte, nous estimons séparément le modèle de régression pour chacun des ensembles de données. soient :

SCR_1 : somme des carrés des résidus pour le premier ensemble.

SCR_2 : somme des carrés des résidus pour le deuxième ensemble.

SCR_1/σ^2 : Suit une distribution χ^2 à $(n_1 - m - 1)$ ddl.

SCR_2/σ^2 : Suit une distribution χ^2 à $(n_2 - m - 1)$ ddl.

Puisque les deux ensembles de données sont indépendants, alors $(SCR_1 + SCR_2)/\sigma^2$ suit une distribution χ^2 à $(n_1 + n_2 - 2m - 2)$ ddl. On notera $(SCR_1 + SCR_2)$ par $SCRN$ somme des carrés des résidus non restreinte. La somme des carrés des résidus restreinte $SCRR$ est obtenue à partir de la régression avec les données regroupées. (Cela impose la restriction que les paramètres sont les mêmes). Ainsi $SCRR/\sigma^2$ suit une distribution χ^2 à $(n_1 + n_2) - m - 1$ ddl.

$$F^* = \frac{\frac{SCRR - SCRN}{m+1}}{\frac{SCRN}{n_1 + n_2 - 2m - 2}}$$

qui suit une distribution de Fisher avec les degrés de liberté $(m + 1)$ et $(n_1 + n_2 - 2m - 2)$.

3.8.2 Tests Prédictifs de Stabilité

Test de Changement Structurel : Test de Chow

Il s'agit ici de tester une forme différente de contrainte sur les paramètres. Le test de Chow permet de déterminer si les paramètres du modèle se sont modifiés au cours du temps. Lorsqu'il est appliqué sur des données temporelles, il est nécessaire de connaître a priori une date de rupture. Le test indique alors si les paramètres sont identiques avant et après la date de rupture. Sur des données en coupe transversale, c'est-à-dire des données individuelles, ce test est utilisé afin de déterminer si des groupes d'individus sont homogènes ou hétérogènes.

On suppose que les aléas sont homoscédastiques, c'est-à-dire qu'ils ont une variance identique pour chaque observation. Sur données temporelles, soient n_1 la date de rupture, β^1 le vecteur des paramètres du

3 Modèle de Régression Linéaire Multiple

modèle sur la période $1, \dots, n_1$ et β^2 le vecteur des paramètres sur la période $n_1 + 1, \dots, n$, on teste

$$\begin{aligned} H_0 : \quad & \beta^1 = \beta^2 = \beta \\ H_1 : \quad & \beta^1 \neq \beta^2 \end{aligned}$$

Sous l'hypothèse H_0 le modèle s'écrit

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \varepsilon_i = \beta^{1'} X_i + \varepsilon_i \quad i = 1, \dots, n$$

Sous l'hypothèse H_1 les paramètres sont différents selon la période considérée

$$Y_i = \beta_0^1 + \beta_1^1 X_{i1} + \beta_2^1 X_{i2} + \dots + \beta_m^1 X_{im} + \varepsilon_i = \beta^{1'} X_i + \varepsilon_i \quad i = 1, \dots, n_1$$

et

$$Y_i = \beta_0^2 + \beta_1^2 X_{i1} + \beta_2^2 X_{i2} + \dots + \beta_m^2 X_{im} + \varepsilon_i = \beta^{2'} X_i + \varepsilon_i \quad i = n_1 + 1, \dots, n$$

Le modèle est estimé par les MCO d'une part sur l'ensemble de la période (c.-à-d. sur $1, \dots, n$) et d'autre part sur les deux sous-périodes $1, \dots, n_1$ et $n_1 + 1, \dots, n$.

On note

- SCR la somme des carrés des résidus du modèle estimé avec les n observations (les degrés de liberté correspondant est égal à $n - m - 1$),
- SCR^1 la somme des carrés des résidus de l'estimation sur la période $1, \dots, n_1$ (les degrés de liberté correspondant est égal à $n_1 - m - 1$),
- SCR^2 la somme des carrés des résidus de l'estimation sur la période $n_1 + 1, \dots, n$ (les degrés de liberté correspondant est égal à $n - n_1 - m - 1$)
- $SCR^1 + SCR^2$ la somme de SCR^1 et SCR^2 (les degrés de liberté correspondant est égal à $n - 2m - 2$).

Ce test correspond à un test de m contraintes, sous l'hypothèse nulle

$$\begin{aligned} \beta_0^1 &= \beta_0^2 = \beta_0 \\ \beta_1^1 &= \beta_1^2 = \beta_1 \\ &\vdots \\ \beta_m^1 &= \beta_m^2 = \beta_m \end{aligned}$$

3 Modèle de Régression Linéaire Multiple

et la statistique de Fisher s'écrit

$$F^* = \frac{\frac{SCR - (SCR^1 + SCR^2)}{(n - m - 1) - (n - 2m - 2)}}{\frac{SCR^1 + SCR^2}{n - 2m - 2}} = \frac{\frac{SCR - (SCR^1 + SCR^2)}{m + 1}}{\frac{SCR^1 + SCR^2}{n - 2m - 2}}$$

Cette statistique est distribuée suivant une loi de Fisher à $m + 2$ et $n - 2m - 3$ degrés de liberté.

Si $F^* < F_{m+1; n-2m-2}^\alpha$ alors l'hypothèse H_0 n'est pas rejetée et on retient une estimation par les MCO avec l'ensemble des observations. Le test s'applique de manière identique sur données individuelles pour tester l'homogénéité du comportement de groupes d'individus.

Tests de Stabilité Temporelle Basés Sur les Résidus Récursifs

Le test précédent suppose la date de rupture n_1 connue a priori. Mais il existe également des tests de stabilité temporelle qui permettent de déterminer les dates de rupture. Ces tests sont basés sur l'utilisation des résidus récursifs. Dans le modèle de régression linéaire on suppose que les paramètres (β, σ^2) ne varient pas au cours du temps

$$Y_i = X'_i \beta + \varepsilon_i \quad i = 1, \dots, n \\ \text{var}(\varepsilon_i) = \sigma^2$$

on suppose

$$\beta^1 = \dots = \beta^t = \dots = \beta^n = \beta \\ \sigma^{2_1} = \dots = \sigma^{2_t} = \dots = \sigma^{2_n} = \sigma^2$$

avec (β^t, σ^{2_t}) les paramètres du modèle à la période t . Les tests de stabilité temporelle développés par Brown, Durbin et Evans (1975) testent l'hypothèse nulle que le vecteur β est identique à chaque période conditionnellement à l'hypothèse $\sigma^{2_1} = \dots = \sigma^{2_t} = \dots = \sigma^{2_n} = \sigma^2$

$$H_0 : \beta^1 = \dots = \beta^t = \dots = \beta^n = \beta \\ H_1 : \beta^t \neq \beta \quad \text{ou} \quad \sigma^{2_t} \neq \sigma^2$$

Il existe notamment deux tests très souvent utilisés, le test du CUSUM et le test du CUSUM of Square. Ces deux tests utilisent les résidus récursifs.

3 Modèle de Régression Linéaire Multiple

Les Résidus Récursifs

Pour calculer les résidus récursifs du modèle, on doit estimer par la méthode des MCO les paramètres β avec un nombre d'observations qui varie de m à n . Soient

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_r \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1m} \\ 1 & X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \vdots & & \\ 1 & X_{r1} & X_{r2} & \cdots & X_{rm} \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_r \end{pmatrix}$$

avec r le nombre d'observations, $r = m + 1, \dots, n$. Soit e_r , l'erreur de prévision ex-post sur l'observation Y_r

$$e_r = Y_r - \hat{Y}_r = Y_r - X'_r \hat{\beta}^{r-1}$$

avec $\hat{\beta}^{m-1}$ l'estimateur des MCO lorsque la taille de l'échantillon est égale à $r-1$.

Sous l'hypothèse de stabilité $Y_r = X'_r + \varepsilon_r$ l'erreur de prévision e_r s'écrit

$$\begin{aligned} e_r &= X'_r \beta + \varepsilon_r - X'_r \hat{\beta}^{r-1} \\ &= \varepsilon_r - X'_r (\hat{\beta}^{r-1} - \beta) \end{aligned}$$

Or l'espérance et la matrice de variance-covariance de $\hat{\beta}^{r-1}$ sont données par

$$\begin{aligned} E(\hat{\beta}^{r-1}) &= \beta \\ var(\hat{\beta}^{r-1}) &= E[(\hat{\beta}^{r-1} - \beta)(\hat{\beta}^{r-1} - \beta)'] = \sigma^2 (X'_{r-1} X_{r-1})^{-1} \end{aligned}$$

d'où

$$\begin{aligned} E(e_r) &= 0 \\ var(e_r) &= \sigma^2 \left(1 + X'_r (X'_{r-1} X_{r-1})^{-1} X_r \right) \end{aligned}$$

Les résidus récursifs notés w_r sont donnés par

$$w_r = \frac{e_r}{\sqrt{1 + X'_r (X'_{r-1} X_{r-1})^{-1} X_r}}$$

Sous l'hypothèse de stabilité, les résidus récursifs sont distribués selon une loi normale

$$w_r \sim N(0, \sigma^2)$$

3 Modèle de Régression Linéaire Multiple

Test du CUSUM

Le test du CUSUM consiste à calculer la série cumulée des w_r

$$W_r = \sum_{j=m+1}^r \frac{w_j}{\hat{\sigma}} \quad r = m+1, \dots, n$$

avec

$$\hat{\sigma}^2 = \frac{\sum_{r=m+1}^n (w_r - \bar{w})^2}{n - m - 1} \quad \text{et} \quad \bar{w} = \frac{\sum_{r=m+1}^n w_r}{n - m}$$

Sous l'hypothèse de stabilité, W_r a une moyenne nulle et la région de confiance du test est donnée par

$$\Pr[-C_\alpha \leq W_r \leq C_\alpha] = 1 - \alpha$$

avec $C_\alpha = a(n-m)^{1/2} + 2a(r-m)(n-m)^{-1/2}$. La valeur de a dépend du risque de première espèce α du test. Pour $\alpha = 0.01, 0.05, 0.10$, a est égal à 1.143, 0.948, 0.850 respectivement.

Afin de faciliter la lecture des résultats du test effectué pour chaque statistique $W_r (r = m+1, \dots, n)$, on représente graphiquement la série W_r et la région de confiance donnée par l'intervalle $[-C_\alpha, C_\alpha]$.

Pour $r = m$, $C_\alpha = a(n-m)^{1/2}$ et pour $r = n$, $C_\alpha = 3a(n-m)^{1/2}$. La région de confiance du test pour les points $r = m+1, \dots, n-1$ est représentée par la droite reliant ces deux points.

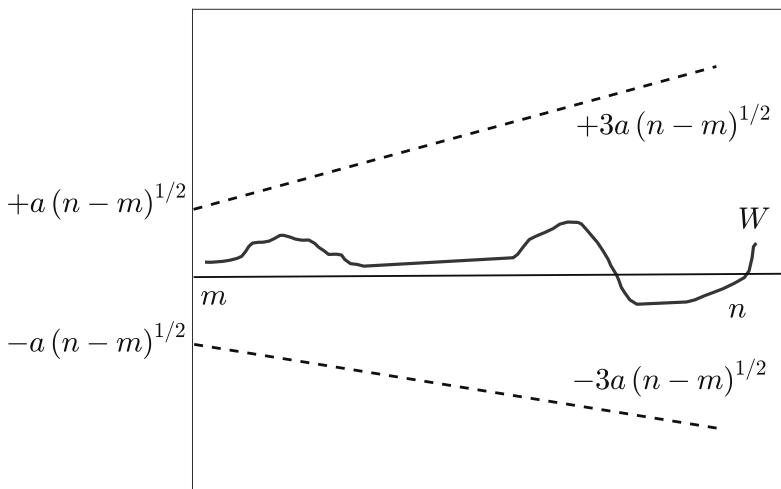


FIGURE 3.1 – Test du CUSUM

3 Modèle de Régression Linéaire Multiple

Dans la figure (3.1), l'espace entre les deux droites donne la région de confiance du test du CUSUM. Si W_r reste dans cet intervalle quel que soit $r = m + 1, \dots, n$, comme sur le graphique, alors l'hypothèse de stabilité est retenue. Sinon, si à une date donnée, W_r va au delà d'une des droites, alors on considère qu'à cette date il y a une rupture et que les paramètres ne sont pas stables.

Test du CUSUM of Square

Le test du CUSUM of Square est similaire au précédent. Il consiste à représenter graphiquement la série S_r

$$S_r = \frac{\sum_{j=m+1}^r w_j^2}{\sum_{j=m+1}^n w_j^2} \quad r = m + 1, \dots, n$$

On peut montrer que sous l'hypothèse de stabilité, S_r a une distribution Bêta d'espérance

$$E(S_r) = \frac{r - m}{n - m}$$

Comme dans le test du CUSUM, la région critique du test peut être représentée graphiquement comme dans la figure (3.2). Les deux bornes sont données par $E(S_r) \pm c_0$ où c_0 ⁵ est déterminé pour un niveau de significativité en fonction de m et n .

3.8.3 Test de Spécification de Ramsey (*Ramsey's RESET Test*)

Ramsey a proposé un test général d'erreur de spécification appelé RESET (Regression Specification Error Test)⁶. Ce dernier porte sur la pertinence de la forme fonctionnelle du modèle, telle que :

- l'omission d'une variable explicative dans le modèle ;
- la corrélation entre la variable explicative et le terme d'erreur ;
- une relation fonctionnelle non adaptée (passage aux logarithmes, fonctions inverses...) entre la variable à expliquer et les variables explicatives..etc.

5. Les valeurs de c_0 sont données dans la table de Durbin et peuvent être trouvées dans Harvey (1999), *The Econometric Analysis of Time Series*, 2nd edition, Cambridge, Mass., MIT Press.

6. **J. B. Ramsey**, “*Tests for Specification Errors in Classical Linear Least Squares Regression Analysis*,” Journal of the Royal Statistical Society, series B, vol. 31, 1969, pp. 350–371

3 Modèle de Régression Linéaire Multiple

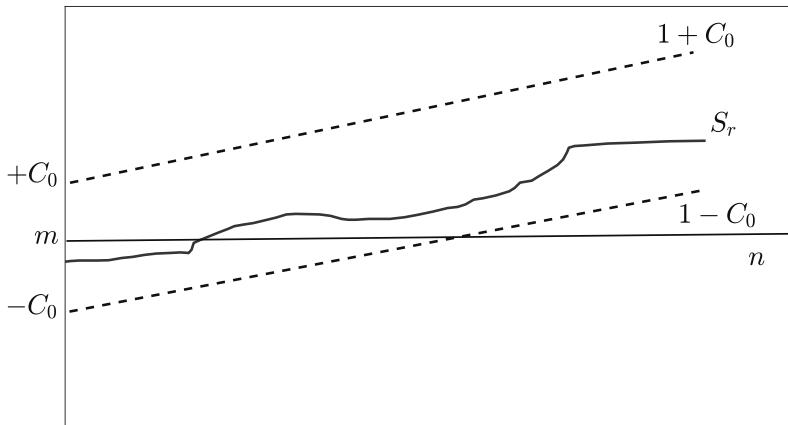


FIGURE 3.2 – **Test du CUSUM of Square**

Nous n’illustrerons ici la version la plus simple du test. Pour mieux comprendre le test, nous poursuivons avec un exemple explicatif.

Exemple :

Supposons un *modèle linéaire* qui lie la varibale Y = coût total et X = output (nombre d’actiles produit par exemple).

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Maintenant, si nous traçons les résidus e_i obtenus à partir de cette régression contre \hat{Y}_i , nous obtenons par exemple l’image illustrée à la figure (3.3). Bien que $\sum_{i=1}^n e_i$ et $\sum_{i=1}^n e_i \hat{Y}_i$ soient nécessairement nuls, les résidus de cette figure montrent un schéma dans lequel leur moyenne change systématiquement avec \hat{Y}_i . Cela suggérerait que si nous introduisons \hat{Y}_i sous une forme ou une autre, comme des variables explicatives dans l’équation précédente, celà devrait augmenter R^2 . Et si l’augmentation de R^2 est statistiquement significative (c.-à-.d sur la base d’un test de Fisher), cela suggérerait que la fonction de coût linéaire précédente a été mal spécifiée. C’est essentiellement l’idée de RESET. Les étapes du test RESET sont les suivantes :

- (1) A partir du modèle choisi, (par exemple le modèle linéaire précédent), on obtient les \hat{Y}_i .
- (2) On reprend l’équation du modèle précédent en introduisant \hat{Y}_i sous une certaine forme, par exemple, en tant que régresseur(s) supplémentaire(s). La figure (3.3) montre qu’il existe une relation curvi-

3 Modèle de Régression Linéaire Multiple

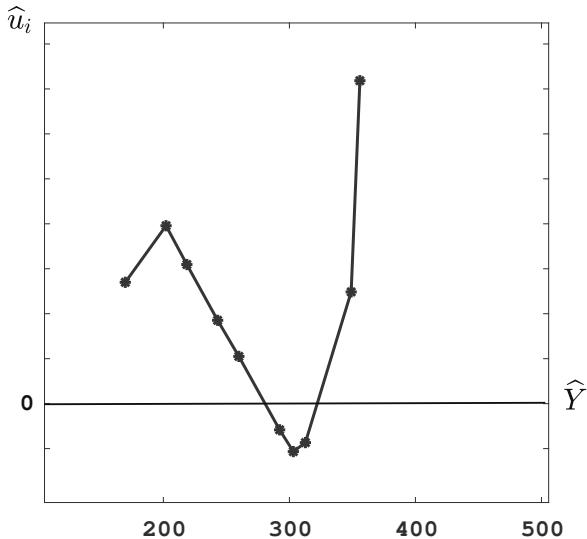


FIGURE 3.3 – **Résidus e_i et \hat{Y}_i à partir de la fonction de coût linéaire**

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

ligne entre e_i et \hat{Y}_i , ce qui suggère que l'on peut introduire \hat{Y}_i^2 et \hat{Y}_i^3 en tant que régresseurs supplémentaires. Ainsi, nous estimons le modèle

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \hat{Y}_i^2 + \varphi_3 \hat{Y}_i^3 + e_i$$

- (3) Soit $R_{nouveau}^2$ le coefficient de détermination obtenu à partir de la nouvelle équation de régression et R_{ancien}^2 celui obtenu à partir de l'équation initiale. Ensuite, nous pouvons utiliser le test de Fisher pour déterminer si l'augmentation de R^2 à l'aide de la nouvelle équation est statistiquement significatif.

$$F^* = \frac{\frac{R_{nouveau}^2 - R_{ancien}^2}{\text{nombre des nouveaux régresseurs}}}{\frac{1 - R_{nouveau}^2}{(n - \text{nombre des paramètres dans le nouveau modèle})}}$$

- (4) Si la valeur F^* calculée est significative, par exemple, au niveau de 5%, on peut accepter l'hypothèse selon laquelle le modèle initial est mal spécifié.

Un des avantages de RESET est qu'il est facile à appliquer car il n'est pas nécessaire de spécifier quel est le modèle alternatif. Mais c'est aussi son inconvénient, car savoir qu'un modèle est mal défini ne nous aide

3 Modèle de Régression Linéaire Multiple

pas nécessairement à choisir une meilleure alternative.

Remarque :

En pratique, le test RESET peut ne pas être particulièrement efficace pour détecter une alternative spécifique au modèle proposé et son utilité réside dans le fait qu'il sert d'indicateur général indiquant que quelque chose ne va pas. Pour cette raison, un test tel que RESET est parfois décrit comme un test d'une "spécification mauvaise", par opposition à un test de spécification. Cette distinction est plutôt subtile, mais l'idée de base est qu'un test de spécification examine un aspect particulier d'une équation donnée, en gardant à l'esprit les hypothèses nulle et alternative. Un test de "spécification mauvaise", en revanche, peut détecter une gamme de solutions de remplacement et indiquer que quelque chose ne va pas avec la sous l'hypothèse nulle, sans nécessairement indiquer clairement quelle hypothèse alternative est appropriée.

3.9 Estimation de la Réponse Moyenne et Prévision de Nouvelles Observations

3.9.1 Estimation d'Intervalle de $E(Y_h)$

Pour des valeurs données de X_1, \dots, X_m notées X_{h1}, \dots, X_{hm} , la réponse moyenne est notée $E(Y_h)$. Nous définissons le vecteur X_h :

$$X_h = \begin{pmatrix} 1 \\ X_{h1} \\ \vdots \\ X_{hm} \end{pmatrix}_{(m+1) \times 1} \quad (3.76)$$

de sorte que la réponse moyenne à estimer est :

$$E(Y_h) = X'_h \beta \quad (3.77)$$

La réponse moyenne estimée correspondante à X_h , notée \hat{Y}_h est :

$$\hat{Y}_h = X'_h \hat{\beta} \quad (3.78)$$

Cet estimateur est sans biais :

$$E(\hat{Y}_h) = X'_h \beta = E(Y_h) \quad (3.79)$$

et sa variance est :

$$\sigma^2(\hat{Y}_h) = \sigma^2 X'_h (X' X)^{-1} X_h \quad (3.80)$$

3 Modèle de Régression Linéaire Multiple

Cette variance peut être exprimée en fonction de la matrice de variance-covariance des coefficients de régression estimés :

$$\sigma^2(\hat{Y}_h) = X_h' \sigma^2(\hat{\beta}) X_h \quad (3.81)$$

Notons à partir de (3.81), la variance $\sigma^2(\hat{Y}_h)$ est fonction des variances $\sigma^2(\hat{\beta}_k)$ des coefficients de régression et des covariances $\sigma(\beta_k, \beta_{k'})$ entre des paires de coefficients de régression, comme dans la régression linéaire simple. La variance estimée $\hat{\sigma}^2(\hat{Y}_h)$ est donnée par :

$$\hat{\sigma}^2(\hat{Y}_h) = \hat{\sigma}^2 \times \left(X_h' (X' X)^{-1} X_h \right) = X_h' \hat{\sigma}^2(\hat{\beta}) X_h \quad (3.82)$$

Les limites de confiance à $(1 - \alpha)$ pour $E(Y_h)$ sont :

$$\hat{Y}_h \pm t_{n-m-1}^{\alpha/2} \hat{\sigma}(\hat{Y}_h) \quad (3.83)$$

3.9.2 Prévision de Nouvelle Observation $Y_{h(nouvelle)}$

Les limites de prédiction à $(1 - \alpha)$ pour une nouvelle observation $Y_{h(nouvelle)}$ correspondant à X_h , les valeurs spécifiées des variables X sont :

$$\hat{Y}_h \pm t_{n-m-1}^{\alpha/2} \hat{\sigma}(préd) \quad (3.84)$$

ou encore

$$\hat{Y}_h \pm t_{n-m-1}^{\alpha/2} \sqrt{\hat{\sigma}^2 \times \left(1 + X_h' (X' X)^{-1} X_h \right)}$$

où

$$\hat{\sigma}^2(préd) = \hat{\sigma}^2 + \hat{\sigma}^2(\hat{Y}_h) = \hat{\sigma}^2 \times \left(1 + X_h' (X' X)^{-1} X_h \right) \quad (3.85)$$

3.9.3 Prédiction de la Moyenne de p Nouvelles Observations à X_h

Lorsque p nouvelles observations doivent être sélectionnées aux mêmes niveaux X_h et que leur moyenne $\bar{Y}_{h(nouvelle)}$ doit être prédite, les limites de prédiction à $(1 - \alpha)$ sont les suivantes :

$$\bar{Y}_h \pm t_{n-m-1}^{\alpha/2} \hat{\sigma}(moypréd) \quad (3.86)$$

3 Modèle de Régression Linéaire Multiple

où

$$\hat{\sigma}^2(\text{moypréd}) = \frac{\hat{\sigma}^2}{p} + \hat{\sigma}^2(\hat{Y}_h) = \hat{\sigma}^2\left(\frac{1}{p} + X_h' (X'X)^{-1} X_h\right) \quad (3.87)$$

3.9.4 Précautions Concernant des Extrapolations Cachées

Lors de l'estimation d'une réponse moyenne ou de la prévision d'une nouvelle observation dans une régression multiple, il convient de veiller tout particulièrement à ce que l'estimation ou la prévision ne sorte pas du cadre du modèle. Le danger, bien sûr, est que le modèle peut ne pas être approprié s'il est étendu en dehors de la région des observations. En régression multiple, il est particulièrement facile de perdre la trace de cette région puisque les niveaux de X_1, \dots, X_m définissent conjointement la région. Ainsi, on ne peut pas simplement regarder les plages de chaque variable prédictive.

Considérons la figure (3.4), où la région ombrée est la région des observations pour une application de régression multiple avec deux variables explicatives et le point entouré représente les valeurs (X_{h1}, X_{h2}) pour lesquelles une prédiction doit être effectuée. Le point encerclé se situe dans les plages des variables prédictives X_1 et X_2 individuellement, mais se trouve bien en dehors de la région commune des observations. Il est facile de repérer cette extrapolation quand il n'y a que deux variables prédictives, mais cela devient beaucoup plus difficile lorsque le nombre de variables prédictives est grand.

3.10 Variables Indicatrices (Dummy Variables)

De nombreuses variables explicatives sont de nature qualitative. Par exemple, le chef de famille peut être un homme ou une femme, blanc ou non, employé ou chômeur. Dans ce cas, on code ces variables comme " H " pour les hommes et " F " pour les femmes, ou changer cette variable qualitative en une variable quantitative appelée *FEMME* qui prend la valeur "0" pour le mâle et "1" pour la femelle. Cela pose évidemment la question : "pourquoi ne pas avoir une variable *HOMME* qui prend la valeur 1 pour l'homme et 0 pour la femme ?" En fait, la variable *HOMME* serait exactement $(1 - FEMME)$. En d'autres termes, le 0 et le 1 peuvent être considérés comme un commutateur, qui se déclenche quand il est 1 et éteint quand il est 0. Supposons que nous sommes intéressés par les gains des ménages, dénotés par *REVENU*, alors que *HOMME* et

3 Modèle de Régression Linéaire Multiple

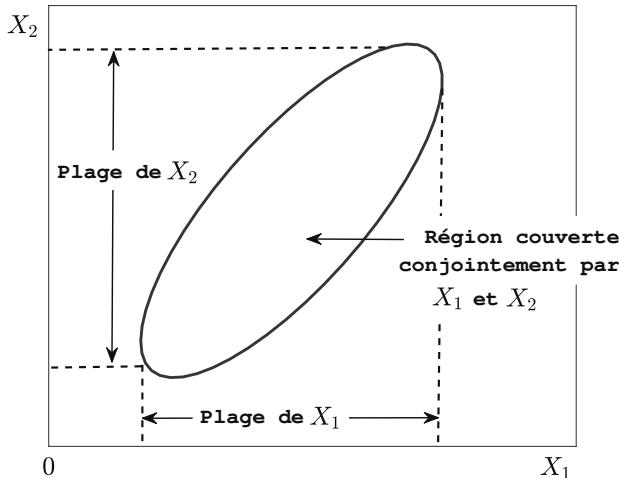


FIGURE 3.4 – **Région des observations couverte par X_1 et X_2 conjointement, comparée aux plages de X_1 et X_2 individuelles**

FEMME sont les seules variables explicatives disponibles :

$$REVENU = \beta_H HOMME + \beta_F FEMME + \varepsilon \quad (3.88)$$

donne $\hat{\beta}_H$ = “gains moyens des hommes dans l’échantillon” et $\hat{\beta}_F$ = “gains moyens des femelles dans l’échantillon”. Remarquez qu’il n’y a pas d’interception dans (3.88), c’est à cause de ce qu’on appelle dans la littérature “trappe à variables indicatrices”⁷ (Dummy variables trap). En résumé, il y aura une parfaite multicolinéarité entre *HOMME*, *FEMME* et la constante. En fait, $HOMME+FEMME = 1$. Certains chercheurs peuvent choisir d’inclure l’interception et d’exclure l’une des variables muettes du sexe, disons *HOMME*, alors

$$REVENU = \beta_0 + \beta FEMME + \varepsilon \quad (3.89)$$

et les estimations des MCO donnent $\hat{\beta}_0 = gains moyens des hommes dans l’échantillon = \hat{\beta}_H$, tandis que $\hat{\beta} = \hat{\beta}_F - \hat{\beta}_H = différence entre les gains moyens des femmes et des hommes dans l’échantillon$. La régression (3.89) est plus populaire lorsqu’on s’intéresse à la comparaison des gains entre hommes et femmes et à l’obtention d’une régression gains

7. Le piège des variables indicatrices est un scénario dans lequel les variables indépendantes sont multicolinéaires - un scénario dans lequel deux variables ou plus sont fortement corrélées ; en termes simples, une variable peut être prédite à partir des autres.

3 Modèle de Régression Linéaire Multiple

moyens $(\hat{\beta}_F - \hat{\beta}_H)$ ainsi que le test de savoir si cette différence est statistiquement différente de zéro. Ce serait simplement la statistique t de Student sur $\hat{\beta}$ dans (3.89). D'un autre côté, si l'on veut estimer séparément les gains moyens des hommes et des femmes, alors le modèle (3.88) devrait être celui à considérer. Dans ce cas, le test t pour $\hat{\beta}_F - \hat{\beta}_H = 0$ impliquerait d'autres calculs non directement donnés par la régression dans (3.88) mais similaires aux calculs donnés dans l'exemple 3 de la sous-section 3.7.5.

Que se passe-t-il lorsqu'une autre variable qualitative est incluse, pour représenter une autre classification des individus dans l'échantillon, disons, par exemple, la race ? S'il y a trois groupes de race dans l'échantillon, *BLANC*, *NOIR* et *HISPANIQUE*. On pourrait créer une variable fictive pour chacune de ces classifications. Par exemple, *BLANC* prendra la valeur 1 lorsque l'individu est blanc et 0 lorsque l'individu n'est pas blanc. Notez que la trappe à variables indicatrices n'autorise pas l'inclusion des trois catégories puisqu'elles se résument à 1. De plus, même si l'interception est supprimée, une fois *HOMME* et *FEMME* inclus, la multicolinéarité parfaite est toujours présente car $HOMME + FEMME = BLANC + NOIR + HISPANIQUE$. Par conséquent, une catégorie de la race devrait être abandonnée. Suits (1984)⁸ soutient que le chercheur devrait utiliser l'omission de la catégorie des variables indicatrices à son avantage, en interprétant les résultats et en gardant à l'esprit le but de l'étude. Par exemple, si on veut comparer les gains entre les sexes en gardant la race comme constante, l'omission de *HOMME* ou *FEMME* est naturelle, alors que si l'on s'intéresse à la différence de race dans les gains en maintenant le sexe, il convient d'omettre une de ses variables. Quelle que soit la variable omise, elle devient la catégorie de base pour laquelle les autres gains sont comparés. La plupart des chercheurs préfèrent garder une interception, bien que les packages de régression permettent une option sans interception. Dans ce cas, il faut omettre une catégorie de chacune des classifications de race et de sexe. Par exemple, si *HOMME* et *BLANC* sont omis :

$$REVENU = \beta_0 + \beta_F FEMME + \beta_N NOIR + \beta_S HISPANIQUE + \varepsilon \quad (3.90)$$

En supposant que l'erreur u vérifie toutes les hypothèses classiques, et en prenant les valeurs espérées des deux côtés de (3.90), on peut voir que $\beta_0 = la\ valeur\ espérée\ des\ gains\ de\ la\ catégorie\ omise\ qui\ est\ "hommes\ blancs"$. Pour cette catégorie, tous les autres commutateurs sont désacti-

8. Suits, D. (1984), “*Dummy Variables : Mechanics vs Interpretation*”, Review of Economics and Statistics, 66 : 132–139.

3 Modèle de Régression Linéaire Multiple

vés. De même, $\beta_0 + \beta_F$ est la valeur espérée des gains des “femmes blanches”, puisque le commutateur *FEMALE* est activé. On peut conclure que $\beta_F = \text{différence dans la valeur espérée des gains entre les femmes blanches et les hommes blancs}$. De même, on peut montrer que $\beta_0 + \beta_N$ est la rémunération espérée des “hommes noirs” et $\beta_0 + \beta_F + \beta_N$ est le revenu espéré des “femmes noires”. Par conséquent, β_F représente la différence entre les gains espérés chez les femmes noires et les hommes noirs. En fait on peut montrer que β_F représente la différence dans les gains espérés entre les femmes hispaniques et les hommes hispaniques. En d’autres termes, β_F représente le différentiel des gains espérés entre les femmes et les hommes détenant une constante de race. De même, on peut montrer que β_N est la différence des gains espérés entre les noirs et les blancs en tenant une constante de sexe, et β_S est l’écart entre les gains espérés entre les hispaniques et les blancs ayant une constante sexuelle. La clé principale de l’interprétation des coefficients des variables indicatrices est de pouvoir d’activer et de désactiver les bons commutateurs et d’écrire les espérances correctes.

La régression réelle contiendra d’autres variables quantitatives et qualitatives, comme

$$\begin{aligned} REVENU &= \beta_0 + \beta_F FEMME + \beta_N NOIR + \beta_S HISPANIQUE \\ &\quad + \gamma_1 EXP + \gamma_2 EXP^2 + \gamma_3 EDUC + \gamma_4 SYND + \varepsilon \end{aligned} \tag{3.91}$$

où *EXP* sont les années d’expérience professionnelle, *EDUC* sont les années d’éducation, et *SYND* est égale à 1 si l’individu appartient à un syndicat et 0 autrement. EXP^2 est la valeur au carré de *EXP*. Encore une fois, on peut interpréter les coefficients de ces régressions en activant ou en désactivant les commutateurs appropriés. Par exemple, γ_4 est interprété comme la différence de revenus attendue entre les membres de syndicat et ceux qui ne le sont pas, tout en maintenant constantes toutes les autres variables incluses dans (3.91). Halvorsen et Palmquist (1980)⁹ ont mis l’accent sur le sujet d’interprétation des coefficients des variables indicatrices lorsque la variable dépendante est en Log. Par exemple, si l’équation des gains est semi-logarithmique :

$$\log(Revenus) = \beta_0 + \beta SYND + \gamma EDUC + \varepsilon \tag{3.92}$$

puis $\gamma = \%$ de variation des revenus pour une année d’éducation supplémentaire.

9. **Halvorsen, R. and R. Palmquist (1980)**, “The Interpretation of Dummy Variables in Semilogarithmic Equations” American Economic Review, 70 : 474–475.

3 Modèle de Régression Linéaire Multiple

mentaire, en maintenant l'affiliation syndicale constante. Mais, qu'en est-il des revenus pour l'adhésion syndicale ? Si on laisse $Y_1 = \log(Revenus)$ quand l'individu appartient à un syndicat, et $Y_0 = \log(Revenus)$ quand l'individu n'appartient pas, alors $g = \% \text{ des revenus dû à l'adhésion syndicale} = (e^{Y_1} - e^{Y_0}) / e^{Y_0}$. De manière équivalente, on peut écrire que $\log(1 + g) = Y_1 - Y_0 = \beta$, ou $g = e^\beta - 1$. En d'autres termes, il ne faut pas se précipiter pour conclure que β a la même interprétation que γ . En fait, la variation en % des revenus due à l'affiliation syndicale est $e^\beta - 1$ et non β . L'erreur impliquée dans l'utilisation de $\hat{\beta}$ plutôt que $e^\beta - 1$ pour estimer g pourrait être substantielle, en particulier si $\hat{\beta}$ est grand. Par exemple, lorsque $\hat{\beta} = 0.5, 0.75, 1$; $\hat{g} = e^{\hat{\beta}} - 1 = 0.65, 1.12, 1.72$, respectivement. Kennedy (1981) note que si $\hat{\beta}$ est un estimateur non biaisé de β , \hat{g} n'est pas nécessairement non biaisé pour g . Cependant, la consistance de $\hat{\beta}$ implique la consistance de \hat{g} . Si l'on suppose des erreurs distribuées log-normales, alors $E(e^{\hat{\beta}}) = e^{\beta+0.5\text{Var}(\hat{\beta})}$. Sur la base de ce résultat, Kennedy (1981) suggère d'estimer g par $\tilde{g} = e^{\hat{\beta}+0.5\widehat{\text{var}}(\hat{\beta})}$, où $\widehat{\text{var}}(\hat{\beta})$ est une estimation consistante de $\text{var}(\hat{\beta})$.

Une autre utilisation des variables indicatrices consiste à prendre en compte des facteurs saisonniers, c'est-à-dire comprenant 3 variables indicatrices saisonnières, la saison omise devenant la base de comparaison¹⁰. Par exemple :

$$Ventes = \beta_0 + \beta_H Hiver + \beta_P Printemps + \beta_A Automne + \gamma_1 Prix + \varepsilon \quad (3.93)$$

la saison omise étant la saison d'été, et si (3.93) modélise les ventes d'unités de climatisation, alors β_A est la différence des ventes attendues entre les saisons d'automne et d'été, maintenant le prix d'une unité de climatisation constant. Si elles sont des unités de chauffage, on pourra changer la saison de base pour la comparaison. Une autre utilisation des variables indicatrices est pour les années de guerre, où la consommation n'est pas à son niveau normal, en raison du rationnement. Envisager d'estimer la fonction de consommation suivante

$$C_i = \beta Y_i + \delta GUERRE_i + \varepsilon_i \quad i = 1, \dots, n \quad (3.94)$$

où C_i indique la consommation réelle par habitant, Y_i indique le revenu personnel disponible réel par habitant, et $GUERRE_i$ est une variable

10. Il existe des moyens plus sophistiqués de désaisonnalisation que d'introduire des variables indicatrices saisonnières, voir Judge et al. (1985).

3 Modèle de Régression Linéaire Multiple

indicatrice prenant la valeur 1 si c'est une période de guerre et 0 sinon. Notez que les années de guerre n'affectent pas la pente de la ligne de consommation par rapport au revenu, seulement l'interception. L'ordonnée à l'origine est β_0 dans les années de non-guerre et $\beta_0 + \delta$ dans les années de guerre. En d'autres termes, la propension marginale du revenu est la même dans les années de guerre et de non-guerre, seul le niveau de consommation est différent.

Bien sûr, on peut simuler d'autres années inhabituelles comme des périodes de grève, des années de catastrophes naturelles, des tremblements de terre, des inondations, ou des chocs externes, par exemple l'année de crise financière 2008. Si cette indication ne comprend qu'une année comme 2008, alors la variable indicatrice pour 2008, appelons-la D08, prend la valeur 1 pour 2008 et zéro sinon. L'inclusion de D08 comme variable supplémentaire dans la régression a pour effet de supprimer l'observation de 2008 des objectifs d'estimation, et les estimations de coefficients de régression résultantes sont exactement les mêmes que celles obtenues en excluant l'observation de 2008 et sa variable indicatrice correspondante.

Effets d'Interaction

Jusqu'à présent, les variables indicatrices ont été utilisées pour décaler l'interception de la régression en maintenant les pentes constantes. On peut également utiliser les variables indicatrices pour déplacer les pentes en les laissant interagir avec les variables explicatives. Par exemple, considérons l'équation suivante des revenus :

$$REVENU = \beta_0 + \beta_F FEMME + \beta EDUC + \varepsilon \quad (3.95)$$

Dans cette régression, seule l'interception passe des hommes aux femmes. Le rendement d'une année supplémentaire d'éducation est simplement β , ce qui est supposé être le même pour les hommes que pour les femmes. Mais si nous introduisons maintenant la variable d'interaction ($FEMME \times EDUC$), alors la régression devient :

$$\begin{aligned} REVENU = & \beta_0 + \beta_F FEMME + \beta EDUC \\ & + \gamma (FEMME \times EDUC) + \varepsilon \end{aligned} \quad (3.96)$$

Dans ce cas, le rendement d'une année supplémentaire d'éducation dépend du sexe de l'individu.

En fait, $\frac{\partial (REVENU)}{\beta (EDUC)} = \beta + \gamma (FEMME) = \beta$ si masculin et $\beta + \gamma$

3 Modèle de Régression Linéaire Multiple

si féminin. Notez que la variable d'interaction = $EDUC$ si l'individu est une femme et 0 si l'individu est un homme.

L'estimation (3.96) équivaut l'estimation de deux équations de revenu, une pour les hommes et une autre pour les femmes, séparément. La seule différence est que (3.96) impose la même variance entre les deux groupes, alors que des régressions séparées n'imposent pas l'hypothèse d'égalité, même restrictive, de la variance. Cette configuration est idéale pour tester l'égalité des pentes, l'égalité des interceptions, ou l'égalité des deux interceptions et des pentes entre les sexes. Cela peut être fait avec le test de Fisher. En fait, pour l'hypothèse " H_0 ; égalité des pentes", étant donné des interceptions différentes, la somme des carrés des résidus du modèle réduit (restreint) est obtenue à partir de (3.95), tandis que la somme des carrés des résidus du modèle complet (non restreint) est obtenue à partir de (3.96).

3.11 Exemple Empirique

RBAG Studios exploite des studios de portrait dans 21 villes de taille moyenne. Ces studios sont spécialisés dans les portraits d'enfants. La société envisage une expansion dans d'autres villes de taille moyenne et souhaite déterminer si les ventes (Y) dans une communauté peuvent être prédictes à partir du nombre de personnes âgées de 16 ans ou moins dans la communauté (X_1) et du revenu personnel disponible par habitant dans la communauté (X_2). Le tableau (3.2) présente des données sur ces variables pour l'année la plus récente pour les 21 villes dans lesquelles RBAG Studios est désormais présente. Les ventes sont exprimées en milliers de dirhams et sont étiquetées Y ; le nombre de personnes âgées de 16 ans ou moins est exprimé en milliers de personnes et est étiqueté X_1 pour la population cible; et le revenu personnel disponible par habitant est exprimé en milliers de dirhams et libellé X_2 pour le revenu disponible.

Le modèle de régression de premier ordre :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad (3.97)$$

3 Modèle de Régression Linéaire Multiple

	X_1	X_2	Y	\hat{Y}	e
1	68.5	16.7	174.4	187.184	-12.7841
2	45.2	16.8	164.4	154.229	10.1706
3	91.3	18.2	244.2	234.396	9.8037
4	47.8	16.3	154.6	153.329	1.2715
5	46.9	17.3	181.6	161.385	20.2151
6	66.1	18.2	207.5	197.741	9.7586
7	49.5	15.9	152.8	152.055	0.7449
8	52	17.2	163.2	167.867	-4.6666
9	48.9	16.6	145.4	157.738	-12.3382
10	38.4	16	137.2	136.846	0.354
11	87.9	18.3	241.9	230.387	11.5126
12	72.8	17.1	191.1	197.185	-6.0649
13	88.4	17.4	232	222.686	9.3143
14	42.9	15.8	145.3	141.518	3.7816
15	52.5	17.8	161.1	174.213	-13.1132
16	85.7	18.4	209.7	228.124	-18.4239
17	41.3	16.5	146.4	145.747	0.653
18	51.7	16.3	144	159.001	-15.0013
19	89.6	18.1	232.6	230.987	1.613
20	82.7	19.1	224.1	230.316	-6.216
21	52.3	16	166.5	157.064	9.4356

TABLE 3.2 – **Données de base**

Calculs de Base

Les matrices X et Y de l'exemple RBAG Studios sont les suivantes :

$$X = \begin{pmatrix} 1 & 68.5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \vdots & \vdots & \vdots \\ 1 & 52.3 & 16.0 \end{pmatrix} \quad Y = \begin{pmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{pmatrix} \quad (3.98)$$

3 Modèle de Régression Linéaire Multiple

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-68.85707	60.01695	-1.147294	0.2663
X1	1.45456	0.211782	6.868201	0.0000
X2	9.3655	4.063958	2.304527	0.0333
R-squared	0.916746	Mean dependent var	181.9048	
Adjusted R-squared	0.907496	S.D. dependent var	36.1913	
Sum squared resid	2180.927	Mean squared resid	121.1626	
Sum squared regr	24015.282	Mean squared regr	12007.6411	

TABLE 3.3 – Résultat de la Régression (Logiciel Eviews)

Nous exigeons :

$$\begin{aligned}
 X'X &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 68.5 & 45.2 & \cdots & 52.3 \\ 16.7 & 16.8 & \cdots & 16.0 \end{pmatrix} \begin{pmatrix} 1 & 68.5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \vdots & \vdots & \vdots \\ 1 & 52.3 & 16.0 \end{pmatrix} \\
 &= \begin{pmatrix} 21.0 & 1302.4 & 360.0 \\ 1302.4 & 87707.9 & 22609.2 \\ 360.0 & 22609.2 & 6190.3 \end{pmatrix} \\
 X'Y &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 68.5 & 45.2 & \cdots & 52.3 \\ 16.7 & 16.8 & \cdots & 16.0 \end{pmatrix} \begin{pmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{pmatrix} = \begin{pmatrix} 3820 \\ 249643 \\ 66073 \end{pmatrix} \\
 (X'X)^{-1} &= \begin{pmatrix} 29.7289 & 0.0722 & -1.9926 \\ 0.0722 & 0.00037 & -0.0056 \\ -1.9926 & -0.0056 & 0.1363 \end{pmatrix}
 \end{aligned}$$

Fonction de régression estimée

Les estimations des moindres carrés $\hat{\beta}$ sont facilement obtenues à l'aide de nos calculs de base :

$$\hat{\beta} = (X'X)^{-1} X'Y = \begin{pmatrix} 29.7289 & 0.0722 & -1.9926 \\ 0.0722 & 0.00037 & -0.0056 \\ -1.9926 & -0.0056 & 0.1363 \end{pmatrix} \begin{pmatrix} 3820 \\ 249643 \\ 66073 \end{pmatrix}$$

3 Modèle de Régression Linéaire Multiple

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} -68.857 \\ 1.455 \\ 9.366 \end{pmatrix}$$

et la fonction de régression estimée :

$$\hat{Y} = -68.857 + 1.455X_1 + 9.366X_2$$

Cette fonction de régression estimée indique que les ventes moyennes devraient augmenter de 1.455 mille dirhams lorsque la population cible augmentera de 1000 personnes âgées de 16 ans ou moins, à revenu personnel disponible par habitant constant, et que les ventes moyennes devraient augmenter de 9.366 mille dirhams lorsque le revenu par habitant augmente de 1000 dirhams, en maintenant la population cible constante. Le tableau (3.3) contient la sortie de régression multiple par le logiciel Eviews pour l'exemple de RBAG Studios. Les coefficients de régression estimés sont indiqués dans la colonne intitulée COEFFICIENT.

Valeurs Estimées et Résidus

Pour examiner la pertinence du modèle de régression pour les données disponibles, nous avons besoin des valeurs ajustées \hat{Y}_i et des résidus $e_i = Y_i - \hat{Y}_i$. On a :

$$\hat{Y} = X\hat{\beta}$$

$$\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_{21} \end{pmatrix} = \begin{pmatrix} 1 & 68.5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \vdots & \vdots & \vdots \\ 1 & 52.3 & 16.0 \end{pmatrix} \begin{pmatrix} -68.857 \\ 1.455 \\ 9.366 \end{pmatrix} = \begin{pmatrix} 187.2 \\ 154.2 \\ \vdots \\ 157.1 \end{pmatrix}$$

De plus on a :

$$e = Y - \hat{Y}$$

$$\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{21} \end{pmatrix} = \begin{pmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{pmatrix} - \begin{pmatrix} 187.2 \\ 154.2 \\ \vdots \\ 157.1 \end{pmatrix} = \begin{pmatrix} -12.8 \\ 10.2 \\ \vdots \\ 9.4 \end{pmatrix}$$

Analyse de la Variance

$$Y'Y = \begin{pmatrix} 174.4 & 164.4 & \cdots & 166.5 \end{pmatrix} \begin{pmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{pmatrix} = 721072.40$$

3 Modèle de Régression Linéaire Multiple

$$\left(\frac{1}{n}\right) Y' JY = \frac{1}{21} \begin{pmatrix} 174.4 & 164.4 & \dots & 166.5 \end{pmatrix} \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{pmatrix}$$

$$= \frac{(3820)^2}{21} = 694876.19$$

ainsi

$$SCT = Y' Y - \left(\frac{1}{n}\right) Y' JY = 721072.40 - 694876.19 = 26196.21$$

aussi

$$SCR = Y' Y - \hat{\beta}' X' Y$$

$$= 721072.40 - \begin{pmatrix} 3820 \\ -68.857 & 1.455 & 9.366 \end{pmatrix} \begin{pmatrix} 249643 \\ 66073 \end{pmatrix}$$

$$= 721072.40 - 718891.47 = 2180.93$$

Enfin, on déduit :

$$SCE = SCT - SCR = 26196.21 - 2180.93 = 24015.28$$

Test de la Relation de Régression

Pour vérifier si les ventes sont liées à la population cible et au revenu disponible par habitant :

$$H_0 : \beta_1 = \beta_2 = 0$$

H_1 : Il existe au moins un de ces coefficients non nul

nous utilisons des statistiques empirique de Fisher

$$F^* = \frac{MCE}{MCR} = \frac{12007.64}{121.1626} = 99.1$$

Pour $\alpha = 0.05$, nous avons $F^* = 99.1 > F_{2;18}^{0,05} = 3.55$ (on deux degrés de liberté 2 et 18). Nous concluons l'hypothèse H_1 postulant que les ventes sont liées à la population cible et au revenu disponible par habitant. La valeur P pour ce test est (0.0000), comme indiqué dans la sortie de Eviews (nommée Prob.).

Coefficient de Détermination Multiple

3 Modèle de Régression Linéaire Multiple

Pour notre exemple on a

$$R^2 = \frac{SCE}{SCT} = \frac{24015.28}{26196.21} = 0.917$$

Ainsi, lorsque les deux variables prédictives, population cible et revenu disponible par habitant, sont prises en compte, la variation des ventes est réduite de 91.7%. Le coefficient de détermination multiple est indiqué dans la sortie d'Eviews intitulée R-Squared. La sortie indique également le coefficient de détermination multiple ajusté, libellé dans la sortie Adjusted R-Squared. Notez que l'ajustement du nombre de variables prédictives dans le modèle n'a eu qu'un faible effet ici sur R^2 .

Estimation des Paramètres de Régression

RBAG Studios ne s'intéresse pas au paramètre β_0 car il dépasse largement le cadre du modèle. Il est souhaitable d'estimer β_1 et β_2 conjointement avec le coefficient de confiance 0.90. Nous utiliserons les limites de confiance simultanées.

Premièrement, nous avons besoin de la matrice de variance-covariance estimée $\hat{\sigma}^2(\hat{\beta})$:

$$\hat{\sigma}^2(\hat{\beta}) = MCR(X'X)^{-1} = \hat{\sigma}^2 \times (X'X)^{-1}$$

$$\begin{aligned}\hat{\sigma}^2(\hat{\beta}) &= 121.1626 \begin{pmatrix} 29.7289 & 0.0722 & -1.9926 \\ 0.0722 & 0.00037 & -0.0056 \\ -1.9926 & -0.0056 & 0.1363 \end{pmatrix} \\ &= \begin{pmatrix} 3602 & 8.748 & -241.43 \\ 8.748 & 0.0448 & -0.679 \\ -241.43 & -0.679 & 16.514 \end{pmatrix}\end{aligned}$$

Les deux variances estimées dont nous avons besoin sont :

$$\hat{\sigma}^2(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}^2 = 0.0448 \quad \text{ou} \quad \hat{\sigma}(\hat{\beta}_1) = 0.212$$

$$\hat{\sigma}^2(\hat{\beta}_2) = \hat{\sigma}_{\hat{\beta}_2}^2 = 16.514 \quad \text{ou} \quad \hat{\sigma}(\hat{\beta}_2) = 4.06$$

Pour un seuil de confiance $\alpha = 0.05$, les deux intervalles de confiance

3 Modèle de Régression Linéaire Multiple

simultanés sont donc ($t_{18}^{0.025} = 2.101$) :

$$\hat{\beta}_1 \pm t_{18}^{0.025} \hat{\sigma}^2 (\hat{\beta}_1) \Rightarrow 1.455 \pm 2.101 (0.212) \Rightarrow 1.01 \leq \beta_1 \leq 1.90$$

$$\hat{\beta}_2 \pm t_{18}^{0.025} \hat{\sigma}^2 (\hat{\beta}_2) \Rightarrow 9.366 \pm 2.101 (4.06) \Rightarrow 0.84 \leq \beta_2 \leq 17.9$$

Il est à noter que les intervalles de confiance simultanés suggèrent que β_1 et β_2 sont tous deux positifs, ce qui est conforme aux attentes théoriques selon lesquelles les ventes devraient augmenter avec une population cible plus élevée et un revenu disponible par habitant plus élevé, l'autre variable étant maintenue constante.

Estimation de la Réponse Moyenne

RBAG Studios aimerait estimer avec un intervalle de confiance de 95% les ventes (moyennes) attendues dans les villes avec une population cible $X_{h1} = 65.4$ mille personnes âgées de 16 ans ou moins et un revenu disponible par habitant $X_{h2} = 17.6$ mille dirhams. Nous définissons :

$$X_h = \begin{pmatrix} 1 \\ 65.4 \\ 17.6 \end{pmatrix}$$

L'estimation ponctuelle des ventes moyennes est :

$$\hat{Y}_h = X'_h \hat{\beta} = \begin{pmatrix} 1 & 65.4 & 17.6 \end{pmatrix} \begin{pmatrix} -68.857 \\ 1.455 \\ 9.366 \end{pmatrix} = 191.10$$

La variance estimée est la suivante :

$$\begin{aligned} \hat{\sigma}^2 (\hat{Y}_h) &= X'_h \hat{\sigma}^2 (\hat{\beta}) X_h \\ &= \begin{pmatrix} 1 & 65.4 & 17.6 \end{pmatrix} \begin{pmatrix} 3602 & 8.748 & -241.43 \\ 8.748 & 0.0448 & -0.679 \\ -241.43 & -0.679 & 16.514 \end{pmatrix} \begin{pmatrix} 1 \\ 65.4 \\ 17.6 \end{pmatrix} \\ &= 7.656 \end{aligned}$$

ou

$$\hat{\sigma} (\hat{Y}_h) = 2.77$$

L'intervalle de confiance pour $E(Y_h)$ est donc :

$$\hat{Y}_h \pm t_{18}^{0.025} \hat{\sigma} (\hat{Y}_h) \Rightarrow 191.10 \pm 2.101 (2.77) \Rightarrow 185.3 \leq E(Y_h) \leq 196.9$$

3 Modèle de Régression Linéaire Multiple

Ainsi, avec un coefficient de confiance de 0.95, nous estimons que les ventes moyennes dans les villes dont la population cible est de 65.4 mille personnes âgées de 16 ans ou moins et dont le revenu disponible par habitant est de 17.6 mille dirhams se situent entre 185.3 et 196.9 mille dirhams. RBAG Studios considère que cet intervalle de confiance fournit des informations suffisamment précises sur les ventes (moyennes) attendues dans les communautés de cette taille et de ce niveau de revenu, aux fins de la planification.

Version Algébrique de la Variance Estimée $\hat{\sigma}^2(\hat{Y}_h)$

$$\hat{\sigma}^2(\hat{Y}_h) = X_h' \hat{\sigma}^2(\hat{\beta}) X_h \quad (3.99)$$

il en résulte pour le cas de deux variables prédictives dans un modèle de premier ordre :

$$\begin{aligned} \hat{\sigma}^2(\hat{Y}_h) &= \hat{\sigma}^2(\hat{\beta}_0) + X_{h1}^2 \hat{\sigma}^2(\hat{\beta}_1) + X_{h2}^2 \hat{\sigma}^2(\hat{\beta}_2) + 2X_{h1}\hat{\sigma}(\hat{\beta}_0, \hat{\beta}_1) \\ &\quad + 2X_{h2}\hat{\sigma}(\hat{\beta}_0, \hat{\beta}_2) + 2X_{h1}X_{h2}\hat{\sigma}(\hat{\beta}_1, \hat{\beta}_2) \end{aligned}$$

Intervalles de Prédition pour les Nouvelles Observations

Dans le cadre d'un éventuel programme d'expansion, RBAG Studios souhaiterait prévoir les ventes de deux nouvelles villes, présentant les caractéristiques suivantes :

	Ville A	Ville B
X_{h1}	65.4	53.1
X_{h2}	17.6	17.7

Des intervalles de prévision avec un coefficient de confiance de 95% sont souhaités. Notez que les deux nouvelles villes ont des caractéristiques qui s'inscrivent bien dans le schéma des 21 villes sur lesquelles est basée l'analyse de régression.

Pour la ville A, nous utilisons les résultats obtenus lors de l'estimation des ventes moyennes, car les niveaux des variables prédictives sont les mêmes ici. Nous avons d'avant :

$$\hat{Y}_h = 191.10 \quad \hat{\sigma}^2(\hat{Y}_h) = 7.656 \quad \hat{\sigma}^2 = 121.1626$$

Par conséquent

$$\hat{\sigma}^2(\text{préd}) = \hat{\sigma}^2 + \hat{\sigma}^2(\hat{Y}_h) = 121.1626 + 7.656 = 128.82$$

3 Modèle de Régression Linéaire Multiple

encore

$$\hat{\sigma}(\text{prééd}) = 11.35$$

De manière similaire, on obtient pour la ville B :

$$\hat{Y}_h = 174.15 \quad \hat{\sigma}(\text{prééd}) = 11.93$$

Les intervalles de prévision avec un coefficient de confiance de 95% :

$$\text{Ville A} \quad 191.10 \pm 2.101 (11.35) \Rightarrow 167.3 \leq Y_{h(\text{nouvelle})} \leq 214.9$$

$$\text{Ville B} \quad 174.15 \pm 2.101 (11.93) \Rightarrow 149.1 \leq Y_{h(\text{nouvelle})} \leq 199.2$$

Avec un coefficient de confiance de 0.95%, nous prévoyons que les ventes dans les deux villes se situeront dans les intervalles indiqués. RBAG Studios considère que ces limites de prévision sont quelque peu utiles pour la planification, mais préféreraient des intervalles plus rapprochés pour prévoir les ventes dans une ville particulière. Un cabinet de conseil a été engagé pour déterminer s'il est possible de trouver des variables prédictives supplémentaires ou alternatives conduisant à des intervalles de prévision plus courts.

Notons par ailleurs que même si le coefficient de détermination multiple, $R^2 = 0.917$, est élevé, les limites de prédition ici ne sont pas pleinement satisfaisantes. Cela sert à rappeler qu'une valeur élevée de R^2 n'indique pas nécessairement que des prédictions précises peuvent être faites.

3.12 Modèle de Régression Multiple Standardisé

Une forme normalisée du modèle général de régression multiple est utilisée pour contrôler les erreurs d'arrondi dans les calculs d'équations normales et pour permettre des comparaisons des coefficients de régression estimés en unités communes.

Erreurs d'Arrondi dans les Calculs d'Equations Normales

Les résultats des calculs d'équations normales peuvent être sensibles à l'arrondissement des données aux étapes intermédiaires des calculs. Lorsque le nombre de variables X est petit, disons - trois ou moins - les effets d'arrondi peuvent être contrôlés en portant un nombre suffisant de chiffres dans les calculs intermédiaires. En effet, la plupart des programmes de régression informatique utilisent l'arithmétique à double précision dans tous les calculs pour contrôler les effets d'arrondi. Néanmoins, avec un grand nombre de variables X , de graves effets d'arrondi

3 Modèle de Régression Linéaire Multiple

peuvent survenir malgré l'utilisation de nombreux chiffres dans les calculs intermédiaires.

Les erreurs d'arrondi ont tendance à entrer dans les calculs d'équations normales principalement lorsque l'inverse de $X'X$ est pris en considération. Bien entendu, toute erreur dans $(X'X)^{-1}$ peut être amplifiée dans le calcul de $\hat{\beta}$ et d'autres statistiques ultérieures. Le danger des erreurs d'arrondi dans $(X'X)^{-1}$ est particulièrement grand lorsque (1) $X'X$ a un déterminant qui est proche de zéro et / ou (2) les éléments de $X'X$ diffèrent sensiblement par ordre de grandeur. La première condition survient lorsque certaines ou toutes les variables X sont fortement corrélées.

La deuxième condition se pose lorsque les variables X ont des amplitudes sensiblement différentes de sorte que les entrées dans la matrice $X'X$ couvrent une large plage, disons, de 15 à 49,000,000. Une solution à cette condition consiste à transformer les variables et à reparamétriser ainsi le modèle de régression en modèle de régression standardisé (ou normalisé).

La transformation pour obtenir le modèle de régression normalisé, appelée *transformation de corrélation*, fait que toutes les entrées de la matrice $X'X$ pour les variables transformées se situent entre -1 et 1 inclus, de sorte que le calcul de la matrice inverse devient sujet aux moins d'erreurs d'arrondi et ce en raison d'ordres de grandeur différents de ceux des variables d'origine.

Transformation de Corrélation

L'utilisation de la transformation de corrélation aide à contrôler les erreurs d'arrondi et, en exprimant les coefficients de régression dans les mêmes unités, peut être utile lorsque ces coefficients sont comparés. Nous décrirons d'abord la transformation de corrélation puis le modèle de régression standardisé résultant.

La transformation de corrélation est une simple modification de la standardisation habituelle d'une variable. La normalisation d'une variable implique le centrage et la mise à l'échelle de la variable. *Le centrage* consiste à prendre la différence entre chaque observation et la moyenne des observations pour la variable ; *la mise à l'échelle* consiste à exprimer les observations centrées en unités de l'écart-type des observations pour la variable. Ainsi, la normalisation de la variable de réponse Y et des variables prédictives X_1, \dots, X_m est la suivante :

3 Modèle de Régression Linéaire Multiple

$$\frac{Y_i - \bar{Y}}{\hat{\sigma}_Y} \quad \text{et} \quad \frac{X_{ik} - \bar{X}_k}{\hat{\sigma}_k} \quad (k = 1, \dots, m) \quad (3.100)$$

où \bar{Y} et \bar{X}_k sont les moyennes respectives des observations Y et X_k , et $\hat{\sigma}_Y$ et $\hat{\sigma}_k$ sont les écarts types respectifs définis comme suit :

$$\hat{\sigma}_Y = \sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n-1}} \quad \text{et} \quad \hat{\sigma}_k = \sqrt{\frac{\sum_i (X_{ik} - \bar{X}_k)^2}{n-1}} \quad (k = 1, \dots, m) \quad (3.101)$$

La transformation de corrélation est une fonction simple des variables standardisées dans (7.43a, b) :

$$Y_i^* = \frac{1}{\sqrt{n-1}} \left(\frac{Y_i - \bar{Y}}{\hat{\sigma}_Y} \right) \quad \text{et} \quad X_{ik} = \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \bar{X}_k}{\hat{\sigma}_k} \right) \quad (k = 1, \dots, m) \quad (3.102)$$

Modèle de Régression Standardisé

Le modèle de régression avec les variables transformées Y^* et X_k^* telles que définies par la transformation de corrélation dans (3.102) est appelé *un modèle de régression standardisé* et se présente comme suit :

$$Y_i^* = \beta_1^* X_{i1}^* + \cdots + \beta_m^* X_{im}^* + e_i^* \quad (3.103)$$

La raison pour laquelle il n'y a pas de paramètre d'interception dans le modèle de régression normalisé (3.103) est que les moindres carrés ou les calculs de maximum de vraisemblance conduiraient toujours à un terme d'interception estimé à zéro si un paramètre d'interception était présent dans le modèle.

Il est facile de montrer que les paramètres $\beta_1^*, \dots, \beta_m^*$ dans le modèle de régression normalisé et les paramètres d'origine $\beta_0, \beta_1, \dots, \beta_m^*$ dans le modèle de régression multiple ordinaire (3.6) sont liés comme suit :

$$\beta_k = \left(\frac{\hat{\sigma}_Y}{\hat{\sigma}_k} \right) \beta_k^* \quad (k = 1, \dots, m) \quad (3.104)$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \cdots - \beta_m \bar{X}_m \quad (3.105)$$

Nous voyons que les coefficients de régression standardisés β_k^* et les coefficients de régression d'origine $\beta_k (k = 1, \dots, m)$ sont liés par des

3 Modèle de Régression Linéaire Multiple

facteurs d'échelle simples impliquant des ratios d'écart-types.

Matrice $X'X$ des Variables Transformées

Afin de pouvoir étudier la nature particulière de la matrice $X'X$ et les équations normales des moindres carrés lorsque les variables ont été transformées, il faut décomposer en deux matrices la matrice de corrélation ci-dessous qui contient tous les coefficients de corrélation par paire entre les variables de réponse et prédictives Y, X_1, X_2, \dots, X_m .

$$\begin{bmatrix} 1 & r_{Y1} & r_{Y2} & \cdots & r_{1m} \\ r_{Y1} & 1 & r_{12} & \cdots & r_{1m} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{Ym} & r_{1m} & r_{2m} & \cdots & 1 \end{bmatrix}$$

- (1) La première matrice, notée r_{XX} , est appelée *matrice de corrélation* des variables X . Elle a pour éléments les coefficients de corrélation simple entre toutes les paires des variables X . Cette matrice est définie comme suit :

$$r_{XX} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & & 1 \end{bmatrix} \quad (3.106)$$

La matrice de corrélation r_{XX} est symétrique ; rappelez-vous que $r_{kk'} = r_{k'k}$.

- (2) La deuxième matrice, notée r_{YX} est un vecteur contenant les coefficients de corrélation simple entre la variable de réponse Y et chacune des variables X , notée à nouveau par r_{Y1}, r_{Y2} , etc. :

$$r_{YX} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Ym} \end{bmatrix} \quad (3.107)$$

On peut considérer maintenant la matrice $X'X$ pour les variables transformées dans le modèle de régression standardisé (3.103). La matrice X est donc :

3 Modèle de Régression Linéaire Multiple

$$X_{n \times m} = \begin{bmatrix} X_{11}^* & \cdots & X_{1m}^* \\ X_{21}^* & \cdots & X_{2m}^* \\ \vdots & & \vdots \\ X_{n1}^* & \cdots & X_{nm}^* \end{bmatrix} \quad (3.108)$$

N'oubliez pas que le modèle de régression standardisé (3.103) ne contient pas de terme d'interception. On peut montrer que la matrice $X'X$ pour les variables transformées est simplement la matrice de corrélation des variables X définie en (3.106) ;

$$X'X = r_{XX} \quad (3.109)$$

La matrice $X'X$ des variables transformées étant constituée de coefficients de corrélation entre les variables X , tous ses éléments sont compris entre -1 et 1 et sont donc du même ordre de grandeur. Comme nous l'avons souligné précédemment, cela peut être très utile pour contrôler les erreurs d'arrondi lors de l'inversion de la matrice $X'X$.

Remarque

Nous illustrons que la matrice $X'X$ pour les variables transformées est la matrice de corrélation des variables X en considérant deux entrées dans la matrice :

(1) Dans le coin supérieur gauche de $X'X$, nous avons :

$$\sum (X_{i1}^*)^2 = \sum \left(\frac{X_{i1} - \bar{X}_1}{\sqrt{n-1}\hat{\sigma}_1} \right)^2 = \frac{\sum (X_{i1} - \bar{X}_1)^2}{(n-1)\hat{\sigma}_1^2} \quad (3.110)$$

(2) Dans la première ligne, et la deuxième colonne de $X'X$, nous avons :

$$\begin{aligned} \sum X_{i1}^* X_{i2}^* &= \sum \left(\frac{X_{i1} - \bar{X}_1}{\sqrt{n-1}\hat{\sigma}_1} \right) \left(\frac{X_{i2} - \bar{X}_2}{\sqrt{n-1}\hat{\sigma}_2} \right) \\ &= \frac{1}{n-1} \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\hat{\sigma}_1 \hat{\sigma}_2} \\ &= \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\left[\sum (X_{i1} - \bar{X}_1)^2 \sum (X_{i2} - \bar{X}_2)^2 \right]^{1/2}} \end{aligned} \quad (3.111)$$

3 Modèle de Régression Linéaire Multiple

Or, cela est équivalent à r_{12} , le coefficient de corrélation entre X_1 et X_2 .

Coefficients de Régression Standardisés Estimés

Les équations normales des moindres carrés pour le modèle de régression multiple ordinaire :

$$X'X\hat{\beta} = X'Y \quad (3.112)$$

et les estimateurs des moindres carrés :

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (3.113)$$

On peut montrer que pour les variables transformées, $X'Y$ devient :

$$X'Y = r_{YX} \quad (3.114)$$

où r_{YX} est défini en (3.107) comme le vecteur des coefficients de corrélation simple entre Y et chaque variable X . Il résulte maintenant à partir de (3.109) et (3.114) que les équations normales des moindres carrés et les estimateurs des coefficients de régression du modèle de régression normalisé (3.103) sont les suivants :

$$r_{XX}\hat{\beta} = r_{YX} \quad (3.115)$$

$$\hat{\beta} = r_{XX}^{-1}r_{YX} \quad (3.116)$$

où

$$\hat{\beta}_{m \times 1} = \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \vdots \\ \hat{\beta}_m^* \end{bmatrix} \quad (3.117)$$

Les coefficients de régression $\hat{\beta}_1^*, \dots, \hat{\beta}_m^*$ sont souvent appelés *coefficients de régression standardisés*.

Le retour aux coefficients de régression estimés pour le modèle de régression (3.6) dans les variables d'origine est accompli en utilisant les relations :

$$\hat{\beta}_k = \left(\frac{\hat{\sigma}_Y}{\hat{\sigma}_k} \right) \hat{\beta}_k^* \quad (k = 1, \dots, m) \quad (3.118)$$

3 Modèle de Régression Linéaire Multiple

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \cdots - \hat{\beta}_m \bar{X}_m \quad (3.119)$$

Remarque

Lorsqu'il y a deux variables explicatives (c-à-d $m = 2$) dans le modèle de régression, nous pouvons facilement voir la forme algébrique des coefficients de régression normalisés. Nous avons :

$$r_{XX} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \quad (3.120)$$

$$r_{YX} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \end{bmatrix} \quad (3.121)$$

$$r_{XX}^{-1} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \quad (3.122)$$

Par conséquent, et à partir de (3.116) nous obtenons :

$$\hat{\beta} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \begin{bmatrix} r_{Y1} \\ r_{Y2} \end{bmatrix} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} r_{Y1} & -r_{12}r_{Y2} \\ r_{Y2} & -r_{12}r_{Y1} \end{bmatrix} \quad (3.123)$$

Ainsi

$$\hat{\beta}_1^* = \frac{r_{Y1} - r_{12}r_{Y2}}{1 - r_{12}^2} \quad \text{et} \quad \hat{\beta}_2^* = \frac{r_{Y2} - r_{12}r_{Y1}}{1 - r_{12}^2} \quad (3.124)$$

3.13 Formes Fonctionnelles des Modèles de Régression

Dans cette section, nous examinons certains modèles de régression couramment utilisés qui peuvent être non linéaires dans les variables mais sont linéaires dans les paramètres ou qui peuvent être créés par des transformations appropriées des variables. Nous discutons en particulier des modèles de régression suivants :

1. Le Modèle Log-Linéaire
2. Modèles Semilog
3. Modèles Réciproques
4. Le Modèle Réciproque Logarithmique

Nous discutons des particularités de chaque modèle, quand ils sont appropriés et comment ils sont estimés. Chaque modèle est illustré avec des exemples appropriés.

3 Modèle de Régression Linéaire Multiple

3.13.1 Comment Mesurer l'Élasticité : le Modèle Log-Linéaire

Considérez le modèle suivant, appelé *Modèle de Régression Exponentielle* :

$$Y_i = \beta_0 X_i^{\beta_1} e^{\varepsilon_i} \quad (3.125)$$

qui peut être exprimé alternativement comme¹¹

$$\ln Y_i = \ln \beta_0 + \beta_1 \ln X_i + \varepsilon_i \quad (3.126)$$

où \ln = log naturel (c.-à-d., log à la base e , et où $e = 2.718$).¹²

On peut présenter le modèle 3.126 sous forme :

$$\ln Y_i = \alpha + \beta_1 \ln X_i + \varepsilon_i \quad (3.127)$$

où $\alpha = \ln \beta_0$, ce modèle est linéaire dans les paramètres α et β_1 , linéaire dans les logarithmes des variables Y et X et peut être estimé par MCO. En raison de cette linéarité, ces modèles sont appelés modèles « log-log », « double-log » ou « log-linéaire ».

Si les hypothèses du modèle de régression linéaire classique sont remplies, les paramètres de modèle (3.127) peuvent être estimés par la méthode des MCO en considérant le modèle :

$$Y_i^* = \alpha + \beta_1 X_i^* + \varepsilon_i \quad (3.128)$$

où $Y_i^* = \ln Y_i$ et $X_i^* = \ln X_i$. Les estimateurs MCO $\hat{\alpha}$ et $\hat{\beta}_1$ obtenus seront les meilleurs estimateurs linéaires non biaisés de α et β_1 , respectivement.

Une caractéristique intéressante du modèle log-log, qui l'a rendu populaire dans les travaux appliqués, est que le coefficient de pente β_1 mesure l'**élasticité** de Y par rapport à X , c'est-à-dire la variation en pourcentage de Y pour un (petit) pourcentage donné de variation en X .¹³ Ainsi,

11. Notez ces propriétés des logarithmes : (1) $\ln(AB) = \ln(A) + \ln(B)$, (2) $\ln(A/B) = \ln(A) - \ln(B)$ et (3) $\ln(A^k) = k \ln(A)$ en supposant que A et B sont positifs et k est une constante.

12. En pratique, on peut utiliser des logarithmes communs, c'est-à-dire log à la base 10. La relation entre le log naturel et le log commun est la suivante : $\ln_e X = 2.3026 \log_{10} X$. Par convention, \ln signifie logarithme naturel et \log signifie logarithme de la base 10 ; par conséquent, il n'est pas nécessaire d'écrire explicitement les indices e et 10.

13. Le coefficient d'élasticité est défini par $(dY/Y) / (dX/X) = [(dY/dX) / (X/Y)]$. Les lecteurs familiarisés avec le calcul différentiel verront facilement que β_1 est en fait le coefficient d'élasticité.

Note technique : Le lecteur intéressé par le calcul remarquera que $d(\ln X)/dX = 1/X$ ou $d(\ln X)/dX = 1/X$, c'est-à-dire pour des modifications in-

3 Modèle de Régression Linéaire Multiple

si Y représente la quantité d'un produit demandée et X son prix unitaire, β_1 mesure l'élasticité-prix de la demande, qui est un paramètre d'intérêt économique considérable. Si la relation entre la quantité demandée et le prix est celle illustrée à la figure (3.5a), la transformation à double log, illustrée à la figure (3.5b), donnera alors l'estimation de l'élasticité-prix ($-\beta_1$).

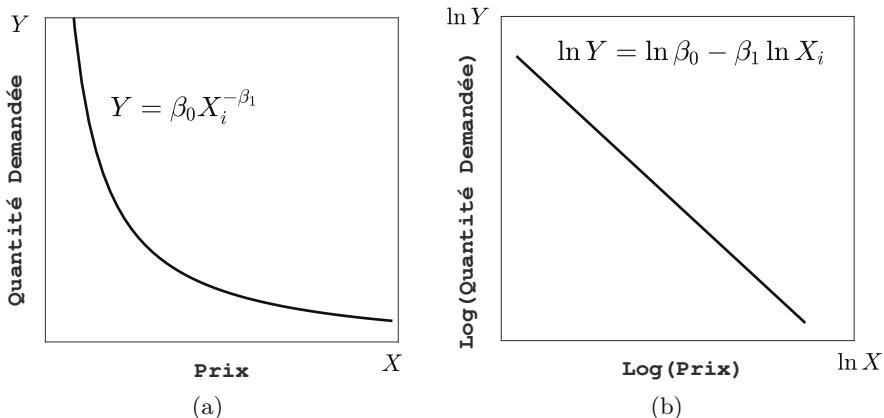


FIGURE 3.5 – Modèle à Elasticité Constante

On peut noter deux particularités du modèle log-linéaire : Le modèle suppose que le coefficient d'élasticité entre Y et X , β_1 , reste constant tout au long (pourquoi ?), D'où la connotation *modèle à élasticité constante*.¹⁴ En d'autres termes, comme le montre la figure (3.5b), la variation de

finiment petites (notez l'opérateur différentiel d) le changement de $\ln X$ est égal au changement relatif ou proportionnel de X . En pratique, cependant, si le changement de X est faible, cette relation peut s'écrire comme suit : changement dans $\ln X$ = changement relatif dans X , où (=) signifie approximativement. Ainsi, pour de petits changements,

$$(\ln X_t - \ln X_{t-1}) = (X_t - X_{t-1}) / X_{t-1} = \text{Changement relatif dans } X$$

Incidemment, le lecteur devrait noter ces termes qui apparaîtront fréquemment : (1) changement absolu, (2) changement relatif ou proportionnel, et (3) changement en pourcentage ou taux de croissance en pourcentage. Ainsi, $(X_t - X_{t-1})$ représente le changement absolu, $(X_t - X_{t-1}) / X_{t-1} = (X_t / X_{t-1} - 1)$ est un changement relatif ou proportionnel et $[(X_t - X_{t-1}) / X_{t-1}] 100$ est le pourcentage de variation ou le taux de croissance. X_t et X_{t-1} sont, respectivement, les valeurs actuelle et précédente de la variable X .

14. Un modèle à élasticité constante donnera une variation totale constante du chiffre d'affaires pour une variation en pourcentage du prix donnée, quel que soit le niveau absolu du prix. Les lecteurs doivent opposer ce résultat aux conditions d'élasticité impliquées par une simple fonction de demande linéaire, $Y_i = \beta_0 + \beta_1 X_i +$

3 Modèle de Régression Linéaire Multiple

$\ln Y$ par unité de variation en $\ln X$ (c'est-à-dire l'élasticité, β_1) reste la même, quel que soit le cas où nous mesurons l'élasticité. Une autre caractéristique du modèle est que bien que $\hat{\alpha}$ et $\hat{\beta}_1$ soient des estimations non biaisées de α et β_1 , β_0 (le paramètre entrant dans le modèle initial) lorsqu'il est estimé comme $\hat{\beta}_1 = \text{antilog}(\hat{\alpha})$ est en soi un estimateur biaisé. Cependant, dans la plupart des problèmes pratiques, le terme d'interception a une importance secondaire et il n'est pas à craindre d'obtenir son estimation non biaisée.

Dans le modèle à une variable explicative, la façon la plus simple de décider si le modèle log-linéaire convient aux données est de tracer la dispersion de $\ln Y_i$ par rapport à $\ln X_i$ et de voir si les points de dispersion se trouvent approximativement sur une ligne droite, comme dans la figure (3.5b).

Remarque :

Le lecteur doit être conscient de la distinction entre un changement de pourcentage et un changement de point de pourcentage. Par exemple, le taux de chômage est souvent exprimé sous forme de pourcentage, disons, le taux de chômage de 6%. Si ce taux passe à 8%, nous disons que la variation en points de pourcentage du taux de chômage est de 2, alors que la variation en pourcentage du taux de chômage est de $(8 - 6) / 6$, soit environ 33%. Donc, il faut être prudent lors du traitement des changements de pourcentage et de point de pourcentage, car les deux sont des concepts très différents.

Exemple : Dépenses en Biens Durables par Rapport aux Dépenses Totales de Consommation Personnelle

Le tableau (3.4) présente les données sur les dépenses totales de consommation personnelle (PCEXP), les dépenses en biens durables (EXP-DUR), les dépenses en biens non durables (EXPNONDUR) et les dépenses en services (EXPSERVICES), toutes mesurées en 2000 milliards de dirhams.¹⁵

Supposons que nous souhaitions trouver l'élasticité des dépenses en biens durables par rapport aux dépenses totales de consommation personnelle. En comparant le logarithme des dépenses en biens durables avec le logarithme des dépenses totales de consommation personnelle, vous constaterez que la relation entre les deux variables est linéaire. Par conséquent,

ε_i . Cependant, une fonction linéaire simple donne un changement constant de quantité par un changement de prix unitaire. Comparez cela avec ce que le modèle log-linéaire implique pour une variation donnée du prix.

15. Les biens durables comprennent les véhicules et les pièces, les meubles et l'équipement ménager ; les biens non durables comprennent les aliments, les vêtements, l'essence et le pétrole, le charbon ; et les services comprennent le logement, l'électricité et l'eau, le transport et les soins médicaux.

3 Modèle de Régression Linéaire Multiple

le modèle à double log peut être approprié. Les résultats de la régression sont les suivants :

Année ou Trimestre	EXPSERVICES	EXPDUR	EXPNONDUR	PCEXP
2013-I	4143.30	971.40	2072.50	7184.90
2013-II	4161.30	1009.80	2084.20	7249.30
2013-III	4190.70	1049.60	2123.00	7352.90
2013-IV	4220.20	1051.40	2132.50	7394.30
2014-I	4268.20	1067.00	2155.30	7479.80
2014-II	4308.40	1071.40	2164.30	7534.40
2014-III	4341.50	1093.90	2184.00	7607.10
2014-IV	4377.40	1110.30	2213.10	7687.10
2015-I	4395.30	1116.80	2241.50	7739.40
2015-II	4420.00	1150.80	2268.40	7819.80
2015-III	4454.50	1175.90	2287.60	7895.30
2015-IV	4476.70	1137.90	2309.60	7910.20
2016-I	4494.50	1190.50	2342.80	8003.80
2016-II	4535.40	1190.30	2351.10	8055.00
2016-III	4566.60	1208.80	2360.10	8111.20

TABLE 3.4 – Dépenses Totales Personnelles

$$\begin{aligned}
 \ln \widehat{\text{EXP}DUR}_t &= -7.5417 & +1.6266 \ln PCEX_t & \quad (3.129) \\
 \hat{\sigma} &= (0.7161) & (0.0800) \\
 t &= (-10.5309)^* & (20.3152) \\
 R^2 &= 0.9695
 \end{aligned}$$

où * indique que la valeur p est extrêmement petite.

Comme le montrent ces résultats, l'élasticité d'EXPDUR par rapport au PCEX est d'environ 1.63, ce qui suggère que si les dépenses personnelles totales augmentent de 1%, en moyenne, les dépenses en biens durables augmentent d'environ 1.63%. Les dépenses consacrées aux biens durables sont donc très sensibles aux variations des dépenses de consommation personnelle. C'est l'une des raisons pour lesquelles les producteurs de biens durables surveillent attentivement l'évolution du revenu personnel et des dépenses de consommation personnelle.

3.13.2 Modèles Semilog : Modèles Log-Lin et Lin-Log

Comment Mesurer le Taux de Croissance : le Modèle Log-Lin

Les économistes, les hommes d'affaires et les gouvernements sont souvent intéressés par le taux de croissance de certaines variables économiques, telles que la population, le PNB, la masse monétaire, l'emploi, la productivité et le déficit commercial.

Supposons que nous voulions connaître le taux de croissance de la dépense de consommation personnelle en services pour les données du tableau (3.4). Soit Y_t la dépense réelle en services au moment t et Y_0 la valeur initiale de la dépense en services (c'est-à-dire la valeur à la fin de 2012-IV). Vous vous souvenez peut-être de la formule d'intérêt composé bien connue de votre cours des mathématiques financières.

$$Y_t = Y_0 (1 + r)^t \quad (3.130)$$

où r est le taux de croissance composé (c'est-à-dire dans le temps) de Y . En prenant le logarithme naturel de l'équation (3.130), nous pouvons écrire

$$\ln Y_t = \ln Y_0 + t \ln (1 + r) \quad (3.131)$$

Soit maintenant

$$\beta_0 = \ln Y_0 \quad \text{et} \quad \beta_1 = \ln (1 + r) \quad (3.132)$$

donc l'équation (3.131) devient

$$\ln Y_t = \beta_0 + \beta_1 t \quad (3.133)$$

En ajoutant le terme d'erreur à l'équation (3.133), nous obtenons¹⁶

$$\ln Y_t = \beta_0 + \beta_1 t + \varepsilon_i \quad (3.134)$$

Ce modèle ressemble à tout autre modèle de régression linéaire du fait aussi que les paramètres β_0 et β_1 sont linéaires. La seule différence est qu'on a le logarithme de Y dans la variable dépendante et que le régresseur est le «temps», qui prendra les valeurs 1, 2, 3, etc.

Les modèles comme (3.134) sont appelés *Modèles Semilog* car une seule variable (dans ce cas, la variable dépendante) apparaît dans la forme logarithmique. À des fins descriptives, un modèle dans lequel variable dépendante est logarithmique sera appelé *Modèle Log-Lin*. Le modèle

16. Nous ajoutons le terme d'erreur car la formule de calcul des intérêts composés ne sera pas exacte.

3 Modèle de Régression Linéaire Multiple

dans lequel la variable dépendante est linéaire mais le (ou les) régresseur est logarithmique est appelé *Modèle Lin-Log*.

Avant de présenter les résultats de la régression, examinons les propriétés du modèle (3.133). Dans ce modèle, le coefficient de pente mesure le changement proportionnel ou relatif constant de Y pour un changement absolu donné de la valeur du régresseur (dans ce cas, la variable t), c'est-à-dire¹⁷

$$\beta_1 = \frac{\text{Changement relatif de } Y}{\text{Changement absolu du régresseur}} \quad (3.135)$$

Si nous multiplions le changement relatif de Y par 100, l'équation (3.135) donnera alors le changement en pourcentage ou le taux de croissance en Y pour un changement absolu en X , le régresseur. C'est-à-dire 100 fois β_1 donne le taux de croissance de Y ; 100 fois β_1 est connue dans la littérature comme la *Semi-élasticité* de Y par rapport à X . (*Question : pour obtenir l'élasticité, que devons-nous faire ?*)

Exemple : *Le Taux de Croissance des Dépenses en Services*

Pour illustrer le modèle de croissance (3.134), considérons les données sur les dépenses en services données dans le tableau (3.4). Les résultats de la régression dans le temps (t) sont les suivants :

$$\begin{aligned} \widehat{\ln EXS}_t &= 8.3226 & +0.00705t & \quad (3.136) \\ s &= (0.0016) & (0.0008) & \\ t &= (5201.625)^* & (39.1667)^* & \\ R^2 &= 0.9919 & & \end{aligned}$$

Remarque : EXS signifie les dépenses en services et * indique que la valeur p est extrêmement petite.

Selon l'interprétation de l'équation (3.136), au cours de la période trimestrielle allant de 2013-I à 2016-III, les dépenses en services ont augmenté au taux (trimestriel) de 0.705%. En gros, cela correspond à un taux de croissance annuel de 2.82%. Depuis $8.3226 = \log$ de EXS au début de la période d'étude, en prenant son antilog, nous obtenons 4115.96 (milliards de dhs) comme valeur de départ de EXS (c'est-à-dire la va-

17. En utilisant le calcul différentiel, on peut montrer que $\beta_1 = d(\ln Y)/dX = (1/Y)(dY/dX) = (dY/Y)/dX$, qui n'est autre que l'équation (3.108). Pour de petits changements dans Y et X , cette relation peut être approximée par

$$\frac{(Y_t - Y_{t-1})/Y_{t-1}}{(X_t - X_{t-1})}$$

Remarque : Ici, $X = t$

3 Modèle de Régression Linéaire Multiple

leur du début de 2013). La droite de régression obtenue dans l'équation (3.136) est représentée à la figure (3.6).

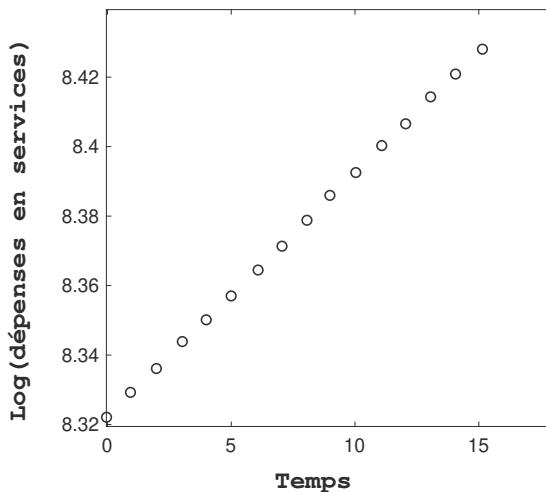


FIGURE 3.6 – Droite de Régression

Le Modèle Lin–Log

Contrairement au modèle de croissance que nous venons de traiter, dans lequel nous voulions trouver le pourcentage de croissance de Y pour un changement absolu de X , supposons maintenant que nous voulions trouver le changement absolu de Y pour un changement de pourcentage de X . Un modèle capable de réaliser cela peut être écrit comme :

$$Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i \quad (3.137)$$

Pour des fins descriptives, nous appelons ce modèle un Modèle Lin–Log. Interprétons le coefficient de pente β_1 .¹⁸ Comme d'habitude,

$$\beta_1 = \frac{\text{Changement en } Y}{\text{Changement en } \ln X} = \frac{\text{Changement en } Y}{\text{Changement relatif en } X} \quad (3.138)$$

La deuxième étape découle du fait qu'*un changement dans le log d'un nombre est un changement relatif*.

18. Encore une fois, en utilisant le calcul différentiel, nous avons $\frac{dY}{dX} = \beta_2 \left(\frac{1}{X} \right) \Rightarrow \beta_2 = dY (dX/X) = \text{Equation}(3.137)$

3 Modèle de Régression Linéaire Multiple

Symboliquement, nous avons

$$\beta_1 = \frac{\Delta Y}{\Delta X/X} \quad (3.139)$$

ou

$$\Delta Y = \beta_1 (\Delta X/X) \quad (3.140)$$

Cette équation stipule que la variation absolue de Y ($= \Delta Y$) est égale à la pente multipliée par la modification relative de X . Si cette dernière est multipliée par 100, alors l'équation (3.140) donne la variation absolue de Y pour une variation en pourcentage de X . Ainsi, si $(\Delta X/X)$ change de 0.01 unité (ou 1%), la variation absolue de Y est 0.01 (β_1); si, dans une application, on trouve que $\beta_1 = 500$, la variation absolue de Y est $(0.01)(500) = 5$. Par conséquent, lorsque la régression (3.137) est estimée par MCO, n'oubliez pas de multiplier la valeur du coefficient estimé de la pente par 0.01 ou, ce qui revient au même, de la diviser par 100. *Si vous ne le tenez pas en considération, votre interprétation dans une application sera très trompeuse.*

La question pratique est : Quand un modèle Lin–Log comme (3.137) est utile ? Une application intéressante a été trouvée dans les modèles appelés « Modèle de Dépenses d'Engel », du nom du statisticien allemand Ernst Engel, 1821–1896. Engel a postulé que “*les dépenses totales consacrées à l'alimentation ont tendance à augmenter en progression arithmétique à mesure que les dépenses totales augmentent en progression géométrique*”.

Remarque :

On peut noter que parfois, la transformation logarithmique est utilisée pour réduire l'hétéroscédasticité ainsi que l'asymétrie. De nombreuses variables économiques montrent en commun une asymétrie positive (par exemple, répartition de la taille des entreprises ou répartition du revenu ou de la richesse) et elles sont hétéroscédastiques. Une transformation logarithmique de telles variables réduit à la fois l'asymétrie et l'hétéroscédasticité. C'est pourquoi les économistes du travail utilisent souvent les logarithmes des salaires dans la régression des salaires sur, par exemple, la scolarité qui est mesurée par le nombre d'années d'études.

3.13.3 Modèles Réciproques

Les modèles du type suivant sont appelés « modèles réciproques »

$$Y_i = \beta_0 + \beta_1 \left(\frac{1}{X_i} \right) + \varepsilon_i \quad (3.141)$$

3 Modèle de Régression Linéaire Multiple

Bien que ce modèle soit non linéaire dans X car il entre inversement ou réciproquement, le modèle est linéaire dans β_0 et β_1 et est donc un modèle de régression linéaire.¹⁹

Ce modèle présente les caractéristiques suivantes : lorsque X augmente indéfiniment, le terme $\beta_1(1/X)$ s'approche de zéro (*remarque* : β_1 est une constante) et Y s'approche de la valeur limite ou *asymptotique* β_1 . Par conséquent, les modèles tels que (3.141) intègrent une asymptote ou une valeur limite que la variable dépendante prendra lorsque la valeur de la variable X augmente indéfiniment. Certaines formes de la courbe correspondant à l'équation (3.141) sont illustrés à la figure (3.7).

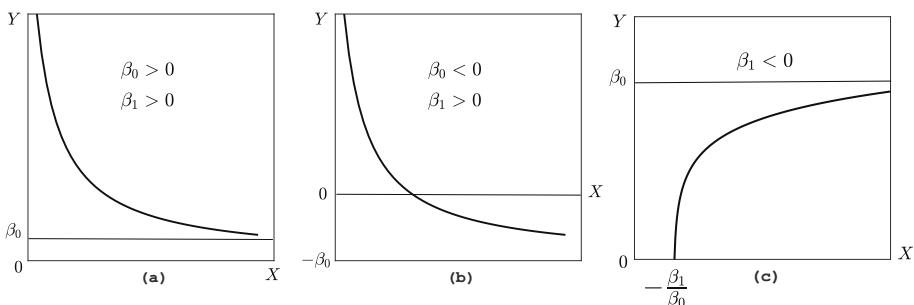


FIGURE 3.7 – **Modèles Réciproque** $Y_i = \beta_0 + \beta_1 \left(\frac{1}{X_i} \right)$

Modèle Réciproque Logarithmique

Nous concluons notre discussion sur les modèles réciproques en considérant le modèle réciproque logarithmique, qui prend la forme suivante :

$$\ln Y_i = \beta_0 - \beta_1 \left(\frac{1}{X_i} \right) + \varepsilon_i \quad (3.142)$$

Sa forme est celle illustrée à la figure (3.8). Comme le montre cette figure, initialement Y augmente à un taux croissant (c'est-à-dire que la courbe est initialement convexe) puis augmente à un taux décroissant (c'est-à-dire que la courbe devient concave).²⁰ Un tel modèle peut donc

19. Si on pose $X_i^* = (1/X_i)$, alors l'équation (3.141) sera linéaire dans les paramètres ainsi que dans les variables Y et X .

20. On peut montrer que $\frac{d}{dX} (\ln Y) = -\beta_1 \left(-\frac{1}{X^2} \right) = \beta_1 \left(\frac{1}{X^2} \right)$

mais $\frac{d}{dX} (\ln Y) = \frac{1}{Y} \frac{dY}{dX}$

En faisant cette substitution, on obtient $\frac{dY}{dX} = \beta_1 \frac{Y}{X^2}$

3 Modèle de Régression Linéaire Multiple

être approprié pour modéliser une fonction de production à court terme. Rappelez-vous de la microéconomie, lorsque le travail et le capital sont les intrants d'une fonction de production et si nous maintenons l'input de capital constant tout en augmentant l'intrant de travail, la relation production-travail à court terme ressemblera à la figure (3.8).

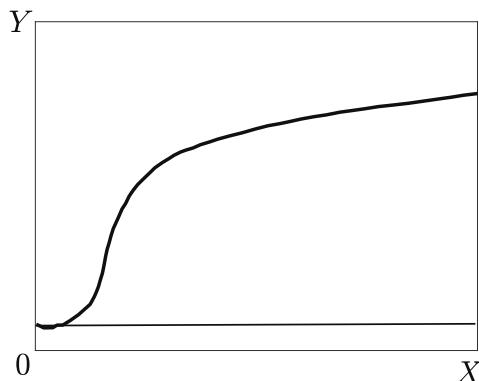


FIGURE 3.8 – **Modèle Réciproque Logarithmique**

3.14 Choix de la Forme Fonctionnelle

Dans la section précédente, nous avons montré plusieurs formes fonctionnelles qu'un modèle empirique peut prendre. Le choix d'une forme fonctionnelle particulière peut être relativement facile dans le cas d'une seule variable explicative, car nous pouvons tracer le couple d'observations et avoir une idée approximative du modèle approprié. Le choix devient beaucoup plus difficile lorsque nous considérons le modèle de régression multiple impliquant plusieurs régresseurs. Il est indéniable que le choix d'un modèle d'estimation empirique requiert beaucoup de compétences et d'expérience. Mais certaines directives peuvent être proposées :

- (1) La théorie sous-jacente peut suggérer une forme fonctionnelle particulière.
- (2) Parmis les bonnes pratiques est de connaître le taux de changement (c'est-à-dire la pente) de la régression par rapport au régresseur ainsi que de déterminer l'élasticité de la régression par rapport au régresseur. Pour les différents modèles examinés dans ce chapitre, nous fournissons les formules nécessaires pour les coefficients de pente et d'élasticité des différents modèles du tableau (3.5). La

3 Modèle de Régression Linéaire Multiple

connaissance de ces formules nous aidera à comparer les différents modèles.

Modèle	Équation	Pente $(= \frac{dY}{dX})$	Élasticité $(= \frac{dY}{dX} \frac{X}{Y})$
Linéaire	$Y = \beta_0 + \beta_1 X$	β_1	$\beta_1 \left(\frac{X}{Y} \right)^*$
Log-Linéaire	$\ln Y = \beta_0 + \beta_1 \ln X$	$\beta_1 \left(\frac{Y}{X} \right)$	β_1
Log-Lin	$\ln Y = \beta_0 + \beta_1 X$	$\beta_1 (Y)$	$\beta_1 (X)^*$
Lin-Log	$Y = \beta_0 + \beta_1 \ln X$	$\beta_1 \left(\frac{1}{X} \right)$	$\beta_1 \left(\frac{1}{Y} \right)^*$
Réiproque	$Y = \beta_0 + \beta_1 \left(\frac{1}{X} \right)$	$-\beta_1 \left(\frac{1}{X^2} \right)$	$-\beta_1 \left(\frac{1}{XY} \right)^*$
Réiproque Logarithmique	$\ln Y = \beta_0 - \beta_1 \left(\frac{1}{X} \right)$	$\beta_1 \left(\frac{Y}{X^2} \right)$	$\beta_1 \left(\frac{1}{X} \right)^*$

Remarque : * indique que l'élasticité varie en fonction de la valeur prise par X ou Y ou les deux. Lorsqu'aucune valeur de X et Y n'est spécifiée, dans la pratique, très souvent, ces élasticités sont mesurées au travers valeurs moyennes de ces variables, à savoir X et Y .

TABLE 3.5 – Les Coefficients de Pente et d'Élasticité des Différents Modèles

- (3) Les coefficients du modèle choisi devraient répondre à certaines attentes a priori. Par exemple, si nous considérons la demande de véhicules automobiles en fonction du prix et de certaines autres variables, nous devrions nous attendre à un coefficient négatif pour la variable de prix.
- (4) En général, il ne faut pas surestimer le R^2 dans la mesure où plus le R^2 est élevé, meilleur est le modèle. La valeur de R^2 augmente à mesure que nous ajoutons davantage de régresseurs au modèle. Ce qui est plus important, c'est le fondement théorique du modèle choisi, les signes des coefficients estimés et leur signification statistique. Si un modèle est bon sur ces critères, un modèle avec une valeur de R^2 inférieure peut être tout à fait acceptable.
- (5) Dans certaines situations, il peut s'avérer difficile de choisir une forme fonctionnelle particulière, auquel cas nous pouvons utiliser les transformations dites de Box-Cox.

3 Modèle de Régression Linéaire Multiple

Exemple : Modèle de Régression de Box-Cox

Considérons le modèle de régression suivant

$$Y_i^\lambda = \beta_0 + \beta_1 X_i + \varepsilon_i \quad Y > 0 \quad (3.143)$$

où λ est un paramètre qui peut être négatif, nul ou positif. Puisque Y est élevé à la puissance λ , nous aurons plusieurs transformations de Y , en fonction de la valeur de λ .

L'équation (3.143) est connue sous le nom de modèle de régression de « Box-Cox », par référence aux statisticiens Box et Cox. En fonction de la valeur de λ , nous avons les modèles de régression suivants, présentés sous forme de tableau (3.6) :

Valeur de λ	Modèle de Régression
1	$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
2	$Y_i^2 = \beta_0 + \beta_1 X_i + \varepsilon_i$
0.5	$\sqrt{Y_i} = \beta_0 + \beta_1 X_i + \varepsilon_i$
0	$\ln Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
-0.5	$\frac{1}{\sqrt{Y_i}} = \beta_0 + \beta_1 X_i + \varepsilon_i$
-1.0	$\frac{1}{Y_i} = \beta_0 + \beta_1 X_i + \varepsilon_i$

TABLE 3.6 – Modèles de Régression Selon la Valeur de λ

3.15 Annexe

Cas de Deux Variables Prédictives

Pour le cas deux variables prédictives, on a la fonction de régression suivante :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad (3.144)$$

$$\text{Min } S = \text{Min} \sum_{i=1}^n e_i^2 = \text{Min} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2})^2 \quad (3.145)$$

La ministration de S passe par le calcul de sa dérivée par rapport aux inconnus β_0 , $\hat{\beta}_1$ et $\hat{\beta}_2$.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \quad (3.146)$$

3 Modèle de Régression Linéaire Multiple

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^n Y_i X_{i1} \right) \left(\sum_{i=1}^n X_{i2}^2 \right) - \left(\sum_{i=1}^n Y_i X_{i2} \right) \left(\sum_{i=1}^n X_{i1} X_{i2} \right)}{\left(\sum_{i=1}^n X_{i1}^2 \right) \left(\sum_{i=1}^n X_{i2}^2 \right) - \left(\sum_{i=1}^n X_{i1} X_{i2} \right)^2} \quad (3.147)$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^n Y_i X_{i2} \right) \left(\sum_{i=1}^n X_{i1}^2 \right) - \left(\sum_{i=1}^n Y_i X_{i1} \right) \left(\sum_{i=1}^n X_{i1} X_{i2} \right)}{\left(\sum_{i=1}^n X_{i1}^2 \right) \left(\sum_{i=1}^n X_{i2}^2 \right) - \left(\sum_{i=1}^n X_{i1} X_{i2} \right)^2} \quad (3.148)$$

Variances et erreurs des estimateurs MCO

$$var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}_1^2 \sum_{i=1}^n X_{i2}^2 + \bar{X}_2^2 \sum_{i=1}^n X_{i1}^2 - 2\bar{X}_1 \bar{X}_2 \sum_{i=1}^n X_{i1} X_{i2}}{\sum_{i=1}^n X_{i1}^2 \sum_{i=1}^n X_{i2}^2 - \left(\sum_{i=1}^n X_{i1} X_{i2} \right)^2} \right] \quad (3.149)$$

$$var(\hat{\beta}_1) = \sigma^2 \frac{\sum_{i=1}^n X_{i2}^2}{\left(\sum_{i=1}^n X_{i1}^2 \right) \left(\sum_{i=1}^n X_{i2}^2 \right) - \left(\sum_{i=1}^n X_{i1} X_{i2} \right)^2} \quad (3.150)$$

ou encore

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n X_{i1}^2 (1 - r_{12}^2)} \quad (3.151)$$

ou r_{12} est le coefficient de corrélation linéaire simple entre X_1 et X_2 .

$$var(\hat{\beta}_2) = \sigma^2 \frac{\sum_{i=1}^n X_{i1}^2}{\left(\sum_{i=1}^n X_{i1}^2 \right) \left(\sum_{i=1}^n X_{i2}^2 \right) - \left(\sum_{i=1}^n X_{i1} X_{i2} \right)^2} \quad (3.152)$$

ou encore

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n X_{i2}^2 (1 - r_{12}^2)} \quad (3.153)$$

$$cov(\hat{\beta}_1, \hat{\beta}_2) = \frac{-r_{12}\sigma^2}{(1 - r_{12}^2) \sqrt{\sum_{i=1}^n X_{i1}^2} \sqrt{\sum_{i=1}^n X_{i2}^2}} \quad (3.154)$$

Dans toutes ces formules, σ^2 est la variance (homoscédastique) des per-

3 Modèle de Régression Linéaire Multiple

turbations ε_i . On peut vérifier qu'un estimateur non biaisé de σ^2 est donné par

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-3} \quad (3.155)$$

avec

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n Y_i^2 - \hat{\beta}_1 \sum_{i=1}^n Y_i X_{i1} - \hat{\beta}_2 \sum_{i=1}^n Y_i X_{i2} \quad (3.156)$$

Preuve :

on a

$$e_i = Y_i - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} \quad (3.157)$$

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (e_i e_i) = \sum_{i=1}^n e_i (Y_i - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2}) = \sum_{i=1}^n e_i Y_i \quad (3.158)$$

$$\text{car } \sum_{i=1}^n e_i X_{i1} = \sum_{i=1}^n e_i X_{i2} = 0$$

alors

$$\begin{aligned} \sum_{i=1}^n e_i^2 &= \sum_{i=1}^n e_i Y_i \\ &= \sum_{i=1}^n Y_i (Y_i - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2}) \\ &= \sum_{i=1}^n Y_i^2 - \hat{\beta}_1 \sum_{i=1}^n Y_i X_{i1} - \hat{\beta}_2 \sum_{i=1}^n Y_i X_{i2} \end{aligned} \quad (3.159)$$

3.16 Exercices

Exercice 1

Considérons les données du tableau ci-dessous

Y	X ₁	X ₂
1	1	2
3	2	1
8	3	-3

À partir de ces données, estimatez les régressions suivantes :

3 Modèle de Régression Linéaire Multiple

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_{1i} \quad (1)$$

$$Y_i = \beta'_0 + \beta'_1 X_{2i} + \varepsilon_{2i} \quad (2)$$

$$Y_i = \beta''_0 + \beta''_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (3)$$

Remarque : Estimez uniquement les coefficients et non les erreurs.

(a) Est-ce que $\beta_1 = \beta''_1$? Pourquoi?

(b) Est-ce que $\beta'_1 = \beta_2$? Pourquoi?

Quelle conclusion importante tirez-vous de cet exercice?

Solution

Pour les deux régressions (1) et (2) on procède par la méthode des MCO avec une seule variable explicative on obtient alors :

$$\hat{\beta}_0 = -3.00 \quad \hat{\beta}_1 = 3.50$$

$$\hat{\beta}'_0 = 4.00 \quad \hat{\beta}'_1 = -1.357$$

Pour la régression (3) on procède par un calcul matriciel pour déterminer $\hat{\beta} = (X'X)^{-1}X'Y$. On a

$$X = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 3 & -3 \end{pmatrix} \quad Y = \begin{pmatrix} 1 \\ 3 \\ 8 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 19 & -9.3333 & -3.3333 \\ -9.3333 & 4.6667 & 1.6667 \\ -3.3333 & 1.6667 & 0.6667 \end{pmatrix}$$

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ &= \begin{pmatrix} 19 & -9.3333 & -3.3333 \\ -9.3333 & 4.6667 & 1.6667 \\ -3.3333 & 1.6667 & 0.6667 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 2 & 1 & -3 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 8 \end{pmatrix} \end{aligned}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}''_0 \\ \hat{\beta}''_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 2.00 \\ 1.00 \\ -1.00 \end{pmatrix}$$

(a) Non, étant donné que le modèle (3) est le modèle réel, $\hat{\beta}_1$ est un estimateur biaisé de β''_1 .

3 Modèle de Régression Linéaire Multiple

(b) Non, $\hat{\beta}'_1$ est un estimateur biaisé de β_2 , pour la même raison que dans (a).

La leçon à tirer est qu'une mauvaise spécification d'une équation peut conduire à une estimation biaisée des paramètres du modèle réel.

Exercice 2

A partir des données suivantes, estimez les coefficients de régression, la somme des erreurs estimée et les valeurs de R^2 ajustés et non ajustés :

$$\begin{aligned}\bar{Y} &= 367.693 & \bar{X}_1 &= 402.760 & \bar{X}_2 &= 8.0 & n &= 15 \\ \sum_{i=1}^n (Y_i - \bar{Y})^2 &= 66042.269 & \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 &= 84855.096 \\ \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2 &= 280.00 & \sum_{i=1}^n (Y_i - \bar{Y})(X_{i1} - \bar{X}_1) &= 74778.346 \\ \sum_{i=1}^n (Y_i - \bar{Y})(X_{i2} - \bar{X}_2) &= 4250.90 & \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) &= 4796.00\end{aligned}$$

Solution

on a

$$\begin{aligned}& \cdot \sum_{i=1}^n (Y_i - \bar{Y})(X_{i1} - \bar{X}_1) = \sum_{i=1}^n Y_i X_{i1} - n \bar{Y} \bar{X}_1 \\ & \sum_{i=1}^n Y_i X_{i1} = 74778.346 + 15(367.693)(402.760) = 2296158.836 \\ & \cdot \sum_{i=1}^n (Y_i - \bar{Y})(X_{i2} - \bar{X}_2) = \sum_{i=1}^n Y_i X_{i2} - n \bar{Y} \bar{X}_2 \\ & \sum_{i=1}^n Y_i X_{i2} = 4250.90 + 15(367.693)(8) = 48374.06 \\ & \cdot \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) = \sum_{i=1}^n X_{i1} X_{i2} - n \bar{X}_1 \bar{X}_2 \\ & \sum_{i=1}^n X_{i1} X_{i2} = 4796 + 15(402.760)(8) = 53127.2 \\ & \cdot \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 = \sum_{i=1}^n X_{i1}^2 - n \bar{X}_1^2 \\ & \sum_{i=1}^n X_{i1}^2 = 84855.096 + 15(402.760)^2 = 2518089.36 \\ & \cdot \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2 = \sum_{i=1}^n X_{i2}^2 - n \bar{X}_2^2 \\ & \sum_{i=1}^n X_{i2}^2 = 280 + 15(8)^2 = 1240 \\ & \cdot \sum_{i=1}^n Y_i^2 = 66042.269 + 15(367.693)^2 = 2094014.403\end{aligned}$$

3 Modèle de Régression Linéaire Multiple

On remplace ces statistiques dans les formules (3.145) et (3.146) de $\hat{\beta}_1$ et $\hat{\beta}_2$:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^n Y_i X_{i1}\right) \left(\sum_{i=1}^n X_{i2}^2\right) - \left(\sum_{i=1}^n Y_i X_{i2}\right) \left(\sum_{i=1}^n X_{i1} X_{i2}\right)}{\left(\sum_{i=1}^n X_{i1}^2\right) \left(\sum_{i=1}^n X_{i2}^2\right) - \left(\sum_{i=1}^n X_{i1} X_{i2}\right)^2} = 0.924406$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^n Y_i X_{i2}\right) \left(\sum_{i=1}^n X_{i1}^2\right) - \left(\sum_{i=1}^n Y_i X_{i1}\right) \left(\sum_{i=1}^n X_{i1} X_{i2}\right)}{\left(\sum_{i=1}^n X_{i1}^2\right) \left(\sum_{i=1}^n X_{i2}^2\right) - \left(\sum_{i=1}^n X_{i1} X_{i2}\right)^2} = -0.594415$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 = 0.134559$$

on a d'après (3.154) $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n Y_i^2 - \hat{\beta}_1 \sum_{i=1}^n Y_i X_{i1} - \hat{\beta}_2 \sum_{i=1}^n Y_i X_{i2}$. Avec les données précédentes on obtient :

$$\begin{aligned} \sum_{i=1}^n e_i^2 &= 185.6649 \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^n e_i^2}{12} = 15.47207 \end{aligned}$$

Coefficient de Détermination :

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{185.6649}{66042.269} = 0.99718$$

Coefficient de Détermination Ajusté :

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \left[\frac{n-1}{n-(m+1)} \right] = 0.99971$$

Exercice 3

Est-il possible d'obtenir les résultats ci-dessous à partir d'un ensemble de données ?

- (a) $r_{12} = 0.9$, $r_{Y2} = -0.2$, $r_{Y1} = 0.8$
- (b) $r_{Y1} = 0.6$, $r_{12} = -0.9$, $r_{2Y} = -0.5$
- (c) $r_{1Y} = 0.01$, $r_{Y2} = 0.66$, $r_{12} = -0.7$

3 Modèle de Régression Linéaire Multiple

Solution

(a) **Non.** Car si on remplaçant les données dans la formule de R^2 ;

$$R^2 = \frac{r_{Y1}^2 + r_{Y2}^2 - 2r_{Y1}r_{Y2}r_{12}}{1 - r_{12}^2}$$

$$\text{on trouve } R^2 = \frac{(0.8)^2 + (-0.2)^2 - 2(0.8)(-0.2)(0.9)}{1 - (0.9)^2} = 5.09473,$$

ce qui est logiquement impossible, car $0 < R^2 < 1$.

(b) **Oui.** Car lorsqu'on remplace les données dans la formule de R^2 on trouve : $R^2 = 0.3684210$, ce qui est possible.

(c) **Oui.** Car lorsqu'on remplace les données dans la formule de R^2 on trouve : $R^2 = 0.8724311$, ce qui est possible.

Exercice 4

Considérez le modèle suivant :

$$Y_i = \beta_0 + \beta_1 Education + \beta_2 Années d'Expérience + \varepsilon_i$$

Supposons que vous omettiez la variable “Années d’Expérience”. Quels types de problèmes ou de biais attendriez-vous ? Expliquez verbalement.

Solution

Si vous omettez la variable “Années d’Expérience” X_2 du modèle, le coefficient d’éducation X_1 sera biaisé, la nature du biais dépendant de la corrélation entre X_1 et X_2 . La somme des carrés des résidus, et R^2 seront tous affectés à la suite de cette omission. Ceci est un exemple du biais de la variable omise.

Exercice 5

Considérez le modèle de régression linéaire à deux variables explicatives.

(a) Supposons que vous multipliez toutes les valeurs de X_1 par 2.

Quel sera l’effet de ce redimensionnement, le cas échéant, sur les estimations des paramètres et leurs erreurs standard ?

(b) Maintenant, au lieu de (a), supposons que vous multipliez toutes les valeurs de Y par 2. Quel en sera l’effet, le cas échéant, sur les paramètres estimés et leurs erreurs standard ?

3 Modèle de Régression Linéaire Multiple

Solution

- (a) et (b) Si vous multipliez X_1 par 2, vous pouvez vérifier à l'aide des équations (3.101) et (3.102) que les pentes ne sont pas affectées. D'autre part, si vous multipliez Y par 2, les pentes ainsi que les constantes et leurs erreurs sont tous multipliés par 2. Gardez toujours à l'esprit les unités dans lesquelles les régressions et les régresseurs sont mesurés.

Exercice 6

Considérez les modèles suivants.

$$\text{Modèle A : } Y_i = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \varepsilon_{i1}$$

$$\text{Modèle B : } (Y_i - X_{i1}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_{i2}$$

- (a) Les estimateurs des MCO de α_0 et β_0 seront-ils les mêmes ? Pourquoi ?
(b) Les estimateurs des MCO de α_2 et β_2 seront-ils les mêmes ? Pourquoi ?
(c) Quelle est la relation entre α_1 et β_1 ?
(d) Pouvez-vous comparer les R^2 des deux modèles ? Pourquoi ?

Solution

- (a) Réécrire le Modèle B comme :

$$\begin{aligned} Y_i &= \beta_0 + (1 + \beta_1) X_{i1} + \beta_2 X_{i2} + \varepsilon_{i2} \\ &= \beta_0 + \beta'_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_{i2} \quad \text{où } \beta'_1 = (1 + \beta_1) \end{aligned}$$

Par conséquent, les deux modèles sont similaires. Oui, les constatations dans les deux modèles sont les mêmes.

- (b) Les estimations MCO du coefficient de la variable X_2 dans les deux modèles seront les mêmes.
(c) $\beta'_1 = (1 + \beta_1) = \alpha_2$
(d) Non, car les variables dépendantes dans les deux modèles sont différentes.

Exercice 7

Supposons que vous estimiez la fonction de consommation

$$Y_i = \alpha_0 + \alpha_1 X_i + \varepsilon_{i1}$$

3 Modèle de Régression Linéaire Multiple

et la fonction d'épargne

$$Z_i = \beta_0 + \beta_1 X_i + \varepsilon_{i2}$$

où Y = Consommation, Z = Épargnes, X = Revenus et $X = Y + Z$, que le Revenu est égal à la Consommation plus les Épargnes.

- (a) Quelle est la relation, le cas échéant, entre α_1 et β_1 ? Montrer les calculs.
- (b) La somme des carrés des résidus, SCR , sera-t-elle la même pour les deux modèles? Expliquer.
- (c) Pouvez-vous comparer les R^2 des deux modèles? Pourquoi?

Solution

- (a) Selon les MCO, on obtient :

$$\begin{aligned}\hat{\alpha}_1 &= \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^n ((X_i + Z_i) - (\bar{X} + \bar{Z}))(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}\end{aligned}$$

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} - \frac{\sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = 1 - \hat{\beta}_1$$

En d'autres termes, la pente de la régression de l'épargne sur le revenu β_1 (c'est-à-dire la propension marginale à épargner) correspond à (1 – la pente de la régression de la consommation sur le revenu α_1). (c'est-à-dire la propension marginale à consommer). Autrement dit, la somme des deux propensions marginales est égale à 1, ce qui devrait être le cas car le revenu total égal au total des dépenses de consommation et de l'épargne totale. Incidemment, notez que $\hat{\alpha}_0 = -\hat{\beta}_0$.

- (b) **Oui.** La somme des carrés des résidus, SCR pour la fonction de consommation est : $\sum_{i=1}^n (Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 X_i)^2$

Remplacez maintenant $(X_i - Y_i)$ par Z_i , $\hat{\alpha}_0 = -\hat{\beta}_0$ et $\hat{\alpha}_1 = (1 - \hat{\beta}_1)$

3 Modèle de Régression Linéaire Multiple

et vérifiez que les deux SCR sont identiques.

- (c) **Non**, car les variables dépendantes dans les deux modèles sont différentes.

Exercice 8

Régression à travers l'origine. Considérons la régression suivante à travers l'origine :

$$Y_i = \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + e_i$$

- (a) Comment feriez-vous pour estimer les inconnues ?
(b) Est-ce que $\sum_{i=1}^n e_i$ sera zéro pour ce modèle ? Pourquoi ?
(c) Est-ce que $\sum_{i=1}^n e_i X_{i1} = \sum_{i=1}^n e_i X_{i2} = 0$ pour ce modèle ?
(d) Quand utiliseriez-vous un tel modèle ?

Solution

- (a) Les équations normales seront :

$$\begin{aligned}\sum_{i=1}^n Y_i X_{i1} &= \hat{\beta}_1 \sum_{i=1}^n X_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n X_{i1} X_{i2} \\ \sum_{i=1}^n Y_i X_{i2} &= \hat{\beta}_1 \sum_{i=1}^n X_{i1} X_{i2} + \hat{\beta}_2 \sum_{i=1}^n X_{i2}^2\end{aligned}$$

- (b) Non, pour la même raison que le cas d'une seule variable explicative.
(c) Oui, ces conditions sont toujours valables.
(d) Cela dépendra de la théorie sous-jacente.

Exercice 9

La demande de roses. Le tableau (3.7) présente des données trimestrielles sur ces variables :

Y = quantité de roses vendues, douzaines

X_1 = prix moyen de vente en gros des roses, en Dhs/douzaine

X_2 = prix moyen de vente en gros des œillets, en Dhs/douzaine

X_3 = revenu moyen familial hebdomadaire disponible, Dhs/semaine

X_4 = la variable tendancielle prenant les valeurs 1, 2, etc., pour la période 2014–III à 2018–II.

3 Modèle de Régression Linéaire Multiple

Vous êtes invité à prendre en compte les fonctions de demande suivantes :

$$Y_t = \alpha_0 + \alpha_1 X_{t1} + \alpha_2 X_{t2} + \alpha_3 X_{t3} + \alpha_4 X_{t4} + \varepsilon_t$$

$$\ln Y_t = \beta_0 + \beta_1 \ln X_{t1} + \beta_2 \ln X_{t2} + \beta_3 \ln X_{t3} + \beta_4 \ln X_{t4} + \varepsilon_t$$

- (a) Estimer les paramètres du modèle linéaire et interpréter les résultats.
- (b) Estimer les paramètres du modèle log-linéaire et interpréter les résultats.
- (c) β_1 , β_2 et β_3 donnent respectivement les élasticités de la demande par rapport au prix propres, aux prix croisés et au revenu. Quels sont leurs signes a priori ? Les résultats correspondent-ils aux attentes a priori ?
- (d) Comment calculez-vous les élasticités prix propres, prix croisés et revenu du modèle linéaire ?
- (e) Sur la base de votre analyse, quel modèle choisiriez-vous et pourquoi ?

Année et Trimestre	Y	X_1	X_2	X_3	X_4
2014-III	11484	2.26	3.49	158.11	1
-IV	9348	2.54	2.85	173.36	2
2015-I	8429	3.07	4.06	165.26	3
-II	10079	2.91	3.64	172.92	4
-III	9240	2.73	3.21	178.46	5
-IV	8862	2.77	3.66	198.62	6
2016-I	6216	3.59	3.76	186.28	7
-II	8253	3.23	3.49	188.98	8
-III	8038	2.6	3.13	180.49	9
-IV	7476	2.89	3.2	183.33	10
2017-I	5911	3.77	3.65	181.87	11
-II	7950	3.64	3.6	185	12
-III	6134	2.82	2.94	184	13
-IV	5868	2.96	3.12	188.2	14
2018-I	3160	4.24	3.58	175.67	15
-II	5872	3.69	3.53	188	16

TABLE 3.7 – **Données**

3 Modèle de Régression Linéaire Multiple

Solution

(a) Modèle Linéaire

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10816.04	5988.348	1.8061	0.0983
X1	-2227.704	920.4657	-2.4201	0.0340
X2	1251.141	1157.021	1.0813	0.3027
X3	6.282986	30.62166	0.2051	0.8412
X4	-197.3999	101.5612	-1.9436	0.0780
R-squared	0.834699	Mean dependent var		7645
Adjusted R-squared	0.774590	S.D. dependent var		2042.814
S.E. of regression	969.8744	Akaike info criterion		16.842
Sum squared resid	10347220	Schwarz criterion		17.083

$$\hat{Y}_t = 10816.04 - 2227.704X_{t1} + 1251.141X_{t2} + 6.283X_{t3} - 197.399X_{t4}$$

Dans ce modèle, les coefficients de pente mesurent le taux de variation de Y par rapport à la variable considérée.

(b) Modèle Log-Linéaire

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.626824	6.148262	0.101951	0.9206
LOG(X1)	-1.273555	0.526649	-2.418224	0.0341
LOG(X2)	0.937305	0.659191	1.421902	0.1828
LOG(X3)	1.712976	1.200843	1.426478	0.1815
LOG(X4)	-0.181597	0.127893	-1.419907	0.1833
R-squared	0.777953	Mean dependent var		8.902209
Adjusted R-squared	0.697208	S.D. dependent var		0.306877
S.E. of regression	0.168864	Akaike info criterion		-0.469145
Sum squared resid	0.313664	Schwarz criterion		-0.227711

$$\ln \hat{Y}_t = 0.627 - 1.274 \ln X_{t1} + 0.937 \ln X_{t2} + 1.713 \ln X_{t3} - 0.182 \ln X_{t4}$$

Dans ce modèle, tous les coefficients de pente partielle sont des élasticités partielles de Y par rapport à la variable considérée.

- (c) L'élasticité-prix propre devrait être négative, l'élasticité croisée des prix devrait être positive pour les produits de substitution et négative pour les produits complémentaires, et l'élasticité des revenus devrait être positive, car les roses sont des produits normaux.

3 Modèle de Régression Linéaire Multiple

- (d) La formule générale d'élasticité pour une équation linéaire est la suivante :

$$\text{Élasticité} = \frac{\partial Y}{\partial X_i} \frac{\bar{X}_i}{\bar{Y}}, \text{ où } X_i \text{ est le régresseur en question.}$$

C'est-à-dire que pour un modèle linéaire, l'élasticité peut être calculée aux valeurs moyennes.

- (e) Les deux modèles donnent des résultats similaires. Un des avantages du modèle log-linéaire est que les coefficients de pente donnent des estimations directes de l'élasticité (constante) de la variable pertinente par rapport au régresseur considéré. Mais gardez à l'esprit que les R^2 's des deux modèles ne sont pas directement comparables.

Exercice 10

La demande de poulet, 1990-2012. Pour étudier la consommation de poulet par habitant, on a les données du tableau (3.8), où Y = consommation réelle par habitant de poulets, en kg

X_1 = revenu réel disponible par habitant, en dhs

X_2 = prix réel du poulet au kg,

X_3 = prix réel de la dinde au kg,

X_4 = prix réel du boeuf au kg,

X_5 = prix réel composé des substituts de poulet par kg, ce qui correspond à la moyenne pondérée des prix au kg de la dinde et du bœuf, le poids étant la consommation relative du bœuf et de la dinde dans la consommation totale du bœuf et de la dinde.

Considérons maintenant les fonctions de demande suivantes :

$$\ln Y_t = \alpha_0 + \alpha_1 \ln X_{t1} + \alpha_2 \ln X_{t2} + \varepsilon_t \quad (1)$$

$$\ln Y_t = \gamma_0 + \gamma_1 \ln X_{t1} + \gamma_2 \ln X_{t2} + \gamma_3 \ln X_{t3} + \varepsilon_t \quad (2)$$

$$\ln Y_t = \lambda_0 + \lambda_1 \ln X_{t1} + \lambda_2 \ln X_{t2} + \lambda_3 \ln X_{t4} + \varepsilon_t \quad (3)$$

$$\ln Y_t = \theta_0 + \theta_1 \ln X_{t1} + \theta_2 \ln X_{t2} + \theta_3 \ln X_{t3} + \theta_4 \ln X_{t4} + \varepsilon_t \quad (4)$$

$$\ln Y_t = \beta_0 + \beta_1 \ln X_{t1} + \beta_2 \ln X_{t2} + \beta_3 \ln X_{t5} + \varepsilon_t \quad (5)$$

3 Modèle de Régression Linéaire Multiple

Date	Y	X_1	X_2	X_3	X_4	X_5
1990	27.8	397.5	42.2	50.7	78.3	65.8
1991	29.9	413.3	38.1	52	79.2	66.9
1992	29.8	439.2	40.3	54	79.2	67.8
1993	30.8	459.7	39.5	55.3	79.2	69.6
1994	31.2	492.9	37.3	54.7	77.4	68.7
1995	33.3	528.6	38.1	63.7	80.2	73.6
1996	35.6	560.3	39.3	69.8	80.4	76.3
1997	36.4	624.6	37.8	65.9	83.9	77.2
1998	36.7	666.4	38.4	64.5	85.5	78.1
1999	38.4	717.8	40.1	70	93.7	84.7
2000	40.4	768.2	38.6	73.2	106.1	93.3
2001	40.3	843.3	39.8	67.8	104.8	89.7
2002	41.8	911.6	39.7	79.1	114	100.7
2003	40.4	931.1	52.1	95.4	124.1	113.5
2004	40.7	1021.5	48.9	94.2	127.6	115.3
2005	40.1	1165.9	58.3	123.5	142.9	136.7
2006	42.7	1349.6	57.9	129.9	143.6	139.2
2007	44.1	1449.4	56.5	117.6	139.2	132
2008	46.7	1575.5	63.7	130.9	165.5	132.1
2009	50.6	1759.1	61.6	129.8	203.3	154.4
2010	50.1	1994.2	58.9	128	219.6	174.9
2011	51.7	2258.1	66.4	141	221.6	180.8
2012	52.9	2478.7	70.4	168.2	232.6	189.4

TABLE 3.8 – **Données**

Selon la théorie microéconomique, la demande d'un produit dépend généralement du revenu réel du consommateur, du prix réel du produit et des prix réels des produits concurrents ou complémentaires. Compte tenu de ces considérations, répondez aux questions suivantes.

- (a) Quelle fonction de demande parmi celles données ici choisiriez-vous et pourquoi ?
- (b) Comment interpréteriez-vous les coefficients de $\ln X_{t1}$ et $\ln X_{t2}$ dans ces modèles ?
- (c) Quelle est la différence entre les spécifications (2) et (4) ?
- (d) Quels problèmes prévoyez-vous si vous adoptez la spécification (4) ? (Noter : les prix du bœuf et de la dinde sont inclus avec le prix du poulet.)
- (e) Étant donné que la spécification (5) comprend le prix composite

3 Modèle de Régression Linéaire Multiple

- du bœuf et de la dinde, préféreriez-vous la fonction demande (5) à la fonction (4) ? Pourquoi ?
- (f) La dinde et/ou le bœuf sont-ils des concurrents ou des produits de substitution au poulet ? Comment le sais-tu ?
- (g) On suppose que la fonction (5) est la fonction de demande «correcte». Estimez les paramètres de ce modèle, obtenez leurs erreurs-types et les coefficients R^2 , \bar{R}^2 et R^2 modifié. Interprétez vos résultats.
- (h) Supposons maintenant que vous exécutez le modèle «incorrect» (2). Évaluez les conséquences de cette erreur de spécification en considérant les valeurs de γ_1 et γ_2 par rapport à β_1 et β_2 , respectivement.

Solution

- (a) Le modèle (5) semble être le meilleur car il inclut toutes les variables économiquement pertinentes, y compris le prix réel composite des substituts de poulet, ce qui devrait contribuer à atténuer le problème de multicolinéarité qui peut exister dans le modèle (4) entre le prix du bœuf et prix de la dinde. Le modèle (1) ne contient pas de bonne information de substitution et les modèles (2) et (3) ont une bonne information de substitution.
- (b) Le coefficient de $\ln X_{t1}$ représente l'élasticité du revenu ; le coefficient de $\ln X_{t2}$ représente l'élasticité par rapport au prix.
- (c) Le modèle (2) considère uniquement la dinde comme produit de substitution, tandis que le modèle (4) considère à la fois la dinde et le bœuf.
- (d) Il peut y avoir un problème de multicolinéarité entre le prix du bœuf et le prix de la dinde.
- (e) Oui. Cela pourrait atténuer le problème de la multicolinéarité.
- (f) Ils doivent être des produits de substitution car ils font concurrence au poulet en tant que produit de consommation alimentaire.
- (g) Résultat de l'estimation

3 Modèle de Régression Linéaire Multiple

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.029865	0.118682	17.10338	0.0000
LOG(X1)	0.481286	0.068188	7.058251	0.0000
LOG(X2)	-0.350628	0.079394	-4.416310	0.0003
LOG(X5)	-0.061035	0.129960	-0.469645	0.6440
R-squared	0.980303	Mean dependent var		3.663887
Adjusted R-squared	0.977193	S.D. dependent var		0.187659
S.E. of regression	0.028340	Akaike info criterion		-4.132296
Sum squared resid	0.015260	Schwarz criterion		-3.934819

$$\ln \hat{Y}_t = 2.030 + 0.481 \ln X_{t2} - 0.35 \ln X_{t3} - 0.061 \ln X_{t6}$$

$$\hat{\sigma} = (0.119) \quad (0.068) \quad (0.079) \quad (0.130)$$

$$R^2 = 0.980 \quad \bar{R}^2 = 0.977 \quad R^2 \text{ ajusté} = 0.810$$

L'élasticité du revenu et l'élasticité propre du prix ont les bons signes.

- (h) L'estimation du modèle (2) aurait pour conséquence que les estimateurs risquent d'être biaisés en raison d'une mauvaise spécification du modèle.

Exercice 11

Reprendons la fonction de demande de roses de l'exercice 9. Limiter vos considérations à la spécification logarithmique,

- (a) Quelle est l'élasticité estimée de la demande par rapport au prix (par exemple, l'élasticité par rapport au prix des roses) ?
- (b) Est-ce qu'elle est statistiquement significative (c.-à.-d. significativement différent de zéro) ?
- (c) Si oui, est-ce qu'elle significativement différente de (-1) ?
- (d) A priori, quels sont les signes attendus de X_2 (prix des œillets) et de X_3 (revenu) ? Les résultats empiriques sont-ils en accord avec ces attentes ?
- (e) Si les coefficients de X_2 et X_3 sont statistiquement non significatifs, quelles peuvent en être les raisons ?

Solution

- (a) L'élasticité-prix propre est de -1.274 (voir le tableau des résultats)

3 Modèle de Régression Linéaire Multiple

- (b) Calcul du t^* empirique de Student

$$t_{\hat{\beta}_1}^* = \frac{-1.274 - 0}{0.527} = -2.4174$$

Enfin $|t_{\hat{\beta}_1}^*| > t_{n-5}^{\alpha/2} = t_{11}^{0,025} = 2.201$ la pente β_1 est significativement différente de 0.

Encore, la valeur p d'avoir une telle statistique t sous l'hypothèse nulle est d'environ 0.034 (voir le tableau des résultats), ce qui est faible. Par conséquent, nous rejetons l'hypothèse selon laquelle l'élasticité des prix est nulle.

- (c) Encore une fois, en utilisant la formule standard, nous obtenons :

$$t_{\hat{\beta}_1}^* = \frac{-1.274 - (-1)}{0.527} = -0.5199$$

Enfin $|t_{\hat{\beta}_1}^*| < t_{n-5}^{\alpha/2} = t_{11}^{0,025} = 2.201$ la pente β_1 n'est pas significativement différente de (-1) .

- (d) Les deux signes devraient être positifs, bien qu'aucune de ces variables ne soit statistiquement significative.
- (e) Notre échantillon est peut-être trop petit pour détecter l'importance statistique des prix à la consommation sur la demande de roses ou du revenu sur la demande de roses. De plus, les dépenses en roses peuvent représenter une si petite partie du revenu total que l'on peut ne pas remarquer l'impact du revenu sur la demande en roses.

Exercice 12

Soit le modèle ci-dessous à trois variables explicatives :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

Nous disposons des données du tableau (3.9).

- (1) Mettre le modèle sous forme matricielle puis estimer les paramètres du modèle.
- (2) Calculer l'estimation de la variance de l'erreur ainsi que les écarts types de chacun des coefficients.
- (3) Calculer le R^2 et le R^2 corrigé.
- (4) Les variables explicatives sont-elles significativement contributives pour expliquer la variable endogène ?

3 Modèle de Régression Linéaire Multiple

- (5) Le coefficient β_1 est-il significativement inférieur à 1 ?
- (6) Les coefficients β_1 et β_2 sont-ils simultanément et significativement différents de 1 et -0.5 ?
- (7) Quel est l'intervalle de confiance pour la variance de l'erreur ?

i	Y	X_1	X_2	X_3
1	15	4	35	96
2	17	3	33	107
3	13	5	33	129
4	19	8	37	120
5	17	9	32	104
6	22	10	31	131
7	24	10	22	107
8	22	7	23	122
9	24	7	31	103
10	19	10	28	138
11	22	6	22	136
12	24	11	21	147
13	28	14	25	149
14	24	9	19	155

TABLE 3.9 – **Données**

Solution

(1)

$$X = \begin{pmatrix} 1 & 4 & 35 & 96 \\ 1 & 3 & 33 & 107 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 9 & 19 & 155 \end{pmatrix} \quad Y = \begin{pmatrix} 15 \\ 17 \\ \vdots \\ 24 \end{pmatrix}$$

En appliquant la méthode présentée dans l'exemple de RBAG Studios dans la section 3.11

on obtient les résultats suivants :

Les estimations des moindres carrés $\hat{\beta} : \hat{\beta} = (X'X)^{-1} X'Y$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 29.545 \\ 0.8019 \\ -0.3813 \\ -0.0371 \end{pmatrix}$$

3 Modèle de Régression Linéaire Multiple

et la fonction de régression estimée :

$$\hat{Y} = 29.545 + 0.8019X_1 - 0.3813X_2 - 0.0371X_3$$

(2) Valeurs Estimées et Résidus

Pour examiner la pertinence du modèle de régression pour les données disponibles, nous avons besoin des valeurs ajustées \hat{Y}_i et des résidus $e_i = Y_i - \hat{Y}_i$. On a :

$$\hat{Y} = X\hat{\beta}$$

$$e = Y - \hat{Y}$$

Analyse de la Variance

$$SCT = Y'Y - \left(\frac{1}{n}\right)Y'JY = 226.85745$$

$$SCR = e'e = Y'Y - \hat{\beta}'X'Y = 67.44767$$

Enfin, on déduit :

$$SCE = SCT - SCR = 159.40978$$

Estimation des variances des Paramètres de Régression

La matrice de variance-covariance estimée $\hat{\sigma}^2(\hat{\beta})$:

$$\hat{\sigma}^2(\hat{\beta}) = MCR(X'X)^{-1}$$

$$\text{avec } MCR = \frac{SCR}{n - m - 1} = \frac{67.44767}{14 - 3 - 1} = 6.744767$$

$$\begin{aligned} \hat{\sigma}^2(\hat{\beta}) &= 6.744767 \begin{pmatrix} 12.6717 & -0.0229 & -0.1831 & -0.0585 \\ -0.0229 & 0.0132 & 0.0012 & -0.0009 \\ -0.1831 & 0.0012 & 0.0036 & 0.0006 \\ -0.0585 & -0.0009 & 0.0006 & 0.0004 \end{pmatrix} \\ &= \begin{pmatrix} 85.4677 & -0.1545 & -1.2350 & -0.3946 \\ -0.1545 & 0.0891 & 0.0081 & -0.0063 \\ -1.2350 & 0.0081 & 0.0245 & 0.0039 \\ -0.3946 & -0.0063 & 0.0039 & 0.0027 \end{pmatrix} \end{aligned}$$

3 Modèle de Régression Linéaire Multiple

Les deux variances estimées dont nous avons besoin sont :

$$\hat{\sigma}^2(\hat{\beta}_0) = \hat{\sigma}_{\hat{\beta}_0}^2 = 85.4677 \quad \text{ou} \quad \hat{\sigma}(\hat{\beta}_0) = 9.24487$$

$$\hat{\sigma}^2(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}^2 = 0.0891 \quad \text{ou} \quad \hat{\sigma}(\hat{\beta}_1) = 0.29849$$

$$\hat{\sigma}^2(\hat{\beta}_2) = \hat{\sigma}_{\hat{\beta}_2}^2 = 0.0245 \quad \text{ou} \quad \hat{\sigma}(\hat{\beta}_2) = 0.15652$$

$$\hat{\sigma}^2(\hat{\beta}_3) = \hat{\sigma}_{\hat{\beta}_3}^2 = 0.0027 \quad \text{ou} \quad \hat{\sigma}(\hat{\beta}_3) = 0.05193$$

(3) Coefficient de Détermination Multiple

Pour notre exemple on a

$$R^2 = \frac{SCE}{SCT} = \frac{159.40978}{226.85745} = 0.7026869$$

Coefficient de Détermination Ajusté :

$$\bar{R}^2 = 1 - (1 - R^2) \left[\frac{n - 1}{n - (m + 1)} \right] = 0.613492$$

(4) On calcule les trois ratios de Student pour les comparer à la valeur lue dans la table pour un seuil de 5%

$|t_{\hat{\beta}_1}^*| = \left| \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right| = \left| \frac{0.8019}{0.29849} \right| = 2.6865 > t_{10}^{0.025} = 2.228 \rightarrow \beta_1 \neq 0$, la variable explicative X_1 est contributive à l'explication de Y .

$|t_{\hat{\beta}_2}^*| = \left| \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} \right| = \left| \frac{-0.3813}{0.15652} \right| = 2.4361 > t_{10}^{0.025} = 2.228 \rightarrow \beta_2 \neq 0$, la variable explicative X_2 est contributive à l'explication de Y .

$|t_{\hat{\beta}_3}^*| = \left| \frac{\hat{\beta}_3}{\hat{\sigma}_{\hat{\beta}_3}} \right| = \left| \frac{-0.0371}{0.05193} \right| = 0.7144 < t_{10}^{0.025} = 2.228 \rightarrow \beta_3 = 0$, la variable explicative X_3 n'est pas contributive à l'explication de Y .

(5) Il s'agit d'un Test Unilatéral de la pente $\hat{\beta}_1$:

$$\begin{aligned} H_0; \beta_1 &= 1 \\ H_1; \beta_1 &> 1 \end{aligned}$$

$$\text{ainsi } t_{\hat{\beta}_1}^* = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0.8019 - 1}{0.29849} = -0.66367$$

3 Modèle de Régression Linéaire Multiple

Enfin $|t_{\hat{\beta}_1}^*| < t_{n-m-1}^\alpha = t_{10}^{0.05} = 1.812$. La pente β_1 n'est pas significativement supérieur à 1.

- (6) Le test d'hypothèse est le suivant :

$$\begin{aligned} H_0: \quad \beta_1 = 1, \quad \beta_2 = -0.5 \\ H_1: \quad \beta_1 \neq 1, \quad \beta_2 \neq -0.5 \end{aligned}$$

Ici, le modèle réduit serait :

$$(R): Y_i - X_{i1} + 0.5X_{i2} = \beta_0 + \beta_3X_{i3} + \varepsilon_i \quad \text{Modèle réduit}$$

Notez la nouvelle variable dépendante ($Y - X_1 + 0.5X_2$) dans le modèle réduit, puisque β_1X_1 et β_2X_2 sont des constantes connues sous H_0 . Nous utilisons ensuite la statistique de test linéaire générale F^* de Fisher présenté dans l'équation (3.55) avec 2 et $n-4$ degrés de liberté.

$$F^* = \frac{\frac{SCR(R) - SCR(C)}{dll(R) - dll(C)}}{\frac{SCR(C)}{dll(C)}} = \frac{\frac{SCR(R) - SCR(C)}{2}}{\frac{SCR(C)}{n-4}}$$

$$F^* = \frac{\frac{75.7066 - 67.4476}{2}}{\frac{67.4476}{10}} = 0.61225 < F_{2;10}^{0.05} = 4.10$$

Alors, on accepte l'hypothèse H_0 . Les données ne sont pas compatibles avec la possibilité que les coefficients β_1 et β_2 soient simultanément et respectivement égaux à 1 et -0.5.

- (7) L'intervalle de confiance de la variance de l'erreur à un seuil $(1 - \alpha)\% = 95\%$ ($\alpha = 0.05$) est calculé pour 10 degrés de liberté :

$$\begin{aligned} IC &= \left[\frac{(n-m-1)\hat{\sigma}^2}{\chi_{0.025}^2}; \frac{(n-m-1)\hat{\sigma}^2}{\chi_{0.975}^2} \right] \\ &= \left[\frac{10 \times 6.744767}{20.48}; \frac{10 \times 6.744767}{3.25} \right] = [3.2933; 20.7531] \end{aligned}$$

3 Modèle de Régression Linéaire Multiple

Exercice 13

Reprenons les données du tableau (3.9) de l'exercice 12, dont nous rappelons les résultats de l'estimation du modèle :

$$\begin{aligned} Y_i &= 29.545 + 0.8019X_{i1} - 0.3813X_{i2} - 0.0371X_{i3} + e_i \\ &\quad (9.2447) \quad (0.2985) \quad (0.1565) \quad (0.0519) \\ R^2 &= 0.7026 \\ n &= 14 \end{aligned}$$

- (1) L'ajout des variables explicatives X_2 et X_3 améliore-t-il significativement la qualité de l'estimation par rapport à X_1 seul ?
- (2) Peut-on considérer le modèle (à trois variables explicatives) comme stable sur l'ensemble de la période, ou doit-on procéder à deux estimations, l'une de la période 1 à 7, et l'autre de la période 8 à 14 ?
- (3) Un économiste suggère que dans ce modèle $\beta_1 = 1$ et $\beta_2 = \beta_3$, qu'en pensez-vous ?

Solution

On applique tout d'abord le test de Fisher afin de tester la signification globale de la régression à trois variables X_1 , X_2 et X_3 .

$$F^* = \frac{R^2/m}{(1-R^2)/(n-m-1)} = \frac{0.7026/3}{(1-0.7026)/10} = 7.8749 > F_{3;10}^{0.05} = 3.71$$

La régression est globalement significative. L'hypothèse H_0 de nullité de tous les coefficients est rejetée.

- (1) Test d'ajout de variables

1ère étape : Calcul de la variabilité totale, expliquée et résiduelle sur le modèle complet avec les 3 variables explicatives X_1 , X_2 et X_3 .

$$SCT = Y'Y - \left(\frac{1}{n}\right)Y'JY = 226.85745$$

$$SCR = e'e = Y'Y - \hat{\beta}'X'Y = 67.44767$$

$$SCE = SCT - SCR = 159.40978$$

2ème étape : Calcul de la variabilité totale, expliquée et résiduelle sur le modèle à une seule variable explicative X_1 . Le modèle estimé

3 Modèle de Régression Linéaire Multiple

est le suivant :

$$Y_i = 1.0118X_{i1} + 12.5475 + e_i$$

Ecart Type = (0.28138)

$$R^2 = 0.5186$$

$$\hat{\sigma} = 2.9183$$

$$n = 14$$

On clacul d'abord

$$SCR^1 = e' e = (n - 2) \times \hat{\sigma}^2 = 109.1983$$

à partir du coefficient de détermination R^2 , nous déduisons SCR et SCE^{21} :

$$SCT^1 = SCR^1 / (1 - R^2) = 226.8348 \text{ et } SCE^1 = 117.6365$$

Le test d'hypothèses est le suivant :

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_1 : \text{Il existe au moins un des deux coefficients non nul}$$

Ce test se ramène à un test par analyse de la variance : le fait d'ajouter des variables explicatives dans un modèle entraîne automatiquement une augmentation de SCE (et donc une diminution de SCR) ; on souhaite donc tester que la différence entre SCE et SCE^1 soit significativement positive (ou bien que la différence entre SCR^1 et SCR soit significativement positive, il s'agit du même test). On compare donc la différence par rapport à la somme des carrés la plus faible, soit ici SCR . On procède au test de Fisher dans 3ème étape.

3ème étape : Calcul de F^* empirique de Fisher

$$F^* = \frac{\frac{SCE - SCE^1}{m - m^1}}{\frac{SCR}{n - m - 1}} = \frac{\frac{159.4097 - 117.6365}{3 - 1}}{\frac{67.44767}{14 - 3 - 1}} = 3.09 < F_{2;10}^{0.05} = 4.10$$

21. Sauf si la ou les variables ajoutées sont orthogonales à la variable à expliquer, SCE reste alors identique. Ce cas est évidemment rare.

3 Modèle de Régression Linéaire Multiple

ou encore

$$F^* = \frac{\frac{SCR^1 - SCR}{m - m^1}}{\frac{SCR}{n - m - 1}} = \frac{\frac{109.1983 - 67.44767}{3 - 1}}{\frac{67.44767}{14 - 3 - 1}} = 3.09$$

Avec m = nombre de variables explicatives du modèle complet et m^1 = nombre de variables explicatives du modèle sans l'ajout des variables X_2 et X_3 . Nous acceptons l'hypothèse H_0 , il n'y a donc pas de différence significative entre les deux variances expliquées, l'ajout des variables explicatives X_2 et X_3 n'améliore pas de manière significative le pouvoir explicatif du modèle et au seuil de 5%.

(2) Le modèle est-il stable sur la totalité de la période ?

Soit le modèle estimé sur une seule période :

$$Y_i = \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + \hat{\beta}_0 + e_i \quad \text{pour } i = 1, \dots, 14$$

ou le modèle estimé sur deux sous-périodes :

$$Y_i = \hat{\beta}_1^1 X_{i1} + \hat{\beta}_2^1 X_{i2} + \hat{\beta}_3^1 X_{i3} + \hat{\beta}_0^1 + e_i \quad \text{pour } i = 1, \dots, 7$$

$$Y_i = \hat{\beta}_1^2 X_{i1} + \hat{\beta}_2^2 X_{i2} + \hat{\beta}_3^2 X_{i3} + \hat{\beta}_0^2 + e_i \quad \text{pour } i = 8, \dots, 14$$

Le test d'hypothèses jointes est alors le suivant :

$$H_0 : \begin{pmatrix} \beta_0 = \beta_0^1 = \beta_0^2 \\ \beta_1 = \beta_1^1 = \beta_1^2 \\ \beta_2 = \beta_2^1 = \beta_2^2 \\ \beta_3 = \beta_3^1 = \beta_3^2 \end{pmatrix}$$

Ce test de stabilité des coefficients (test de Chow) se ramène à la question suivante : existe-t-il une différence significative entre la somme des carrés des résidus SCR de l'ensemble de la période et l'addition de la somme des carrés des résidus calculée à partir des deux sous-périodes ($SCR^1 + SCR^2$) ?

En effet, dans le cas d'une réponse négative, cela signifie que le fait de scinder en deux échantillons n'améliore pas la qualité du modèle, donc qu'il est stable sur la totalité de la période.

1ère étape : Estimation du modèle sur chacune des deux sous-

3 Modèle de Régression Linéaire Multiple

périodes²² et calcul des sommes des carrés de résidus.

Résultat de la 1ère sous-période d'estimation, n(1, ..., 7)

$$Y_i = 23.483 + 0.774X_{i1} - 0.2932X_{i2} - 0.0125X_{i3} + e_i$$

(9.2447)	(0.5290)	(0.3136)	(0.10)
----------	----------	----------	--------

$$R^2 = 0.69256$$

$$n = 7$$

$$\hat{\sigma} = 3.01759$$

Les sommes des carrés correspondantes sont :

$$SCT^1 = 88.85496, SCR^1 = 27.31757, SCE^1 = 61.537393$$

Résultat de la 2ème sous-période d'estimation, n(8, ..., 14)

$$Y_i = 52.062 + 1.228X_{i1} - 0.6208X_{i2} - 0.1843X_{i3} + e_i$$

(27.5295)	(0.6852)	(0.5223)	(0.1528)
-----------	----------	----------	----------

$$R^2 = 0.54385$$

$$n = 7$$

$$\hat{\sigma} = 2.62817$$

Les sommes des carrés correspondantes sont :

$$SCT^2 = 45.42779, SCR^2 = 20.72189, SCE^2 = 24.7059$$

2ème étape : Calcul de la statistique empirique de Fisher

$$F^* = \frac{\frac{SCR - (SCR^1 + SCR^2)}{(n - m - 1) - (n - 2m - 2)}}{\frac{SCR^1 + SCR^2}{n - 2m - 2}} = \frac{\frac{SCR - (SCR^1 + SCR^2)}{m + 1}}{\frac{SCR^1 + SCR^2}{n - 2m - 2}}$$

$$F^* = \frac{\frac{67.44767 - (27.31757 + 20.72189)}{4}}{\frac{27.31757 + 20.72189}{6}} = 0.606 < F_{4;6}^{0.05} = 4.53$$

Nous acceptons l'hypothèse H_0 , ainsi, les coefficients sont significativement stables sur l'ensemble de la période²³.

1. Test de l'hypothèse $\beta_1 = 1$ et $\beta_2 = \beta_3$,

22. Les deux sous-périodes peuvent être de longueur inégale, cependant elles doivent impérativement recouvrir la totalité des observations de la période.

23. **Remarque :** Attention, en cas d'hétéroscédasticité, le test de Chow est biaisé dans le sens d'une surestimation du seuil de rejet du test, nous rejetons trop souvent l'hypothèse H_0 .

3 Modèle de Régression Linéaire Multiple

Si l'hypothèse précédente est vérifiée, le modèle :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i$$

peut s'écrire comme suit :

$$Y_i = \beta_0 + 1X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i$$

ou encore

$$\begin{aligned} Y_t - X_{i1} &= \beta_0 + \beta_2 (X_{i2} + X_{i3}) + e_i \\ Z_i &= \beta_0 + \beta_2 V_i + e_i \end{aligned}$$

Le tableau (3.10) présente les nouvelles variables Z_i et V_i .

i	$Z_i = Y_i - X_{i1}$	$V_i = X_{i2} + X_{i3}$
1	11	131
2	14	140
3	8	162
4	11	157
5	8	136
6	12	162
7	14	129
8	15	145
9	17	134
10	9	166
11	16	158
12	13	168
13	14	174
14	15	174

TABLE 3.10 – **Données des nouvelles variables** Z_i et V_i

L'estimation des deux coefficients du modèle conduit aux résultats

3 Modèle de Régression Linéaire Multiple

suivants :

$$\begin{aligned}
 Z_i &= 14.3447 - 0.01115V_i + e_i \\
 \text{Ecart Type} &= (7.8970) \quad (0.05149) \\
 R^2 &= 0.003896 \\
 n &= 14 \\
 \hat{\sigma} &= 3.0109
 \end{aligned}$$

Les sommes des carrés correspondantes sont :

$$SCT^1 = 109.2142, SCR^1 = 108.7888, SCE^1 = 0.425498$$

Enfin, l'hypothèse à tester est donc :

$$H_0 : \text{les restrictions sont toutes vérifiées } (SCR^1 = SCR)$$

$$H_1 : \text{il existe au moins une restriction non vérifiée } (SCR^1 \neq SCR)$$

$$\begin{aligned}
 F^* &= \frac{\frac{SCR^1 - SCR}{(n - m' - 1) - (n - m - 1)}}{\frac{SCR}{n - m - 1}} = \frac{\frac{SCR^1 - SCR}{m - m'}}{\frac{SCR}{n - m - 1}} \\
 F^* &= \frac{\frac{108.7888 - 67.44767}{2}}{\frac{67.44767}{10}} = 3.06468 < F_{2;10}^{0.05} = 4.10
 \end{aligned}$$

Nous acceptons l'hypothèse H_0 , ainsi, les contraintes envisagées sur les coefficients sont compatibles avec les données.

Exercice 14

On reprend les données de l'exercice 12.

- (1) Estimer le modèle à deux variables explicatives X_1 et X_2 (Car on a montré que la variable X_3 n'est pas significative) ;

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$$

- (2) Calculer la prévision ponctuelle et l'intervalle de prédiction à 95% pour les périodes 15 et 16, sachant que $X_{15,1} = 4$, $X_{16,1} = 7$ et $X_{15,2} = 27$, $X_{16,2} = 40$

3 Modèle de Régression Linéaire Multiple

Solution

- (1) Résultats de l'estimation du modèle à deux variables explicatives X_1 et X_2 :

$$Y_i = 24.1312 + 0.7149X_{i1} - 0.3281X_{i2} + e_i$$

Ecart Type = (5.1653) (0.2662) (0.1345)

$$R^2 = 0.68754$$

$$n = 14$$

$$\hat{\sigma} = 2.5385$$

On peut remarquer que les t de Student sont supérieurs à $t_{11}^{0.025} = 2.201$, les coefficients β_1 et β_2 sont significativement différents de 0.

- (2) Les prévisions ponctuelles estimées pour les deux périodes 15 et 16 sont directement calculées à partir du modèle estimé :

Pour la période 15 :

$$\begin{aligned}\widehat{Y}_{15} &= X'_{15}\widehat{\beta} = \begin{pmatrix} 1 & X_{15,1} & X_{15,2} \end{pmatrix} \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 4 & 27 \end{pmatrix} \begin{pmatrix} 24.1312 \\ 0.7149 \\ -0.3281 \end{pmatrix} = 18.1321\end{aligned}$$

Pour la période 16 :

$$\begin{aligned}\widehat{Y}_{16} &= X'_{16}\widehat{\beta} = \begin{pmatrix} 1 & X_{16,1} & X_{16,2} \end{pmatrix} \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 7 & 40 \end{pmatrix} \begin{pmatrix} 24.1312 \\ 0.7149 \\ -0.3281 \end{pmatrix} = 16.011\end{aligned}$$

Intervalles de prédiction :

Les variances estimées sont les suivantes :

$$\hat{\sigma}^2(\widehat{Y}_h) = X'_h \hat{\sigma}^2(\widehat{\beta}) X_h = \hat{\sigma}^2 X'_h (X' X)^{-1} X_h$$

3 Modèle de Régression Linéaire Multiple

$$\begin{aligned}\hat{\sigma}^2(\hat{Y}_{15}) &= (2.5385)^2 \times (1 \ 4 \ 27) \begin{pmatrix} 4.1404 & -0.16 & -0.0992 \\ -0.16 & 0.011 & 0.0025 \\ -0.0992 & 0.0025 & 0.0028 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \\ 27 \end{pmatrix} \\ &= 1.7870\end{aligned}$$

$$\begin{aligned}\hat{\sigma}^2(\hat{Y}_{16}) &= (2.5385)^2 \times (1 \ 7 \ 40) \begin{pmatrix} 4.1404 & -0.16 & -0.0992 \\ -0.16 & 0.011 & 0.0025 \\ -0.0992 & 0.0025 & 0.0028 \end{pmatrix} \begin{pmatrix} 1 \\ 7 \\ 40 \end{pmatrix} \\ &= 2.7277\end{aligned}$$

L'intervalle de prédiction pour la période 15 est :

$$Y_{15(nouvelle)} = \left[\hat{Y}_{15} \pm t_{n-3}^{0.025} \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2(\hat{Y}_{15})} \right]$$

$$\begin{aligned}Y_{15(nouvelle)} &= \left[18.1321 \pm 2.201 \sqrt{6.4439 + 1.7870} \right] \\ &= [11.8175; 24.4466]\end{aligned}$$

L'intervalle de prédiction pour la période 16 est :

$$Y_{16(nouvelle)} = \left[\hat{Y}_{16} \pm t_{n-3}^{0.025} \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2(\hat{Y}_{16})} \right]$$

$$\begin{aligned}Y_{16(nouvelle)} &= \left[16.011 \pm 2.201 \sqrt{6.4439 + 2.7277} \right] \\ &= [9.3453; 22.6766]\end{aligned}$$

On remarque que les intervalles de la prévision semblent assez larges, cependant il convient de souligner que la distribution de probabilité suit une loi de Student et qu'à ce titre la valeur la plus probable demeure la valeur centrale (la prévision estimée) et que la probabilité d'apparition diminue lorsque l'on s'éloigne de cette valeur centrale.

Exercice 15

On reprend les données de l'exercice 12. On considère le modèle de régression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$$

3 Modèle de Régression Linéaire Multiple

- (1) On cherche à partir d'une régression récursive de tester un éventuel changement structurel en examinant la stabilité des coefficients et en procédant aux tests CUSUM.
- (2) Etablir le test de spécifications de Ramsey.

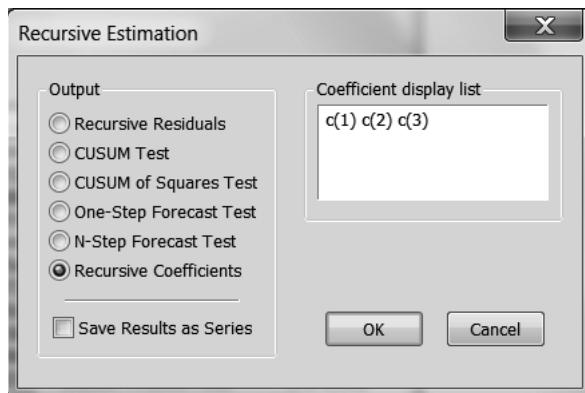
Solution

- (1) Le test d'un éventuel changement structurel, partant d'une estimation récursive (test de la stabilité des coefficients : tests de CUSUM).

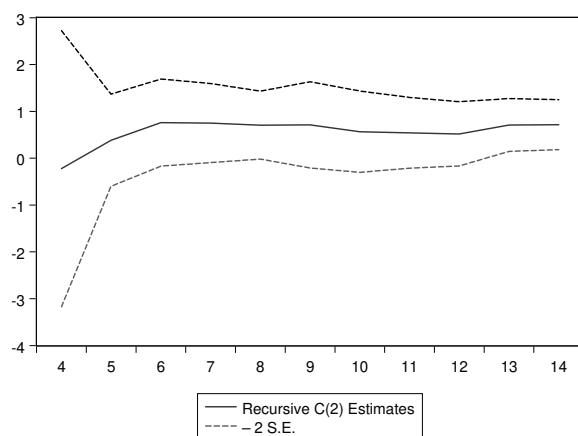
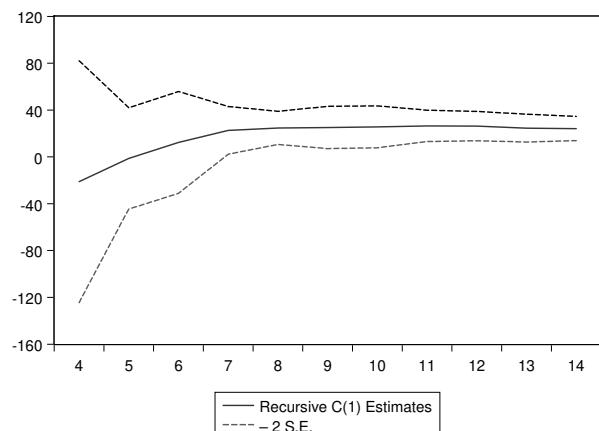
Sur **EViews**, faire : $Y c X1 X2$

Dans l'output des résultats, suivre : View/Stability Tests/Recursive Estimates (OLS only)... : la figure ci-dessous complète la procédure et les résultats des instructions données :

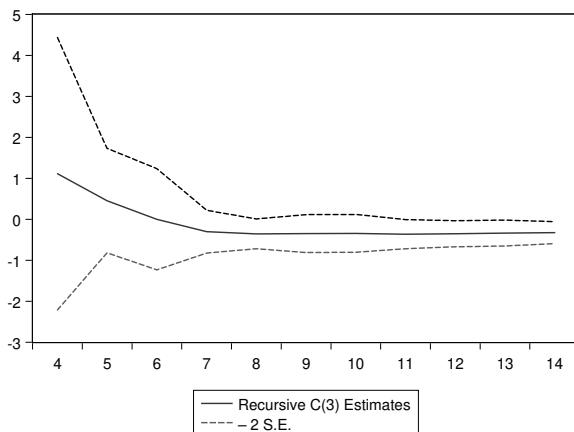
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24.13122	5.165337	4.6717	0.0007
X1	0.714896	0.266264	2.6849	0.0212
X2	-0.328113	0.134561	-2.4383	0.0329
R-squared	0.68754	Mean dependent var		20.714
Adjusted R-squared	0.630729	S.D. dependent var		4.1773
S.E. of regression	2.538501	Akaike info criterion		4.8884
Sum squared resid	70.88389	Schwarz criterion		5.0253
F-statistic	12.10223	Durbin-Watson stat		3.0780
Prob(F-statistic)	0.001665			



3 Modèle de Régression Linéaire Multiple

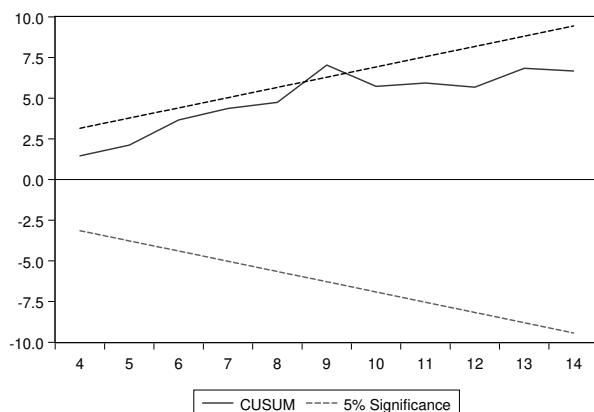


3 Modèle de Régression Linéaire Multiple

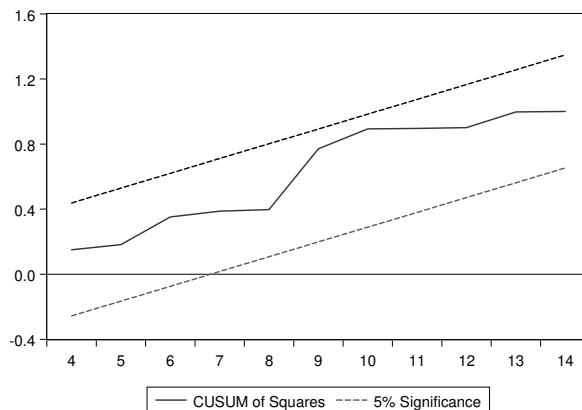


Commentaires : Nos trois coefficients sont à l'intérieur de l'intervalle de confiance, ce qui amène à conclure en faveur de la stabilité des paramètres. Compte tenu du degré de liberté très faible, l'estimation de l'écart type des coefficients (et donc l'intervalle de confiance) pour les trois premiers calculs n'est pas significatif.

« CUSUM Test » et « CUSUM of Squares Test »



3 Modèle de Régression Linéaire Multiple



Commentaires : les statistiques CUSUM et CUSUM of Squares évoluent à l'intérieur de l'intervalle de confiance, sauf un léger débordement pour la 9 ème dans CUSUM. Ainsi, nous concluons en faveur de l'absence d'un changement structurel.

- (2) Comme expliqué auparavant, le test de Ramsey est conduit de la manière suivante :

- Calcul des \hat{Y}_i à partir de : $\hat{Y}_i = 24.13 + 0.7148X_{i1} - 0.3281X_{i2}$
- Élever au carré la série à expliquer : \hat{Y}_i^2 , nous ne testons qu'une spécification non linéaire de type quadratique
- Estimer le modèle en ajoutant comme variable explicative supplémentaire \hat{Y}_i^2 :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varphi_2 \hat{Y}_i^2 + e_i$$

- Tester l'hypothèse $H_0 : \varphi_2 = 0$ contre $H_1 : \varphi_2 \neq 0$. La probabilité critique du coefficient φ_2 est égale à 0.827. Nous acceptons l'hypothèse H_0 , le modèle est donc correctement spécifié.

Sur **EViews**, dans l'output des résultats, suivre : View/Stability Tests/Ramsey RESET Test... Number of fitted : 1 :

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	15.14525	40.60157	0.373021	0.7169
X1	0.260542	2.053636	0.126869	0.9016
X2	-0.119662	0.944027	-0.126757	0.9016
FITTED ²	0.015484	0.069342	0.223308	0.8278
F-statistic	7.387881			
Prob(F-statistic)	0.006767			

3 Modèle de Régression Linéaire Multiple

Commentaires : le modèle est bien spécifié (la variable non linéaire du type quadratique est non significative et la probabilité associée à la statistique de Fisher calculée est $> 5\%$).

Exercice 16

On considère les statistiques ci-dessous obtenues à partir de $n = 95$ observations sur trois variables Y , X_1 et X_2

$$r_{X_1, X_2}^2 = 0.47 \quad r_{Y, X_1}^2 = 0.72 \quad r_{Y, X_2}^2 = 0.80 \quad \sigma_Y^2 = 800 \quad \bar{Y} = 13$$

- (1) Après avoir effectué la régression de Y sur X_1 , on a obtenu : $\hat{Y} = 9X_1 - 7.45$. Le coefficient β_1 de X_1 est-il significativement différent de 0 ?
- (2) Après avoir effectué la régression de Y sur X_2 , on a obtenu : $\hat{Y} = 5X_2 + 9$. Le coefficient β_2 de X_2 est-il significativement différent de 0 ?
- (3) Calculer les coefficients du modèle : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$, et le coefficient de corrélation multiple.
- (4) Les coefficients β_1 et β_2 sont-ils significativement différents de 0 ? La régression est-elle globalement significative ?

Solution

- (1) La variance estimée du coefficient de régression est donnée par :

$$\widehat{\sigma}_{\beta_1}^2 = \frac{\widehat{\sigma}^2}{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2}$$

Dans le cadre de la régression simple, il y a égalité entre corrélation simple et corrélation multiple, alors :

$$\begin{aligned} r_{Y, X_1}^2 &= \frac{\text{cov}(Y, X_1)^2}{\text{var}(Y) \text{var}(X_1)} = \frac{\left[\sum_{i=1}^n (Y_i - \bar{Y})(X_{i1} - \bar{X}_1) \right]^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2} \\ &= R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 0.72 \end{aligned}$$

3 Modèle de Régression Linéaire Multiple

$$\text{or } \text{var}(Y) = 800 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n}$$

$$= \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{95} \Rightarrow \sum_{i=1}^n (Y_i - \bar{Y})^2 = 76000^{24}$$

alors

$$\sum_{i=1}^n e_i^2 = (1 - 0.72) \times 76000 = 21280 \Rightarrow \hat{\sigma}^2 = 21280 / (n - 2) = 228.817$$

Calcul de $\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2$

On a

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_{i1} - \bar{X}_1)}{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2} = 9$$

et $r_{Y,X_1}^2 = \frac{\left[\sum_{i=1}^n (Y_i - \bar{Y})(X_{i1} - \bar{X}_1) \right]^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2} = 0.72$

Puisque $\sum_{i=1}^n (Y_i - \bar{Y})^2 = 76000$, alors

$$\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 = 675.5 \text{ et } \sum_{i=1}^n (Y_i - \bar{Y})(X_{i1} - \bar{X}_1) = 6080$$

soit $\text{var}(X_1) = 7.11$ et $\text{cov}(Y, X_1) = 64$

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2} = \frac{228.817}{675.5} = 0.3387$$

$$\Rightarrow t_{\hat{\beta}_1}^* = \frac{9}{0.5820} = 15.46 > t_{93}^{0.025} = 1.96$$

Enfin, le coefficient β_1 de régression de Y sur X_1 est significativement différent de 0.

- (2) On poursuit la même méthode de la question (1) cette fois pour le cas de la régression de Y sur X_2 . On trouve les résultats suivants :

$$\sum_{i=1}^n e_i^2 = (1 - 0.80) \times 76000 = 15200 \Rightarrow \hat{\sigma}^2 = \frac{15200}{n - 2} = 163.44$$

$$\sum_{i=1}^n (X_{i2} - \bar{X}_2)^2 = 2432 \text{ et } \sum_{i=1}^n (Y_i - \bar{Y})(X_{i2} - \bar{X}_2) = 12160$$

24. Ici, nous utilisons la formule de la variance d'une population (division par n) et non la formule de la variance d'un échantillon (division par $n - 1$).

3 Modèle de Régression Linéaire Multiple

soit $\text{var}(X_2) = 25.6$ et $\text{cov}(Y, X_2) = 128$

$$\begin{aligned}\hat{\sigma}_{\hat{\beta}_2}^2 &= \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_{i2} - \bar{X}_2)^2} = \frac{163.44}{2432} = 0.06720 \\ \Rightarrow t_{\hat{\beta}_2}^* &= \frac{5}{0.2592} = 19.287 > t_{93}^{0.025} = 1.96\end{aligned}$$

Enfin, le coefficient β_2 de régression de Y sur X_2 est significativement différent de 0.

- (3) Calcul des coefficients du modèle $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$

On peut calculer les estimateurs de β_1 et β_2 comme déjà fait dans la section 3.3

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) \end{pmatrix}^{-1} \times \begin{pmatrix} \text{cov}(X_1, Y) \\ \text{cov}(X_2, Y) \end{pmatrix}$$

Calcul de $\text{cov}(X_1, X_2)$:

$$\begin{aligned}r_{X_1, X_2}^2 &= \frac{\text{cov}(X_1, X_2)^2}{\text{var}(X_1) \text{var}(X_2)} = \frac{\left[\sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) \right]^2}{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2} \\ &= 0.47\end{aligned}$$

alors $\text{var}(X_1) = 7.11$ et $\text{var}(X_2) = 25.6 \Rightarrow \text{cov}(X_1, X_2) = 9.2492$

$$\begin{aligned}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \begin{pmatrix} 7.11 & 9.2492 \\ 9.2492 & 25.6 \end{pmatrix}^{-1} \times \begin{pmatrix} 64 \\ 128 \end{pmatrix} \\ &= \begin{pmatrix} 0.2654 & -0.0959 \\ -0.0959 & 0.0737 \end{pmatrix} \times \begin{pmatrix} 64 \\ 128 \end{pmatrix} = \begin{pmatrix} 1.6389 \\ 3.8687 \end{pmatrix}\end{aligned}$$

On déduit $\hat{\beta}_0$ à partir de la relation $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$
on a $\bar{Y} = 13$ et $\bar{X}_1 = 2.272$ car $\bar{Y} = 9\bar{X}_1 - 7.45$ (d'après le modèle 1 de la question 1)
et $\bar{X}_2 = 0.8$ car $\bar{Y} = 5\bar{X}_2 + 9$ (d'après le modèle 2 de la question 2)

Alors $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 = 6.181$

Coefficient de Corrélation Multiple

Si on raisonne sur base des données centrées, le coefficient de dé-

3 Modèle de Régression Linéaire Multiple

terminisation sera :

$$R^2 = \frac{\widehat{Y}'\widehat{Y}}{Y'Y} = \frac{\widehat{\beta}'X'Y}{Y'Y} = \frac{\begin{pmatrix} 1.6389 & 3.8687 \end{pmatrix} \begin{pmatrix} 6080 \\ 12160 \end{pmatrix}}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

$$= \frac{57007.90}{76000} = 0.7501$$

Le coefficient de corrélation multiple est égal à $r = \sqrt{R^2} = 0.86608$

(4) Test de significativité des coefficients β_1 et β_2

on a

$$\widehat{\sigma}^2(\widehat{\beta}) = \widehat{\sigma}^2 \times (X'X)^{-1} = \widehat{\sigma}^2 \times \frac{1}{n} \times \begin{pmatrix} var(X_1) & cov(X_1, X_2) \\ cov(X_2, X_1) & var(X_2) \end{pmatrix}^{-1}$$

$$\widehat{\sigma}^2 = MCR = \frac{SCR}{n-m-1} = \frac{(1-R^2)SCT}{n-m-1}$$

$$= \frac{(1-R^2) \sum_{i=1}^n (Y_i - \bar{Y})^2}{n-m-1} = \frac{(1-0.7501) \times 76000}{95-2-1} = 206.44$$

alors

$$\widehat{\sigma}^2(\widehat{\beta}) = 206.44 \times \frac{1}{95} \times \begin{pmatrix} 0.2654 & -0.0959 \\ -0.0959 & 0.0737 \end{pmatrix}$$

$$= \begin{pmatrix} 0.5767 & -0.2084 \\ -0.2084 & 0.1601 \end{pmatrix}$$

$$\widehat{\sigma}_{\widehat{\beta}_1}^2 = 0.5767 \rightarrow t_{\widehat{\beta}_1}^* = \frac{1.6389}{\sqrt{0.5767}} = 2.1581 > t_{92}^{0.025} = 1.96$$

$$\widehat{\sigma}_{\widehat{\beta}_2}^2 = 0.1601 \rightarrow t_{\widehat{\beta}_2}^* = \frac{3.8687}{\sqrt{0.1601}} = 9.6687 > t_{92}^{0.025} = 1.96$$

Les coefficients $\widehat{\beta}_1$ et $\widehat{\beta}_2$ sont significativement différents de 0.

Remarque :

On observe des différences par rapport aux valeurs estimées avant sur les coefficients β_1 , β_2 et sur leurs écarts types. Cela est la conséquence de la colinéarité entre X_1 et X_2 , ($r_{X_1, X_2} = 0.6855$).

3 Modèle de Régression Linéaire Multiple

Test de Fisher

$$F^* = \frac{R^2/m}{(1-R^2)/(n-m-1)} = \frac{0.7501/2}{(1-0.7501)/92} \\ = 138.07 > F_{2,92}^{0.05} = 3.10$$

La régression est globalement significative.

Exercice 17

Une étude réalisée sur un échantillon de $n = 23$ entreprises concerne l'étude de l'influence des *heures de travail* X_1 et du *capital utilisé* X_2 sur la production industrielle Y . L'étude menée a abouti au résultats ci-dessous :

$$Y_i = 26.7 + 0.676X_{i1} - 0.461X_{i2} + e_i$$

$$R^2 = 0.915$$

$$n = 23$$

$$SCT = \sum_{i=1}^n (Y_i - \bar{Y})^2 = 319.67$$

$$(X'X)^{-1} = \begin{pmatrix} 0.001 & -0.07 & -0.32 \\ -0.07 & 0.0037 & 0.019 \\ -0.32 & 0.019 & 0.29 \end{pmatrix}$$

- (1) Existe-t-il une influence d'au moins de l'un des deux facteurs (c-à-d les heures de travail et le capital utilisé) ?
- (2) Le coefficient des heures de travail est-il significativement deux fois plus élevé que celui du capital utilisé ?

Solution

- (1) La question se ramène à un test de Fisher dont les hypothèses sont :

$$H_0 : \beta_1 = \beta_2 = 0$$

H_1 : Il existe au moins un de ces coefficients non nul

$$F^* = \frac{R^2/m}{(1-R^2)/(n-m-1)} = \frac{0.915/2}{(1-0.915)/20} \\ = 107.647 > F_{2,20}^{0.05} = 3.49$$

3 Modèle de Régression Linéaire Multiple

Nous rejetons l'hypothèse H_0 , ainsi, il existe au moins un des coefficients non nul.

(2) La question se ramène à un test dont les hypothèses sont :

$$\begin{aligned} H_0 : \beta_1 - 2\beta_2 &= 0 \\ H_1 : \beta_1 - 2\beta_2 &\neq 0 \end{aligned} \quad \text{où} \quad \begin{aligned} H_0 : d &= 0 \\ H_1 : d &\neq 0 \end{aligned}$$

sous H_0 , le ratio $\frac{\hat{d} - 0}{\hat{\sigma}_{\hat{d}}}$ suit une loi de Student à $n - m - 1$ degrés de liberté.

On a

$$var(\hat{d}) = var(\hat{\beta}_1 - 2\hat{\beta}_2) = var(\hat{\beta}_1) + 4var(\hat{\beta}_2) - 4cov(\hat{\beta}_1, \hat{\beta}_2)$$

$$SCR = \sum_{i=1}^n e_i^2 = (1 - R^2) SCT = (1 - 0.915) \times 319.67 = 27.1719$$

$$\begin{aligned} \Rightarrow \hat{\sigma}_{\hat{d}}^2 &= \hat{\sigma}_{\hat{\beta}_1}^2 + 4\hat{\sigma}_{\hat{\beta}_2}^2 - 4cov(\hat{\beta}_1, \hat{\beta}_2) \\ &= 0.001 + 4(0.0037) - 4(-0.07) = 0.2957 \end{aligned}$$

$$t_{\hat{d}}^* = \frac{|\hat{d} - d|}{\hat{\sigma}_{\hat{d}}} = \frac{0.676 + 2 \times 0.461}{\sqrt{0.2957}} = 2.938 > t_{20}^{0.25} = 2.86$$

Nous rejetons, au seuil de 5%, l'hypothèse H_0 .

Remarque :

On peut répondre à cette question par un test d'analyse de la variance en considérant un modèle complet et un modèle réduit comme suit :

$$(C) : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad \text{Modèle complet}$$

$$(R) : Y_i = \beta_0 + \beta_c \left(X_{i1} + \frac{1}{2} X_{i2} \right) + \varepsilon_i \quad \text{Modèle réduit}$$

où β_c désigne le coefficient commun pour β_1 et β_2 sous H_0 et $\left(X_{i1} + \frac{1}{2} X_{i2} \right)$ est la nouvelle variable X correspondante.

Chapitre 4

Violations des Hypothèses Classiques

Dans ce chapitre, nous assouplissons les hypothèses formulées dans le chapitre 2 et nous étudions par la suite l'effet de cela sur l'estimateur MCO. Dans le cas où l'estimateur MCO n'est plus un estimateur viable, nous dérivons un estimateur alternatif et nous proposons quelques tests qui nous permettront de vérifier si une telle hypothèse est violée.

4.1 La Multicolinéarité et Ses Effets

Dans une analyse de régression multiple, la nature et l'importance des relations entre les variables explicatives et la variable de dépendante présentent souvent un intérêt particulier. Voici quelques questions fréquemment posées :

- Quelle est l'importance de l'effet des différentes variables prédictives ?
- Quelle est l'ampleur de l'effet d'une variable prédictive donnée sur la variable dépendante ?
- Est-ce qu'une variable prédictive peut être supprimée du modèle car elle n'a que peu ou pas d'effet sur la variable dépendante ?
- Les variables prédictives non encore incluses dans le modèle devraient-elles être considérées pour une éventuelle inclusion ?

Si les variables prédictives incluses dans le modèle sont (1) non corrélées entre elles et (2) non corrélées avec d'autres variables prédictives liées à la variable de dépendante mais omises du modèle, des réponses relativement simples peuvent être données à ces questions. Malheureusement,

4 Violations des Hypothèses Classiques

dans de nombreuses situations non expérimentales dans les domaines de l'économie et des sciences sociales et biologiques, les variables explicatives ou prédictives ont tendance à être corrélées entre elles et avec d'autres variables liées à la variable de dépendante mais elle ne sont pas incluses dans le modèle. Par exemple, dans une régression des dépenses alimentaires familiales sur les variables explicatives revenu familial, épargne familiale et âge du chef de ménage, les variables explicatives seront corrélées entre elles. En outre, elles seront également corrélées à d'autres variables socioéconomiques non incluses dans le modèle et qui ont une incidence sur les dépenses alimentaires de la famille, telles que la taille de la famille.

Lorsque les variables prédictives sont corrélées entre elles, on dit qu'il existe une *intercorrélation* ou une *multicolinéarité*. (Parfois, ce dernier terme est réservé aux cas où la corrélation entre les variables prédictives est très élevée.) Nous allons explorer divers problèmes liés entre eux, créés par la multicolinéarité parmi les variables prédictives. Cependant, nous examinons d'abord la situation dans laquelle les variables prédictives ne sont pas corrélées.

Le terme "multicolinéarité" a été introduit pour la première fois en 1934 par Ragnar Frisch dans son livre sur l'analyse de confluence et faisait référence à une situation dans laquelle les variables traitées sont sujettes à deux relations ou plus. La multicolinéarité ou les intercorrélations élevées entre les variables explicatives ne doivent pas nécessairement être un problème. Que ce soit un problème ou non, cela dépend d'autres facteurs. Ainsi, le problème de multicolinéarité ne peut pas être entièrement traité en termes d'intercorrélations entre les variables. En outre, différentes paramétrisations des variables donneront des magnitudes différentes de ces intercorrélations.

4.1.1 Variables Prédictives Non Corrélées

Le tableau (4.1) contient les données d'une expérience à petite échelle sur l'effet de la taille de l'équipe de travail (X_1) et du montant de la prime (X_2) sur la productivité de l'équipage (Y). Les variables prédictives X_1 et X_2 ne sont pas corrélées ici, c'est-à-dire $r_{12}^2 = 0$, où r_{12}^2 désigne le coefficient de détermination simple entre X_1 et X_2 . Le tableau (4.2a) contient la fonction de régression ajustée et le tableau d'analyse de la variance lorsque X_1 et X_2 sont inclus dans le modèle. Le tableau (4.2b) contient les mêmes informations lorsque seul X_1 est inclus dans le modèle, et le tableau (4.2c) contient ces informations lorsque seul X_2 figure dans le modèle.

4 Violations des Hypothèses Classiques

Une caractéristique importante à noter dans le tableau (4.2) est le coefficient de régression de X_1 , $\hat{\beta}_1 = 5.375$ qui reste le même, que ce soit dans le cas où seul X_1 soit inclus dans le modèle ou dans le cas où les deux variables prédictives soient incluses. Il en va de même pour $\hat{\beta}_2 = 9.250$. C'est le résultat de la non corrélation des deux variables prédictives.

Productivité de l'équipage Y_i	Taille de l'équipage X_{i1}	Prime d'activité X_{i2}
42	4	2
39	4	2
48	4	3
51	4	3
49	6	2
53	6	2
61	6	3
60	6	3

TABLE 4.1 – **Variables de prédiction non corrélées - Exemple de productivité des équipes de travail**

Ainsi, lorsque les variables prédictives ne sont pas corrélées, les effets qui leur sont attribués par un modèle de régression de premier ordre sont les mêmes, quel que soit les autres variables prédictives incluses dans le modèle. Ceci est un argument fort pour les expériences contrôlées chaque fois que possible, car le contrôle expérimental permet de choisir les niveaux des variables prédictives afin de rendre ces dernières non corrélées.

Une autre caractéristique importante du tableau (4.2) est liée à la somme des carrés des résidus. On remarque à partir du tableau (4.2) que la somme supplémentaire des carrés expliquée $SCE(X_1 | X_2)$ est égale à la somme la somme des carrés expliquée $SCE(X_1)$ lorsque seul X_1 figure dans le modèle de régression :

$$\begin{aligned} SCE(X_1 | X_2) &= SCR(X_2) - SCR(X_1, X_2) \\ &= 248.750 - 17.625 = 231.125 \\ SCE(X_1) &= 231.125 \end{aligned}$$

De même, la somme supplémentaire des carrés $SCE(X_2 | X_1)$ est égale à $SCE(X_2)$, la somme des carrés expliquée lorsque seul X_2 figure dans le modèle de régression :

4 Violations des Hypothèses Classiques

(a) Regression de Y sur X_1 et X_2			
Source de variation	Somme des Carrés	Degrés de liberté (ddl)	Carrés Moyens
X	402.250	2	201.125
Résidu	17.625	5	3.525
Total	419.875	7	

(b) Regression de Y sur X_1			
Source de variation	Somme des Carrés	Degrés de liberté (ddl)	Carrés Moyens
X	231.125	1	231.125
Résidu	188.750	6	31.458
Total	419.875	7	

(c) Regression de Y sur X_2			
Source de variation	Somme des Carrés	Degrés de liberté (ddl)	Carrés Moyens
X	171.125	1	171.125
Résidu	248.750	6	41.458
Total	419.875	7	

TABLE 4.2 – Résultats de régression lorsque les variables prédictives ne sont pas corrélées - Exemple de productivité des équipes de travail

$$\begin{aligned}
 SCE(X_2 | X_1) &= SCR(X_1) - SCR(X_1, X_2) \\
 &= 188.750 - 17.625 = 171.125 \\
 SCE(X_2) &= 171.125
 \end{aligned}$$

En général, lorsque deux variables prédictives ou plus ne sont pas corrélées, la contribution marginale d'une variable prédictive à la réduction de la somme des carrés des résidus lorsque les autres variables prédictives figurent dans le modèle est exactement la même que lorsque cette variable prédictive se trouve seule dans le modèle.

Remarque :

4 Violations des Hypothèses Classiques

Pour montrer que le coefficient de régression de X_1 demeure le même lorsque X_2 est ajouté au modèle de régression dans lequel X_1 et X_2 ne sont pas corrélés, considérons l'expression algébrique suivante de $\hat{\beta}_1$ dans le modèle de régression multiple de premier ordre avec deux variables prédictives :

$$\hat{\beta}_1 = \frac{\frac{\sum_{i=1}^n (X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2} - \left[\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2} \right]^{1/2} r_{Y2} r_{12}}{1 - r_{12}^2} \quad (4.1)$$

où, r_{Y2} désigne le coefficient de corrélation simple entre Y et X_2 , et r_{12} désigne le coefficient de corrélation simple entre X_1 et X_2 .

Si X_1 et X_2 ne sont pas corrélés, $r_{12} = 0$ et (4.1) se réduit à :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2} \quad (4.2)$$

Or, (4.2) est l'estimateur de la pente pour la régression linéaire simple de Y sur X_1 .

Par conséquent, lorsque X_1 et X_2 ne sont pas corrélés, l'ajout de X_2 au modèle de régression ne modifie pas le coefficient de régression pour X_1 ; en conséquence, l'ajout de X_1 au modèle de régression ne modifie pas le coefficient de régression pour X_2 .

4.1.2 Nature du Problème lorsque les Variables Prédictives sont Parfaitement Corrélées

Pour voir la nature essentielle du problème de la multicolinéarité, nous allons utiliser un exemple simple où les deux variables prédictives sont parfaitement corrélées. Les données du tableau (4.3) se rapportent à quatre exemples d'observations concernant une variable dépendante et deux variables prédictives. Il est demandé d'estimer la fonction de régression multiple de premier ordre :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \quad (4.3)$$

4 Violations des Hypothèses Classiques

			Valeurs estimées pour la fonction de régression	
X_{i1}	X_{i2}	Y_i	(4.4)	(4.5)
2	6	23	23	23
8	9	83	83	83
6	8	63	63	63
10	10	103	103	103

Fonctions Estimées

$$\hat{Y} = -87 + X_1 + 18X_2 \quad (4.4)$$

$$\hat{Y} = -7 + 9X_1 + 2X_2 \quad (4.5)$$

TABLE 4.3 – Exemple de Variables Prédictives Parfaitement Corrélates

La fonction de réponse est parfaitement ajustée aux données comme indiqué dans le tableau (4.3).

En effet, on peut montrer qu'une infinité de fonctions de réponse s'adapteront parfaitement aux données du tableau (4.3). La raison en est que les variables prédictives X_1 et X_2 sont parfaitement liées selon la relation :

$$X_2 = 5 + 0.5X_1 \quad (4.6)$$

Notez que les fonctions de réponse ajustées (4.4) et (4.5) sont des surfaces de réponse entièrement différentes, comme on peut le voir sur la figure (4.1). Les deux surfaces de réponse n'ont les mêmes valeurs ajustées que lorsqu'elles se coupent. Cela se produit lorsque X_1 et X_2 suivent la relation (4.6). c'est-à-dire, lorsque $X_2 = 5 + 0.5X_1$.

Ainsi, lorsque X_1 et X_2 sont parfaitement liés comme c'est le cas dans notre exemple, les données ne contiennent aucune composante d'erreur aléatoire, de nombreuses fonctions de réponse différentes conduiront aux mêmes valeurs parfaitement ajustées pour les observations et aussi aux mêmes valeurs ajustées pour les autres combinaisons (X_1, X_2) suivant la relation entre X_1 et X_2 . Pourtant, ces fonctions de réponse ne sont pas les mêmes et conduiront à des valeurs ajustées différentes pour les combinaisons (X_1, X_2) qui ne suivent pas la relation entre X_1 et X_2 .

Les deux implications clés de cet exemple sont :

- (1) La relation parfaite entre X_1 et X_2 n'a pas empêché notre capacité d'obtenir un bon ajustement aux données.
- (2) Étant donné que de nombreuses fonctions de réponse différentes fournissent le même « bon » ajustement, nous ne pouvons interpré-

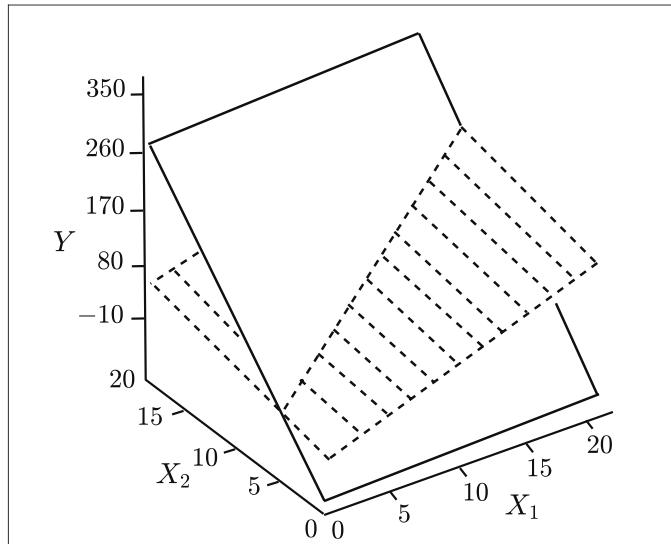


FIGURE 4.1 – Deux plans de réponse qui se croisent lorsque $X_2 = 5 + 0.5X_1$

ter aucun ensemble de coefficients de régression comme reflétant les effets des différentes variables prédictives. Ainsi, dans la fonction de réponse (4.4), $\hat{\beta}_1 = 1$ et $\hat{\beta}_2 = 18$ ne veut pas dire que X_2 est la variable prédictive clé et X_1 joue peu de rôle, car la fonction de réponse (4.5) fournit un ajustement tout aussi bon et ses coefficients de régression ont des grandeurs comparatives opposées.

4.1.3 Effets de la Multicolinéarité

En pratique, nous trouvons rarement des variables prédictives parfaitement liées ou des données qui ne contiennent pas de composante d'erreur aléatoire. Néanmoins, les implications que nous venons de noter pour notre exemple sont toujours pertinentes.

- (1) Le fait que toutes ou certaines variables prédictives soient corrélées entre elles n'empêche généralement pas notre capacité à obtenir un bon ajustement et n'a pas tendance à affecter les inférences sur les réponses moyennes ou les prédictions de nouvelles observations, à condition que ces inférences soient faites dans la région d'observations. (La figure (3.4) illustre le concept de la région des observations pour le cas de deux variables prédictives).
- (2) Dans la vie réelle, le fait que de nombreuses fonctions de régres-

4 Violations des Hypothèses Classiques

sion différentes fournissent de bons ajustements aux données de notre exemple est que les coefficients de régression estimés ont tendance à avoir une grande variabilité d'échantillonnage lorsque les variables prédictives sont fortement corrélées. Ainsi, les coefficients de régression estimés ont tendance à varier considérablement d'un échantillon à l'autre lorsque les variables prédictives sont fortement corrélées. Par conséquent, seules des informations imprécises peuvent être disponibles sur les coefficients de régression réels individuels. En effet, un bon nombre des coefficients de régression estimés individuellement peuvent ne pas être statistiquement significatifs même s'il existe une relation statistique définie entre la variable de réponse et l'ensemble de variables prédictives.

- (3) L'interprétation courante d'un coefficient de régression comme une mesure de la variation de la valeur attendue de la variable de réponse lorsque la variable prédictive donnée est augmentée d'une unité tandis que toutes les autres variables prédictives sont maintenues constantes n'est pas entièrement applicable lorsqu'il existe une multicolinéarité. Il peut être conceptuellement faisable de penser à faire varier une variable prédictive et à maintenir les autres constantes, mais il peut ne pas être possible dans la pratique de le faire pour des variables prédictives qui sont fortement corrélées. Par exemple, pour prédire le rendement des cultures à partir de la quantité de précipitations et des heures d'ensoleillement, la relation entre les deux variables prédictives rend irréaliste d'envisager de varier l'une tout en maintenant l'autre constante. Par conséquent, l'interprétation simple des coefficients de régression comme mesure des effets marginaux est souvent injustifiée avec l'existence des variables prédictives hautement corrélées.

Nous illustrons ces effets de la multicolinéarité en considérant l'exemple ci-dessous.

Exemple

Le tableau (4.4) contient une partie des données pour une étude de la relation entre la quantité de graisse corporelle (Y) et plusieurs variables prédictives possibles, sur la base d'un échantillon de 20 femmes en bonne santé âgées de 25 à 34 ans. Les variables prédictives possibles sont l'épaisseur du pli cutané du triceps (X_1), la circonférence de la cuisse (X_2) et la circonférence du bras (X_3).

Le tableau (4.5) contient les principaux résultats de régression lorsque la graisse corporelle (Y) est régressée (1) sur l'épaisseur du pli cutané du triceps (X_1) seul, (2) sur la circonférence de la cuisse (X_2) seule, (3) sur (X_1) et (X_2) uniquement, et (4) sur les trois variables prédictives.

4 Violations des Hypothèses Classiques

La somme des carrés expliquée lorsque X_1 seul figure dans le modèle est $SCE(X_1) = 352.27$ selon le tableau (4.5a). La somme des carrés des résidus pour ce modèle est $SCR(X_1) = 143.12$. De même, le tableau (4.6c) indique que lorsque X_1 et X_2 sont dans le modèle de régression, la somme des carrés expliquée est $SCE(X_1, X_2) = 385.44$ et la somme des carrés des résidus est $SCR(X_1, X_2) = 109.95$.

Notez que la somme des carrés des résidus lorsque X_1 et X_2 sont dans le modèle ($SCR(X_1, X_2) = 109.95$) est plus petite que celle lorsque le modèle ne contient que X_1 , $SCR(X_1, X_2) = 143.12$. La différence est appelée une *somme supplémentaire des carrés* et sera notée par $SCE(X_2 | X_1)$:

$$\begin{aligned} SCE(X_2 | X_1) &= SCR(X_1) - SCR(X_1, X_2) \\ &= 143.12 - 109.95 = 33.17 \end{aligned}$$

La raison de l'équivalence de la réduction marginale de la somme

Observation i	Quantité de graisse corporelle Y_i	Epaisseur du pli cutané du triceps X_{i1}	Circonférence de la cuisse X_{i2}	Circonférence du bras X_{i3}
1	11.9	19.5	43.1	29.1
2	22.8	24.7	49.8	28.2
3	18.7	30.7	51.9	37
:	:	:	:	:
18	25.4	30.2	58.6	24.6
19	14.8	22.7	48.2	27.1
20	21.1	25.2	51	27.5

TABLE 4.4 – **Données de base - Exemple de graisse corporelle**

des carrés des résidus et de l'augmentation marginale de la somme des carrés expliquée est la base de l'analyse de la variance, et on a $SCT = SCE + SCR$.

Étant donné que SCT mesure la variabilité des observations Y_i et ne dépend donc pas du modèle de régression ajusté, toute réduction de la SCR implique une augmentation identique de la SCE . Nous pouvons considérer d'autres sommes supplémentaires des carrés, comme l'effet marginal de l'ajout de X_3 au modèle de régression lorsque X_1 et X_2 sont déjà dans le modèle. Nous trouvons dans les tableaux (4.6c) et (4.6d) que :

4 Violations des Hypothèses Classiques

(a) Regression de Y sur X_1			
Source de variation	Somme des Carrés	ddl	Carrés Moyens
X	352.27	1	352.27
Résidu	143.12	18	7.95
Total	495.39	19	
Variable	Coefficient Estimé	Ecart type Estimé	t^*
X_1	$\hat{\beta}_1 = 0.8572$	$\hat{\sigma}_{\hat{\beta}_1} = 0.1288$	6.66

(b) Regression de Y sur X_2			
Source de variation	Somme des Carrés	ddl	Carrés Moyens
X	381.97	1	381.97
Résidu	113.42	18	6.30
Total	495.39	19	
Variable	Coefficient Estimé	Ecart type Estimé	t^*
X_2	$\hat{\beta}_2 = 0.8565$	$\hat{\sigma}_{\hat{\beta}_2} = 0.1100$	7.79

TABLE 4.5 – Résultats de régression pour plusieurs modèles ajustés - Exemple de graisse corporelle

$$\begin{aligned} SCE(X_3 | X_1, X_2) &= SCR(X_1, X_2) - SCR(X_1, X_2, X_3) \\ &= 109.95 - 98.41 = 11.54 \end{aligned}$$

ou de manière équivalente

$$\begin{aligned} SCE(X_3 | X_1, X_2) &= SCE(X_1, X_2, X_3) - SCE(X_1, X_2) \\ &= 396.98 - 385.44 = 11.54 \end{aligned}$$

Nous pouvons même considérer l'effet marginal de l'ajout de plusieurs variables, comme l'ajout de X_2 et X_3 au modèle de régression contenant

4 Violations des Hypothèses Classiques

(c) Regression de Y sur X_1 et X_2

$$\hat{Y} = -19.174 + 0.2224X_1 + 0.6594X_2$$

Source de variation	Somme des Carrés	ddl	Carrés Moyens
X	385.44	2	192.72
Résidu	109.95	17	6.47
Total	495.39	19	
Variable	Coefficient Estimé	Ecart type Estimé	t^*
X_1	$\hat{\beta}_1 = 0.2224$	$\hat{\sigma}_{\hat{\beta}_1} = 0.3034$	0.73
X_2	$\hat{\beta}_2 = 0.6594$	$\hat{\sigma}_{\hat{\beta}_2} = 0.2912$	2.26

(d) Regression de Y sur X_1 , X_2 et X_3

$$\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$$

Source de variation	Somme des Carrés	ddl	Carrés Moyens
X	396.98	3	132.33
Résidu	98.41	16	6.15
Total	495.39	19	
Variable	Coefficient Estimé	Ecart type Estimé	t^*
X_1	$\hat{\beta}_1 = 4.334$	$\hat{\sigma}_{\hat{\beta}_1} = 3.016$	1.44
X_2	$\hat{\beta}_2 = -2.857$	$\hat{\sigma}_{\hat{\beta}_2} = 2.582$	-1.11
X_3	$\hat{\beta}_3 = -2.186$	$\hat{\sigma}_{\hat{\beta}_3} = 1.596$	-1.37

TABLE 4.6 – Suite des résultats - Exemple de graisse corporelle

déjà X_1 (voir les tableaux (4.5a) et (4.6d)) :

$$\begin{aligned} SCE(X_2, X_3 | X_1) &= SCR(X_1) - SCR(X_1, X_2, X_3) \\ &= 143.12 - 98.41 = 44.71 \end{aligned}$$

ou de manière équivalente

$$\begin{aligned} SCE(X_2, X_3 | X_1) &= SCE(X_1, X_2, X_3) - SCE(X_1) \\ &= 396.98 - 352.27 = 44.71 \end{aligned}$$

La figure (4.2) contient la matrice du nuage de points et la matrice de corrélation des variables prédictives. Il ressort de la matrice du nuage de points que les variables prédictives X_1 et X_2 sont fortement corrélées ;

4 Violations des Hypothèses Classiques

la matrice de corrélation des variables X montre que le coefficient de corrélation simple est $r_{12} = 0.924$. D'un autre côté, X_3 n'est pas si étroitement lié à X_1 et X_2 individuellement ; la matrice de corrélation montre que les coefficients de corrélation sont $r_{13} = 0.458$ et $r_{23} = 0.085$. (Mais X_3 est fortement corrélé avec X_1 et X_2 ensemble ; le coefficient de détermination multiple lorsque X_3 est régressé sur X_1 et X_2 est de 0.998).

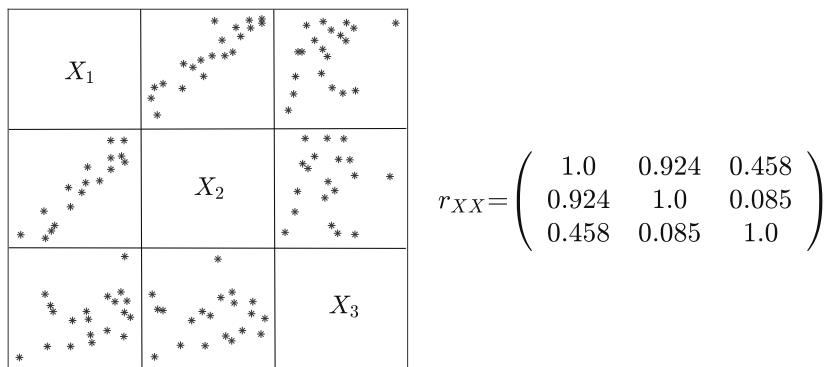


FIGURE 4.2 – Matrice de nuage de points et matrice de corrélation des variables prédictives - Exemple de graisse corporelle

Effets sur les Coefficients de Régression

Il ressort des tableaux (4.5) et (4.6) que le coefficient de régression pour X_1 , l'épaisseur du pli cutané des triceps, varie considérablement en fonction des autres variables incluses dans le modèle :

Variables dans le modèle	$\hat{\beta}_1$	$\hat{\beta}_2$
X_1	0.8572	–
X_2	–	0.8565
X_1, X_2	0.2224	0.6594
X_1, X_2, X_3	4.334	-2.857

L'histoire est la même pour le coefficient de régression pour X_2 . En effet, le coefficient de régression $\hat{\beta}_2$ change même de signe lorsque X_3 est ajouté au modèle qui inclut X_1 et X_2 .

La conclusion importante que nous devons tirer est la suivante : Lorsque les variables prédictives sont corrélées, le coefficient de régression de n'importe quelle variable dépend des autres variables prédictives incluses

4 Violations des Hypothèses Classiques

dans le modèle et de celles qui sont omises. Ainsi, un coefficient de régression ne reflète aucun effet inhérent de la variable prédictive particulière sur la variable de réponse, mais seulement un effet marginal ou partiel, compte tenu des autres variables prédictives corrélées incluses dans le modèle.

Remarque

Pour montrer la façon dont les variables prédictives intercorrélées qui sont omises du modèle de régression peuvent influencer les coefficients de régression dans le modèle de régression, on considère l'exemple suivant : Un analyste avait trouvé dans une régression des ventes des entreprises d'un territoire sur la taille de la population de ce territoire, le revenu par habitant et certaines autres variables prédictives que le coefficient de régression pour la taille de la population était négatif, et cette conclusion était appuyée par un intervalle de confiance pour le coefficient de régression. Un consultant a noté que l'analyste n'avait pas inclus la pénétration du marché du principal concurrent comme variable prédictive dans le modèle. Le concurrent était le plus actif et le plus efficace dans les territoires à forte population, limitant ainsi les ventes de l'entreprise sur ces territoires. Le résultat de l'omission de cette variable prédictive du modèle était un coefficient négatif pour la variable de taille de la population.

Effets sur les Sommes Supplémentaires des Carrés

Lorsque les variables prédictives sont corrélées, la contribution marginale de toute variable prédictive à la réduction de la somme des carrés des résidus varie selon les autres variables déjà présentes dans le modèle de régression, tout comme pour les coefficients de régression. Par exemple, le tableau (4.5) fournit les sommes supplémentaires de carrés suivantes pour X_1 :

$$\begin{aligned}SCE(X_1) &= 352.27 \\SCE(X_1 | X_2) &= 3.47\end{aligned}$$

La raison pour laquelle le $SCE(X_1 | X_2)$ est si petit par rapport au $SCE(X_1)$ est que X_1 et X_2 sont fortement corrélés entre eux et avec la variable dépendante. Ainsi, lorsque X_2 est déjà dans le modèle de régression, la contribution marginale de X_1 à la réduction de la somme des carrés des résidus est relativement faible car X_2 contient une grande partie des mêmes informations que X_1 . La même histoire se trouve dans les tableaux (4.5) et (4.6) pour X_2 . Ici $SCE(X_2 | X_1) = 33.17$, ce qui est

4 Violations des Hypothèses Classiques

beaucoup plus petit que $SCE(X_2) = 381.97$. La conclusion importante est la suivante : Lorsque les variables prédictives sont corrélées, il n'y a pas de somme unique des carrés qui peut être attribuée à toute variable prédictive comme reflétant son effet de réduction de la variation totale de Y . La réduction de la variation totale attribuée à une variable prédictive doit être considérée dans le contexte des autres variables prédictives corrélées déjà incluses dans le modèle.

Remarque

La multicollinéarité affecte également les coefficients de détermination partielle par ses effets sur les sommes supplémentaires des carrés. Remarquez dans le tableau (4.5) pour l'exemple de la graisse corporelle, par exemple, que X_1 est fortement corrélé avec Y :

$$R_{Y1}^2 = \frac{SCE(X_1)}{SCT} = \frac{352.27}{495.39} = 0.71$$

Cependant, le coefficient de détermination partielle entre Y et X_1 lorsque X_2 est déjà dans le modèle de régression, est beaucoup plus petit :

$$R_{Y1|2}^2 = \frac{SCE(X_1 | X_2)}{SCR(X_2)} = \frac{3.47}{113.42} = 0.03$$

Le fait d'avoir un faible coefficient de détermination partielle ici est, comme nous l'avons vu, c'est à cause de la forte corrélation entre X_1 et X_2 et avec la variable de réponse. Par conséquent, X_1 ne fournit que des informations supplémentaires relativement limitées au-delà de celles fournies par X_2 .

La somme supplémentaire des carrés pour une variable prédictive dans un modèle contenant déjà d'autres variables prédictives corrélées ne doit pas nécessairement être plus petite qu'avant que ces autres variables ne soient dans le modèle, comme nous l'avons trouvé dans l'exemple de graisse corporelle. Dans des cas particuliers, elle peut être plus grande. Considérez l'ensemble de données suivant et la matrice de corrélation :

$$\begin{array}{ccc} Y & X_1 & X_2 \\ \hline 20 & 5 & 25 \\ 20 & 10 & 30 \\ 0 & 5 & 5 \\ 1 & 10 & 10 \end{array} \quad \begin{array}{ccc} Y & X_1 & X_2 \\ \hline Y & 1 & 0.026 & 0.976 \\ X_1 & & 1 & 0.243 \\ X_2 & & & 1 \end{array}$$

Ici, Y et X_2 sont fortement corrélés positivement, mais Y et X_1 sont pratiquement non corrélés. De plus, X_1 et X_2 sont moyennement positivement corrélés. La somme supplémentaire des carrés pour X_1 lors-

4 Violations des Hypothèses Classiques

qu'elle est la seule variable du modèle pour cet ensemble de données est $SCE(X_1) = 0.25$, mais lorsque X_2 est déjà dans le modèle, la somme supplémentaire des carrés est $SCE(X_1|X_2) = 18.01$. De même, nous avons pour ces données :

$$SCE(X_2) = 362.49 \quad SCE(X_2|X_1) = 380.25$$

L'augmentation des sommes supplémentaires des carrés avec l'ajout de l'autre variable prédictive dans le modèle est liée à la situation particulière ici où X_1 est pratiquement non corrélé avec Y mais modérément corrélé avec X_2 , qui est à son tour fortement corrélé avec Y . Même ici, le point général est toujours valable, c'est-à-dire que la somme supplémentaire des carrés est affectée par les autres variables prédictives corrélées déjà présentes dans le modèle.

Lorsque $SCE(X_1|X_2) > SCE(X_1)$, comme dans l'exemple qui vient d'être cité, la variable X_2 est parfois appelée « variable de suppression ». Puisque $SCE(X_2|X_1) > SCE(X_2)$ dans l'exemple, la variable X_1 serait également appelée variable de suppression.

Effets sur $\hat{\sigma}_{\beta_k}$

On remarque à partir des tableaux (4.5) et (4.6) de l'exemple de graisse corporelle combien les coefficients de régression estimés $\hat{\beta}_1$ et $\hat{\beta}_2$ deviennent plus imprécis à mesure que des variables prédictives sont ajoutées davantage au modèle de régression :

Variables dans le modèle	$\hat{\beta}_1$	$\hat{\beta}_2$
X_1	0.1288	—
X_2	—	0.1100
X_1, X_2	0.3034	0.2912
X_1, X_2, X_3	3.016	2.582

Encore une fois, le degré élevé de multicolinéarité parmi les variables prédictives est responsable de la variabilité gonflée des coefficients de régression estimés.

Effets sur les Valeurs Ajustées et les Prévisions

On remarque à partir des tableaux (4.5) et (4.6) de l'exemple de graisse corporelle que la multicolinéarité élevée parmi les variables prédictives n'empêche pas la somme des carrés des résidus moyenne MCR , d'être régulièrement réduite à mesure que des variables supplémentaires sont ajoutées au modèle de régression :

4 Violations des Hypothèses Classiques

Variables dans le modèle	MCR
X_1	7.95
X_1, X_2	6.47
X_1, X_2, X_3	6.15

De plus, la précision des valeurs ajustées dans la plage des observations sur les variables prédictives n'est pas érodée avec l'ajout de variables prédictives corrélées dans le modèle de régression. Considérer l'estimation de la graisse corporelle moyenne lorsque la seule variable prédictive dans le modèle est l'épaisseur du pli cutané des triceps (X_1) pour $X_{h1} = 25$. La valeur ajustée et son écart-type estimé sont (calculs non représentés) :

$$\hat{Y}_h = 19.93 \quad \hat{\sigma}_{\hat{Y}_h} = 0.632$$

Lorsque la variable prédictive hautement corrélée “circonférence de la cuisse hautement corrélée (X_2)” est également incluse dans le modèle, la graisse corporelle moyenne estimée et son écart-type estimé sont les suivants pour $X_{h1} = 25$ et $X_{h2} = 50$:

$$\hat{Y}_h = 19.36 \quad \hat{\sigma}_{\hat{Y}_h} = 0.624$$

Ainsi, la précision de la réponse moyenne estimée est tout aussi bonne qu'auparavant, malgré l'ajout de la deuxième variable prédictive qui est fortement corrélée avec la première. Cette stabilité dans la précision de la réponse moyenne estimée s'est produite malgré le fait que l'écart type estimé de $\hat{\beta}_1$ soit devenu beaucoup plus important lorsque X_2 a été ajouté au modèle (tableaux (4.5) et (4.6)). La raison essentielle de la stabilité est que la covariance entre $\hat{\beta}_1$ et $\hat{\beta}_2$ est négative, ce qui joue une forte influence contre l'augmentation de $\hat{\sigma}_{\beta_1}^2$ dans la détermination de la valeur de $\hat{\sigma}_{\hat{Y}_h}^2$ comme indiqué dans la section 3.11 du chapitre 3 (précisement dans la version algébrique de $\hat{\sigma}^2(\hat{Y}_h)$).

Lorsque les trois variables prédictives sont incluses dans le modèle, la graisse corporelle moyenne estimée et son écart-type estimé sont les suivants pour $X_{h1} = 25$, $X_{h2} = 50$ et $X_{h3} = 29$:

$$\hat{Y}_h = 19.19 \quad \hat{\sigma}_{\hat{Y}_h} = 0.621$$

Ainsi, l'ajout de la troisième variable prédictive, qui est fortement corrélée avec les deux premières variables prédictives ensemble, n'affecte pas non plus matériellement la précision de la réponse moyenne estimée.

Effets sur les Tests Simultanés de β_k

Un abus non rare dans l'analyse des modèles de régression multiple consiste à examiner la statistique t dans (3.57) :

$$t_{\hat{\beta}_k}^* = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}}$$

pour chaque coefficient de régression à tour de rôle il faut décider si $\beta_k = 0$ pour $k = 1, \dots, m$. Même si une procédure d'inférence simultanée est utilisée, et souvent elle ne l'est pas, des problèmes subsistent lorsque les variables prédictives sont fortement corrélées.

Supposons que nous souhaitons tester si $\beta_1 = 0$ et $\beta_2 = 0$ dans l'exemple la graisse corporelle avec deux variables prédictives du tableau (4.6c). En contrôlant le niveau de signification à 0.05, nous exigeons avec la méthode de Bonferroni¹ que chacun des deux tests de Student soit effectué avec un niveau de signification 0.025. Par conséquent, nous avons besoin de $t_{17}^{0.9875} = 2.46$. Étant donné que les deux statistiques de Student t^* du tableau (4.6c) ont des valeurs absolues qui ne dépassent pas 2.46, nous conclurons des deux tests distincts que $\beta_1 = 0$ et $\beta_2 = 0$. Pourtant, le test de Fisher approprié pour $H_0 : \beta_1 = \beta_2 = 0$ conduirait à accepter H_1 stipulant que les deux coefficients ne sont pas égaux à zéro. Cela peut être remarqué dans le tableau (4.6c), où nous trouvons $F^* = MCE/MCR = 192.72/6.47 = 29.8$, ce qui dépasse de loin $F_{2;17}^{0.05} = 3.59$.

La raison de ce résultat paradoxal est que chaque test de Student est un test marginal, comme nous l'avons vu dans (3.54) du point de vue de l'approche générale du test linéaire. Ainsi, une petite $SCE(X_1 | X_2)$ indique ici que X_1 ne fournit pas beaucoup d'informations supplémentaires au-delà de X_2 qui est déjà dans le modèle ; par conséquent, nous sommes amenés à la conclusion que $\beta_1 = 0$. De même, nous sommes amenés à conclure $\beta_2 = 0$ parce que $SCE(X_2 | X_1)$ est petite, indiquant que X_2 ne fournit pas beaucoup plus d'informations supplémentaires lorsque X_1 est déjà dans le modèle. Mais les deux tests des effets marginaux de X_1

1. La Méthode de Bonferroni permet de contrôler le niveau de confiance simultané de la totalité d'un ensemble d'intervalles de confiance. Il est important de prendre en considération le niveau de confiance simultané lors de l'examen de plusieurs intervalles de confiance, car la probabilité qu'au moins l'un des intervalles de confiance ne contienne pas le paramètre de population est plus grande pour un ensemble d'intervalles que pour n'importe quel intervalle unique. Pour contrecarrer ce taux d'erreur plus élevé, la méthode de Bonferroni ajuste le niveau de confiance de chacun des intervalles de manière à ce que le niveau de confiance simultané obtenu soit égal à la valeur que vous spécifiez.

et X_2 conjointement ne sont pas équivalents à tester s'il existe une relation de régression entre Y et les deux variables prédictives. La raison en est que le modèle réduit pour chacun des tests séparés contient l'autre variable prédictive, tandis que le modèle réduit pour tester si $\beta_1 = 0$ et $\beta_2 = 0$ ne contiendrait aucune variable prédictive. Le test de Fisher approprié montre qu'il existe ici une relation de régression définie entre Y et X_1 et X_2 .

Le même paradoxe se rencontrerait dans le tableau (4.6d) pour le modèle de régression à trois variables prédictives si trois tests simultanés sur les coefficients de régression étaient effectués au niveau de signification 0.05.

Remarques

- (1) Un déterminant proche de zéro de $X'X$ est une source potentielle d'erreurs d'arrondis graves dans les calculs des équations. Une multicolinéarité forte a pour effet de rendre ce déterminant proche de zéro. Ainsi, en cas d'une multicolinéarité forte, les coefficients de régression peuvent être sujets à de grandes erreurs d'arrondis ainsi qu'à de grandes variances d'échantillonnage. Par conséquent, il est judicieux d'utiliser la transformation de corrélation dans les calculs d'équations normales lorsque la multicolinéarité est présente.
- (2) Tout comme les fortes corrélations entre les variables prédictives tendent à rendre les coefficients de régression estimés imprécis (c.-à-d. erratiques d'un échantillon à l'autre), les coefficients de corrélation partielle entre la variable de réponse et chaque variable prédictive tendent à devenir erratiques d'un échantillon à l'autre lorsque les variables prédictives sont fortement corrélées.
- (3) L'effet des intercorrélations entre les variables prédictives sur les écarts-types des coefficients de régression estimés est facilement visible lorsque les variables du modèle sont transformées au moyen de la transformation de corrélation. Considérons le modèle de premier ordre avec deux variables prédictives :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i \quad (4.7)$$

Ce modèle avec les variables transformées en (3.102) devient :

$$Y_i^* = \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + e_i^* \quad (4.8)$$

La matrice $(X'X)^{-1}$ de ce modèle standardisé est donnée par :

4 Violations des Hypothèses Classiques

$$(X' X)^{-1} = r_{XX}^{-1} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \quad (4.9)$$

Par conséquent, la matrice variance-covariance des coefficients de régression estimés est donnée à partir de (3.50) et (4.9) comme :

$$\sigma^2 (\hat{\beta}) = (\sigma^*)^2 r_{XX}^{-1} = (\sigma^*)^2 \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \quad (4.10)$$

où $(\sigma^*)^2$ est la variance du terme d'erreur pour le modèle standarisé (4.8). On voit que les coefficients de régression estimés $\hat{\beta}_1^*$ et $\hat{\beta}_2^*$ ont ici la même variance :

$$\sigma^2 (\hat{\beta}_1^*) = \sigma^2 (\hat{\beta}_2^*) = \frac{(\sigma^*)^2}{1 - r_{12}^2} \quad (4.11)$$

et que chacune de ces variances augmente à mesure que la corrélation entre X_1 et X_2 augmente. En effet, à mesure que X_1 et X_2 approchent de la corrélation parfaite (c'est-à-dire que r_{12}^2 s'approche de 1), les variances de $\hat{\beta}_1^*$ et $\hat{\beta}_2^*$ deviennent plus grandes sans limite.

- (4)** Dans notre discussion sur les tests simultanés des coefficients de régression et dans le cas où il est possible qu'un ensemble de variables prédictives soit lié à la variable de réponse, lesdits tests individuels sur les coefficients de régression mèneront à la conclusion qu'ils sont égaux à zéro en raison de la multicolinéarité entre les variables prédictives. Ce résultat apparemment paradoxal est également possible dans des circonstances particulières lorsqu'il n'y a pas de multicolinéarité parmi les variables prédictives. Pourtant, il est peu probable que les circonstances particulières se retrouvent dans la pratique.

4.1.4 Diagnostic de Multicolinéarité - Facteur d'Inflation de la Variance

Lorsque nous avons discuté de la multicolinéarité, nous avons noté certains problèmes clés qui surviennent généralement lorsque les variables prédictives considérées pour le modèle de régression sont fortement corrélées entre elles :

4 Violations des Hypothèses Classiques

- (1) L'ajout ou la suppression d'une variable prédictive modifie les coefficients de régression.
- (2) La somme supplémentaire des carrés associée à une variable prédictive varie en fonction des autres variables prédictives déjà incluses dans le modèle.
- (3) Les écarts-types estimés des coefficients de régression deviennent grands lorsque les variables prédictives du modèle de régression sont fortement corrélées entre elles.
- (4) Les coefficients de régression estimés individuellement peuvent ne pas être statistiquement significatifs même s'il existe une relation statistique définie entre la variable de réponse et l'ensemble de variables prédictives.

Ces problèmes peuvent également survenir en l'absence d'une multicolinéarité substantielle, mais uniquement dans des circonstances inhabituelles qui ne sont pas susceptibles d'être trouvées dans la pratique. Nous considérons d'abord certains diagnostics informels de la multicolinéarité, puis un diagnostic formel très utile, le *Facteur d'Inflation de la Variance*.

Diagnostics Informels

Les indications de la présence d'une multicolinéarité grave sont fournies par les diagnostics informels suivants :

- (1) Des changements importants dans les coefficients de régression estimés lorsqu'une variable prédictive est ajoutée ou supprimée, ou lorsqu'une observation est modifiée ou supprimée.
- (2) Des résultats non significatifs dans les tests individuels sur les coefficients de régression pour les variables prédictives les plus importantes.
- (3) Des coefficients de régression estimés avec un signe algébrique contraire à celui attendu des considérations théoriques ou de l'expérience antérieure.
- (4) De grands coefficients de corrélation simple entre des paires de variables prédictives dans la matrice de corrélation r_{XX} .
- (5) Un coefficient R^2 élevé mais avec peu de ratios t significatifs.

Exemple

Nous considérons à nouveau l'exemple de graisse corporelle du tableau (4.4), cette fois avec les trois variables prédictives : l'épaisseur du

4 Violations des Hypothèses Classiques

pli cutané des triceps (X_1), la circonférence des cuisses (X_2) et la circonférence des bras (X_3). Nous avons déjà noté que les variables prédictives “l'épaisseur du pli cutané du triceps” et “la circonférence de la cuisse” sont fortement corrélées. Nous avons également noté d'importants changements dans les coefficients de régression estimés et leurs écarts-types estimés lorsqu'une variable a été ajoutée, aussi des résultats non significatifs dans les tests individuels sur les variables importantes anticipées et un coefficient négatif estimé lorsqu'un coefficient positif était attendu. Ce sont toutes des indications informelles qui suggèrent une multicolinéarité sérieuse parmi les variables prédictives.

Remarque

Les méthodes informelles qui viennent d'être décrites présentent des limites importantes. Ils ne fournissent pas des mesures quantitatives de l'impact de la multicolinéarité et ils peuvent ne pas identifier la nature de la multicolinéarité. Par exemple, si les variables prédictives X_1 , X_2 et X_3 ont de faibles corrélations par paires, alors l'analyse des coefficients de corrélation simple peut ne pas révéler l'existence des relations entre les groupes des variables prédictives comme par exemple l'existence d'une corrélation élevée entre X_1 et une combinaison linéaire de X_2 et X_3 .

Une autre limitation des méthodes de diagnostic informel est que parfois le comportement observé peut se produire sans multicolinéarité.

Facteur d'Inflation de la Variance

Une méthode formelle de détection de la présence d'une multicolinéarité consiste à utiliser des facteurs d'inflation de la variance. Ces facteurs donnent dans quelle mesure les variances des coefficients de régression estimés sont gonflées par rapport à celles dans le cas où les variables prédictives ne sont pas linéairement liées.

Pour comprendre l'importance des facteurs d'inflation de la variance, nous commençons par la précision des coefficients de régression estimés (MCO), qui est mesurée par leurs variances. Nous savons à partir de (3.48) que la matrice variance-covariance des coefficients de régression estimés est :

$$\sigma^2 (\hat{\beta}) = \sigma^2 \times (X' X)^{-1} \quad (4.12)$$

Pour mesurer l'impact de la multicolinéarité, il est utile de travailler avec le modèle de régression standardisé (3.103), qui est obtenu en transformant les variables au moyen de la transformation de corrélation (3.102). Lorsque le modèle de régression standardisé est ajusté, les coefficients de régression estimés $\hat{\beta}_k^*$ sont des coefficients standardisés qui sont liés aux coefficients de régression estimés pour les variables non

4 Violations des Hypothèses Classiques

transformées selon (3.104) et (3.105). La matrice variance-covariance des coefficients de régression standardisés estimés est obtenue à partir de (4.12) en utilisant le résultat de (3.109), qui indique que la matrice $X'X$ pour les variables transformées est la matrice de corrélation des variables X soit r_{XX} . On obtient donc :

$$\sigma^2(\hat{\beta}^*) = (\sigma^*)^2 \times r_{XX}^{-1} = (\sigma^*)^2 \begin{bmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & & 1 \end{bmatrix}^{-1} \quad (4.13)$$

où r_{XX} est la matrice des coefficients de corrélation simple par paire des variables X , et $(\sigma^*)^2$ est la variance du terme d'erreur pour le modèle transformé. Notez que à partir de (4.13), la variance de $\hat{\beta}_k^*$ ($k = 1, \dots, m$) est égale à ce qui suit, en posant $(VIF)_k$ le k ième élément diagonal de la matrice r_{XX}^{-1} :

$$\sigma^2(\hat{\beta}^*) = (\sigma^*)^2 (VIF)_k \quad (4.14)$$

L'élément diagonal (VIF) est appelé *facteur d'inflation de la variance* (VIF) pour $\hat{\beta}_k^*$. On peut montrer que ce facteur d'inflation de la variance est égal à :

$$(VIF)_k = (1 - R_k^2)^{-1} \quad k = 1, 2, \dots, m \quad (4.15)$$

où R_k^2 est le coefficient de détermination multiple lorsque X_k est régressé sur les $m - 1$ autres variables du modèle. Nous avons donc :

$$\sigma^2(\hat{\beta}_k^*) = \frac{(\sigma^*)^2}{1 - R_k^2} \quad (4.16)$$

Nous avons présenté en (4.11) les résultats spéciaux pour $\sigma^2(\hat{\beta}_k^*)$ lorsque $m = 2$, et où $R_k^2 = r_{12}^2$ est le coefficient de détermination simple entre X_1 et X_2 .

Le facteur d'inflation de la variance (VIF) est égal à 1 lorsque $R_k^2 = 0$, c'est-à-dire lorsque X_k n'est pas lié linéairement aux autres variables X . Lorsque $R_k^2 \neq 0$, alors (VIF) est supérieur à 1, indiquant une variance gonflée pour $\hat{\beta}_k^*$; en raison des intercorrélations entre les variables X . Lorsque X_k a une association linéaire parfaite avec les autres variables X du modèle de sorte que $R_k^2 = 1$, alors $(VIF)_k$ et $\sigma^2(\hat{\beta}_k^*)$ sont non bornés.

Utilisations Diagnostiques

La plus grande valeur de VIF parmi toutes les variables X est souvent utilisée comme indicateur de la gravité de la multicolinéarité. Une valeur VIF maximale supérieure à 10 est souvent considérée comme une indication que la multicolinéarité peut influencer indûment les estimations des moindres carrés.

La moyenne des valeurs VIF fournit également des informations sur la gravité de la multicolinéarité en termes de distance entre les coefficients de régression normalisés $\hat{\beta}_k^*$ estimés et les valeurs réelles β_k^* . On peut montrer que la valeur attendue de la somme de ces erreurs au carré $(\hat{\beta}_k^* - \beta_k^*)^2$ est donnée par :

$$E \left\{ \sum_{k=1}^m (\hat{\beta}_k^* - \beta_k^*)^2 \right\} = (\sigma^*)^2 \sum_{k=1}^m (VIF)_k \quad (4.17)$$

Ainsi, des valeurs VIF élevées entraînent, en moyenne, de plus grandes différences entre les coefficients de régression normalisés estimés et réels.

Lorsqu'aucune variable X n'est liée linéairement aux autres dans le modèle de régression, $R_k^2 = 0$; par conséquent, $(VIF)_k \equiv 1$, leur somme est m , et la valeur attendue de la somme des erreurs quadratiques est :

$$E \left\{ \sum_{k=1}^m (\hat{\beta}_k^* - \beta_k^*)^2 \right\} = (\sigma^*)^2 m \quad \text{quand } (VIF)_k \equiv 1 \quad (4.18)$$

Un rapport des résultats en (4.17) et (4.18) fournit des informations utiles sur l'effet de la multicolinéarité sur la somme des carrés des erreurs :

$$\frac{(\sigma^*)^2 \sum (VIF)_k}{(\sigma^*)^2 m} = \frac{\sum (VIF)_k}{m} \quad (4.19)$$

Notez que ce rapport est simplement la moyenne des valeurs VIF , à noter par (\bar{VIF}) :

$$(\bar{VIF}) = \frac{\sum_{k=1}^m (VIF)_k}{m} \quad (4.20)$$

Les valeurs moyennes de VIF considérablement supérieures à 1 indiquent de graves problèmes de multicolinéarité.

Exemple

4 Violations des Hypothèses Classiques

Le tableau (4.7) contient les coefficients de régression normalisés estimés et les valeurs VIF pour l'exemple de graisse corporelle avec trois variables prédictives (les calculs sont non illustrés). Le maximum des valeurs VIF est 708.84 et leur valeur moyenne est $(\bar{VIF}) = 459.26$. Ainsi, la somme attendue des carrés des erreurs dans la régression normalisés des moindres carrés est près de 460 fois plus grande qu'elle ne le serait si les variables X n'étaient pas corrélées. De plus, les trois valeurs VIF dépassent largement 10, ce qui indique à nouveau qu'il existe de graves problèmes de multicolinéarité.

Variable	$\hat{\beta}_k^*$	$(VIF)_k$
X_1	4.2637	708.84
X_2	-2.9287	564.34
X_3	-1.5614	104.61
Maximum $(VIF)_k = 708.84$		$(\bar{VIF}) = 459.26$

TABLE 4.7 – Facteurs d’Inflation de la Variance - Exemple de graisse corporelle avec trois variables prédictives

Il est intéressant de noter que $(VIF)_3 = 105$ malgré le fait que r_{13}^2 et r_{23}^2 (voir figure (??)) ne sont pas grands. Voici un exemple où X_3 est étroitement lié à X_1 et X_2 ensemble ($R_3^2 = 0.99$), même si les coefficients de détermination simple par paire ne sont pas importants. L'examen des corrélations par paires ne révèle pas cette multicolinéarité.

- (1) Certains logiciels informatiques utilisent l'inverse du facteur d'inflation de la variance pour détecter les cas où une variable X ne devrait pas être autorisée dans le modèle de régression ajusté en raison d'une interdépendance excessivement élevée entre cette variable et les autres variables X du modèle. Les limites de tolérance pour $1/(VIF)_k = 1 - R_k^2$ fréquemment utilisées sont 0.01, 0.001 ou 0.0001, en dessous desquelles la variable n'est pas entrée dans le modèle.
- (2) Une limitation des facteurs d'inflation de variance pour détecter les multicolinéarités est qu'ils ne peuvent pas faire la distinction entre plusieurs multicolinéarités simultanées.

4.1.5 Mesures Correctives

Losqu'on a une présence forte de la multicolinéarité ? Nous avons deux choix : (1) ne rien faire ou (2) suivre quelques mesures correctives

4 Violations des Hypothèses Classiques

comme :

Information a Priori

Avoir des informations sur les relations qui peuvent exister entre les variables explicatives. Cela pourrait provenir de travaux empiriques antérieurs dans lesquels le problème de colinéarité est moins grave ou de la théorie pertinente.

Exemple : On considère le modèle de consommation suivant

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

où Y = consommation, X_1 = revenu et X_3 = richesse. Comme indiqué précédemment, les variables « revenu » et « richesse » ont tendance à être très colinéaires. Supposons a priori que nous croyons que $\beta_2 = \alpha\beta_1$. On peut alors exécuter la régression suivante :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \alpha\beta_1 X_{i2} + \varepsilon_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

avec $X_i = X_{i1} + \alpha X_{i2}$.

Combinaison de Données Transversales et de Séries Chronologiques

Une technique d'information a priori est la combinaison de données transversales et de séries chronologiques, appelée « *pooling the data* ».

Suppression d'Une ou de Plusieurs Variables et Biais de Spécification

Face à une multicolinéarité sévère, l'une des choses les plus « simples » à faire est de supprimer l'une des variables colinéaires. Ainsi, dans notre exemple précédent de consommation – revenu – richesse, la suppression d'une variable du modèle, peut mener à un biais de spécification ou une erreur de spécification. Le biais de spécification provient d'une spécification incorrecte du modèle utilisé dans l'analyse. Ainsi, si la théorie économique dit que le revenu et la richesse devraient être inclus tous les deux dans le modèle expliquant les dépenses de consommation, la suppression de la variable de richesse constituerait un biais de spécification.

Transformation des Variables

Supposons que nous disposions des séries chronologiques sur les dépenses de consommation, les revenus et la richesse. L'une des raisons

4 Violations des Hypothèses Classiques

de la multicolinéarité élevée entre le revenu et la richesse dans ces données est qu'au fil du temps, les deux variables ont tendance à évoluer dans la même direction. On peut procéder par une première différence de l'équation de régression afin de minimiser cette dépendance.

si la relation est de la forme suivante :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

L'équation de première différence à estimer sera :

$$Y_t - Y_{t-1} = \beta_1 (X_{t1} - \rho X_{t-1,1}) + \beta_2 (X_{t2} - \rho X_{t-1,2}) + (\varepsilon_t - \rho \varepsilon_{t-1})$$

Données Supplémentaires ou Nouvelles

Étant donné que la multicolinéarité est une caractéristique de l'échantillon, il est possible que dans un autre échantillon impliquant les mêmes variables, la colinéarité ne soit pas aussi grave que dans le premier échantillon. Parfois, l'augmenter de la taille de l'échantillon (si possible) peut atténuer le problème de colinéarité.

Autres Méthodes pour Remédier à la Multicolinéarité

Des techniques statistiques multivariées telles que l'analyse factorielle et l'analyse en composantes principales (ACP) ou techniques telles que la régression de crête (Ridge Regression) sont souvent utilisées pour «résoudre» le problème de la multicolinéarité.

4.2 Autocorrélation des Erreurs

La violation de l'hypothèse 3 signifie que les erreurs sont corrélées, c'est-à-dire $E(\varepsilon_i \varepsilon_j) = \sigma_{ij} \neq 0$, pour $i \neq j$, et $i, j = 1, 2, \dots, n$. Puisque ε_i a une moyenne nulle, alors $E(\varepsilon_i \varepsilon_j) = \text{cov}(\varepsilon_i, \varepsilon_j)$ notée σ_{ij} . Cette corrélation est plus susceptible de se produire dans les séries chronologiques que dans les études transversales.

Lorsque les données sont classées par ordre chronologique, l'erreur sur une période peut affecter l'erreur sur la (ou) les périodes suivantes (ou autres). Il est très probable qu'il y aura des intercorrélations entre les observations successives, en particulier lorsque l'intervalle est court, comme les fréquences quotidiennes, hebdomadaires ou mensuelles, par rapport à un ensemble de données transversales. Par exemple, une augmentation inattendue de la confiance des consommateurs peut amener

une équation de la fonction de consommation à sous-estimer la consommation pendant deux périodes ou plus. Dans les données transversales, le problème de l'autocorrélation est moins susceptible d'exister parce que nous pouvons facilement changer la disposition des données sans altérer significativement les résultats. (Ce n'est pas vrai dans le cas de l'autocorrélation spatiale, mais cela dépasse le cadre de ce texte.)

4.2.1 Qu'est-ce qui Cause l'Autocorrélation ?

Le premier facteur qui peut provoquer une autocorrélation est l'omission des variables. Supposons que Y_t soit lié à X_{t1} et X_{t2} mais et par erreur, nous n'incluons pas X_{t2} dans notre modèle. L'effet de X_{t2} sera capturé par les perturbations ε_t . Si X_{t2} , comme dans de nombreuses séries chronologiques économiques, dépend de $X_{t-1,2}$, $X_{t-2,2}$ et ainsi de suite, cela conduira à une corrélation inévitable entre ε_t et ε_{t-1} , ε_{t-2} et ainsi de suite, donc les variables omises peuvent être une cause d'autocorrélation.

L'autocorrélation peut également se produire en raison d'une mauvaise spécification du modèle. Supposons que Y_t soit connecté à X_{t1} avec une relation quadratique $Y = \beta_0 + \beta_1 X_{t1}^2 + \varepsilon_t$, mais nous supposons et estimons à tort une droite $Y = \beta_0 + \beta_1 X_{t1} + \varepsilon_t$. Ensuite, le terme d'erreur obtenu à partir de la spécification linéaire dépendra de X_{t1}^2 . Si X_{t1} augmente ou diminue avec le temps, ε_t fera de même, indiquant une autocorrélation.

Un troisième facteur est les erreurs systématiques de mesure. Supposons qu'une entreprise met à jour son inventaire à une période donnée ; si une erreur systématique se produit dans sa mesure, le stock d'inventaire cumulé présentera des erreurs de mesure accumulées. Ces erreurs apparaîtront comme une procédure autocorrélée.

Nous changeons les indices i et j par t et s pour les observations en séries temporelles $t, s = 1, 2, \dots, T$ et la taille de l'échantillon sera notée T plutôt que n . Ce terme de covariance est symétrique, de sorte que $\sigma_{12} = E(\varepsilon_1 \varepsilon_2) = E(\varepsilon_2 \varepsilon_1) = \sigma_{21}$. Par conséquent, seuls $T(T - 1)/2$ termes σ_{ts} distincts doivent être estimés. Par exemple, si $T = 3$, alors σ_{12} , σ_{13} et σ_{23} sont les termes de covariance distincts. Cependant, il est inutile d'estimer $T(T - 1)/2$ covariances σ_{ts} avec seulement T observations. Par conséquent, plus de structure sur ces σ_{ts} doit être imposée. Une hypothèse populaire est que les ε_t suivent un processus autorégressif de premier ordre noté $AR(1)$:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad t = 1, 2, \dots, T \quad (4.21)$$

4 Violations des Hypothèses Classiques

où $|\rho| < 1$ et $u_t \sim IID(0, \sigma_u^2)$ est indépendant et identiquement distribué. Il est autorégressif car u_t est lié à sa valeur décalée u_{t-1} .

Il est évident que la taille de ρ déterminera la force de la corrélation sérielle, et nous pouvons différencier trois cas :

- (a) Si ρ est nul, alors nous n'avons pas de corrélation sérielle, car $\varepsilon_t = u_t$ et donc un terme d'erreur IID .
- (b) Si ρ se rapproche de $+1$, la valeur de l'observation précédente de l'erreur (u_{t-1}) devient plus importante dans la détermination de la valeur du terme d'erreur actuel (u_t) et donc l'existence d'une grande corrélation sérielle positive. Dans ce cas, l'observation actuelle du terme d'erreur a tendance à avoir le même signe que l'observation précédente du terme d'erreur (c'est-à-dire négatif conduira au négatif et le positive conduira au positive). C'est ce qu'on appelle une corrélation sérielle positive. La figure (4.3) montre comment les résidus d'un cas de corrélation sérielle positive apparaissent.
- (c) Si ρ s'approche de -1 , la force de la corrélation sérielle sera évidemment très élevée. Cette fois, cependant, nous avons une corrélation sérielle négative. Une corrélation sérielle négative implique qu'il existe un comportement semblable à celui d'une dent de scie dans le diagramme temporel des termes d'erreur. Les signes des termes d'erreur ont tendance à passer du négatif au positif et vice versa dans les observations consécutives. La figure (4.4) illustre le cas d'une corrélation sérielle négative. En général, en économie, une corrélation sérielle négative est beaucoup moins susceptible de se produire qu'une corrélation sérielle positive.

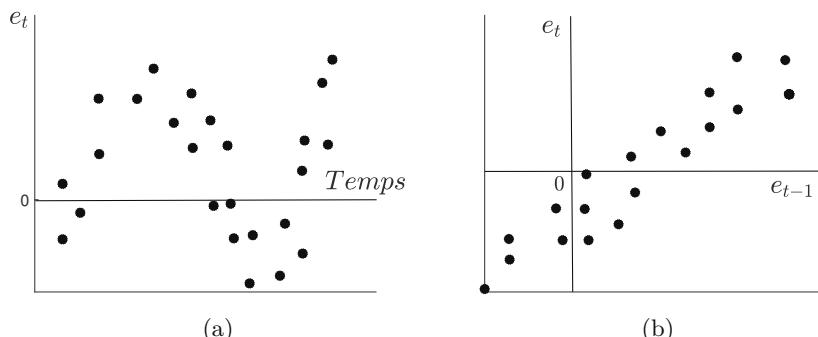


FIGURE 4.3 – Corrélation Sérielle Positive

4 Violations des Hypothèses Classiques

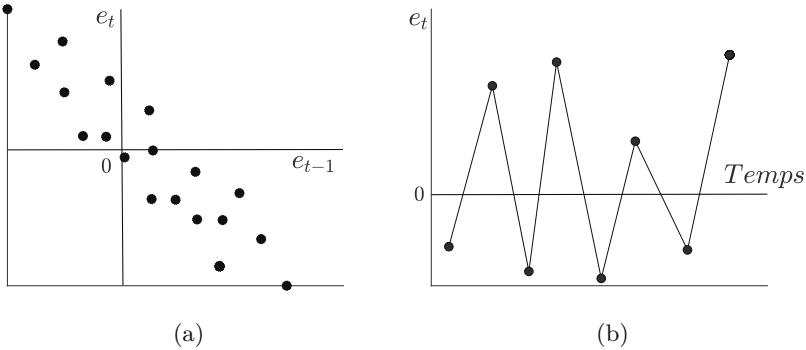


FIGURE 4.4 – Corrélation Sérielle Négative

On peut aussi écrire à partir de (4.21), pour la période $t - 1$

$$\varepsilon_{t-1} = \rho \varepsilon_{t-2} + u_{t-1} \quad (4.22)$$

puis on remplace (4.22) dans (4.21) pour obtenir

$$\varepsilon_t = \rho^2 \varepsilon_{t-2} + \rho u_{t-1} + u_t \quad (4.23)$$

Notez que la puissance de ρ et l'indice de ε ou u sont toujours sommés à t . Par substitution continue de cette forme, on obtient finalement

$$\varepsilon_t = \rho^t \varepsilon_0 + \rho^{t-1} u_1 + \dots + \rho u_{t-1} + u_t \quad (4.24)$$

Cela signifie que ε_t est une fonction des valeurs actuelles et passées de u_t et ε_0 où ε_0 est la valeur initiale de ε_t . Si ε_0 a une moyenne nulle, alors ε_t a une moyenne nulle. Cela découle de (4.24) en prenant les espérances. Aussi, depuis (4.21)

$$var(\varepsilon_t) = \rho^2 var(\varepsilon_{t-1}) + var(u_t) + 2\rho cov(\varepsilon_{t-1}, u_t) \quad (4.25)$$

En utilisant (4.24), ε_{t-1} est une fonction de u_{t-1} , des valeurs passées de u_{t-1} et de ε_0 . Puisque ε_0 est indépendant des u 's et que les u 's ne sont pas eux-mêmes corrélés en série, alors ε_{t-1} est indépendant de u_t . Cela signifie que $cov(\varepsilon_{t-1}, u_t) = 0$. De plus, pour que ε_t soit homoscédastique, $var(\varepsilon_t) = var(\varepsilon_{t-1}) = \sigma_\varepsilon^2$ et (4.25) se réduit à $\sigma_\varepsilon^2 = \rho^2 \sigma_\varepsilon^2 + \sigma_u^2$, ce qui, une fois résolu pour σ_ε^2 , donne :

$$\sigma_\varepsilon^2 = \frac{\sigma_u^2}{1 - \rho^2} \quad (4.26)$$

4 Violations des Hypothèses Classiques

Par conséquent, $\varepsilon_0 \sim (0, \sigma_u^2 / (1 - \rho^2))$ pour que les ε 's aient une moyenne nulle et des perturbations homoscédastiques. En multipliant (4.21) par ε_{t-1} et en prenant les espérances, on obtient

$$E(\varepsilon_t \varepsilon_{t-1}) = \rho E(\varepsilon_{t-1}^2) + E(\varepsilon_{t-1} u_t) = \rho \sigma_\varepsilon^2 \quad (4.27)$$

Puisque $E(\varepsilon_{t-1}^2) = \sigma_\varepsilon^2$ et $E(\varepsilon_{t-1} u_t) = 0$, alors $cov(\varepsilon_t, \varepsilon_{t-1}) = \rho \sigma_\varepsilon^2$, et le coefficient de corrélation entre ε_t et ε_{t-1} est

$$correl(\varepsilon_t, \varepsilon_{t-1}) = cov(\varepsilon_t, \varepsilon_{t-1}) / \sqrt{var(\varepsilon_t) var(\varepsilon_{t-1})} = \rho \sigma_\varepsilon^2 / \sigma_\varepsilon^2 = \rho$$

Puisque ρ est un coefficient de corrélation, cela signifie que $-1 \leq \rho \leq 1$. En général, on peut montrer que

$$cov(\varepsilon_t, \varepsilon_s) = \rho^{|t-s|} \sigma_\varepsilon^2 \quad t, s = 1, 2, \dots, T \quad (4.28)$$

Cela signifie que la corrélation entre ε_t et ε_{t-r} est ρ^r , qui est une fraction élevée à une puissance entière, c'est-à-dire que la corrélation se dégrade entre les perturbations plus elles sont éloignées. Ceci est raisonnable en économie et peut être la raison pour laquelle cette forme autorégressive (4.21) est si populaire. Il convient de noter que ce n'est pas la seule forme qui corrélera les perturbations dans le temps mais il y a d'autres formes comme le processus de Moyenne Mobile (MA) et les processus Autorégressifs Moyennes Mobiles (ARMA).

4.2.2 Conséquences pour les Moindres Carrés Ordinaires

Comment l'estimateur des MCO est-il affecté par la violation de l'hypothèse d'absence d'autocorrélation des erreurs ?. *L'estimateur des MCO est toujours non biaisé et cohérent* puisque ces propriétés reposent sur les hypothèses 1 et 5 et n'ont rien à voir avec l'hypothèse 3. Pour la régression linéaire simple, en utilisant (2.5), la variance de $\hat{\beta}$ est maintenant

$$\begin{aligned} var(\hat{\beta}) &= var\left(\sum_{t=1}^T w_t \varepsilon_t\right) = \sum_{t=1}^T \sum_{s=1}^T w_t w_s cov(\varepsilon_t, \varepsilon_s) \\ &= \frac{\sigma_\varepsilon^2}{\sum_{t=1}^T (X_t - \bar{X})^2} + \sum_{t \neq s} w_t w_s \rho^{|t-s|} \sigma_\varepsilon^2 \end{aligned} \quad (4.29)$$

4 Violations des Hypothèses Classiques

$$\text{où } w_t = \frac{\sum_{t=1}^T (X_t - \bar{X}) \varepsilon_t}{\sum_{t=1}^T (X_t - \bar{X})^2} \text{ et } \text{cov}(\varepsilon_t, \varepsilon_s) = \rho^{|t-s|} \sigma_\varepsilon^2 \text{ comme expliqué dans (4.28).}$$

Notez que le premier terme de (4.29) est la variance habituelle de $\hat{\beta}$ dans le cas classique. Le deuxième terme de (4.29) apparaît en raison de la corrélation entre les ε_t . Par conséquent, la variance des MCO calculée (c.-à-d. $\hat{\sigma}^2 / \sum_{t=1}^T (X_t - \bar{X})^2$), est une mauvaise estimation de la variance de $\hat{\beta}$ pour deux raisons. Premièrement, une mauvaise formule utilisée pour la variance, c'est-à-dire $\sigma_\varepsilon^2 / \sum_{t=1}^T (X_t - \bar{X})^2$ au lieu de celle en (4.29). Cette dernière dépend de ρ à travers le terme supplémentaire dans (4.29). Deuxièmement, on peut montrer que $E(\hat{\sigma}^2) \neq \sigma_\varepsilon^2$ et impliquera ρ ainsi que σ_ε^2 .

Par conséquent, $\hat{\sigma}^2$ n'est pas un estimateur sans biais pour σ_ε^2 et $\hat{\sigma}^2 / \sum_{t=1}^T (X_t - \bar{X})^2$ est une estimation biaisée de $\text{var}(\hat{\beta})$. L'ampleur de ce biais dépendent de ρ et du régresseur. En fait, si ρ est positif et que les $(X_t - \bar{X})$ sont eux-mêmes positivement autocorrélés, alors $\hat{\sigma}^2 / \sum_{t=1}^T (X_t - \bar{X})^2$ sous-estime la vraie variance de $\hat{\beta}$. Cela signifie que l'intervalle de confiance pour β est plus serré qu'il ne devrait l'être et la statistique t^* de Student pour $H_0 : \beta = 0$ est exagérée. Comme dans le cas hétéroscédastique, mais pour des raisons complètement différentes, toute inférence basée sur $\text{var}(\hat{\beta})$ rapportée par la régression standarde sera trompeuse si les ε_t 's sont corrélés en série.

Newey et West (1987) ont suggéré une simple hétéroscédasticité et une matrice covariance avec une autocorrélation cohérente pour l'estimateur des MCO sans spécifier la forme fonctionnelle de la corrélation en série. L'idée de base étend le remplacement de White (1980) des variances hétéroscédastiques par des résidus des MCO au carré et en incluant en outre les produits des résidus des moindres carrés $e_t e_{t-s}$ pour $s = 0, \pm 1, \dots, \pm p$ où p est l'ordre maximal de corrélation en série nous sont prêts à assumer. La cohérence de cette procédure repose sur le fait que p est très petit par rapport au nombre d'observations T . Ceci est cohérent avec les spécifications de corrélation en série populaires considérées dans ce chapitre où l'autocorrélation diminuent rapidement lorsque j augmente. Newey et West (1987) permettent aux termes de covariance d'ordre supérieur de recevoir des poids décroissants. Cette option de Newey-West pour l'estimateur des moindres carrés est dis-

4 Violations des Hypothèses Classiques

ponible à l'aide du logiciel EViews. Andrews (1991) prévient contre le manque de fiabilité de telles corrections d'erreur standard dans certaines circonstances. Wooldridge (1991) montre qu'il est possible de construire des statistiques de Fisher robustes corrélées en série pour effectuer les tests d'hypothèses.

4.2.3 Détection de l'Autocorrélation

Une façon simple de détecter l'autocorrélation consiste à examiner si les diagrammes résiduels en fonction du temps et le diagramme de dispersion de e_t contre e_{t-1} présentent des profils similaires à ceux présentés dans les figures (4.3) et (4.4). Dans de tels cas, nous disons que nous avons des preuves de corrélation sérielle positive si le modèle est similaire à celui de la figure (4.3), et de corrélation sérielle négative s'il est similaire à celle de la figure (4.4). Un exemple avec des données réelles est donné ci-dessous.

Exemple Empirique

Le tableau (4.8) donne les dépenses de consommation personnelle réelle (Y) et le revenu personnel disponible réel (X) pour une période de 49 ans. La figure (4.5) trace les valeurs réelles, ajustées et résiduelles à l'aide d'EViews (Le graphe est obtenu à partir de la barre de menu de la fenêtre des résultats de la régression de Y sur X dans le tableau (4.9)). Cela montre une corrélation sérielle positive avec une chaîne de résidus positifs suivie d'une chaîne de résidus négatifs suivie de résidus positifs.

Le Test de Durbin-Watson

Le test statistique le plus fréquemment utilisé pour la présence d'une corrélation en série est le test de Durbin-Watson (DW) (1950), qui est valide lorsque les hypothèses suivantes sont remplies :

- (a) Le modèle de régression comprend une constante ;
- (b) La corrélation sérielle est supposée être de premier ordre seulement ;
- (c) L'équation n'inclut pas une variable dépendante retardée comme une variable explicative.

Considérer le modèle :

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \dots + \beta_m X_{tm} + \varepsilon_t \quad (4.30)$$

où

4 Violations des Hypothèses Classiques

Y = Dépenses de consommation personnelle réelle

X = Revenu personnel disponible réel

Année	Y	X	Année	Y	X	Année	Y	X
1971	8776	9685	1988	13919	15738	2005	19593	21493
1972	8837	9735	1989	14364	16128	2006	20082	21812
1973	8873	9901	1990	14837	16704	2007	20382	22153
1974	9170	10227	1991	15030	16931	2008	20835	22546
1975	9412	10455	1992	14816	16940	2009	21365	23065
1976	9839	11061	1993	14879	17217	2010	22183	24131
1977	10331	11594	1994	14944	17418	2011	23050	24564
1978	10793	12065	1995	15656	17828	2012	23862	25472
1979	10994	12457	1996	16343	19011	2013	24215	25697
1980	11510	12892	1997	17040	19476	2014	24632	26238
1981	11820	13163	1998	17570	19906	2015	25073	26566
1982	11955	13563	1999	17994	20072	2016	25750	27274
1983	12256	14001	2000	18554	20740	2017	26290	27403
1984	12868	14512	2001	18898	21120	2018	26835	28098
1985	13371	15345	2002	19067	21281	2019	27319	28614
1986	13148	15094	2003	18848	21109			
1987	13320	15291	2004	19208	21548			

TABLE 4.8 – Données de Consommation

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad |\rho| < 1 \quad (4.31)$$

Ensuite sous l'hypothèse nulle $H_0 : \rho = 0$ le test de DW implique les étapes suivantes :

Etape 1 Estimer le modèle en utilisant MCO et obtenir les résidus e_t .

Etape 2 Calculez la statistique du test de DW donnée par :

$$d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2} \quad (4.32)$$

Etape 3 Construire le tableau de DW celui dans la figure (4.6), en remplaçant par d_U , d_L , $4-d_U$ et $4-d_L$ calculés à partir du tableau des valeurs critiques DW donné dans les annexes. Noter que le tableau des valeurs critiques est fonction de m , qui est le nombre

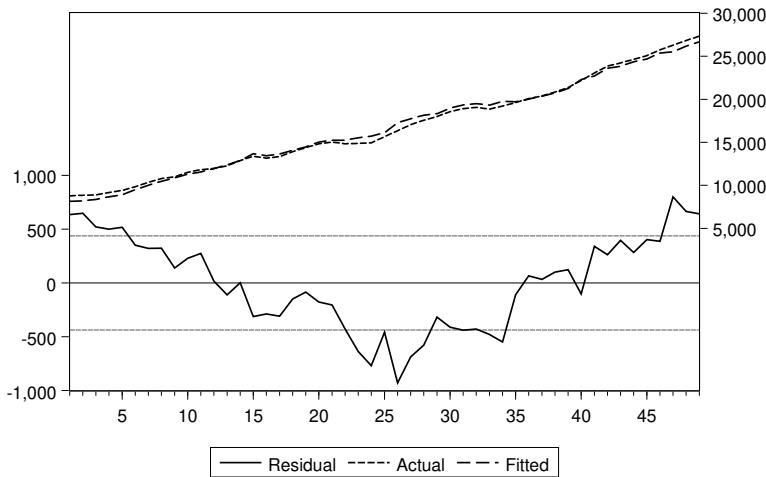


FIGURE 4.5 – Graphique des Résidus : Régression de la Consommation

de variables explicatives.

Etape 4 Pour tester la corrélation sérielle positive, les hypothèses sont les suivantes :

$$H_0; \rho = 0 \text{ pas d'autocorrélation}$$

$$H_a; \rho > 0 \text{ autocorrélation positive}$$

- 1 Si $d \leq d_L$, nous rejetons H_0 et nous affirmons l'existence d'une corrélation sérielle positive.
- 2 Si $d \geq d_U$, nous exceptons H_0 et il n'y a donc pas de corrélation sérielle positive.
- 3 Dans le cas particulier où $d_L < d < d_U$, le test n'est pas concluant.

Etape 5 Pour tester la corrélation sérielle négative, les hypothèses sont les suivantes :

$$H_0; \rho = 0 \text{ pas d'autocorrélation}$$

$$H_a; \rho < 0 \text{ autocorrélation négative}$$

- 1 Si $d \geq 4 - d_L$, nous rejetons H_0 et nous affirmons l'existence d'une corrélation sérielle négative.

4 Violations des Hypothèses Classiques

Dependent Variable : Y

Method : Least Squares

Sample (adjusted) : 1971 2019

Included observations : 49

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1343.314	219.5614	-6.118168	0.0000
X	0.979228	0.011392	85.96093	0.0000
R-squared	0.993680	Mean dependent var	16749.10	
Adjusted R-squared	0.993545	S.D. dependent var	5447.060	
S.E. of regression	437.6277	Akaike info criterion	15.04057	
Sum squared resid	9001348.	Schwarz criterion	15.11779	
Log likelihood	-366.4941	Hannan-Quinn criter.	15.06987	
F-statistic	7389.281	Durbin-Watson stat	0.180503	
Prob(F-statistic)	0.000000			

TABLE 4.9 – Régression de la Consommation sur le Revenu

- 2 Si $d \leq 4 - d_U$, nous exceptons H_0 et il n'y a donc pas de corrélation sérielle négative.
- 3 Dans le cas particulier où $4 - d_U < d < 4 - d_L$, le test n'est pas concluant.

La raison de l'inconvénient du test de DW est que la petite taille de l'échantillon pour la statistique DW dépend des variables X et elle est généralement difficile à déterminer. Une procédure de test préférée est le test LM , qui sera décrit après.

Règle importante pour le Test de DW

À partir des résidus estimés, nous pouvons obtenir une estimation de ρ comme :

$$\hat{\rho} = \frac{\sum_{t=2}^T e_t e_{t-1}}{\sum_{t=1}^T e_t^2} \quad (4.33)$$

Il est montré en annexe de ce chapitre que la statistique DW est approximativement égale à $d = 2(1 - \hat{\rho})$. Parce que ρ par définition varie de -1 à 1 , la plage de d sera de 0 à 4 . Par conséquent, nous pouvons avoir trois cas différents :

4 Violations des Hypothèses Classiques

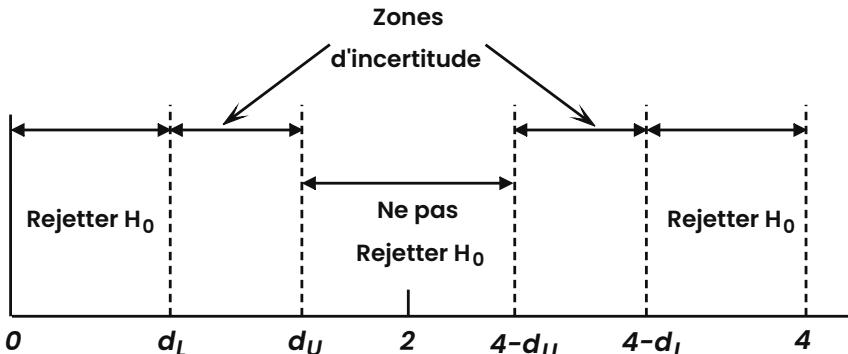


FIGURE 4.6 – **Test de Durbin Watson**

- (a) $\rho = 0 ; d = 2$: par conséquent, une valeur de d proche de 2 indique qu'il n'y a aucune preuve de corrélation sérielle.
- (b) $\rho \simeq 1 ; d \simeq 0$: une forte autocorrélation positive signifie que ρ sera proche de +1, et donc d aura des valeurs très faibles (proches de zéro) pour une autocorrélation positive.
- (c) $\rho \simeq -1 ; d \simeq 4$: de même, lorsque ρ est proche de -1, alors d sera proche de 4, indiquant une forte corrélation sérielle négative.

D'après cette analyse, nous pouvons voir que, en règle générale, lorsque la statistique du test de DW est très proche de 2, nous n'avons pas de corrélation sérielle.

Le Test de DW avec EViews et Stata

EViews rapporte la statistiques du test de DW directement dans le diagnostic de chaque sortie de régression. Les résultats de régression Stata ne contiennent pas automatiquement la statistique DW, mais cela peut être obtenu très facilement en utilisant la commande ci-après (la commande doit être saisie et exécutée immédiatement après avoir obtenu les résultats de régression si vous souhaitez tester pour l'autocorrélation) :

```
estat dwatson
```

Le résultat est signalé dans la fenêtre de résultats de Stata. Par conséquent, pour les deux logiciels, le seul travail qui reste à faire pour le chercheur est de construire le tableau avec les valeurs critiques et de vérifier s'il existe une corrélation sérielle et de quel type elle est. Un exemple est donné ci-dessous.

Exemple Empirique

4 Violations des Hypothèses Classiques

À partir des résultats de régression obtenus dans l'exemple précédent (déttection graphique de l'autocorrélation), nous observons que la statistique DW est égale à 0.18. Trouver les valeurs critiques d_L et d_U au niveau de signification de 1% pour $n = 49$ et $m = 1$ dans la table de DW dans les annexes et en les plaçant dans le tableau de DW, nous avons les résultats indiqués dans la figure (4.7). Étant donné que $d = 0.18$ est inférieur à $d_L = 1.245$, il existe des preuves solides d'une corrélation sérielle positive.

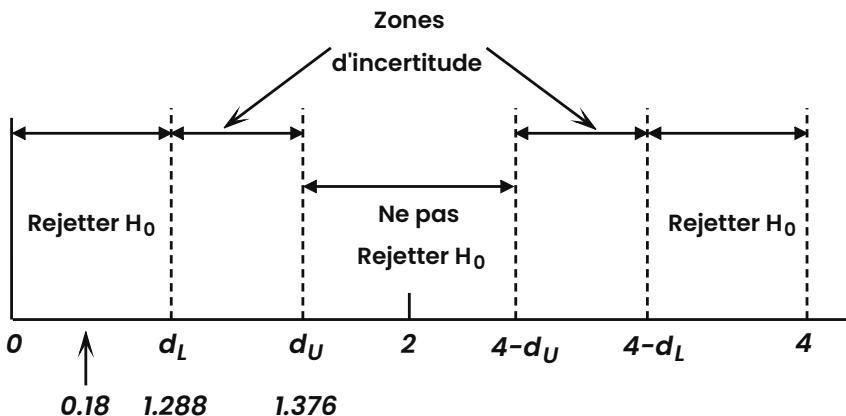


FIGURE 4.7 – **Test de Durbin Watson**

Le Test de Breusch-Godfrey LM

Le test de DW présente plusieurs inconvénients qui rendent son utilisation inappropriée dans divers cas. Par exemple (a) il peut donner des résultats non concluants ; (b) il n'est pas applicable lorsqu'une variable dépendante décalée est utilisée ; et (c) il ne peut pas prendre en compte des ordres plus élevés de corrélation sérielle.

Pour ces raisons, Breusch (1978) et Godfrey (1978) ont développé un test LM qui peut s'adapter à tous les cas ci-dessus. Considérez le modèle :

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \dots + \beta_m X_{tm} + \varepsilon_t \quad (4.34)$$

où

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \dots + \rho_p \varepsilon_{t-p} + u_t \quad (4.35)$$

Le test Breusch-Godfrey LM combine ces deux équations :

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \dots + \beta_m X_{tm} + \dots + \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \dots + \rho_p \varepsilon_{t-p} + u_t \quad (4.36)$$

et donc les hypothèses nulle et alternative sont :

$H_0 : \rho_1 = \rho_2 = \dots = \rho_p = 0$ pas d'autocorrélation

H_a : au moins l'un des ρ_s n'est pas nul, donc une corrélation sérielle

Les étapes de réalisation du test sont les suivantes :

Etape 1 Estimer le modèle (4.34) en utilisant MCO et obtenir les résidus e_t .

Etape 2 Exécutez le modèle de régression suivant avec le nombre de retards utilisés (ρ) étant déterminé selon l'ordre de corrélation sérielle à tester.

$$e_t = \alpha_0 + \alpha_1 X_{t1} + \dots + \alpha_R X_{tR} + \alpha_{R+1} e_{t-1} + \dots + \alpha_{R+p} e_{t-p} \quad (4.37)$$

Etape 3 Calculer la statistique $LM = (n - p) R^2$ à partir de la régression exécutée à l'étape 2. Si cette statistique LM est supérieure à la valeur critique χ_p^2 pour un niveau de signification donné, l'hypothèse nulle est rejetée et nous concluons qu'une corrélation sérielle est présente. Notez que le choix de p est arbitraire. Cependant, la périodicité des données (trimestrielles, mensuelles, hebdomadaires, etc.) suggérera souvent la taille de p .

Le Test de Breusch-Godfrey avec EViews et Stata

Après avoir estimé une équation de régression dans EViews, afin d'effectuer le test de Breusch-Godfrey LM, nous passons à partir de la fenêtre des résultats d'estimation à **View / Residual Tests / Serial Correlation LM**. EViews demande le nombre de retards à inclure dans le test, et après avoir spécifié cela, les résultats du test sont obtenus. L'interprétation est comme d'habitude.

Dans Stata, la commande utilisée pour obtenir les résultats du test de Breusch-Godfrey est :

```
estat bgodfrey , lags(number)
```

4 Violations des Hypothèses Classiques

où (nombre) doit être remplacé par le nombre de retards que nous voulons tester pour l'auto-corrélation. Par conséquent, si nous voulons tester le quatrième ordre d'autocorrélation, la commande est :

```
estat bgodfrey , lags(4)
```

De même pour les autres commandes, nous changeons simplement le nombre entre parenthèses.

Exemple Empirique

Poursuivant l'exemple de la relation entre la consommation et le revenu disponible, nous procédons en testant la corrélation sérielle du premier ordre parce que nous avons des données annuelles. Pour tester cette corrélation sérielle, nous utilisons le test Breusch-Godfrey LM. Dans la fenêtre des résultats de la régression estimée, accédez à **View / Residual Tests / Serial Correlation LM** et spécifiez 1 comme nombre de retards. Les résultats de ce test sont présentés dans le tableau (4.10). Nous pouvons voir dans les premières colonnes que les valeurs à la fois de

Breusch-Godfrey Serial Correlation LM Test				
F-statistic	168.9023	Prob. F(1,46)	0.0000	
Obs*R-squared	38.51151	Prob. Chi-Square(1)	0.0000	
Dependent Variable : RESID				
Method : Least Squares				
Sample : 1 49				
Included observations : 49				
Presample missing value lagged residuals set to zero.				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-54.41017	102.7650	-0.529462	0.5990
R	0.003590	0.005335	0.673044	0.5043
RESID(-1)	0.909272	0.069964	12.99624	0.0000
R-squared	0.785949	Mean dependent var	-5.34E-13	
Adjusted R-squared	0.776643	S.D. dependent var	433.0451	
S.E. of regression	204.6601	Akaike info criterion	13.53985	
Sum squared resid	1926746.	Schwarz criterion	13.65567	
Log likelihood	-328.7263	Hannan-Quinn criter.	13.58379	
F-statistic	84.45113	Durbin-Watson stat	2.116362	
Prob(F-statistic)	0.000000			

TABLE 4.10 – Test de Breusch-Godfrey

la statistique LM et de la statistique F sont assez élevées, ce qui suggère

4 Violations des Hypothèses Classiques

le rejet de l'hypothèse nulle d'absence de corrélation sérielle de premier ordre. Il est également clair que cela est dû au fait que les p -values sont très petites (inférieures à 0.05 pour un intervalle de confiance à 95%). Par conséquent, une corrélation sérielle est définitivement présente.

Test h de Durbin en Présence de Variables Dépendantes Décalées

Nous avons mentionné précédemment, dans les hypothèses de ce test, que ce test n'est pas applicable lorsque le modèle de régression inclut des variables dépendantes retardées comme aussi pour les variables explicatives. Par conséquent, si le modèle examiné a la forme :

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \dots + \beta_m X_{tm} + \gamma Y_{t-1} + \varepsilon_t \quad (4.38)$$

le test DW n'est pas valide.

Durbin (1970) a conçu une statistique de test qui peut être utilisée pour de tels modèles, et cette statistique h a la forme suivante :

$$h = \left(1 - \frac{d}{2}\right) \sqrt{\frac{n}{1 - n\sigma_\gamma^2}} \quad (4.39)$$

où n est le nombre d'observations, d est la statistique DW régulière définie dans l'équation (4.32) et σ_γ^2 est la variance estimée du coefficient de la variable dépendante décalée. Pour les grands échantillons, cette statistique suit une distribution normale. Les étapes impliquées dans le test- h sont les suivantes :

Etape 1 Estimer l'équation (4.38) par MCO pour obtenir les résidus et calculer la statistique DW donnée par l'équation (4.38). (Comme nous l'avons noté précédemment, en pratique, cette étape utilisant EViews n'implique que l'estimation de l'équation par MCO. EViews fournit la statistique DW dans ses diagnostics de régression).

Etape 2 Calculez la statistique h donnée dans l'équation (4.39).

Etape 3 Les hypothèses nulle et alternative sont :

$$H_0; \rho = 0 \text{ pas d'autocorrélation}$$

$$H_a; \rho \neq 0 \text{ présence d'autocorrélation}$$

Etape 4 Comparer la statistique h avec la valeur critique (pour les grands échantillons et pour $\alpha = 0.05$, $z = \pm 1.96$). Si la statistique

4 Violations des Hypothèses Classiques

h dépasse la valeur critique, alors H_0 est rejetée et nous concluons qu'il existe une corrélation sérielle (voir également la figure (4.8)).

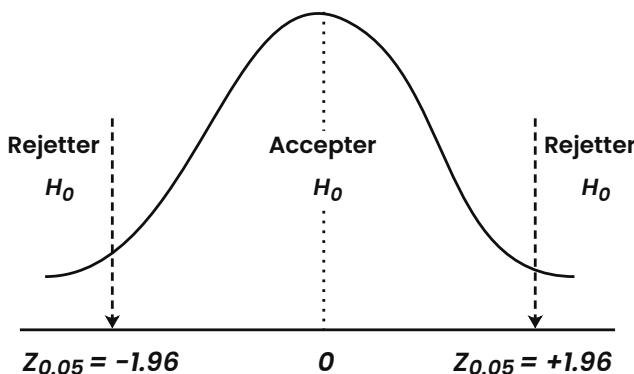


FIGURE 4.8 – **Test h de Durbin, affiché graphiquement**

Le Test h de Durbin avec Eviews et Stata

EViews signale uniquement le test DW, indépendamment du fait qu'une variable dépendante retardée soit utilisée ou non comme régresseur. Par conséquent, l'étape 2 est nécessaire pour calculer la statistique h . Un exemple de test h utilisant EViews est donné ci-dessous.

Dans Stata, après avoir estimé la régression avec la variable dépendante décalée, nous devons utiliser la commande de test DW :

estat dwatson

suivi du calcul de la statistique h comme décrit dans l'étape 2. Un exemple informatique utilisant EViews est donné ci-dessous. Il est très facile de produire les mêmes résultats avec Stata.

Exemple Empirique

Si nous voulons estimer le modèle de régression suivant :

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 Y_{t-1} + \varepsilon_t$$

qui comprend une variable dépendante décalée, nous savons que le test DW n'est plus valide. Ainsi, dans ce cas, nous devons utiliser le test h de Durbin ou le test LM . Exécution du modèle de régression se fait par la commande :

ls y c x y(-1)

4 Violations des Hypothèses Classiques

Dependent Variable : Y

Method : Least Squares

Sample (adjusted) : 1972 2019

Included observations : 48 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-254.5241	155.2906	-1.639019	0.1082
R	0.211505	0.068310	3.096256	0.0034
Y(-1)	0.800004	0.070537	11.34159	0.0000
R-squared	0.998367	Mean dependent var	16915.21	
Adjusted R-squared	0.998294	S.D. dependent var	5377.825	
S.E. of regression	222.1108	Akaike info criterion	13.70469	
Sum squared resid	2219995.	Schwarz criterion	13.82164	
Log likelihood	-325.9126	Hannan-Quinn criter.	13.74889	
F-statistic	13754.09	Durbin-Watson stat	0.969327	
Prob(F-statistic)	0.000000			

TABLE 4.11 – Résultats de la Régression avec une Variable Dépendante Décalée

nous obtenons les résultats indiqués dans le tableau (4.11).

La statistique DW est égale à 0.969327, et à partir de cela, nous pouvons obtenir la statistique h calculée à partir de la formule :

$$h = \left(1 - \frac{d}{2}\right) \sqrt{\frac{n}{1 - n\sigma_{\gamma}^2}}$$

où σ_{γ}^2 est la variance estimée du coefficient de la variable dépendante décalée $Y(-1) = (0.070537)^2 = 0.0049754$

En tapant la commande suivante dans Eviews, nous obtenons la valeur de la statistique h :

```
scalar h=(1-0.969327/2)*(48/(1-48*0.070537^2))^(.5)
```

et en double-cliquant sur le scalaire h , nous pouvons voir la valeur :

```
scalar h=4.0923101
```

et donc parce que $h > z - critique = 1.96$, nous rejetons l'hypothèse H_0 et nous concluons que ce modèle souffre d'une corrélation sérielle.

En appliquant le test LM pour cette équation de régression en cliquant sur **View / Residual Tests / Serial Correlation LM Test** et en spécifiant l'ordre de décalage égal à 1 (en tapant 1 dans la case appropriée),

4 Violations des Hypothèses Classiques

nous obtenons les résultats indiqués dans le tableau (4.12). D'après ces résultats, il est à nouveau clair qu'il y a corrélation sérielle dans ce modèle.

Breusch-Godfrey Serial Correlation LM Test				
F-statistic	17.62096	Prob. F(1,44)	0.0001	
Obs*R-squared	13.72595	Prob. Chi-Square(1)	0.0002	
Dependent Variable : RESID				
Method : Least Squares				
Sample : 1972 2019				
Included observations : 48				
Presample missing value lagged residuals set to zero.				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-100.5944	134.8513	-0.745965	0.4597
X	0.068067	0.060585	1.123501	0.2673
Y(-1)	-0.070717	0.062588	-1.129887	0.2646
RESID(-1)	0.555620	0.132362	4.197733	0.0001
R-squared	0.285957	Mean dependent var	7.95E-13	
Adjusted R-squared	0.237273	S.D. dependent var	217.3337	
S.E. of regression	189.8068	Akaike info criterion	13.40955	
Sum squared resid	1585171.	Schwarz criteron	13.56548	
Log likelihood	-317.8291	Hannan-Quinn criter.	13.46847	
F-statistic	5.873655	Durbin-Watson stat	1.741955	
Prob(F-statistic)	0.001832			

TABLE 4.12 – Résultats du Test de Breusch–Godfrey LM

4.2.4 Correction de l'Autocorrélation

Étant donné que la présence d'autocorrélation nous fournit des estimateurs MCO inefficaces, il est important d'avoir des moyens de corriger nos estimations. Deux cas différents sont présentés dans les deux sous-sections suivantes.

Lorsque ρ est connu

Considérer le modèle :

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \dots + \beta_m X_{tm} + \varepsilon_t \quad (4.40)$$

4 Violations des Hypothèses Classiques

où ε_t est autocorrélé et nous supposons qu'il suit une corrélation sérielle de premier ordre, comme cité dans l'équation (4.21).

Si l'équation (4.40) est valable pour la période t , elle le sera également pour la période $t - 1$, donc :

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1,1} + \beta_2 X_{t-1,2} + \dots + \beta_m X_{t-1,m} + \varepsilon_{t-1} \quad (4.41)$$

En multipliant les deux côtés de l'équation (4.41) par ρ , on obtient :

$$\rho Y_{t-1} = \beta_0 \rho + \beta_1 \rho X_{t-1,1} + \beta_2 \rho X_{t-1,2} + \dots + \beta_m \rho X_{t-1,m} + \rho \varepsilon_{t-1} \quad (4.42)$$

et en soustrayant l'équation (4.42) de l'équation (4.40), nous obtenons :

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1 (X_{t1} - \rho X_{t-1,1}) + \beta_2 (X_{t2} - \rho X_{t-1,2}) + \dots + \beta_m (X_{tm} - \rho X_{t-1,m}) + (\varepsilon_t - \rho \varepsilon_{t-1}) \quad (4.43)$$

ou

$$Y_t^* = \beta_0^* + \beta_1 X_{t1}^* + \beta_2 X_{t2}^* + \dots + \beta_m X_{tm}^* + u_t \quad (4.44)$$

où $Y_t^* = Y_t - \rho Y_{t-1}$, $\beta_0^* = \beta_0 (1 - \rho)$ et $X_{tk}^* = (X_{tk} - \rho X_{t-1,k})$.

Cette transformation, connue sous le nom de « transformation de Cochrane-Orcutt (1949) », réduit les perturbations aux erreurs classiques. Par conséquent, l'application de la méthode des MCO sur la régression résultante rend les estimations BLU, c'est-à-dire, exécuter une régression de $Y_t^* = (Y_t - \rho Y_{t-1})$ sur une constante et $X_{tk}^* = (X_{tk} - \rho X_{t-1,k})$ pour $t = 1, 2, \dots, T$. Notez que nous avons perdu une observation par décalage, et les estimateurs résultants sont BLUE uniquement pour les combinaisons linéaires des $(T - 1)$ observations en Y ².

Prais et Winsten (1954) tirent les estimateurs BLU pour les combinaisons linéaires des T observations en Y . Cela implique de recapturer l'observation initiale comme suit : **(i)** Y_1 et X_{1k} doivent être transformés pour la première observation, comme suit :

2. Il faut prendre en considération l'ordre des observations lorsqu'on applique la transformation Cochrane-Orcutt à des données en coupe. Les données en séries chronologiques ont un ordre naturel qu'on observe pas généralement dans les données transversales. Par conséquent, il faut être prudent lors de l'application de la transformation Cochrane-Orcutt aux données de section car elle n'est pas invariante à l'ordre des observations.

$$Y_1^* = Y_1 \sqrt{1 - \rho^2} \quad \text{et} \quad X_{1k}^* = X_{1k} \sqrt{1 - \rho^2} \quad (4.45)$$

(ii) ajouter cette observation initiale transformée aux observations transformées de Cochrane-Orcutt pour $t = 2, \dots, T$ et exécuter la régression sur les T observations plutôt que sur les $(T - 1)$ observations. Noter que

$$Y_1^* = \sqrt{1 - \rho^2} Y_1 \quad \text{et} \quad Y_t^* = Y_t - \rho Y_{t-1} \quad \text{pour } t = 2, \dots, T \quad (4.46)$$

De même, $X_{1k}^* = \sqrt{1 - \rho^2} X_1$ et $X_{tk}^* = X_{tk} - \rho X_{t-1,k}$ pour $t = 2, \dots, T$.

La transformation qui a généré Y_t^* , β_0^* et X_{tk}^* est connue sous le nom de « Quasi-Différenciation » ou « Différenciation Généralisée ». Noter que le terme d'erreur dans l'équation (4.44) satisfait toutes les hypothèses classiques de la régression linéaire. Donc, si ρ est connu, nous pouvons appliquer les MCO à l'équation (4.44) et obtenir des estimations BLUE. Un exemple d'utilisation de la différenciation généralisée est fourni ci-dessous.

Approche de Différenciation Généralisée avec Eviews

Exemple Empirique

Nous procédons avec les données de l'exemple de consommation précédent. Nous estimons dans EViews l'équation de régression suivante :

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

en tapant dans la ligne de commande EViews :

```
ls y c x
```

La régression MCO donne :

$$\begin{aligned} Y_t &= -1343.31 + 0.979 X_t + \text{résidus} \\ &\quad (0.062) \quad (0.011) \end{aligned}$$

Après avoir estimé la régression, nous stockons les résidus de la régression dans un vecteur en tapant la commande :

```
genr res01=resid
```

4 Violations des Hypothèses Classiques

En exécutant une régression de res01 sur res01(-1), nous obtenons les résultats indiqués dans le tableau (4.13), et à partir desquels le coefficient ρ est égal à 0.9059.

On procède par la commande ci-dessous dans Eviews :

```
ls res01 res01(-1)
```

Dependent Variable : RES01				
Method : Least Squares				
Sample (adjusted) : 1972 2019				
Included observations : 48 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
RES(-1)	0.905943	0.061943	14.62536	0.0000
R-squared	0.819679	Mean dependent var	-13.23939	
Adjusted R-squared	0.819679	S.D. dependent var	427.4886	
S.E. of regression	181.5299	Akaike info criterion	13.26133	
Sum squared resid	1548796.	Schwarz criterion	13.30031	
Log likelihood	-317.2719	Hannan-Quinn criter.	13.27606	
Durbin-Watson stat	2.408838			

TABLE 4.13 – Régression des Résidus sur Leurs Valeurs Décalées

Afin de transformer les variables pour la première observation, nous devons entrer les commandes suivantes dans la fenêtre de commande dans EViews :

```
scalar rho=c(1) → Enregistre l'estimation du coefficient  $\rho$ 
smpl 1:1 1:1 → Définit l'échantillon comme étant uniquement
                  la première observation
genr y_star=((1-rho^2)^(0.5))*y
genr x_star=((1-rho^2)^(0.5))*x
genr beta0_star=((1-rho^2)^(0.5))
```

où les trois commandes génèrent les variables suivies et la dernière commande crée la nouvelle constante.

Pour transformer les variables des observations 2 à 49, nous devons taper les commandes suivantes dans la fenêtre de commande EViews :

```
smpl 2:1 49:1
genr y_star=y-rho*y(-1)
genr x_star=x-rho*x(-1)
genr beta0_star=1-rho
```

4 Violations des Hypothèses Classiques

Et pour estimer l'équation (4.44), nous devons d'abord modifier l'échantillon pour toutes les observations puis exécuter la régression, ci-dessous les commandes à executer dans Eviews :

```
smp1 1:1 49:1
ls y_star beta0_star x_star
```

Les résultats de cette régression figurent dans le tableau (4.14)

Dependent Variable : Y_STAR				
Method : Least Squares				
Sample (adjusted) : 1971 2019				
Included observations : 48 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
BETA0_STAR	-743.4074	600.8505	-1.237259	0.2221
X_STAR	0.956751	0.029600	32.32255	0.0000
R-squared	0.928428	Mean dependent var	1977.190	
Adjusted R-squared	0.926906	S.D. dependent var	678.6280	
S.E. of regression	183.4736	Akaike info criterion	13.30198	
Sum squared resid	1582140	Schwarz criterion	13.37920	
Log likelihood	-323.8985	Hannan-Quinn criter.	13.33127	
Durbin-Watson stat	2.263429			

TABLE 4.14 – Les résultats de la Régression de Différenciation Généralisée

Lorsque ρ n'est pas connu

Il est évident que les estimateurs BLU résultants impliqueront ρ et sont donc différents des estimateurs habituels des MCO, sauf dans le cas où $\rho = 0$. Par conséquent, l'estimation des MCO n'est plus BLUE. De plus, nous devons connaître ρ pour obtenir les estimateurs BLU. Dans le travail appliqué, ρ n'est pas connu et doit être estimé³. Plusieurs procédures ont été élaborées, dont deux sont les plus populaires et les plus importantes : (a) Procédure itérative Cochrane-Orcutt ; et (b) Procédure de recherche Hildreth-Lu.

La Méthode Cochrane-Orcutt (1949)

3. Tant que $\hat{\rho}$ est une estimation cohérente de ρ , il s'agit d'une condition suffisante pour que les estimations correspondantes des β 's à l'étape suivante soient asymptotiquement efficaces.

4 Violations des Hypothèses Classiques

Cochrane et Orcutt (1949) ont développé une procédure itérative qui peut être présentée à travers les étapes suivantes :

Etape 1 Exécuter une régression de l'équation (4.40) et obtenir les résidus e_i de cette équation de régression.

Etape 2 Estimer le coefficient de corrélation sérielle de premier ordre ρ par MCO à partir de $e_t = \rho e_{t-1} + u_t$.

Etape 3 Transformer les variables d'origine en $Y_t^* = Y_t - \hat{\rho}Y_{t-1}$, $\beta_0^* = \beta_0(1 - \hat{\rho})$ et $X_{tk}^* = (X_{tk} - \hat{\rho}X_{t-1,k})$ pour $t = 2, \dots, T$ et $Y_1^* = Y_1\sqrt{1 - \hat{\rho}}$ et $X_{t1}^* = X_{t1}\sqrt{1 - \hat{\rho}}$ pour $t = 1$.

Etape 4 Exécuter la régression en utilisant les variables transformées et trouver les résidus de cette régression. Comme nous ne savons pas que le $\hat{\rho}$ obtenu à l'étape 2 est la «meilleure» estimation de ρ , on revient pour répéter les étapes 2 à 4 pour plusieurs fois jusqu'à ce que le critère d'arrêt ci-dessous soit respecté.

Critère d'arrêt : La procédure itérative peut être arrêtée lorsque les estimations de ρ à partir de deux itérations successives ne diffèrent pas plus que de certaines valeurs présélectionnées (très petites), telles que 0.001. Le $\hat{\rho}$ final est utilisé pour obtenir les estimations de l'équation (4.44). En général, la procédure itérative converge rapidement et ne nécessite pas plus de 3 à 6 itérations.

EViews utilise une méthode non linéaire itérative pour estimer les résultats de différenciation généralisés avec des erreurs AR(1) (Erreurs Autorégressives d'ordre 1) en présence d'une corrélation sérielle. Étant donné que la procédure est itérative, elle nécessite un certain nombre de répétitions pour atteindre la convergence, qui est indiquée dans les résultats d'EViews. Les estimations de cette méthode itérative peuvent être obtenues en ajoutant simplement les termes d'erreur AR(1) à la fin de la spécification de l'équation. Donc, si nous avons un modèle avec les variables Y et X , la commande de régression linéaire simple est :

```
ls y c x
```

Si nous savons que les estimations souffrent d'une corrélation sérielle d'ordre 1, les résultats peuvent être obtenus grâce au processus itératif en utilisant la commande :

```
ls y c x ar(1)
```

EViews fournit les résultats de la manière habituelle concernant la constante et le coefficient de la variable X , ainsi qu'une estimation pour ρ , qui sera le coefficient du terme AR(1). Le tableau (4.15) montre les

4 Violations des Hypothèses Classiques

Dependent Variable : Y

Method : ARMA Maximum Likelihood (BFGS)

Sample : 1971 2019

Included observations : 49

Convergence achieved after 7 iterations

Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-342.8355	2287.443	-0.149877	0.8815
X	0.940135	0.071876	13.07997	0.0000
AR(1)	0.946243	0.056591	16.72069	0.0000
SIGMASQ	31655.28	6484.395	4.881763	0.0000
R-squared	0.998911	Mean dependent var	16749.10	
Adjusted R-squared	0.998838	S.D. dependent var	5447.060	
S.E. of regression	185.6585	Akaike info criterion	13.40987	
Sum squared resid	1551109.	Schwarz criterion	13.56431	
Log likelihood	-324.5419	Hannan-Quinn criter.	13.46846	
F-statistic	13757.56	Durbin-Watson stat	2.327582	
Prob(F-statistic)	0.000000			
Inverted AR Roots	.95			

TABLE 4.15 – Les résultats de la de Procédure itérative

résultats de l'application de la méthode sur les données de l'exemple numérique précédent. Il a fallu 7 itérations pour obtenir des résultats convergents. De plus, le coefficient AR(1) (qui est en fait le ρ) est égal à 0.9462. Le cas ici pourrait par exemple être affecté par la fréquence trimestrielle des données. Si nous ajoutons un terme AR(4) en utilisant la commande :

```
ls y c x ar(1) ar(4)
```

Procédure Itérative de Cochrane-Orcutt avec Stata

En considérant toujours les données de l'exemple précédent, la procédure itérative de Cochrane-Orcutt (1949) executée avec Stata moyennant la commande ci-dessous, donne les résultats donnés dans le tableau (4.16) :

```
prais y x, corc
```

4 Violations des Hypothèses Classiques

Cochrane-Orcutt AR(1) regression - iterated estimates

Source	SS	df	MS	Number of obs	= 48
				F(1, 46)	= 689.89
Model	23168597	1	23168597	Prob > F	= 0.0000
Residual	1544819.27	46	33583.0277	R-squared	= 0.9375
Total	24713416.3	47	525817.367	Adj R-squared	= 0.9361
				Root MSE	= 183.26
y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x	0.996136	0.0379253	26.27	0.000	0.91979 1.0724
_cons	-1723.689	859.2143	-2.01	0.051	-3453.19 5.8191
rho	0.8879325				
Durbin-Watson statistic (original)				0.180503	
Durbin-Watson statistic (transformed)				2.447750	

TABLE 4.16 – **La procédure Itérative de Cochrane-Orcutt avec Stata**

La Procédure de recherche de Hildreth-Lu (1960)

Hildreth et Lu (1960) ont développé une méthode alternative à la procédure itérative Cochrane-Orcutt. Leur méthode comprend les étapes suivantes :

Etape 1 Choisir une valeur pour ρ (disons ρ_1), et pour cette valeur, transformer le modèle comme dans l'équation (4.44) et estimez-le par les MCO.

Etape 2 À partir de l'estimation de l'étape 1, obtenir les résidus e_t et la somme des carrés des résidus $SCR(\rho_1)$. Choisir ensuite une valeur différente de ρ (soit ρ_2) et répétez les étapes 1 et 2.

Etape 3 En faisant varier ρ de -1 à $+1$ d'une manière systématique prédéterminée (disons avec des pas de longueur 0.05), nous pouvons obtenir une série de valeurs pour $SCR(\rho_i)$. Nous choisissons le ρ pour lequel SCR est minimale et aussi l'équation (4.44), qui a été estimée en utilisant le ρ choisi comme solution optimale.

Cette procédure est très complexe et implique de nombreux calculs. EViews fournit des résultats très rapidement avec la méthode itérative Cochrane-Orcutt (comme nous l'avons montré ci-dessus), et est généralement préféré en cas d'autocorrélation.

4.3 Hétéroscédasticité : Que se passe-t-il si la variance d'erreur n'est pas constante ?

Une hypothèse importante du modèle de régression linéaire classique (hypothèse 2) est que les erreurs ε_i apparaissant dans la fonction de régression sont homoscédastiques ; c'est-à-dire qu'ils ont tous la même variance. Dans cette section, nous examinons la validité de cette hypothèse et découvrons ce qui se passe si cette dernière n'est pas remplie.

4.3.1 La Nature de l'Hétéroscédasticité

Comme indiqué précédemment, l'une des hypothèses importantes du modèle de régression linéaire classique est que la variance de chaque terme d'erreur ε_i est constante et égale à $\text{var}(\varepsilon_i) = \sigma^2$. Il s'agit de l'hypothèse d'**homoscédasticité**, (égal (homo), écart (scédasticité)), c'est-à-dire de variance égale. Schématiquement, dans le modèle de régression à une variable explicative, l'homoscédasticité peut être montrée comme dans la figure (2.8). Comme le montre la figure, la variance conditionnelle de Y_i (qui est égale à celle de ε_i), reste la même quelles que soient les valeurs prises par la variable X .

En revanche, lorsque la variance conditionnelle de Y_i augmente à mesure que X augmente. Ici, les variances de Y_i ne sont pas les mêmes. Il existe donc une hétéroscédasticité. Symboliquement,

$$\text{var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma_i^2$$

Remarquez l'indice de σ^2 , qui nous rappelle que les variances conditionnelles de ε_i (= variances conditionnelles de Y_i) ne sont plus constantes.

Il existe plusieurs raisons pour lesquelles les variances de ε_i peuvent être variables, on cite :

- (1) En suivant les modèles d'apprentissage des erreurs (*the error-learning models*), au fur et à mesure que les gens apprennent, leurs erreurs de comportement deviennent plus petites au fil du temps ou le nombre d'erreurs devient plus cohérent. Dans ce cas, σ_i^2 devrait diminuer. À titre d'exemple, considérons le nombre d'erreurs de frappe commises au cours d'une période donnée lors d'un test de dactylographie. A mesure que le nombre d'heures de pratique de frappe augmente, le nombre moyen d'erreurs de frappe ainsi que leurs écarts diminuent.
- (2) À mesure que les revenus augmentent, les gens ont plus de choix quant à la disposition de leurs revenus. Par conséquent, σ_i^2 est susceptible d'augmenter avec le revenu. Ainsi, dans la régression de

4 Violations des Hypothèses Classiques

l'épargne sur le revenu, il est probable que σ_i^2 augmente avec le revenu parce que les gens ont plus de choix quant à leur comportement d'épargne. De même, les sociétés dont les bénéfices sont plus importants devraient généralement afficher une plus grande variabilité dans leurs politiques de dividendes que les sociétés dont les bénéfices sont inférieurs.

- (3) À mesure que les techniques de collecte de données s'améliorent, σ_i^2 est susceptible de diminuer.
- (4) Une hétéroscédasticité peut également résulter de la présence de valeurs **aberrantes**. Une observation aberrante est une observation qui est très différente (soit très petite ou très grande) par rapport aux observations de l'échantillon. L'inclusion ou l'exclusion d'une telle observation, surtout si la taille de l'échantillon est petite, peut modifier considérablement les résultats de l'analyse de régression.
- (5) Une autre source d'hétéroscédasticité provient lorsque le modèle de régression n'est pas correctement spécifié par exemple lorsque certaines variables importantes sont omises du modèle. Ainsi, dans la fonction de demande d'un produit, si nous n'incluons pas les prix des produits complémentaires ou en concurrence avec le produit en question, les résidus provenant de la régression peuvent donner l'impression distincte que la variance de l'erreur peut ne pas être constante. Mais si les variables omises sont incluses dans le modèle, cette impression peut disparaître.
- (6) Une autre source d'hétéroscédasticité est l'**asymétrie** dans la distribution d'un ou plusieurs régresseurs inclus dans le modèle. Des exemples sont des variables économiques telles que le revenu, la richesse et l'éducation. Il est bien connu que la répartition des revenus et la richesse dans la plupart des sociétés est inégale, ainsi la part essentielle des revenus appartient au plus haut niveau de la hiérarchie d'une société.
- (7) Autres sources d'hétéroscédasticité : Comme le note David Hendry, l'hétéroscédasticité peut également survenir en raison de (1) une transformation incorrecte des données et (2) une forme fonctionnelle incorrecte (par exemple, des modèles linéaires ou log-linéaires).

Il est à noter que le problème de l'hétéroscédasticité est probablement plus fréquent dans les données transversales que dans les données en séries chronologiques. Dans les données transversales, on traite généralement des membres d'une population à un moment donné, tels que les

4 Violations des Hypothèses Classiques

consommateurs individuels ou leurs familles, les entreprises, les industries ou les subdivisions géographiques telles que l'État, le pays, la ville, etc. En outre, ces membres peuvent être de tailles différentes, comme des petites, moyennes ou grandes entreprises ou des revenus faibles, moyens ou élevés. Dans les données en séries chronologiques, en revanche, les variables ont tendance à être de l'ordre de grandeur similaire, car on recueille généralement les données pour la même entité sur une période de temps. Les exemples sont le produit national brut (PNB), les dépenses de consommation, l'épargne ou l'emploi pour une telle période.

4.3.2 Estimation des MCO en Présence d'Hétéroscédasticité

Qu'advient-il des estimateurs des MCO et leurs variances si nous introduisons l'hétéroscédasticité en laissant $E(\varepsilon_i^2) = \sigma_i^2$ mais en maintenant toutes les autres hypothèses du modèle classique ? Pour répondre à cette question, revenons au modèle à deux variables :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

En appliquant la formule habituelle, l'estimateur des MCO de β_1 est :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2} \quad (4.47)$$

mais sa variance est maintenant donnée par l'expression suivante :

$$var(\hat{\beta}_1) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \sigma_i^2}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \right]^2} = \frac{\sigma_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (4.48)$$

ce qui est évidemment différente de la formule de variance habituelle obtenue sous l'hypothèse d'homoscédasticité, à savoir,

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (4.49)$$

Bien sûr, si $\sigma_i^2 = \sigma^2$ pour chaque i , les deux formules seront identiques. (Pourquoi ?)

Rappelons que $\hat{\beta}_1$ est le meilleur estimateur linéaire sans biais (BLUE)

si les hypothèses du modèle classique, y compris l'homoscédasticité, se vérifient. Est-il toujours BLUE lorsque nous laissons tomber l'hypothèse d'homoscédasticité et la remplaçons par l'hypothèse d'hétéroscédasticité ? Il est facile de prouver que $\hat{\beta}_1$ est toujours linéaire et non biaisé. En fait, comme le montre la sous-section 2.3.3, pour établir le caractère non biaisé de $\hat{\beta}_1$, il n'est pas nécessaire que les erreurs (ε_i) soient homoscélastiques. En fait, la variance de ε_i qu'elle soit homoscélastique ou hétéroscédastique, ne joue aucun rôle dans la détermination de la propriété de non biais. Rappelons qu'à la sous-section 2.3.3, nous avons montré que $\hat{\beta}_1$ est un estimateur cohérent selon les hypothèses du modèle de régression linéaire classique. Bien que nous ne le prouverons pas, on peut montrer que $\hat{\beta}_1$ est un estimateur cohérent malgré l'hétéroscédasticité ; c'est-à-dire que lorsque la taille de l'échantillon augmente indéfiniment, le $\hat{\beta}_1$ estimé converge vers sa vraie valeur. De plus, on peut également montrer que dans certaines conditions (appelées conditions de régularité), $\hat{\beta}_1$ est *asymptotiquement distribué normalement*. Bien sûr, ce que nous avons dit à propos de $\hat{\beta}_1$ vaut également pour d'autres paramètres d'un modèle de régression multiple.

Étant donné que $\hat{\beta}_1$ est toujours linéaire et non biaisé, est-il «efficace» ou «meilleur» ? Autrement dit, a-t-il une variance minimale dans la classe des estimateurs sans biais ? Et est-ce que la variance donnée dans (4.48) est minimale ? La réponse est non aux deux questions : $\hat{\beta}_1$ n'est plus le meilleur et la variance minimale n'est pas donnée dans (4.48). Alors qu'est-ce que le BLUE en présence d'hétéroscédasticité ? La réponse est donnée dans la section suivante.

4.3.3 La Méthode des Moindres Carrés Généralisés (MCG)

Pourquoi l'estimateur des MCO habituel de β_1 donné dans (4.47) n'est-il pas le meilleur, bien qu'il soit toujours sans biais ?

Pour répondre à cette question et illustrer un exemple d'hétéroscédasticité susceptible d'être rencontrée dans l'analyse transversale, considérons l'exemple ci-après.

Exemple

Le tableau (4.17) donne des données sur la rémunération par employé dans 10 industries manufacturières de biens non durables, classées selon la taille de l'emploi de l'entreprise ou de l'établissement pour une année. Le tableau présente également des chiffres de la productivité moyenne pour neuf catégories d'emplois.

Comme le montre le tableau (4.17), il existe une variabilité considérable des rémunérations entre les catégories d'emploi. Si nous devions

4 Violations des Hypothèses Classiques

Industries	Taille de l'Emploi (Nombre Moyen d'Employés)								
	1-4	5-9	10-19	20-49	50-99	100-249	250-499	500-999	1000-2499
Indutrie 1	2994	3295	3565	3907	4189	4486	4676	4968	5342
Indutrie 2	1721	2057	3336	3320	2980	2848	3072	2969	3822
Indutrie 3	3600	3657	3674	3437	3340	3334	3225	3163	3168
Indutrie 4	3494	3787	3533	3215	3030	2834	2750	2967	3453
Indutrie 5	3498	3847	3913	4135	4445	4885	5132	5342	5326
Indutrie 6	3611	4206	4695	5083	5301	5269	5182	5395	5552
Indutrie 7	3875	4660	4930	5005	5114	5248	5630	5870	5876
Indutrie 8	4616	5181	5317	5337	5421	5710	6316	6455	6347
Indutrie 9	3538	3984	4014	4287	4221	4539	4721	4905	5481
Indutrie 10	3016	3196	3149	3317	3414	3254	3177	3346	4067
Rémunération	3396	3787	4013	4104	4146	4241	4388	4538	4843
Moyenne									
Ecart Type	742.2	851.4	727.8	805.06	929.9	1080.6	1241.2	1307.7	1110.7
Productivité	9355	8584	7962	8275	8389	9418	9795	10281	11750
Moyenne									

TABLE 4.17 – Rémunération par Employé dans les Industries Manufacturières Selon la Taille de l'emploi de l'Entreprise

régresser la rémunération par employé en fonction de la taille de l'emploi, nous aimeraisons tirer parti du fait qu'il existe une variabilité interclasse considérable des gains. Idéalement, nous aimeraisons concevoir le schéma d'estimation de telle manière que les observations provenant des populations à plus grande variabilité reçoivent moins de poids que celles provenant de populations à plus faible variabilité. En examinant le tableau (4.17), nous voudrions pondérer les observations provenant des classes d'emploi 10 – 19 et 20 – 49 plus fortement que celles provenant des classes d'emploi comme 5 – 9 et 250 – 499, car les premières sont plus étroitement regroupées autour de leurs valeurs moyennes que ce dernier, nous permettant ainsi d'estimer plus précisément la fonction de régression de la population (FRP).

Malheureusement, la méthode des MCO habituelle ne suit pas cette stratégie et n'utilise donc pas les «informations» contenues dans la variabilité inégale de la variable dépendante Y (c-à-d la rémunération des employés), cependant elle attribue un poids ou une importance égale

4 Violations des Hypothèses Classiques

à chaque observation. Mais une méthode d'estimation, connue sous le nom de « moindres carrés généralisés (MCG) », prend explicitement en compte ces informations et elle est donc capable de produire des estimateurs BLUE. Cette méthode sera expliquée plus loin dans la section concernée par la correction de l'hétéroscédasticité.

Moindres Carrés Généralisés (ou pondérés)

Considérer le modèle à deux variables suivant :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (4.50)$$

pour faciliter la manipulation algébrique de l'équation, nous écrivons

$$Y_i = \beta_0 X_{i0} + \beta_1 X_i + \varepsilon_i \quad (4.51)$$

où $X_{i0} = 1$ pour chaque i . Le lecteur peut constater que ces deux formulations sont identiques.

Supposons maintenant que les variances hétéroscédastiques σ_i^2 sont connues. Divisez l'équation (4.51) par σ_i pour obtenir

$$\frac{Y_i}{\sigma_i} = \beta_0 \left(\frac{X_{i0}}{\sigma_i} \right) + \beta_1 \left(\frac{X_i}{\sigma_i} \right) + \left(\frac{\varepsilon_i}{\sigma_i} \right) \quad (4.52)$$

qui, pour faciliter l'écriture, nous écrivons

$$Y_i^* = \beta_0^* X_{i0}^* + \beta_1^* X_i^* + \varepsilon_i^* \quad (4.53)$$

où les variables avec le signe (*) sont les variables originales divisées par (les connues) σ_i . Nous utilisons la notation β_0^* et β_1^* , pour les paramètres du modèle transformé, et ce pour les distinguer des paramètres des MCO habituels β_1 et β_2 .

Quel est le but de transformer le modèle original? Pour voir cela, notez la propriété suivante du terme d'erreur transformé ε_i^* :

$$\begin{aligned} \text{var}(\varepsilon_i^*) &= E(\varepsilon_i^*)^2 = E\left(\frac{\varepsilon_i}{\sigma_i}\right)^2 \text{ puisque } E(\varepsilon_i^*) = 0 \\ &= \frac{1}{\sigma_i^2} E(\varepsilon_i^2) \text{ puisque } \sigma_i^2 \text{ est connue} \\ &= \frac{1}{\sigma_i^2} (\sigma_i^2) = 1 \text{ puisque } E(\varepsilon_i^2) = \sigma_i^2 \end{aligned} \quad (4.54)$$

Autrement dit, la variance du terme d'erreur transformée ε_i^* est main-

4 Violations des Hypothèses Classiques

tenant homoscédastique. Puisque nous retenons toujours les autres hypothèses du modèle classique, la conclusion selon laquelle ε^* est homoscédastique suggère que si nous appliquons les MCO au modèle transformé (4.52), il produira des estimateurs qui sont BLUES. Enfin, les β_0^* et β_1^* estimés sont maintenant BLUES et ne sont pas les estimateurs des MCO $\hat{\beta}_0$ et $\hat{\beta}_1$.

L'application de cette procédure de transformation des variables d'origine de telle sorte que les variables transformées satisfont aux hypothèses du modèle classique, suivie par l'application des MCO est connue comme la méthode des moindres carrés généralisés (MCG). En bref, la méthode des MCG est celle des MCO appliquée sur les variables transformées qui satisfont aux hypothèses des moindres carrés ordinaires. Les estimateurs ainsi obtenus sont appelés estimateurs des MCG, et ce sont des estimateurs qui sont BLUES.

Les mécanismes réels d'estimation de $\hat{\beta}_0^*$ et $\hat{\beta}_1^*$ sont les suivants. Tout d'abord, nous notons l'exemple de la fonction de régression (FRP) de l'équation (4.53)

$$\frac{Y_i}{\sigma_i} = \hat{\beta}_0^* \left(\frac{X_{i0}}{\sigma_i} \right) + \hat{\beta}_1^* \left(\frac{X_i}{\sigma_i} \right) + \left(\frac{\varepsilon_i}{\sigma_i} \right)$$

où

$$Y_i^* = \hat{\beta}_0^* X_{i0}^* + \hat{\beta}_1^* X_i^* + \varepsilon_i^* \quad (4.55)$$

Maintenant, pour obtenir les estimateurs MCG, nous minimisons

$$\sum_{i=1}^n e_i^{2*} = \sum_{i=1}^n \left(Y_i^* - \hat{\beta}_0^* X_{i0}^* - \hat{\beta}_1^* X_i^* \right)^2$$

c'est-à-dire

$$\sum_{i=1}^n \left(\frac{e_i}{\sigma_i} \right)^2 = \sum_{i=1}^n \left[\left(\frac{Y_i}{\sigma_i} \right) - \hat{\beta}_0^* \left(\frac{X_{i0}}{\sigma_i} \right) - \hat{\beta}_1^* \left(\frac{X_i}{\sigma_i} \right) \right] \quad (4.56)$$

La minimisation de cette fonction passe par le calcul des dérivés partielles par rapport à β_0^* et β_1^* , nous obtenons ainsi

$$\hat{\beta}_1^* = \frac{\left(\sum_{i=1}^n w_i \right) \left(\sum_{i=1}^n w_i X_i Y_i \right) - \left(\sum_{i=1}^n w_i X_i \right) \left(\sum_{i=1}^n w_i Y_i \right)}{\left(\sum_{i=1}^n w_i \right) \left(\sum_{i=1}^n w_i X_i^2 \right) - \left(\sum_{i=1}^n w_i X_i \right)^2} \quad (4.57)$$

4 Violations des Hypothèses Classiques

et sa variance est donnée par

$$var(\hat{\beta}_1^*) = \frac{\sum_{i=1}^n w_i}{\left(\sum_{i=1}^n w_i\right) \left(\sum_{i=1}^n w_i X_i^2\right) - \left(\sum_{i=1}^n w_i X_i\right)^2} \quad (4.58)$$

où $w_i = \frac{1}{\sigma_i^2}$.

Différence entre MCO et MCG

Rappelons que dans les MCO nous minimisons

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \quad (4.59)$$

mais dans les MCG, nous minimisons l'expression (4.56), qui peut être également écrite comme

$$\sum_{i=1}^n w_i e_i^{2*} = \sum_{i=1}^n w_i (Y_i - \hat{\beta}_0^* X_{i0} - \hat{\beta}_1^* X_i)^2 \quad (4.60)$$

où $w_i = \frac{1}{\sigma_i^2}$.

Ainsi, dans les MCG, nous minimisons une somme pondérée des carrés des résidus avec w_i agissant comme poids, mais dans les MCO nous minimisons une somme non pondérée des carrés des résidus ou (ce qui revient au même) également pondérée (*SCR*). Comme le montre l'équation (4.56), dans les MCG, le poids attribué à chaque observation est inversement proportionnel à son σ_i , c'est-à-dire que les observations provenant d'une population avec un plus grand σ_i auront un poids relativement plus petit et celles d'une population avec un plus petit σ_i auront un poids proportionnellement plus important dans la minimisation de *SCR* en (4.60).

4.3.4 Conséquences de l'Hétéroscédasticité pour les Estimateurs des MCO

Approche Générale

Considérons le modèle de régression linéaire classique :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \varepsilon_i \quad k = 1, \dots, m \quad (4.61)$$

Si le terme d'erreur ε_i dans cette équation est hétéroscédastique, les conséquences pour les estimateurs $\hat{\beta}$ des MCO peuvent être résumées comme suit :

- (1) Les estimateurs $\hat{\beta}$ des MCO sont toujours sans biais et cohérents. En effet, aucune des variables explicatives n'est corrélée avec le terme d'erreur. Ainsi, une équation correctement spécifiée qui ne souffre que de la présence d'hétéroscédasticité nous donnera des valeurs de $\hat{\beta}$ qui sont relativement bonnes.
- (2) L'hétéroscédasticité affecte la distribution des $\hat{\beta}$ en augmentant les variances des distributions et en rendant ainsi les estimateurs des MCO inefficaces (car ils violent la propriété de convergence ou variance minimale). Pour comprendre cela, considérons la figure (4.9), qui montre la distribution d'un estimateur $\hat{\beta}$ avec et sans hétéroscédasticité. Il est évident que l'hétéroscédasticité ne provoque pas de biais car $\hat{\beta}$ est centré autour de β (donc $E(\hat{\beta}) = \beta$), mais l'élargissement de la distribution ne la rend plus efficace.
- (3) L'hétéroscédasticité affecte également les variances (et donc les erreurs types (ou erreurs standards) également) des $\hat{\beta}$ estimés. En fait, la présence d'une hétéroscédasticité amène la méthode des MCO à sous-estimer les variances (et les erreurs standard), et conduit à des valeurs plus élevées que prévu des statistiques t de Student et F de Fisher. Par conséquent, l'hétéroscédasticité a un grand impact sur le test d'hypothèse : les statistiques de Student et de Fisher ne sont plus fiables pour le test d'hypothèse car elles nous conduisent à rejeter trop souvent l'hypothèse nulle.

Effet sur la Matrice Variance-Covariance des Termes d'Erreur

Il est utile de voir comment l'hétéroscédasticité affectera la forme de la matrice de variance-covariance des termes d'erreur du modèle classique de régression multiple linéaire. En raison des hypothèses 2 et 3, la matrice variance-covariance des erreurs ressemble à :

$$E(\varepsilon\varepsilon') = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 I \quad (4.62)$$

4 Violations des Hypothèses Classiques

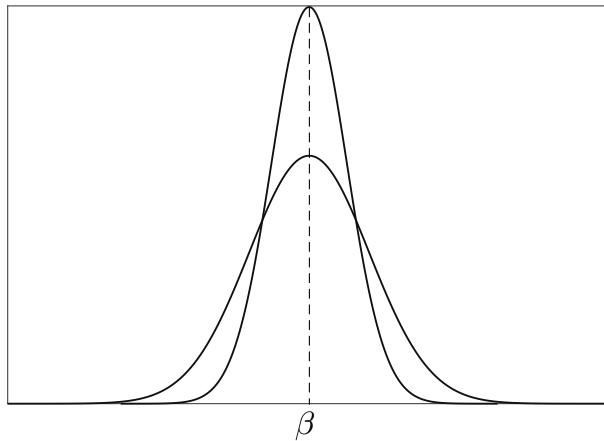


FIGURE 4.9 – L’Effet de l’Hétéroscléasticité sur un Paramètre Estimé

L’hypothèse 2 n’est plus valable en présence d’hétéroscléasticité, donc la matrice variance-covariance des résidus sera la suivante :

$$E(\varepsilon \varepsilon') = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} = \Omega \quad (4.63)$$

Effet sur les Estimateurs des MCO du Modèle de Régression Multiple

La matrice variance-covariance des estimateurs $\hat{\beta}$ des MCO est donnée par :

$$\begin{aligned} cov(\hat{\beta}) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E\left\{ \left[(X'X)^{-1} X' \varepsilon \right] \left[(X'X)^{-1} X' \varepsilon \right]' \right\} \\ &= E\left[(X'X)^{-1} X' \varepsilon \varepsilon' X (X'X)^{-1} \right] \\ &= (X'X)^{-1} X' E(\varepsilon \varepsilon') X (X'X)^{-1} \\ &= (X'X)^{-1} X' \Omega X (X'X)^{-1} \end{aligned} \quad (4.64)$$

qui est totalement différente de l'expression classique $\sigma^2 (X' X)^{-1}$, et Ω désigne la nouvelle matrice variance-covariance. Les formules de l'équation (4.64) forment la base de ce que l'on appelle souvent une inférence « *robuste* », à savoir le calcul des erreurs standards et les statistiques de Student qui sont correctes même lorsque certaines des hypothèses des MCO sont violées ; une forme particulière est supposée pour la matrice Ω et l'équation (4.64) est utilisée pour calculer une matrice de covariance corrigée.

4.3.5 Détection de l'Hétéroscédasticité

La Méthode Informelle

De manière informelle, et dans le cas de la régression linéaire simple, l'hétéroscédasticité peut facilement être détectée par une simple inspection du diagramme de dispersion. Cependant, cela ne peut pas être fait dans le cas de régression multiple. Dans ce cas, des informations utiles sur la présence possible d'hétéroscédasticité peuvent être fournies en traçant les carrés des résidus par rapport à la variable dépendante et /ou variables explicatives.

Il existe des cas dans lesquels des informations utiles sur l'hétéroscédasticité peuvent être déduites de ce type de modèle de graphique. Les schémas possibles sont présentés dans les figures (4.10) à (4.12). Dans la figure (4.10), il n'existe pas de schéma systématique entre les deux variables, ce qui suggère qu'il s'agit d'un modèle qui ne souffre pas d'hétéroscédasticité. Dans la figure (4.11a), il existe un schéma clair qui suggère une hétéroscédasticité, dans la figure (4.11b), il existe une relation linéaire claire entre Y_i (ou X_i) et e_i^2 , tandis que les figures (4.12a) et (4.12b) présentent une relation quadratique. La connaissance de la relation entre les deux variables peut être très utile car elle permet de transformer les données de manière à éliminer l'hétéroscédasticité.

Les Méthodes Formelles

Le Test de Breusch-Pagan LM

Breusch et Pagan (1979) mis au point un test du multiplicateur de Lagrange (LM) pour détecter hétéroscédasticité. Dans le modèle suivant :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \varepsilon_i \quad k = 1, \dots, m \quad (4.65)$$

4 Violations des Hypothèses Classiques

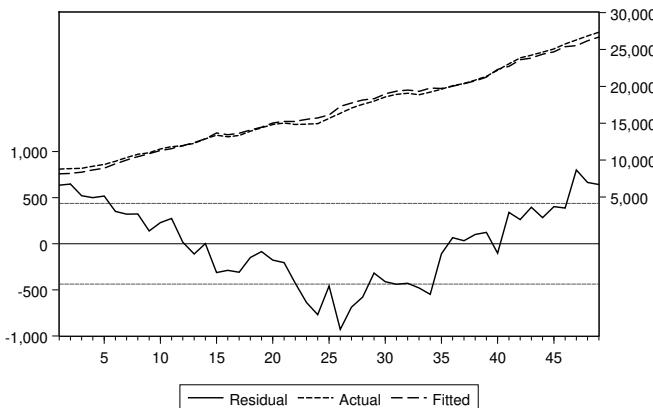


FIGURE 4.10 – Une Distribution « Saine » des Carrés des Résidus

$\text{var}(\varepsilon_i) = \sigma_i^2$. Le test de Breusch-Pagan comprend les étapes suivantes :

Etape 1 Exécuter une régression de l'équation (4.65) et obtenir les résidus e_i de cette équation de régression.

Etape 2 Exécuter la régression auxiliaire suivante :

$$e_i^2 = a_0 + a_1 Z_{i1} + a_2 Z_{i2} + \dots + a_p Z_{ip} + v_i \quad (4.66)$$

où $Z_{i,m}$ est un ensemble de variables censées déterminer la variance du terme d'erreur. (Généralement, pour $Z_{i,m}$, les variables explicatives de l'équation de régression d'origine sont utilisées, c'est-à-dire les X 's)

Etape 3 Formuler les hypothèses nulle et alternative. L'hypothèse nulle d'homoscédasticité est :

$$H_0 : a_0 = a_1 = \dots = a_p = 0 \quad (4.67)$$

tandis que l'hypothèse alternative est qu'au moins l'un des a 's est différent de zéro et qu'au moins l'un des Z 's affecte la variance des résidus, qui sera différente pour différents t .

Etape 4 Calculer la statistique $LM = nR^2$, où n est le nombre d'observations utilisées pour estimer la régression auxiliaire à l'étape 2, et R^2 est le coefficient de détermination de cette régression. La statistique LM suit la distribution χ^2 avec $p - 1$ degrés de liberté.

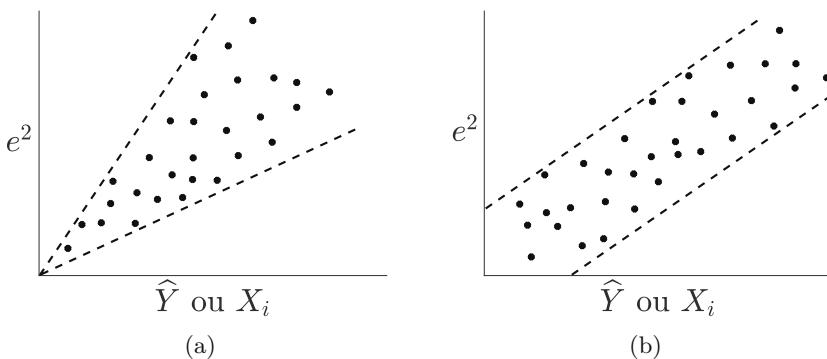


FIGURE 4.11 – Indications de la Présence d'Hétéroscédasticité

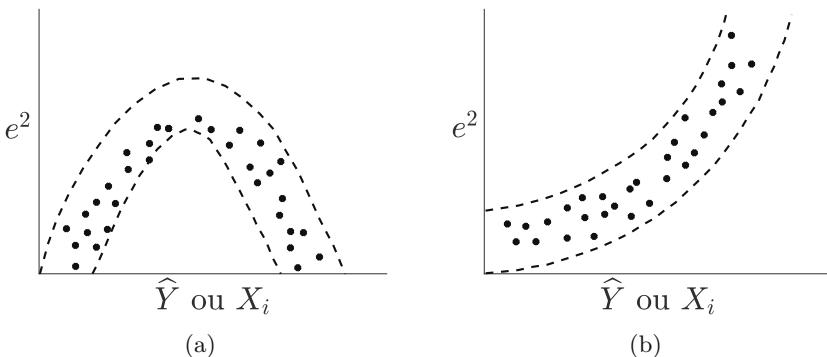


FIGURE 4.12 – Relations Non Linéaires Menant à l'Hétéroscédasticité

Etape 5 Lorsque la statistique LM est supérieure à la valeur critique ($LM_{stat} > \chi^2_{p-1,\alpha}$), on rejette l'hypothèse nulle en concluant qu'il existe des preuves significatives d'hétéroscléasticité. Alternative-
ment, calculez la $p-value$ et rejetez l'hypothèse nulle si la $p-value$ est inférieure au niveau de signification α (généralement $\alpha = 0.05$).

Dans ce cas (et comme dans tous les autres tests LM que nous examinerons plus tard), l'équation auxiliaire fait une hypothèse explicite sur la forme d'hétéroscédasticité qui peut être attendue dans les données. Il existe trois autres tests LM , qui introduisent différentes formes de régressions auxiliaires, suggérant différentes formes fonctionnelles concernant la relation des résidus au carré et les variables explicatives.

Le Test de Breusch-Pagan LM avec Eviews

Le test de Breusch-Pagan peut être effectué avec EViews comme suit. Le modèle de régression doit d'abord être estimé avec la méthode des MCO à l'aide de la commande :

```
ls y c x1 x2 x3...xm
```

où **y** est la variable dépendante et **x1** à **xm** sont les variables explicatives. Ensuite, la commande *generate* (*genr*) est utilisée pour obtenir les résidus :

```
genr et = resid
```

Noter qu'il est important de taper et d'exécuter cette commande immédiatement après avoir obtenu les résultats de l'équation afin que le vecteur *resid* ait le résidu de l'équation estimée précédemment. Ici **et** est utilisé pour les termes d'erreur de ce modèle. Les résidus au carré sont ensuite calculés comme suit :

```
genr etsq = et^2
```

et l'estimation de la régression auxiliaire est obtenue à partir de la commande :

```
ls etsq c z1 z2 z3...zp
```

Pour obtenir la statistique *LM*, on calcul $LM = n \times R^2$, où n est le nombre d'observations et R^2 est le coefficient de détermination de la régression auxiliaire.

Enfin, des conclusions sont tirées de la comparaison entre *LM* critique et la statistique *LM*.

Remarque

EViews inclut déjà une routine pour exécuter le test d'hétéroscédatitité de Breusch-Pagan. Pour appliquer ce test, il suffit d'estimer la relation de régression, puis cliquer sur :

```
View / Residual Diagnostics / Heteroskedasticity Test /
Breusch-Pagan-Godfrey.
```

Le Test de Breusch-Pagan LM avec Stata

Pour effectuer le test de Breusch-Pagan avec Stata, le modèle de régression doit d'abord être estimé avec la méthode des MCO à l'aide de la commande :

4 Violations des Hypothèses Classiques

regress y c x1 x2 x3...xm

où **y** est la variable dépendante et **x1** à **xm** sont les variables explicatives. Les résidus sont obtenus en utilisant la commande *predict* comme suit :

predict et , residual

où **et** représente les résidus. Les résidus au carré sont ensuite calculés comme suit :

g etsq = et^2

et l'estimation de la régression auxiliaire est obtenue à partir de la commande :

regress etsq c z1 z2 z3...zp

Pour obtenir la statistique LM , on calcul $LM = n \times R^2$, où n est le nombre d'observations et R^2 est le coefficient de détermination de la régression auxiliaire.

Enfin, des conclusions sont tirées de la comparaison entre LM critique et la statistique LM .

Le Test de Glesjer LM

Le test de Glesjer (1969) se base sur les étapes étapes de test de Breusch-Pagan à l'exception de l'étape 2, qui implique une équation de régression auxiliaire différente.

Etape 1 Exécuter une régression de l'équation (4.65) et obtenir les résidus e_i de cette équation de régression.

Etape 2 Exécuter la régression auxiliaire suivante :

$$|e_i| = a_0 + a_1 Z_{i1} + a_2 Z_{i2} + \dots + a_p Z_{ip} + v_i \quad (4.68)$$

Etape 3 Formuler les hypothèses nulle et alternative. L'hypothèse nulle d'homoscédasticité est :

$$H_0 : a_0 = a_1 = \dots = a_p = 0 \quad (4.69)$$

tandis que l'hypothèse alternative est qu'au moins l'un des a 's est différent de zéro.

4 Violations des Hypothèses Classiques

Etape 4 Calculer la statistique $LM = nR^2$, où n est le nombre d'observations utilisées pour estimer la régression auxiliaire à l'étape 2, et R^2 est le coefficient de détermination de cette régression. La statistique LM suit la distribution χ^2 avec $p - 1$ degrés de liberté.

Etape 5 Lorsque la statistique LM est supérieure à la valeur critique ($LM_{stat} > \chi^2_{p-1,\alpha}$), on rejette l'hypothèse nulle en concluant qu'il existe des preuves significatives d'hétéroscélasticité. Alternative-ment, calculez la $p-value$ et rejetez l'hypothèse nulle si la $p-value$ est inférieure au niveau de signification α (généralement $\alpha = 0.05$).

Le Test de Glesjer LM avec Eviews

Le test de Glesjer peut être effectué avec EViews comme suit. Le modèle de régression doit d'abord être estimé avec la méthode des MCO à l'aide de la commande :

```
ls y c x1 x2 x3...xm
```

où **y** est la variable dépendante et **x1** à **xm** sont les variables explicatives. Ensuite, la commande *generate* (genr) est utilisée pour obtenir les résidus :

```
genr et = resid
```

Noter qu'il est important de taper et d'exécuter cette commande immédiatement après avoir obtenu les résultats de l'équation afin que le vecteur **resid** ait le résidu de l'équation estimée précédemment. Ici **et** est utilisé pour les termes d'erreur de ce modèle. La valeur absolue des résidus est ensuite calculée comme suit :

```
genr abset = abs(et)
```

et l'estimation de la régression auxiliaire est obtenue à partir de la commande :

```
ls abset c z1 z2 z3...zp
```

Pour obtenir la statistique LM , on calcul $LM = n \times R^2$, où n est le nombre d'observations et R^2 est le coefficient de détermination de la régression auxiliaire.

Enfin, des conclusions sont tirées de la comparaison entre LM critique et la statistique LM .

Remarque

EViews inclut déjà une routine pour exécuter le test d'hétéroscélasticité de Glejser. Pour appliquer ce test, il suffit d'estimer la relation de régression, puis cliquer sur **View / Residual Diagnostics / Heteroskedasticity Test / Glejser**.

Le Test de Glesjer LM avec Stata

Pour effectuer le test de Glesjer avec Stata, le modèle de régression doit d'abord être estimé avec la méthode des MCO à l'aide de la commande :

```
regress y x1 x2 x3...xm
```

où **y** est la variable dépendante et **x1** à **xm** sont les variables explicatives. Les résidus sont obtenus en utilisant la commande *predict* comme suit :

```
predict et , residual
```

où **et** représente les résidus. La valeur absolue des résidus est ensuite calculée comme suit :

```
g abset = abs(et)
```

et l'estimation de la régression auxiliaire est obtenue à partir de la commande :

```
regress abset z1 z2 z3...zp
```

Pour obtenir la statistique *LM*, on calcul $LM = n \times R^2$, où n est le nombre d'observations et R^2 est le coefficient de détermination de la régression auxiliaire.

Enfin, des conclusions sont tirées de la comparaison entre *LM* critique et la statistique *LM*.

Le Test de Harvey-Godfrey LM

Harvey (1976) et Godfrey (1978) ont développé le test suivant :

Etape 1 Exécuter une régression de l'équation 4.65 et obtenir les résidus e_i de cette équation de régression.

Etape 2 Exécuter la régression auxiliaire suivante :

$$\ln(e_i^2) = a_0 + a_1 Z_{i1} + a_2 Z_{i2} + \dots + a_p Z_{ip} + v_i \quad (4.70)$$

Etape 3 Formuler les hypothèses nulle et alternative. L'hypothèse nulle d'homoscédasticité est :

$$H_0 : a_0 = a_1 = \dots = a_p = 0 \quad (4.71)$$

tandis que l'hypothèse alternative est qu'au moins l'un des a 's est différent de zéro.

4 Violations des Hypothèses Classiques

Etape 4 Calculer la statistique $LM = nR^2$, où n est le nombre d'observations utilisées pour estimer la régression auxiliaire à l'étape 2, et R^2 est le coefficient de détermination de cette régression. La statistique LM suit la distribution χ^2 avec $p - 1$ degrés de liberté.

Etape 5 Lorsque la statistique LM est supérieure à la valeur critique ($LM_{stat} > \chi^2_{p-1,\alpha}$), on rejette l'hypothèse nulle en concluant qu'il existe des preuves significatives d'hétéroscédasticité. Alternative-ment, calculez la $p-value$ et rejetez l'hypothèse nulle si la $p-value$ est inférieure au niveau de signification α (généralement $\alpha = 0.05$).

Le Test de Harvey-Godfrey avec Eviews

Le test de Harvey-Godfrey peut être effectué avec EViews comme suit. Le modèle de régression doit d'abord être estimé avec la méthode des MCO à l'aide de la commande :

```
ls y c x1 x2 x3...xm
```

où **y** est la variable dépendante et **x1** à **xm** sont les variables explicatives. Ensuite, la commande *generate* (genr) est utilisée pour obtenir les résidus :

```
genr et = resid
```

Noter qu'il est important de taper et d'exécuter cette commande immédiatement après avoir obtenu les résultats de l'équation afin que le vecteur **resid** ait le résidu de l'équation estimée précédemment. Ici **et** est utilisé pour les termes d'erreur de ce modèle. Les résidus au carré sont ensuite calculés comme suit :

```
genr etsq = et^2
```

et l'estimation de la régression auxiliaire est obtenue à partir de la commande :

```
ls log(etsq) c z1 z2 z3...zp
```

Pour obtenir la statistique LM , on calcul $LM = n \times R^2$, où n est le nombre d'observations et R^2 est le coefficient de détermination de la régression auxiliaire.

Enfin, des conclusions sont tirées de la comparaison entre LM critique et la statistique LM .

Remarque

EViews inclut déjà une routine pour exécuter le test d'hétéroscédasticité de Harvey-Godfrey. Pour appliquer ce test, il suffit d'estimer la relation de régression, puis cliquer sur **View / Residual Diagnostics / Heteroskedasticity Test / Harvey**.

Le Test de Harvey-Godfrey avec Stata

Après avoir obtenu les résidus au carré comme décrit dans les tests précédents, le logarithme des résidus au carré doit être également obtenu. Ceci est effectué dans Stata à l'aide de la commande suivante :

```
g letsq = log(etsq)
```

où **letsq** représente le logarithmique des résidus au carré. La régression auxiliaire pour le test de Harvey-Godfrey dans Stata est :

```
regress letsq z1 z2 z3...zp
```

Pour obtenir la statistique LM , on calcul $LM = n \times R^2$, où n est le nombre d'observations et R^2 est le coefficient de détermination de la régression auxiliaire.

Enfin, des conclusions sont tirées de la comparaison entre LM critique et la statistique LM .

Le Test de Park LM

Park (1966) a développé un test alternatif, comprenant les étapes suivantes :

Etape 1 Exécuter une régression de l'équation (4.65) et obtenir les résidus e_i de cette équation de régression.

Etape 2 Exécuter la régression auxiliaire suivante :

$$\ln(e_i^2) = a_0 + a_1 \ln Z_{i1} + a_2 \ln Z_{i2} + \dots + a_p \ln Z_{ip} + v_i \quad (4.72)$$

Etape 3 Formuler les hypothèses nulle et alternative. L'hypothèse nulle d'homoscédasticité est :

$$H_0 : a_0 = a_1 = \dots = a_p = 0 \quad (4.73)$$

tandis que l'hypothèse alternative est qu'au moins l'un des a 's est différent de zéro et qu'au moins l'un des Z 's affecte la variance des résidus, qui sera différente pour différents t .

Etape 4 Calculer la statistique $LM = nR^2$, où n est le nombre d'observations utilisées pour estimer la régression auxiliaire à l'étape 2, et R^2 est le coefficient de détermination de cette régression. La statistique LM suit la distribution χ^2 avec $p - 1$ degrés de liberté.

4 Violations des Hypothèses Classiques

Etape 5 Lorsque la statistique LM est supérieure à la valeur critique ($LM_{stat} > \chi^2_{p-1,\alpha}$), on rejette l'hypothèse nulle en concluant qu'il existe des preuves significatives d'hétérosécédasticité. Alternative-ment, calculez la $p-value$ et rejetez l'hypothèse nulle si la $p-value$ est inférieure au niveau de signification α (généralement $\alpha = 0.05$).

Le Test de Park avec Eviews

Le test de Park peut être effectué avec EVViews comme suit. Le modèle de régression doit d'abord être estimé avec la méthode des MCO à l'aide de la commande :

```
ls y c x1 x2 x3...xm
```

où **y** est la variable dépendante et **x1** à **xm** sont les variables explicatives. Ensuite, la commande *generate* (genr) est utilisée pour obtenir les résidus :

```
genr et = resid
```

Noter qu'il est important de taper et d'exécuter cette commande immédiatement après avoir obtenu les résultats de l'équation afin que le vecteur **resid** ait le résidu de l'équation estimée précédemment. Ici **et** est utilisé pour les termes d'erreur de ce modèle. Les résidus au carré sont ensuite calculés comme suit :

```
genr etsq = et^2
```

et l'estimation de la régression auxiliaire est obtenue à partir de la commande :

```
ls log(etsq) c log(z1) log(z2) log(z3)...log(zp)
```

Pour obtenir la statistique LM , on calcul $LM = n \times R^2$, où n est le nombre d'observations et R^2 est le coefficient de détermination de la régression auxiliaire.

Enfin, des conclusions sont tirées de la comparaison entre LM critique et la statistique LM .

Le Test de Park avec Stata

Le test de Park peut être effectué avec Stata d'une manière similaire aux autres tests déjà décrits, en utilisant la régression auxiliaire suivante :

```
regress letsq lz1 lz2 lz3...lzp
```

ce qui nécessite simplement que nous obtenions d'abord le logarithme des variables **z1**, ..., **zp** avec la commande *generate* (g) dans Stata.

Critique des Tests LM

Une critique évidente de tous les tests LM décrits ci-dessus est qu'ils nécessitent une connaissance préalable de ce qui pourrait être à l'origine de l'hétéroscédasticité capturée sous la forme de l'équation auxiliaire. Des modèles alternatifs ont été proposés et sont présentés ci-dessous.

Le Test de Goldfeld-Quandt

Goldfeld et Quandt (1965) ont proposé un autre test basé sur l'idée que si les variances des résidus sont les mêmes pour toutes les observations (c'est-à-dire homoscédastiques), alors la variance pour une partie de l'échantillon devrait être la même que la variance pour un autre. Pour appliquer le test, il est nécessaire d'identifier une variable à laquelle la variance des résidus est principalement liée (cela peut être fait avec des graphiques des résidus par rapport aux variables explicatives). Les étapes du test Goldfeld-Quandt sont les suivantes :

Etape 1 Identifier une variable étroitement liée à la variance du terme d'erreur et classer (ou ordonner) les observations de cette variable par ordre décroissant (de la valeur la plus élevée à la valeur la plus faible).

Etape 2 Divisez l'échantillon ordonné en deux sous-échantillons de taille égale en omettant c observations centrales (la valeur de c doit être approximativement égale au quart du nombre d'observations totales), de sorte que les deux sous-échantillons contiendront $\frac{1}{2}(n - c)$ observations. Le premier contiendra les valeurs les plus élevées et le second les plus faibles.

Etape 3 Exécuter une régression des MCO de Y sur X utilisée à l'étape 1 pour chaque sous-échantillon et obtenir SCR pour chaque équation.

Etape 4 Calculez la statistique F comme suit :

$$F^* = \frac{SCR_1}{SCR_2} \quad (4.74)$$

où SCR avec la plus grande valeur est dans le numérateur. La statistique F^* est distribuée avec $F_{\frac{(n-c)}{2}-k-1; \frac{(n-c)}{2}-k-1}$ degrés de liberté (k étant le nombre des variables explicatives).

Etape 5 Rejeter l'hypothèse nulle d'homoscédasticité si F empirique $> F$ -critique.

4 Violations des Hypothèses Classiques

Si les termes d'erreur sont homoscédastiques, la variance des résidus sera la même pour chaque échantillon, de sorte que le rapport est égal à l'unité. Si le rapport est significativement plus grand, l'hypothèse nulle auquelle les variances sont égales sera rejetée. La valeur de c est choisie arbitrairement et devrait généralement se situer entre 1/6 et 1/3 des observations.

Le problème du test de Goldfeld-Quandt est qu'il ne prend pas en compte les cas où l'hétéroscédasticité est causée par plusieurs variables et qu'il n'est pas toujours adapté aux données en séries chronologiques. Cependant, il s'agit d'un modèle très populaire pour le cas de régression simple (avec une seule variable explicative).

Le Test de Goldfeld-Quandt avec Eviews

Pour effectuer le test de Goldfeld-Quandt avec EViews, les données doivent d'abord être triées dans l'ordre décroissant en fonction de la variable supposée être à l'origine de l'hétéroscédasticité. Pour ce faire, cliquez sur **Procs / Sort Series**, entrez le nom de la variable (dans ce cas X) dans la boîte de dialogue de clé de tri et cochez «décroissant» pour l'ordre de tri. L'échantillon est ensuite divisé en deux sous-échantillons différents et la régression de Y sur X est exécutée pour les deux sous-échantillons afin d'obtenir les SCR . Les commandes suivantes sont utilisées :

smpl start end ls y c x scalar scr1=@ssr	et	smpl start end ls y c x scalar scr2=@ssr
---	----	---

La statistique F est ensuite calculée, donnée par SCR_1/SCR_2 avec la commande suivante :

```
genr F_GQ=scr1/scr2
```

et en comparaison avec à la valeur F critique donnée par :

```
scalar f_crit=@qfdist(.95,n1-k-1,n2-k-1)
```

Alternativement, la $p - value$ peut être obtenue avec la commande ci-dessous pour tirer des conclusions pour le test :

```
scalar p_value=1-@fdist(.05,n1-k,n2-k)
```

Le Test de Goldfeld-Quandt avec Stata

Pour effectuer le test de Goldfeld-Quandt avec Stata, les données doivent d'abord être triées dans l'ordre décroissant en fonction de la variable supposée être à l'origine de l'hétérosécédasticité (dans ce cas X) en utilisant la commande ci-dessous :

```
sort x
```

L'échantillon est ensuite divisé en deux sous-échantillons différents et la régression de Y sur X est exécutée pour les deux sous-échantillons afin d'obtenir les SCR . Les commandes suivantes sont utilisées (en supposant ici un échantillon de 100 observations au total, l'échantillon étant divisé en 40 premières (1 – 40) et les 40 dernières (61 – 100), laissant 20 observations intermédiaires (41 – 60) hors de l'estimation) :

Pour le 1^{er} échantillon

```
regress y x in 1/40
scalar scr1 = e(rmse)^2
scalar df_scr1 = e(df_r)
```

Pour le 2^{ème} échantillon

```
regress y x in 61/100
scalar scr2 = e(rmse)^2
scalar df_scr2 = e(df_r)
```

La statistique F de Goldfeld-Quandt est ensuite calculée avec la commande suivante :

```
scalar FGQ = scr1/scr1
```

puis on compare avec la valeur F critique donnée par la commande suivante :

```
scalar Fcrit = invFtail(df_scr1,df_scr2,.05)
```

Alternativement, la $p - value$ est donnée par la commande suivante :

```
scalar pvalue = Ftail(df_scr1,df_scr2,FGQ)
```

Les résultats sont affichés en entrant les commandes ci-dessous dans la zone des commandes :

```
scalar list FGQ pvalue Fcrit
```

Le Test de White

White (1980) a développé un test plus général d'hétérosécédasticité qui élimine les problèmes apparus lors des tests précédents. Le test de White est également un test LM, mais il présente les avantages comme : (a) Il ne suppose aucune détermination préalable de l'hétérosécédasticité, (b)

4 Violations des Hypothèses Classiques

contrairement au test de Breusch-Pagan, il ne dépend pas de l'hypothèse de normalité, et (c) il propose un choix particulier pour les Z 's dans la régression auxiliaire.

Les étapes impliquées dans le test de White sont décrites ci-dessous. On suppose un modèle à deux variables explicatives :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad (4.75)$$

Etape 1 Exécuter une régression de l'équation (4.75) et obtenir les résidus e_i de cette équation de régression.

Etape 2 Exécuter la régression auxiliaire suivante :

$$e_i^2 = a_0 + a_1 X_{i1} + a_2 X_{i2} + a_3 X_{i1}^2 + a_4 X_{i2}^2 + a_5 X_{i1} X_{i2} + v_i \quad (4.76)$$

Autrement dit, régresser les résidus au carré respectivement sur une constante, toutes les variables explicatives, les variables explicatives au carré et leurs produits croisés respectifs.

Etape 3 Formuler les hypothèses nulle et alternative. L'hypothèse nulle d'homoscédasticité est :

$$H_0 : a_0 = a_1 = \cdots = a_p = 0 \quad (4.77)$$

tandis que l'hypothèse alternative est qu'au moins l'un des a 's est différent de zéro.

Etape 4 Calculer la statistique $LM = nR^2$, où n est le nombre d'observations utilisées pour estimer la régression auxiliaire à l'étape 2, et R^2 est le coefficient de détermination de cette régression. La statistique LM suit la distribution χ^2 avec p degrés de liberté.

Etape 5 Lorsque la statistique LM est supérieure à la valeur critique ($LM_{stat} > \chi^2_{p,\alpha}$), on rejette l'hypothèse nulle en concluant qu'il existe des preuves significatives d'hétéroscédasticité. Alternative-ment, calculez la p -value et rejetez l'hypothèse nulle si la p -value est inférieure au niveau de signification α (généralement $\alpha = 0.05$).

Le Test de White avec Eviews

EViews inclut déjà une routine pour exécuter le test d'hétéroscédasticité de White. Après avoir obtenu les résultats des MCO, cliquez sur

View / Residual Diagnostics / Heteroskedasticity Tests.

4 Violations des Hypothèses Classiques

Une nouvelle fenêtre s'ouvre qui comprend divers tests, parmi lesquels le **test de White**. Notez que EViews offre la possibilité d'inclure ou d'exclure des termes croisés en cliquant (où ne pas cliquant) sur le bouton « **Include White cross terms** ». Dans les deux cas, EViews fournit les résultats de l'équation de régression auxiliaire estimée, ainsi que le test LM et sa $p - value$ respective.

Le Test de White avec Stata

Le test de White peut être effectué avec Stata comme suit. Premièrement, Le modèle de régression doit d'abord être estimé (En supposant ici pour simplifier qu'il n'y a que deux variables explicatives X_1 et X_2) :

```
regress y x1 x2
```

Les résidus sont obtenus en utilisant la commande *predict* comme suit :

```
predict et , residual
```

où **et** représente les résidus. Les résidus au carré sont ensuite calculés comme suit :

```
g etsq = et^2
```

et l'estimation de la régression auxiliaire est obtenue à partir de la commande :

```
regress etsq c x1 x2 x1^2 x2^2 x1*x2
```

Pour obtenir la statistique LM , on calcul $LM = n \times R^2$, où n est le nombre d'observations et R^2 est le coefficient de détermination de la régression auxiliaire.

Enfin, des conclusions sont tirées de la comparaison entre LM critique et la statistique LM .

Test ARCH d'Engle⁴

Jusqu'à présent, nous avons recherché la présence d'une autocorrélation dans les termes d'erreur d'un modèle de régression. Engle (1982) a introduit un nouveau concept permettant à l'autocorrélation de se produire dans la variance des termes d'erreur, plutôt que dans les termes

4. Ce test ne s'applique que dans le cas d'une série chronologique et donc dans cette section les variables sont indexées par t .

4 Violations des Hypothèses Classiques

d'erreur eux-mêmes. Pour capturer cette autocorrélation, Engle a développé le Modèle **ARCH** (**Hétéroscédasticité Conditionnelle Auto-régressive**), l'idée clé derrière laquelle est que la variance de e_t dépend de la taille du terme d'erreur au carré décalé d'une période (soit e_{t-1}^2).

Plus analytiquement, considérons le modèle de régression suivant :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \varepsilon_i \quad (4.78)$$

et supposons que la variance du terme d'erreur suit un processus ARCH(1) :

$$\text{var}(e_t) = \sigma_t^2 = \gamma_0 + \gamma_1 e_{t-1}^2 \quad (4.79)$$

S'il n'y a pas d'autocorrélation dans $\text{var}(e_t)$, alors γ_1 doit être nul et donc $\sigma_t^2 = \gamma_0$. Il y a donc une variance constante (homoscédastique).

Le modèle peut facilement être étendu pour des effets ARCH(p) d'ordre supérieur :

$$\text{var}(e_t) = \sigma_t^2 = \gamma_0 + \gamma_1 e_{t-1}^2 + \gamma_2 e_{t-2}^2 + \dots + \gamma_p e_{t-p}^2 \quad (4.80)$$

Ici, l'hypothèse nulle est :

$$H_0 : \gamma_1 = \gamma_2 = \dots = \gamma_p = 0 \quad (4.81)$$

c'est-à-dire qu'aucun effet ARCH n'est présent. Les étapes du test ARCH sont les suivantes :

Etape 1 Exécuter une régression de l'équation (4.78) et obtenir les résidus e_i de cette équation de régression.

Etape 2 Régresser les résidus au carré (e_t^2) contre une constante, e_{t-1}^2 , $e_{t-2}^2, \dots, e_{t-p}^2$ (la valeur de p sera déterminée par l'ordre de ARCH(p) sujet de test).

Etape 3 Calculer la statistique $LM = (n - p) R^2$, à partir de la régression dans l'étape 2. Si $LM > \chi_p^2$ pour un niveau de signification donné, on rejette l'hypothèse nulle (c'est-à-dire aucun effet ARCH) et on conclue que des effets ARCH sont effectivement présents.

The Test ARCH-LM avec EViews et Stata

Après avoir estimé une équation de régression avec EViews, cliquez sur **View / Residual Diagnostics / Heteroskedasticity Tests**. Une nouvelle fenêtre apparaît, qui comprend divers tests possibles (notez ici que cette fenêtre offre la possibilité de faire les tests examinés ci-dessus d'une manière différente). Parmi les différentes possibilités, choisissez le test

4 Violations des Hypothèses Classiques

ARCH, spécifiez le nombre de retards que nous voulons utiliser et cliquez sur OK pour obtenir les résultats du test. Celles-ci sont interprétées de la manière habituelle.

Enfin, dans Stata, après avoir estimé un modèle de régression, le test ARCH-LM peut être effectué à l'aide du menu **Statistics** et en choisissant :

**Statistics / Linear models and related / Regression
Diagnostics / Specification tests**

Sélectionnez dans la liste "Test for ARCH effects in the residuals (archlm test – time series only)", et spécifier le nombre de retards à tester. Les résultats apparaissent immédiatement dans la fenêtre des résultats. Une manière plus simple et beaucoup plus rapide consiste à utiliser la commande suivante :

```
estat archlm , lag(number)
```

où **number** doit être remplacé par le nombre de retards à tester pour les effets ARCH. Par conséquent, pour tester quatre termes résiduels carrés décalés, tapez :

```
estat archlm , lags(4)
```

De même, pour les autres ordres de décalage, modifiez le nombre entre parenthèses.

4.3.6 Correction de l'Hétéroscédasticité

Si une hétéroskédasticité est trouvée, il y a deux façons de procéder. Premièrement, le modèle peut être réestimé d'une manière qui reconnaît pleinement la présence du problème et qui impliquerait l'application de la méthode des moindres carrés généralisée (ou pondérée). Cela produirait alors un nouvel ensemble d'estimations de paramètres qui serait plus efficace que celles des MCO et un ensemble correct de covariances et des statistiques empiriques t . Alternativement, nous pouvons reconnaître que même si l'estimateur des MCO n'est plus le meilleur, il est toujours cohérent et le vrai problème est que les covariances et les statistiques t sont tout simplement erronées. Nous pouvons ensuite corriger les covariances et les statistiques t en les basant sur un ensemble de formules telles que l'équation (4.64). Bien sûr, cela ne changera pas les estimations réelles des paramètres, qui resteront moins qu'efficaces.

Lorsque σ_i^2 est Connue : La Méthode des Moindres Carrés Pondérés

Comme nous l'avons vu dans la sous-section 4.3.3, si σ_i^2 est connue, la méthode la plus simple pour corriger l'hétéroscléasticité est au moyen des moindres carrés pondérés, car les estimateurs ainsi obtenus sont BLUE.

Lorsque σ_i^2 est Inconnue

Un problème pratique majeur avec les moindres carrés généralisés (MCG) et les moindres carrés pondérés (MCP) est que σ_i^2 est inconnue et donc l'équation (4.53) ne peut pas être estimée sans faire des hypothèses explicites sur la structure de σ_i^2 .

Cependant, s'il existe une croyance préalable sur la structure de σ_i^2 , alors MCG et MCP seront valables dans la pratique. Considérons l'équation (4.50) où on a :

$$var(\varepsilon_i) = \sigma_i^2 = \sigma^2 Z_i^2 \quad (4.82)$$

où Z_i est une variable dont les valeurs sont connues pour tout i . La division de chaque terme dans l'équation (4.50) par Z_i donne :

$$\frac{Y_i}{Z_i} = \beta_0 \left(\frac{1}{Z_i} \right) + \beta_1 \left(\frac{X_{i1}}{Z_i} \right) + \beta_2 \left(\frac{X_{i2}}{Z_i} \right) + \cdots + \beta_m \left(\frac{X_{mi}}{Z_i} \right) + \left(\frac{\varepsilon_i}{Z_i} \right) \quad (4.83)$$

ou

$$Y_i^* = \beta_0^* X_{i0}^* + \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \cdots + \beta_m^* X_{mi}^* + \varepsilon_i^* \quad (4.84)$$

où les termes marqués d'une étoile désignent des variables divisées par Z_i . Dans ce cas :

$$var(\varepsilon_i^*) = var\left(\frac{\varepsilon_i}{Z_i}\right) = \sigma^2 \quad (4.85)$$

Le problème d'hétéroscléasticité a été résolu à partir du modèle d'origine. Notez, cependant, que cette équation n'a pas de terme constant ; la constante de la régression d'origine (β_0 dans l'équation (4.78)) devient le coefficient sur X_i^* dans l'équation (4.84). Il faut être prudent dans l'interprétation des coefficients, surtout lorsque Z_i est une variable explicative dans le modèle original équation (4.50). Si, par exemple, $Z_i = X_{i2}$, alors :

$$\frac{Y_i}{Z_i} = \beta_0 \left(\frac{1}{Z_i} \right) + \beta_1 \left(\frac{X_{i1}}{Z_i} \right) + \beta_2 \left(\frac{X_{i2}}{Z_i} \right) + \cdots + \beta_m \left(\frac{X_{mi}}{Z_i} \right) + \left(\frac{\varepsilon_i}{Z_i} \right) \quad (4.86)$$

4 Violations des Hypothèses Classiques

ou

$$\frac{Y_i}{Z_i} = \beta_0 \left(\frac{1}{Z_i} \right) + \beta_1 \left(\frac{X_{i1}}{Z_i} \right) + \beta_2 + \cdots + \beta_m \left(\frac{X_{mi}}{Z_i} \right) + \left(\frac{\varepsilon_i}{Z_i} \right) \quad (4.87)$$

Si cette forme des MCP est utilisée, alors les coefficients obtenus doivent être interprétés très soigneusement. Notez que β_2 est maintenant le terme constant de l'équation (4.84), alors qu'il s'agissait d'un coefficient de pente dans l'équation (4.50), et β_0 est maintenant un coefficient de pente dans l'équation (4.84). L'effet de X_{2i} dans l'équation (4.50) peut donc être recherché en examinant l'ordonnée à l'origine dans l'équation (4.84); l'autre cas peut être abordé de la même manière.

Méthodes d'Estimation Cohérente de l'Hétéroscléasticité

White (1980) a proposé une méthode pour obtenir des estimateurs cohérents des variances et des covariances des estimateurs MCO. Cependant, plusieurs logiciels informatiques, y compris EViews, sont désormais capables de calculer les variances et les erreurs standard corrigées de l'hétéroscléasticité de White.

Correction de l'Hétéroscléasticité avec Eviews

Si, dans un exemple donné avec deux variables explicatives, les tests montrent des preuves d'hétéroscléasticité, alors des méthodes d'estimation alternatives sont nécessaires à la place des MCO.

Cependant, nous savons qu'en raison de l'hétéroscléasticité, les erreurs types des estimations du coefficient MCO sont incorrectes. Pour obtenir les estimations d'erreur standard corrigées de White, cliquer sur **Quick / Estimate Equation**, puis sur le bouton **Options** situé en haut à droite de la fenêtre **Equation Specification**. Dans la fenêtre **Estimation Options** qui s'ouvre, cliquer sur la case **covariance method** et choisir **Huber-White**, enfin sur **OK**. De retour à la fenêtre **Equation Specification**, entrer l'équation de régression requise en tapant :

y c x1 x2

puis cliquer sur **OK**. Les résultats obtenus seront ceux indiqués dans un tableau où les erreurs standard de White ne sont plus les mêmes que celles du cas des MCO simple, bien que les coefficients soient, bien sûr, identiques. Les erreurs standard corrigées de White fournissent ainsi une meilleure estimation (plus précise).

4 Violations des Hypothèses Classiques

Alternativement, EViews nous permet également d'utiliser la méthode des moindres carrés pondérés ou généralisés. En supposant que la variable à l'origine de l'hétéroscléasticité est la variable **x2** (ou en notation mathématique en supposant que) :

$$var(\varepsilon_i) = \sigma_i^2 = \sigma_{X_2}^2 \quad (4.88)$$

alors la variable de poids sera $1/\sqrt{X_2}$. Pour ce faire, cliquez sur **Quick / Estimate Equation** puis sur **Options**, cette fois dans la case **Weights**, on choisit **Inverse std. dev.**, puis en entrant la variable de pondération $1/\sqrt{X_2}$ dans la case **Weight series** en tapant :

x2^{-0.5}

De retour à la fenêtre **Equation Specification**, entrer l'équation de régression requise en tapant :

y c x1 x2

puis cliquez sur **OK**.

Les résultats de cette méthode sont clairement différents de la simple estimation MCO. Le lecteur peut les utiliser pour calculer et comparer les erreurs standard et les intervalles de confiance pour.

De même, dans Stata, afin d'obtenir des résultats corrigés de l'hétéroscléasticité à travers les moindres carrés pondérés ou généralisés, accédez à **Statistics / Linear models and related / Linear regression** pour obtenir la fenêtre de dialogue **regress - linear regression**. Compléter les variables dépendantes et explicatives dans l'onglet **Model**, tandis que dans l'onglet **Weights**, cocher le bouton **Analytic weights** et spécifier le poids souhaité (dans ce cas, il est $1/X_2$) dans la case. Cliquez sur **OK** pour obtenir les résultats corrigés de l'hétéroscléasticité. Alternativement, cela peut être fait plus simplement en utilisant la commande :

regress y x1 x2 [aweight = 1/x2]

Remarque

La correction de l'hétéroscléasticité peut être réalisée en changeant la forme spécifique du modèle de régression par exemple en passant à un modèle Log (voir la section 3.14).

4.4 Questions et Exercices

Questions

- Définir ce que c'est une multicolinéarité et expliquer ses conséquences sur des estimations des MCO.
- Considérer le modèle suivant :

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + \varepsilon_t$$

supposons que X_4 est une combinaison linéaire parfaite de X_2 . Montrer que dans ce cas, il est impossible d'obtenir des estimations MCO.

- Nous savons à partir du chapitre 3 que $\hat{\beta} = (X'X)^{-1}X'Y$. Qu'arrive-t-il à $\hat{\beta}$ lorsqu'il y a une colinéarité parfaite entre les X 's ? Comment sauriez-vous s'il existe une colinéarité parfaite ?
- Expliquez ce qu'est le *VIF* et son utilisation.
- Montrez comment détecter une multicolinéarité possible dans un modèle de régression.

Exercice 1

Le tableau (4.18) contient des données trimestrielles sur les importations (*IMP*), le Produit Intérieur Brut (*PIB*), l'Indice des Prix à la Consommation (*IPC*) et l'Indice des Prix à la Production (*IPP*) d'un pays.

- (1) Utiliser les données du tableau pour estimer modèle suivant :

$$\ln IMP_t = \beta_0 + \beta_1 \ln PIB_t + \beta_2 \ln IPC_t + \varepsilon_t$$

- (2) Calculer la matrice de corrélation des quatres variables. Vérifier s'il existe une multicolinéarité dans les données.

Solution

- (1) Estimation de l'équation avec le logarithme des importations comme variable dépendante et les logarithmes du PIB et de l'IPC comme variables explicatives peut être fait en tapant dans la ligne de commande de EViews :

```
ls log(imp) c log(pib) log(ipc)
```

4 Violations des Hypothèses Classiques

	IMP	PIB	IPC	IPP		IMP	PIB	IPC	IPP
2010(T1)	44805	92.8	120.4	80.8	2014(T2)	48456	96.7	144.5	95.9
2010(T2)	44664	93.2	126	82.6	2014(T3)	48047	98.1	144.6	96.3
2010(T3)	43852	92.1	128.1	83.6	2014(T4)	50077	98.9	145.5	96.9
2010(T4)	43056	91.6	130.1	85	2015(T1)	49297	99.3	146.8	98.8
2011(T1)	41653	91.5	130.8	86	2015(T2)	51140	99.7	149.5	99.8
2011(T2)	41504	90.9	133.6	87.6	2015(T3)	52097	100.2	149.9	100.4
2011(T3)	41845	90.8	134.2	87.9	2015(T4)	52687	100.8	150.1	100.9
2011(T4)	42470	91	135.5	88.3	2016(T1)	54859	101.8	150.9	102.2
2012(T1)	43440	90.9	136.2	89.1	2016(T2)	55293	102.2	152.8	102.6
2012(T2)	45035	90.7	139.1	90.3	2016(T3)	56297	102.8	153.1	102.6
2012(T3)	44928	91.2	139	90.4	2016(T4)	57512	103.4	154	103.2
2012(T4)	45476	91.6	139.6	91	2017(T1)	58272	104.6	154.9	103.4
2013(T1)	46011	92.1	138.7	92.4	2017(T2)	61187	105.7	156.9	103.4
2013(T2)	45252	92.6	140.9	93.9	2017(T3)	61664	106.6	158.4	103.8
2013(T3)	46295	93.5	141.3	94.3	2017(T4)	64017	107.6	159.7	103.9
2013(T4)	47049	94.5	141.8	94.6	2018(T1)	64144	108.1	160.2	104
2014(T1)	47971	95.5	142	95.3	2018(T2)	65406	108.6	163.2	104.4

TABLE 4.18 – **Données**

nous obtenons les résultats indiqués dans le tableau (4.19) :

- (2) La matrice de corrélation des quatres variables peut être obtenue à partir d'EViews en cliquant sur **Quick / Group Statistics / Correlations**. EViews nous demande de définir la liste des séries que nous voulons inclure dans le groupe et nous tapons :

imp pib ipc ipp

puis cliquez sur **OK**. Les résultats sont présentés dans le tableau (4.20) :

Nous constatons à partir de la matrice de corrélation que les corrélations entre les variables sont très élevées, mais les corrélations les plus élevées sont entre *IPC* et *IPP*, (0.981983).

A partir des résultats de régression dans le tableau (4.19), on constate que le R^2 de cette dernière est très élevé, et les deux variables semblent être positives, le log(*PIB*) étant également très significatif. Le log(*IPC*) n'est que marginalement significatif.

Cependant, l'estimation du modèle en incluant également le log(*IPP*) se fait en tapant sur la ligne de commande de EViews :

ls log(imp) c log(pib) log(ipc) log(ipp)

4 Violations des Hypothèses Classiques

Dependent Variable : LOG(IMP)

Method : Least Squares

Sample : 1 34

Variable	Coefficient	Std. Error	t-	Prob.
	Statistic			
C	0.631870	0.344368	1.834867	0.0761
LOG(PIB)	1.926936	0.168856	11.41172	0.0000
LOG(IPC)	0.274276	0.137400	1.996179	0.0548
R-squared	0.966057	Mean dependent var	10.81363	
Adjusted R-squared	0.963867	S.D. dependent var	0.138427	
S.E. of regression	0.026313	Akaike info criterion	-4.353390	
Sum squared resid	0.021464	Schwarz criterion	-4.218711	
Log likelihood	77.00763	Hannan-Quinn criter.	-4.307461	
F-statistic	441.1430	Durbin-Watson stat	0.475694	
Prob(F-statistic)	0.000000			

TABLE 4.19 – Résultats de la Régression

nous obtenons les résultats indiqués dans le tableau (4.21). Désormais, le log (*IPC*) est hautement significatif, tandis que le log (*IPP*) (qui est fortement corrélé avec log (*IPC*) et devrait donc avoir plus ou moins le même effet sur le log (*IMP*)) est négatif et très significatif. Ceci, bien sûr, est dû à l'inclusion des deux indices de prix dans la même spécification d'équation, en raison du problème de la multicolinéarité.

En estimant l'équation cette fois sans log (*IPC*) mais avec log (*IPP*), nous obtenons les résultats dans le tableau (4.22), qui montrent que log (*IPP*) est positif et non significatif. Il est clair que l'importance du log (*IPP*) dans la spécification ci-dessus était le résultat de la relation linéaire qui relie les deux variables de prix.

Les conclusions de cette analyse sont :

- La corrélation entre les variables explicatives était très élevée
- Les erreurs-types ou les *t* empiriques des coefficients estimés ont changé d'une estimation à une autre.
- La stabilité des coefficients estimés était également assez problématique, les coefficients négatifs et positifs étant estimés pour la même variable avec d'autres spécifications.

4 Violations des Hypothèses Classiques

	IMP	PIB	IPC	IPP
IMP	1.000000	0.979713	0.916331	0.883530
PIB	0.979713	1.000000	0.910961	0.899851
IPC	0.916331	0.910961	1.000000	0.981983
IPP	0.883530	0.899851	0.981983	1.000000

TABLE 4.20 – Matrice de Corrélation

Dependent Variable : LOG(IMP)

Method : Least Squares

Sample : 1 34

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.213906	0.358425	0.596795	0.5551
LOG(PIB)	1.969713	0.156800	12.56198	0.0000
LOG(IPC)	1.025473	0.323427	3.170645	0.0035
LOG(IPP)	-	0.305218	-2.524894	0.0171
		0.770644		
R-squared	0.972006	Mean dependent var	10.81363	
Adjusted R-squared	0.969206	S.D. dependent var	0.138427	
S.E. of regression	0.024291	Akaike info criterion	-4.487253	
Sum squared resid	0.017702	Schwarz criterion	-4.307682	
Log likelihood	80.28331	Hannan-Quinn criter.	-4.426014	
F-statistic	347.2135	Durbin-Watson stat	0.608648	
Prob(F-statistic)	0.000000			

TABLE 4.21 – Résultats de la Régression

Dans ce cas, il est clair que la multicolinéarité est fortement présente car nous avons inclus deux variables de prix qui sont assez fortement corrélées.

Exercice 2

Le tableau contient des données pour les variables suivantes, I = Investissement, Y = Revenu et R = Taux d'intérêt d'un pays.

- (1) Estimer une équation de régression qui a l'investissement comme variable dépendante et le revenu et le taux d'intérêt comme variables explicatives.
- (2) Vérifiez l'autocorrélation en utilisant à la fois les méthodes infor-

4 Violations des Hypothèses Classiques

Dependent Variable : LOG(IMP)

Method : Least Squares

Sample : 1 34

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.685704	0.370644	1.850031	0.0739
LOG(PIB)	2.093849	0.172585	12.13228	0.0000
LOG(IPP)	0.119566	0.136062	0.878764	0.3863
R-squared	0.962625	Mean dependent var	10.81363	
Adjusted R-squared	0.960213	S.D. dependent var	0.138427	
S.E. of regression	0.027612	Akaike info criterion	-4.257071	
Sum squared resid	0.023634	Schwarz criterion	-4.122392	
Log likelihood	75.37021	Hannan-Quinn criter.	-4.211142	
F-statistic	399.2113	Durbin-Watson stat	0.448237	
Prob(F-statistic)	0.000000			

TABLE 4.22 – Résultats de la Régression

melles et formelles (tests) que nous avons couvertes tout au long de ce chapitre.

- (3) Si l'autocorrélation existe, utiliser la procédure itérative Cochrane-Orcutt pour résoudre ce problème.

Solution

- (1) Estimation de l'équation de régression de la variable $I = \text{Investissement}$ sur les variables $Y = \text{Revenu}$ et $R = \text{Taux d'intérêt}$ peut être fait en tapant dans la ligne de commande de EViews :

ls i c y r

nous obtenons les résultats indiqués dans le tableau (4.24) :

- (2) La figure (4.13) trace les valeurs réelles, ajustées et résiduelles à l'aide d'EViews (Le graphe est obtenu à partir de la barre de menu de la fenêtre des résultats de la régression de Y sur X dans le tableau (4.24)).

Test de DW

À partir des résultats de régression obtenus dans la question précédente (détection graphique de l'autocorrélation), nous observons

4 Violations des Hypothèses Classiques

	Y	I	R		Y	I	R
1	8.58	11.53	18.12	16	23.74	21.96	17.51
2	10.47	13.25	11.07	17	25.74	23.07	16.42
3	8.34	10.87	8.98	18	24.21	25.67	7.42
4	10.64	10.46	17.01	19	25.21	26.15	15.47
5	9.65	15.09	16.26	20	26.21	25.56	19.16
6	12.01	17.49	13.78	21	28.6	28.12	5.47
7	13.45	17.77	19.95	22	30.58	24.21	9.51
8	14.2	16.11	18.73	23	25.98	21.51	7.44
9	13.83	10.66	9.53	24	26.86	22.93	19.91
10	14.45	10.59	13.79	25	31.32	32.3	7.94
11	16.57	9.32	19.31	26	32.93	24.6	21.35
12	18.02	11	15.19	27	32.1	30.44	8.65
13	18.38	15.03	12.4	28	33.29	32.51	11.12
14	20.42	15.09	16.48	29	35.59	29.49	21.68
15	20.99	22.7	5.93	30	33.86	31.18	15.82

TABLE 4.23 – **Données**

que la statistique DW est égale à 0.85215. Trouver les valeurs critiques d_L et d_U au niveau de signification de 1% pour $n = 49$ et $m = 2$ dans la table de DW dans les annexes et on les place dans le tableau de DW. Nous avons les résultats indiqués dans la figure (4.13). Étant donné que $d = 0.85215$ est inférieur à $d_L = 1.134$, il existe des preuves solides d'une corrélation sérielle positive.

Test de Breusch-Godfrey LM

Pour tester cette corrélation sérielle, nous utilisons le test Breusch-Godfrey LM. Dans la fenêtre des résultats de la régression estimée, accédez à **View / Residual Tests / Serial Correlation LM** et spécifiez 1 comme nombre de retards. Les résultats de ce test sont présentés dans le tableau

Nous pouvons voir dans les premières colonnes que les valeurs à la fois de la statistique LM et de la statistique F sont assez élevées, ce qui suggère le rejet de l'hypothèse nulle d'absence de corrélation sérielle de premier ordre. Il est également clair que cela est dû au fait que les p -values sont très petites (inférieures à 0.05 pour un intervalle de confiance à 95%). Par conséquent, une corrélation sérielle est définitivement présente.

4 Violations des Hypothèses Classiques

Dependent Variable : I

Method : Least Squares

Sample : 1 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.224938	2.510894	2.479172	0.0197
Y	0.769911	0.071791	10.72442	0.0000
R	-0.184196	0.126416	-1.457068	0.1566
R-squared	0.816282	Mean dependent var	20.22200	
Adjusted R-squared	0.802673	S.D. dependent var	7.495569	
S.E. of regression	3.329642	Akaike info criterion	5.338246	
Sum squared resid	299.3358	Schwarz criterion	5.478366	
Log likelihood	-77.07369	Hannan-Quinn criter.	5.383071	
F-statistic	59.98221	Durbin-Watson stat	0.852153	
Prob(F-statistic)	0.000000			

TABLE 4.24 – Résultats de la Régression

(3) Correction de l'Autocorrélation

Approche de Différenciation Généralisée avec Eviews

On va appliquer la même démarche déjà expliquée pour cette méthode et ce pour exécuter une estimation des MCO avec des variables modifiées :

```
ls beta0_star r_star y_star
```

le résultat de l'estimation est indiqué dans le tableau (4.26) :

Procédure itérative de Cochrane-Orcutt

La procédure itérative de Cochrane-Orcutt (1949) exécutée avec Stata moyennant la commande ci-dessous, donne les résultats donnés dans le tableau (4.27) :

```
prais i r y, corc
```

Exercice 3

Le tableau (4.28) présente des données sur les dépenses de consommation (Y) par rapport au revenu (X) pour un échantillon représentatif de 30 familles. La réorganisation nécessaire des données pour l'application des tests est également présentée dans le tableau.

4 Violations des Hypothèses Classiques

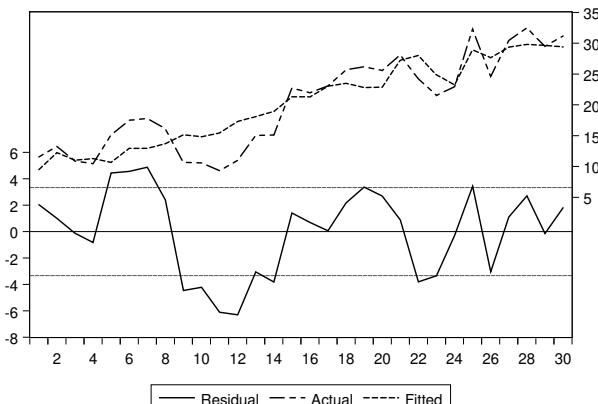


FIGURE 4.13 – Graphique des Résidus

On demande de procéder aux tests de détection d'hétérosécédasticité suivants :

- (1) Test de Goldfeld-Quandt ;
- (2) Test de Glesjer ;
- (3) Test de White ;
- (4) En cas d'hétérosécédasticité, d'en corriger les effets.

Solution

(1) Test de Goldfeld-Quandt

Etape 1 Identifier une variable étroitement liée à la variance du terme d'erreur et classer (ou ordonner) les observations de cette variable par ordre décroissant (de la valeur la plus élevée à la valeur la plus faible).

Etape 2 Divisez l'échantillon ordonné en deux sous-échantillons de taille égale en omettant c observations centrales, de sorte que les deux sous-échantillons contiendront $\frac{1}{2}(n - c)$ observations ($c = 10$, la valeur de c doit être approximativement égale au quart du nombre d'observations totales). Le premier échantillon contiendra dans notre cas les 30 valeurs les plus élevées et le second les 30 valeurs les plus faibles.

4 Violations des Hypothèses Classiques

Breusch-Godfrey Serial Correlation LM Test

F-statistic	13.04955	Prob. F(1,26)	0.0013
Obs*R-squared	10.02538	Prob. Chi-Square(1)	0.0015

Dependent Variable : RESID

Method : Least Squares

Sample : 1 30

Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.365569	2.121810	0.643587	0.5255
Y	-0.007865	0.059735	-0.131664	0.8963
R	-0.082614	0.107576	-0.767957	0.4494
RESID(-1)	0.595067	0.164728	3.612417	0.0013
R-squared	0.334179	Mean dependent var	-2.35E-15	
Adjusted R-squared	0.257354	S.D. dependent var	3.212775	
S.E. of regression	2.768671	Akaike info criterion	4.998178	
Sum squared resid	199.3040	Schwarz criterion	5.185004	
Log likelihood	-70.97266	Hannan-Quinn criter.	5.057945	
F-statistic	4.349852	Durbin-Watson stat	1.695546	
Prob(F-statistic)	0.013046			

TABLE 4.25 – **Test de Breusch-Godfrey**

Etape 3 Exécuter une régression des MCO de Y sur X utilisée à l'étape 1 pour chaque sous-échantillon et obtenir SCR pour chaque équation.

Le résultat de la régression du premier sous-échantillon est :

$$\hat{Y}_i = 78.12107 + 0.268185X_i \\ (2.391585) \quad (0.013095) \quad R^2 = 0.937416 \quad SCR_1 = 471.5663$$

Le résultat de la régression du deuxième sous-échantillon est :

$$\hat{Y}_i = 76.04923 + 0.366185X_i \\ (2.390487) \quad (0.031505) \quad R^2 = 0.828323 \quad SCR_2 = 160.1689$$

Etape 4 Calculez la statistique F comme suit :

$$F^* = \frac{SCR_1}{SCR_2} = \frac{471.5663}{160.1689} = 2.94418$$

4 Violations des Hypothèses Classiques

Dependent Variable : I_STAR

Method : Least Squares

Sample : 2 30

Included observations : 29

Variable	Coefficient	Std. Error	t-Statistic	Prob.
BETA1_STAR	7.241045	3.337342	2.169704	0.0394
R_STAR	-0.293100	0.080636	-3.634844	0.0012
Y_STAR	0.787463	0.131353	5.995034	0.0000
R-squared	0.635559	Mean dependent var	9.255682	
Adjusted R-squared	0.607525	S.D. dependent var	4.276520	
S.E. of regression	2.679146	Akaike info criterion	4.906570	
Sum squared resid	186.6234	Schwarz criterion	5.048015	
Log likelihood	-68.14527	Durbin-Watson stat	1.547610	

TABLE 4.26 – Les résultats de la Régression de Différenciation Généralisée

où SCR avec la plus grande valeur est dans le numérateur. La statistique F est distribuée avec $F_{\frac{(n-c)}{2} - k - 1; \frac{(n-c)}{2} - k - 1}^{\alpha}$ degrés de liberté dans notre exemple on a $F_{28;28}^{0.05} = 1.88$.

Étant donné que $F^* > F_{28;28}^{0.05}$, nous pouvons conclure qu'il existe une hétérosécédasticité dans la variance d'erreur.

(2) Test de Glesjer

Le test de Glesjer (1969) se base sur les étapes étapes de test de Breusch-Pagan à l'exception de l'étape 2, qui implique une équation de régression auxiliaire différente.

Etape 1 Exécuter une régression de Y sur X à l'aide des commandes ci-dessous et obtenir les résidus e_i de cette équation de régression :

```
ls y c x
```

Ensuite, la commande *generate* (*genr*) est utilisée pour obtenir les résidus :

```
genr et = resid
```

Ici **et** est utilisé pour les termes d'erreur de ce modèle. La valeur absolue des résidus est ensuite calculée comme suit :

```
genr abset = abs(et)
```

4 Violations des Hypothèses Classiques

Cochrane-Orcutt AR(1) regression - iterated estimates

Source	SS	df	MS	Number of obs F(1, 26)	= 29 = 19.83
Model	283.6556	2	141.82784	Prob > F	= 0.0000
Residual	185.9630	26	7.152426	R-squared	= 0.6040
Total	469.6187	28	16.77209	Adj R-squared	= 0.5736
				Root MSE	= 2.6744
y	Coef.	Std. Err.	t	P> t 	[95% Conf. Interval]
r	-0.295775	0.078671	-3.76	0.001	-0.45748 -0.13406
y	0.784853	0.144227	5.44	0.000	0.48839 1.08131
_cons	7.329872	3.658536	2.00	0.056	-0.19035 14.8501
rho	0.6146382				
Durbin-Watson statistic (original)				0.852153	
Durbin-Watson statistic (transformed)				1.608128	

TABLE 4.27 – La procédure Itérative de Cochrane-Orcutt avec Stata

Etape 2 Exécuter la régression auxiliaire suivante :

$$|e_i| = a_0 + a_1 X_{i1} + v_i \quad (4.89)$$

l'estimation de la régression auxiliaire est obtenue à partir de la commande :

ls abset c x

Etape 3 Formuler les hypothèses nulle et alternative. L'hypothèse nulle d'homoscédasticité est :

$$H_0 : a_0 = a_1 = 0 \quad (4.90)$$

tandis que l'hypothèse alternative est qu'au moins l'un des a 's est différent de zéro.

Les résultats de la régression auxiliaire (4.89) sont indiqués dans le tableau :

Etape 4 La statistique $LM = nR^2 = 81 \times 0.092537 = 7.49549$, où $n = 81$ est le nombre d'observations utilisées pour estimer la régression auxiliaire à l'étape 2, et $R^2 = 0.092537$ est le

4 Violations des Hypothèses Classiques

Y							X						
Obs													
1	96	22	101	42	106	62	107	1	49	22	74	42	95
2	97	23	111	43	109	63	114	2	55	23	95	43	102
3	97	24	105	44	106	64	114	3	55	24	81	44	95
4	105	25	111	45	105	65	117	4	70	25	95	45	93
5	96	26	110	46	108	66	122	5	53	26	92	46	100
6	105	27	110	47	108	67	122	6	70	27	92	47	100
7	97	28	110	48	107	68	122	7	55	28	92	48	98
8	98	29	90	49	120	69	122	8	62	29	52	49	130
9	98	30	112	50	109	70	148	9	62	30	103	50	115
10	107	31	103	51	109	71	160	10	80	31	84	51	115
11	103	32	103	52	109	72	121	11	73	32	84	52	115
12	113	33	111	53	109	73	121	12	92	33	102	53	115
13	113	34	111	54	133	74	110	13	92	34	102	54	180
14	103	35	102	55	125	75	110	14	73	35	81	55	160
15	100	36	106	56	115	76	121	15	66	36	90	56	130
16	103	37	106	57	102	77	165	16	73	37	90	57	96
17	106	38	109	58	109	78	140	17	78	38	102	58	115
18	113	39	109	59	104	79	147	18	92	39	102	59	100
19	106	40	120	60	105	80	157	19	78	40	130	60	100
20	109	41	106	61	120	81	130	20	90	41	95	61	145
21	110							21	92				

TABLE 4.28 – Données

coefficient de détermination de cette régression. La statistique LM suit la distribution χ^2 avec p degrés de liberté.

Etape 5 on a $LM_{stat} = 7.49549 < \chi^2_{1,0.05} = 3.84$, on rejette l'hypothèse nulle en concluant qu'il existe des preuves significatives d'hétéroscédasticité.

(3) Test de White

Les étapes impliquées dans le test de White sont décrites ci-dessous. On suppose un modèle à une variable explicative :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (4.91)$$

Etape 1 Exécuter une régression de l'équation (4.91) et obtenir les résidus e_i de cette équation de régression.

Etape 2 Exécuter la régression auxiliaire suivante :

4 Violations des Hypothèses Classiques

Dependent Variable : ABSET

Method : Least Squares

Sample : 1 81

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.388535	0.551710	2.516783	0.0139
X	0.012003	0.004229	2.838291	0.0058
R-squared	0.092537	Mean dependent var	2.798484	
Adjusted R-squared	0.081050	S.D. dependent var	2.253516	
S.E. of regression	2.160262	Akaike info criterion	4.402718	
Sum squared resid	368.6720	Schwarz criterion	4.461840	
Log likelihood	-176.3101	Hannan-Quinn criter.	4.426439	
F-statistic	8.055894	Durbin-Watson stat	1.428557	
Prob(F-statistic)	0.005763			

TABLE 4.29 – Résultats de la Régression

$$e_i^2 = a_0 + a_1 X_i + a_2 X_i^2 + v_i \quad (4.92)$$

Autrement dit, régresser les résidus au carré respectivement sur une constante, la variable explicative et la variable explicative au carré.

Etape 3 Formuler les hypothèses nulle et alternative. L'hypothèse nulle d'homoscédasticité est :

$$H_0 : a_0 = a_1 = a_2 = 0$$

tandis que l'hypothèse alternative est qu'au moins l'un des a 's est différent de zéro.

Etape 4 Calculer la statistique $LM = nR^2$, où n est le nombre d'observations utilisées pour estimer la régression auxiliaire à l'étape 2, et R^2 est le coefficient de détermination de cette régression. La statistique LM suit la distribution χ^2 avec 2 degrés de liberté.

Les résultats de la régression en (4.91) sont indiqués dans le tableau (4.30).

EViews inclut déjà une routine pour exécuter le test d'hétérosécédasticité de White. Après avoir obtenu les résultats des MCO,

4 Violations des Hypothèses Classiques

Dependent Variable : Y

Method : Least Squares

Sample : 1 81

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	84.37846	0.926909	91.03210	0.0000
X	0.239027	0.007105	33.64332	0.0000
R-squared	0.934758	Mean dependent var	112.4568	
Adjusted R-squared	0.933932	S.D. dependent var	14.12006	
S.E. of regression	3.629380	Akaike info criterion	5.440382	
Sum squared resid	1040.619	Schwarz criterion	5.499504	
Log likelihood	-218.3355	Hannan-Quinn criter.	5.464102	
F-statistic	1131.873	Durbin-Watson stat	1.305051	
Prob(F-statistic)	0.000000			

TABLE 4.30 – Résultats de la Régression

cliquez sur **View / Residual Diagnostics / Heteroskedasticity Tests**. Une nouvelle fenêtre s'ouvre qui comprend divers tests, parmi lesquels le **test de White**. Notez que EViews offre la possibilité d'inclure ou d'exclure des termes croisés en cliquant (où ne pas cliquant) sur le bouton « **Include White cross terms** ».

Le résultat du test de White (c'est-à-dire la régression en (4.92)) est indiqué dans le tableau (4.31).

Etape 5 Selon les résultats indiqués dans le tableau (4.31), $LM stat = nR^2 = 8.513505 > \chi^2_{2,0.05} = 5.991$, on rejette l'hypothèse nulle en concluant qu'il existe des preuves significatives d'hétéroscé-dastitité.

(4) Correction de l'Hétéroscé-dastitité :

On va procéder avec Eviews. La variable de poids sera $1/\sqrt{X}$. Pour ce faire, cliquez sur **Quick / Estimate Equation** puis sur **Options**, cette fois dans la case **Weights**, on choisit **Inverse std. dev.**, puis en entrant la variable de pondération $1/\sqrt{X}$ dans la case **Weight series** en tapant :

$$x^{(-.5)}$$

De retour à la fenêtre **Equation Specification**, entrer l'équation de régression requise en tapant :

4 Violations des Hypothèses Classiques

Heteroskedasticity Test : White				
F-statistic	4.580511	Prob. F(2,27)		0.0132
Obs*R-squared	8.513470	Prob. Chi-Square(2)		0.0142
Scaled explained SS	9.279553	Prob. Chi-Square(2)		0.0097
Dependent Variable : RESID^2				
Method : Least Squares				
Sample : 1 30				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5.747031	11.36784	-0.505552	0.6146
X^2	-0.000265	0.000473	-0.559639	0.5773
X	0.196628	0.160526	1.224896	0.2243
R-squared	0.105105	Mean dependent var	12.84715	
Adjusted R-squared	0.082159	S.D. dependent var	19.56985	
S.E. of regression	18.74871	Akaike info criterion	8.736460	
Sum squared resid	27418.09	Schwarz criterion	8.825144	
Log likelihood	-350.8266	Hannan-Quinn criter.	8.772041	
F-statistic	4.580511	Durbin-Watson stat	1.512700	
Prob(F-statistic)	0.013156			

TABLE 4.31 – Résultats du Test de White

y c x

puis cliquez sur OK.

Les résultats de l'estimation sont indiqués dans le tableau (4.32)

On peut vérifier cette correction par le test de Breusch-Pagan. Les résultats du test sont indiqués dans le tableau (4.33).

Selon les résultats indiqués dans le tableau. Il parrait clairement que le modèle n'est plus hétéroscédastique.

4.5 Annexe

Démonstration

On peut développer la statistique du test DW donnée dans l'équation (4.32) pour obtenir :

4 Violations des Hypothèses Classiques

Dependent Variable : Y

Method : Least Squares

Sample : 1 81

Weighting series : X^(-.5)

Weight type : Inverse standard deviation (EViews default scaling)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	84.78714	0.924581	91.70330	0.0000
X	0.235548	0.008583	27.44373	0.0000
Weighted Statistics				
R-squared	0.905066	Mean dependent var	110.0313	
Adjusted R-squared	0.903864	S.D. dependent var	11.98283	
S.E. of regression	3.382671	Akaike info criterion	5.299590	
Sum squared resid	903.9545	Schwarz criterion	5.358712	
Log likelihood	-212.6334	Hannan-Quinn criter.	5.323310	
F-statistic	753.1585	Durbin-Watson stat	1.424498	
Prob(F-statistic)	0.000000	Weighted mean dep.	108.0559	
Unweighted Statistics				
R-squared	0.934560	Mean dependent var	112.4568	
Adjusted R-squared	0.933731	S.D. dependent var	14.12006	
S.E. of regression	3.634883	Sum squared resid	1043.778	
Durbin-Watson stat	1.361488			

TABLE 4.32 – Correction de l'Hétéroscédasticité (Résultats de l'Estimation)

$$d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2} = \frac{\sum_{t=2}^T e_t^2 + \sum_{t=2}^T e_{t-1}^2 - 2 \sum_{t=2}^T e_t e_{t-1}}{\sum_{t=1}^T e_t^2}$$

Du fait que e_t sont généralement petits, les sommes de 2 à T ou de 2 à $T - 1$ seront approximativement égales à la sommation de 1 à T . Donc :

$$\sum_{t=2}^T e_t^2 \simeq \sum_{t=2}^T e_{t-1}^2 \simeq \sum_{t=1}^T e_t^2$$

et nous avons d'après l'équation précédente :

4 Violations des Hypothèses Classiques

Heteroskedasticity Test : Breusch-Pagan-Godfrey				
F-statistic	0.005281	Prob. F(1,79)	0.9423	
Obs*R-squared	0.005414	Prob. Chi-Square(1)	0.9413	
Scaled explained SS	0.005270	Prob. Chi-Square(1)	0.9421	
Dependent Variable : WGT_RESID^2				
Method : Least Squares				
Sample : 1 81				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.57424	8.257412	1.280576	0.2041
X*WGT	0.005465	0.075203	0.072669	0.9423
R-squared	0.000067	Mean dependent var	11.15993	
Adjusted R-squared	-0.012591	S.D. dependent var	16.06535	
S.E. of regression	16.16616	Akaike info criterion	8.428099	
Sum squared resid	20646.25	Schwarz criterion	8.487222	
Log likelihood	-339.3380	Hannan-Quinn criter.	8.451820	
F-statistic	0.005281	Durbin-Watson stat	1.697778	
Prob(F-statistic)	0.942253			

TABLE 4.33 – **Test de Breusch-Pagan**

$$d \simeq 1 + 1 - \frac{2 \sum_{t=2}^T e_t e_{t-1}}{\sum_{t=1}^T e_t^2}$$

or nous avons d'après l'équation (4.33) que $\hat{\rho} = 2 \sum_{t=2}^T e_t e_{t-1} / \sum_{t=1}^T e_t^2$, on aura :

$$d \simeq 2 - 2\rho = \simeq 2(1 - \rho)$$

Enfin, puisque ρ prend des valeurs de +1 à -1, alors d prendra des valeurs de 0 à 4.

Chapitre 5

Modèles à Équations Simultanées

Les modèles les plus importants en économie sont de nature simultanée. L'offre et la demande, par exemple, sont évidemment simultanées. Étudier la demande de poulet sans regarder également l'offre de poulet, c'est tenter de manquer des liens importants et ainsi commettre des erreurs importantes. Presque toutes les principales approches de la macroéconomie, en partant des modèles keynésiens de demande agrégée jusqu'aux schémas d'anticipations rationnelles, sont intrinsèquement simultanées. Même les modèles qui semblent intrinsèquement à équation unique se révèlent souvent beaucoup plus simultanés que vous ne le pensez. Le prix du logement, par exemple, est considérablement affecté par le niveau d'activité économique, le taux d'intérêt en vigueur dans les actifs alternatifs et un certain nombre d'autres variables déterminées simultanément.

Tout cela ne signifierait pas grand-chose pour les économétriciens si ce n'était le fait que l'estimation des systèmes d'équations simultanées avec MCO cause un certain nombre de difficultés qui ne sont pas rencontrées avec les équations simples. Le plus important, l'hypothèse classique 5, qui stipule que toutes les variables explicatives ne doivent pas être corrélées avec le terme d'erreur, est violée dans les modèles simultanés. Principalement à cause de cela, les estimateurs des MCO sont biaisés dans les modèles simultanés.

Si nous avons une causalité bidirectionnelle dans une fonction, cela implique que la fonction ne peut pas être traitée isolément comme un modèle d'équation unique, mais elle appartient à un système plus large

5 Modèles à Équations Simultanées

d'équations qui décrit les relations entre toutes les variables pertinentes. Si $Y = f(X)$, mais aussi $X = f(Y)$, nous ne sommes pas autorisés à utiliser un modèle à équation unique pour la description de la relation entre Y et X . Nous devons utiliser un modèle à équations multiples, qui comprendrait des équations distinctes dans lesquelles Y et X apparaissent comme des variables endogènes, bien qu'elles puissent apparaître comme explicatives dans d'autres équations du modèle. Un système décrivant la dépendance conjointe des variables est appelé un *système d'équations simultanées*.

Étant donné la nature des phénomènes économiques, il est presque certain que toute équation appartient à un système plus large d'équations simultanées. Plusieurs exemples illustreront la signification des relations simultanées et la violation de l'hypothèse 5 des moindres carrés ordinaires, ce qui crée ce qu'on appelle le *biais d'équations simultanées*.

5.1 Le Biais Simultané

Exemple 1 : Considérons le modèle Keynésien simple suivant :

$$C_t = \beta_0 + \beta_1 R_t + \varepsilon_t \quad t = 1, 2, \dots, T \quad (5.1)$$

$$R_t = C_t + I_t \quad (5.2)$$

où C_t désigne la consommation, R_t désigne le revenu disponible et I_t désigne l'investissement autonome. Il s'agit d'un système de deux équations simultanées, également appelées équations structurelles, la deuxième équation étant une identité. La première équation peut être estimée par MCO donnant :

$$\hat{\beta}_1 = \frac{\sum_{t=1}^T (R_t - \bar{R})(C_t - \bar{C})}{\sum_{t=1}^T (R_t - \bar{R})^2} \quad \text{et} \quad \hat{\beta}_0 = \bar{C} - \hat{\beta}_1 \bar{R} \quad (5.3)$$

Puisque I_t est autonome, il s'agit d'une variable exogène déterminée en dehors du système, tandis que C_t et R_t sont des variables *endogènes* déterminées par le système. Résolvons pour R_t et C_t en termes de constante I_t et de la constante. Les deux équations résultantes sont connues sous le nom d'*équations de forme réduite* :

5 Modèles à Équations Simultanées

$$C_t = \frac{\beta_0}{1 - \beta_1} + \frac{\beta_1 I_t}{1 - \beta_1} + \frac{\varepsilon_t}{1 - \beta_1} \quad (5.4)$$

$$R_t = \frac{\beta_0}{1 - \beta_1} + \frac{I_t}{1 - \beta_1} + \frac{\varepsilon_t}{1 - \beta_1} \quad (5.5)$$

Ces équations expriment chaque variable endogène en termes des variables exogènes et des termes d'erreur. Notez que C_t et R_t sont tous deux fonction de ε_t , et donc toutes les deux sont corrélées avec ε_t . En fait, $R_t - E(R_t) = \varepsilon_t / (1 - \beta_1)$, et

$$\text{cov}(R_t, \varepsilon_t) = E[(R_t - E(R_t)) \varepsilon_t] = \frac{\sigma_\varepsilon^2}{1 - \beta_1} \geq 0 \quad \text{si } 0 \leq \beta_1 \leq 1 \quad (5.6)$$

Cela tient parce que $\varepsilon_t \sim (0, \sigma_\varepsilon^2)$ et I_t est exogène et indépendant du terme d'erreur. L'équation (5.6) montre que le régresseur dans le côté droit de l'équation (5.1) est corrélé avec le terme d'erreur. Cela fait que les estimations de s MCO sont biaisées et *incohérentes*. En fait, à partir de l'équation (5.1) on a

$$C_t - \bar{C} = \beta_1 (R_t - \bar{R}) + (\varepsilon_t - \bar{\varepsilon})$$

et en substituant cette expression dans (5.3), on obtient

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{t=1}^T (R_t - \bar{R})(\varepsilon_t - \bar{\varepsilon})}{\sum_{t=1}^T (R_t - \bar{R})^2} \quad (5.7)$$

D'après (5.7), il est clair que $E(\hat{\beta}_1) \neq \beta_1$, puisque la valeur attendue du deuxième terme n'est pas nécessairement nulle. De plus, en utilisant (5.5) on obtient

$$R_t - \bar{R} = [(I_t - \bar{I}) + (\varepsilon_t - \bar{\varepsilon})] / (1 - \beta_1)$$

En posant $m_{RR} = \sum_{t=1}^T (R_t - \bar{R})^2 / T$, nous obtenons

$$m_{yy} = (m_{II} + 2m_{I\varepsilon} + m_{\varepsilon\varepsilon}) / (1 - \beta_1)^2 \quad (5.8)$$

$$\text{où } m_{II} = \frac{\sum_{t=1}^T (I_t - \bar{I})^2}{T}, \quad m_{I\varepsilon} = \frac{\sum_{t=1}^T (I_t - \bar{I})(\varepsilon_t - \bar{\varepsilon})}{T}$$

5 Modèles à Équations Simultanées

et $m_{\varepsilon\varepsilon} = \frac{\sum_{t=1}^T (\varepsilon_t - \bar{\varepsilon})^2}{T}$. Aussi on a

$$m_{R\varepsilon} = \frac{m_{I\varepsilon} + m_{\varepsilon\varepsilon}}{1 - \beta_1} \quad (5.9)$$

En utilisant le fait que $\text{plim } m_{I\varepsilon} = 0$ et $\text{plim } m_{\varepsilon\varepsilon} = \sigma_\varepsilon^2$, nous obtenons

$$\text{plim } \hat{\beta}_1 = \beta_1 + \text{plim} \left(\frac{m_{R\varepsilon}}{m_{RR}} \right) = \beta_1 + \frac{\sigma_\varepsilon^2 (1 - \beta_1)}{\text{plim } m_{II} + \sigma_\varepsilon^2}$$

ce qui montre que $\hat{\beta}_1$ surestime β_1 si $0 \leq \beta_1 \leq 1$.

Exemple 2 : Considérons un modèle simple d'offre et de demande

$$Q_t^d = \beta_0 + \beta_1 P_t + \varepsilon_{t1} \quad (5.10)$$

$$Q_t^o = \alpha_0 + \alpha_1 P_t + \varepsilon_{t2} \quad (5.11)$$

$$Q_t^d = Q_t^o = Q_t \quad t = 1, 2, \dots, T \quad (5.12)$$

où Q_t^d et Q_t^o représentent respectivement la quantité demandée et la quantité offerte de marchandise.

En remplaçant la condition d'équilibre (5.12) dans (5.10) et (5.11), nous obtenons

$$Q_t = \beta_0 + \beta_1 P_t + \varepsilon_{t1} \quad (5.13)$$

$$Q_t = \alpha_0 + \alpha_1 P_t + \varepsilon_{t2} \quad t = 1, 2, \dots, T \quad (5.14)$$

Pour l'équation de la demande (5.13), le signe de β_1 devrait être négatif, tandis que pour l'équation de l'offre (5.14), le signe de α_1 devrait être positif. Cependant, nous n'observons qu'une seule paire d'équilibre (Q_t, P_t) et celles-ci ne sont pas appelées quantités et prix de demande ou d'offre. Lorsque nous exécutons la régression MCO de Q_t sur P_t , nous ne savons pas ce que nous estimons, la demande ou l'offre ? En fait, toute combinaison linéaire de (5.13) et (5.14) ressemble exactement à (5.13) ou (5.14). Il y aura une constante, un prix et un terme d'erreur. Étant donné que la demande ou l'offre ne peut être distinguée de cette forme "hypride", nous avons ce que l'on appelle un *problème d'identification*. Si l'équation de la demande (ou l'équation de l'offre) semblait différente de cette forme "hypride", alors cette équation particulière serait identifiée. Le problème d'identification sera traité dans la section suivante. Nous examinons pour l'instant les propriétés des estimations MCO de

5 Modèles à Équations Simultanées

l'équation de la demande. Il est bien connu que

$$\hat{\beta}_1 = \frac{\sum_{t=1}^T (Q_t - \bar{Q})(P_t - \bar{P})}{\sum_{t=1}^T (P_t - \bar{P})^2} = \beta + \frac{\sum_{t=1}^T (P_t - \bar{P})(\varepsilon_{t1} - \bar{\varepsilon}_1)}{\sum_{t=1}^T (P_t - \bar{P})^2} \quad (5.15)$$

Cet estimateur est sans biais selon que le dernier terme dans (5.15) a une espérance nulle. Afin de trouver cette espérance, nous résolvons les équations structurelles dans (5.13) et (5.14) pour Q_t et P_t

$$Q_t = \frac{\beta_0\alpha_1 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1\varepsilon_{t1} - \beta_1\varepsilon_{t2}}{\alpha_1 - \beta_1} \quad (5.16)$$

$$P_t = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\varepsilon_{t1} - \varepsilon_{t2}}{\alpha_1 - \beta_1} \quad (5.17)$$

Les équations (5.16) et (5.17) sont connus comme les équations de forme réduite. Noter que Q_t et P_t sont fonction des deux erreurs ε_1 et ε_2 . Par conséquent, P_t est corrélé avec ε_{t1} . En effet,

$$P_t - \bar{P} = \frac{\varepsilon_{t1} - \bar{\varepsilon}_1}{\alpha_1 - \beta_1} - \frac{\varepsilon_{t2} - \bar{\varepsilon}_2}{\alpha_1 - \beta_1} \quad (5.18)$$

et

$$\text{plim} \sum_{t=1}^T (P_t - \bar{P})(\varepsilon_{t1} - \bar{\varepsilon}_1) / T = \frac{\sigma_{11} - \sigma_{12}}{\alpha_1 - \beta_1} \quad (5.19)$$

$$\text{plim} \sum_{t=1}^T \frac{(P_t - \bar{P})^2}{T} = \frac{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}{(\alpha_1 - \beta_1)^2} \quad (5.20)$$

où $\sigma_{ij} = \text{cov}(\varepsilon_{ti}, \varepsilon_{tj})$ pour $i, j = 1, 2$; et $t = 1, \dots, T$. Par conséquent, à partir de (5.15)

$$\text{plim} \hat{\beta}_1 = \beta_1 + \frac{(\sigma_{11} - \sigma_{12})(\alpha_1 - \beta_1)}{(\sigma_{11} + \sigma_{22} - 2\sigma_{12})} \quad (5.21)$$

et le dernier terme n'est pas nécessairement nul, ce qui implique que $\hat{\beta}_1$ n'est pas un estimateur cohérent pour β_1 . De même, on peut montrer que l'estimateur des MCO pour α_1 n'est pas cohérent. Ce *biais simultané* est à nouveau dû à la corrélation de la variable de droite (prix) avec le terme d'erreur ε_1 . Cette corrélation pourrait être due au fait que P_t est

5 Modèles à Équations Simultanées

une fonction de ε_{t2} , à partir de (5.17), aussi ε_{t2} et ε_{t1} sont corrélés, ce qui rend P_t corrélé avec ε_{t1} . Alternativement, P_t est une fonction de Q_t à partir de (5.13) ou (5.14), et Q_t est une fonction de ε_{t1} à partit de (5.13), faisant de P_t une fonction de ε_{t1} . Intuitivement, si un choc de demande (c'est-à-dire un changement de ε_{t1}) déplace la courbe de demande, la nouvelle intersection de la demande et de l'offre détermine un nouveau prix et quantité d'équilibre. Ce nouveau prix est donc affecté par la variation de ε_{t1} et il est corrélé avec lui.

En général, chaque fois qu'une variable de droite est corrélée avec le terme d'erreur, les estimations des MCO sont biaisées et incohérentes. Nous appelons cela un *problème d'endogénéité*.

5.2 Le Problème d'Identification

On va considerer le même modèle précédent d'offre et de demande, cette fois comme suit :

$$Q_t^d = \beta_0 + \beta_1 P_t + \beta_2 Y_t + \varepsilon_{t1} \quad (5.22)$$

$$Q_t^o = \alpha_0 + \alpha_1 P_t + \varepsilon_{t2} \quad (5.23)$$

$$Q_t^d = Q_t^o = Q_t \quad t = 1, 2, \dots, T \quad (5.24)$$

À partir des équations (5.22) et (5.23), on résout pour P_t pour obtenir :

$$P_t = \frac{\beta_0 - \alpha_0}{\beta_1 - \alpha_1} + \frac{\beta_2}{\beta_1 - \alpha_1} Y_t + \frac{\varepsilon_{t1} - \varepsilon_{t2}}{\beta_1 - \alpha_1} \quad (5.25)$$

qui peut être réécrite sous forme :

$$P_t = \pi_0 + \pi_1 Y_t + \nu_{t1} \quad (5.26)$$

en substituant l'équation (5.26) dans l'équation (5.23), nous obtenons :

$$\begin{aligned} Q &= \alpha_0 + \alpha_1 (\pi_0 + \pi_1 Y_t + \nu_{t1}) + \varepsilon_{t2} \\ &= \alpha_0 + \alpha_1 \pi_0 + \alpha_1 \pi_1 Y_t + \alpha_1 \nu_{t1} + \varepsilon_{t2} \\ &= \pi_2 + \pi_3 Y_t + \nu_{t2} \end{aligned} \quad (5.27)$$

Maintenant, les équations (5.23) et (5.27) spécifient chacune des variables endogènes en termes uniquement des variables exogènes, des paramètres du modèle et des termes d'erreur stochastique. Ces deux équations sont appelées *équations de forme réduite* et les π 's sont appelées

5 Modèles à Équations Simultanées

paramètres de forme réduite. En général, des équations de forme réduites peuvent être obtenues en résolvant pour chacune des variables endogènes en termes des variables exogènes, des paramètres inconnus et des termes d'erreur.

Les équations de forme réduites n'expriment les variables endogènes qu'en fonction des variables exogènes. Par conséquent, il est possible d'appliquer MCO à ces équations pour obtenir des estimations cohérentes et efficaces des paramètres de forme réduite (les π 's).

La question ici est de savoir si nous pouvons obtenir des estimations cohérentes (les β 's et les α 's) en revenant en arrière et en résolvant ces paramètres. La réponse est qu'il y a trois situations possibles :

- (1) il n'est pas possible de revenir de la forme réduite à la forme structurelle ;
- (2) il est possible de revenir en arrière d'une manière unique ; ou
- (3) il y a plus d'une façon de revenir en arrière.

Ce problème de pouvoir (ou de ne pas pouvoir) revenir en arrière et de déterminer des estimations des paramètres structurels à partir d'estimateurs des coefficients de forme réduite est appelé problème d'identification.

La première situation (impossible de revenir en arrière) est appelée *sous - identification*, la deuxième situation (le cas unique) est appelée *identification exacte* et la troisième situation (où il y a plus d'une voie) est appelée *sur-identification*.

Conditions d'Identification

Il y a deux conditions requises pour qu'une équation soit identifiée : la *condition d'ordre* et la *condition de rang*. Les deux conditions sont d'abord décrites, puis des exemples sont donnés pour illustrer leur utilisation.

la Condition d'Ordre

Soit D le nombre de variables endogènes dans le système, et M le nombre de variables manquantes dans l'équation considérée (celles-ci peuvent être endogènes, exogènes ou variables endogènes retardées). Ensuite, la condition d'ordre stipule que :

- (a) si $M < D - 1$; l'équation est sous-identifiée ;
- (b) si $M = D - 1$; l'équation est exactement identifiée ; et
- (c) si $M > D - 1$; l'équation est sur-identifiée.

5 Modèles à Équations Simultanées

La condition d'ordre est nécessaire mais pas suffisante. Nous entendons par là que si cette condition n'est pas vérifiée, alors l'équation n'est pas identifiée, mais si cette condition est vérifiée, nous ne pouvons pas être certains qu'elle est identifiée, nous devons donc encore utiliser la condition de rang pour conclure.

la Condition de Rang

Pour la condition de rang, nous devons d'abord construire un tableau avec une colonne pour chaque variable et une ligne pour chaque équation. Pour chaque équation, mettez un \checkmark dans la colonne si la variable qui correspond à cette colonne est incluse dans l'équation, sinon mettez un 0. Cela donne un tableau de \checkmark et 0 pour chaque équation. Ensuite, pour une équation particulière :

- (a) supprimer la ligne de l'équation sujet de l'examen ;
- (b) écrire les éléments restants de chaque colonne pour laquelle il y a un zéro dans l'équation examinée ; et
- (c) En considérant le tableau résultant : s'il y a au moins $(D - 1)$ lignes et colonnes qui ne sont pas toutes des zéros, alors l'équation est identifiée ; sinon, elle n'est pas identifiée.

La condition de rang est nécessaire et suffisante, mais la condition d'ordre est nécessaire pour indiquer si l'équation est exactement identifiée ou sur-identifiée.

Exemples de Procédure d'Identification

Exemple 1 : Modèle de l'Offre et de la Demande

Considérons le modèle de l'offre et de la demande décrit dans les équations (5.22), (5.23) et (5.24). Construire d'abord un tableau avec une colonne pour chaque variable et une ligne pour chacune des trois équations :

	Q^d	Q^o	P	Y
Équation 1	\checkmark	0	\checkmark	\checkmark
Équation 2	0	\checkmark	\checkmark	0
Équation 3	\checkmark	\checkmark	0	0

Ici, nous avons trois variables endogènes (Q^d , Q^o et P), donc $D = 3$ et $D - 1 = 2$.

Considérer maintenant la condition d'ordre. Pour la fonction de demande, le nombre de variables exclues est 1, donc $M = 1$, et parce que

5 Modèles à Équations Simultanées

$M < D - 1$, la fonction de demande n'est pas identifiée. Pour la fonction de l'offre, $M = 1$ et parce que $M = D - 1$, la fonction de l'offre est exactement identifiée.

En procédant à la condition de rang, nous devons vérifier uniquement la fonction d'offre (car nous avons vu que la demande n'est pas identifiée). Le tableau résultant (après avoir supprimé les colonnes Q^o et P et la ligne de l'Équation 2) sera donné comme suit :

	Q^d	Q^o	P	Y		Q^d	Y
Équation 1	✓	0	✓	✓			
Équation 2	0	✓	✓	0	Équation 1	✓	✓
Équation 3	✓	✓	0	0	Équation 3	✓	0

La question est, y a-t-il au moins $D - 1 = 2$ lignes et colonnes qui ne sont pas toutes des zéros ? La réponse est "oui", et donc la condition de rang est satisfaite et la fonction d'offre est en effet exactement identifiée.

Exemple 2 : Modèle Macroéconomique

Considérons le modèle macroéconomique simple pour une économie décrit par les équations ci-dessous :

$$C_t = \beta_0 + \beta_1 Y_t \quad (5.28)$$

$$I_t = \alpha_0 + \alpha_1 Y_t + \alpha_2 R_t \quad (5.29)$$

$$Y_t = C_t + I_t + G_t \quad (5.30)$$

où C_t désigne la consommation, Y_t est le PIB, I_t représente l'investissement, R_t désigne le taux d'intérêt et G_t représente les dépenses publiques. Ici, C_t , I_t et Y_t sont des variables endogènes, tandis que R_t et G_t sont des variables exogènes. Tout d'abord, créer un tableau avec cinq colonnes (une pour chaque variable) et trois lignes (une pour chaque équation) :

	C	Y	I	R	G
Équation 1	✓	✓	0	0	0
Équation 2	0	✓	✓	✓	0
Équation 3	✓	✓	✓	0	✓

Le tableau montre que pour l'équation 1, $M = 3$ (I , R et G sont exclus) tandis que $D = 3$ et donc, $M > D - 1$, donc la fonction de consommation semble être sur-identifiée. De même, pour l'équation 2, $M = D - 1$, elle apparaît donc exactement identifiée.

En utilisant la condition de rang pour la fonction de consommation, nous avons (après exclusion des colonnes C et Y et de la ligne de l'équation 1) le tableau suivant :

5 Modèles à Équations Simultanées

	<i>I</i>	<i>R</i>	<i>G</i>
Équation 2	✓	✓	0
Équation 3	✓	0	✓

Donc, il y a $D - 1 = 2$ lignes et colonnes sans éléments entièrement nuls et donc elle est sur-identifiée. Pour la fonction d'investissement (après exclusion des colonnes *I*, *Y* et *R* et la ligne de l'équation 2), nous avons :

	<i>C</i>	<i>G</i>
Équation 1	✓	0
Équation 3	✓	✓

Encore une fois, il y a $D - 1 = 2$ lignes et colonnes sans éléments entièrement nuls, de sorte que la condition de rang est à nouveau satisfaite et nous concluons que la fonction d'investissement est effectivement identifiée.

5.3 Estimation des Modèles à Équations Simultanées

La question de l'identification est étroitement liée au problème de l'estimation des paramètres structurels dans un modèle à équations simultanées. Ainsi, lorsqu'une équation n'est pas identifiée, une telle estimation n'est pas possible. Cependant, dans les cas d'identification exacte ou de sur-identification, il existe des procédures qui nous permettent d'obtenir des estimations des paramètres structurels. Ces procédures sont différentes des MCO simples et ce afin d'éviter le biais de simultanéité présenté précédemment.

En général, en cas d'identification exacte, la méthode appropriée est la méthode dite des Moindres Carrés Indirects (MCI), tandis que dans les cas d'équations sur-identifiées, la méthode des Doubles Moindres Carrés (DMC) est celle qui est la plus utilisée.

5.3.1 Estimation d'une Équation Identifiée avec Précision : La Méthode MCI

Cette méthode ne peut être utilisée que lorsque les équations du modèle d'équations simultanées sont identifiées avec précision. La procédure MCI comprend ces trois étapes :

Etape 1 Trouver les équations de forme réduite ;

Etape 2 Estimer les paramètres de forme réduite en appliquant MCO simple aux équations de forme réduite ; et

Etape 3 Obtenir des estimations uniques des paramètres structurels à partir des estimations des paramètres de l'équation de forme réduite à l'étape 2.

Les estimations MCO des paramètres de forme réduite ne sont pas biaisées, mais lorsqu'elles sont transformées, les estimations des paramètres structurels qu'elles fournissent sont cohérentes. Dans les rares cas où toutes les équations de forme structurelle sont identifiées avec précision, MCI fournit des estimations cohérentes et asymptotiquement normales.

Cependant, la méthode MCI n'est pas couramment utilisée, pour deux raisons :

- (1) La plupart des modèles à équations simultanées ont tendance à être sur-identifiés ; et
- (2) Si le système a plusieurs équations, la résolution de la forme réduite puis de la forme structurelle peut être très fastidieuse. Une alternative est la méthode MCDE.

5.3.2 Estimation d'une Équation Sur-Identifiée : La Méthode DMC

L'idée de base de la méthode DMC est de remplacer le régresseur endogène stochastique (qui est corrélé avec le terme d'erreur et provoque le biais) par un régresseur non stochastique et par conséquent indépendant du terme d'erreur. Cela implique les deux étapes suivantes (d'où les *Doubles Moindres Carrés*) :

Etape 1 Régresser chaque variable endogène qui est également un régresseur, sur toutes les variables endogènes et endogènes retardées dans l'ensemble du système en utilisant MCO simple (cela équivaut à estimer les équations de forme réduite) et obtenir les valeurs ajustées des variables endogènes de ces régressions (\hat{Y}).

Etape 2 Utiliser les valeurs ajustées de l'étape 1 comme proxys ou instruments pour les régresseurs endogènes dans les équations originales (forme structurelle).

les R^2 's des équations estimées dans l'étape 1 devraient être relativement élevés. Il s'agit de garantir que \hat{Y} et Y sont fortement corrélés et que, par conséquent, \hat{Y} est un bon instrument pour Y . Un avantage de la méthode DMC est que, pour les équations qui sont exactement identifiées, elle donnera des estimations identiques à celles obtenues à partir de MCI, tandis que la méthode DMC est également appropriée même pour les équations sur-identifiées.

5 Modèles à Équations Simultanées

Exemple : le Modèle IS–LM

On considère le modèle IS – LM suivant :

$$R_t = \beta_0 + \beta_1 M_t + \beta_2 Y_t + \beta_3 M_{t-1} + \varepsilon_{1t} \quad (5.31)$$

$$Y_t = \alpha_0 + \alpha_1 R_t + \alpha_2 I_t + \varepsilon_{2t} \quad (5.32)$$

où R désigne le taux d'intérêt, M désigne la masse monétaire, Y désigne le PIB et I désigne les dépenses d'investissement. Dans ce modèle, R et Y sont les variables endogènes et M et I représentent les variables exogènes. On peut montrer que l'équation (5.31) est exactement identifiée et l'équation (5.32) est sur-identifiée.

Nous voulons estimer le modèle et, comme la deuxième équation est sur-identifiée, nous devrons utiliser la méthode DMC. Les données de cet exemple se trouvent dans le tableau (5.1).

Estimation des Équations Simultanées avec Eviews

On considère les données de l'exemple précédent. Pour estimer une équation à l'aide de DMC, vous avez la première option, celle d'accéder à **Quick / Estimate Equation** et, dans la fenêtre **Equation Specification**, changer la méthode des MCO par défaut **LS - Least Squares (NLS and ARMA)** à **TSLS - Two-stage Least Squares (TSNLS and ARMA)** et puis spécifier l'équation à estimer dans la première case et la liste des instruments dans la seconde. La deuxième option est de taper la commande suivante dans EVViews :

```
tsls r c m y m(-1) @ c m i m(-1)
```

La commande écrite avant le symbole @ représente l'équation à estimer, et celle écrite après le symbole @ représente les noms des variables qui doivent être utilisées comme instruments. Les résultats de cette estimation sont donnés dans le tableau (5.2).

L'équation du taux d'intérêt peut être considérée comme la relation LM. Le coefficient de Y est très petit et positif (mais non significatif), suggérant que la fonction LM est très plate, tandis que les augmentations de la masse monétaire réduisent le taux d'intérêt. De plus, R^2 est très petit, ce qui suggère qu'il manque des variables dans l'équation.

Pour estimer la deuxième équation (considérée comme la relation IS), on tape la commande suivante :

```
TSLS y c r i @ c m i m(-1)
```

5 Modèles à Équations Simultanées

Les résultats sont présentés dans le tableau (5.3).

En interprétant ces résultats, nous pouvons voir que le revenu et le taux d'intérêt sont liés négativement, comme le suggère la théorie, et le revenu est assez sensible aux changements du taux d'intérêt. En outre, un changement dans les investissements entraînerait un déplacement de la fonction vers la droite, comme le suggère la théorie. Le R^2 de cette spécification est assez élevé.

Estimation des Équations Simultanées avec Stata

Dans Stata, pour estimer un modèle à équations simultanées, la commande est la suivante :

```
reg3 (1ère equation) (2ème equation) , 2sls
```

où, dans les premières parenthèses, nous mettons la première équation à estimer et, la deuxième équation dans les secondes parenthèses. Le **2sls** dans la ligne de commande indique que la méthode d'estimation devrait être la méthode des doubles moindres carrés. Par conséquent, pour l'exemple précédent de IS-LM, la commande est :

```
reg3 (r = m y L.m) (y = r i) , 2sls
```

Les résultats de cette commande sont présentés dans le tableau (5.4) et sont très similaires à celle obtenus avec EViews.

5 Modèles à Équations Simultanées

	I	M	R	Y
1	20239	9276	9.49	125023
2	19894	9571	9.38	126308
3	19580	9836	11.91	126959
4	20148	10136	11.78	126358
5	20368	10420	13.79	131564
6	20689	10782	13.82	128596
7	21102	11013	16.49	129565
8	20402	11151	16.97	128696
9	19722	11376	16.32	126157
10	19349	11566	14.86	125671
11	18925	11649	13.45	124227
12	17631	11946	11.87	123666
13	17927	12018	12.24	123750
14	18040	11893	15.72	125537
15	17853	11925	15.35	125361
16	17943	11911	12.87	125499
17	18834	11993	12.62	126893
18	19396	12144	9.92	127106
19	19493	12302	9.96	127795
20	19801	12550	10.49	130397
21	19405	12702	9.51	130602
22	19599	12850	9.17	131897
23	20684	13038	9.04	133389
24	21363	13230	8.56	134716
25	21643	13401	9.06	134283
26	21807	13591	10.24	134273
27	22099	13746	9.33	135737

TABLE 5.1 – **Données Pour le Modèle IS-LM**

5 Modèles à Équations Simultanées

Dependent Variable : R
 Method : Two-Stage Least Squares
 Sample (adjusted) : 2 27
 Included observations : 26 after adjustments
 Instrument specification : C M I M(-1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	25.59570	21.53009	1.188834	0.2472
M	-0.013355	0.006166	-2.166028	0.0414
Y	6.46E-05	0.000215	0.299801	0.7671
M(-1)	0.011689	0.005697	2.051649	0.0523
R-squared	0.378091	Mean dependent var	12.10462	
Adjusted R-squared	0.293286	S.D. dependent var	2.665914	
S.E. of regression	2.241135	Sum squared resid	110.4991	
F-statistic	4.669841	Durbin-Watson stat	0.543456	
Prob(F-statistic)	0.011343	Second-Stage SSR	107.3120	
J-statistic	1.99E-34	Instrument rank	4	

TABLE 5.2 – Estimation DMC de l’Équation R(LM)

Dependent Variable : Y
 Method : Two-Stage Least Squares
 Sample (adjusted) : 2 27
 Included observations : 26 after adjustments
 Instrument specification : C M I M(-1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	99402.89	9308.308	10.67894	0.0000
R	-747.5648	273.7889	-2.730443	0.0119
I	1.938498	0.362461	5.348149	0.0000
R-squared	0.730834	Mean dependent var	128653.9	
Adjusted R-squared	0.707428	S.D. dependent var	3651.339	
S.E. of regression	1975.007	Sum squared resid	89714994	
F-statistic	34.45775	Durbin-Watson stat	1.285755	
Prob(F-statistic)	0.000000	Second-Stage SSR	64491551	
J-statistic	5.051549	Instrument rank	4	
Prob(J-statistic)	0.024604			

TABLE 5.3 – Estimation DMC de l’Équation Y(IS)

5 Modèles à Équations Simultanées

Two-stage least-squares regression

Equation	Obs	Parms	RMSE	"R-sq"	F-Stat	P
	Coef.	Std. Err.	t	P> t 	[95% Conf.	Interval]
r	26	3	2.241135	0.3781	4.67	0.0063
y	26	2	1975.007	0.7308	34.46	0.0000
r						
m	-0.0133551	0.0061657	-2.17	0.036	-0.0257735	-0.0009367
y	0.0000646	0.0002153	0.30	0.766	-0.0003691	0.0004982
m						
L1.	0.0116888	0.0056973	2.05	0.046	0.0002139	0.0231637
_cons	25.5957	21.53009	1.19	0.241	-17.76813	68.95953
y						
r	-747.5648	273.7889	-2.73	0.009	-1299.004	-196.1257
i	1.938498	0.3624614	5.35	0.000	1.208463	2.668532
_cons	99402.89	9308.308	10.68	0.000	80655	118150.8

Endogenous variables : r y

Exogenous variables : m L.m i

TABLE 5.4 – Résultats de l'Estimation DMC avec Stata

Annexes

A1. La Loi Normale Centrée Réduite

Exemple :

$$P(0 \leq Z \leq 1.96) = 0.4750$$

$$P(Z \geq 1.96) = 0.5 - 0.4750 = 0.025$$

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.004	0.008	0.012	0.016	0.0199	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	0.091	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	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	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	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	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	0.334	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	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	0.437	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	0.4454	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	0.475	0.4756	0.4761	0.4767
2	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	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
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	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	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	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	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	0.498	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	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499

A2. Table de la Loi de Fisher-Snedecor $\alpha = 5\%$

ddl1 ddl2	1	2	3	4	5	10	15	20	25	30
1	161.45	199.50	215.71	224.58	230.16	241.88	245.95	248.01	249.26	250.09
2	18.513	19.000	19.164	19.247	19.296	19.396	19.429	19.446	19.456	19.462
3	10.128	9.552	9.277	9.117	9.013	8.786	8.703	8.660	8.634	8.617
4	7.709	6.944	6.591	6.388	6.256	5.964	5.858	5.803	5.769	5.746
5	6.608	5.786	5.409	5.192	5.050	4.735	4.619	4.558	4.521	4.496
6	5.987	5.143	4.757	4.534	4.387	4.060	3.938	3.874	3.835	3.808
7	5.591	4.737	4.347	4.120	3.972	3.637	3.511	3.445	3.404	3.376
8	5.318	4.459	4.066	3.838	3.687	3.347	3.218	3.150	3.108	3.079
9	5.117	4.256	3.863	3.633	3.482	3.137	3.006	2.936	2.893	2.864
10	4.965	4.103	3.708	3.478	3.326	2.978	2.845	2.774	2.730	2.700
11	4.844	3.982	3.587	3.357	3.204	2.854	2.719	2.646	2.601	2.570
12	4.747	3.885	3.490	3.259	3.106	2.753	2.617	2.544	2.498	2.466
13	4.667	3.806	3.411	3.179	3.025	2.671	2.533	2.459	2.412	2.380
14	4.600	3.739	3.344	3.112	2.958	2.602	2.463	2.388	2.341	2.308
15	4.543	3.682	3.287	3.056	2.901	2.544	2.403	2.328	2.280	2.247
16	4.494	3.634	3.239	3.007	2.852	2.494	2.352	2.276	2.227	2.194
17	4.451	3.592	3.197	2.965	2.810	2.450	2.308	2.230	2.181	2.148
18	4.414	3.555	3.160	2.928	2.773	2.412	2.269	2.191	2.141	2.107
19	4.381	3.522	3.127	2.895	2.740	2.378	2.234	2.155	2.106	2.071
20	4.351	3.493	3.098	2.866	2.711	2.348	2.203	2.124	2.074	2.039
21	4.325	3.467	3.072	2.840	2.685	2.321	2.176	2.096	2.045	2.010
22	4.301	3.443	3.049	2.817	2.661	2.297	2.151	2.071	2.020	1.984
23	4.279	3.422	3.028	2.796	2.640	2.275	2.128	2.048	1.996	1.961
24	4.260	3.403	3.009	2.776	2.621	2.255	2.108	2.027	1.975	1.939
25	4.242	3.385	2.991	2.759	2.603	2.236	2.089	2.007	1.955	1.919
26	4.225	3.369	2.975	2.743	2.587	2.220	2.072	1.990	1.938	1.901
27	4.210	3.354	2.960	2.728	2.572	2.204	2.056	1.974	1.921	1.884
28	4.196	3.340	2.947	2.714	2.558	2.190	2.041	1.959	1.906	1.869
29	4.183	3.328	2.934	2.701	2.545	2.177	2.027	1.945	1.891	1.854
30	4.171	3.316	2.922	2.690	2.534	2.165	2.015	1.932	1.878	1.841
40	4.085	3.232	2.839	2.606	2.449	2.077	1.924	1.839	1.783	1.744
60	4.001	3.150	2.758	2.525	2.368	1.993	1.836	1.748	1.690	1.649
100	3.936	3.087	2.696	2.463	2.305	1.927	1.768	1.676	1.616	1.573

A2. Table de la Loi de Fisher-Snedecor $\alpha = 1\%$ (suite)

ddl1 ddl2	1	2	3	4	5	10	15	20	25	30
1	4052.18	4999.5	5403.35	5624.58	5763.65	6055.85	6157.28	6208.73	6239.82	6260.65
2	98.503	99.000	99.166	99.249	99.299	99.399	99.433	99.449	99.459	99.466
3	34.116	30.817	29.457	28.710	28.237	27.229	26.872	26.690	26.579	26.505
4	21.198	18.000	16.694	15.977	15.522	14.546	14.198	14.020	13.911	13.838
5	16.258	13.274	12.060	11.392	10.967	10.051	9.722	9.553	9.449	9.379
6	13.745	10.925	9.780	9.148	8.746	7.874	7.559	7.396	7.296	7.229
7	12.246	9.547	8.451	7.847	7.460	6.620	6.314	6.155	6.058	5.992
8	11.259	8.649	7.591	7.006	6.632	5.814	5.515	5.359	5.263	5.198
9	10.561	8.022	6.992	6.422	6.057	5.257	4.962	4.808	4.713	4.649
10	10.044	7.559	6.552	5.994	5.636	4.849	4.558	4.405	4.311	4.247
11	9.646	7.206	6.217	5.668	5.316	4.539	4.251	4.099	4.005	3.941
12	9.330	6.927	5.953	5.412	5.064	4.296	4.010	3.858	3.765	3.701
13	9.074	6.701	5.739	5.205	4.862	4.100	3.815	3.665	3.571	3.507
14	8.862	6.515	5.564	5.035	4.695	3.939	3.656	3.505	3.412	3.348
15	8.683	6.359	5.417	4.893	4.556	3.805	3.522	3.372	3.278	3.214
16	8.531	6.226	5.292	4.773	4.437	3.691	3.409	3.259	3.165	3.101
17	8.400	6.112	5.185	4.669	4.336	3.593	3.312	3.162	3.068	3.003
18	8.285	6.013	5.092	4.579	4.248	3.508	3.227	3.077	2.983	2.919
19	8.185	5.926	5.010	4.500	4.171	3.434	3.153	3.003	2.909	2.844
20	8.096	5.849	4.938	4.431	4.103	3.368	3.088	2.938	2.843	2.778
21	8.017	5.780	4.874	4.369	4.042	3.310	3.030	2.880	2.785	2.720
22	7.945	5.719	4.817	4.313	3.988	3.258	2.978	2.827	2.733	2.667
23	7.881	5.664	4.765	4.264	3.939	3.211	2.931	2.781	2.686	2.620
24	7.823	5.614	4.718	4.218	3.895	3.168	2.889	2.738	2.643	2.577
25	7.770	5.568	4.675	4.177	3.855	3.129	2.850	2.699	2.604	2.538
26	7.721	5.526	4.637	4.140	3.818	3.094	2.815	2.664	2.569	2.503
27	7.677	5.488	4.601	4.106	3.785	3.062	2.783	2.632	2.536	2.470
28	7.636	5.453	4.568	4.074	3.754	3.032	2.753	2.602	2.506	2.440
29	7.598	5.420	4.538	4.045	3.725	3.005	2.726	2.574	2.478	2.412
30	7.562	5.390	4.510	4.018	3.699	2.979	2.700	2.549	2.453	2.386
40	7.314	5.179	4.313	3.828	3.514	2.801	2.522	2.369	2.271	2.203
60	7.077	4.977	4.126	3.649	3.339	2.632	2.352	2.198	2.098	2.028
100	6.895	4.824	3.984	3.513	3.206	2.503	2.223	2.067	1.965	1.893

A3. Table de La loi du Chi-Deux $\chi^2_{\alpha;n}$

n	$\alpha = 0.9$	0.8	0.7	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01
1	0.02	0.06	0.15	1.64	0.71	1.07	1.64	2.71	3.84	5.41	6.63
2	0.21	0.45	0.71	3.22	1.83	2.41	3.22	4.61	5.99	7.82	9.21
3	0.58	1.01	1.42	4.64	2.95	3.66	4.64	6.25	7.81	9.84	11.34
4	1.06	1.65	2.19	5.99	4.04	4.88	5.99	7.78	9.49	11.67	13.28
5	1.61	2.34	3.00	7.29	5.13	6.06	7.29	9.24	11.07	13.39	15.09
6	2.20	3.07	3.83	8.56	6.21	7.23	8.56	10.64	12.59	15.03	16.81
7	2.83	3.82	4.67	9.80	7.28	8.38	9.80	12.02	14.07	16.62	18.48
8	3.49	4.59	5.53	11.03	8.35	9.52	11.03	13.36	15.51	18.17	20.09
9	4.17	5.38	6.39	12.24	9.41	10.66	12.24	14.68	16.92	19.68	21.67
10	4.87	6.18	7.27	13.44	10.47	11.78	13.44	15.99	18.31	21.16	23.21
11	5.58	6.99	8.15	14.63	11.53	12.90	14.63	17.28	19.68	22.62	24.72
12	6.30	7.81	9.03	15.81	12.58	14.01	15.81	18.55	21.03	24.05	26.22
13	7.04	8.63	9.93	16.98	13.64	15.12	16.98	19.81	22.36	25.47	27.69
14	7.79	9.47	10.82	18.15	14.69	16.22	18.15	21.06	23.68	26.87	29.14
15	8.55	10.31	11.72	19.31	15.73	17.32	19.31	22.31	25.00	28.26	30.58
16	9.31	11.15	12.62	20.47	16.78	18.42	20.47	23.54	26.30	29.63	32.00
17	10.09	12.00	13.53	21.61	17.82	19.51	21.61	24.77	27.59	31.00	33.41
18	10.86	12.86	14.44	22.76	18.87	20.60	22.76	25.99	28.87	32.35	34.81
19	11.65	13.72	15.35	23.90	19.91	21.69	23.90	27.20	30.14	33.69	36.19
20	12.44	14.58	16.27	25.04	20.95	22.77	25.04	28.41	31.41	35.02	37.57
21	13.24	15.44	17.18	26.17	21.99	23.86	26.17	29.62	32.67	36.34	38.93
22	14.04	16.31	18.10	27.30	23.03	24.94	27.30	30.81	33.92	37.66	40.29
23	14.85	17.19	19.02	28.43	24.07	26.02	28.43	32.01	35.17	38.97	41.64
24	15.66	18.06	19.94	29.55	25.11	27.10	29.55	33.20	36.42	40.27	42.98
25	16.47	18.94	20.87	30.68	26.14	28.17	30.68	34.38	37.65	41.57	44.31
26	17.29	19.82	21.79	31.79	27.18	29.25	31.79	35.56	38.89	42.86	45.64
27	18.11	20.70	22.72	32.91	28.21	30.32	32.91	36.74	40.11	44.14	46.96
28	18.94	21.59	23.65	34.03	29.25	31.39	34.03	37.92	41.34	45.42	48.28
29	19.77	22.48	24.58	35.14	30.28	32.46	35.14	39.09	42.56	46.69	49.59
30	20.60	23.36	25.51	36.25	31.32	33.53	36.25	40.26	43.77	47.96	50.89

A4. Table de La loi de Student avec n degrés de liberté, Quantiles d'ordre $1 - \alpha$

n	0.25	0.20	0.15	0.10	0.05	0.025	0.010	0.005	0.0025	0.0010	0.0005
1	1.00	1.38	1.96	3.08	6.31	12.71	31.82	63.66	127.32	318.31	636.62
2	0.816	1.061	2.282	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.924	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.778	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.699	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.650	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.617	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.592	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.574	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.559	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.548	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.538	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.530	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.523	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.517	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.512	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.508	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.504	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.500	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.497	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.494	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.492	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.489	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.487	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.485	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.483	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.482	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.480	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.479	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.477	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
35	0.682	0.852	1.472	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
40	0.681	0.851	1.468	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.462	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.458	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
100	0.677	0.845	1.451	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390

A5. Table de Durbin-Watson $\alpha = 5\%$

	1	2	3	4	5	6	7	8	9	10
	dL	dU								
6	0.61	1.4	—	—	—	—	—	—	—	—
7	0.7	1.356	0.467	1.896	—	—	—	—	—	—
8	0.763	1.332	0.559	1.777	0.367	2.287	—	—	—	—
9	0.824	1.32	0.629	1.699	0.455	2.128	0.296	2.588	—	—
10	0.879	1.32	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.38	2.506
13	1.01	1.34	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.39
14	1.045	1.35	0.905	1.551	0.767	1.779	0.632	2.03	0.505	2.296
15	1.077	1.361	0.946	1.543	0.814	1.75	0.685	1.977	0.562	2.22
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157
17	1.133	1.381	1.015	1.536	0.897	1.71	0.779	1.9	0.664	2.104
18	1.158	1.391	1.046	1.535	0.933	1.696	0.82	1.872	0.71	2.06
19	1.18	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023
20	1.201	1.411	1.1	1.537	0.998	1.676	0.894	1.828	0.792	1.991
21	1.221	1.42	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.94
23	1.257	1.437	1.168	1.543	1.078	1.66	0.986	1.785	0.895	1.92
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902
25	1.288	1.454	1.206	1.55	1.123	1.654	1.038	1.767	0.953	1.886
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873
27	1.316	1.469	1.24	1.556	1.162	1.651	1.084	1.753	1.004	1.861
28	1.328	1.476	1.255	1.56	1.181	1.65	1.104	1.747	1.028	1.85
29	1.341	1.483	1.27	1.563	1.198	1.65	1.124	1.743	1.05	1.841
30	1.352	1.489	1.284	1.567	1.214	1.65	1.143	1.739	1.071	1.833

A5. Table de Durbin-Watson $\alpha = 5\%$ (suite)

	1	2	3	4	5	6	7	8	9	10										
	dL	dU																		
31	1.363	1.496	1.297	1.57	1.229	1.65	1.16	1.735	1.09	1.825	1.02	1.92	0.95	2.018	0.879	2.12	0.81	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.65	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.73	1.127	1.813	1.061	1.9	0.994	1.991	0.927	2.085	0.861	2.181	0.796	2.281
34	1.393	1.514	1.333	1.58	1.271	1.652	1.208	1.728	1.144	1.808	1.079	1.891	1.015	1.978	0.95	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.16	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.876	1.053	1.957	0.991	2.041	0.93	2.127	0.868	2.216
37	1.419	1.53	1.364	1.59	1.307	1.655	1.249	1.723	1.19	1.795	1.131	1.87	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.197
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.97	2.098	0.912	2.18
39	1.435	1.54	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.99	2.085	0.932	2.164
40	1.442	1.544	1.391	1.6	1.338	1.659	1.285	1.721	1.23	1.786	1.175	1.854	1.12	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.43	1.615	1.383	1.666	1.336	1.72	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.93	1.156	1.986	1.11	2.044
55	1.528	1.601	1.49	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.17	2.01
60	1.549	1.616	1.514	1.652	1.48	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.85	1.298	1.894	1.26	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.37	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.838	1.369	1.874	1.337	1.91	1.305	1.948
75	1.598	1.652	1.571	1.68	1.543	1.709	1.515	1.739	1.487	1.77	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.56	1.715	1.534	1.743	1.507	1.772	1.48	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.6	1.696	1.575	1.721	1.55	1.747	1.525	1.774	1.5	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.42	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.78	1.55	1.803	1.528	1.826	1.506	1.85	1.484	1.874	1.462	1.898
150	1.72	1.747	1.706	1.76	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.846	1.608	1.862	1.593	1.877
200	1.758	1.779	1.748	1.789	1.738	1.799	1.728	1.809	1.718	1.82	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

A5. Table de Durbin-Watson $\alpha = 1\%$

	1	2	3	4	5	6	7	8	9	10
	dL	dU								
6	0.39	1.142	—	—	—	—	—	—	—	—
7	0.435	1.036	0.294	1.676	—	—	—	—	—	—
8	0.497	1.003	0.345	1.489	0.229	2.102	—	—	—	—
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	—	—
10	0.604	1.001	0.466	1.333	0.34	1.733	0.23	2.193	0.15	2.69
11	0.653	1.01	0.519	1.297	0.396	1.64	0.286	2.03	0.193	2.453
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.28
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.15
14	0.776	1.054	0.66	1.254	0.547	1.49	0.441	1.757	0.343	2.049
15	0.811	1.07	0.7	1.252	0.591	1.465	0.487	1.705	0.39	1.967
16	0.844	1.086	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803
19	0.928	1.133	0.835	1.264	0.742	1.416	0.65	1.583	0.561	1.767
20	0.952	1.147	0.862	1.27	0.774	1.41	0.684	1.567	0.598	1.736
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691
23	1.017	1.186	0.938	1.29	0.858	1.407	0.777	1.535	0.699	1.674
24	1.037	1.199	0.959	1.298	0.881	1.407	0.805	1.527	0.728	1.659
25	1.055	1.21	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645
26	1.072	1.222	1	1.311	0.928	1.41	0.855	1.517	0.782	1.635
27	1.088	1.232	1.019	1.318	0.948	1.413	0.878	1.514	0.808	1.625
28	1.104	1.244	1.036	1.325	0.969	1.414	0.901	1.512	0.832	1.618
29	1.119	1.254	1.053	1.332	0.988	1.418	0.921	1.511	0.855	1.611
30	1.134	1.264	1.07	1.339	1.006	1.421	0.941	1.51	0.877	1.606

A5. Table de Durbin-Watson $\alpha = 1\%$ (suite)

	1	2	3	4	5	6	7	8	9	10
	dL	dU								
31	1.147	1.274	1.085	1.345	1.022	1.425	0.96	1.509	0.897	1.601
32	1.16	1.283	1.1	1.351	1.039	1.428	0.978	1.509	0.917	1.597
33	1.171	1.291	1.114	1.358	1.055	1.432	0.995	1.51	0.935	1.594
34	1.184	1.298	1.128	1.364	1.07	1.436	1.012	1.511	0.954	1.591
35	1.195	1.307	1.141	1.37	1.085	1.439	1.028	1.512	0.971	1.589
36	1.205	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.987	1.587
37	1.217	1.322	1.164	1.383	1.112	1.446	1.058	1.514	1.004	1.585
38	1.227	1.33	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.584
39	1.237	1.337	1.187	1.392	1.137	1.452	1.085	1.517	1.033	1.583
40	1.246	1.344	1.197	1.398	1.149	1.456	1.098	1.518	1.047	1.583
45	1.288	1.376	1.245	1.424	1.201	1.474	1.156	1.528	1.111	1.583
50	1.324	1.403	1.285	1.445	1.245	1.491	1.206	1.537	1.164	1.587
55	1.356	1.428	1.32	1.466	1.284	1.505	1.246	1.548	1.209	1.592
60	1.382	1.449	1.351	1.484	1.317	1.52	1.283	1.559	1.248	1.598
65	1.407	1.467	1.377	1.5	1.346	1.534	1.314	1.568	1.283	1.604
70	1.429	1.485	1.4	1.514	1.372	1.546	1.343	1.577	1.313	1.611
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.586	1.34	1.617
80	1.465	1.514	1.44	1.541	1.416	1.568	1.39	1.595	1.364	1.624
85	1.481	1.529	1.458	1.553	1.434	1.577	1.411	1.603	1.386	1.63
90	1.496	1.541	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636
95	1.51	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.641
100	1.522	1.562	1.502	1.582	1.482	1.604	1.461	1.625	1.441	1.647
150	1.611	1.637	1.598	1.651	1.584	1.665	1.571	1.679	1.557	1.693
200	1.664	1.684	1.653	1.693	1.643	1.704	1.633	1.715	1.623	1.725

Bibliographie

- Aljandali, A. and Tatahi, M. (2018)** *Economic and Financial Modelling with EViews : A Guide for Students and Professionals*. Springer International Publishing.
- Baltagi, B.H. (1995)** *Econometric Analysis of Panel Data*. New York : John Wiley.
- Baltagi, B.H. and J.M. Griffin (1997)** ‘*Pooled Estimators vs their Heterogeneous Counterparts in the Context of Dynamic Demand for Gasoline*’, *Journal of Econometrics*, 77.
- Bourbonnais, R. and Terraiza, M. (2008)** *Analyse des séries temporelles : Applications à l'économie et à la gestion*. Dunod.
- Chow, G. (1960)** ‘*Tests of Equality between Sets of Coefficients in Two Linear Regressions*’, *Econometrica*, 28.
- Cochrane, D. and G. Orcutt (1949)** ‘*Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms*’, *Journal of the American Statistical Association*, 44.
- Damodar N. Gujarati, Dawn C. Porter (2008)** *Basic Econometrics*. New York : McGraw-Hill.
- Durbin, J. (1970)** ‘*Testing for Serial Correlation in Least Squares Regression – When Some of the Variables are Lagged Dependent Variables*’, *Econometrica*, 38.
- Durbin, J. and G. Watson (1950)** ‘*Testing for Serial Correlation in Least Squares Regression I*’, *Biometrika*, 37.
- Engle, R.F. (1982)** ‘*Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation*’, *Econometrica*, 50.
- Engle, R.F. (1995)** *ARCH Selected Readings (Advanced Texts in Econometrics)*. Oxford University Press.
- Glesjer, H. (1961)** ‘*A New Test for Multiplicative Heteroskedasticity*’, *Journal of the American Statistical Association*, 60.
- Godfrey, L.G. (1978)** ‘*Testing for Higher Order Serial Correlation in Regression Equations when the Regressions Contain Lagged*

- Dependent Variables*', *Econometrica*, 46.
- Goldfeld, S. and R. Quandt (1965)** 'Some Tests for Homoscedasticity', *Journal of the American Statistical Association*, 60.
- Gujarati, D.N. and Porter, D.C. (2009)** *Basic Econometrics*. 5th Edition, McGraw Hill Inc., New York.
- Harvey (1999)**, *The Econometric Analysis of Time Series*, 2nd edition, Cambridge, Mass., MIT Press.
- Koutsoyiannis, A. (2011)** *Theory of Econometrics*. Palgrave Macmillan Limited.
- Maddala, G.S and S. Wu (1999)** 'A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test', *Oxford Bulletin of Economics and Statistics*, special issue, 61.
- Maddala, G.S. (2001)** *Introduction to Econometrics*, 3rd ed. London : John Wiley.
- Racicot, F. E. et Théoret, R. (2001)**, *Traité d'économétrie financière, Modélisation Financière*. Presses de l'Université du Québec.
- Stock, J. H., & Watson, M. W. (2015)** *Introduction to Econometrics*, Third Update, Global Edition. Pearson Education Limited.

Index

- échantillonnage, 18, 54, 73, 122, 230, données centrées, 118, 218
240 données de panel, 19, 20, 22
économie mathématique, 11 données transversales, 17, 18, 20,
22, 247–249, 266, 274
ARCH, 298, 299 doubles moindres carrés, 330, 331,
ARMA, 252, 332 333
autocorrélation, 248, 249, 252–254, Durbin-Watson, 254, 258, 259, 343–
258, 259, 261, 265, 272, 297, 346
298, 306, 307, 309
biais simultané, 322, 325
causalité, 20, 21, 32, 57, 138, 321
ceteris paribus, 21, 22
Cochrane-Orcutt, 266, 267, 269, 271,
272, 307, 309, 313
coeffcient de corrélation, 52, 57, 58,
60, 66, 96, 97, 100–103, 127,
137–140, 168–170, 184, 216,
218, 219, 227, 234, 240, 242–
244, 252, 270
coeffcient de détermination, 58, 65,
86, 88, 91, 95, 96, 126, 127,
133, 135–137, 139, 148, 161,
162, 165, 188, 202, 205, 224,
234, 236, 244, 284, 286–292,
296, 297, 314, 315
coupe instantanée, 33
coupes transversales, 19, 20, 22, 141
différenciation généralisée, 267, 269,
309, 312
données centrées, 118, 218
données de panel, 19, 20, 22
données transversales, 17, 18, 20,
22, 247–249, 266, 274
doubles moindres carrés, 330, 331,
333
Durbin-Watson, 254, 258, 259, 343–
346
erreur de prévision, 66, 67, 144
Erreur de Type I, 76, 77, 84, 85
Erreur de Type I, 76
Erreur de Type II, 76, 77, 84–86
estimateur sans biais, 42, 44–46, 51,
118, 253, 275, 276
hétéroscédasticité, 179, 207, 253, 273–
276, 278, 280–283, 285, 286,
288, 290, 292–296, 298–302,
310, 312, 314–316, 318
homoscédasticité, 120, 273, 275, 276,
284, 287, 289, 291, 293, 296,
313, 315
intervalle de confiance, 53, 67, 74,
84, 86, 88, 92, 93, 102, 130,
163, 164, 200, 203, 214, 215,
235, 253, 262, 308
intervalle de prédition, 67
intervalle de prévision, 86
loi de Fisher, 63, 88, 143, 339, 340

INDEX

- loi de Student, 54, 55, 58, 88, 134, 211, 221, 342
loi Normale, 55, 144, 338
méthodes statistiques, 11
matrice Hat, 119, 121
modèle économétrique, 12–14, 16, 21, 22
modèle économique, 13, 14, 16, 22
modèle de régression exponentielle, 172
modèle log-linéaire, 171–174, 193–195
modèles réciproques, 171, 179, 180
modèles semilog, 171, 176
moindres carrés, 35, 36, 39, 40, 42, 47, 58, 60, 61, 68, 72, 79, 80, 117, 129, 159, 167, 168, 170, 200, 245, 246, 253, 276, 278, 279, 299, 300, 302, 322, 330, 331, 333
moindres carrés indirects, 330
multicolinéarité, 152, 153, 197, 223, 224, 227, 229, 230, 237, 240–243, 245–248, 303, 305, 306
multiplicateur de Lagrange, 283
prévision, 14, 66, 67
prévision ponctuelle, 66, 67
processus autorégressif, 249, 252
régression linéaire multiple, 113, 126
régression linéaire simple, 23, 33, 61, 87, 90, 93, 94, 96, 98, 109, 110, 115, 125, 126, 128, 132, 150, 227, 252, 270, 283
résidus récursifs, 143, 144
séries chronologiques, 18, 19, 22, 247–249, 266, 274, 275, 294
test bilatéral, 54, 56, 87, 96, 108
test de Breusch-Godfrey, 259–261, 308, 311
test de Breusch-Pagan, 283, 284, 286, 287, 296, 312, 317, 319
test de Chow, 141, 206, 207
test de Durbin-Watson, 254
test de Fisher, 86, 88, 89, 91, 93, 97–99, 128, 131, 134, 141, 147, 148, 157, 204, 205, 220, 239, 240
test de Glesjer, 287–289, 310, 312
test de Goldfeld-Quandt, 293–295, 310
test de Harvey-Godfrey, 289–291
test de Park, 291, 292
test de Ramsey, 146, 212, 215
test de White, 295–297, 310, 314, 316, 317
test du CUSUM, 143, 145, 146
test du CUSUM of Square, 143, 146, 147
test unilatéral, 55, 56, 90, 95, 99, 101, 104, 107, 202
vraisemblance, 50, 51, 57, 80, 121–123, 129, 167

Revue « Marché et Organisations »

Le but de la revue est de promouvoir la recherche originale sur les relations de plus en plus étroites qui se tissent entre le marché et les organisations. Les acteurs économiques de taille, de puissance et de pouvoir différents dont les intérêts peuvent être convergents, complémentaires ou, le plus souvent, antagoniques, ont tendance à organiser les marchés. La raison du marché, pourtant, est la référence stratégique pour l'entreprise ainsi que pour les institutions publiques de décision économique.

Numéros parus :

- N°1 : Artisanat. La modernité réinventée, 2006
N°2 : La petite entreprise, elle a tout d'une grande. De l'accompagnement aux choix stratégiques, 2006
N°3 : Tourisme et Innovation. La force créative des loisirs, 2007
N°4 : Le travail. Formes récentes et nouvelles questions, 2007
N°5 : Les universités et l'innovation. L'enseignement et la recherche dans l'économie des connaissances, 2007
N°6 : Entrepreneuriat et accompagnement. Outils, actions et paradigmes nouveaux, 2008
N°7 : Développement durable des territoires. Economie sociale, environnement et innovations, 2008
N°8 : Développement durable et responsabilité sociale des acteurs, 2009
N°9 : Gouvernance : exercices de pouvoir, 2009
N°10 : Le travail collaboratif. Une innovation générique, 2009
N°11 : Economie sociale et solidaire. Nouvelles trajectoires d'innovations, 2010
N°12 : Relations à la marque et marques de la relation. Regards croisés sur le management relationnel de la marque, 2010
N°13 : Les contrats au service de la recherche ?, 2011
N°14 : Le potentiel économique de l'Afrique subsaharienne, 2011
N°15 : Management de la distribution, 2012
N°16 : Territoire Vert. Entreprises, Institutions, Innovations, 2012
N°17 : Eco-conception, conception et innovation. Les nouveaux défis de l'entreprise, 2013
N°18 : Intelligence économique : entreprises et territoires, 2013
N°19 : La finance globale. Un monde fini, 2013
N°20 : La crise du « développement », 2014
N°21 : La Chine innove. Politiques publiques et stratégies d'entreprise, 2014
N°22 : Go East! Nouvelles Economies de Marché, 2015
N°23 : L'économie du changement, 2015
N°24 : Le temps des artisans. Permanences et mutations, 2015
N°25 : Innovations de proximité et esprit d'entreprise, 2016
N°26 : L'avenir des économies du Maghreb : entre inertie structurelle et envie de rupture, 2016/3 (n° 27)
N°27 : Le sport aux frontières du marché du travail, 2016
N°28 : Dynamiques internationales des entreprises, 2017
N°29 : L'univers du risque, 2017
N°30 : Les petites entreprises dans l'histoire industrielle, 2017
N°31 : Economie sociale et *social business* ? Au défi d'entreprendre et se financer, 2018
N°32 : Education et fertilisation des économies africaines, 2018
N°33 : L'entrepreneuriat innovant dans les pays du Maghreb, 2018
N°34 : L'entrepreneuriat scientifique : institutions et innovation, 2019
N°35 : Le champ économique de la culture, 2019
N°36 : Economie sociale et solidaire ? Modèles d'innovation et modes de gouvernance, 2019
N°37 : Rethinking Luxury Business, 2020
N°38 : La santé connectée. Nouvelles technologies et réorganisation des soins, 2020
N°39 : Options stratégiques et systèmes complexes d'innovation, 2020
N°40 : Les services : intégration des systèmes productifs et lien social, 2021
N°41 : Crise pandémique, crise globale. Réalité et controverses théoriques, 2021
N°42 : La crise de la covid-19. Le marché et les organisations en perspective, 2021
N°43 : (Re)faire système. *Bug & reset*, 2022

Collection « L'esprit économique »

fondée par Sophie Boutillier et Dimitri Uzunidis en 1996

Dernières parutions

► Série *Economie et Innovation*

- T. MICHAUD, *La réalité virtuelle : de la science-fiction à l'innovation*, 2018.
B. LAPERCHE, M. LIMA, E. SEULLIET, B. TROUSSE (dir.), *Les écosystèmes d'innovation. Regards croisés des acteurs clés*, 2019.
S. MONNERIE, *L'économie digitale révolutionne le monde. Innovation et total design*, 2019.
T. MICHAUD, *Le projet spatial européen, entre pragmatisme et imagination*, 2020.
J.-P. SCHMITT, N. RAVIDAT, *Management conjoint de la technologie et de la personne. Pour performer et croître*, 2020.
Y. EL YAHYAOUI, *Economie des plateformes numériques. Captation de la valeur, pouvoir de marché et communs collaboratifs*, 2021.

► Série *Le Monde en Question*

- F. FAURE, *Empowerment et finance. La démocratie au service de l'investissement socialement responsable*, 2021.
A. SGHAIER, *La finance islamique. Aspects critiques de la finance conventionnelle*, 2021.
H. AGOURRAME, *La réglementation bancaire internationale. Genèse, évolution et impact sur le « business model » des banques*, 2021.
A. AMINE, C. GALLOUJ, *Consommateurs et pratiques de consommation au Maroc*, 2021.
T. MARINOVA, *Economie sociale et solidaire dans les pays des Balkans. Bulgarie, Roumanie, Serbie : quels enseignements ?* 2021.
T. GARCENOT, *La finance contre le dérèglement climatique. Politiques monétaires et enjeux géopolitiques de la finance verte*, 2021.
A. VALLERAY, *Mondialisation mon amour. Avatars d'une utopie*, 2021.

► Série *Krisis*

- M. RAMEAUX, *Le tao de l'économie. Du bon usage de l'économie de marché*, 2020.
B. TONGLET, *Innovation et dynamique discontinue au bas Moyen Age. Pour une autre approche de l'histoire économique*, 2020.
J.-C. MOREL, *La nécessaire valorisation de la richesse ou intégrer le capital humain au bilan*, 2020.
V. LAURE VAN BAMBEKE, *La valeur du travail humain. Essai sur la refondation de l'expression monétaire de la valeur-travail*, 2021.
A. FONTAINE, *L'économie est l'opium du peuple. Une fausse science*, 2021.
A. RODRIGUEZ-HERRERA, *Travail, valeur et prix. Reprise et clôture d'un débat centenaire (1885-1985) à la lumière des textes marxiens*, 2021.

► Série *Clichés*

- S. BOUTILLIER, D. UZUNIDIS (dir.), *La Russie européenne. Du passé composé au futur antérieur*, 2008.
R. VOLPI, *La négociation. Pain, paix, liberté*, 2009.

► Série *Cours Principaux*

- O. ESNEU, *Le droit du transport routier de marchandises (TRM)*, 2018.
G. DUTHIL, *Economie publique et métamorphoses sociales*, 2019.
R. SANTENAC, *Le cycle des crises financières. Une étude approfondie des causes, des impacts et de la fréquence*, 2020.
B. LANDAIS, *Macroéconomie efficace. Croissance et crises*, 2020.
A. MBULI LANDU, *Le memo d'un comptable. Approche par le SYSCOHADA révisé*, 2021.

► Série *L'économie formelle*

- A. EL ALAOUI, *Modèle d'équilibre général calculable. Présentation théorique et application empirique*, 2020.
H. AMAAZOUL, *Responsabilité sociétale des entreprises et performance financière. Étude appliquée aux entreprises marocaines*, 2021.
M. BERTONECHE, *Chroniques économiques de notre temps*, 2021.
E. CARREY, H. LANDIER, *Après la guerre contre la Covid : de l'économie financière à l'entrepreneuriat social*, 2021.
J. LATREILLE, *Sale temps pour l'économie marchande*, 2021.

STRUCTURES ÉDITORIALES DU GROUPE L'HARMATTAN

L'HARMATTAN ITALIE

Via degli Artisti, 15
10124 Torino
harmattan.italia@gmail.com

L'HARMATTAN HONGRIE

Kossuth l. u. 14-16.
1053 Budapest
harmattan@harmattan.hu

L'HARMATTAN SÉNÉGAL

10 VDN en face Mermoz
BP 45034 Dakar-Fann
senharmattan@gmail.com

L'HARMATTAN CONGO

67, boulevard Denis-Sassou-N'Gesso
BP 2874 Brazzaville
harmattan.congo@yahoo.fr

L'HARMATTAN CAMEROUN

TSINGA/FECAFOOT
BP 11486 Yaoundé
inkoukam@gmail.com

L'HARMATTAN MALI

ACI 2000 - Immeuble Mgr Jean Marie Cisse
Bureau 10
BP 145 Bamako-Mali
mali@harmattan.fr

L'HARMATTAN BURKINA FASO

Achille Somé – tengnule@hotmail.fr

L'HARMATTAN GUINÉE

Almamya, rue KA 028 OKB Agency
BP 3470 Conakry
harmattanguinee@yahoo.fr

L'HARMATTAN TOGO

Djidjole – Lomé
Maison Amela
face EPP BATOME
ddamela@aol.com

L'HARMATTAN CÔTE D'IVOIRE

Résidence Karl – Cité des Arts
Abidjan-Cocody
03 BP 1588 Abidjan
espace_harmattan.ci@hotmail.fr

NOS LIBRAIRIES EN FRANCE

LIBRAIRIE INTERNATIONALE

16, rue des Écoles
75005 Paris
librairie.internationale@harmattan.fr
01 40 46 79 11
www.librairieharmattan.com

LIBRAIRIE DES SAVOIRS

21, rue des Écoles
75005 Paris
librairie.sh@harmattan.fr
01 46 34 13 71
www.librairieharmattansh.com

LIBRAIRIE LE LUCERNAIRE

53, rue Notre-Dame-des-Champs
75006 Paris
librairie@lucernaire.fr
01 42 22 67 13

Introduction à l'Économétrie Appliquée

Cette *Introduction à l'Économétrie Appliquée* est conçue pour un premier cours d'économétrie de premier cycle. Pour que l'économétrie soit pertinente dans un cours d'introduction, des applications intéressantes doivent faire appel à la théorie qui doit elle aussi correspondre aux applications.

Cet ouvrage fournit aux étudiants de premier cycle un aperçu des techniques économétriques utilisées par les économistes aujourd'hui. Le texte se concentre sur des techniques économétriques standards et aussi sur les développements récents dans le domaine.

Un aspect très utile de cet ouvrage est l'utilisation de données pour illustrer l'application des différentes techniques. Cette approche rend le texte vivant et très pertinent pour les étudiants et les chercheurs. Avec la grande disponibilité des progiciels économétriques tels que EViews, Stata, etc., les lecteurs peuvent acquérir une expérience pratique en les maniant dans certains exercices fournis dans le texte. Il est essentiel que les lecteurs comprennent les principes sous-jacents qui guident l'utilisation de la gamme de procédures des tests statistiques produites par ces programmes économétriques. L'objectif de ce livre est de préparer les étudiants à effectuer des travaux économétriques appliqués.

Moad El kharrim est professeur à la faculté des sciences économiques et gestion de l'Université Abdelmalek Essaâdi à Tétouan. Il y enseigne l'économétrie appliquée et la modélisation statistique appliquée en économie et finance en Licence et Master. Coordonnateur de master Gestion Informatique de l'Entreprise, ses travaux de recherche portent sur l'économétrie financière, la gestion de portefeuille et choix optimal d'actifs financiers.

ISBN : 978-2-14-026703-1

36 €

