

# Table des matières

<b>Table des matières</b>	<b>3</b>
<b>1 (re)Prise en main et algèbre</b>	<b>5</b>
1.1 Travail personnel	8
<b>2 Mesure de la liaison entre une variable et un ensemble de variables</b>	<b>13</b>
2.1 Etude du jeu de données Poids-Naissance	14
2.1.1 Régression linéaire simple	14
2.1.2 Régression linéaire multiple	19
2.2 Travail personnel	27
<b>3 Analyse en Composantes Principales</b>	<b>33</b>
3.1 Pratique de l'ACP	33
3.1.1 Notions de base	34
3.1.2 Calcul	35
3.1.3 Visualisation et interprétation	38
3.1.4 Biplot	53
3.2 Exemple sur les variétés d'eaux minérales	61
3.2.1 ACP avec d'autres fonctions de R	63
3.2.2 ACP "à la main"	63
3.2.3 Interprétation des résultats	64
3.3 Effet taille	65
3.4 Etude des données sur les pays l'OCDE	65
3.5 Devoir	66
<b>4 Analyse Factorielle des Correspondances (AFC)</b>	<b>81</b>
4.1 Pratique de l'AFC	81
4.1.1 Calcul	82
4.1.2 Visualisation et interprétation	86
4.1.3 Eléments supplémentaires	103
4.1.4 Filtrer les résultats	107
4.1.5 Outliers	108

4.1.6	Exportation des résultats	108
4.1.7	Résumé	108
4.1.8	Autres lectures	109
4.2	Exemple numérique	109
4.3	Etude des liens entre des catégories socio-professionnelles et le type d'hébergement	111
4.4	Devoir	112
<b>Bibliographie</b>		<b>114</b>
<b>5</b>	<b>Analyse Factorielle des Correspondances Multiples (AFCM ou ACM)</b>	<b>117</b>
5.1	Mise en oeuvre	117
5.1.1	Calcul	118
5.1.2	Visualisation et interprétation	121
5.1.3	Eléments supplémentaires	131
5.1.4	Filtrer des résultats	133
5.1.5	Exportation des résultats	135
5.1.6	Résumé	136
5.1.7	Autres lectures	136
5.2	Devoir	136
<b>Bibliographie</b>		<b>139</b>
<b>6</b>	<b>Classification</b>	<b>141</b>
6.1	Exemple numérique	141
6.2	Etude des données sur l'OCDE	141
6.3	Etude des données	141
6.4	Devoir	142

## Chapitre 2

# Mesure de la liaison entre une variable et un ensemble de variables

Dans la plupart des situations, nous sommes amenés à étudier la relation entre une variable d'intérêt  $Y$  (souvent quantitative) et une ou plusieurs variable(s)  $X_1, X_2, \dots, X_k$ , avec pour objectif d'expliquer les variations de la variable d'intérêt. La variable  $Y$  est appelée variable "à expliquer" (ou parfois variable dépendante, réponse, output, ...), et les variables  $X_1, X_2, \dots, X_k$  sont dites "explicatives" et représentent, en épidémiologie par exemple, les facteurs de risque ou de confusion. L'utilisation des méthodes d'analyse multivariée, et plus particulièrement des modèles de régression linéaire, permet donc :

- de prendre en compte simultanément plusieurs facteurs pouvant expliquer la variation ou la distribution de la variable  $Y$  ;
- d'étudier le rôle de modification d'effet ou de confusion d'un ou de plusieurs facteur(s) ;
- de prédire les valeurs ou la distribution de la variable à expliquer connaissant les valeurs des variables explicatives.

Comme pour le TP précédent, on peut commencer par :

1. Dans le répertoire `TP_M1MIASHS_SSD_AD` qui est déjà créé.
2. Lancer ensuite R et modifier le répertoire de travail en allant dans `Fichier -> Changer le Répertoire Courant` et en choisissant le répertoire `Bureau/TP_M1MIASHS_SSD_AD` qui a été créé.
3. Ouvrir une fenêtre d'éditeur `Fichier -> Nouveau Script`.

4. Sauver le fichier dans le répertoire courant sous le nom TP1.R : **Fichier -> Sauver sous**
5. Pour les différentes questions, on peut utiliser un “copier-coller” à partir de ce document. *Il est fortement recommandé de saisir toutes les commandes dans la fenêtre ouverte de l’éditeur.* Pour exécuter les commandes saisies, il suffit de les sélectionner avec la souris et d’appuyer simultanément sur les touches **Ctrl** et **R**.
6. Pour inclure des commentaires dans le programme, ce qui est fortement recommandé, utiliser le caractère **#**. Tout ce qui suit le caractère **#** sera négligé lors de l’exécution.
7. Penser à sauvegarder régulièrement le contenu du fichier TP1.R en appuyant sur les touches **Ctrl** et **S**.

## 2.1 Etude du jeu de données Poids-Naissance

Un échantillon de dossiers d’enfants a été saisi. Ce sont des enfants vus lors d’une visite en 1ère section de maternelle en 1996 – 1997 dans des écoles de Bordeaux (Gironde, France). L’échantillon est constitué de 152 enfants âgés de 3 ou 4 ans.

Nous allons travailler sur le jeu de donnée Poids-Naissance (cf. fichier `Poids_naissance.txt`). Il s’agit ici d’expliquer la variabilité du poids de naissance de l’enfant en fonction des caractéristiques de la mère, de ses antécédents et de son comportement pendant la grossesse. La variable à expliquer est le poids de naissance de l’enfant (variable quantitative **BWT**, exprimée en grammes) et les facteurs étudiés (variables explicatives) sont : Age de la mère, Poids de la mère lors du dernier cycle menstruel, “Race” de la mère, Tabagisme durant la grossesse, Nombre d’antécédents de prématurité, Antécédents d’hypertension, Présence d’irritabilité utérine, Nombre de visites à un médecin durant le premier trimestre de la grossesse, Poids de naissance et Poids de naissance inférieur ou égal à 2500 g.

1. Lecture des données.
2. Le poids de la mère étant exprimé en **livres**, nous commençons par effectuer une transformation du `data.frame` des données pour recoder cette variable en kilogrammes (1 livre = 0.45359237 kg). Convertir dans les données les poids des mamans en *kg*.

### 2.1.1 Régression linéaire simple

Le tableau ci-dessous présente les principales fonctions à utiliser afin d’effectuer une régression linéaire simple entre la variable à expliquer  $Y$  et la variable explicative  $X$ . On cherche à “expliquer” les variations d’une variable quantitative  $Y$  (par exemple, le poids de naissance de l’enfant, noté **BWT**) par une variable explicative  $X$  également quantitative (par exemple, le poids noté **LWT**). Le modèle :

$$\text{BWT}_i = \beta_0 + \beta_1 \text{LWT}_i + \varepsilon_i, i = 1, \dots, n,$$

Description	Unité ou Codage	Variable
Age de la mère	Années	AGE
Poids de la mère lors du dernier cycle menstruel	Livres	LWT
Race de la mère	1=Blanche ; 2=Noire ; 3= Autre	Race
Tabagisme durant la grossesse	Oui=1 ; Non =0	SMOKE
Nombre d'antécédents de prématurité	0=Non ; 1 = Un ; 2 =Deux ; etc	PTL
Antécédents d'hypertension	Oui=1 ; Non =0	HT
Presence d'irritabilité utérine	Oui=1 ; Non =0	UI
Nombre de visites à un médecin durant le premier trimestre de la grossesse	O=Aucune ; 1=Une ; etc.	FVT
Poids de naissance	Grammes	BWT
Poids de naissance inférieur ou égal à 2500 g	Oui=1 ; Non =0	LOW

TABLE 2.1 – Variables et codage du jeu de données : Poids\_naissance (txt, xls, ...)

Instruction R	Description
<code>plot(Y~X)</code>	Graphe du nuage de points
<code>lm(Y~ X)</code>	Estimation du modèle linéaire
<code>summary(lm(Y~ X))</code>	Description des résultats du modèle
<code>abline(lm(Y~ X))</code>	Trace la droite estimée
<code>confint(lm(Y~ X))</code>	Intervalle de confiance des paramètres de régression
<code>predict()</code>	Fonction permettant d'obtenir des prédictions
<code>plot(lm(Y~ X))</code>	Analyse graphique des résidus

TABLE 2.2 – Liste des principales fonctions R permettant l'analyse d'une régression linéaire simple

où les  $\varepsilon_i$  sont des variables aléatoire centrées et de variance,  $\sigma^2$ , constante pour tout  $i$ .

**Remarque 2.1.1.** L'hypothèse gaussienne du bruit  $\varepsilon$  permet d'obtenir la loi des estimateurs et ainsi d'effectuer des tests d'hypothèses sur les paramètres du modèle. Cependant, cette hypothèse n'est pas très importante puisque l'on peut s'en passer quand le nombre de données est important.

### 1. Ajustement sur des données

- (a) *Inspection graphique* : Afin d'étudier la relation entre le poids de naissance de l'enfant et l'âge de la mère, nous pouvons commencer par tracer le nuage des points (poids de l'enfant (en g) versus poids de la mère (en kg)) grâce à l'instruction `plot(BWT~ LWT)`. Que remarque-t-on ?
- (b) *Estimation des paramètres* : La fonction R permettant de réaliser cette opération est la fonction `lm()` (abregé de *linear models*). Le paramètre principal de cette fonction est une formule, symbolisée par un tilde  $\sim$ , qui permet de préciser le sens de la relation entre BWT et LWT.
  - i. Donner les estimations MCO des coefficients  $\beta_0$  et  $\beta_1$ .
  - ii. Représenter la droite de régression sur le nuage de points.
- (c) *Tests sur les paramètres* : Il est bon de noter que la fonction `lm()` permet une analyse complète du modèle linéaire et qu'on peut récupérer un résumé des calculs liés au jeu de données en utilisant la fonction `summary()`.

**2. Tableau d'analyse de la variance** En régression linéaire simple, le test de Fisher (dont la statistique se lit dans `F-statistic`) est équivalent au test de Student associé à la pente de la régression. On a la relation suivante `F-statistic=t2` et les  $p$ -valeurs des tests sont égales. Le test de Fisher est souvent associé à une table d'analyse de la variance qu'on peut avoir en utilisant la fonction `anova()`.

### 3. Interprétation des résultats

- (a) Le test associé à l'intercept  $\beta_0$  du modèle est significatif ( $p$ -valeur  $< 0.05$ ), il est donc conseillé de garder l'intercept ( $\beta_0$ ) dans le modèle. Toutefois, l'intercept de cette régression n'a aucun sens. Il pourrait donc être plus judicieux de considérer une régression sur la variable poids de la mère, préalablement centrée. Dans ce cas-ci,  $\beta_0$  représenterait le poids moyen des enfants pour des mères ayant un poids égal à la moyenne des poids des mères observées. Essayer la commande `lm(y~ x-1)`
- (b) La relation linéaire entre BWT et LWT est démontrée par le résultat du test de Student sur le coefficient  $\beta_1$ . La  $p$ -valeur  $< 0.05$  nous indique une relation linéaire significative entre le poids de l'enfant et le poids de la mère.
- (c) Le pourcentage de variance expliqué par la régression ( $r^2$ ) vaut 0.035. Ce qui veut dire que seulement 3.5% de la variabilité du poids de l'enfant est expliquée par le

• Call	Un rappel e la formule utilisée dans le modèle
• Residuals	Une analyse descriptive des résidues $\hat{\varepsilon}_i = \hat{y}_i - y_i$ . On verra, par la suite, l'intérêt des résidues pour valider les hypothèses du modèle de régression
• Coefficients	Ce tableau comprend quatre colonnes :
Estimate	Correspond aux estimations des parametres de la droite de régression
Std. Error	Correspond à l'estimation de l'écart-type des estimateurs de la droite de régression
t value	Correspond à la réalisation de la statistique du test de Student associée aux hypothèses $H_0 : \beta_i = 0$ et $H_1 : \beta_i \neq 0$
Pr(>  t )	Correspond à la $p$ -valeur du test de Student
• Signif. codes	Symboles de niveau de significativité
• Residual standard error	Une estimation de l'écart-type du bruit $\sigma$ est fournie ainsi que le degré de liberté associé $n - 2$
• Multiple R-Squared	Valeur du coefficient de détermination $r^2$ (pourcentage de variance expliqué par la régression).
• Adjusted R-Squared	$r_a^2$ ajusté (qui n'a pas grand intérêt en régression linéaire simple).
• F-Statistic	Correspond à la réalisation du test de Fisher associé aux hypothèses $H_0 : \beta_1 = 0$ et $H_1 : \beta_1 \neq 0$ . Nous y trouvons les degrés de liberté associés (1 et $n - 2$ ) ainsi que la $p$ -valeur.

TABLE 2.3 – Descriptions des différentes informations contenues dans la sortie `summary`

poids de la mère. Il sera donc utile d'introduire dans le modèle d'autres variables explicatives (régression linéaire multiple) afin d'améliorer le pouvoir prédictif du modèle.

- (d) L'estimation de la pente indique que la différence entre les poids moyens de naissance des bébés de mères ayant un écart de poids d'un kilogramme est de 9.765 g.

#### 4. Intervalle de confiance et de prédiction pour une nouvelle valeur

Considérons une nouvelle observation  $x_0$  de la variable  $X$  pour laquelle nous n'avons pas observé la valeur  $y_0$  correspondante de la variable à expliquer  $Y$ . Cette valeur  $y_0$  inconnue, car non observée, est une réalisation de la variable aléatoire  $Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0$ . Le *préviseur* (ou *prédicteur*) de  $Y_0$  pour la nouvelle valeur  $x_0$  est donné par :

$$\hat{y}_0^p = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

La fonction permettant de définir l'intervalle de prévision et l'intervalle de confiance pour une nouvelle valeur  $x_0$  est `predict()`.

- (a) Calculer la prédiction pour le poids d'un bébé dont la mère a un poids de `lwt = 56 kg`.
- (b) Donner un intervalle de confiance de la valeur moyenne du poids des bébés pour un poids de la mère de 56 kg.

- (c) En considérant la série de nouvelles valeurs de poids de la mère
- ```
x <- seq(min(L WT),max(BWT),length=50)
```
- représenter l'intervalle de confiance et l'intervalle de prévision.

## 5. Analyse des résidus

### (a) Vérification des hypothèses du modèle

L'analyse des résidus consiste à examiner si les hypothèses de base du modèle linéaire sont violées. Différents graphiques tracés à l'aide des résidus permettent de détecter assez facilement si les hypothèses sur les erreurs  $\varepsilon_i$  ne sont pas respectées :

- Tracé de l'histogramme des résidus pour détecter la non-normalité. Le graphique **QQ-plot** est une autre approche. On peut également appliquer le test de non-normalité de Jarque et Bera sur les résidus disponible dans le *package* **tseries**.
- Tracé des résidus  $\hat{\varepsilon}_i$  en fonction des valeurs prédites  $\hat{y}_i$ . En effet, lorsque les hypothèses associées au modèle sont correctes, les résidus et les valeurs prédites sont non corrélées. Par conséquent, le tracé de ces points ne devrait pas avoir de structure particulière. Ce type de tracé donne aussi des indications sur la validité des hypothèses de linéarité, ainsi que sur l'homogénéité de la variance de l'erreur. On devrait observer sur le graphique de  $\hat{\varepsilon}_i$  versus  $\hat{y}_i$  une répartition uniforme des résidus suivant une bande horizontale de part et d'autre de l'axe des abscisses.
- Pour l'hypothèse d'indépendance des  $Y_i$ , si les valeurs de  $Y_i$  sont mesurées chez des individus différents sans aucune relation entre eux, l'hypothèse d'indépendance est en principe valide. En revanche, si les valeurs de  $Y_i$  représentent des mesures d'une certaine quantité au cours du temps (par exemple chaque mois), l'hypothèse d'indépendance peut ne pas être valide. On peut alors vérifier l'hypothèse d'indépendance en examinant l'autocorrélation des résidus soit graphiquement en portant sur un graphique les résidus successifs  $\hat{\varepsilon}_i$  soit en utilisant des tests statistiques comme, par exemple, le test de Durbin-Watson (fonction **dwtest()** dans le *package* **Intest**).

### (b) Exemple d'analyse des résidus

Malgré que le pouvoir prédictif, pour le modèle étudié sur les poids de naissance, est faible ( $r^2 = 3.5\%$ ), nous allons faire une étude des résidus.

Tout d'abord, examinons l'hypothèse de normalité :

```
par(mfrow=c(1,2))
hist(residuals(modele1), main="Histogramme")
qqnorm(resid(modele1),datax=TRUE) # Attention : quantiles normalises
en ordonnee
```



- i. Tracer l'histogramme des résidus et le `QQ-plot`.
- ii. Faire un test de Jarque Bera. Conclure.
- iii. Afin d'examiner le graphe des résidus en fonction des valeurs prédites, tracer le nuage correspondant. Conclure.

### 2.1.2 Régression linéaire multiple

Il s'agit d'étudier les variations d'une variable quantitative  $Y$  (variable dite à expliquer ou dépendante, supposée aléatoire) en fonction de  $p$  ( $p > 1$ ) variables explicatives  $X_1, X_2, \dots, X_p$  (variables aussi dites indépendantes). Les variables explicatives peuvent être uniquement quantitatives, uniquement qualitatives (auquel cas on retombe sur l'ANOVA), ou un mélange de variables quantitatives et qualitatives. Dans ce dernier cas, le modèle de régression linéaire multiple est aussi appelé ANCOVA.

#### 1. Objectif et modèle

Le modèle de régression linéaire multiple s'écrit  $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$ , où  $\varepsilon$  représente le terme de perturbation aléatoire (bruit) du modèle souvent supposé gaussien d'espérance nulle et de variance  $\sigma^2$ , et indépendant des  $X_j$ . Les paramètres (inconnus) du modèle de régression sont  $\beta_0, \beta_1, \dots, \beta_p$  et  $\sigma^2$ .

Le tableau ci-dessous présente les principales fonctions à utiliser afin d'effectuer une régression linéaire multiple entre la variable à expliquer  $Y$  et les variables explicatives  $X_1, \dots, X_k$ .

| Instruction R                     | Description                                                  |
|-----------------------------------|--------------------------------------------------------------|
| <code>pairs()</code>              | Inspection graphique                                         |
| <code>lm(Y ~ X1+X2+...+Xk)</code> | Estimation du modèle linéaire multiple                       |
| <code>summary(lm())</code>        | Description des résultats du modèle                          |
| <code>confint(lm())</code>        | Intervalle de confiance des paramètres de régression         |
| <code>predict()</code>            | Fonction permettant d'obtenir des prédictions                |
| <code>plot(lm())</code>           | Analyse graphique des résidus                                |
| <code>anova(mod1, mod2)</code>    | Test de Fisher partiel                                       |
| <code>X1 : X2</code>              | Interaction entre $X_1$ et $X_2$                             |
| <code>vif()</code>                | Calcul du VIF (disponible dans le package <code>car</code> ) |

TABLE 2.4 – Liste des principales fonctions R permettant l'analyse d'une régression linéaire multiple

- 2. **Ajustement sur des données** Le jeu de données provenant du modèle ci-dessus, pour les  $n$  individus, peut s'écrire sous la forme  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , où  $\mathbf{y}$  est le vecteur des  $n$  réponses,  $\boldsymbol{\beta}$  le vecteur des  $p$  paramètres de régression,  $\mathbf{X}$  est la matrice  $(n, p+1)$  des covariables (dont la 1ère colonne ne comporte que des 1) et  $\boldsymbol{\varepsilon}$  est le vecteur des

$n$  erreurs.

L'estimateur de  $\beta$  est donné par :

$$\hat{\beta} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}$$

**Remarque 2.1.2.** On suppose que la matrice  $\mathbf{X}$ , de type  $(n, p + 1)$ , est de plein rang ( $\text{rang}(\mathbf{X}) = p + 1 < n$ ). Le but est que la matrice  $\mathbf{X}^t \mathbf{X}$  soit inversible (car  $\text{rang}(\mathbf{X}^t \mathbf{X}) = p + 1$ ).

Afin d'illustrer ce modèle, nous considérons le même exemple et démarche que dans le livre de Drouilhet et al. (2012). Il s'agit d'étudier la régression du poids de naissance de l'enfant (BWT qu'on notera  $Y$  : variable à expliquer) en fonction de l'âge de la mère (AGE qu'on notera  $X_1$ ), de son poids (LWT qu'on notera  $X_2$ ) et de son statut tabagique durant la grossesse (SMOKE qu'on notera  $X_3$ ). Du coup, nous avons l'équation de la régression suivante :

$$\mathbb{E}(\text{BWT} | \text{AGE, LWT, SMOKE}) = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{LWT} + \beta_3 \text{SMOKE}$$

- (a) *Inspection graphique* : Afin de visualiser la relation entre la variable à expliquer et chacune des variables explicatives, et aussi de juger de la corrélation entre les variables explicatives, voire de diagnostiquer un problème potentiel de colinéarité entre les variables explicatives, tracer un diagramme de dispersion de toutes les paires de ces variables.
- (b) *Estimation des paramètres* : Comme en régression linéaire simple, estimer le modèle par la fonction `lm`.
- (c) Utiliser la commande `summary` pour faire des tests sur les différents paramètres. Conclure.

Rappelons que les valeurs réalisées des statistiques des tests de Student associés aux hypothèses  $H_0 : \beta_i = 0$  versus  $H_1 : \beta_i \neq 0$ , se trouvent dans la colonne `t value`, les  $p$ -valeurs associées dans la colonne `Pr(>|t|)`. `Residual standard error` fournit l'estimation de  $\sigma$  ainsi que le nombre de degrés de liberté associés  $n - p - 1$ . On trouve enfin le coefficient de détermination  $r^2$  (`Multiple R-squared`) ainsi qu'une version ajustée (`Adjusted R-squared`). Enfin, on trouve la réalisation du test de Fisher global (`F-statistic`) et sa  $p$ -valeur- $p$  (`p-value`) associée.

- (d) Utiliser la commande `anova` pour obtenir la table d'analyse de la variance.
- (e) Interpréter les résultats de l'étude sur les poids de naissance.
- (f) Aurait-on bien pu conclure en utilisant la commande `confint` ?
- (g) Supposons qu'une mère est âgée de 23 ans, pèse 57 kg et fume. Faire une prédiction du poids de naissance de son enfant, et son intervalle de prévision ainsi que l'intervalle de confiance du poids moyen des enfants dont les mères ont les mêmes caractéristiques.

- (h) *Test d'une sous-hypothèse linéaire partiel* : Le test de Fisher partiel permet de tester l'apport d'un sous-ensemble de variables explicatives dans un modèle qui en contient déjà d'autres. Considérons par exemple les deux modèles suivants :

$$\text{BWT} = \beta_0 + \beta_1 \text{LWT} + \varepsilon \quad \text{et} \quad \text{BWT} = \beta_0 + \beta_1 \text{LWT} + \beta_2 \text{AGE} + \beta_3 \text{SMOKE} + \varepsilon$$

Décrire les hypothèses puis faire le test de Fisher afin de tester l'apport simultané des variables AGE et SMOKE dans le modèle. Conclure.

### 3. Cas des variables qualitatives à plus de deux modalités

Le cas des variables explicatives binaires ne pose pas de problème comme nous avons pu le voir dans l'exemple pour la variable SMOKE. L'utilisation d'une telle variable dans un modèle de régression linéaire multiple équivaut à comparer les moyennes de la variable à expliquer  $Y$  dans les deux groupes définis par la variable qualitative binaire. Cette comparaison est ajustée sur les autres variables explicatives. Cependant, pour une variable qualitative à plus de deux modalités, il est nécessaire d'introduire des variables indicatrices (*dummy variables*) afin de comparer les moyennes de  $Y$  dans les différents groupes définis par les modalités de la variable explicative qualitative. Une variable indicatrice d'un groupe ou d'une modalité est une variable binaire qui prend la valeur 1 pour ce groupe et 0 pour les autres groupes.

Prenons l'exemple sur le poids de naissance de l'enfant en fonction de la "race"/origine/couleur de la mère et de son poids. La variable RACE est codée avec trois modalités : 1 pour blanche, 2 pour noire et 3 pour autre. Dans cet exemple, on peut donc définir trois variables indicatrices RACE1, RACE2, RACE3. Utiliser la fonction `factor()` dans l'instruction `lm()` pour ajuster le modèle. Interpréter.

**Remarque 2.1.3.** Ne jamais introduire dans le modèle l'indicatrice du groupe de référence : pour une variable qualitative à  $p$  modalités, il ne faut introduire que  $p - 1$  variables indicatrices.

4. **Interaction entre les variables** On dit donc qu'il y a interaction entre deux variables explicatives  $X_1$  et  $X_2$  si l'association entre l'une des variables et la variable à expliquer  $Y$  n'est pas la même selon les valeurs de l'autre variable. Cela correspond à la notion de modification d'effet en épidémiologie, par exemple.

Supposons qu'on s'intéresse à l'association entre  $X_1$  et  $Y$ , et que l'on veuille savoir si l'effet de  $X_1$  sur  $Y$  est différent selon les valeurs de  $X_2$ . Il suffit pour cela d'introduire dans le modèle les variables  $X_1$  et  $X_2$  ainsi qu'une troisième variable  $X_3$  définie comme le produit de  $X_1$  et  $X_2$  :  $X_3 = X_1 \times X_2$  (appelée terme d'interaction entre  $X_1$  et  $X_2$ ). Pour fixer les idées, supposons que l'on veuille déterminer si l'effet d'une variable  $X_1$  quantitative est modifié par une variable  $X_2$  binaire (prenant les valeurs 0 et 1). On considère alors le modèle :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 + \varepsilon$$

Dans le groupe  $X_2 = 0$ , le modèle s'écrit :  $Y = \beta_0 + \beta_1 X_1 + \varepsilon$  et l'effet de  $X_1$  est mesuré par  $\beta_1$ . Dans le groupe  $X_2 = 1$ , le modèle s'écrit  $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \varepsilon$  et l'effet de  $X_1$ , est mesuré par  $\beta_1 + \beta_3$ .

Il y a interaction entre  $X_1$  et  $X_2$  (ou modification de l'effet de  $X_1$  par  $X_2$ ) si l'effet de  $X_1$  est différent dans les groupes  $X_2 = 0$  et  $X_2 = 1$  ; autrement dit si  $\beta_3 \neq 0$ . Il suffit donc d'effectuer un test individuel de Student sur  $\beta_3$  :  $H_0 : \beta_3 = 0$  versus  $H_1 : \beta_3 \neq 0$ . Si l'on accepte  $H_1$ , on garde le terme d'interaction dans le modèle : il y a modification d'effet.

- (a) Etudier l'effet de l'âge de la mère sur le poids de naissance de l'enfant par le statut tabagique de la mère.
- (b) Proposer un modèle plus souple, permettant d'avoir des effets différents dans les deux groupes  $\text{SMOKE} = 0$  et  $\text{SMOKE} = 1$ .

**Remarque 2.1.4.** Pour introduire un terme d'interaction entre deux variables  $X_1$  et  $X_2$  dans un modèle linéaire, il suffit de taper  $\mathbf{X}_1 : \mathbf{X}_2$ . Noter que l'écriture  $\mathbf{X}_1 * \mathbf{X}_2$  dans une formule correspond à l'écriture  $\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_1 : \mathbf{X}_2$

- (c) Peut-on visualiser graphiquement la différence entre ces deux modèles ?
- (d) Tester la significativité du terme d'interaction. Conclure.

## 5. Problème de la colinéarité

Lorsque plusieurs variables explicatives apportent le même type d'information, plusieurs phénomènes peuvent apparaître :

- qualité des estimations perturbée (variance très grande)
- valeurs des coefficients contradictoires (signes opposés)
- coefficients devenant non significatifs

C'est ce que l'on nomme *le problème de colinéarité*.

Le critère utilisé pour juger de la colinéarité entre les variables explicatives est le facteur d'inflation de la variance **VIF** (variance inflation factor) :  $\frac{1}{1 - r_j^2}$  où  $r_j^2$  ne désigne rien d'autre que le coefficient de corrélation multiple au carré (coefficient de détermination) lorsque l'on régresse la  $j$ ème variable explicative  $\mathbf{x}_j$  sur l'ensemble des autres régresseurs.

Utiliser `vif()`, du package `car`, sur le modèle de régression de **BWT** sur **LWT** et **AGE**.

**Remarque 2.1.5.** Il n'est pas étonnant que lorsqu'il n'y a que deux régresseurs, les **VIF** soient identiques. En fait, il n'y a que deux régresseurs et on aurait pu analyser cette colinéarité graphiquement, mais l'on comprend aisément son intérêt lorsque l'on a un grand nombre de régresseurs.

## 6. Sélection de variables

Parmi le grand nombre de variables explicatives potentielles, il s'agit de sélectionner celles qui sont le plus à même d'expliquer  $Y$ . Cela permet d'économiser le nombre de

prédicteurs (et ainsi d'obtenir un modèle parcimonieux) et d'obtenir un bon pouvoir prédictif en éliminant les variables redondantes qui augmentent le facteur d'inflation de la variance (VIF).

Plus le nombre de paramètres augmente (nombre important de variables explicatives), plus l'ajustement aux données est bon ( $r^2$  proche de 1). En contre-partie, l'estimation des paramètres est détériorée (la variance des estimateurs augmente) à cause des problèmes de colinéarité.

Nous présentons brièvement dans cette sous-section quelques méthodes de sélection de variables disponibles avec le logiciel R. Celles-ci sont illustrées sur quelques variables du jeu de données "Poids à la naissance".

Les variables explicatives considérées sont : LWT, AGE, UI, SMOKE, HT et deux variables recodées FTV1 et PTL1. On note FVT1 = 1 s'il y a eu au moins une visite chez le médecin, et FVT1 = 0 sinon. De même, on note PTL1 = 1 s'il y a au moins un antécédent de prématurité, et PTL1 = 0 sinon.

- (a) *La méthode du meilleur sous-ensemble (best subset)* : Lorsque le nombre  $p$  de variables explicatives n'est pas trop grand, on peut étudier toutes les possibilités. Un algorithme efficace (voir [19] et [20]) permet ainsi d'aller jusqu'à une trentaine de variables ; il s'agit de la procédure *leaps and bounds*. A  $p$  fixe, on choisira le modèle de régression qui fournit le  $r^2$  le plus grand. Pour deux modèles de régression ayant un nombre différent de variables explicatives, on peut choisir celle qui fournit le  $r_a^2$  ajusté le plus grand.
  - i. Utiliser la fonction `leaps()` qui est disponible dans le package `leaps` de R, pour donner le meilleur modèle au sens du  $r_a^2$  ajusté.  
**Remarque 2.1.6.** Il est possible d'obtenir d'autres critères de sélection que le  $r_a^2$  ajusté grâce au paramètre `method` de la fonction `leaps()`. Par exemple, `method="Cp"` utilise le critère bien connu du  $C_p$  de Mallows [26].
  - ii. Une autre fonction très intéressante disponible dans le package `leaps` est la fonction `regsubsets()`. Elle permet par exemple au moyen de son paramètre `force.in` de spécifier une ou plusieurs variables qui seront incluses dans tous les modèles comparés. Un exemple est donné ici en faisant le choix du meilleur modèle par le critère BIC (*bayesian information criterion*). Le meilleur modèle est celui ayant obtenu la valeur du critère BIC la plus petite. Donner le meilleur modèle au sens du BIC de l'exemple précédent.  
**Remarque 2.1.7.** Il est possible d'obtenir d'autres critères de sélection grâce au paramètre `scale` utilisé dans la fonction `plot()` appliquée à un objet de classe `regsubsets`.
- (b) *La méthode pas à pas ascendante (forward selection)* : La régression pas à pas ascendante (ou méthode par additions successives) est une méthode itérative. Elle

consiste à sélectionner à chaque étape la variable explicative la plus significative (au seuil  $\alpha$ ) lorsque l'on régresse  $Y$  sur toutes les variables explicatives sélectionnées aux étapes précédentes et la nouvelle variable choisie, tant que l'apport marginal de cette dernière est significatif.

- i. Montrer le fonctionnement de cette procédure à l'aide de la fonction `add1()` pour un seuil  $\alpha = 0.05$ .
  - ii. Quand plus aucune variable n'est significative, donner le modèle final.
- (c) *La méthode pas à pas descendante (backward selection)* : Cette méthode est aussi appelée régression par éliminations successives. On part cette fois du modèle complet et on élimine à chaque étape la variable ayant la plus petite valeur pour la statistique du test de Student ( $p$ -valeur la plus grande) en valeur absolue, à condition qu'il soit non significatif (au seuil  $\alpha$  choisi).
- i. Montrer le fonctionnement de cette procédure à l'aide de la fonction `drop1()` pour un seuil  $\alpha = 0.05$ .
  - ii. Quand plus aucune variable n'est significative, donner le modèle final.
- (d) *La méthode pas à pas (stepwise)* : Cet algorithme est un perfectionnement de la méthode ascendante. Il consiste à effectuer en plus, à chaque étape, des tests du type Student ou Fisher pour ne pas introduire une variable non significative et pour éliminer éventuellement des variables déjà introduites qui ne seraient plus informatives compte tenu de la dernière variable sélectionnée. L'algorithme s'arrête quand on ne peut plus ajouter ni retrancher de variables.

A l'aide de la fonction `step()` effectuer la méthode pas à pas (en utilisant à chaque étape de la procédure une sélection par le critère AIC (*an information criterion*)).

**Remarque 2.1.8.** Le jeu de données sur les poids à la naissance a été adopté ici uniquement pour illustrer l'utilisation des fonctions R qui sont propres aux méthodes de sélection automatique alors que la bonne stratégie sur ce jeu de données particulier aurait plutôt consisté à procéder "manuellement". En effet, il est important de noter que diverses méthodes de sélection automatique peuvent ne pas conduire aux mêmes choix de variables explicatives à retenir dans le modèle final. Elles ont l'avantage d'être faciles à utiliser, et de traiter le problème de la sélection de variables de façon systématique. En revanche, l'inconvénient majeur est que les variables sont retenues ou éliminées du modèle sur la base de critères uniquement statistiques, sans tenir compte de l'objectif de l'étude. On aboutit généralement à un modèle qui peut être satisfaisant sur le plan purement statistique, alors que les variables retenues ne sont pas les plus pertinentes pour comprendre et interpréter les données de l'enquête.

## 7. Analyse des résidues

Nous présentons ici quelques éléments sur l'analyse des résidus permettant de vérifier les hypothèses du modèle et de détecter des valeurs possiblement atypiques ou aberrantes.

- (a) *Validation des hypothèses du modèle* : En régression linéaire simple, nous avons déjà évoqué l'analyse des résidus pour éprouver les hypothèses du modèle de régression. Deux graphiques ont été présentés permettant d'invalidier l'hypothèse de normalité des erreurs et l'hypothèse d'homoscédasticité des erreurs.

Reprenons l'exemple sur les poids à la naissance des enfants. Nous étudions la validité des hypothèses pour le modèle suivant :

$$\begin{aligned} \text{BWT} = & \beta_0 + \beta_1 \text{SMOKE} + \beta_2 \text{AGE} + \beta_3 \text{LWT} + \beta_4 \text{RACE2} \\ & + \beta_5 \text{RACE3} + \beta_6 \text{UI} + \beta_7 \text{HT} + \beta_8 \text{SMOKE} \times \text{AGE} + \varepsilon \end{aligned}$$

- i. Tracer les graphiques des résidus versus les valeurs ajustées.
  - ii. Il peut aussi être utile de représenter les résidus en fonction de chaque variable explicative (6 graphes). Ce type de graphique permet de détecter s'il existe une relation entre le terme d'erreur et les variables explicatives et ainsi de vérifier l'hypothèse d'indépendance entre les erreurs et les variables explicatives. Outre le fait que ce genre de graphique permet de vérifier l'hypothèse d'indépendance entre les erreurs et les variables explicatives, il est aussi utile afin de visualiser des points qui seraient potentiellement atypiques.
- (b) *Points atypiques et/ou influents* : Un point atypique ou aberrant est un point s'écartant des autres. Il peut être visualisé sur le graphique des résidus versus les valeurs prédites (ou versus l'une des variables explicatives), comme un point très éloigné. Plusieurs types de résidus sont alors définis :

- i. *les résidus standardisés*  $t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$  avec  $h_{ii}$  le "levier" (défini ultérieurement). Ces résidus sont obtenus par la fonction `rstandard()`. Les résidus standardisés sont "principalement" compris entre  $-2$  et  $2$ , mais ils sont dépendants
- ii. *les résidus studentisés*

$$t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}} = t_i \sqrt{\frac{n-p-2}{n-p-1-t_i^2}}$$

où  $\sigma_{(-i)}$  est l'estimation résiduelle, obtenue sans l'utilisation de l'observation  $i$ . Les résidus studentisés sont obtenus via la fonction `rstudent()`. Une observation sera considérée comme aberrante lorsque  $|t_i^*| > t_{n-p-2,0.975}$  où  $t_{n-p-2,0.975}$  est le quantile d'ordre 0.975 d'une loi de Student à  $n-p-2$  degrés de liberté.

- i. Commenter les résultats obtenus par les commandes suivantes :

```

res.stud <- rstudent(rnodelefinal) # Calcul des residus studentises.
seuil.stud <- qt(0.975,189-822) # Calcul du seuil par la loi de Student
cond <- res.stud<(-seuil.stud) | res.stud > seuil.stud # Liste des
individus susceptibles d'être consideres comme aberrants.
id.student <- ID[cond]
val.ajust <- fitted(modelefinal)
plot(res.stud~val.ajust, xlab="Valeurs ajustées",
ylab="Résidus studentisés")
abline(h=c(-seuil.stud, seuil.stud))
text(val.ajust[cond],res.stud[cond],id.student,col="red",pos=1)

```

- ii. Un autre moyen pour étudier les points atypiques est la notion de “points leviers”. Le levier pour l’observation  $i$  (noté  $h_{ii}$ ) est la valeur lue sur la diagonale de la matrice  $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$  (dite *hat matrix*). Cette mesure intervient principalement dans la variance des résidus :  $var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$ . Un levier dépassant  $2(p+1)/n$  peut être considéré comme important. Un  $h_{ii}$  élevé indique que la  $i$ ème observation est éloignée du centre de gravité. Les  $h_{ii}$  stockés dans le vecteur `levier`, sont obtenus par :

```

levier <- hatvalues(modelefinal)
# Equivalent à hat(model.matrix(modelefinal))
# On aurait pu également utiliser les deux instructions suivantes :
atyp <- influence.measures(modelefinal)
levier <- atyp$infmtat[, "hat"]

```

- iii. Pour détecter les valeurs atypiques au sens du levier, on peut taper :

```

seuil.levier <- 2*(8+1)/189
atyp.levier <- ID[levier>seuil.levier] # Liste des individus qui
ont un levier important :
atyp.levier

```

D’autres mesures diagnostiques sont aussi utilisées pour inspecter les valeurs atypiques et déterminer l’influence de ces valeurs sur le modèle de régression réalisé :

- i. La distance de Cook `cooks.distance()`, qui permet de mesurer l’influence de l’observation  $i$  sur l’estimation de l’ensemble des paramètres de régression par le calcul de la valeur  $C_i$ . Une forte valeur de  $C_i$ , indique que la  $i$ ème observation est influente (1 est parfois considéré comme limite). La suppression de cette observation peut entraîner de grosses modifications sur l’équation de la régression.
- ii. La distance de Welsh-Kuh ou Dffits `dffits()`. Elle permet de calculer  $Dffits_i$ .



Une forte valeur de  $|Dffits_i|$  indique une influence de l'observation  $i$  sur l'estimation  $\hat{y}_i$ , et permet donc d'affirmer que cette observation est influente sur les résultats de la régression. Dans la pratique, on considère qu'une observation peut être influente dès que  $|Dffits_i| \geq 2\sqrt{\frac{p+1}{n}}$ .

- iii. La mesure Dfbetas `dfbetas()`. Elle permet de calculer  $Dfbetas_{j,i}$ . Cette quantité mesure l'influence de l'observation  $i$  sur l'estimation du  $j$ ème coefficient. Pour des jeux de données de petite taille ou de taille modérée, une valeur supérieure à 1 semble suspecte. Pour de gros jeux de données, on considérera que l'observation  $i$  est suspecte si  $|Dfbetas_{j,i}| > 2/\sqrt{n}$  pour au moins un  $j$ .

Pour plus de détails sur ces mesures diagnostiques, nous conseillons la lecture de [2], [5] ou [10].

## 2.2 Travail personnel

### Exercice 19. (*Exemple du cours*).

Cet exemple concerne le tableau de petite taille étudié en cours (croisant classe d'âge et diplôme). Cela permet de détailler les calculs que nécessite la mise en oeuvre des notions du cours, et donc de mieux comprendre les règles d'interprétation des résultats des techniques abordée.

1. Ecrire une `data.frame` comportant les données.  

```
data <- data.frame(BEPC = c(15,10,15,40), BAC = c(12,18,5,35), Licence
= c(3,4,8,15), Total = c(30,32,28,90))
rownames(data) <- c("Plus de 50 ans", "Entre 30 et 50 ans", "Moins de 30
ans", "Total")
data
```
2. Donner le tableau de contingence des fréquences.
3. Donner la matrice des profils lignes et celle des profils colonnes.
4. Effectuer un test d'indépendance des deux caractères statistiques.
5. Les deux caractères statistiques sont-ils indépendants pour un seuil  $\alpha = 0.05$  ?

**Exercice 20.** Considérons l'exemple fictif discuté au cours. Pour une population d'effectif  $n = 1000$  on a mesuré les deux variables qualitatives "Couleur des yeux" et "Etat matrimonial". Les résultats sont résumés sous la forme d'un tableau de contingence :

1. Créer une variable tableau à l'aide des commandes :  

```
tableau <- matrix(c(290,410,110,190), ncol=2, byrow=TRUE)
colnames(tableau) <- c("Bleu","Brun")
rownames(tableau) <- c("Celib","Marie")
tableau <- as.table(tableau)
```

2. Afficher le contenu de la variable `tableau`. Afficher une représentation graphique à l'aide de la commande `barplot(tableau)`.

3. Exécuter les commandes suivantes et comprendre leur signification :

```
n <- margin.table(tableau)
m1 <- margin.table(tableau,1)
m2 <- margin.table(tableau,2)
prop.table(tableau)
```

4. Le test du chi-deux :

- (a) Créer un tableau `tab0` à l'aide des commandes :

```
tab0 <- as.array(m1) %*% t(as.array(m2))/n
tab0 <- as.table(tab0)
```

Quelle est sa signification ?


- (b) Exécuter les commandes : `summary(tableau)` et `summary(tab0)`

Explication : **Chisq** donne la valeur de la distance  $\chi^2$  à l'indépendance, **df** est le nombre de degrés de liberté, et **p-value** est la probabilité que le  $\chi^2$  d'un échantillon indépendant dépasse la valeur calculée. En général, on considère les échantillons comme non indépendants si **p-value** est inférieur à un certain seuil 0.05, par exemple.

5. Créer un tableau dans lequel tous les individus aux yeux bleus sont mariés et tous les autres sont célibataires, et effectuer le test du chi-deux.

6. Applications

- (a) Appliquer le test du chi-deux au tableau de contingence suivant,
- (b) Appliquer le test du chi-deux à quelques échantillons statistiques de R, par exemple `HairEyeColor`, `Titanic` et `UCBAdmissions`.

**Exercice 21.** L'objectif de cet exercice est de manipuler les commandes de bases permettant d'effectuer une régression linéaire sous R, mais également de contrôler les résultats obtenus, sélectionner les modèles et effectuer des représentations graphiques .

1. Afin de pouvoir conserver certaines données si on le souhaite, effectuer un changement de répertoire de travail.
2. On va utiliser un jeu de données simple déjà implanté :

```
data()
```

```
data(cars)
```

```
cars
```

Que représente ce jeu de données (attention aux unités) ?

```
names(cars)
```

Comment doit-on appeler les variables ?

```
dim(cars)
```

Quelle est la taille de la matrice ?

---

1. Il est inutile de se dépêcher de taper les commandes du TP en vitesse (et de prendre une avance artificielle sans rien comprendre à leur signification ... surtout pour ceux qui ne viennent qu'irrégulièrement en cours.

`plot(cars)` Est-ce que la représentation graphique est adaptée ?

La régression linéaire multidimensionnelle est obtenue très simplement par la commande `lm`.

?`lm` Noter au passage que `lm` ne sert pas uniquement à la régression linéaire mais permet également d'implémenter d'autres modèles (analyse de la variance (ANOVA) notamment).

Noter aussi qu'il est nécessaire de spécifier le modèle dans la syntaxe de `lm`. Un coup d'oeil aux exemples, qui suivant, permet de comprendre que la formule se met sous la forme :

$$Y \sim X_1 + X_2 + \dots + X_p$$

On y va

```
reg<-lm(dist ~ speed,cars)
```

`reg` Pas terrible. On aurait pu dire bien des choses encore

Faisons connaissance avec l'objet que nous venons de construire

```
attributes(reg)
```

Sa carte d'identité

La fonction `lm` est tellement importante que l'on a créé une classe pour elle toute seule.

On voit également dans la rubrique `names` que l'objet `reg` contient un certain nombre d'informations qui ne sont manifestement pas affichées par la commande `reg`.

`summary(reg)` : C'est nettement mieux.

*Il est crucial de comprendre toutes les rubriques qui apparaissent ici.* Elles constituent le coeur de la régression linéaire.

`anova(reg)` : Comparer avec `summary`

Revenir à `names(reg)`. Affichez les différentes valeurs et, avec l'aide en ligne, dire ce qu'elles représentent. Un certain nombre de ces rubriques ne sera d'aucune utilité : `terms`, `call`, `assign`, `effects`. Par contre `xlevel` sera exploitée lors de l'analyse de variance.

`plot(reg)` 4 graphiques on n'en connaît qu'un, éventuellement deux (les graphiques 1 et 3) mais le 2 et le 4 sont sans doute inconnus.

Le graphique 2 (QQ-plot) permet de vérifier l'hypothèse de normalité des résidus : si les points sont à peu près alignés en se confondant avec la première bissectrice des axes, on peut dire que les résidus suivent une loi normale.

Le graphique 4 (Cook's D) permet de repérer les points influents, c'est-à-dire ceux pour qui la régression linéaire est mal (ou pas) adaptée, parce qu'ils se situent trop loin de la droite de régression. Ces points sont repérés par de grandes valeurs du D de Cook.

On peut tracer désormais la droite de régression

```
plot(cars,pch=20,col='blue') abline(reg=reg,col='red')
```

Recommencer avec `abline(reg$coeff,col='yellow')`

Pour la prévision, on a besoin de la commande `predict`.

- (a) Quelle valeur prédite pour une vitesse de 20 ?
- (b) Donner un intervalle de confiance pour cette valeur (avec les options `confidence` puis `prediction`) ?
- (c) Quelles différences sont notables ?  
La solution est 61.06908.

On passe aux rudiments de la sélection de modèles.

- (a) L'exemple `cars` est-il adapté à la sélection de modèles ?
- (b) Se renseignez sur les commandes `update` et `step`.
- (c) Choisir le jeu de données `cpus` dans la librairie `MASS`.
- (d) Pour avoir une idée globale du comportement des variables les unes par rapport aux autres, quelles commande peut-on utiliser ?

Effectuer la régression de `perf` contre toutes les autres variables quantitatives.

Affiner enfin le modèle en sélectionnant pas ou pas vous-même ou automatiquement les variables pertinentes (Vous pouvez également utiliser la doc sur cette librairie très utile et jeter un coup d'oeil aux fonctions `stepAIC`, `addterm`, `dropterm`).

**Exercice 22.** Les données sont extraites d'un recueil de données issu d'une enquête portant sur une population d'enseignants de collèges. Elles ont été modifiées pour les besoins du TP. La plupart des variables sont explicites. Le salaire est exprimé en euros, l'âge et l'ancienneté en années. Le stress, l'estime de soi et la satisfaction au travail sont mesurés sur des échelles allant de 0 à 50 suivant des techniques appropriées.

1. Importer les données du fichier `TP2_Analyse_de_Donnees.xls` :  

```
library(readxl)
data<- read_excel("C :/VOTRE CHEMIN CHEMIN/TP2_Analyse_de_Donnees")
head(data)
```
2. Donner un résumé descriptif standard des données : `names`, `summary`, ...
3. Examiner les résumés et en particulier celui du salaire.
4. **Croisement qualitatif vs qualitatif**
  - (a) Donner les tableaux de contingences (effectifs, fréquences, pourcentages).
  - (b) Donner différentes représentations graphiques du tableau de contingence des effectifs : `balloonplot`, `barplot`, en utilisant l'option `beside=TRUE` et `beside=FALSE`, `mosaicplot`, ...
  - (c) Donner les distributions marginales et comparer avec les distributions univariées.
  - (d) Donner les distributions conditionnelles. En fait, les distributions conditionnelles permettent de voir comment connaissant la modalité d'une variable se répartissent les modalités de l'autre variable.

- (e) Faire un test de chi-deux (avec `chisq.test`) afin d'apprécier la dépendance, ou non, des variables **Sexe** et **EtatCivil**.
  - (f) Récupérer le tableau correspondant à l'indépendance de la question précédente. Calculer le chi-deux par étapes (Tableau des effectifs théoriques correspondant à l'indépendance des variables, Calcul du coefficient de chi-deux par étapes, ...) et comparer aux résultats obtenus à la question précédente.
  - (g) Tester et conclure de deux manières (à partir du coefficient et à partir de la  $p$ -valeur) la liaison entre les variables **Sexe** et **EtatCivil**. Conclure.
5. **Croisement quantitatif vs qualitatif.** On s'intéresse au croisement **Stress** vs **EtatCivil**. On va déterminer s'il existe une relation ou non entre le stress et l'état civil des enseignants interrogés.
- (a) Donner le résumé standard de la variable **Stress**.
  - (b) Représenter et commenter le boxplot de la variable **Stress**.
  - (c) Constituer 5 classes, avec la commande `Nclasse`, de mêmes amplitudes pour la variable **Stress**.
  - (d) Donner les tableaux de contingences de **Stress** vs **EtatCivil** : en effectifs, en fréquences, en pourcentages (arrondir au 100ème près).
  - (e) Donner les boxplots de la variable **Stress** en fonction de la variable **EtatCivil**. Que remarque-t-on ?
  - (f) Donner les histogrammes de la variable **Stress** en fonction de la variable **EtatCivil** (package `lattice`).
  - (g) Donner les résumés de la variable **Stress** en fonction de la variable **EtatCivil**.
  - (h) Calculer le rapport de corrélation `net2` par étapes (Calcul de la variance intra-groupes, Calcul de la variance inter-groupes, le coefficient  $\eta^2$ ).
  - (i) Comparer avec le  $F$  de Fisher (Calcul de l'indicateur de Fisher  $F(\text{Stress}/\text{EtatCivil})$ , Calcul du seuil de signicativité de  $F(\text{Stress}/\text{EtatCivil})$ ). Conclure
6. **Croisement quantitatif vs quantitatif.** On s'intéresse au croisement **Age** vs **Satisfaction**. On va déterminer s'il existe une relation ou non entre l'âge et la satisfaction au travail des enseignants interrogés.
- (a) Donner le résumé standard de la variable **Satisfaction**.
  - (b) Représenter et commenter le boxplot de la variable **Satisfaction**. Commenter.
  - (c) Détecter graphiquement les "outliers" en dehors d'une bande de 2 écarts-types autour de la moyenne. On pourra utiliser la fonction `identify()`
  - (d) Donner le résumé standard de la variable **Age**.
  - (e) Représenter et commenter le boxplot de la variable **Age**. Commenter.
  - (f) Détecter graphiquement les "outliers" en dehors d'une bande de 2 écarts-types autour de la moyenne

- (g) Constituer 5 classes de même amplitude de la variable **Satisfaction**.
  - (h) Constituer 4 classes de même amplitude de la variable **Age**.
  - (i) Donner les tableaux de contingences de **Age** vs **Satisfaction** : en effectifs, en fréquences, en pourcentages (arrondir au 100ème près). Utiliser **Balloonplot** pour voir le tableau de contingence en effectif de **Age** vs **Satisfaction**.
  - (j) Représenter le nuage de points de la variable **Satisfaction** en fonction de la variable **Age**.
  - (k) Détecter dynamiquement un “intrus” et afficher le profil de cet “intrus”. Que remarque-t-on ?
  - (l) Réexaminer plus attentivement les résumés réalisés en début de traitement. Que conclure ?
  - (m) Représenter le nuage de points de la variable **Satisfaction** en fonction de la variable **Age** en distinguant les femmes et les hommes. On pourra utiliser la fonction `split()`
  - (n) Calculer par étapes la covariance entre la variable **Age** et la variable **Satisfaction**. Comparer avec celle donnée par R.
  - (o) Calculer par étapes la corrélation entre la variable **Age** et la variable **Satisfaction**. Comparer avec celle donnée par R.
  - (p) Donner la matrice de corrélation entre les variables numériques (sauf le nombre d’enfants).
  - (q) Représenter cette matrice graphiquement et commenter. On pourra utiliser la fonction `scatterplotMatrix`
7. Faire un rapport sur cette étude.