

Correction Feuille de Travaux Dirigés 4

Modèles de régression linéaire

Exercice 1

1 $\hat{\beta}$ est l'estimateur MCO de β . En effet, nous avons $Y = X\beta + E$ avec $Y = (Y_1, \dots, Y_n)$, $X = (x_1, \dots, x_n)$ et $E = (E_1, \dots, E_n)$.

Ainsi, $\hat{\beta} = (X'X)^{-1} X'Y = \left(\sum_{i=1}^n x_i^2 \right)^{-1} \sum_{i=1}^n x_i Y_i$.

$\tilde{b}$, la réalisation de $\tilde{\beta}$, est la pente de la droite qui passe par l'origine et par

le point moyen $(\bar{x}, \bar{y})$. En effet, $\bar{Y} = \tilde{\beta} \bar{x} \implies \tilde{\beta} = \frac{\bar{Y}}{\bar{x}} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}$.

2 Nous avons $\mathbb{E}(\hat{\beta}) = \frac{\sum_{i=1}^n x_i \mathbb{E}(Y_i)}{\sum_{i=1}^n x_i^2} = \beta$ et $\mathbb{E}(\tilde{\beta}) = \frac{\sum_{i=1}^n \mathbb{E}(Y_i)}{\sum_{i=1}^n x_i} = \beta$.

Par conséquent, $\hat{\beta}$ et $\tilde{\beta}$ sont des estimateurs non biaisés de β .

3 Nous avons $\mathbb{V}(\hat{\beta}) = \frac{1}{\left(\sum_{i=1}^n x_i^2 \right)^2} \sum_{i=1}^n x_i^2 \mathbb{V}(Y_i) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$

et $\mathbb{V}(\tilde{\beta}) = \frac{1}{\left(\sum_{i=1}^n x_i \right)^2} \sum_{i=1}^n \mathbb{V}(Y_i) = \frac{n\sigma^2}{\left(\sum_{i=1}^n x_i \right)^2}$.

Par ailleurs, $\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})^2 \geq 0$.

De ce fait, $\mathbb{V}(\hat{\beta}) \leq \mathbb{V}(\tilde{\beta})$ et $\mathbb{V}(\hat{\beta}) = \mathbb{V}(\tilde{\beta}) \iff x_1 = \dots = x_n$.

Dans la mesure où $\tilde{\beta}$ est un estimateur non biaisé de β linéaire en Y , le résultat précédent était prévisible d'après le théorème de Gauss-Markov.

Exercice 2

1 Nous avons $\tilde{\theta}_1 = (W'W)^{-1} W'Y$.

Par conséquent,

$$\begin{aligned} \mathbb{E}(\tilde{\theta}_1) &= (W'W)^{-1} W' \mathbb{E}(Y) \\ &= (W'W)^{-1} W' (W\theta_1 + Z\theta_2) \\ &= \theta_1 + (W'W)^{-1} W' Z \theta_2 \end{aligned}$$

De ce fait, $\tilde{\theta}_1$ est un estimateur biaisé de β .

2 Le modèle estimé est le modèle sous contrainte suivant :

$$\begin{cases} Y = W\theta_1 + Z\theta_2 + E \\ \theta_2 = 0 \end{cases} \iff \begin{cases} Y = W\theta_1 + Z\theta_2 + E \\ R \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = 0_{p_2} \end{cases}$$

avec $R = (0_{p_2 \times p_1} \mid I_{p_2})$.

a) Les conditions de Kuhn et Tucker permettent d'obtenir le minimum. Nous savons que ces conditions s'obtiennent en égalant à 0 les dérivées par rapport à chacune des composantes de β et λ du Lagrangien $L = (Y - X\beta)'(Y - X\beta) + \lambda'(R\beta - r)$.

Nous obtenons alors :

$$-2X'Y + 2X'X\hat{\beta}^c + R'\hat{\lambda} = 0 \quad (1)$$

$$R\hat{\beta}^c - r = 0 \quad (2)$$

En prémultipliant (1) par $R(X'X)^{-1}$, nous obtenons :

$$-2R(X'X)^{-1} X'Y + 2R\hat{\beta}^c + R(X'X)^{-1} R'\hat{\lambda} = 0.$$

Soit, en utilisant (2),

$$\widehat{\lambda} = 2 \left(R (X'X)^{-1} R' \right)^{-1} \left(R' (X'X)^{-1} X'Y - r \right).$$

En substituant dans (1), nous obtenons :

$$-X'Y + X'X\widehat{\beta}^c + R' \left[R (X'X)^{-1} R' \right]^{-1} \left[R (X'X)^{-1} X'Y - r \right] = 0.$$

De ce fait,

$$\widehat{\beta}^c = \widehat{\beta} + (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} \left[r - R\widehat{\beta} \right].$$

b)

$$\begin{aligned} \mathbb{E}(\widehat{\beta}^c) &= \mathbb{E}(\widehat{\beta}) - (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} \left[R\mathbb{E}(\widehat{\beta}) - r \right] \\ &= \beta - (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} [R\beta - r] \end{aligned}$$

c)

$$\begin{aligned} \mathbb{V}(\widehat{\beta}^c) &= \mathbb{V}(\widehat{\beta}^c - \mathbb{E}(\widehat{\beta}^c)) \\ &= \mathbb{V}\left(\widehat{\beta}^c - \beta + (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} [R\beta - r]\right) \\ &= \mathbb{V}\left(\left(I_p - (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} R\right) (\widehat{\beta} - \beta)\right) \end{aligned}$$

Posons $M = (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} R$.

Par conséquent,

$$\begin{aligned} \mathbb{V}(\widehat{\beta}^c) &= \mathbb{V}\left((I_p - M) (\widehat{\beta} - \beta)\right) \\ &= (I_p - M) \mathbb{V}(\widehat{\beta}) (I_p - M)' \\ &= \sigma^2 (I_p - M) (X'X)^{-1} (I_p - M)' \end{aligned}$$

Par ailleurs, $M^2 = M$ et $(X'X)^{-1} M' = M (X'X)^{-1}$.

D'où

$$\begin{aligned}
\mathbb{V}(\widehat{\beta}^c) &= \sigma^2 (I_p - M) (I_p - M) (X'X)^{-1} \\
&= \sigma^2 (I_p - M) (X'X)^{-1} \\
&= \sigma^2 \left((X'X)^{-1} - (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} R (X'X)^{-1} \right)
\end{aligned}$$

d) Nous avons :

$$\mathbb{V}(\widehat{\beta}) - \mathbb{V}(\widehat{\beta}^c) = \sigma^2 (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} R (X'X)^{-1}.$$

$(X'X)$ est une matrice définie positive (X est de plein rang).

Donc $R (X'X)^{-1} R'$ est une matrice définie positive.

Par conséquent, $(X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} R (X'X)^{-1}$ est une matrice semi-définie positive.

De ce fait, $\mathbb{V}(\widehat{\beta}) - \mathbb{V}(\widehat{\beta}^c)$ est une matrice semi-définie positive et, pour tout $x \in \mathbb{R}^p$, $x' \mathbb{V}(\widehat{\beta}) x \geq x' \mathbb{V}(\widehat{\beta}^c) x$.

4 Dans notre cas, $\beta = (\theta_1, \theta_2)$, $\widehat{\beta}^c = (\widetilde{\theta}_1, 0_{p_2})$ et $\widehat{\beta} = (\widehat{\theta}_1, \widehat{\theta}_2)$.

$$\begin{aligned}
\mathbb{V}(\widehat{\theta}_1) - \mathbb{V}(\widetilde{\theta}_1) &= \mathbb{V} \left((I_{p_1} | 0_{p_2}) \widehat{\beta} \right) - \mathbb{V} \left((I_{p_1} | 0_{p_2}) \widehat{\beta}^c \right) \\
&= (I_{p_1} | 0_{p_2}) \left[\mathbb{V}(\widehat{\beta}) - \mathbb{V}(\widehat{\beta}^c) \right] (I_{p_1} | 0_{p_2})'
\end{aligned}$$

Par conséquent, $\mathbb{V}(\widehat{\theta}_1) - \mathbb{V}(\widetilde{\theta}_1)$ est une matrice semi-définie positive. Ainsi, l'estimateur MCO du faux modèle a une variance inférieure à celle de l'estimateur MCO du vrai modèle contenant des variables explicatives supplémentaires, cependant il est biaisé.

Exercice 3

1 Le modèle peut s'écrire sous la forme

$$Y = X\beta + E$$

$$\text{avec } Y = (\log VA_1, \dots, \log VA_n), X = \begin{bmatrix} 1 & \log L_1 & \log K_1 \\ 1 & \log L_2 & \log K_2 \\ \vdots & \vdots & \vdots \\ 1 & \log L_n & \log K_n \end{bmatrix}, \beta = (\alpha, \lambda, \gamma)$$

et $E = (E_1, \dots, E_n)$.

L'estimateur MCO de β est $\hat{\beta} = (\hat{\alpha}, \hat{\lambda}, \hat{\gamma}) = (X'X)^{-1}X'Y$

Nous avons $\mathbb{V}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$.

Un estimateur non biaisé de σ^2 est

$$\hat{\sigma}^2 = \left(\frac{1}{n-3} \right) \sum_{i=1}^n \left(Y_i - \hat{\alpha} - \hat{\lambda} \log L_i - \hat{\gamma} \log K_i \right)^2.$$

Un estimateur non biaisé de $\mathbb{V}(\hat{\beta})$ est $\widehat{\mathbb{V}(\hat{\beta})} = \hat{\sigma}^2 (X'X)^{-1}$.

2 Notons s^2 la réalisation de $\hat{\sigma}^2$.

Nous avons $\text{SCR} = \sum_{i=1}^n (\log VA_i - a - l \log L_i - g \log K_i)^2$ où a , l et g sont respectivement les réalisations de $\hat{\alpha}$, $\hat{\lambda}$ et $\hat{\gamma}$.

Par conséquent, $s^2 = \frac{\text{SCR}}{n-3} = \frac{148.27}{1655} \simeq 8.959 \times 10^{-2}$.

Notons v la réalisation de $\widehat{\mathbb{V}(\hat{\beta})}$.

$$\text{Nous avons } v = (10)^{-4} \begin{bmatrix} 25.8 & 1.1 & -3 \\ 1.1 & 1.4 & -0.9 \\ -3 & -0.9 & 0.8 \end{bmatrix}.$$

La formule d'analyse de la variance d'une régression est la suivante :

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

avec $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ et $\hat{y}_i = a + l \log L_i - g \log K_i$.

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \text{SCR} : \text{Somme des Carrés Résiduelle};$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \text{SCT} : \text{Somme des Carrés Totale};$$

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \text{SCEM} : \text{Somme de Carrés Expliquée par le Modèle}.$$

Le R^2 est le coefficient d'ajustement du modèle. Il s'agit du pourcentage de somme de carrés expliquée par le modèle dans la somme des carrés totale :

$$R^2 = \frac{\text{SCEM}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}}. \quad (3)$$

La variance empirique de la variable log VA est égale à :

$$\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{\text{SCT}}{n}.$$

D'après l'égalité (3), nous avons $\text{SCT} = \frac{\text{SCR}}{1 - R^2} = \frac{148.27}{1 - 0.945} \simeq 2695.8$.

Par conséquent, la variance empirique de la variable log VA est égale à :

$$\frac{2695.8}{1658} \simeq 1.6.$$

3 Nous avons

$$\hat{\alpha} - \alpha \sim N \left(0, \sigma^2 \left((X'X)^{-1} \right)_{(1,1)} \right),$$

$$\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2(n-3).$$

Par conséquent,

$$\frac{\sqrt{n-3}(\hat{\alpha} - \alpha)}{\sqrt{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 ((X'X)^{-1})_{(1,1)}}} \sim T(n-3).$$

À l'aide de cette variable pivotale, nous construisons un intervalle de confiance de niveau 95 % pour le paramètre α . La densité de la loi de student étant symétrique par rapport à l'axe des ordonnées, nous choisissons un intervalle de confiance à risque symétrique. Nous obtenons :

$$\text{IC}_{95\%}(\alpha) = \left[a + F_{T(1655)}^{-1}(0.025) \frac{\sqrt{0.0288 \times \text{SCR}}}{\sqrt{1655}}; a + F_{T(1655)}^{-1}(0.975) \frac{\sqrt{0.0288 \times \text{SCR}}}{\sqrt{1655}} \right].$$

Nous avons $F_{T(1655)}^{-1}(0.025) = -F_{T(1655)}^{-1}(0.975) \simeq -1.6$, $a = 3.136$ et $\text{SCR} = 148.27$.

Par conséquent,

$$\text{IC}_{95\%}(\alpha) = [3.05; 3.22].$$

De la même manière, nous obtenons

$$\text{IC}_{95\%}(\lambda) = [0.72; 0.76].$$

Nous souhaitons maintenant tester l'hypothèse H_0 selon laquelle $\gamma = 0$ contre l'hypothèse H_1 selon laquelle $\gamma > 0$.

Sous l'hypothèse H_0 , nous avons :

$$T_1 = \frac{\sqrt{n-3} \times \hat{\gamma}}{\sqrt{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 ((X'X)^{-1})_{(3,3)}}} \sim T(n-3).$$

La région critique (région de rejet de l'hypothèse H_0) est telle que $T_1 > k$.

Pour un risque de première espèce égal à 5%, nous obtenons

$$k = F_{T(1655)}^{-1}(0.95) \simeq 1.645.$$

Par conséquent, le test est tel que nous rejetons l'hypothèse H_0 si $t_1 > 1.645$ où t_1 est la réalisation de la variable aléatoire T_1 .

$$\text{Nous avons } t_1 = \frac{\sqrt{1655} \times 0.282}{\sqrt{148.27 \times 0.0009}} \simeq 31.40 > 1.645.$$

Ainsi, nous rejetons l'hypothèse selon laquelle $\gamma = 0$ au seuil de 5%.

4

$$\begin{aligned} \alpha(\mu L)^\lambda (\mu K)^\gamma &= \mu \alpha(L)^\lambda (K)^\gamma, \quad \forall \mu \\ \iff \mu^{\lambda+\gamma} &= \mu, \quad \forall \mu \\ \iff \lambda + \gamma &= 1 \end{aligned}$$

Par conséquent, les rendements d'échelle sont constants si et seulement si $\lambda + \gamma = 1$. Il sont croissants si $\lambda + \gamma > 1$.

Ainsi, nous testons l'hypothèse H_0 selon laquelle $\lambda + \gamma = 1$ contre l'hypothèse H_1 selon laquelle $\lambda + \gamma > 1$.

Sous l'hypothèse H_0 , nous avons :

$$\hat{\lambda} + \hat{\gamma} - 1 \sim N\left(0, \sigma^2 \left(\left((X'X)^{-1} \right)_{(2,2)} + \left((X'X)^{-1} \right)_{(3,3)} + 2 \left((X'X)^{-1} \right)_{(3,2)} \right) \right).$$

Par ailleurs,

$$\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2(n-3).$$

Ainsi,

$$T_2 = \frac{\sqrt{n-3} (\hat{\lambda} + \hat{\gamma} - 1)}{\sqrt{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \left(\left((X'X)^{-1} \right)_{(2,2)} + \left((X'X)^{-1} \right)_{(3,3)} + 2 \left((X'X)^{-1} \right)_{(3,2)} \right)}} \sim T(n-3).$$

La région critique est telle que $T_2 > k$. Pour un risque de première espèce égal à 5%, nous obtenons $k = F_{T(1655)}^{-1}(0.95) \simeq 1.645$. Par conséquent, le test est tel que nous rejetons l'hypothèse H_0 si $t_2 > 1.645$ où t_2 est la réalisation de la variable aléatoire T_2 .

$$\text{Nous avons } t_2 = \frac{\sqrt{1655}(0.738 + 0.282 - 1)}{\sqrt{148.27 \times (0.0016 + 0.0009 - 2 \times 0.001)}} \simeq 2.99.$$

Ainsi, nous rejetons l'hypothèse selon laquelle les rendements d'échelle sont constants au seuil de 5%.

5 Nous souhaitons tester H_0 selon laquelle les trois secteurs sont régis par la même fonction de production contre l'hypothèse H_1 selon laquelle les trois secteurs ne sont pas régis par la même fonction de production.

Le modèle s'écrit :

$$\begin{aligned} \log VA_i &= \alpha_1 + \lambda_1 \log L_i + \gamma_1 \log K_i, & i = 1, \dots, 40 \\ \log VA_i &= \alpha_2 + \lambda_2 \log L_i + \gamma_2 \log K_i, & i = 41, \dots, 64 \\ \log VA_i &= \alpha_3 + \lambda_3 \log L_i + \gamma_3 \log K_i, & i = 65, \dots, 302 \end{aligned}$$

Nous souhaitons tester :

$$\begin{aligned} H_0 : \alpha_1 &= \alpha_2 = \alpha_3 \quad \text{et} \quad \lambda_1 = \lambda_2 = \lambda_3 \quad \text{et} \quad \gamma_1 = \gamma_2 = \gamma_3, \\ H_1 : \overline{H_0}. \end{aligned}$$

Nous avons :

$$SCR_{H_0} = \sum_{i=1}^{302} \left(\log VA_i - \hat{\alpha} - \hat{\lambda} \log L_i - \hat{\gamma} \log K_i \right)^2$$

où $\hat{\alpha}$, $\hat{\lambda}$ et $\hat{\gamma}$ sont respectivement les estimateurs MCO des paramètres α_1 , λ_1 et γ_1 sous l'hypothèse H_0 .

$$\begin{aligned} SCR_{H_1} &= \sum_{i=1}^{40} \left(\log VA_i - \hat{\alpha}_1 - \hat{\lambda}_1 \log L_i - \hat{\gamma}_1 \log K_i \right)^2 \\ &+ \sum_{i=41}^{64} \left(\log VA_i - \hat{\alpha}_2 - \hat{\lambda}_2 \log L_i - \hat{\gamma}_2 \log K_i \right)^2 \\ &+ \sum_{i=65}^{302} \left(\log VA_i - \hat{\alpha}_3 - \hat{\lambda}_3 \log L_i - \hat{\gamma}_3 \log K_i \right)^2 \end{aligned}$$

où, pour tout $i \in \{1, 2, 3\}$, $\hat{\alpha}_i$, $\hat{\lambda}_i$ et $\hat{\gamma}_i$ sont respectivement les estimateurs MCO des paramètres α_i , λ_i et γ_i sous l'hypothèse H_1 .

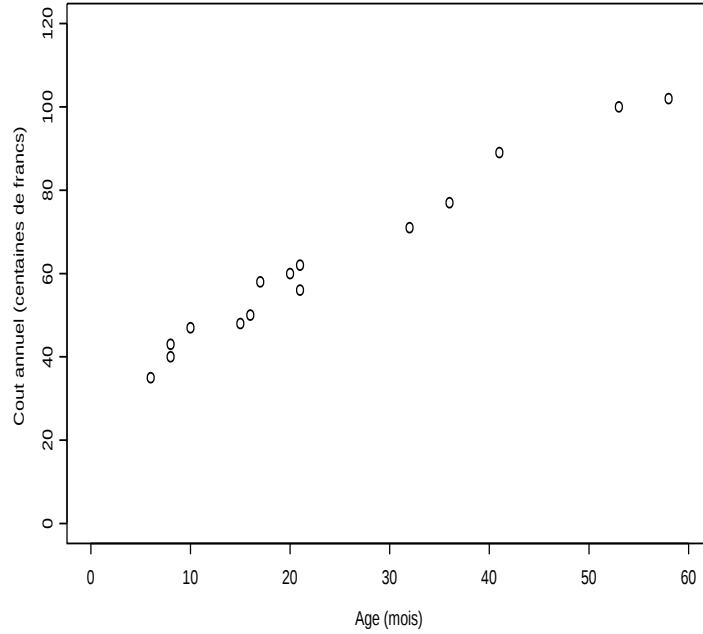
Nous allons appliquer le test de Fisher d'une hypothèse linéaire. Sous H_0 , nous avons $F = \left(\frac{293}{6}\right) \frac{\text{SCR}_{H_0} - \text{SCR}_{H_1}}{\text{SCR}_{H_1}} \sim F(6, 293)$.

La région critique est telle que $F > k'$. Pour un risque de première espèce égal à 5%, nous obtenons $k' = F_{F(6,293)}^{-1}(0.95) \simeq 2.1$. Par conséquent, le test est tel que nous rejetons l'hypothèse H_0 si $f > 2.1$ où f est la réalisation de la variable aléatoire F . Nous avons $f \simeq 44.39 > 2.1$.

Ainsi, nous rejetons l'hypothèse selon laquelle les trois secteurs sont régis par la même fonction de production au seuil de 5%.

Exercice 5

1 Nous avons :



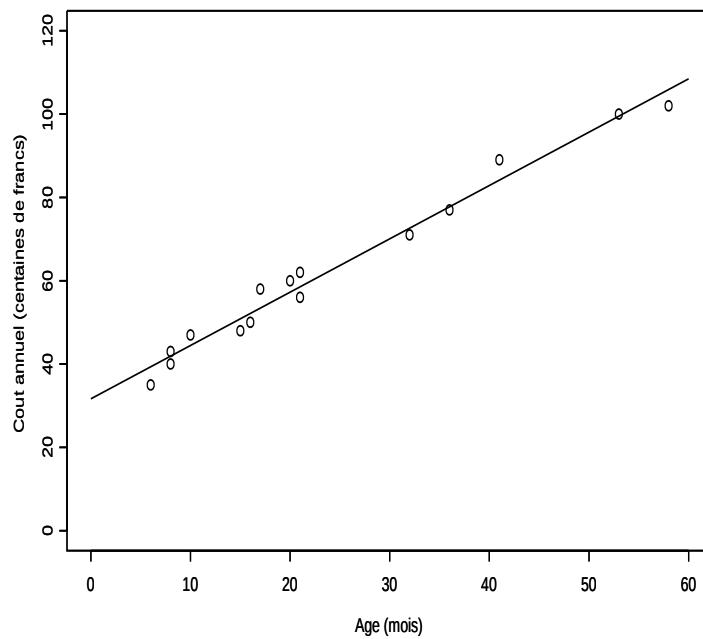
2 Soit $\hat{\alpha}$ et $\hat{\beta}$ les estimateurs MCO des paramètres α et β .

Nous avons $\hat{\beta} = \frac{\sum_{i=1}^{15} x_i Y_i - 15 \bar{x} \bar{Y}}{\sum_{i=1}^{15} x_i^2 - 15 (\bar{x})^2}$ et $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x}$.

Ainsi, $a \simeq 31.67$ et $b \simeq 1.28$.

3 La droite de régression est telle que :

$$y = 31.67 + 1.28x.$$



4 Un prédicteur $\hat{Y}_i$ de Y_i est $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$.

Nous déterminons aisément que $\text{SCR} = \sum_{i=1}^{15} (y_i - \hat{y}_i)^2 = 132.015$.

5 Un estimateur non biaisé de σ^2 est $\hat{\sigma}^2 = \left(\frac{1}{13}\right) \sum_{i=1}^{15} (Y_i - \hat{Y}_i)^2$. Soit s^2 la réalisation de $\hat{\sigma}^2$.

Nous avons $s^2 = \frac{\text{SCR}}{13} = 10.155$.

6 Nous avons $\text{SCT} = \text{SCEM} + \text{SCR}$.

Par conséquent, $\text{SCEM} = 6137.719 - 132.015 = 6005.704$.

7 Le coût de maintenance pour un véhicule utilitaire de 4 ans est estimé à 9311 francs.