Table des matières

Chapitre 3

Régression linéaire multiple

De nombreux problèmes de régression impliquent plusieurs variables exogènes. De telles approches sont appelées *modèles de régression multiple*. La régression linéaire multiple reste une des méthodes statistiques les plus appliquées.

Le modèle de régression linéaire multiple permet de mettre en relation une variable endogène notée Y avec plusieurs variables exogènes nommées $(x_1, x_2, ..., x_p)$. Par exemple, on peut s'intéresser à l'espérance de vie d'un individu en fonction de certaines charactéristiques. Dans cet exemple, l'espérance de vie constitue la variable endogène (Y) et elle peut dépendre de diverses variables telles l'âge (x_1), le poids (x_2), la taille (x_3), le revenu (x_4), le sexe (x_5), la consommation de tabac (x_6), etc.

On constate dans cet exemple que plusieurs types de variables peuvent être utilisées en régression linéaire multiple, à savoir des variables :

- dichotomiques (exp. présence/ absence)
- discrètes (exp. une classe de revenu)
- continues (exp. le poids)
- qualitatives (exp. le sexe).

3.1 Modèle et notation

Le modèle de régression linéaire multiple constitue une généralisation du modèle de régression linéaire simple lorsqu'on considère plusieurs variables explicatives (exogènes).

L'équation du modèle de régression linéaire multiple s'exprime de la manière suivante

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{3.1}$$

où

- $Y_i, i = 1 \ldots, n$, représentent la variable endogène et sont considérés comme aléatoires.

- $x_{ij}, i = 1\ldots, n; j = 1\ldots, p$, dénotent les variables exogènes. Ce sont des nombres connus, non aléatoires. Il est possible de multiplier la variable β_0 par $x_{i0} = 1, i = 1\ldots, n$. Dans ce cas, β_0 represente la constante, appelée également *intercept* en anglais.
- β_0 et $\beta_j, j = 1, \ldots, p$ notent les paramètres du modèle qui sont inconnus et par conséquent doivent être estimés.
- $\epsilon_i, i = 1\ldots, n$ représentent les termes d'erreur et sont des variables aléatoires inconnues.

Il est à souligner que le nombre d'observations doit être supérieur au nombre de paramètres afin d'être en mesure d'estimer ces derniers adéquatement. On veut donc en général que $n > p + 1$.

Les quatre postulats du modèle sont identiques à ceux de la régression linéaire simple, à savoir :

$\mathcal{H}_1$. $\mathbb{E}[\varepsilon_i] = 0$, pour $i = 1, \ldots, n$.

$\mathcal{H}_2$. $\mathbb{V}\text{ar}[\varepsilon_i] = \sigma^2$, pour $i = 1, \ldots, n$.

$\mathcal{H}_3$. $\mathbb{C}\text{ov}[\varepsilon_i, \epsilon_j] = 0$ pour $i \neq j$.

$\mathcal{H}_4$. ε_i est distribué selon une loi normale pour $i = 1, \ldots, n$.

Dans de nombreux problèmes des interactions existent entre les variables explicatives. Ainsi, si on considère par exemple deux variables exogènes, on a que

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Un tel type de modélisation s'insère parfaitement dans le cadre de la régression multiple. Les variables d'interaction sont en fait des produits de variables connues et sont par conséquent connues. Dans l'exemple ci-dessus, la troisième variable explicative x_3 est tout simplement le produit $x_1 x_2$ et nous retrouvons le modèle de régression linéaire multiple défini à l'équation (3.1). D'autres extensions peuvent être considérées comme le modèle de régression polynômiale. Par exemple le modèle polynômial cubique à une variable indépendante s'écrit comme

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i3}^3 + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Il est possible d'écrire ce modèle sous la formulation de l'équation (3.1) en posant $x_2 = x_1^2$ et $x_3 = x_1^3$.

Dans sa version matricielle, le modèle s'écrit comme suit :

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{n\times 1} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}}_{n\times(p+1)} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{(p+1)\times 1} + \underbrace{\begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{n\times 1}$$

soit

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p'} \boldsymbol{\beta}_{p'\times 1} + \boldsymbol{\varepsilon}_{n\times 1} \tag{3.2}$$

où

- $\boldsymbol{Y}$ désigne le vecteur contenant la variable endogène (dimension $n \times 1$)
- $\boldsymbol{X}$ note la matrice d'incidence contenant les variables exogènes (dimension $n \times p'$)
- $\boldsymbol{\beta}$ est le vecteur des coefficients de la régression (dimension $p' \times 1$)
- $\boldsymbol{\varepsilon}$ dénote le vecteur des termes d'erreur (dimension $n \times 1$)
- n note le nombre d'observations
- p indique le nombre de variables exogènes
- $p' = p + 1$ désigne le nombre de variables exogènes + l'ordonnée à l'origine.

3.1.1 Interprétation des paramètres

Dans le modèle de régression linéaire multiple défini à l'équation (3.1), et sous le postulat $\mathcal{H}_1$, β_0 représente la valeur moyenne de Y_i lorsque $x_{i1}, \ldots, x_{ip}$ prennent toutes simultanément la valeur 0. Le paramètre β_j représente la variation dans la moyenne de Y_i lorsque x_j augmente d'une unité et que la valeur de toutes les autres variables exogènes reste constante. La figure 3.1 montre une représentation géométrique du modèle de régression linéaire multiple. Dans le modèle de régression linéaire simple les points Y sont situés autour d'une droite, alors que le modèle de régression linéaire multiple disperse la valeur de la variable endogène Y autour d'un hyperplan. Les coefficients $\beta_j, j = 1, \ldots, p$ représentent la pente de l'hyperplan dans la direction de la variable exogène correspondante.

3.2 Estimation et propriétés des paramètres

Comme dans le cas de la régression linéaire simple, on adopte à nouveau le principe des moindres carrés de Legendre. On cherche par conséquent le vecteur $\boldsymbol{\beta}$ qui minimise la somme des carrés des résidus ($\sum_{i=1}^{n} \hat{\epsilon}_i^2$).

Le vecteur des résidus s'exprime comme

$$\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_n)^\top = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = f(\hat{\boldsymbol{\beta}})$$

Dans le cas où la matrice $\mathbf{X}^\top \boldsymbol{X}$ est inversible (de plein rang) on obtient que

$$\underbrace{\hat{\boldsymbol{\beta}}}_{(p+1)\times 1} = \underbrace{(\mathbf{X}^\top \boldsymbol{X})^{-1}}_{(p+1)\times(p+1)} \underbrace{\mathbf{X}^\top}_{(p+1)\times n} \underbrace{\boldsymbol{Y}}_{n\times 1}. \tag{3.3}$$

Si $\mathbf{X}^\top \boldsymbol{X}$ n'est pas inversible, une ligne de cette matrice est une combinaison linéaire des autres et il est nécessaire de retirer cette variable.

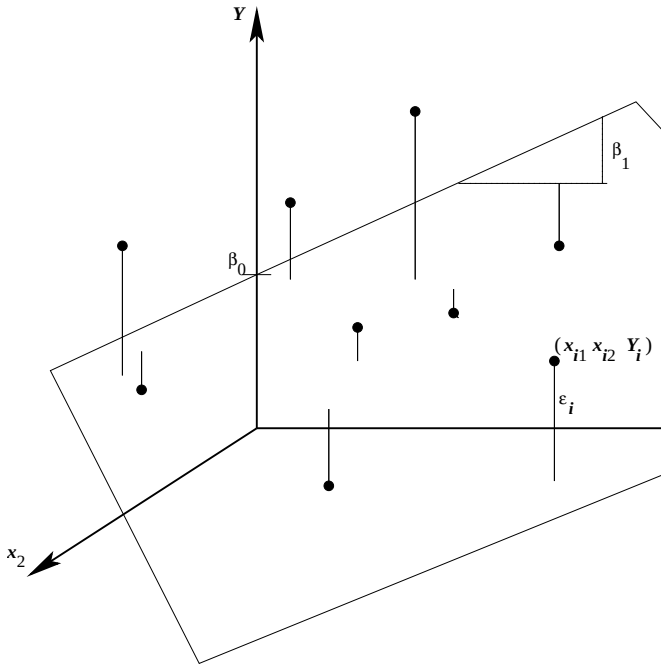


FIGURE 3.1: Représentation géométrique du modèle de régression linéaire multiple.

L'estimateur des moindres carrés est sans biais et sa distribution est une normale multivariée (voir annexe **??** pour un rappel sur la loi normale multivariée)

$$\hat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right).$$

3.3 Distribution de $\hat{Y}$ et $\hat{\varepsilon}$

Voir notes en classe

3.4 Interprétation géométrique de la régression

Dans le modèle de régression linéaire multiple, on cherche à écrire $\boldsymbol{Y}$ comme une combinaison linéaire des colonnes de $\boldsymbol{X}$, plus une erreur $\boldsymbol{\varepsilon}$. Autrement dit, on veut projeter $\boldsymbol{Y}$ dans le sous-espace vectoriel engendré par $\boldsymbol{X}$. Si $\boldsymbol{X}$ est de plein rang (inversible), cet espace aura $p' = p + 1$ dimensions. La matrice chapeau

$$H = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

est la matrice de projection de $\boldsymbol{Y}$ dans l'espace de $\boldsymbol{X}$. Le vecteur résiduel défini par $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\mathbb{I}_n - \boldsymbol{H})\boldsymbol{Y}$ est perpendiculaire à l'espace engendré par $\boldsymbol{X}$. La figure 3.2 représente géométriquement le cas où l'on est en présence de deux variables ($p = 2$).

Si $\boldsymbol{X}$ n'est pas de plein rang, certaines colonnes de $\boldsymbol{X}$ (certaines variables) sont linéairement dépendantes. Dans ce cas, il suffit de retirer une ou plusieurs variables.

Le vecteur des résidus est la projection de $\boldsymbol{Y}$ dans l'espace perpendiculaire à $\boldsymbol{X}$. Toujours en considérant $\boldsymbol{X}$ de plein rang, cet espace est de dimension $(n - p)$.

L'angle droit entre le vecteur des valeurs prédites ($\hat{\boldsymbol{Y}}$) et le vecteur des résidus $\hat{\boldsymbol{\varepsilon}}$ découle du fait que l'on ait choisi $\hat{\boldsymbol{\beta}}$ de façon à minimiser la somme des résidus au carré, c'est-à-dire $\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}$. L'orthogonalité de $\hat{\boldsymbol{Y}}$ et $\hat{\boldsymbol{\varepsilon}}$ revient à dire que $\mathbb{C}\text{ov}[\hat{\boldsymbol{\varepsilon}}, \hat{\boldsymbol{Y}}] = 0$. En effet, des vecteurs orthogonaux sont des vecteurs non corrélés. Dans le cas gaussien, ceci implique de plus que ces vecteurs sont indépendants.

3.5 Estimation sans biais de σ^2 et tests d'hypothèses sur les paramètres

On a vu que du postulat de normalité $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\mathbf{I}_n)$, découle que $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$. Si on s'intéresse à un β_i en particulier, on a alors que

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{v_{ii}}} \sim N(0, 1)$$

où v_{ii} est l'élément (i, i) de la matrice $(\mathbf{X}^\top\mathbf{X})^{-1}$. Si on connaît σ, il est alors possible d'utiliser ce résultat pour construire un intervalle de confiance sur β_i ou pour faire des test d'hypothèses. Or en pratique, σ est

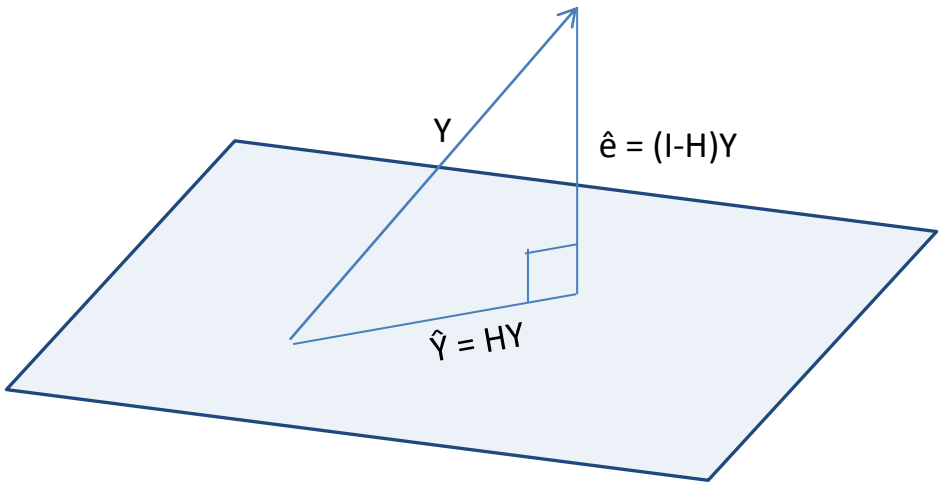


FIGURE 3.2: Représentation géométrique de la régression dans le cas où il y a deux variables explicatives.

généralement inconnu. On l'estimera alors par

$$\hat{\sigma^2} = \frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{n - p'}$$

où $p' = p + 1$ [1].

Afin de comprendre la division par $(n - p')$, il faut connaître le Théorème suivant.

Théorème 3.1. *Supposons que le vecteur aléatoire* $\mathbf{Z}$ *suit une loi normale multivariée centrée réduite* $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ *et que* $\mathbf{W} = \mathbf{AZ}$*, alors* $\mathbf{W}^\top(\mathbf{AA}^\top)^{-1}\mathbf{W} \sim \chi^2_l$ *où* l *est le rang de* $\mathbf{A}$*.*

1. Attention, p' doit être plus petit que le nombre d'observations n.

On sait que $\hat{\boldsymbol{\varepsilon}} \sim N_n\{0, \sigma^2(\mathbf{I}_n - \mathbf{H})\}$ et par conséquent que $\hat{\boldsymbol{\varepsilon}}/\sigma \sim N_n\{0, (\mathbf{I}_n - \mathbf{H})\}$. En posant $\hat{\boldsymbol{\varepsilon}}/\sigma = \mathbf{AZ}$, où $\mathbf{Z} \sim N_n(0, \mathbf{I}_n)$ et $\mathbf{A} = (\mathbf{I}_n - \mathbf{H})$ et en utilisant le Théorème, on trouve

$$\begin{aligned}
\frac{\hat{\boldsymbol{\varepsilon}}}{\sigma}^\top (\mathbf{A}^\top \mathbf{A})^{-1} \frac{\hat{\boldsymbol{\varepsilon}}}{\sigma} &\sim \chi^2_l \\
\frac{\hat{\boldsymbol{\varepsilon}}}{\sigma}^\top \{(\mathbf{I}_n - \mathbf{H})^\top (\mathbf{I}_n - \mathbf{H})\}^{-1} \frac{\hat{\boldsymbol{\varepsilon}}}{\sigma} &\sim \chi^2_l \\
\frac{\hat{\boldsymbol{\varepsilon}}}{\sigma}^\top \mathbf{I}_n^{-1} \frac{\hat{\boldsymbol{\varepsilon}}}{\sigma} &\sim \chi^2_l \\
\frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{\sigma^2} &\sim \chi^2_l
\end{aligned}$$

où $\mathbf{H}$ désigne la matrice chapeau et l est le rang de $(\mathbf{I}_n - \mathbf{H})$. On a vu que le rang de $(\mathbf{I}_n - \mathbf{H}) = n - p'$. On a donc que

$$\frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{\sigma^2} \sim \chi^2_{n-p'}. \tag{3.4}$$

Puisque l'espérance d'une variable aléatoire distribuée selon une loi chi-carré à k degrés de liberté (χ^2_k) est k, on peut en déduire que

$$\begin{aligned}
\mathbb{E}\left[\frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{\sigma^2}\right] &= n - p' \\
\mathbb{E}\left[\frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n - p'}\right] &= \sigma^2.
\end{aligned}$$

On en conclut que $\frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n-p'}$ est un estimateur sans biais de σ^2. De plus, cet estimateur est indépendant de $\hat{\boldsymbol{\beta}}$. Cela découle du Théorème suivant :

Théorème 3.2. *Soit $W = \mathbf{BY}$ et $U = \mathbf{Y}^\top \mathbf{AY}$, avec $\mathbf{Y} \sim \mathcal{N}(\mu, \mathbf{V})$. Alors, W est indépendant de U si et seulement si $\mathbf{BVA} = 0$.*

Par conséquent, on peut prouver que l'estimateur de σ^2 est indépendant de $\hat{\boldsymbol{\beta}}$. Alors, on a

$$\frac{\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{v_{ii}}}}{\sqrt{\frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{\sigma^2(n-p')}}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-p'}/(n-p')}} \sim \mathcal{T}_{n-p'} \tag{3.5}$$

où $\mathcal{T}_{n-p'}$ note la loi de Student à $n-p'$ degrés de liberté. On peut alors utiliser (3.7) pour construire un intervalle de confiance pour $\boldsymbol{\beta}$ ou pour effectuer des tests d'hypothèses. Ainsi l'intervalle de confiance de niveau $(1-\kappa)$ pour β_i s'exprime comme

$$\left[\hat{\beta}_i \pm t_{n-p'}(1-\kappa/2)\sqrt{s^2 v_{ii}}\right] \tag{3.6}$$

3.6 Prévision

Supposons que l'on s'intéresse à une combinaison linéaire de $\boldsymbol{\beta}$, que nous notons θ.

$$\theta = \sum_{i=0}^{p} a_i \beta_i = \mathbf{a}^\top \boldsymbol{\beta}$$

$$a = [a_0 \ a_1 \ ... \ a_p]^\top.$$

On sait que $\mathbf{a}^\top\hat{\boldsymbol{\beta}} \sim N(\mathbf{a}^\top\boldsymbol{\beta}, \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{a})$. Cet estimateur n'est pas biaisé. Selon le théorème de Gauss-Markov, il est le meilleur estimateur non biaisé.

On a que

$$\frac{\frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{v_{ii}}}}{\sqrt{\frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{\sigma^2(n-p')}}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-p'}/(n-p')}} \sim \mathcal{T}_{n-p'}, \tag{3.7}$$

où v_{ii} est l'élément (i,i) de la matrice $(\mathbf{X}^\top\mathbf{X})^{-1}$. Par conséquent, on peut faire un test d'hypothèses sur le paramètre β_i. Si on désre tester $H_0 : \beta_i = 0$, alors la statistique

$$t_{obs,i} = \frac{\hat{\beta}_i\sqrt{n-p'}}{\sqrt{v_{ii}\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}}$$

suit une loi de Student avec $n-p'$ degrés de liberté sous H_0. On rejette l'hypothèse nulle avec un niveau de confiance de $(1-\kappa)$ si $|t_{obs,i}| > t_{n-p'}(1-\kappa/2)$.

L'intervalle de confiance de niveau $(1-\kappa)$ pour β_i s'exprime comme

$$\left[\hat{\beta}_i \pm t_{n-p'}(\kappa/2)\sqrt{s^2 v_{ii}}\right]. \tag{3.8}$$

3.6.1 Théorème de Gauss-Markov

Théorème 3.3. *Le théorème de Gauss-Markov stipule que* $\mathbf{a}^\top\hat{\boldsymbol{\beta}} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ *est le meilleur estimateur non biaisé de* $\mathbf{a}^\top\boldsymbol{\beta}$*, c'est-à-dire qu'il est l'estimateur sans biais dont la variance est la plus petite.*

Démonstration
Supposons que $\mathbf{c}^\top \mathbf{Y}$ est un autre estimateur linéaire quelconque sans biais de $\mathbf{a}^\top \boldsymbol{\beta}$. Montrons que $\mathbb{V}\text{ar}[\mathbf{cY}] \geq \mathbb{V}\text{ar}[\mathbf{a}^\top \hat{\boldsymbol{\beta}}]$. On a que

$$\mathbb{E}[\mathbf{c}^\top \mathbf{Y}] = \mathbf{c}^\top \mathbb{E}[Y] = \mathbf{c}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{a}^\top \boldsymbol{\beta},$$

la dernière égalité provenant du fait qu'on veut un estimateur sans biais. On a donc que $\mathbf{a}^\top = \mathbf{c}^\top \mathbf{X}$. Si on calcule la variance de $\mathbf{c}^\top \mathbf{Y}$, on obtient

$$\mathbb{V}\text{ar}[\mathbf{c}^\top \mathbf{Y}] = \mathbf{c}^\top \mathbb{V}\text{ar}[\mathbf{Y}]\mathbf{c} = \mathbf{c}^\top \sigma^2 I_n \mathbf{c} = \sigma^2 \mathbf{c}^\top \mathbf{c}.$$

La variance de $\mathbf{a}^\top \hat{\boldsymbol{\beta}}$ est

$$\mathbb{V}\text{ar}[\mathbf{a}^\top \hat{\boldsymbol{\beta}}] = \mathbf{a}^\top \mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}]\mathbf{a} = \sigma^2 \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{a}.$$

Si on calcule $\mathbb{V}\text{ar}[\mathbf{c}^\top \mathbf{Y}] - \mathbb{V}\text{ar}[\mathbf{a}^\top \hat{\boldsymbol{\beta}}]$, on obtient un terme nécessairement positif. En effet

$$\begin{aligned}
\mathbb{V}\text{ar}[\mathbf{c}^\top \mathbf{Y}] - \mathbb{V}\text{ar}[\mathbf{a}^\top \hat{\boldsymbol{\beta}}] &= \sigma^2(\mathbf{c}^\top \mathbf{c} - \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{a}) \\
&= \sigma^2\{\mathbf{c}^\top \mathbf{c} - \mathbf{c}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{c}\} \\
&= \sigma^2 \mathbf{c}^\top [I_n - \{\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\}]\mathbf{c} \\
&= \sigma^2 \mathbf{c}^\top (I_n - \mathbf{H})\mathbf{c},
\end{aligned}$$

où $\mathbf{H}$ désigne la matrice chapeau. D'après les propriétés de cette matrice, $I_n - \mathbf{H}$ est une matrice semi-définie positive, ce qui implique que $\sigma^2 \mathbf{c}^\top (I_n - \mathbf{H})\mathbf{c} \geq 0$. Donc $\mathbb{V}\text{ar}[\mathbf{c}^\top \mathbf{Y}] \geq \mathbb{V}\text{ar}[\mathbf{a}^\top \hat{\boldsymbol{\beta}}]$.

□

Supposons que l'on s'intéresse à une combinaison linéaire de $\boldsymbol{\beta}$, que nous notons θ.

$$\theta = \sum_{i=0}^{p} a_i \beta_i = \mathbf{a}^\top \boldsymbol{\beta}$$

$$a = [a_0 \ a_1 \ ... \ a_p]^\top.$$

On sait que $\mathbf{a}^\top \hat{\boldsymbol{\beta}} \sim N(\mathbf{a}^\top \boldsymbol{\beta}, \sigma^2 \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{a})$. Cet estimateur n'est pas biaisé. Selon le théorème de Gauss-Markov, il est le meilleur estimateur non biaisé.

3.6.2 Tests d'hypothèse pour $\mathbf{a}^\top \boldsymbol{\beta}$

Comme on ne connaît pas σ^2, on l'estime par $s^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n-p'}$ et on trouve que

$$\frac{\mathbf{a}^\top \hat{\boldsymbol{\beta}} - \theta}{\sqrt{s^2 \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{a}}} \sim \mathcal{T}_{n-p'}. \tag{3.9}$$

On peut alors utiliser ce résultat pour construire un intervalle de confiance pour θ ou pour tester $H_0 : \theta = \theta^*$.

3.6.3 Inférence sur la valeur moyenne $\mathbb{E}[Y|\mathbf{x}^*]$

On désire estimer la valeur moyenne de la variable endogène pour une combinaison de valeurs ${\mathbf{x}^*}^\top = (x_1^*, \ldots, x_p^*)$ spécifiées des variables exogènes. Un intervalle de confiance à $100(1-\kappa)\%$ pour $\mathbb{E}[Y|\mathbf{x}^*]$ est donné par

$$\left[{\mathbf{x}^*}^\top \hat{\boldsymbol{\beta}} \pm t_{n-p'}(\kappa/2) \sqrt{s^2 {\mathbf{x}^*}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}^*} \right]. \tag{3.10}$$

3.6.4 Inférence sur une prévision de Y étant donné $\mathbf{x}^*$

Nous cherchons maintenant à prédire la valeur même d'une réalisation de la variable endogène pour une combinaison de valeurs $\mathbf{x}^*$ fixées des variables exogènes.

En fait on cherche à estimer $Y = {\mathbf{x}^*}^\top \boldsymbol{\beta} + \varepsilon$. Le théorème de Gauss-Markov recommande d'estimer ${\mathbf{x}^*}^\top \boldsymbol{\beta}$ avec ${\mathbf{x}^*}^\top \hat{\boldsymbol{\beta}}$. Comme $\mathbb{E}[\varepsilon] = 0$, notre meilleur estimé d'une réalisation de ε est 0. Notre estimateur ponctuel de Y à une valeur $\mathbf{x}^*$ donnée sera donc ${\mathbf{x}^*}^\top \hat{\boldsymbol{\beta}} + 0 = {\mathbf{x}^*}^\top \hat{\boldsymbol{\beta}}$. Ensuite, on a que

$$\frac{{\mathbf{x}^*}^\top \hat{\boldsymbol{\beta}} - ({\mathbf{x}^*}^\top \boldsymbol{\beta} + \varepsilon)}{\sqrt{s^2 \left(1 + {\mathbf{x}^*}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}^*\right)}} \sim \mathcal{T}_{n-p'}. \tag{3.11}$$

Ceci nous conduit à l'intervalle de confiance suivant :

Un intervalle de confiance à $100(1-\kappa)\%$ pour Y étant donnée la combinaison des variables exogènes $\mathbf{x}^*$ est donné par

$$\left[{\mathbf{x}^*}^\top \hat{\boldsymbol{\beta}} \pm t_{n-p'}(\kappa/2) \sqrt{s^2 \left(1 + {\mathbf{x}^*}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}^*\right)} \right]. \tag{3.12}$$

3.7 Analyse de la variance

En général, les n valeurs observées, $Y_1, \ldots, Y_n$, de la variable endogène ne sont pas toutes égales, c'est-à-dire que nous observons de la variabilité dans la valeur de la variable endogène. Un des buts d'une analyse de régression consiste à expliquer la plus grande partie possible de cette variabilité à partir des valeurs des variables exogènes. Ainsi, en considérant la décomposition de la variabilité dans la valeur de la variable endogène suivante,

$$\begin{pmatrix} \text{Variabilité de} \\ Y_1, \ldots, Y_n \end{pmatrix} = \begin{pmatrix} \text{Variabilité expliquée par} \\ \text{la variabilité de } \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \end{pmatrix} + \begin{pmatrix} \text{Variabilité inexpliquée} \\ \text{(fluctuation aléatoire)} \end{pmatrix}, \tag{3.13}$$

on voudrait un modèle pour lequel une grande portion de la variabilité est expliquée par la variabilité dans les variables exogènes. Nous pouvons effectivement faire la décomposition proposée en (3.13) :

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i+\hat{Y}_i-\bar{Y})^2\\ &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+2\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})+\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2\\ &= \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2.\end{aligned}$$

On peut refaire les calculs précédents sous forme matricielle et définir les sommes de carrés suivantes :

$$\begin{aligned}SST &= \sum_{i=1}^{n}(Y_i-\bar{Y})^2=\boldsymbol{Y}^\top\boldsymbol{Y}-n\bar{Y}^2\\ SSR &= \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2=\hat{\boldsymbol{Y}}^\top\hat{\boldsymbol{Y}}-n\bar{Y}^2=\hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}-n\bar{Y}^2\\ SSE &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=(\boldsymbol{Y}-\hat{\boldsymbol{Y}})^\top(\boldsymbol{Y}-\hat{\boldsymbol{Y}})=\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}.\end{aligned}$$

On peut donc exprimer (3.13) sous les formes suivantes :

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \tag{3.14}$$

$$\boldsymbol{Y}^\top\boldsymbol{Y}-n\bar{Y}^2 = \left(\hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}-n\bar{Y}^2\right)+\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} \tag{3.15}$$

$$SST = SSR+SSE. \tag{3.16}$$

Les équations (3.14)-(3.16) montrent de façon formelle comment la variabilité dans la valeur des Y_i peut être décomposée en la somme de la variabilitée expliquée par la variabilité dans les variables exogènes et de la variabilité due à la fluctuation aléatoire.

SST : S'il n'y a aucune variabilité dans la valeur des Y_i, alors ces valeurs sont toutes égales à $\bar{Y}$ et $SST = 0$. Plus il y aura de variabilité dans les Y_i, plus leurs valeurs seront éloignées de la moyenne et, donc, plus SST sera grande.

SSR : Si la combinaison des valeurs des variables exogènes est la même pour les n observations, alors toutes les valeurs ajustées $\hat{Y}_i$ prennent la valeur $\bar{Y}$ (à voir en exercice) et donc $SSR = 0$.

SSE : Comme la moyenne des $\hat{\boldsymbol{\varepsilon}}_i$ est toujours 0 (à voir en exercice), alors s'il n'y a aucune variabilité dans les $\hat{\boldsymbol{\varepsilon}}_i$ ces derniers prennent tous la valeur 0 et $SSE = 0$. Ceci voudrait dire que le modèle de régression explique entièrement la valeur des Y_i et qu'il n'y a pas de fluctuation aléatoire.

Pour chaque somme de carrés, on associe un nombre de degrés de liberté. Les degrés de liberté constituent en fait le nombre de termes indépendants dont nous devons connaître la valeur afin de pouvoir calculer la somme de carrés. Par exemple SST possède $n-1$ degrés de liberté, puisque seulement $n-1$ des termes $(Y_1 - \bar{Y}), \ldots, (Y_n - \bar{Y})$ sont indépendants (on sait que leur somme est 0, donc si on connait la valeur de $n-1$ d'entre eux, on peut calculer la valeur du $n$ième).

Les sommes de carrés et leurs degrés de liberté sont en général résumés dans une table d'analyse de la variance (table ANOVA), tel que montré dans le tableau 3.1.

TABLE 3.1: Tableau d'analyse de la variance dans le cas de la régression linéaire multiple présentant la source, le nombre de degrés de liberté, la somme des carrés, le carré moyen et la statistique F de Fisher.

Source	Degrés de liberté	Sommes des carrés	Carrés moyens	F
Modèle	$ddl_1 = p$	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$MSR = SSR/ddl_1$	MSR/MSE
Erreur résiduelle	$ddl_2 = n - p'$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = SSE/ddl_2$	
Totale (corrigée)	$n - 1 = p + n - p'$	$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$		

3.8 Test F de l'importance globale de la régression

Un test d'hypothèse important en régression consiste à tester si au moins une des variables exogènes explique une partie significative de la variabilité dans les Y_i. Ceci revient donc à tester si les données démontrent de l'évidence contre l'hypothèse nulle H_0 : les variables exogènes n'expliquent rien. Mathématiquement, une variable exogène n'explique en rien la valeur de Y_i si le coefficient de régression correspondant prend la valeur 0. On veut donc tester

$$\begin{aligned} H_0 : &\quad \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ \text{contre } H_1 : &\quad \text{au moins un des coefficients n'est pas nul.} \end{aligned} \tag{3.17}$$

Sous H_0, le modèle de régression ne devrait pas expliquer la variabilité dans les Y_i et donc le ratio SSR/SSE devrait prendre une petite valeur. Par contre sous H_1, le modèle de régression devrait expliquer une partie de la variabilité des Y_i et donc le ratio SSR/SSE devrait prendre une grande valeur. Afin de savoir si la valeur du ratio est « petite »ou « grande », on standardise le ratio pour obtenir la statistique F de la table d'analyse de la variance :

$$F = \frac{SSR/p}{SSE/(n-p')} = \frac{SSE/p}{s^2} = MSR/MSE.$$

Sous H_0, cette statistique suit une loi $\mathcal{F}$ de Fisher-Snedecor avec p degrés de liberté au numérateur et $n - p'$ degrés de liberté au dénominateur. On rejette donc H_0 au niveau κ (c'est-à-dire que les données montrent que le modèle n'est pas complètement inutile, ou qu'il existe relation entre la variable endogène et

au moins une des variables exogènes) lorsque la statistique $\mathcal{F}$ est supérieure ou égale au quantile supérieur $F_{p,n-p'}(1-\kappa)$.

3.9 Test de réduction du modèle (Test F partiel)

Il est plutôt rare en pratique que le test $\mathcal{F}$ de Fisher global ne rejette pas l'hypothèse H_0 donnée par (3.17) que le modèle de régression est totalement inutile. Cependant, on veut souvent tester si le modèle peut être réduit. Cela revient à tester si un sous-modèle plus simple explique une partie suffisamment grande de la variabilité dans les Y_i pour qu'il ne soit pas nécessaire d'utiliser le modèle plus complexe.

Le principe de la somme de carrés résiduels additionnelle permet de tester cette hypothèse de façon formelle. L'idée sous-jacente est simple : si les termes qui sont exclus du modèle plus simple expliquent une partie importante de la variabilité dans les Y_i, alors la variabilité due à la fluctuation aléatoire (SSE) apparaitra beaucoup plus importante dans le modèle simple que dans le modèle complet. Il s'agit donc de mesurer si la différence entre les sommes de carrés résiduels des deux modèles est faible ou élevée.

Cette idée se traduit mathématiquement de la manière suivante : supposons le modèle de régression multiple standard

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

On désire tester si un modèle comprenant seulement $k < p$ variables exogènes suffit à expliquer la variabilité dans les Y_i. Si c'est le cas $p-k$ variables exogènes sont superflues. Par souci de simplicité, on suppose que les k variables exogènes en question sont $x_1, \ldots, x_k$. Alors on veut tester

$$\begin{aligned} H_0 :\ Y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \\ H_1 :\ Y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \beta_{k+1} x_{i,k+1} + \cdots + \beta_p x_{ip} + \varepsilon_i. \end{aligned} \tag{3.18}$$

Le modèle donné par H_0 est appelé modèle réduit, alors que l'équation présentée en H_1 se nomme modèle complet. Soit $SSE^{(0)}$ la somme des carrés résiduels obtenue avec le modèle réduit et $SSE^{(1)}$ la somme des carrés résiduels obtenue avec le modèle complet. Alors il est toujours vrai que $SSE^{(1)} \leq SSE^{(0)}$. Cependant, sous H_0, la différence entre $SSE^{(0)}$ et $SSE^{(1)}$ sera faible, alors que cette même différence sera élevée sous H_1. Encore une fois, on standardise le rapport $(SSE^{(0)} - SSE^{(1)})/SSE^{(1)}$ afin d'obtenir une distribution connue qui nous permettra d'identifier si une valeur est « faible »ou « élevée ». On obtient

$$F = \frac{(SSE^{(0)} - SSE^{(1)})/\Delta_{dl}}{SSE^{(1)}/(n-p')} = \frac{SSE^{(0)} - SSE^{(1)}}{\Delta_{dl} s^2_{H_1}}, \tag{3.19}$$

où Δ_{dl} note la différence entre les degrés de liberté de $SSE^{(0)}$ et les degrés de liberté de $SSE^{(1)}$. Sous l'hypothèse nulle H_0 donnée par (3.18), la statistique F en (3.19) suit une distribution $\mathcal{F}$ de Fisher avec Δ_{dl} degrés de liberté au numérateur et $n-p'$ degrés de liberté au dénominateur. On rejette donc H_0 au niveau de confiance κ lorsque $F \geq F_{\Delta_{dl},n-p'}(1-\kappa)$. Il est à noter que $n-p'$ réfère au nombre de degrés de liberté de $SSE^{(1)}$, et que $s^2_{H_1}$ réfère au carré moyen résiduel sous H_1. Comme il arrive parfois que le « modèle complet »constitue un modèle qui n'inclut pas toutes les p variables exogènes, il faut ajuster la

formule (3.19) en conséquence. C'est pourquoi il est probablement mieux de retenir la forme à droite de la seconde égalité en (3.19).

On peut calculer Δ_{dl} comme suit :

$$\Delta_{dl} = (\text{nombre de paramètres du modèle } H_1) - (\text{nombre de paramètres du modèle } H_0).$$

Ainsi, dans le cas présenté en (3.18), on aurait
- Paramètres sous H_1 : $p' = p + 1$,
- Paramètres sous H_0 : $k + 1$,

donc $\Delta_{dl} = p + 1 - k - 1 = p - k$.

La très grande majorité des tests d'hypothèses que nous aurons à effectuer dans ce cours pourrons être exprimés sous la forme H_0 : modèle réduit contre H_1 : modèle complet. Par exemple le test F de la table ANOVA constitue un test de ce type dans lequel le modèle réduit s'écrit simplement comme $Y_i = \beta_0 + \varepsilon_i$. Dans ce cas on peut facilement voir que la statistique F donnée en (3.19) est égale à la statistique F de la table ANOVA (à voir en exercice). De plus, on déduit aussi facilement que $\Delta_{dl} = p$.

3.10 Géométrie de l'analyse de variance et des tests d'hypothèses

La figure 3.3 illustre la géométrie des moindres carrés. Soient
- $\tilde{\boldsymbol{Y}}_1 = \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1$ le modèle selon les moindres carrés en utilisant seulement $\mathbf{X}_1$;
- $\tilde{\boldsymbol{Y}}_2 = \mathbf{X}_2\tilde{\boldsymbol{\beta}}_2$ le modèle selon les moindres carrés en utilisant seulement $\mathbf{X}_2$;
- $\hat{\boldsymbol{Y}} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}\hat{\boldsymbol{\beta}}$ le modèle complet.

Par souci de simplicité, on note

Vecteur :	$\tilde{\boldsymbol{Y}}_1$	$\hat{\boldsymbol{Y}}$	$\hat{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}_1$	$\boldsymbol{Y} - \tilde{\boldsymbol{Y}}_1$	$\boldsymbol{Y} - \hat{\boldsymbol{Y}}$	$\boldsymbol{Y}$
Longueur :	a	b	c	d	e	f

On remarque que, comme les estimateurs ont été obtenus avec la méthode des moindres carrés, $\hat{\boldsymbol{Y}}$ est la projection de $\boldsymbol{Y}$ dans l'espace $(\mathbf{X}_1, \mathbf{X}_2)$. Par conséquent, l'angle $O\hat{\boldsymbol{Y}}\boldsymbol{Y}$ est droit et on a

$$||\boldsymbol{Y}||^2 = ||\hat{\boldsymbol{Y}}||^2 + ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2, \tag{3.20}$$

où $||\cdot||$ représente la norme (longueur) du vecteur. (Plus simplement, $f^2 = d^2 + e^2$.) Ainsi, comme $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,

$$\begin{aligned}
||\boldsymbol{Y}||^2 &= ||\boldsymbol{X}\hat{\boldsymbol{\beta}}||^2 + ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2 \\
\boldsymbol{Y}^\top\boldsymbol{Y} &= (\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} + (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
\Rightarrow \boldsymbol{Y}^\top\boldsymbol{Y} - n\bar{Y}^2 &= \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} + (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - n\bar{Y}_n^2 \\
&= \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} - n\bar{Y}^2 + \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}},
\end{aligned}$$

ce qui est le résultat désiré, c'est-à-dire l'équation de l'analyse de la variance.

On peut aussi mieux comprendre le test F du principe de somme des carrés résiduels additionnelle. En effet, $\tilde{\boldsymbol{Y}}_1$ est obtenu avec la méthode des moindres carrés, c'est donc la projection dans l'espace de $\mathbf{X}_1$ de $\boldsymbol{Y}$. Alors, l'angle $O\tilde{\boldsymbol{Y}}_1\boldsymbol{Y}$ est droit, et on a que

$$||\boldsymbol{Y}||^2 = ||\tilde{\boldsymbol{Y}}_1||^2 + ||\boldsymbol{Y} - \tilde{\boldsymbol{Y}}_1||^2,$$

donc, en utilisant l'équation (3.20),

$$\begin{aligned} ||\hat{\boldsymbol{Y}}||^2 + ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2 &= ||\tilde{\boldsymbol{Y}}_1||^2 + ||\boldsymbol{Y} - \tilde{\boldsymbol{Y}}_1||^2 \\ \hat{\boldsymbol{Y}}^\top\hat{\boldsymbol{Y}} + \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} &= \tilde{\boldsymbol{Y}}_1^\top\tilde{\boldsymbol{Y}}_1 + \tilde{\boldsymbol{\varepsilon}}_1^\top\tilde{\boldsymbol{\varepsilon}}_1 \\ \hat{\boldsymbol{Y}}^\top\hat{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}_1^\top\tilde{\boldsymbol{Y}}_1 &= \tilde{\boldsymbol{\varepsilon}}_1^\top\tilde{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} \end{aligned}$$

Or, $\tilde{\boldsymbol{Y}}_1$ est la projection de $\hat{\boldsymbol{Y}}$ sur l'espace $\mathbf{X}_1$. Donc, l'angle $O\tilde{\boldsymbol{Y}}_1\hat{\boldsymbol{Y}}$ est droit et on a que

$$||\hat{\boldsymbol{Y}}||^2 - ||\tilde{\boldsymbol{Y}}_1||^2 = c^2.$$

Sur la figure 3.3, on voit que c est la distance entre la prévision selon le modèle complet $\hat{\boldsymbol{Y}}$ et la prévision sous le modèle plus simple contenant seulement $\mathbf{X}_1$. Si cette distance est "petite", alors le modèle réduit est une bonne simplification du modèle complet. C'est pourquoi, lors de tests F partiels, on utilise la statistique $\tilde{\boldsymbol{\varepsilon}}_1^\top\tilde{\boldsymbol{\varepsilon}}_1 - \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}$, standardisée par les degrés de liberté et on divise par la norme des résidus dans le modèle complet $\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}$.

3.11 Critères de sélection de modèle

3.11.1 Introduction

En présence d'un très grand nombre de variables exogènes, on court deux dangers : enlever trop de variables ou conserver trop de variables.

Si on conserve trop de variables, on va augmenter inutilement la variance des estimations. Par conséquent, certains effets importants risquent d'être jugés non significatifs. Plus précisément, supposons que le bon modèle soit :

$$Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \epsilon_i. \tag{3.21}$$

Si on ajoute la variable x_{k+1} et qu'on ajuste le modèle :

$$Y_i = \beta_0^* + \beta_1^* x_{1i} + ... + \beta_k^* x_{ki} + \beta_{k+1}^* x_{(k+1)i} + \epsilon_i \tag{3.22}$$

on obtiendra que $\mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}^*] \geq \mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}]$. Autrement dit, la prévision faite pour un $\boldsymbol{x}^*$ donné sera moins précise avec le modèle de l'équation (3.22) qu'avec celui de l'équation (3.21).

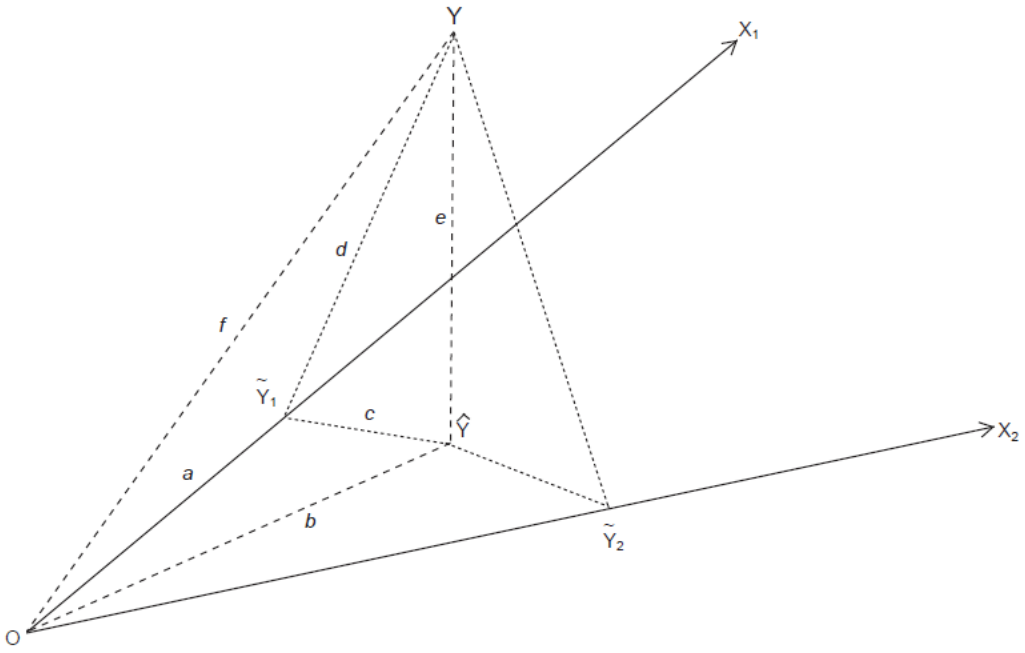


FIGURE 3.3: Illustration de l'analyse de la variance généralisée. Source : David Stephens, notes de cours MATH 523, Université McGill

À l'opposé, si on enlève trop de variables, le modèle sera biaisé. Pour illustrer le biais, prenons par exemple les deux modèles suivants exprimés sous forme matricielle :

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}, \tag{3.23}$$

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \tag{3.24}$$

où β_1 comporte $p' = p + 1$ paramètres et β_2 inclut $m - p'$ paramètres. Supposons que le bon modèle soit celui de l'équation (3.24) et qu'on choisisse plutôt le modèle de l'équation (3.23), on aura alors

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}_1] &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\mathbb{E}[\boldsymbol{Y}] \\
&= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2) \\
&= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2.
\end{aligned}$$

On a donc que $\hat{\boldsymbol{\beta}}_1$ est un estimateur biaisé. Si l'on s'intéresse plutôt à la prévision au point $\boldsymbol{x}^*$, le modèle

à p' paramètres nous donnerait

$$\begin{aligned}\hat{\boldsymbol{Y}}^* &= \boldsymbol{x}_1^{*'}\hat{\boldsymbol{\beta}}_1 \\ \mathbb{E}[\hat{\boldsymbol{Y}}^*|\boldsymbol{x}^*] &= \boldsymbol{x}_1^{*'}\mathbb{E}[\hat{\boldsymbol{\beta}}_1] \\ &= \boldsymbol{x}_1^{*'}\{\boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2\}\end{aligned}$$

alors que le modèle à m paramètres conduirait à

$$\begin{aligned}\hat{\boldsymbol{Y}}^* &= \boldsymbol{x}_1^{*'}\hat{\boldsymbol{\beta}}_1^* + \boldsymbol{x}_2^{*'}\hat{\boldsymbol{\beta}}_2^* \\ \mathbb{E}[\hat{\boldsymbol{Y}}^*|\boldsymbol{x}^*] &= \boldsymbol{x}_1^{*'}\boldsymbol{\beta}_1 + \boldsymbol{x}_2^{*'}\boldsymbol{\beta}_2.\end{aligned}$$

Le biais dans la prévision serait donc

$$\begin{aligned}\text{biais}[\hat{\boldsymbol{Y}}^*|\boldsymbol{x}^*] &= \boldsymbol{x}_1^{*'}\{\boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2\} - \boldsymbol{x}_1^{*'}\boldsymbol{\beta}_1 - \boldsymbol{x}_2^{*'}\boldsymbol{\beta}_2 \\ &= [\boldsymbol{x}_1^{*'}(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2 - \boldsymbol{x}_2^{*'}]\boldsymbol{\beta}_2.\end{aligned}$$

En résumé, un trop grand nombre de variables mène à une variance trop grande et un trop petit nombre de variables conduit à une estimation biaisée. Il est donc important de trouver le nombre adéquat de variables. Pour ce faire, on a souvent recours à l'une des techniques de sélection de variables pour choisir un sous-ensemble raisonnable de variables.

3.11.2 Critères de comparaison classiques

Comme la décomposition de la variabilité totale peut nous inciter à le penser, un bon modèle de régression expliquera en général une partie importante de la variabilité. Un premier critère mesurant la qualité d'un modèle de régression peut donc être défini comme suit :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}. \tag{3.25}$$

Puisque $SST = SSR + SSE$ et que les trois sommes de carrés sont positives, alors $0 \leq R^2 \leq 1$. Si le modèle de régression explique la totalité de la variabilité dans Y, alors $SSR = SST$, $SSE = 0$ et donc $R^2 = 1$. à l'opposé, si le modèle n'explique rien, alors $SSR = 0 = R^2$. Ainsi, plus la valeur de R^2 est grande et approche 1, plus le modèle de régression explique une grande partie de la variabilité de la variable endogène. La statistique R^2 est parfois appelée *coefficient de détermination.*

Bien que facilement interprété et naturellement attrayant, le coefficient de détermination souffre de quelques problèmes qui font qu'il ne peut être utilisé pour comparer n'importe quels modèles de régression l'un avec l'autre. L'inconvénient principal réside dans le fait que dès que l'on ajoute un terme à un modèle de régression, SSR ne peut pas diminuer. Comme SST ne dépend pas du modèle, elle reste inchangée. Donc l'ajout d'un terme au modèle se traduit forcément par un R^2 plus grand. Ainsi, si le "vrai" modèle générant les données est $Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$ et que l'on ajuste ce modèle ainsi que le modèle $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$,

il est fort probable que le R^2 du second modèle soit supérieur à celui du "vrai" modèle. Bien que l'ajout de termes superflus ne crée pas de biais dans le modèle, ces termes excédentaires réduisent en général la précision dans les prévisions et il est souhaitable de les éviter.

Afin de pallier ce problème, nous pouvons utiliser le *coefficient de détermination ajusté* défini comme :

$$R_a^2 = 1 - \frac{MSE}{MS_{tot}} = 1 - \frac{SSE/(n-p')}{SST/(n-1)} = 1 - (n-1)\frac{s^2}{SST}. \tag{3.26}$$

Avec R_a^2, l'ajout d'une variable exogène peut aussi résulter en une diminution de la statistique. Comme on peut le voir dans la dernière égalité de l'équation (3.26), comparer deux modèles sur la base de leur R_a^2 est équivalent à comparer deux modèles sur la base de leur estimation de la variance des termes d'erreur s^2, puisque SST et n ne dépendent pas du modèle choisi.

Il est bon de noter que peu importe si on utilise R^2 ou R_a^2, on ne peut pas vraiment se fier sur ces statistiques pour comparer des modèles employant des transformations différentes de la variable endogène, ou un modèle avec et un modèle sans ordonnée à l'origine.

3.11.3 Méthodes basées sur la puissance de prévision

Comme les prévisions sont très sensibles au choix de modèle, il serait souhaitable de définir les critères de qualité d'un modèle selon son habileté à prédire de nouvelles observations de façon adéquate. Malheureusement, nous voulons en général utiliser toutes nos données afin d'estimer les paramètres de la façon la plus précise possible, donc nous ne pouvons pas tester notre modèle sur de nouvelles observations, à moins d'user d'un peu d'ingéniosité !

Le principe de validation croisée

La validation croisée sert à mesurer la capacité d'un modèle donné à bien prédire de nouvelles observations. Un algorithme classique de validation croisée consiste en

1. Enlever la $i$ème observation du jeu de données.
2. Estimer les paramètres du modèle à partir des $n-1$ données restantes.
3. Prédire Y_i à partir de $\boldsymbol{x}_i$ et du modèle obtenu en 2. Dénoter cette valeur prédite $\hat{Y}_{i,-i}$.
4. Répéter les étapes 1-3 pour chaque i, $i = 1, \ldots, n$.
5. Calculer la somme des carrés des erreurs de prévisions $PRESS = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i,-i})^2$.

Le critère $PRESS$ permet de comparer entre eux tous les modèles utilisant la même transformation pour la variable endogène. évidemment, plus le critère $PRESS$ est faible, plus le modèle prédit bien. Le critère $PRESS$ peut aussi servir à définir un *coefficient de détermination de prévision* de la manière suivante :

$$R_p^2 = 1 - \frac{PRESS}{SST} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_{i,-i})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}. \tag{3.27}$$

Encore une fois, une valeur de R_p^2 approchant 1 est signe d'un modèle qui prédit bien, tandis qu'une valeur approchant 0 indique un modèle qui prédit mal.

Les résidus PRESS

Le $i$ème résidu $PRESS$ est le $i$ème terme de la somme des carrés des erreurs de prévision, c.-à-d. $\hat{\epsilon}_{i,-i} = Y_i - \hat{Y}_{i,-i}$. Pour comparer des modèles, on peut donc faire appel à la somme $PRESS$, mais il est aussi bon d'examiner chaque résidu $PRESS$ de façon individuelle. En effet, un modèle précis dont une seule des prévisions est mauvaise pourrait être indûment pénalisé par la somme $PRESS$.

La pertinence théorique des résidus $PRESS$ devrait maintenant être claire. Cependant la tâche de les calculer semble très fastidieuse ! Heureusement, les propriétés de la matrice chapeau rendent ce calcul très simple.

Théorème 3.4. *Le ième résidu PRESS peut être obtenu à l'aide du ième résidu ordinaire et de l'élément en position i, i de la matrice chapeau grâce à la relation*

$$\hat{\epsilon}_{i,-i} = \frac{\hat{\epsilon}_i}{1 - h_{ii}}.$$

La conséquence très pratique du théorème 3.4 est que les résidus $PRESS$ et la statistique $PRESS$ peuvent être obtenus sans avoir à effectuer les n régressions de l'algorithme de validation croisée ! Le corollaire 3.5 livre l'expression de la statistique $PRESS$:

Corollaire 3.5.

$$PRESS = \sum_{i=1}^{n} \left(\frac{\hat{\epsilon}_i}{1 - h_{ii}} \right)^2.$$

Au chapitre 3 on a vu que h_{ii} détermine la longueur de l'intervalle de prévision. Comme $0 \leq h_{ii} \leq 1$, une valeur de h_{ii} près de 1 indique un intervalle de prévision large. Ceci est tout-à-fait en accord avec la formule du théorème 3.4, puisque le dénominateur de $\hat{\epsilon}_{i,-i}$ approche 0 lorsque h_{ii} approche 1, et donc $\hat{\epsilon}_{i,-i}$ (l'erreur de prévision) approche l'infini et la valeur de Y_i est difficile à prédire.

3.11.4 Le C_p de Mallows

On a vu précédemment qu'un trop grand nombre de variables cause un problème de variance et qu'un nombre insuffisant de variables induit un problème de biais. Un compromis serait alors de tenter de minimiser l'erreur quadratique moyenne (EQM) de prévision.

$$\begin{aligned} \mathrm{EQM}[\hat{Y}(\boldsymbol{x}^*)] &= \mathbb{V}\mathrm{ar}[\hat{Y}(\boldsymbol{x}^*)] + \mathrm{biais}^2[\hat{Y}(\boldsymbol{x}^*)] \\ &= \sigma^2 \boldsymbol{x}_1^{*\prime} (\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{x}_1^* + \left\{ [\boldsymbol{x}_1^{*\prime} (\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1' \boldsymbol{X}_2 - \boldsymbol{x}_2^{*\prime}] \boldsymbol{\beta}_2 \right\}^2 \end{aligned}$$

Pour minimiser l'EQM, on fait face à deux difficultés : premièrement, il on ne sait pas quel x^* choisir, deuxièmement, l'EQM dépend de β_2, que l'on ne connaît pas. Mallows a donc proposé un autre critère, qui permet de régler le premier problème. On minimise

$$\sum_{i=1}^{n} \frac{\mathrm{EQM}[\hat{Y}(x_i)]}{\sigma^2}.$$

La division par σ^2 ne joue aucun rôle mais permettra une simplification commode.

Calcul de la portion variance de l'EQM

$$\begin{aligned}
\sum_{i=1}^{n} \frac{\mathbb{V}\mathrm{ar}[\hat{Y}(x_i)]}{\sigma^2} &= \sum_{i=1}^{n} \frac{x'_{1i}\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}_1]x_{1i}}{\sigma^2} \\
&= \sum_{i=1}^{n} x'_{1i}(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}x_{1i} \\
&= \sum_{i=1}^{n} \mathrm{tr}[x'_{1i}(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}x_{1i}] \\
&= \sum_{i=1}^{n} \mathrm{tr}[x_{1i}x'_{1i}(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}] \\
&= \mathrm{tr}\left[(\sum_{i=1}^{n} x_{1i}x'_{1i})(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}\right] \\
&= \mathrm{tr}\left[(\boldsymbol{X}'_1\boldsymbol{X}_1)(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}\right] \\
&= \mathrm{tr}(I'_p) = p'
\end{aligned}$$

Dans le développement ci-dessus, nous avons utilisé le fait que la trace d'un nombre réel est le nombre lui-même et que la $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

Calcul de la portion biais de l'EQM La portion biais peut être estimée facilement une fois que l'on réalise que

$$\mathbb{E}[s_p^2] = \sigma^2 + \frac{\sum_{i=1}^{n} \mathrm{biais}[\hat{Y}(x_i)]^2}{n - p'}$$

où s_p^2 est la variance calculée dans le modèle à p variables. Si σ^2 était connu, on pourrait estimer $\sum_{i=1}^{n} \mathrm{biais}[\hat{Y}(x_i)]^2$ par $(s_p^2 - \sigma^2)(n-p)$. Puisque σ^2 est inconnu, on l'estime par $\hat{\sigma}^2$, la variance calculée dans le plus grand modèle dont on dispose. Le critère proposé par Mallows est donc le suivant :

$$C_p = p' + \frac{(s_p^2 - \hat{\sigma}^2)(n - p')}{\hat{\sigma^2}}.$$

En pratique, un modèle pour lequel $C_p \approx p'$ (ou $C_p - p' \approx 0$) est un bon modèle. Une façon simple de choisir un modèle lorsque k variables exogènes sont disponibles est de calculer le C_p des 2^{k+1} modèles possibles et de prendre celui pour lequel $C_p - p'$ est le plus faible. Mais en général, on utilise C_p pour cerner un petit ensemble de modèles raisonnables, et ensuite on utilise d'autres critères ($PRESS$, R_a^2, etc.) pour départager ces modèles. On peut réécrire le C_p de la façon suivante :

$$C_p = \frac{SSE}{\hat{\sigma^2}} + 2p' - n$$

3.11.5 Critère d'information d'Akaike

Le critère d'information d'Akaike a été proposé par Hirotsugu Akaike (1974), un statisticien japonais. Ce critère est de plus en plus utilisé dans la pratique et permet d'évaluer la qualité de l'ajustement d'un modèle. Il est défini comme suit :

$$\text{AIC} = -2\ln L(\hat{\boldsymbol{\beta}}, \sigma^2_{\text{ML}}) + 2p',$$

où

$$\begin{aligned} -2\ln L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_{\text{ML}}) &= \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{\hat{\sigma}^2_{\text{ML}}} + n\ln(2\pi\hat{\sigma}^2_{\text{ML}}) \\ &= n + n\ln(2\pi\hat{\sigma}^2_{\text{ML}}) \\ &= n + n\ln(2\pi) + n\ln(SSE/n), \end{aligned}$$

car $\hat{\sigma}^2_{\text{ML}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/n = SSE/n$.
En laissant de côté la constante, $n + n\ln(2\pi)$, la statistique AIC s'obtient alors simplement par l'expression suivante

$$\text{AIC} = n\ln(SSE/n) + 2p'.$$

L'AIC prend en compte à la fois la qualité des prédictions du modèle et sa complexité. En effet, un modèle qui s'ajuste bien sur les observations est associé à une faible valeur de SSE et donc à une faible valeur de $n\ln(SSE/n)$. La complexité du modèle est prise en compte en ajoutant le double du nombre de paramètres ($2p'$) à cette dernière quantité. Un bon modèle est donc associé à une faible valeur d'AIC.

3.11.6 Méthodes algorithmiques

Ces méthodes sont moins recommandées que l'utilisation du C_p, mais comme 2^{p+1} croit rapidement en fonction de p, il peut parfois être impossible de calculer les critères C_p, R^2_a, etc. pour tous les modèles possibles. C'est pour ces raisons, et aussi dû à un manque de puissance informatique à l'époque, que des méthodes algorithmiques ont été inventées.

La méthode d'inclusion, «forward»

Cette méthode est associable à un paradigme d'achat. La procédure est la suivante :

1. On commence en prenant uniquement β_0.
2. On cherche ensuite, parmi toutes les variables, la variable la plus payante, c'est-à-dire celle qui permet de réduire le plus le terme d'erreur (SSE).
3. On effectue un test pour savoir si la variable mérite d'être «achetée» (ajoutée au modèle). La statistique du test la suivante :

$$F_{obs} = \frac{SSE_{\text{Petit modèle}} - SE_{\text{Grand modèle}}}{SSE_{\text{Grand modèle}}/ll}$$

où le grand modèle est le modèle qui inclut la variable ajoutée à l'étape 2, le petit modèle est celui qui ne l'inclut pas et ll est le nombre de degrés de liberté du grand modèle $(n - p')$. On considère que ça vaut la peine d'ajouter la variable lorsque $F_{obs} > F_{1,ll}1 - \alpha$.

4. Si la variable vaut la peine d'être ajoutée au modèle, on l'ajoute et on recommence l'étape 2 avec une autre variable. On répète la procédure jusqu'à ce qu'aucune variable ne vaille la peine d'être achetée.

Typiquement, on pose le seuil α à $0,5$. Ce seuil ne doit pas être trop petit sans quoi aucune variable n'est sélectionnés.

La méthode d'exclusion, «backward»

La méthode d'exclusion est assimilable à un paradigme de vente. On procède à l'inverse de la méthode d'inclusion.

1. On commence avec le modèle complet, qui contient toutes les variables
2. On cherche à «vendre» la variable la moins payante, c'est-à-dire celle qui augmente le moins le terme d'erreur.
3. On effectue le test

$$F_{obs} = \frac{SSE_{\text{Petit modèle}} - SSE_{\text{Grand modèle}}}{SSE_{\text{Grand modèle}}/ll}$$

et on considère que ça ne vaut pas la peine de garder la variable choisie à l'étape 2 si $F_{obs} < F_{1,ll}(1-\alpha)$

4. Si la variable est rejetée, on recommence avec une autre variable jusqu'à ce qu'aucune variable ne vaille la peine d'être retirée.

Ici, on fixe généralement $\alpha = 0,10$. Il ne faut pas que α soit trop élevé, sans quoi on ne «vend» aucune variable. **Remarque importante** : Une erreur commune commise en pratique consiste à ajuster le modèle complet et d'éliminer ensuite d'un seul coup toutes les variables dont le seuil est trop élevé. Ceci est fortement déconseillé, car après l'élimination d'une ou deux variables, certaines variables qui ne semblaient pas du tout importantes peuvent soudain le devenir !

La méthode pas à pas, «stepwise»

Il s'agit d'un mélange des méthodes d'inclusion et d'exclusion. On commence par la méthode d'inclusion. Après avoir ajouté une variable, on effectue la méthode d'exclusion. À chaque étape, on se remet en question et on rejette les variables qui n'en valent pas la peine avant d'en ajouter de nouvelles.

Exemple 3.6. *Prenons un cas avec trois variables et $n = 10$ pour illustrer les méthodes de sélection de variables dites algorithmiques (inclusion, exclusion et pas à pas). Le tableau*

Variables du modèle	*SSE*
x_0	*30*
x_1	*25*
x_2	*24*
x_3	*20*
x_1, x_2	*6*
x_1, x_3	*15*
x_2, x_3	*12*
x_1, x_2, x_3	*4*

Si on procède par inclusion :

1. *On commence avec la variable x_3. On obtient $F = \frac{30-20}{20/(10-2)} = 4 > F_{0,5;1;8}$ et on «achète» la variable.*
2. *On choisit ensuite la variable x_2. On obtient $F = \frac{20-12}{12/(10-3)} = 4,67 > F_{0,5;1;7}$ et on «achète» la variable.*
3. *Enfin choisit la variable x_3. On obtient $F = \frac{12-4}{4/(10-4)} = 12 > F_{0,5;1;6}$ et on «achète» la variable.*

La méthode d'inclusion conduit donc à retenir toutes les variables.

Si on procède par exclusion :

1. *On retire x_3. On obtient $F = \frac{6-4}{4/(10-4)} = 3 < F_{0,10;1;6}$. On rejette la variable x_3.*
2. *On retire x_1. On obtient $F = \frac{24-6}{6/(10-3)} = 21 > F_{0,10;1;7}$. On ne rejette pas la variable x_1.*

La méthode par exclusion conduit à ne retenir que les variables x_1 et x_3. On peut faire le même exercice avec la méthode pas à pas et le résultat serait, dans ce cas précis, le même que le résultat obtenu par exclusion.

3.11.7 Le graphique des variables ajoutées

Lorsqu'on envisage d'ajouter une variable $\tilde{X}$, on peut vérifier si cette variable apporte quelque chose de neuf avec le graphique des variables ajoutés. Supposons le modèle $Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p_i} x_{1p_i} + \epsilon_i$. La portion non expliquée du modèle est ϵ_i. L'apport de $\tilde{X}$ est la partie de $\tilde{X}$ qui n'est pas expliquée par une combinaison linéaire de $X_1, ..., X_{p_i}$. On peut donc régresser

$$\tilde{x}_i \quad = \quad \delta_0 + \delta_1 x_{i1} + ... + \delta p_i x_{ip_i} + \tilde{\epsilon_i}.$$

Plus les résidus de cette régression sont grands, plus $\tilde{X}$ ajoute une dimension importante. Pour savoir si elle est utile, on regarde le graphique de $\hat{\epsilon}_i$ en fonction de $\hat{\tilde{\epsilon}}_i$. La figure 3.4 présente deux cas possible.

Il est à noter que la pente de la droite qui passe par les points du graphique de la variable ajoutée $\tilde{X}$ est $\beta_{\tilde{X}}$. Le graphique permet en outre de voir si l'effet de $\tilde{X}$ est linéaire ou non.

3.12 Régression avec variables qualitatives

Voir notes en classe.

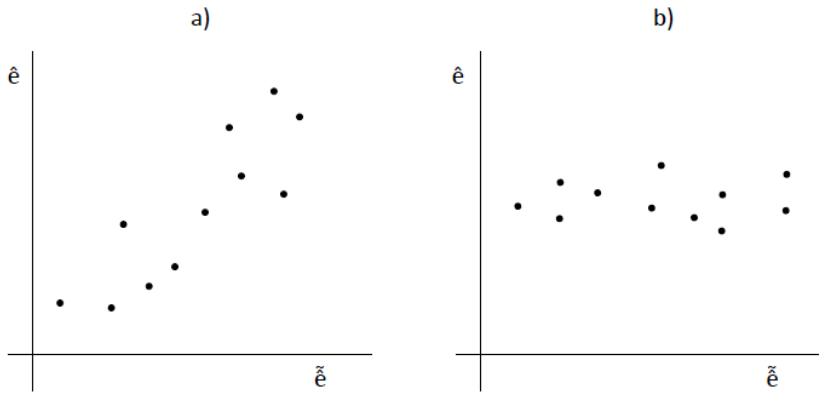


FIGURE 3.4: Exemple de graphiques de variables ajoutées. Le graphique a) illustre un cas où la variable ajoutée est utile. Le graphique b) illustre un cas où la variable ajoutée n'est pas utile.

3.12.1 Exemple avec variable polytomique

Soit x_{i1}, le numéro de lot du produit i, x_{i2}, la concentration de sel dans le produit i et Y_i, l'indice de qualité du produit i. Les variables Y_i et x_{i2} sont des variables continues, alors que x_{i1} est un variable polytomique prenant une des valeurs $\{1, 2, 3, 4\}$. À l'aide d'un graphique de $\mathbb{E}[Y_i]$ en fonction de x_{i2}, interpréter les coefficients du modèle

$$Y_i = \beta_0 + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_2 x_{i2} + \varepsilon_i,$$

où

$$x_{i11} = \begin{cases} 1, & x_{i1} = 1 \\ 0, & x_{i1} \neq 1 \end{cases} \qquad x_{i12} = \begin{cases} 1, & x_{i1} = 2 \\ 0, & x_{i1} \neq 2 \end{cases}$$

$$x_{i13} = \begin{cases} 1, & x_{i1} = 3 \\ 0, & x_{i1} \neq 3. \end{cases}$$

Test de l'effet d'une variable polytomique

Pour tester si une variable polytomique possède un effet significatif sur la valeur moyenne de la variable endogène, il s'agit de tester si plusieurs coefficients sont simultanément égaux à zéro à l'aide d'un test F. Ainsi, si nous voulons tester si le groupe d'où provient l'item i a un effet dans l'exemple, il faut tester $H_0 : \beta_{11} = \beta_{12} = \beta_{13} = 0$. (Notez que ceci revient à tester si la ligne de régression est la même pour les 4 groupes, c'est-à-dire que sur la Figure 3.5, les 4 lignes sont superposées.)

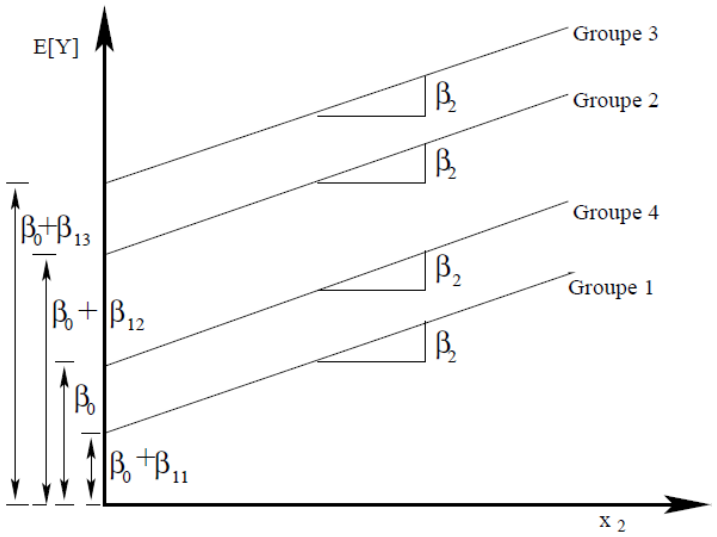


FIGURE 3.5: Espérance de la variable endogène Y en fonction de la variable exogène x_2 pour chacun des groupes 1 à 4. Comme on peut le voir, les β_0 et β_{1j} déterminent les ordonnées à l'origine, tandis que β_2 détermine la pente des droites de régression.

En général, pour tester si une variable polytomique prenant C valeurs possibles a un effet significatif, il faut tester si $C-1$ coefficients sont simultanément égaux à zéro.

Sélection de modèle

La règle utilisée en sélection de modèle, quel que soit le critère de sélection, est que les β correspondant à une même variable polytomique soient ou bien tous exclus du modèle, ou bien tous inclus dans le modèle. Ainsi dans l'exemple, si nous utilisons un critère de sélection qui nous dit que le meilleur modèle est $Y_i = \beta_0 + \beta_{11}x_{11i} + \beta_{13}x_{13i} + \varepsilon_i$, alors on ajoute $\beta_{12}x_{12i}$ à ce modèle car β_{11}, β_{12} et β_{13} sont tous reliés à la variable « lot »et doivent donc ou tous être dans le modèle, ou tous être exclus du modèle, et comme au moins un de ces β semble important, nous les incluons tous.

3.13 Multicolinéarité

Depuis le début du cours, nous avons supposé que la matrice de schéma $\boldsymbol{X}$ est une matrice de plein rang, c'est-à-dire qu'il existe une unique inverse à la matrice $\boldsymbol{X}'\boldsymbol{X}$, ce qui possède comme conséquence qu'il n'existe qu'un seul vecteur $\hat{\boldsymbol{\beta}}$ qui miminise la somme des carrés résiduels SS_{res}. Parfois certaines colonnes

de X sont reliées linéairement entre elles ou sont près de l'être, dans ce cas nous sommes en présence de multicolinéarité.

Définition 3.7. *Il y a multicolinéarité exacte lorsqu'une variable est la combinaison linéaire d'une ou plusieurs autres variables, c'est-à-dire lorsqu'il existe une combinaison de constantes $d_1, \ldots, d_p$ (pas toutes égales à zéro) telle que*

$$\sum_{j=1}^{p} d_j X_j = \mathbf{0}.$$

La multicolinéarité au sens large se produit lorsqu'il existe une combinaison de constantes $d_1, \ldots, d_p$ (pas toutes égales à zéro) telle que

$$\sum_{j=1}^{p} d_j X_j \cong \mathbf{0}.$$

Dans le cas de multicolinéarité exacte, la matrice $(\boldsymbol{X'X})$ n'est pas inversible et la régression n'est pas possible. La multicolinéarité exacte ne représente pas un problème : elle se détecte facilement et il suffit d'enlever la ou les variables en cause pour régler le problème. La multicolinéarité au sens large est plus fréquente et plus difficile à détecter. Elle pose les problèmes suivants :

1. Instabilité de $(\boldsymbol{X'X})^{-1}$, ce qui entraîne qu'une petite variation en $\boldsymbol{Y}$ peut se traduire par de très grands changements en $\hat{\boldsymbol{\beta}}$ et en $\hat{\boldsymbol{Y}}$.
2. On observe des $\hat{\beta}_i$ de signe inattendu ou contre-intuitif.
3. Les variances des $\hat{\beta}_i$ et des $\hat{Y}_i$ sont très grandes.
4. Les méthodes de sélection de variables de concordent pas.
5. Certains paramètres peuvent se révéler non-significatifs alors que la corrélation des variables correspondantes avec Y est très grande.

3.13.1 Détection de la multicolinéarité

La matrice de corrélations

À première vue, on pourrait être tenté de calculer la matrice des corrélations entre les variables exogènes pour détecter la multicolinéarité. Soit

$$\boldsymbol{x}_j^* = \frac{\boldsymbol{x}_j - \bar{\boldsymbol{x}}_j}{s_j} \quad j \in \{1, \ldots, p\},$$

où $\bar{x}_j = \sum_{i=1}^{n} x_{ji}/n, \boldsymbol{x}_j = (\bar{x}_j, \ldots, \bar{x}_j)'$ et $s_j = \sqrt{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2/(n-1)}$. La matrice des coefficients de corrélation échantillonnaux est alors donnée par $\boldsymbol{X}^{*'}\boldsymbol{X}^*$

$$\boldsymbol{X}^{*'}\boldsymbol{X}^* = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

où

$$\boldsymbol{X}^* = \left(\frac{\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1}{s_1} \quad \frac{\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2}{s_2} \dots \frac{\boldsymbol{x}_p - \bar{\boldsymbol{x}}_p}{s_p} \right),$$

et

$$r_{jk} = \sum_{i=1}^{n} \frac{(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{s_j s_k},\ j \in \{1, \dots, p\}, k \in \{1, \dots, p\}$$

Si deux variables exogènes sont linéairement reliées, leur coefficient de corrélation devrait être élevé. Cependant deux problèmes majeurs empêchent les coefficients de corrélation d'être des bons outils de diagnostic : (i) il est difficile de dire ce qu'est une large valeur de la corrélation et (ii) dans une grande proportion des cas, la multicolinéarité est induite par une dépendance linéaire entre plus de deux variables. Pour voir que le point (ii) constitue un problème, il est possible de créer des exemples où p variables sont parfaitement linéairement dépendantes, mais où les coefficients de corrélations de toutes les paires de variables sont inférieurs à $1/(p-1)$. Ainsi, même si la corrélation entre deux variables est indésirable, cela ne représente qu'un cas particulier de multicolinéarité.

3.13.2 Le facteur d'inflation de la variance (VIF)

Une approche plus sensible consiste à évaluer le degré de dépendance linéaire de chaque variable exogène sur les autres variables exogènes. Ainsi pour la j$^{\text{ème}}$ variable exogène, on peut mesurer ce niveau de dépendance en effectuant une régression linéaire avec la j$^{\text{ème}}$ variable exogène comme variable réponse et les $p-1$ variables exogènes restantes comme variables explicatives. Le coefficient de détermination de cette régression, dénoté R_j^2, mesurera la proportion de la variabilité de la j$^{\text{ème}}$ variable exogène qui est expliquée de façon linéaire par les autres variables exogènes. Si R_j^2 est grand, on a vraisemblablement un problème.

Le facteur d'inflation de la variance (en anglais, VIF) est défini de la façon suivante :

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \tag{3.28}$$

Si $R_j^2 \to 1$, $VIF_j \to \infty$.

Le nom « facteur d'inflation de la variance »vient du fait que la variance de $\hat{\beta}_i$ s'exprime en fonction du VIF comme suit :

$$\mathbb{V}\text{ar}[\hat{\beta}_j] = \frac{\sigma^2}{(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)_{jj}} \text{VIF}_j$$

On peut donc voir qu'une grande valeur pour VIF_j signifie une valeur proche de 1 pour R_j^2, et donc une forte dépendance linéaire des variables exogènes. De plus, on peut voir l'effet de cette multicolinéarité sur la variance des estimateurs des coefficients. Plusieurs auteurs suggèrent un VIF > 10 comme point où l'on doit commencer à considérer la présence de multicolinéarité aux conséquences importantes. Cependant, les VIF à eux seuls ne sont pas un diagnostic complet. Parmi les points faibles des VIF on dénote l'incapacité

de détecter des multicolinéarités impliquant la colonne de 1 de la matrice de schéma, l'incapacité de cerner le nombre de quasi dépendances linéaires présentes dans les données et finalement on n'a jamais réussi à cerner une valeur précise pour le VIF où l'on doit vraiment commencer à s'inquiéter (10 est une valeur *ad hoc*).

3.13.3 Les solutions possibles à la multicolinéarité

La solution la plus simple est de retirer une ou plusieurs variables exogènes. Typiquement, on retranchera les variables ayant le plus grand VIF, une à la fois. Dans la pratique, ce retrait des variables doit se faire avec précautions, car c'est d'abord l'interprétation qui compte. Il est également possible parfois de combiner des variables exogènes redondantes.

Remarque 3.8. *La régression ridge et la régression basée sur les composantes principales sont deux méthodes statistiques qui permettent de traiter la multicolinéarité.*

3.14 Analyse des résidus et test pour manque d'ajustement

L'analyse des résidus a été couverte en très grande partie au Chapitre 2. Dans ce chapitre, nous ne faisons que réviser brièvement les notions du Chapitre 2 et nous introduisons quelques nouvelles procédures.

Définition 3.9. *Les résidus sont contenus dans le vecteur* $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$*, c'est-à-dire que le* i*ème résidu est* $Y_i - \hat{Y}_i$*.*

Quelques résultat :

1. $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (I - H)\boldsymbol{Y}$;
2. $Var(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(I - H)$
3. $r_i = \hat{\varepsilon}_i / \sqrt{s^2(1 - h_{ii})}$ = le $i$ème résidu studentisé
4. $\hat{\varepsilon}_{i,-i} = \hat{\varepsilon}_i / (1 - h_{ii})$ = le $i$ème résidu PRESS

La plupart des graphiques du Chapitre 2 sont toujours très utiles pour la validation des postulats du modèle et s'interprètent de la même façon qu'auparavant. Nous traiterons dans les sections suivantes quelques étapes de validation d'un modèle de régression linéaire multiple.

3.15 Vérification de la linéarité

Les graphique de $\hat{\boldsymbol{\varepsilon}}$ *versus* x_j permet de cerner des problèmes de linéarité. Ces graphiques devraient ne montrer aucune tendance, avec les résidus symétriques autour de zéro en ordonnée et une variabilité qui ne varie pas beaucoup en fonction de la valeur de $x_j i$.

Toutefois, **les graphiques à variable ajoutée** ($\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{Y}|\boldsymbol{X}_{-j}}$ vs $\hat{\boldsymbol{\varepsilon}}_{x_j|\boldsymbol{X}_{-j}}$) sont mieux adaptés au contexte de la régression multiple pour détecter le manque de linéarité. Ici $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{Y}|\boldsymbol{X}_{-j}}$ représente le vecteur des résidus

obtenu en régressant $\boldsymbol{Y}$ sur toutes les variables exogènes du modèle **excepté** x_j, et $\hat{\boldsymbol{\varepsilon}}_{x_j|\boldsymbol{X}_{-j}}$ est le vecteur des résidus de la régression avec x_j comme variable endogène et toutes les variables explicatives sauf x_j comme variables exogènes. Si x_j est vraiment une variable qui entre linéairement dans le modèle, le graphique devrait donner une droite de pente β_j passant par l'origine. Si x_j est « inutile », le graphique devrait ressembler à un graphique de résidus ordinaires. Si x_j est utile mais entre dans le modèle de façon non linéaire, le graphique devrait avoir l'allure d'une courbe plutôt que d'une droite. Ce graphique permet donc de détecter un manque de linéarité en x_j de la moyenne de $\boldsymbol{Y}$.

Un test pour manque d'ajustement (lack-of-fit) sera présenté plus loin dans le chapitre.

3.16 Homogénéité des variances

Le graphique r_i vs $\hat{Y}_i$ sert principalement à détecter l'hétéroscédasticité. Ce probème apparait lorsque les observations sont distribuées sous une forme d'entonnoir ouvert vers la gauche ou vers la droite. Comme le graphique doit être symétrique autour de zéro en ordonnée, ce graphique peut aussi servir à détecter un manque de linéarité ou encore à détecter la présence d'observations extrêmes.

3.17 Normalité des erreurs

Le graphique Q-Q plot normal permet de vérifier la normalité des résidus. Une ligne droite est bon signe, une ligne courbe est signe que les résidus peuvent provenir d'une distribution à queues épaisses ou une distribution asymétrique.

Quand le postulat de normalité n'est pas rencontré, il peut être utile d'utiliser la technique de transformation de Box-Cox pour déterminer la meilleure transformation de la variable endogène.

3.18 Indépendance entre les observations

Lorsque les données ont été recueillies de façon chronologique, il peut y avoir une certaine dépendance entre les observations. Le graphique $\hat{\epsilon}_i$ vs i permet de détecter la présence d'autocorrélation dans les résidus. Les points de ce graphique apparaissent alors sous forme de grappes avec de grands et petits résidus en alternance.

Le test de Durbin-Watson est un test très populaire qui permet de détecter la présence d'autocorrélation positive entre les résidus.

3.18.1 Conséquences potentielles de l'autocorrélation

Que se passe-t-il si une autocorrélation est présente entre les résidus, mais que nous n'en tenons pas compte ?

Toutes nos procédures inférentielles (tests d'hypothèses, intervalles de confiances, etc.) dépendent du fait que la variance des termes d'erreur est $\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}] = \sigma^2 I$. En présence d'autocorrélation ($\mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_{i+j}) \neq 0$),

les éléments en dehors de la diagonale de $\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}]$ ne sont plus tous égaux à zéro, et donc nos estimateurs de la variance de $\hat{\boldsymbol{\beta}}$ sont erronés. Il s'en suit que les statistiques pour les tests t de Student sont erronées et donc de mauvaises conclusions quant à la pertinence de certaines variables exogènes pourront être obtenues.

3.19 Test pour manque d'ajustement (lack-of-fit)

Ce test permet de vérifier si le modèle proposé s'ajuste bien aux données. Il constitue un bon complément à l'analyse des résidus. Malheureusement, ce test n'est faisable que lorsque nous avons plusieurs observations dont la (les) variable(s) exogène(s) a (ont) exactement la (les) même(s) valeur(s).
Idée : On peut estimer σ^2 (l'erreur pure) grâce à la variabilité de la variable endogène parmi les observations ayant une même combinaison de variables exogènes. On peut donc séparer notre somme de carrés résiduels en une partie due à l'erreur pure et une partie due au manque d'ajustement du modèle.

Exemple 3.10. *Soit le jeu de données suivant :*

Obs	Y_i	x_1	x_2	*Obs*	Y_i	x_1	x_2
1	*8.74*	*2*	*1*				
2	*7.10*	*2*	*0*	*3*	*9.85*	*2*	*0*
4	*12.1*	*5*	*1*	*5*	*13.1*	*5*	*1*
6	*14.1*	*5*	*0*				
7	*18.6*	*8*	*1*	*8*	*15.0*	*8*	*1*
9	*16.9*	*8*	*0*				

Les données 2 et 3, 4 et 5 et 7 et 8 ont les mêmes combinaisons de valeurs pour leurs variables exogènes. Un test de manque d'ajustement est donc possible.

Supposons que le jeu de données compte m combinaisons distinctes des valeurs des variables exogènes. Soit n_i, le nombre d'observations de la $i$ème combinaison de variables exogènes, $i = 1, \ldots, m$. Soit Y_{ij} la $j$ème observation de la variable endogène pour la $i$ème combinaison de variables exogènes, $j = 1, \ldots, n_i$. Soit $\bar{Y}_i = \sum_j Y_{ij}/n_i$, la valeur moyenne de la variable endogène pour la $i$ème combinaison de variables exogènes.

Exemple 3.11 (suite). *On a $m = 6$ combinaisons de valeurs différentes pour (x_{1i}, x_{2i}) :*

$$\{(2,0),(2,1),(5,0),(5,1),(8,0),(8,1)\}.$$

On a $n_1 = 2$ observations avec $(x_{1i}, x_{2i}) = (2,0)$, $n_2 = 1$ observations avec $(x_{1i}, x_{2i}) = (2,1)$, et ainsi de suite jusqu'à $n_6 = 2$ observations avec $(x_{1i}, x_{2i}) = (8,1)$. Finalement, $\bar{Y}_1 = 8.475$, $\bar{Y}_2 = 8.74$, ... $\bar{Y}_6 = 16.8$.

La variabilité dans les Y_i peut maintenant être estimée à partir des données ayant les mêmes combinaisons pour les variables exogènes, et ce même sans modèle :

$$s^2 = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{\sum_{n_i - 1}}.$$

Quand un modèle de régression est donné, alors la somme des carrés résiduels est décomposée en sommes de carrés provenant de deux sources :

$$\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\hat{Y}_i)^2}_{Sc_{res}} = \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2}_{SC_{res:pure}} + \underbrace{\sum_{i=1}^{m} n_i(\bar{Y}_i-\hat{Y}_i)^2}_{SC_{manque\ ajust.}}.$$

Si le modèle de régression est bon, alors l'erreur pure explique une partie importante de la somme des carrés résiduels et la somme due au manque d'ajustement du modèle devrait être de faible valeur. On rejetterait donc H_0 : le modèle est adéquat en faveur de H_1 : le modèle ne s'ajuste pas bien aux données pour de grandes valeurs de

$$F = \frac{SC_{manque\ ajust.}/(m-p')}{SC_{res:pure}/(n-m)},$$

où "grandes valeurs signifie $F \geq F_{m-p',n-m}(1-\alpha)$.

Exemple 3.12 (suite). *Si on prend le modèle $Y_{ij} = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$, alors on obtient*

Residual	*df*	*SS*	*MS*	*Fvalue*	*Pr > F*
Lack of Fit	*1*	*0.322672*	*0.322672*	*0.16*	*0.7055*
Pure Error	*6*	*12.314733*	*2.052456*		
Total Error	*7*	*12.637406*	*1.805344*		

En effet, pour les degrés de liberté on a $m - p' = 3 - 2 = 1$ et $n - m = 9 - 3 = 6$. Il ne semble pas y avoir de problème de manque d'ajustement dans notre exemple. Bien entendu, avec le très petit nombre de données observé, très peu d'hypothèses peuvent être rejetées !!

3.20 Hétéroscédasticité et régression pondérée

Les procédures étudiées jusqu'à présent sont adéquates seulement si $\varepsilon_1, \ldots, \varepsilon_n$ sont indépendants et $\varepsilon_i \sim N(0, \sigma^2)$, ou de façon équivalente si $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I)$. Que doit-on faire lorsque $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2))$, où $\sigma_i^2 \neq \sigma^2$ pour tout i ? Autrement dit, que faire en cas d'hétéroscédasticité ?

La méthode des moindres carrés pondérés consiste à trouver $\hat{\boldsymbol{\beta}}$ tel que

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

est minimale, où on pose $w_i \propto 1/\sigma_i^2$. Il est relativement simple de démontrer que dans ce cas,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} X^\top \boldsymbol{V}^{-1} \boldsymbol{Y},$$

où $V = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. La variance de $\hat{\boldsymbol{\beta}}$ est

$$\begin{aligned} \mathbb{V}\text{ar}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{V}^{-1}\, \boldsymbol{V}\, \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}. \end{aligned}$$

On peut donc tester $H_0 : \beta_j = \beta_{j,0}$ en utilisant la statistique

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{(\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}_{j,j}}}.$$

Parfois on utilise $\boldsymbol{V} = \boldsymbol{V}^* \sigma^2$ et on estime σ^2 par

$$s^2 = \frac{\sum_{i=1}^n w_i (Y_i - \hat{Y}_i)^2}{n - p'}.$$

Une table d'analyse de variance peut aussi être construite en utilisant la décomposition

$$\underbrace{\boldsymbol{Y}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}}_{SC_{tot:W}} = \underbrace{\hat{\boldsymbol{\beta}}\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}}_{SC_{reg:W}} + \underbrace{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})' \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}_{SC_{res:W}}.$$

Les degrés de liberté sont calculés de la même façon qu'en régression linéaire ordinaire.

Le choix des poids w_i est en général difficile, surtout en régression linéaire multiple. En régression linéaire simple, on peut faire un graphe des paires r_i, x_i, et voir comment la variance des résidus studentisés évolue en fonction de x_i. Si le graphe est en forme d'entonnoir ouvert vers la droite, alors on pourra prendre $w_i = 1/x_i$, $w_i = 1/\sqrt{x_i}$, $w_i = x_i^{-2}$, etc. Si l'entonnoir ouvre vers la gauche, alors on peut prendre $w_i = x_i$, $w_i = \sqrt{x_i}$, $w_i = x_i^2$, etc. Notez que ce choix des poids est relativement simple à effectuer en régression linéaire simple, mais virtuellement impossible à faire en régression linéaire multiple. Il est donc recommandé d'essayer de tranformer Y pour stabiliser la variance des résidus. Si cette transformation ne règle pas le problème, alors on peut essayer la régression pondérée.

3.21 Données aberrantes et influentes

Définition 3.13. *Les données aberrantes sont des valeurs qui ne cadrent pas avec le modèle*

Définition 3.14. *Les données influentes sont des observations qui, à elles seules, peuvent avoir un impact sérieux sur les conclusions d'une analyse de régression.*

Parmi les exemples « d'impact sérieux », on peut citer

- Des $\hat{\beta}$ négatifs quand on sait qu'ils devraient être positifs.
- Une variable exogène qui devrait être importante considérée non significative.
- Un modèle, bien que scientifiquement raisonnable, est incapable de prédire correctement.

3.21.1 Sources de l'influence

L'influence d'une donnée provient souvent d'une combinaison de deux sources :

Source 1 : Donnée dont la valeur de $\boldsymbol{x}_i$ est loin du centre (centroïde) de toutes les données.

Source 2 : Donnée ayant une erreur produisant une valeur extrême de Y.

3.21.2 Diagnostics : Résidus et la matrice chapeau

Une combinaison des sources 1 et 2 peut rendre une donnée influente.
Pour mesurer la contribution de la source 1, il faut mesurer la distance d'une donnée au centroïde de toutes les données. Pour la $i$ème donnée, cette distance est une fonction de h_{ii}, le $i$ème élément diagonal de la matrice chapeau $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, qui peut s'exprimer comme $h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$.

Définition 3.15. *On dit que la $i$ème donnée possède beaucoup de levier lorsque cette donnée est située loin du centroïde des données, c'est-à-dire lorsque la valeur de h_{ii} est élevée.*

Qu'est-ce qu'une grande valeur de h_{ii} ? Il est difficile de répondre à cette question de façon générale. Cependant nous savons que $\text{tr}(H) = \sum_{i=1}^{n} h_{ii} = p$, où p est le nombre de paramètres dans le modèle. Donc un h_{ii} moyen prend une valeur de p/n et un h_{ii} qui se démarque nettement de cette valeur signifie une donnée à grand levier. Belsley et al. (1980) suggèrent de retenir les valeurs de $h_{ii} > 2p/n$.

Un grand levier n'implique pas nécessairement une influence. Tout diagnostic d'influence doit être basé sur une combinaison de h_{ii} et R_i.

Définition 3.16. *Le $i$ème résidu R-Student est donné par*

$$T_i = \frac{\hat{\epsilon}_i}{s_{-i}\sqrt{1-h_{ii}}} \sim \mathcal{T}_{n-p-2},$$

où

$$s_{-i} = \sqrt{\frac{(n-p)s^2 - \hat{\epsilon}_i^2/(1-h_{ii})}{n-p-1}}.$$

L'indice $(-i)$ indique que la valeur (dans cas l'écart-type) a été calculée sans l'observation i. Comme on peut le voir, ce résidu est une fonction (croissante) à la fois en h_{ii} et en $\hat{\epsilon}_i$.

Belsley et al. (1980) prônent de prendre les valeurs de $|T_i| > t_{n-p-1}(1-\alpha^*/2)$, où $\alpha^* = \alpha/n$ (le seuil corrigé de Bonferroni) comme un bon indicateur d'une valeur influente. D'autres statisticiens suggèrent de retenir les valeurs de $|T_i| > 2$. En résumé, nous devons prendre garde aux observations pour lesquelles $|T_i|$ et/ou h_{ii} prennent une grande valeur.

3.21.3 Mesures d'influence

Les h_{ii} et T_i nous disent que la $i$ème donnée est peut-être influente, mais son influence sur les estimés et les prévisions n'a pas été calculée. Or pour mesurer si une donnée est influente ou non, on évalue son impact sur la prévision ($\hat{\boldsymbol{Y}}$) et sur l'estimation de β. Les statistiques utilisées sont présentées dans les sections subséquentes.

Influence sur la prévision

La statistique suivante mesure l'influence de la $i$ème donnée dans la prévision $\hat{Y}_i$ de Y_i :

$$DFFIT_i = \frac{\hat{Y}_i - \hat{Y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}} = T_i\sqrt{\frac{h_{ii}}{1-h_{ii}}},$$

où $\hat{Y}_{i,-i}$ désigne la valeur de la prévision de Y_i basée sur un jeu de donnée qui ne contient pas la $i$ème donnée, c'est-à-dire $\hat{Y}_{i,-i} = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{-i}$. Il faut se rappeler ici que $\mathbb{V}\text{ar}[Y_i] = \sigma^2 h_{ii}$. Les différences dans la statistique $DFFIT_i$ sont standardisées de sorte qu'elles s'expriment en multiple de l'écart-type pertinent.

Nous analysons les $DFFIT_i$ de la même façon que nous nous nous penchons sur les T_i, c'est-à-dire que des grandes valeurs de $|DFFIT_i|$ constituent un signe que la $i$ème donnée possède une grande influence sur la valeur de la prévision. Belsley et al. (1980) indiquent qu'il faut retenir les valeur de $|DFFIT_i| > 2\sqrt{p/n}$ alors que d'autres statisticiens proposent simplement de retenir les valeurs de $|DFFIT_i| > 2$.

Influence sur les estimations des paramètres

La statistique suivante mesure l'influence de la $i$ème observation sur la valeur de l'estimateur de β_j :

$$DFBETA_{ji} = \frac{\hat{\beta}_j - \hat{\beta}_{j,-i}}{s_{-i}\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}}}.$$

Il faut se souvenir ici que $\mathbb{V}\text{ar}[\hat{\beta}_j] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}$. Les différences au numérateur de la statistique sont par conséquent à nouveau standardisées. Encore une fois, les grandes valeurs de $DFBETA_{ji}$ sont révélateurs que la $i$ème donnée possède une grande influence sur l'estimation de β_j. Belsley et al. (1980) proposent de considérer les valeurs de $|DFBETA_{ji}| > 2/\sqrt{n}$ alors que d'autres statisticiens suggèrent simplement de retenir les valeurs de $|DFBETA_{ji}| > 2$.
Le signe de $DFBETA_{ji}$ est un bon indicateur. En effet, si un certain $\hat{\beta}_j$ est négatif alors qu'on se serait attendu à une valeur positive, alors un $DFBETA_{ji}$ très négatif indique que le « coupable »pourrait être l'observation i.

Comme nous obtenons p valeurs de $DFBETA$ pour chaque observation, cette mesure devient peu pratique lorsque nous sommes en présence d'un jeu de données ayant plusieurs variables exogènes. Pour

simplifier la recherche de valeurs influentes, Dennis Cook a proposé la statistique suivante, appelée distance de Cook, qui permet de combiner les $DFBETA$ d'une même observation :

$$\begin{aligned} C_i &= \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})}{p\hat{\sigma}^2} \\ &= \frac{(\hat{Y}_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{-i})^2}{p\hat{\sigma}^2} \\ &= \frac{\hat{\epsilon}_i^2}{(1-h_{ii})^2}\frac{h_{ii}}{\hat{\sigma}^2 p} \end{aligned}$$

Une grande valeur de C_i indique une observation ayant une influence prépondérante sur les β. Fox (1991) suggère de porter notre attention sur les valeurs de $C_i > \frac{4}{n-p}$. Il s'agit par la suite d'examiner les $DFBETA$ pour cette observation pour savoir quel(s) $\hat{\beta}$ est (sont) affecté(s).

Influence sur la performance des estimateurs

Une donnée peut aussi avoir un impact néfaste sur la variance des estimateurs. Soit $GV = ||Var(\hat{\boldsymbol{\beta}})||$, le déterminant de la matrice de variance de $\hat{\boldsymbol{\beta}}$; on appelle parfois GV variance généralisée. L'influence de la $i$ème donnée sur la variance de $\hat{\boldsymbol{\beta}}$ peut être calculée comme suit :

$$\begin{aligned} COVRATIO_i &= \frac{||(\boldsymbol{X}'_{-i}\boldsymbol{X}_{-i})^{-1}s^2_{-i}||}{||(\boldsymbol{X}'\boldsymbol{X})^{-1}s^2||} \\ &= \frac{s^{2p}_{-i}}{s^{2p}}\left(\frac{h_{ii}}{1-h_{ii}}\right). \end{aligned}$$

Comme la seconde ligne l'illustre, une donnée possédant beaucoup de levier donnera un $COVRATIO > 1$, alors qu'une donnée aberrante donnera un $COVRATIO < 1$.
Belsley et al. (1980) proposent de retenir notre attention sur les valeurs telles que $|COVRATIO - 1| > 3p/n$, c'est-à-dire sur les observations telle que $COVRATIO < 1 - 3p/n$ ou $COVRATIO > 1 + 3p/n$.

3.21.4 Que faire avec les données influentes ?

Si une donnée est influente, il faut prendre tous les moyens possibles pour essayer de comprendre comment cette donnée a été obtenue et s'assurer qu'il n'y a pas d'erreur qui aurait pu survenir lors de la mesure ou collection des données ou lors de la saisie des données. Si l'influence provient d'un grand levier, alors il serait souhaitable d'aller chercher d'autres données afin de combler le vide entre le centroïde des données et la donnée influente.

En général, on doit s'abstenir d'éliminer les données influentes d'un jeu de données. Quelques exceptions :

– Erreur de mesure ou lors de la saisie (corriger la donnée est préférable à son élimination).

– Valeurs mesurées scientifiquement peu plausibles.
– Donnée très peu représentative du reste des données et de la population en général.

Il est à noter que souvent, une transformation de variable peut grandement réduire l'influence de certaines données.

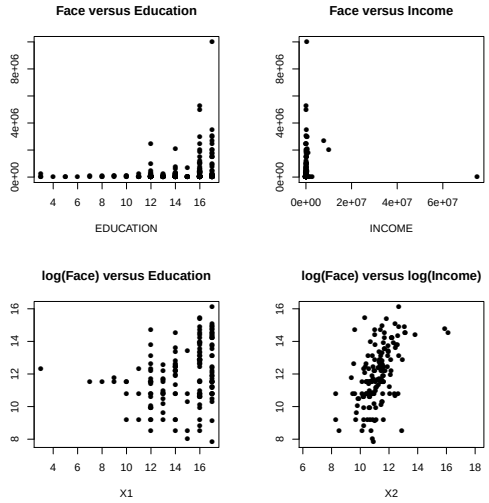
3.22 Exemple : Assurance vie temporaire

Cet exemple est tiré de Frees (2009). Comme toutes les compagnies, les compagnies d'assurance vie sont toujours en quête de nouveaux moyens pour amener leurs produits sur le marché. Les personnes impliquées dans le développement de produits (actuaires, agents de marketing, etc.) veulent comprendre "Qui achète de l'assurance et combien en achètent-ils ?" Les caractéristiques des clients actuels de la compagnie sont facilement disponibles en fouillant dans la base de données de la compagnie. Pour augmenter sa part de marché, la compagnie de s'intéresser aux clients potentiels, c'est-à-dire ceux qui n'ont pas d'assurance avec la compagnie.

On a accès au données du *Survey of Consumer Finances* (SCF), un échantillon représentatif de la population des états-Unis contenant des infos sur les actifs, dettes, revenus et autres caractéristiques démographiques. On utilise un échantillon aléatoire de 275 maisonnées avec des revenus et un montant d'assurance positifs.

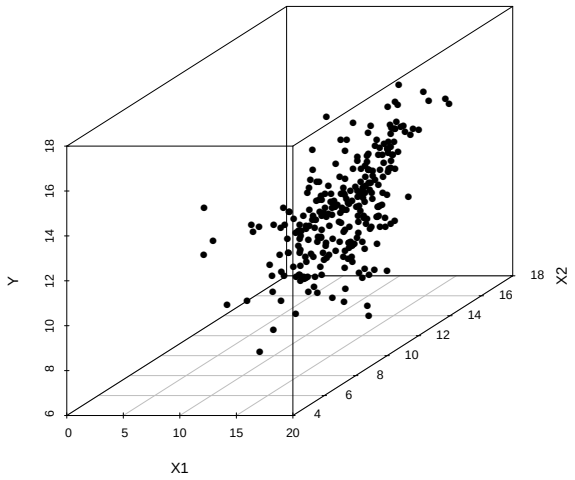
On s'intéresse donc à modéliser le montant d'assurance (FACE amount) qui représente le "besoin" en assurance. On utilise d'abord les variables explicatives du nombre d'années d'éducation du revenu annuel.

FIGURE 3.6: Graphiques univariés



Comme illustré sur les graphiques 3.6 et 3.7, la relation est linéaire seulement lorsqu'on prend le loga-

FIGURE 3.7: Graphique en nuage de points



rithme naturel des variables monétaires FACE et INCOME. Le modèle considéré est donc

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

où

- $Y = \log(FACE)$ est la variable réponse
- $x_1 = EDUCATION$ est une variable explicative discrète
- $x_2 = \log(INCOME)$ est une variable explicative continue
- ε est le terme d'erreur, on suppose que $\varepsilon \sim N(0, \sigma^2)$
- β_0, β_1 et β_2 sont les paramètres du modèle, à estimer

3.22.1 Estimation des paramètres

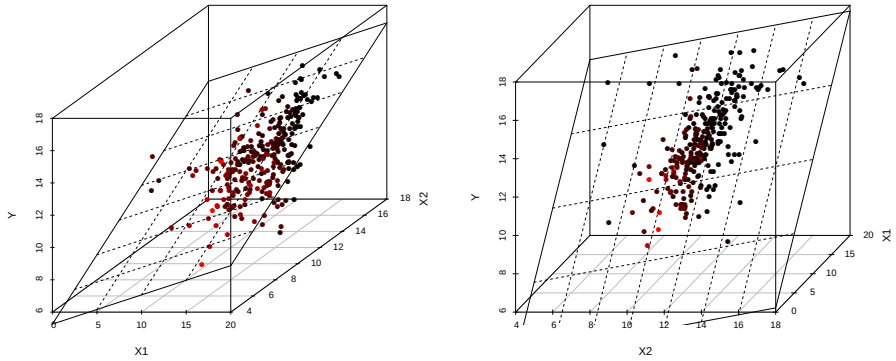
L'estimation des paramètres est effectuée avec la fonction `lm()` en `R`. On trouve les résultats suivants :

```
> summary(modele)
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1266 -1.0284  0.1817  0.9185  5.3403

Coefficients:
```

FIGURE 3.8: Plan de régression




```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.96235    0.87676   3.379 0.000835 ***
X1           0.18103    0.04003   4.523 9.11e-06 ***
X2           0.57392    0.07879   7.284 3.50e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.587 on 272 degrees of freedom
Multiple R-squared: 0.2859,Adjusted R-squared: 0.2806
F-statistic: 54.44 on 2 and 272 DF,  p-value: < 2.2e-16
```

La Figure 3.8 montre le plan de régression estimé selon deux angles différents.

3.22.2 Interprétation des paramètres

On a trouvé que $\hat{\beta}_0 = 2.96235$, $\hat{\beta}_1 = 0.18103$ et $\hat{\beta}_2 = 0.57392$.

- Dans ce cas, on ne peut pas vraiment interpréter β_0 puisque le nombre d'années d'éducation est toujours supérieur à 0.
- Lorsque le nombre d'années de scolarité augmente de 1 et que le revenu reste constant, l'espérance du log du montant nominal d'assurance augmente de 0.18103.
- Lorsque le log du revenu augmente de 1 et que le nombre d'année de scolarité reste constant, l'espérance du log du montant nominal d'assurance augmente de 0.5739.

Ces interprétations ne parlent pas beaucoup !

On peut essayer de réécrire le modèle pour mieux comprendre ces paramètres.

$$\begin{aligned}\log(FACE) &= \beta_0 + \beta_1 x_1 + \beta_2 \log(INCOME) + \varepsilon \\ FACE &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 \log(INCOME) + \varepsilon\right) \\ &= e^{\beta_0} e^{\beta_1 x_1} INCOME^{\beta_2} e^{\varepsilon}\end{aligned}$$

Ce modèle multiplicatif est plus simple à interpréter. On a donc

$$\begin{aligned}\widehat{FACE} &= e^{\hat{\beta}_0} e^{\hat{\beta}_1 x_1} INCOME^{\hat{\beta}_2} \mathrm{E}[\mathrm{e}^{\varepsilon}] \\ &= 19.34 \exp(0.18 x_1) INCOME^{0.57} \times \text{erreur}\end{aligned}$$

Si le nombre d'année d'éducation augmente d'une année, on a

$$\begin{aligned}\widehat{FACE} &= 19.34 \exp(0.18(x_1+1)) INCOME^{0.57} \times \text{erreur} \\ &= 19.34 \exp(0.18 x_1) INCOME^{0.57} \times \text{erreur} \times \exp(0.18) \\ &= 19.34 \exp(0.18 x_1) INCOME^{0.57} \times \text{erreur} \times 1.1984\end{aligned}$$

Cela signifie que l'impact d'augmenter le nombre d'années de scolarité de 1, en tenant le revenu constant, est une augmentation d'environ 20% sur le montant nominal d'assurance.

Si le revenu augmente de 10%, on a

$$\begin{aligned}\widehat{FACE} &= 19.34 \exp(0.18 x_1)(1.1 INCOME)^{0.57} \times \text{erreur} \\ &= 19.34 \exp(0.18 x_1) INCOME^{0.57} \times \text{erreur} \times 1.1^{0.57} \\ &= 19.34 \exp(0.18 x_1) INCOME^{0.57} \times \text{erreur} \times 1.056\end{aligned}$$

Cela signifie que l'impact d'augmenter le revenu de 10%, en tenant le nombre d'année de scolarité constant, est une augmentation d'environ 5.6% sur le montant nominal d'assurance.

3.22.3 Intervalles de confiance pour les prévisions

On s'intéresse à :

1. un individu ayant un revenu de 65000$ et 14 ans de scolarité,
2. un individu avec un revenu de 100000$ et 14 ans de scolarité,
3. un individu avec un revenu de 65000$ et 16 ans de scolarité.

```
> newdat <- data.frame(X1=c(14,14,16),X2=log(c(65000,100000,65000)))
> lnmoy <- predict(modele,newdat,interval="confidence")
> exp(lnmoy)[,2:3]
       lwr      upr
1 116392.1 171000.5
2 147185.7 221717.6
3 162046.5 253377.9
> lnmoy <- predict(modele,newdat,interval="prediction")
> exp(lnmoy)[,2:3]
       lwr     upr
1 6169.921 3225828
2 7894.189 4133884
3 8843.518 4642838
```

3.22.4 Inclusion de plus de variables dans le modèle

Dans la base de données, on a de l'information supplémentaire qui pourrait être incluse dans le modèle. Les variables à considérer sont :

- `GENDER` : Sexe de l'assuré, 0 si femme, 1 si homme
- `AGE`
- `NUMHH` : Nombre de personne dans la famille
- `CHARITY` : Dons de charité
- On s'intéresse aussi à l'intéraction entre le nombre d'années de scolarité et le revenu, puisqu'on peut s'attendre à ce qu'un individu ayant plus d'années de scolarité gagne un plus gros revenu.

On trouve

```
> modtot <- lm(Y~X1*X2+GENDER+AGE+NUMHH+CHARITY,data=dat)
> summary(modtot)
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.678e+00  6.813e+00  -1.274 0.203866
X1           1.008e+00  4.380e-01   2.301 0.022178 *
X2           1.502e+00  6.304e-01   2.383 0.017885 *
GENDER       7.915e-01  2.821e-01   2.805 0.005397 **
AGE         -4.024e-03  7.838e-03  -0.513 0.608058
NUMHH        2.336e-01  6.776e-02   3.447 0.000659 ***
CHARITY      7.789e-06  2.376e-06   3.278 0.001184 **
X1:X2       -7.348e-02  4.044e-02  -1.817 0.070349 .
---

Residual standard error: 1.482 on 267 degrees of freedom
Multiple R-squared: 0.3886,Adjusted R-squared: 0.3726
F-statistic: 24.24 on 7 and 267 DF,  p-value: < 2.2e-16
```

3.22.5 Analyse de la variance

```
> anova(modtot)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 140.55 140.549 64.0087 3.796e-14 ***
X2          1 133.58 133.583 60.8363 1.400e-13 ***
GENDER      1  34.26  34.262 15.6035 0.0001000 ***
AGE         1   7.41   7.414  3.3767 0.0672358 .
NUMHH       1  30.13  30.128 13.7209 0.0002577 ***
CHARITY     1  19.44  19.440  8.8536 0.0031931 **
X1:X2       1   7.25   7.249  3.3012 0.0703491 .
Residuals 267 586.27   2.196
```

Ce n'est pas le tableau auquel on s'attendait ! En fait, `anova()` donne l'analyse de la variance détaillée. Il faut sommer les lignes correspondant aux variables explicatives pour trouver SSR.

Le tableau ANOVA est donc :

Source	Df	Sum Sq	Mean Sq	F value	Pr($> F$)
Regression	7	372.62	53.23211	24.24049	1.862564e-25
Error	267	586.27	2.196		

3.22.6 Test F Partiel

On désire tester si le modèle qui contient seulement les variables `EDUCATION` et `INCOME` est une simplification adéquate du modèle complet.

H_0 : Le modèle est `EDUCATION+INCOME`

H_1 : Les modèle est `EDUCATION*INCOME+GENDER+AGE+NUMHH+CHARITY`

La statistique est

$$F_{obs} = \frac{(SSE^{(0)} - SSE^{(1)})/\Delta_{dl}}{MSE^{(1)}} = \frac{(684.76 - 586.27)/5}{2.196} = 8.97.$$

Le quantile supérieur à 5% de la loi $\mathcal{F}(6, 267)$ est 2.132619. Comme $F_{obs} > 2.13$, on rejette l'hypothèse nulle en faveur de l'hypothèse alternative. Cependant, peut-être que nous pourrions enlever moins de variables à la fois...

On désire tester si la variable `AGE` devrait être conservée dans le modèle.

$H_0 : \beta_{AGE} = 0$

$H_1 : \beta_{AGE} \neq 0$

On doit ajuster le modèle sous H_0 pour trouver $SSE^{(0)}$:

```
> mod0 <- update(modtot,~.-AGE)
> anova(mod0)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 140.55 140.549  64.185 3.492e-14 ***
X2          1 133.58 133.583  61.004 1.293e-13 ***
GENDER      1  34.26  34.262  15.646 9.781e-05 ***
NUMHH       1  37.04  37.039  16.915 5.205e-05 ***
CHARITY     1  18.83  18.827   8.598  0.003656 **
X1:X2       1   7.79   7.787   3.556  0.060411 .
Residuals 268 586.85   2.190
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La statistique est

$$F_{obs} = \frac{(SSE^{(0)} - SSE^{(1)})/\Delta_{dl}}{MSE^{(1)}} = \frac{(586.85 - 586.27)/1}{2.196} = 0.2641.$$

Le quantile supérieur à 5% de la loi $\mathcal{F}(1, 267)$ est 3.876522. Comme $F_{obs} < 3.87$, on ne peut pas rejeter l'hypothèse nulle que $\beta_{AGE} = 0$. Cela signifie que la variable explicative AGE n'est pas nécessaire dans le modèle et qu'elle n'apporte pas vraiment d'informations sur le besoin en assurance.

Il y a peut-être d'autres variables que nous pourrions enlever pour simplifier le modèle. Toutefois, il serait fastidieux de faire tous les tests F partiels possibles !

3.22.7 Intervalles de confiance pour les prévisions

On rappelle le modèle

$$\log(FACE) = \beta_0 + \beta_1 EDUCATION + \beta_2 \log(INCOME) + \varepsilon.$$

On s'intéresse à :

1. un individu ayant un revenu de 65000$ et 14 ans de scolarité,
2. un individu avec un revenu de 100000$ et 14 ans de scolarité,
3. un individu avec un revenu de 65000$ et 16 ans de scolarité.

Les I.C. pour les prévisions sont calculées en `R` :

```
> newdat <- data.frame(X1=c(14,14,16),X2=log(c(65000,100000,65000)))
> lnmoy <- predict(modele,newdat,interval="confidence")
> exp(lnmoy)[,2:3]
       lwr      upr
1 116392.1 171000.5
2 147185.7 221717.6
3 162046.5 253377.9
> lnmoy <- predict(modele,newdat,interval="prediction")
> exp(lnmoy)[,2:3]
       lwr     upr
1 6169.921 3225828
2 7894.189 4133884
3 8843.518 4642838
```

3.22.8 Inclusion de plus de variables dans le modèle

Dans la base de données, on a de l'information supplémentaire qui pourrait être incluse dans le modèle. Les variables à considérer sont :

- `GENDER` : Sexe de l'assuré, 0 si femme, 1 si homme
- `AGE`
- `NUMHH` : Nombre de personne dans la famille
- `CHARITY` : Dons de charité
- On s'intéresse aussi à l'intéraction entre le nombre d'années de scolarité et le revenu, puisqu'on peut s'attendre à ce qu'un individu ayant plus d'années de scolarité gagne un plus gros revenu.

On trouve :

```
> modtot <- lm(Y~X1*X2+GENDER+AGE+NUMHH+CHARITY,data=dat)
> summary(modtot)
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.678e+00  6.813e+00  -1.274 0.203866
X1           1.008e+00  4.380e-01   2.301 0.022178 *
X2           1.502e+00  6.304e-01   2.383 0.017885 *
GENDER       7.915e-01  2.821e-01   2.805 0.005397 **
AGE         -4.024e-03  7.838e-03  -0.513 0.608058
NUMHH        2.336e-01  6.776e-02   3.447 0.000659 ***
CHARITY      7.789e-06  2.376e-06   3.278 0.001184 **
X1:X2       -7.348e-02  4.044e-02  -1.817 0.070349 .
---

Residual standard error: 1.482 on 267 degrees of freedom
Multiple R-squared: 0.3886,Adjusted R-squared: 0.3726
F-statistic: 24.24 on 7 and 267 DF,  p-value: < 2.2e-16
```

3.22.9 Analyse de la variance

```
> anova(modtot)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 140.55 140.549 64.0087 3.796e-14 ***
X2          1 133.58 133.583 60.8363 1.400e-13 ***
GENDER      1  34.26  34.262 15.6035 0.0001000 ***
AGE         1   7.41   7.414  3.3767 0.0672358 .
NUMHH       1  30.13  30.128 13.7209 0.0002577 ***
CHARITY     1  19.44  19.440  8.8536 0.0031931 **
X1:X2       1   7.25   7.249  3.3012 0.0703491 .
Residuals 267 586.27   2.196
```

Ce n'est pas le tableau auquel on s'attendait ! En fait, `anova()` donne l'analyse de la variance détaillée. Il faut sommer les lignes correspondant aux variables explicatives pour trouver SSR.

Le tableau ANOVA est donc :

Source	Df	Sum Sq	Mean Sq	F value	$\Pr(>F)$
Regression	7	372.62	53.23211	24.24049	1.862564e-25
Error	267	586.27	2.196		

3.22.10 Test F partiel

On désire tester si le modèle qui contient seulement les variables `EDUCATION` et `INCOME` est une simplification adéquate du modèle complet.

H_0 : Le modèle est `EDUCATION+INCOME`

H_1 : Les modèle est `EDUCATION*INCOME+GENDER+AGE+NUMHH+CHARITY`

La statistique est

$$F_{obs} = \frac{(SSE^{(0)} - SSE^{(1)})/\Delta_{dl}}{MSE^{(1)}} = \frac{(684.76 - 586.27)/5}{2.196} = 8.97.$$

Le quantile supérieur à 5% de la loi $\mathcal{F}(6, 267)$ est 2.132619. Comme $F_{obs} > 2.13$, on rejette l'hypothèse nulle en faveur de l'hypothèse alternative. Cependant, peut-être que nous pourrions enlever moins de variables à la fois...

On désire tester si la variable `AGE` devrait être conservée dans le modèle.

$H_0 : \beta_{AGE} = 0$

$H_1 : \beta_{AGE} \neq 0$

On doit ajuster le modèle sous H_0 pour trouver $SSE^{(0)}$:

```
> mod0 <- update(modtot,~.-AGE)
> anova(mod0)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 140.55 140.549  64.185 3.492e-14 ***
X2          1 133.58 133.583  61.004 1.293e-13 ***
GENDER      1  34.26  34.262  15.646 9.781e-05 ***
NUMHH       1  37.04  37.039  16.915 5.205e-05 ***
CHARITY     1  18.83  18.827   8.598  0.003656 **
X1:X2       1   7.79   7.787   3.556  0.060411 .
Residuals 268 586.85   2.190
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La statistique est

$$F_{obs} = \frac{(SSE^{(0)} - SSE^{(1)})/\Delta_{dl}}{MSE^{(1)}} = \frac{(586.85 - 586.27)/1}{2.196} = 0.2641.$$

Le quantile supérieur à 5% de la loi $\mathcal{F}(1, 267)$ est 3.876522. Comme $F_{obs} < 3.87$, on ne peut pas rejeter l'hypothèse nulle que $\beta_{AGE} = 0$. Cela signifie que la variable explicative AGE n'est pas nécessaire dans le modèle et qu'elle n'apporte pas vraiment d'informations sur le besoin en assurance.

Il y a peut-être d'autres variables que nous pourrions enlever pour simplifier le modèle. Toutefois, il serait fastidieux de faire tous les tests F partiels possibles !

3.22.11 Sélection de variables

Avec le package `leaps` en `R`, on peut connaître les "meilleurs" modèles pour chaque nombre de paramètres :

```
library(leaps)
Xmat <- cbind(X1,X2,X1X2=X1*X2,GENDER,AGE,NUMHH,CHARITY)
allsubsets<-regsubsets(Xmat,Y,nbest=2)
stats <- summary(allsubsets)
```

```
> stats$which
  (Intercept)    X1    X2  X1X2 GENDER   AGE NUMHH CHARITY
1        TRUE FALSE FALSE  TRUE  FALSE FALSE FALSE   FALSE
```

```
1        TRUE FALSE  TRUE FALSE  FALSE FALSE FALSE   FALSE
2        TRUE FALSE FALSE  TRUE  FALSE FALSE  TRUE   FALSE
2        TRUE FALSE FALSE  TRUE   TRUE FALSE FALSE   FALSE
3        TRUE FALSE FALSE  TRUE  FALSE FALSE  TRUE    TRUE
3        TRUE FALSE FALSE  TRUE   TRUE FALSE  TRUE   FALSE
4        TRUE FALSE FALSE  TRUE   TRUE FALSE  TRUE    TRUE
4        TRUE  TRUE  TRUE FALSE  FALSE FALSE  TRUE    TRUE
5        TRUE  TRUE  TRUE FALSE   TRUE FALSE  TRUE    TRUE
5        TRUE FALSE FALSE  TRUE   TRUE  TRUE  TRUE    TRUE
6        TRUE  TRUE  TRUE  TRUE   TRUE FALSE  TRUE    TRUE
6        TRUE  TRUE  TRUE FALSE   TRUE  TRUE  TRUE    TRUE
7        TRUE  TRUE  TRUE  TRUE   TRUE  TRUE  TRUE    TRUE
```

À partir de cela, on peut comparer les meilleurs modèles pour chaque nombre de paramètre selon les critères de sélection de modèle PRESS, C_p de Mallows, R^2 ajusté, AIC et BIC.

```
> npar <- rowSums(stats$which)
> cbind(npar,df=275-npar,Cp=stats$cp,adjR2=stats$adjr2,
+bic=stats$bic,aic=stats$bic+(2-log(275))*npar)
  npar  df        Cp     adjR2        bic        aic
1    2 273 52.038999 0.2575645  -71.67224  -78.90579
1    2 273 64.311054 0.2293598  -61.41871  -68.65226
2    3 272 23.328732 0.3256754  -93.52635 -104.37666
2    3 272 32.878863 0.3036458  -84.68596  -95.53628
3    4 271 15.201408 0.3466343  -97.60551 -112.07259
3    4 271 15.684055 0.3455169  -97.13558 -111.60266
4    5 270  8.310388 0.3648756 -100.79229 -118.87615
4    5 270 13.673472 0.3524128  -95.44832 -113.53217
5    6 269  7.809905 0.3683468  -97.70304 -119.40367
5    6 269  9.737106 0.3638517  -95.75295 -117.45358
6    7 268  6.263636 0.3742923  -95.71120 -121.02859
6    7 268  9.301230 0.3671808  -92.60330 -117.92070
7    8 267  8.000000 0.3725683  -90.36583 -119.30000
```

On peut aussi utiliser les méthodes algorithmiques :

```
>library(car)
>modbackward <- stepAIC(modtot,modtot,direction="backward")
>modstepwise <- stepAIC(modtot,modtot,direction="both")
```

On obtient le même résultat avec les méthodes “backward” et “stepwise”.

```
> modbackward
Call:
lm(formula = Y ~ X1 + X2 + GENDER + NUMHH + CHARITY + X1:X2,
    data = dat)
Coefficients:
(Intercept)           X1           X2       GENDER        NUMHH
 -9.234e+00    1.031e+00    1.535e+00    7.770e-01    2.453e-01
  CHARITY        X1:X2
7.721e-06   -7.572e-02

> modstepwise
Call:
lm(formula = Y ~ X1 + X2 + GENDER + NUMHH + CHARITY + X1:X2,
    data = dat)
Coefficients:
(Intercept)           X1           X2       GENDER        NUMHH
 -9.234e+00    1.031e+00    1.535e+00    7.770e-01    2.453e-01
     CHARITY      X1:X2
   7.721e-06 -7.572e-02
```

3.22.12 Graphiques des variables ajoutées

Les graphiques de variables ajoutées peuvent être calculées directement en R avec la fonction suivante. Le résultat est présenté dans la Figure 3.9.

```
library(car)
avPlots(modtot)
```

3.22.13 Multicolinéarité

Le modèle que nous avons choisi présente un problème de multicolinéarité, dû à l'inclusion de l'intéraction entre l'éducation et le revenu :

```
> cor(Xmat,method="pearson")
                  X1          X2       X1X2       GENDER          AGE       NUMHH
X1         1.00000000 0.34270358 0.87477045 -0.044498132  0.091636944 -0.06352920
X2         0.34270358 1.00000000 0.75172774  0.229230443  0.045229132  0.17933542
X1X2       0.87477045 0.75172774 1.00000000  0.084434822  0.094680496  0.04447216
GENDER    -0.04449813 0.22923044 0.08443482  1.000000000  0.005044888  0.29694101
AGE        0.09163694 0.04522913 0.09468050  0.005044888  1.000000000 -0.31828356
NUMHH     -0.06352920 0.17933542 0.04447216  0.296941007 -0.318283557  1.00000000
CHARITY    0.15977611 0.39736590 0.32603836  0.091430865  0.072946776  0.11119192
```

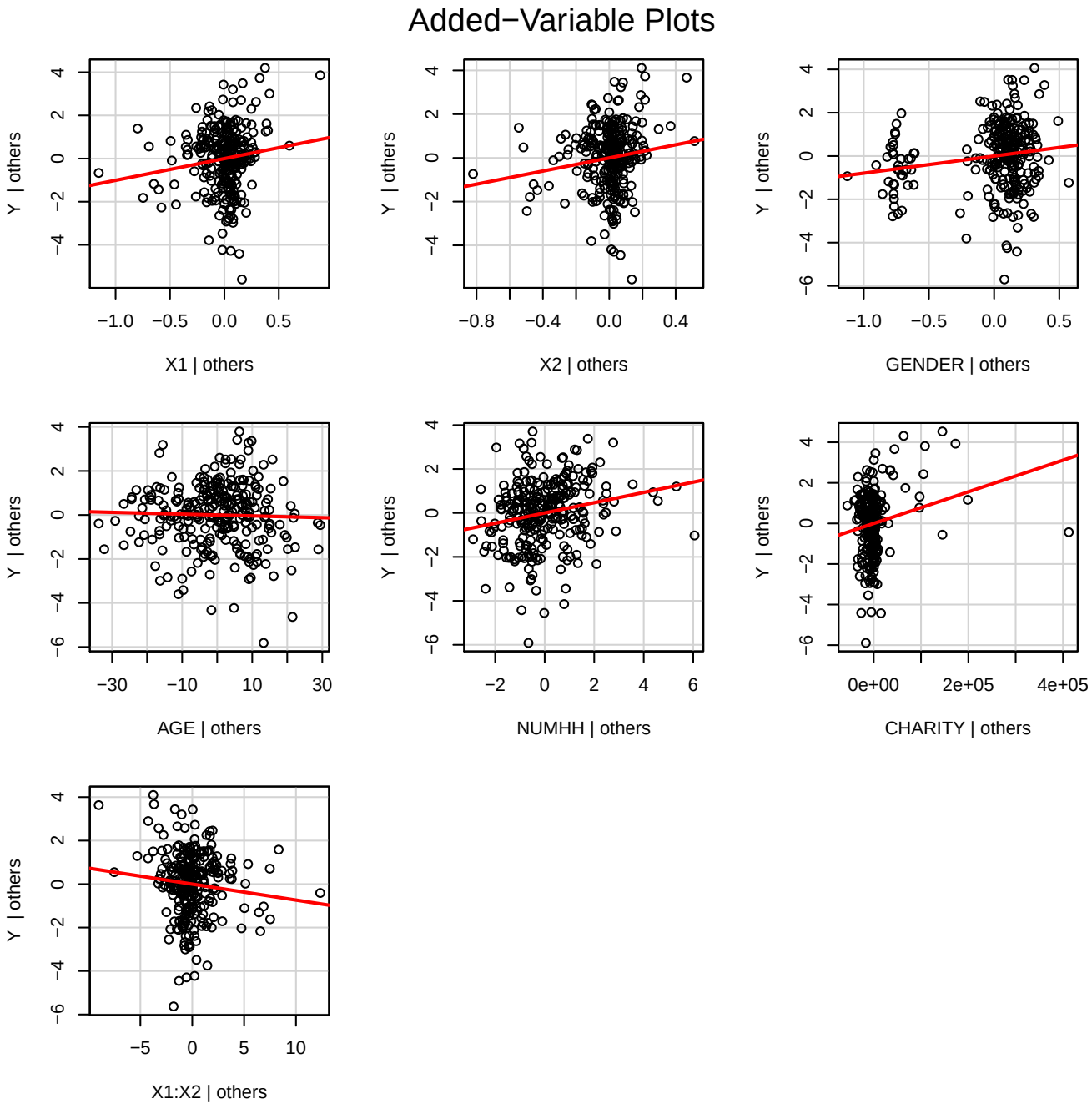


FIGURE 3.9: Graphiques des variables ajoutées

```
           CHARITY
X1       0.15977611
X2       0.39736590
X1X2     0.32603836
GENDER   0.09143087
AGE      0.07294678
NUMHH    0.11119192
CHARITY  1.00000000
```

```
> vif(modstepwise)
        X1         X2     GENDER      NUMHH    CHARITY      X1:X2
153.913115  82.290516   1.149294   1.131123   1.244584 313.255995
```

Le modèle final serait donc le modèle `EDUCATION+INCOME+GENDER+NUMHH+CHARITY` :

```
> modfin <- update(modstepwise,~.-X1:X2)
> vif(modfin)
      X1       X2   GENDER    NUMHH  CHARITY
1.167829 1.414947 1.148879 1.127508 1.191272
> summary(modfin)

Call:
lm(formula = Y ~ X1 + X2 + GENDER + NUMHH + CHARITY, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6982 -0.8362  0.1556  0.8836  4.0737

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.324e+00  8.880e-01   3.743 0.000222 ***
X1          2.137e-01  3.808e-02   5.611 4.99e-08 ***
X2          3.644e-01  8.250e-02   4.417 1.45e-05 ***
GENDER      7.871e-01  2.816e-01   2.795 0.005567 **
NUMHH       2.521e-01  6.389e-02   3.946 0.000102 ***
CHARITY     6.796e-06  2.329e-06   2.918 0.003816 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.487 on 269 degrees of freedom
Multiple R-squared: 0.3799,Adjusted R-squared: 0.3683
F-statistic: 32.96 on 5 and 269 DF,  p-value: < 2.2e-16
```

Bibliographie

Belsley, D. A., Kuh, E. et Welsch, R. E. (1980). *Regression diagnostics : identifying influential data and sources of collinearity*. John Wiley & Sons, New York-Chichester-Brisbane. Wiley Series in Probability and Mathematical Statistics.

Fox, J. (1991). *Regression Diagnostics*. Sage Publications, Newbury Park, California. Quantitative Applications in the Social Sciences Series No. 79.

Frees, E. W. (2009). *Regression modeling with actuarial and financial applications*. Cambridge University Press.