Université Rennes 1
Master 1 Mathématiques année 2022-2023
Nathalie Krell

Fiche de TD n°4 : Modèle de régression linéaire

1 Modèle de régression linéaire simple

Définition :Un modèle de régression linéaire simple est défini par une équation de la forme :

$$\forall i \in \{1, ..., n\} \quad Y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

Dans un premier temps on suppose uniquement que $\mathbb{E}(\epsilon_i) = 0 \ \ \forall i \in \{1, ..., n\}$ et $Cov(\epsilon_i, \epsilon_j) = \delta_{i,j}\ \sigma^2 \ \ \forall (i,j) \in \{1, ..., n\}^2$.

[a] Ecrire le modèle sous la forme d'un modèle linéaire.

[b] Soient $\hat{\beta}_1$ et $\hat{\beta}_2$ les estimateurs des moindres carrés. Que valent-ils ?

[c] Montrer que parmi les estimateurs sans biais linéaires en Y, les estimateurs $\hat{\beta}_j$ sont de variances minimales.

[d] Soit $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ et $\hat{\epsilon}_i = y_i - \hat{y}_i \ \forall i \{1, ..., n\}$.
Montrer que la statistique $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2$ est un estimateur sans biais de σ^2.

[e] On suppose dans toute la suite de l'exercice que les ϵ_i sont i.i.d de loi $\mathcal{N}(0, \sigma^2)$. Quel est l'estimateur du maximum de vraisemblance de $\theta = (\beta_1, \beta_2, \sigma^2)$?

[f] Soit

$$\begin{array}{llll} c & = \frac{-\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2} & \hat{\sigma}^2 & = \frac{1}{n-2} \sum \hat{\epsilon}_i^2 \\ \sigma_1^2 & = \sigma^2 \left(\frac{\sum x_i^2}{\sum (x_i - \bar{x})^2} \right) & \hat{\sigma}_1^2 & = \hat{\sigma}^2 \left(\frac{\sum x_i^2}{\sum (x_i - \bar{x})^2} \right) \\ \sigma_2^2 & = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} & \hat{\sigma}_2^2 & = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} \end{array}$$

Que peut-on dire de ces quantités ? Y-a-t-il un lien avec les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$?

[g] Montrer la proposition suivante :

Proposition 1 *(a)* $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 V)$ *où* $\beta = (\beta_1, \beta_2)'$ *et* $V = \frac{1}{\sigma^2} \begin{pmatrix} \sigma_1^2 & c \\ c & \sigma_1^2 \end{pmatrix}$.

(b) $\frac{n-2}{\sigma^2} \hat{\sigma}^2 \sim \chi^2_{n-2}$ *loi du* χ^2 *à* $n-2$ *degrés de liberté.*

(c) $\hat{\beta}$ *et* $\hat{\sigma}^2$ *sont indépendants.*

[h] Montrer la proposition suivante :

Proposition 2 *(a)* $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_1} \sim \mathcal{T}_{n-2}$.

(b) $\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_2} \sim \mathcal{T}_{n-2}$.

(c) $\frac{1}{2\hat{\sigma}^2} (\hat{\beta} - \beta)' V^{-1} (\hat{\beta} - \beta) \sim \mathcal{F}^2_{n-2}$.

[i] En déduire des intervalles de confiance des différents paramètres.

2 On considère le modèle de régression linéaire simple

$$\forall i \in \{1, ..., n\} \quad Y_i = \beta_1 + \beta_2 x_i + \epsilon_i.$$

Soit un échantillon $(x_i, y_i)_{1 \leq i \leq 100}$ de statistiques résumées

$$\sum_{i=1}^{100} x_i = 0 \quad \sum_{i=1}^{100} x_i^2 = 400 \quad \sum_{i=1}^{100} x_i y_i = 100 \quad \sum_{i=1}^{100} y_i = 100 \quad \sum_{i=1}^{100} x_i = 0 \quad \hat{\sigma}^2 = 1$$

[a] Exprimer les intervalles de confiance à 95% pourβ_1 et β_2.

[b] Donner l'équation de la région de confiance à 95% de (β_1, β_2). (Rappel : l'ensemble des points (x, y) tels que $\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} \leq 1$ est l'intérieur d'une ellipse centrée en (x_0, y_0), dont les axes sont parallèles à ceux des abscisses et des ordonnées, et de sommets $(x_0 \pm a, 0)$ et $(0, y_0 \pm b)$.)

3 **Modèle de régression linéaire simple**
Nous considérons le modèle statistique suivante :

$$\forall i \in \{1, ..., n\} \quad Y_i = \beta x_i + \epsilon_i$$

On suppose que $\mathbb{E}(\epsilon_i) = 0 \quad \forall i \in \{1, ..., n\}$ et $Cov(\epsilon_i, \epsilon_j) = \delta i, j \; \sigma^2 \quad \forall (i, j) \in \{1, ..., n\}^2$.

[a] En revenant à la définition des moindres carrés, montrer que l'estimateur des moindres carrés de β vaut

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}.$$

[b] Montrer que la droite passant par l'origine et le centre de gravité du nuage de points est $y = \beta^* x$, avec

$$\beta^* = \frac{\sum Y_i}{\sum x_i}.$$

[c] Montrer que $\hat{\beta}$ et β^* sont tous deux des estimateurs sans biais de β.

[d] En utilisant l'inégalité de Cauchy-Schwarz, montrer que $Var(\beta^*) > Var(\hat{\beta})$ sauf dans le cas où tous les x_i sont égaux. Ce résultat était-il prévisible ?

4 **Régression sur variables centrées** Nous considérons le modèle de régression linéaire

$$Y = X\beta + \epsilon \tag{1}$$

où $Y \in \mathbb{R}^n$, X est une matrice de taille $n * p$ de rang p, $\beta \in \mathbb{R}^p$ et $\epsilon \in \mathbb{R}^n$. La première colonne de X est le vecteur constant $\mathbb{1}$. X peut donc s'écrire $X = [\mathbb{1}, Z]$ où $Z = [X_2, ..., X_p]$ est la matrice $n * (p-1)$ des $(p-1)$ derniers vecteurs colonnes de X. Le modèle peut donc s'écrire sous la forme :

$$Y = \beta_1 \mathbb{1} + Z\beta_{(1)} + \epsilon$$

où β_1 est la première coordonnée du vecteur β et $\beta_{(1)}$ représente le vecteur β privé de sa première coordonnée.

[a] Donner $P_{\mathbb{1}}$, matrice de projection orthogonale sur le sous-espace engendré par le vecteur $\mathbb{1}$.

[b] En déduire la matrice de projection orthogonale $P_{\mathbb{1}^\perp}$ sur le sous-espace $\mathbb{1}^\perp$ orthogonal au vecteur $\mathbb{1}$.

[c] Calculer $P_{\mathbb{1}^\perp} Z$.

[d] En déduire que l'estimateur de β des Moindres Carrés Ordinaires du modèle (1) peut être obtenu en minimisant par les MCO le modèle suivant :

$$\tilde{Y} = \tilde{Z}\beta_{(1)} + \eta,$$

où $\tilde{Y} = P_{\mathbb{1}^\perp} Y$ et $\tilde{Z} = P_{\mathbb{1}^\perp} Z$.

5 **Modèle à deux variables explicatives** On considère le modèle de régression suivant :

$$\forall i \in \{1, ..., n\} \quad Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i.$$

Les $x_{i,j}$ sont des variables exogènes du modèle, les ϵ_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. En posant :

$$X = \begin{pmatrix} 1 & x_{1,2} & x_{1,3} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,2} & x_{n,3} \end{pmatrix} \quad \text{et} \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

on a observé

$$X'X = \begin{pmatrix} 30 & 20 & 0 \\ 20 & 20 & 0 \\ 0 & 0 & 10 \end{pmatrix}, \qquad X'Y = \begin{pmatrix} 15 \\ 20 \\ 10 \end{pmatrix}, \qquad Y'Y = 59.5.$$

[a] Déterminer la valeur de n, la moyenne des $x_{i,3}$, le coefficient de corrélation des $x_{i,2}$ et des $x_{i,3}$.

[b] Estimer β_1, β_2,β_3, σ^2 par la méthode des moindres carrés ordinaires.

6

[a] (a) On considère le modèle de régression suivant :

$$Y = X\beta + \epsilon.$$

où $Y \in \mathbb{R}^n$, X est une matrice de taille $n * p$ de rang p, $\beta \in \mathbb{R}^p$ et $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

(b) Qu'appelle-t-on estimateur des moindres carrés $\hat{\beta}$ de β ? Rappeler sa formule.

(c) Quelle est l'interprétation géométrique de $\hat{Y} = X\hat{\beta}$ (faites un dessin) ?

(d) Rappeler espérances et matrices de covariance de $\hat{\beta}$, $\hat{Y}$ et $\hat{\epsilon}$.

[b] (a) Nous considérons dorénavant un modèle avec 4 variables explicatives (la première variable étant la constante). Nous avons observé :

$$X'X = \begin{pmatrix} 100 & 20 & 0 & 0 \\ 20 & 20 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad X'Y = \begin{pmatrix} -60 \\ 20 \\ 10 \\ 1 \end{pmatrix}, \qquad Y'Y = 159.$$

(b) Estimer β et σ^2.

(c) Donner un estimateur de la variance de $\hat{\beta}$.

(d) Donner un intervalle de confiance pour β_2, au niveau 95%.