Régression linéaire

Prof. Armel Fabrice Yodé
Laboratoire de Mathématiques Appliquées et Informatique (L.M.A.I.)
UFR Mathématique et Informatique
Université de Cocody-Abidjan, Côte d'Ivoire
yafevrard@yahoo.fr

11 octobre 2019

Nous avons confiance en Dieu ; que tous les autres apportent des justificatifs. [Edwards Deming, Professeur de statistique, 1900-1993]

Chapitre 1

Introduction

1.1 Le Modèle de régression linéaire

Un modèle est une simple description d'un état ou d'un processus. Le modèle de régression linéaire est un modèle statistique défini par

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \tag{1.1.1}$$

où

- Y est une variable aléatoire réelle appelée variable à expliquer ou variable réponse ou variable dépendante ou variable endogène ;
- $X_1, \ldots, X_p$ sont des variables réelles également observées appelées variables explicatives ou prédicteurs ou variables exogènes ;
- $\beta_0, \ldots, \beta_p$ sont des paramètres non observés ;
- ε est le terme d'erreur ; c'est une variable aléatoire réelle non observée.

Le modèle de régression linéaire est dit simple si $p = 1$, c'est à dire,

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

Le modèle de régression linéaire est dit multiple si $p > 1$.

Exemple 1. Quelques exemples de problèmes :

1. le salaire en fonction de certaines caractéristiques socio-démographiques telles que l'ancienneté dans l'entreprise (en années), le nombre d'années d'études après le bac ;
2. le budget consacré à la consommation des ménages en fonction du revenu du ménage, du nombre de personnes par ménage ;

La régression linéaire a trois objectifs essentiels :

- mesurer l'impact ou l'effet de $X_1, \ldots, X_p$ sur Y
- prédire Y connaissant $X_1, \ldots, X_p$.
- parmi les variables $X_1, \ldots, X_p$, identifier celles qui expliquent de manière efficace (avec précision) la variable Y.

Remarque 1. Le modèle est linéaire en $\beta_0, \ldots, \beta_p$. Par exemple, le modèle ci-dessous

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \ln(X_2) + \varepsilon$$

est linéaire.

Chapitre 2

Régression linéaire simple

2.1 Modélisation

Définition 1. Le modèle de régression linéaire simple est défini par une équation de la forme

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad 1 \leq i \leq n.$$

Nous faisons les hypothèses suivantes :

$$\begin{cases} (\mathcal{H}_0) : \text{les variables } X_i \text{ sont non aléatoires} \\ (\mathcal{H}_1) : \mathbb{E}(\varepsilon_i) = 0 \ \ \forall i \in \{1, \ldots, n\} \\ (\mathcal{H}_2) : Cov(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \\ (\mathcal{H}_3) : \varepsilon_1, \ldots, \varepsilon_n \text{ sont indépendantes et } \varepsilon_i \hookrightarrow \mathcal{N}(0, \sigma^2) \text{ pour tout } i \in \{1, \ldots, n\} \end{cases}$$

Remarque 2.
- l'hypothèse $(\mathcal{H}_1)$ signifie que les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont de moyenne nulle, ou autrement dit, on ne se trompe pas en moyenne ;
- l'hypothèse $(\mathcal{H}_2)$ signifie que les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont non corrélés ($cov(\varepsilon_i, \varepsilon_j) = 0 \, \forall i \neq j$) et de même variance (Homoscédasticité) ($var(\varepsilon_i) = \sigma^2$, $\forall i = 1, \ldots, n$) ;
- l'hypothèse $(\mathcal{H}_3)$ que les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont gaussiennes, centrées et de même variance
- $(\mathcal{H}_3)$ implique $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$

2.1.1 Démarche de la régression

- **Vérifier la possibilté d'une liaison linéaire entre Y et X**

- **nuage de points** : réaliser un graphique cartésien, dont l'abscisse représente X et l'ordonnée Y. Dans ce repère, chaque individu i est représenté par un point de coordonnées (X_i, Y_i). L'ensemble des individus constitue un nuage de points dont la forme révèle la liaison entre les deux variables.

- le **coefficient de corrélation linéaire** a pour objet de quantifier l'allure plus ou moins linéaire d'un nuage de points. Il est défini par

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}}.$$

  ρ est toujours compris entre -1 et 1, valeurs atteintes lorsque la liaison linéaire est parfaite ; $\rho = 0$ ne signifie pas que X et Y sont indépendantes ; dans ce cas, X et Y sont dites linéairement indépendantes ou non corrélées.

- **Estimation des paramètres β_0, β_1 et σ^2 :** on utilisera la méthode des moindres carrés ordinaires ou la méthode du maximum de vraisemblance selon la nature des hypothèses sur les erreurs.

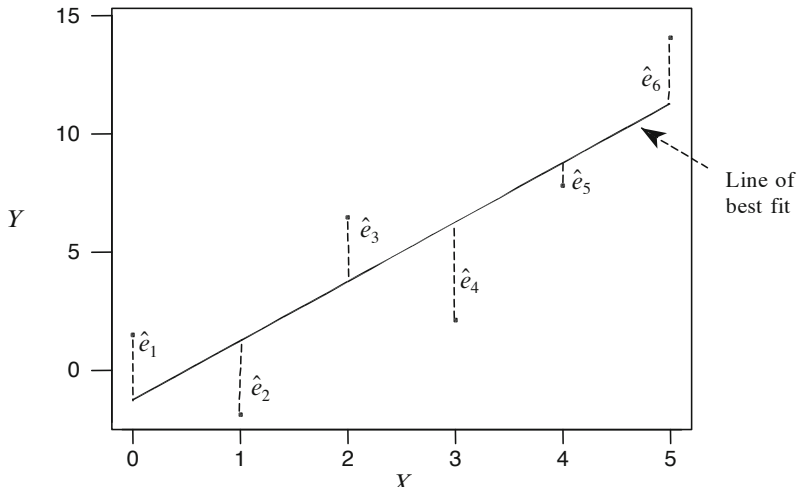
- **Validation du modèle :** cette étape permet de vérifier la validité des hypothèses du modèle ; coefficient de détermination, validité marginale de Student, analyse des résidus.

2.2 Estimateurs des moindres carrés

2.2.1 Définitions

Définition 2. On appelle estimateurs des moindres carrés de β_0 et β_1, les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ obtenus par minimisation de la quantité

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2.$$



Le vecteur $(\hat{\beta}_0, \hat{\beta}_1)'$ qui minimise $S(\beta_0, \beta_1)$ vérifie la condition du premier ordre

$$\begin{cases} \dfrac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \displaystyle\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \dfrac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \displaystyle\sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \end{cases}$$

On en déduit alors que

$$\hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} X_i Y_i - n \overline{X}_n \overline{Y}_n}{\displaystyle\sum_{i=1}^{n} X_i^2 - n \overline{X}_n^2} = \frac{\displaystyle\sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\displaystyle\sum_{i=1}^{n} (X_i - \overline{X}_n)^2}$$

$$\hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1 \overline{X}_n.$$

La condition de second ordre permet de vérifier aisément que $(\hat{\beta}_0, \hat{\beta}_1)'$ minimise $S(\beta_0, \beta_1)$.

2.2.2 Propriétés des estimateurs

On remarque que les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ peuvent se mettre sous la forme :

$$\hat{\beta}_1 = \sum_{i=1}^{n} \Big(\frac{X_i - \overline{X}_n}{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2} \Big) Y_i$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \Big(\frac{1}{n} - \frac{\overline{X}_n (X_i - \overline{X}_n)}{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2} \Big) Y_i.$$

Nous déduisons que $\hat{\beta}_0$ (respectivement $\hat{\beta}_1$) s'écrit comme une fonction linéaire des Y_i.

Proposition 1. *Sous les hypothèses $(\mathcal{H}_0)$, $(\mathcal{H}_1)$, $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 et β_1 respectivement i.e. $\mathbb{E}(\hat{\beta}_0) = \beta_0$ et $\mathbb{E}(\hat{\beta}_1) = \beta_1$.*

Proposition 2. *Sous les hypothèses $(\mathcal{H}_0)$, $(\mathcal{H}_1)$, $(\mathcal{H}_2)$, nous avons :*

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

$$var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}_n^2}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} \right).$$

Proposition 3. *Sous les hypothèses $(\mathcal{H}_0)$, $(\mathcal{H}_1)$, $(\mathcal{H}_2)$, nous avons :*

$$cov(\hat{\beta}_1, \overline{Y}_n) = 0.$$

Théorème 1. ***(Gauss-Markov)****Sous les hypothèses $(\mathcal{H}_0)$, $(\mathcal{H}_1)$, $(\mathcal{H}_2)$, parmi les estimateurs sans biais, fonctions linéaires des Y_i, les estimateurs des moindres carrés ordinaires $\hat{\beta}_0$ et $\hat{\beta}_1$ sont optimaux.*

Démonstration. Soit $\tilde{\beta}_1$ un autre estimateur sans biais de β_1, s'écrivant comme une fonction linéaire des Y_i i.e.

$$\tilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i \quad \mathbb{E}(\tilde{\beta}_1) = \beta_1.$$

Alors $\sum_{i=1}^{n} a_i = 0$ et $\sum_{i=1}^{n} a_i X_i = 1$. Nous pouvons écrire alors

$$\begin{aligned} var(\tilde{\beta}_1) &= var(\tilde{\beta}_1 - \hat{\beta}_1 + \hat{\beta}_1) \\ &= var(\tilde{\beta}_1 - \hat{\beta}_1) + var(\hat{\beta}_1) + 2Cov(\tilde{\beta}_1 - \hat{\beta}_1, \hat{\beta}_1). \end{aligned}$$

En vérifiant $Cov(\tilde{\beta}_1 - \hat{\beta}_1, \hat{\beta}_1) = 0$, on en déduit

$$var(\tilde{\beta}_1) \geq var(\hat{\beta}_1).$$

□

Définition 3. La droite de regression est déterminée par la formule

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Remarque 3. - Si nous évaluons la droite aux points X_i ayant servi à estimer les paramètres, nous obtenons des $\hat{Y}_i$ appelées **valeurs ajustées.**
- si nous évaluons la droite en des points n'ayant pas servi à l'estimation des paramètres, les valeurs obtenues seront appelées **valeurs prévues** ou **prévisions**
- La droite de régression passe par le centre de gravité $(\overline{X}_n, \overline{Y}_n)$.

2.2.3 Résidus et variance résiduelle

Les résidus $\hat{\varepsilon}_i$ sont les estimateurs des erreurs inconnus ε_i.

Définition 4. Les résidus sont définis par :

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i.$$

où $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ avec $1 \leq i \leq n$.

Proposition 4. *Dans un modèle de regression linéaire simple, nous avons :*

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$$

Démonstration. Car

$$\hat{\varepsilon}_i = (X_i - \overline{X}_n)(\beta_1 - \hat{\beta}_1) + \varepsilon_i - \overline{\varepsilon}_n.$$

□

Proposition 5. *La statistique* $\hat{\sigma}^2 = \dfrac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$ *est un estimateur sans biais de* σ^2.

Démonstration. Des égalités suivantes

$$\hat{\varepsilon}_i = \beta_0 + \beta_1 X_i + \varepsilon_i - \overline{Y}_n - \hat{\beta}_1 (X_i - \overline{X}_n)$$
$$\beta_0 = \overline{Y}_n - \beta_1 \overline{X}_n - \overline{\varepsilon}_n$$

nous déduisons

$$\hat{\varepsilon}_i = (X_i - \overline{X}_n)(\beta_1 - \hat{\beta}_1) + \varepsilon_i - \overline{\varepsilon}_n.$$

Alors

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = (\beta_1 - \hat{\beta}_1)^2 \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 + \sum_{i=1}^{n} (\varepsilon_i - \overline{\varepsilon}_n)^2 + 2(\beta_1 - \hat{\beta}_1) \sum_{i=1}^{n} (X_i - \overline{X}_n)(\varepsilon_i - \overline{\varepsilon}_n)$$
$$= (\beta_1 - \hat{\beta}_1)^2 \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 + \sum_{i=1}^{n} (\varepsilon_i - \overline{\varepsilon}_n)^2 - 2(\beta_1 - \hat{\beta}_1)^2 \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

En passant à l'espérance, on déduit le résultat. □

2.3 Modèle linéaire Gaussien

On suppose que les hypothèses $(\mathcal{H}_0)$ et $(\mathcal{H}_3)$ sont vérifiées.

2.3.1 Estimateur du maximum de vraisemblance

L'hypothèse $(\mathcal{H}_3)$ implique que $\varepsilon_1, \ldots, \varepsilon_n$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Par suite $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ suit une loi $\mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$. Ainsi, la vraisemblance de l'échantillon $(Y_1, \ldots, Y_n)$ s'écrit

$$\begin{aligned} L(Y_1, \ldots, Y_n, \beta_0, \beta_1) &= \prod_{i=1}^{n} \Big[\frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \Big) \Big] \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \Big). \end{aligned}$$

La log-vraisemblance est

$$\ln\Big(L(Y_1, \ldots, Y_n, \beta_0, \beta_1) \Big) = - \ln\Big((2\pi)^{\frac{n}{2}} \sigma^n \Big) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Maximiser la log-vraisemblance revient minimiser

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Les estimateurs du maximum de vraisemblance de β_0 et β_1 coïncident avec ceux des moindres carrés ordinaires.

2.3.2 Propriétés des estimateurs du maximum de vraisemblance

On note :

$$\hat{\sigma}^2_{\hat{\beta}_0} = \hat{\sigma}^2 \Big(\frac{1}{n} + \frac{\overline{X}_n^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \Big) \tag{2.3.1}$$

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}. \tag{2.3.2}$$

On note aussi :

- $\chi^2(n)$, la loi de Khi-deux à n degrés de liberté
- $T(n)$, la loi de Student à n degrés de liberté.

Proposition 6. *Sous les hypothèses $(\mathcal{H}_0)$ et $(\mathcal{H}_3)$, nous avons les résultats suivants :*

- *$\hat{\beta}_1$ et $\overline{Y}_n$ sont indépendantes.*
- *le vecteur $(\hat{\beta}_0, \hat{\beta}_1)'$ et la variable $\hat{\sigma}^2$ sont indépendants.*

Proposition 7. *Sous les hypothèses $(\mathcal{H}_0)$ et $(\mathcal{H}_3)$, nous avons les résultats suivants :*

- $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \hookrightarrow \chi^2(n-2)$.
- $\dfrac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \hookrightarrow T(n-2)$.
- $\dfrac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \hookrightarrow T(n-2)$.

2.4 Intervalles de confiance

Nous proposons des intervalles de confiance des paramètre β_0 et β_1.

Proposition 8. *Un intervalle de confiance de β_0 de niveau $1-\alpha$ est donné par :*

$$[\hat{\beta}_0 - t_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_0}, \hat{\beta}_0 + t_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_0}]$$

où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de $T(n-2)$, où $\hat{\sigma}_{\hat{\beta}_0}$ est défini par (2.3.1).

Démonstration. **La preuve en exercice.** □

Proposition 9. *Un intervalle de confiance de β_1 de niveau $1-\alpha$ est donné par :*

$$[\hat{\beta}_1 - t_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_1}]$$

où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de $T(n-2)$, où $\hat{\sigma}_{\hat{\beta}_1}$ est défini par (2.3.2).

Démonstration. **La preuve en exercice.** □

Proposition 10. *Un intervalle de confiance de σ^2 est donné par*

$$\left[\frac{(n-2)\hat{\sigma}^2}{b}, \frac{(n-2)\hat{\sigma}^2}{a}\right]$$

où a est le quantile d'ordre α_1 et b est le quantile d'ordre $1-\alpha_2$ de $\chi^2(n-2)$ avec $\alpha = \alpha_1 + \alpha_2$.

Démonstration. **La preuve en exercice.** □

Nous proposons ensuite un intervalle de confiance de la droite de régression $\beta_0 + \beta_1 x_*$.

Proposition 11. *Un intervalle de confiance pour $\beta_0 + \beta_1 x^*$ est donné par*

$$\left[\hat{y}_* \overset{+}{-} t_{1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \overline{X}_n)^2}{\sum_{i=1}^n (x^* - \overline{X}_n)^2}}\right]$$

où

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

Démonstration. **La preuve en exercice.** □

2.5 Tests de validité marginale

Nous considérons le test de l'hypothèse

$$H_0 : \beta_1 = 0 \quad \text{contre} \quad H_1 : \beta_1 \neq 0.$$

Si H_1 est rejetée, on dira que le coefficient β_1 n'est pas significatif. Dans le cas contraire, on dira que le coefficient est significatif et que la variable X influe sur la variable Y.
La région critique du test est donnée par

$$W = \left\{ \left| \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right| > C \right\}$$

où C est une constante à déterminer et où $\hat{\sigma}_{\hat{\beta}_0}$ est défini par (2.3.2).

Sous H_0, d'après la Proposition 7, $\dfrac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$ suit la loi de Student $T(n-2)$. Alors, nous rejetons l'hypothèse H_0 si $C = t_{1-\frac{\alpha}{2}}$ où C le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student $T(n-2)$.

2.6 Critère de comparaison des modèles

2.6.1 Equation de l'Analyse de la variance

Un modèle est bon si $\hat{Y}_i$ sont proches des vraies valeurs Y_i.

- $SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ (Somme des carrés totale)

- $SCE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ (Somme des carrés expliquée)
- $SCR = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$ (Somme des carrés résiduelle)

Equation de l'analyse de la variance : $SCT = SCE + SCR$.

2.6.2 Coefficients de détermination

Définition 5. Le coefficient de détermination R^2 est défini par :

$$R^2 = \frac{SCE}{SCT}$$

c'est à dire la part de la variabilité expliquée par le modèle sur la variabilité totale

Remarque 4.
- R^2 est le carrédu coefficient de corrélation.
- $0 \leq R^2 \leq 1$
- Si $R^2 = 1$, le modèle explique tout i.e. $Y_i = \beta_0 + \beta_1 X_i$.
- Si $R^2 = 0$ i.e $\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = 0$ et donc que $\hat{Y}_i = \bar{Y}$, le modèle de regression linéaire est inadapté (absence de liaison linéaire).

2.7 Prévision

La valeur pour laquelle nous effectuons la précision n'a pas servi dans le calcul des estimateurs. Soit X_{n+1} cette valeur. Nous voulons prédire Y_{n+1}. Le modèle indique que $Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1}$ avec $\mathbb{E}(\varepsilon_{n+1}) = 0$, $var(\varepsilon_{n+1}) = \sigma^2$ et $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \ldots, n$. Nous pouvons prédire Y_{n+1} gràce au modèle estimé :

$$\hat{Y}^p_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1}.$$

Deux types d'erreurs entachent notre prévision :

- l'une due à la non connaissance de ε_{n+1}
- l'autre due à l'estimation des paramètres.

Proposition 12. *(Variance de la prévision Y^p_{n+1})*

$$var(Y^p_{n+1}) = \sigma^2 \left(\frac{1}{n} + \frac{(X_{n+1} - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right).$$

$var(Y_{n+1}^p)$ nous donne une idée de la stabilité de l'estimation. En prévision, on s'interesse généralement à l'erreur que l'on commet entre la vraie valeur à prévoir Y_{n+1} et celle que l'on prévoit Y_{n+1}^p. L'erreur peut être simplement résumée par la différence entre les deux valeurs : erreur de prévision. Cette erreur de prévision permet de quantifier la capacité du modèle à prévoir.

Proposition 13. *(Erreur de prévision)*
L'erreur de prévision définie par $\varepsilon_{n+1}^p = Y_{n+1} - Y_{n+1}^p$ *satisfait les propriétés suivantes :*

$$\mathbb{E}(\varepsilon_{n+1}^p) = 0$$
$$var(\varepsilon_{n+1}^p) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_{n+1} - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right).$$

Remarque 5. La variance augmente lorque X_{n+1} s'éloigne du centre de gravité du nuage de points. Effectuer une prévision lorsque X_{n+1} est "loin" de $\overline{X}$ est donc périleux, la variance de l'erreur de prévision peut être alors très grande.

Proposition 14. *Un intervalle de confiance pour* Y_{n+1} *est donné par*

$$\left[Y_{n+1}^p \stackrel{+}{-} t_{1-\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_{n+1} - \overline{X}_n)^2}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}}\right].$$

Cette formule exprime que plus le point à prévoir est éloigné de $\overline{X}$, plus la variance de la prévision et donc de l'intervalle de confiance seront grandes.

2.8 Régression linéaire simple avec R

La masse monétaire et le revenu national brut (en milliards de francs CFA) de la Côte d'Ivoire sont reproduits dans le tableau ci-dessous (source : Banque mondiale).

Année	Masse Monétaire	Revenu national brut
2000	1646.26	6289.63
2001	1840.12	6386.73
2002	2401.83	6306.89
2003	1759.82	6245.24
2004	1932.57	6403.78
2005	2080.94	6495.91
2006	2294.76	6523.78
2007	2836.59	6625.53
2008	2997.35	6781.55
2009	3511.75	7025.49
2010	4152.21	7194.92
2011	4595.55	6856.40

Établir une relation linéaire dans laquelle la masse monétaire explique explique le revenu national brut.

```
> donnees<-read.table("masse.txt",header=TRUE)
> donnees
   MasseMonetaire     RNB
1         1646.26 6289.63
2         1840.12 6386.73
3         2401.83 6306.89
4         1759.82 6245.24
5         1932.57 6403.78
6         2080.94 6495.91
7         2294.76 6523.78
8         2836.59 6625.53
9         2997.35 6781.55
10        3511.75 7025.49
11        4152.21 7194.92
12        4595.55 6856.40



> modele1<-lm(RNB~MasseMonetaire,data=donnees)
> summary(modele1)

Call:
lm(formula = RNB ~ MasseMonetaire, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-279.31  -36.09   21.12   74.08  194.44

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.844e+03  1.299e+02  45.000 7.07e-13 ***
MasseMonetaire 2.811e-01  4.591e-02   6.123 0.000112 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 148.2 on 10 degrees of freedom
Multiple R-squared:  0.7894,    Adjusted R-squared:  0.7684
F-statistic: 37.49 on 1 and 10 DF,  p-value: 0.0001122
```

```
> plot(donnees,xlab="Masse monétaire")
```

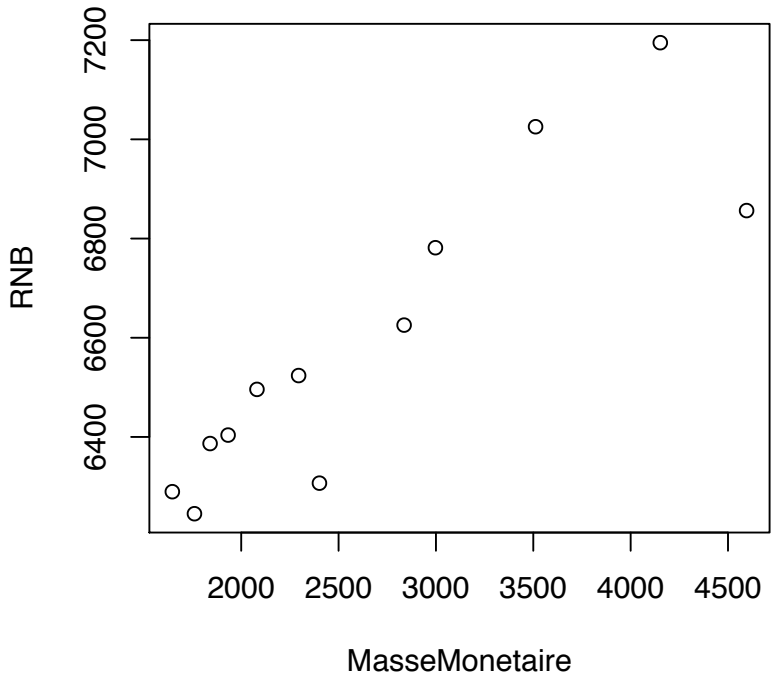


FIGURE 2.1.Nuage de points

Chapitre 3

Régression linéaire multiple

3.1 Modélisation

Le modèle de régression linéaire multiple est défini par l'équation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \ldots, n$$

où

- Y_i est la réponse mesurée pour l'individu i
- X_{ij} est la valeur de X_j pour l'individu i ;
- $\beta_0, \ldots, \beta_p$ sont des paramètres inconnus
- ε_i appelée aléa est une variable aléatoire.

En posant

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1j} & \cdots & X_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & X_{i1} & \cdots & X_{ij} & \cdots & X_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nj} & \cdots & X_{np} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

on obtient la forme matricielle suivante :

$$Y = X\beta + \varepsilon. \tag{3.1.1}$$

Soient les hypothèses suivantes :

($\mathcal{H}_0$) La matrice X n'est pas aléatoire.

($\mathcal{H}_1$) $rang(X) = p + 1$.

($\mathcal{H}_2$) $\mathbb{E}(\varepsilon) = 0$ et $Var(\varepsilon) = \sigma^2 \mathbb{I}_n$ avec $\sigma^2 > 0$.

($\mathcal{H}_3$) $\varepsilon \hookrightarrow \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$ avec $\sigma^2 > 0$.

Remarque 6. Les paramètres inconnus du modèle sont : β et σ^2.

Remarque 7. ($\mathcal{H}_0$) implique que l'on choisit les valeurs des variables explicatives puis on observe Y.
($\mathcal{H}_1$) implique que les colonnes de X forment des vecteurs linéairement indépendants de $\mathbb{R}^n$. Ainsi, nous avons

$$\forall C \in \mathbb{R}^{p+1}, \quad XC = 0 \Rightarrow C = 0;$$

il existe donc un unique vecteur θ associé au modèle (3.1.1) ; de plus, on a $n \geq p + 1$; si l'on avait $rang(X) < p + 1$, cela signifierait qu'il existe au moins une variable explicative qui peut s'écrire comme une combinaison linéaire d'une ou des autres variables explicatives : cette variable explicative serait donc superflue, elle n'apporterait rien à l'explication de Y déjà fournie par les autres variables explicatives.
($\mathcal{H}_2$) implique que les composantes de ε sont centrées, de même variance (homoscédasticité) et non correlées entre elles.
($\mathcal{H}_3$) implique que les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont indépendantes identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$.

3.2 Estimateurs des moindres carrés

3.2.1 Définition

Proposition 15. *Dérivée matricielle*
Pour tout $v, a \in \mathbb{R}^k$, pour toute matrice carrée d'ordre k, nous avons

- $\dfrac{\partial v' a}{\partial v} = \dfrac{\partial a' v}{\partial v} = a$
- $\dfrac{\partial v' M v}{\partial v} = (M + M')v$

Nous supposons vérifier les hypothèses ($\mathcal{H}_0$) et ($\mathcal{H}_1$).

Définition 6. On appelle estimateur des moindres carrés ordinaires $\hat{\beta}$, la valeur de β qui minimise la fonction suivante

$$S(\beta) = (Y - X\beta)'(Y - X\beta)$$

Comme $\varepsilon = Y - X\beta$, on a

$$S(\beta) = \varepsilon^{'}\varepsilon = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}) \right)^2 .$$

Théorème 2. *L'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β est défini par*

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}.$$

Démonstration. Nous avons

$$\begin{aligned} S(\beta) &= (Y - X\beta)^{'}(Y - X\beta) \\ &= Y^{'}Y - Y^{'}X\beta - \beta^{'}X^{'}Y + \beta^{'}X^{'}X\beta \end{aligned}$$

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^{'}Y + 2X^{'}X\beta = 0 \Rightarrow \beta = (X^{'}X)^{-1}X^{'}Y$$

Comme $\dfrac{\partial^2 S(\beta)}{\partial \beta^2} = 2X^{'}X$ est une matrice définie positive, on obtient le résultat. □

3.2.2 Propriétés des estimateurs

On suppose vérifier les hypothèses $\mathcal{H}_0$, $\mathcal{H}_1$ et $\mathcal{H}_2$.

Proposition 16. *$\hat{\beta}$ est un estimateur sans biais de β i.e.*

$$\mathbb{E}(\hat{\beta}) = \beta.$$

Proposition 17. *La matrice de variance-covariance de $\hat{\beta}$*

$$Var(\hat{\beta}) = \sigma^2 (X^{'}X)^{-1}.$$

Théorème 3. *(de Gauss-Markov)*

Parmi les estimateurs sans biais de β de la forme BY, où $B \in \mathcal{M}_{p+1,n}(\mathbb{R})$, $\hat{\beta}$ est optimal, c'est à dire de variance minimale.

Démonstration. Soit BY un autre estimateur linéaire sans biais de β. Puisque $\mathbb{E}(BY) = BX\beta$, on a $BX = \mathbb{I}_{p+1}$. Par suite, $CX = 0$ avec $C = B - (X^{'}X)^{-1}X^{'}$. Ainsi, nous avons

$$\begin{aligned} Var(BY) &= BVar(Y)B^{'} \\ &= [C + (X^{'}X)^{-1}X^{'}]\sigma^2 \mathbb{I}_n [C + (X^{'}X)^{-1}X^{'}]^{'} \\ &= \sigma^2 CC^{'} + Var(\hat{\beta}). \end{aligned}$$

Par suite, $Var(BY) - Var(\hat{\beta})$ est une matrice symétrique positive pour tout $B \in \mathcal{M}_{p+1,n}(\mathbb{R})$ □

3.2.3 Valeurs ajustées, résidus

Définition 7. Le vecteur $\hat{Y} = X\hat{\beta} = \begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix}$ est le vecteur des valeurs ajustées,

où

$$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} X_{ij}\hat{\beta}_j$$

Définition 8. Le vecteur $\hat{\varepsilon} = Y - \hat{Y} = \begin{pmatrix} \hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}$ est appelé vecteur des résidus

estimés.

Posons

$$H = X(X^{'}X)^{-1}X^{'}.$$

La matrice H est appelée la "matrice chapeau" ou "hat matrix".
Nous pouvons écrire alors

$$\hat{Y} = HY \qquad \hat{\varepsilon} = (\mathbb{I} - H)Y.$$

Posons

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Proposition 18. *Sous les hypothèses $\mathcal{H}_0$, $\mathcal{H}_1$ et $\mathcal{H}_2$, la statistique $\hat{\sigma}^2$ est un estimateur sans biais de σ^2.*

Démonstration. Nous avons

$$\mathbb{E}(\hat{\varepsilon}^{'}\hat{\varepsilon}) = \mathbb{E}(tr(\hat{\varepsilon}^{'}\hat{\varepsilon})) = \mathbb{E}(tr(\hat{\varepsilon}\hat{\varepsilon}^{'})) = tr(Var(\hat{\varepsilon})) = tr(\sigma^2(\mathbb{I}-H)) = \sigma^2(n-p-1).$$

□

Nous obtenons ainsi un estimateur de $Var(\hat{\beta}) = \sigma^2(X^{'}X)^{-1}$ en remplaçant σ^2 par son estimateur $\hat{\sigma}^2$:

$$\hat{\sigma}^2_{\hat{\beta}} = \hat{\sigma}^2(X^{'}X)^{-1}.$$

Nous avons donc un estimateur de l'écart-type de l'estimateur $\hat{\beta}_j$ du coefficient β_j :

$$\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{[(X'X)^{-1}]_{j+1,j+1}} \quad j = 0, \ldots, p. \tag{3.2.1}$$

C'est le $(j+1)$-ième coefficient diagonal de la matrice $\hat{\sigma}^2(X'X)^{-1}$; $\hat{\sigma}_{\hat{\beta}_j}$ est un indicateur du caractère plus ou moins stable de l'estimation de β_j.
La source principale d'instabilité dans l'estimation de β est la multicolinéarité (les variables explicatives sont très corrélées entre elles). Comme $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$, si les variables explicatives sont très corrélées entre elles, $X'X$ aura un déterminant proche de 0 et son inverse aura des termes élevés. Les paramètres du modèle seront estimés avec imprécision et les prédictions pourront être entachées d'erreurs considérables même si R^2 a une valeur élevée.

3.3 Modèle linéaire gaussien

Dans cette section, nous supposons vérifier les hypothèses $(\mathcal{H}_0)$, $(\mathcal{H}_1)$ et $(\mathcal{H}_3)$.

3.3.1 Méthode du maximum de vraisemblance

La vraisemblance de l'échantillon est défini par

$$L(Y,\beta,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2\right].$$

Ainsi, nous avons

$$\ln(L(Y,\beta,\sigma^2)) = -\ln((2\pi)^{\frac{n}{2}}\sigma^n) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2.$$

De ce fait, maximer la vraisemblance revient à utiliser la méthode des moindres carrés ordinaires. Ainsi, l'estimateur du maximum de vraisemblance et l'estimateur des moindres carrés ordinaires coïncident.

3.3.2 Intervalles de confiance

Proposition 19. *Nous avons*

- $\hat{\beta}$ *est un vecteur gaussien de moyenne* β *et de variance* $\sigma^2(X^TX)^{-1}$
- $\dfrac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \hookrightarrow \chi^2(n-p-1)$
- $\hat{\beta}$ *et* $\hat{\sigma}^2$ *sont indépendantes*

Proposition 20. *Pour* $j = 0, 1, \ldots, p$, *la variable*

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \hookrightarrow T(n-p-1).$$

Proposition 21. *Un intervalle de confiance de niveau* $1-\alpha$ *pour* β_j, $j = 1, \ldots, p$ *est donné par*

$$\left[\hat{\beta}_j - t^* \hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t^* \hat{\sigma}_{\hat{\beta}_j}\right]$$

où t^* *est le quantile d'ordre* $1-\alpha/2$ *de* $T(n-p-1)$.

Proposition 22. *Un intervalle de confiance de niveau* $1-\alpha$ *pour* σ^2 *est donné par*

$$\left[\frac{(n-p-1)\hat{\sigma}^2}{c_2}, \frac{(n-p-1)\hat{\sigma}^2}{c_1}\right]$$

avec $\mathbb{P}(c_1 \leq Z \leq c_2) = 1-\alpha$ *où* $Z \hookrightarrow \chi^2(n-p-1)$.

3.3.3 Test de validité marginale de Student

Nous considérons le test de l'hypothèse $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$. C'est un test bilateral. La statistique de test est $T = \dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$. Sous H_0,

$$T = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \hookrightarrow T(n-p-1).$$

Nous rejetons H_0 si $|T| > t^*$ où t^* est le quantile d'ordre $1-\alpha/2$ de $T(n-p-1)$. Nous concluons alors que le coefficient X_j est significatif.

3.4 Critère de comparaison de modèle

3.4.1 Equation d'analyse de la variance

Soient

dispersion totale : $SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

dispersion expliquée par le modèle : $SCE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$

dispersion résiduelle : $SCR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$.

Alors, l'équation d'analyse de variance est

$$SCT = SCE + SCR.$$

3.4.2 Coefficient de détermination

Le pourcentage de variabilité dû au modèle se mesure par le coefficient de détermination :

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

- Si $R^2 = 1 \Leftrightarrow Y = X\beta$.
- Si $R^2 \simeq 0 \Leftrightarrow$ résidus élevés$\Leftrightarrow$modèle de régression linéaire inadapté.

Remarque 8. En général, il ne faut pas utiliser le R^2 comme critère de choix de modèle car ce critère va toujours augmenter avec le nombre de variables explicatives. Il peut cependant servir à comparer des modèles ayant le même nombre de variables explicatives.

3.4.3 Coefficient de détermination ajusté

Le R^2 ajusté est défini par

$$R^2_{ad} = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)} = \frac{(n-1)R^2 - p}{n-p-1} = 1 - \frac{n-1}{n-p-1}(1-R^2).$$

3.4.4 Test de validité globale de Fisher

Nous considérons le test de l'hypothèse

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

contre

$$H_1 : \exists j \in \{1, \ldots, p\}, \quad \beta_j \neq 0.$$

Nous testons l'hypothèse que tous les coefficients sont nuls excepté la constante. Sous H_0

$$F = \frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}} = \frac{R^2}{1-R^2}\frac{n-p-1}{p} \hookrightarrow F(p, n-p-1)$$

la loi de Fisher à p et $n-p-1$ dégrés de liberté. On rejette H_0 si $F > F_{1-\alpha}$ où $F_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher et on conclut qu'il existe au moins un paramètre non nul dans le modèle.

3.5 Prévision

Soit une nouvelle valeur $X'_{n+1} = (1, X_{n+1,1}, \ldots, X_{n+1,p})$ et nous voulons prédire Y_{n+1}. Or

$$Y_{n+1} = X'_{n+1}\beta + \varepsilon_{n+1}$$

avec $\mathbb{E}(\varepsilon_{n+1}) = 0$, $var(\varepsilon_{n+1}) = \sigma^2$ et $cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \ldots, n$. La prévision de Y_{n+1} est

$$Y_{n+1}^p = X_{n+1}^{'} \hat{\beta}.$$

Deux types d'erreurs vont entacher la prévision :

- la première due à l'incertitude sur ε_{n+1}
- l'autre due à l'incertitude due à l'estimation.

L'espérance de l'erreur de prévision est $\mathbb{E}(Y_{n+1} - Y_{n+1}^p) = 0$. La variance de l'erreur de prévision est

$$var(Y_{n+1} - Y_{n+1}^p) = \mathbb{E}(Y_{n+1} - Y_{n+1}^p)^2 = \sigma^2(1 + X_{n+1}^{'}(X^{'}X)^{-1}X_{n+1}).$$

Nous retrouvons bien l'incertitude due aux erreurs σ^2 sur laquelle vient s'ajouter l'incertitude de l'estimation.

Théorème 4. *Un intervalle de confiance de niveau* $1-\alpha$ *pour* Y_{n+1} *est donné par*

$$\left[Y_{n+1}^p \pm t^* \hat{\sigma} \sqrt{1 + X_{n+1}^{'}(X^{'}X)^{-1}X_{n+1}} \right]$$

où t^* *est le quantile d'ordre* $1 - \alpha/2$ *de* $T(n-p-1)$.

Chapitre 4

Analyse des résidus

L'analyse des résidus permet d'étudier empiriquement le bien-fondé des hypothèses de la régression linéaire : linéarité, homoscédasticité, normalité des erreurs, etc. Elle permet aussi de repérer des observations éventuellement aberrantes ou des observations qui jouent un rôle important dans la régression. La démarche est généralement graphique.

4.1 Définitions

Le modèle de régression linéaire est défini par

$$Y = X\beta + \varepsilon$$

où

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1j} & \cdots & X_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & X_{i1} & \cdots & X_{ij} & \cdots & X_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nj} & \cdots & X_{np} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Par hypothèse, nous avons ε est un vecteur gaussien d'espérance $\mathbb{E}(\varepsilon) = 0$ et de matrice de variance-covariance $V(\varepsilon) = \sigma^2 I_n$. ε est le vecteur des résidus théoriques appelés aléas ou erreurs dans les chapitres précédents. Les résidus théoriques sont estimés par

$$\hat{\varepsilon} = \begin{pmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix} = Y - \hat{Y}$$

où pour chaque individu i, le résidu est défini par

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i.$$

Remarque 9. $\mathbb{E}(\hat{\varepsilon}) = 0$ et $V(\hat{\varepsilon}) = \sigma^2(I_n - H)$ où

$$H = X(X^{'}X)^{-1}X^{'} = (h_{ij})_{1\leq i,j\leq n}$$

Nous déduisons

$$var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}) \quad 1 \leq i \leq n.$$

Ainsi, pour $i \neq j$, $\hat{\varepsilon}_i$ et $\hat{\varepsilon}_j$ n'ont pas la même variance. Nous introduisons les résidus standardisés

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.$$

afin d'éliminer la non homogénéité des variances des $\hat{\varepsilon}_i$. Ces résidus ne sont pas indépendants.
Les résidus studentisés par validation croisée sont définis par :

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

où $\hat{\sigma}_{(i)}$ est l'estimateur de σ dans le modèle linéaire privé de l'observation i.

Théorème 5. *Sous les hypothèse $(\mathcal{H}_1)$ et $(\mathcal{H}_3)$, et si la suppression de la ligne i ne modifie pas le rang de la matrice alors les résidus studentisés t_i^* suivent la loi de Student à $(n - p - 2)$ dégrés de liberté.*

4.2 Analyse de la normalité

L'hypothèse de normalité sera examinée à l'aide d'un histogramme ou d'un graphique comparant les quantiles des résidus à ces mêmes quantiles sous l'hypothèse de normalité appelé QQ-plot. Si cette hypothèse est respectée, le graphique QQ-plot sera proche de la première bissectrice.

4.3 Indépendance, homoscédasticité et linéarité

Il est recommandé de tracer les résidus studentisés t_i^* en fonction des valeurs ajustées $\hat{Y}_i$ i.e. de tracer le nuage de points $(\hat{Y}_i, t_i^*)$. Si les points se retrouvent à l'intérieur d'un rectangle centré sur l'ordonnée nulle alors les hypothèses d'indépendance et de linéarité sont vérifiées.
Si une structure apparaît (tendance, cone, vagues), l'hypothèse d'homoscédasticité risque fort de ne pas être vérifiée.

4.4 Multicolinéarité

4.5 Individus extrêmes

La régression est sensible aux individus extrêmes. Ils peuvent considérablement influencer la valeur des paramètres de la régression.

4.5.1 Jacknife

On estime les paramètres de la régression en retirant l'individu i pour voir si celui-ci influence beaucoup ou non la régression.

4.5.2 Leverage ou effet levier

L'effet levier associé à l'individu i est défini par

$$h_{ii} = \frac{1}{n} + \frac{X_i - \bar{X}_n}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}.$$

h_{ii} permet d'estimer dans quelle mesure un individu contribue à lui seul à la régression. Il faudra se méfier des individus pour lesquels $h_{ii} > \frac{2p}{n}$.

4.5.3 Distance de Cook

Définition 9. La distance de Cook est donnée par

$$D_{Cook_i} = \frac{h_{ii}}{(p+1)(1-h_{ii})^2}\frac{\hat{\varepsilon}_i^2}{\hat{\sigma}^2}$$

Une observation influente est une observation qui, enlevée, conduit à une grande variation dans l'estimation des coefficients i.e. à une distance de Cook élevée. Pour juger si la distance D_{Cook_i} est élevée, Cook (1977) propose le seuil $F_{p,n-p-1}(0.1)$ comme souhaitable et le seuil $F_{p,n-p-1}(0.5)$ comme préoccupant.

Chapitre 5

Etude de cas

Dans cet exemple, la variable à expliquer est la consommation des véhicules. Nous avons 4 variables explicatives : le prix, la cylindrée, la puissance et le poids. Les objectifs sont les suivants :

- Etude de la relation linéaire entre la consommation de véhicules et ses caractéristiques que sont le prix, la cylindrée, la puissance et le poids.
- Diagnostic de la régression avec les graphiques des résidus.
- Prévision

Nous utilisons la commande **read.table()** pour importer les données dans le logiciel à partir du fichier texte (.txt).

```
> #Importation des données
> #
> donnees<-read.table("conso_veh.txt",header=TRUE)
> #
> #pour afficher les données
> donnees
                    modele   prix cylindree puissance poids consom
1           Daihatsu.Cuore  11600       846        32   650    5.7
2     Suzuki.Swift.1.0.GLS  12490       993        39   790    5.8
3       Fiat.Panda.Mambo.L  10450       899        29   730    6.1
4           VW.Polo.1.4.60  17140      1390        44   955    6.5
5      Opel.Corsa.1.2i.Eco  14825      1195        33   895    6.8
6         Subaru.Vivio.4WD  13730       658        32   740    6.8
7           Toyota.Corolla  19490      1331        55  1010    7.1
8           Ferrari.456.GT 285000      5474       325  1690   21.3
9           Mercedes.S.600 183900      5987       300  2250   18.7
10      Maserati.Ghibli.GT  92500      2789       209  1485   14.5
11     Opel.Astra.1.6i.16V  25000      1597        74  1080    7.4
12      Peugeot.306.XS.108  22350      1761        74  1100    9.0
```

```
13   Renault.Safrane.2.2.V  36600      2165       101  1500   11.7
14      Seat.Ibiza.2.0.GTI  22500      1983        85  1075    9.5
15         VW.Golt.2.0.GTI  31580      1984        85  1155    9.5
16      Citroen.ZX.Volcane  28750      1998        89  1140    8.8
17 Fiat.Tempra.1.6.Liberty  22600      1580        65  1080    9.3
18     Fort.Escort.1.4i.PT  20300      1390        54  1110    8.6
19   Honda.Civic.Joker.1.4  19900      1396        66  1140    7.7
20           Volvo.850.2.5  39800      2435       106  1370   10.8
21   Ford.Fiesta.1.2.Zetec  19740      1242        55   940    6.6
22     Hyundai.Sonata.3000  38990      2972       107  1400   11.7
23          Lancia.K.3.0.LS 50800      2958       150  1550   11.9
24       Mazda.Hachtback.V  36200      2497       122  1330   10.8
25       Mitsubishi.Galant  31990      1998        66  1300    7.6
26       Opel.Omega.2.5iV6  47700      2496       125  1670   11.3
27         Peugeot.806.2.0  36950      1998        89  1560   10.8
28      Nissan.Primera.2.0  26950      1997        92  1240    9.2
29       Seat.Alhambra.2.0  36400      1984        85  1635   11.6
30     Toyota.Previa.salon  50900      2438        97  1800   12.8
31     Volvo.960.Kombi.aut  49300      2473       125  1570   12.7
>
```

La commande **lm()** permet de faire la régression.

```
> #regression linéaire
> regression<-lm(consom~prix+cylindree+puissance+poids,data=donnees)
> #
> #résultats de la regression
> resultats<-summary(regression)
> resultats

Call:
lm(formula = consom ~ prix + cylindree + puissance + poids, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5677 -0.6704  0.1183  0.5283  1.4361

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.456e+00  6.268e-01   3.919 0.000578 ***
prix         2.042e-05  8.731e-06   2.339 0.027297 *
cylindree   -5.006e-04  5.748e-04  -0.871 0.391797
puissance    2.499e-02  9.992e-03   2.501 0.018993 *
```

```
poids        4.161e-03  8.788e-04   4.734 6.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8172 on 26 degrees of freedom
Multiple R-squared:  0.9546,    Adjusted R-squared:  0.9476
F-statistic: 136.5 on 4 and 26 DF,  p-value: < 2.2e-16

>
```

La commande **confint** fournit les intervalles de confiance de chaque paramètre.

```
> #intervalle de confiance
> IC<-confint(regression,level=0.95)
> IC
                    2.5 %       97.5 %
(Intercept)  1.167851e+00 3.744737e+00
prix         2.474392e-06 3.836669e-05
cylindree   -1.682157e-03 6.809703e-04
puissance    4.455929e-03 4.553302e-02
poids        2.354210e-03 5.966955e-03
```

La fonction **predict()** permet de donner les prévisions mais aussi les intervalles de confiance tant pour le modèle que pour pour les prévisions.

```
> #Prévision pour Z1=(14000,800,35,700) Z2=(170000,6000,270,1870)
 IC pour Y et E(Y)
> prix=c(14000,170000)
> cylindree=c(800,670)
> puissance=c(35,270)
> poids=c(700,1870)
> nouv.donnees=data.frame(prix,cylindree,puissance,poids)
> nouv.donnees
    prix cylindree puissance poids
1  14000       800        35   700
2 170000       670       270  1870
> ICdte=predict.lm(regression,nouv.donnees,interval="confidence")
> ICdte
        fit       lwr       upr
1  6.128921  5.528698  6.729145
2 20.121187 15.197486 25.044887
> ICpred=predict.lm(regression,nouv.donnees,interval="prediction")
> ICpred
        fit       lwr       upr
```

```
1  6.128921  4.345052  7.912791
2 20.121187 14.918808 25.323565
>

> ######GRAPHIQUES des résidus
> par(mfrow=c(2,3))  #subdviser la page en 6 parties
> plot(regression,which=1,sub="",main="")
> plot(regression,which=2,sub="",main="")
> plot(regression,which=3,sub="",main="")
> plot(regression,which=4,sub="",main="")
> plot(regression,which=5,sub="",main="")
> plot(regression,which=6,sub="",main="")
>
```

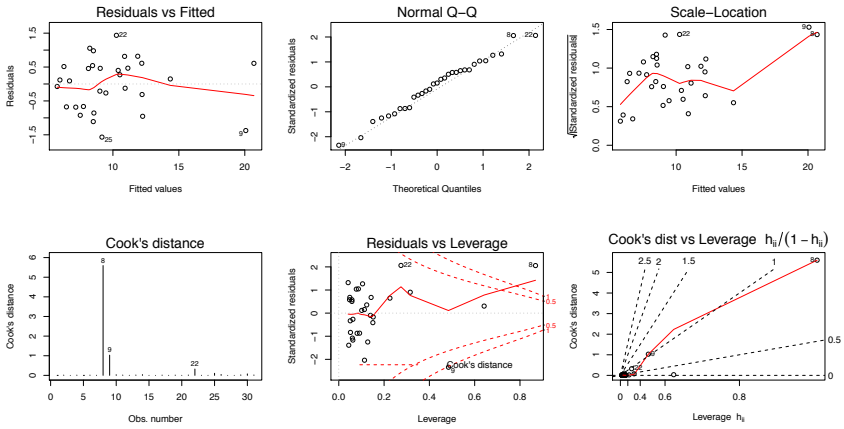


FIGURE 5.1.

Bibliographie

[1] J. M. Azaïs, J. M. Bardet, Le modèle linéaire par l'exemple, Dunod, Paris, 2005.

[2] P. Cornillon, E. Matzner-Lober, Régression avec R, Springer-Verlag France, 2011.