

1 Théorie

Les données proviennent de l'observation d'un échantillon statistique de taille n de vecteurs de \mathbb{R}^{p+1} :

$$(Z_i^1, \dots, Z_i^p, Y_i) \quad i = 1, \dots, n.$$

On cherche à expliquer une variable quantitative Y_i (réponse) par p variables Z_i^1, \dots, Z_i^p dites *variables explicatives* (ou encore *régresseurs*).

On pose $\theta := (\beta_1, \dots, \beta_p, \sigma^2) \in \Theta := \mathbb{R}^p \times \mathbb{R}_+^*$. Le *modèle linéaire* consiste à supposer que, pour tout $\theta \in \Theta$, les variables

$$\varepsilon_i(\theta) = \sigma^{-1} \{Y_i - (\beta_1 Z_i^1 + \dots + \beta_p Z_i^p)\}, \quad i = 1, 2, \dots, n,$$

sont des variables indépendantes et identiquement distribuées. Il est plus pratique dans ce cas de représenter le modèle sous la forme matricielle

$$\mathbf{Y} = \mathbf{Z}\beta + \sigma\varepsilon(\theta), \quad (1)$$

où

$$\mathbf{Y} := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \beta := \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon(\theta) := \begin{bmatrix} \varepsilon_1(\theta) \\ \vdots \\ \varepsilon_n(\theta) \end{bmatrix}, \quad \mathbf{Z} := \begin{bmatrix} Z_1^1 & \dots & Z_1^p \\ Z_2^1 & \dots & Z_2^p \\ \vdots & & \vdots \\ Z_n^1 & \dots & Z_n^p \end{bmatrix};$$

\mathbf{Y} est le vecteur des observations, β est le vecteur des paramètres de régression et \mathbf{Z} est la matrice de régression de taille $n \times p$.

Pour estimer le paramètre $\beta \in \mathbb{R}^p$ dans le modèle de régression linéaire, la *méthode des moindres carrés* consiste à chercher un estimateur $\hat{\beta} \in \mathbb{R}^p$ qui minimise le risque quadratique empirique i.e. minimise la fonction

$$\mathbf{u} \mapsto J_n(\mathbf{u}) := \sum_{i=1}^n (Y_i - u_1 Z_i^1 - \dots - u_p Z_i^p)^2 = \|\mathbf{Y} - \mathbf{Z}\mathbf{u}\|^2, \quad \text{où } \mathbf{u} := \begin{bmatrix} u_1 \\ \vdots \\ u_p \end{bmatrix}.$$

1. Montrer que toute solution $\hat{\mathbf{u}} \in \arg \min_{\mathbf{u} \in \mathbb{R}^p} J_n(\mathbf{u})$ est solution des *équations d'estimation* :

$$\mathbf{Z}^T \mathbf{Y} = \mathbf{Z}^T \mathbf{Z} \mathbf{u},$$

en $\mathbf{u} \in \mathbb{R}^p$.

Nous supposons dans la suite que les hypothèses de *Gauss-Markov* sont vérifiées :

GM1 $n > p$ et la matrice \mathbf{Z} est de rang p .

GM2 les erreurs de régression sont *homoscédastiques*, pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[\varepsilon(\theta)] = 0$ et $\text{Var}_\theta(\varepsilon(\theta)) = \mathbf{I}_n$.

Notons que l'hypothèse (GM1) implique que la matrice de Gram $\mathbf{Z}^T \mathbf{Z}$ est inversible. Nous posons

$$\mathbf{Z}^\# = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T,$$

qui est appelée la *pseudo-inverse* de \mathbf{Z} .

2. Montrer que $\mathbf{Z}^\# \mathbf{Z} = \mathbf{I}_p$ et $\mathbf{Z} \mathbf{Z}^\# = \mathbf{H}$ où \mathbf{H} est le projecteur orthogonal sur l'espace vectoriel engendré par les colonnes de la matrice \mathbf{Z} .
3. Montrer que l'estimateur des moindres carrés est unique et a pour expression :

$$\hat{\beta} := (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} = \mathbf{Z}^\# \mathbf{Y}. \quad (2)$$

L'estimateur $\hat{\beta}$ est dit *linéaire*, car il est obtenu en calculant une combinaison linéaire des observations Y_1, \dots, Y_n .

4. Montrer que l'estimateur des moindres carrés est un estimateur sans biais de β .
5. Montrer que la matrice de covariance de cet estimateur est donnée par :

$$\text{Var}_\theta(\hat{\beta}) = \sigma^2 (\mathbf{Z}^T \mathbf{Z})^{-1}. \quad (3)$$

Soit \mathbf{B} une matrice déterministe de taille $p \times n$; on pose $\tilde{\beta} := \mathbf{B} \mathbf{Y}$.

6. Montrer que l'estimateur $\tilde{\beta}$ est sans biais si et seulement si $\mathbf{B} \mathbf{Z} = \mathbf{I}_p$.

Dans la suite du problème, nous supposons que $\mathbf{B} \mathbf{Z} = \mathbf{I}_p$.

7. Montrer que pour tout $\theta \in \Theta$, la matrice de covariance de l'estimateur $\tilde{\beta}$ est

$$\text{Var}_\theta(\tilde{\beta}) = \sigma^2 \mathbf{B} \mathbf{B}^T.$$

8. Montrer que pour tout $\theta \in \Theta$,

$$\mathbb{E}_\theta \left[(\tilde{\beta} - \beta)(\hat{\beta} - \beta)^T \right] = \sigma^2 \mathbf{B} (\mathbf{Z}^\#)^T = \sigma^2 (\mathbf{Z}^T \mathbf{Z})^{-1}.$$

Si A et B sont deux matrices symétriques $p \times p$, nous notons $A \succeq B$ si et seulement si, pour tout $x \in \mathbb{R}^p$, $x^T A x \geq x^T B x$.

9. Montrer que pour tout $\theta \in \Theta$, $\text{Var}_\theta(\tilde{\beta}) \succeq \text{Var}_\theta(\hat{\beta})$.

Nous appelons $\hat{\mathbf{Y}} = \mathbf{Z} \hat{\beta}$ la *prédiction* des observations \mathbf{Y} . En observant que $\mathbf{Z} \hat{\beta} = \mathbf{H} \mathbf{Y}$, la prédiction $\hat{\mathbf{Y}}$ est la projection orthogonale de \mathbf{Y} sur l'espace engendré par les colonnes de la matrice de régression \mathbf{Z} . Nous appelons les *résidus de régression* les composantes du vecteur

$$\mathbf{Y} - \mathbf{Z} \hat{\beta} = \mathbf{Y} - \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} = (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}. \quad (4)$$

Considérons la statistique définie comme la somme des carrés des résidus (appelée *Sum of Squared Errors of prediction* ou *SSE* dans la littérature anglo-saxonne) :

$$\text{SSE} := \|\mathbf{Y} - \mathbf{H} \mathbf{Y}\|^2. \quad (5)$$

10. Montrer que pour tout $\theta \in \Theta$:

$$\|(\mathbf{I}_n - \mathbf{H}) \varepsilon(\theta)\|^2 = \text{Tr}((\mathbf{I}_n - \mathbf{H}) \varepsilon(\theta) \varepsilon^T(\theta) (\mathbf{I}_n - \mathbf{H})).$$

11. En déduire que

$$\hat{\sigma}^2 := (n - p)^{-1} \text{SSE}$$

est un estimateur sans biais de la variance σ^2 .

12. Montrer que, pour tout $\theta \in \Theta$,

$$\mathbb{E}_\theta[\hat{\mathbf{Y}}(\mathbf{Y} - \hat{\mathbf{Y}})^T] = 0.$$

Considérons finalement la somme des carrés de régression (ou *régression sum of squares*, RSS) :

$$\text{RSS} := \|\mathbf{H}\mathbf{Y}\|^2. \quad (6)$$

13. Montrer que

$$\text{TSS} := \|\mathbf{Y}\|^2 = \text{RSS} + \text{SSE}. \quad (7)$$

On suppose maintenant que \mathbf{Y} est l'observation canonique d'un modèle gaussien

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{N(\mathbf{Z}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) : \boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*\}).$$

14. Déterminer l'estimateur du maximum de vraisemblance du paramètre $\boldsymbol{\theta}$. Qu'en déduire ?
 15. Déterminer la distribution de l'estimateur des moindres carrés $\hat{\boldsymbol{\beta}}$.
 16. Déterminer la distribution de $\hat{\sigma}^2$.
 17. Montrer que $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}^2$ sont indépendants.
 18. Soit $\mathbf{x} \in \mathbb{R}^p$. Montrer que

$$\frac{\mathbf{x}^T \hat{\boldsymbol{\beta}} - \mathbf{x}^T \boldsymbol{\beta}}{\hat{\sigma} \sqrt{\mathbf{x}^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{x}}}$$

suit une loi de Student à $(n - p)$ degrés de liberté.

19. Déterminer un intervalle de confiance bilatéral de niveau de couverture $1 - \alpha$ pour $\boldsymbol{\beta}^T \mathbf{x}$ pour $\alpha \in]0, 1[$.
 20. Construire un test de l'hypothèse

$$H_0 : \boldsymbol{\beta}^T \mathbf{x} = 0, \quad \text{contre} \quad H_1 : \boldsymbol{\beta}^T \mathbf{x} \neq 0$$

de niveau α .

21. Déterminer la p -valeur de ce test.
 22. Soit R une matrice de taille $q \times p$ de rang $q \leq p$. Montrer que

$$\frac{1}{q\hat{\sigma}^2} \{R(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^T [R(\mathbf{Z}^T \mathbf{Z})^{-1} R^T]^{-1} \{R(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}$$

suit une loi de Fisher $q, n - p$ degrés de liberté.

23. Déterminer une région de confiance pour le vecteur (β_1, β_2) .

2 Pratique

La variable à expliquer est la concentration en ozone notée O_3 et les variables explicatives sont la température notée T_{12} , le vent noté V_x et la nébulosité notée N_{e12} . Nous rajouterons dans la matrice de régression le vecteur constant $(1, \dots, 1)^T$ que nous appelons *intercept*.

1. Estimer l'estimateur des moindres carrés du paramètre.
2. Déterminer les intervalles de confiance bilatères à 95% pour chaque valeur des paramètres.
3. Visualiser les régions de confiance à 95 % pour (β_1, β_2) et (β_1, β_3) .

Nous cherchons à répondre aux questions suivantes :

- (i) est-ce que la valeur de O_3 est influencée par V_x ?
 - (ii) y a-t-il un effet nébulosité ?
 - (iii) est-ce que la valeur de O_3 est influencée par V_x ou T_{12} ?
4. Formuler les différentes questions comme des tests d'hypothèses.
 5. Construire des procédures de tests pour ces trois hypothèses.
 6. Conclure.