1 Théorie

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

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

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

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

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

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

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

où

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

$\mathbf{Y}$ 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

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

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 $\mathrm{Var}_\theta(\varepsilon(\theta)) = \mathrm{I}_n$.

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 $\mathbf{Z}^\#\mathbf{Z} = \mathrm{I}_p$ et $\mathbf{Z}\mathbf{Z}^\# = \mathrm{H}$ où 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 :

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

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 :

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

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 $\mathbf{B}\mathbf{Z} = \mathrm{I}_p$.

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

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

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

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 $x \in \mathbb{R}^p$, $x^TAx \geq x^TBx$.

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

Nous appelons $\hat{\mathbf{Y}} = \mathbf{Z}\hat{\beta}$ la *prédiction* des observations $\mathbf{Y}$. En observant que $\mathbf{Z}\hat{\beta} = H\mathbf{Y}$, 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

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

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) :

$$\mathrm{SSE} := \|\mathbf{Y} - \mathrm{H}\mathbf{Y}\|^2. \tag{5}$$

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

$$\|(\mathrm{I}_n - \mathrm{H})\varepsilon(\theta)\|^2 = \mathrm{Tr}\left((\mathrm{I}_n - \mathrm{H})\varepsilon(\theta)\varepsilon^T(\theta)(\mathrm{I}_n - \mathrm{H})\right).$$

11. En déduire que

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

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) :
$$\text{RSS} := \|\,\mathrm{H}\mathbf{Y}\|^2. \tag{6}$$

13. Montrer que
$$\text{TSS} := \|\mathbf{Y}\|^2 = \text{RSS} + \text{SSE}. \tag{7}$$

On suppose maintenant que $\mathbf{Y}$ est l'observation canonique d'un modèle gaussien
$$(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathrm{N}(\mathbf{Z}\beta, \sigma^2\mathrm{I}_n)\,:\ \theta = (\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*\}).$$

14. Déterminer la matrice de Fisher et la borne de Cramer-Rao.
15. Déterminer l'estimateur du maximum de vraisemblance du paramètre θ. Qu'en déduire ?
16. Déterminer la distribution de l'estimateur des moindres carrés $\hat{\beta}$.
17. Déterminer la distribution de $\hat{\sigma}^2$.
18. Montrer que $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants.
19. Soit $\mathbf{x} \in \mathbb{R}^p$. Montrer que
$$\frac{\mathbf{x}^T\hat{\beta} - \mathbf{x}^T\beta}{\hat{\sigma}\sqrt{\mathbf{x}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{x}}}$$
suit une loi de Student à $(n-p)$ degrés de liberté.
20. Déterminer un intervalle de confiance bilatéral de niveau de couverture $1-\alpha$ pour $\beta^T\mathbf{x}$ pour $\alpha \in]0,1[$.
21. Construire un test de l'hypothèse
$$H_0\,:\ \beta^T\mathbf{x} = 0, \qquad \text{contre} \qquad H_1\,:\ \beta^T\mathbf{x} \neq 0$$
de niveau α.
22. Déterminer la p-valeur de ce test.
23. Soit R une matrice de taille $q \times p$ de rang $q \leq p$. Montrer que
$$\frac{1}{q\hat{\sigma}^2}\{R(\hat{\beta}-\beta)\}^T[R(\mathbf{Z}^T\mathbf{Z})^{-1}R^T]^{-1}\{R(\hat{\beta}-\beta)\}$$
suit une loi de Fisher q,$n-p$ degrés de liberté.
24. Déterminer une région de confiance pour le vecteur (β_1, β_2).

2 Pratique

Nous allons maintenant illustrer l'utilisation de ces résultats en analysant des données journalières de concentration en ozone. Nous conseillons d'utiliser `Jupyter` et pour la régression linéaire, la fonction `R` `lm` (il y a beaucoup de tutoriels intéressants en ligne sur la régression linéaire que nous vous conseillons de consulter).

La variable à expliquer est la concentration en ozone notée `O3` et les variables explicatives sont la température notée `T12`, le vent noté `Vx` et la nébulosité notée `Ne12`. Nous rajouterons dans la matrice de régression le vecteur constant $(1,\cdots,1)^T$ 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 `O3` est influencée par `Vx` ?

(ii) y a-t-il un effet nébulosité ?

(iii) est-ce que la valeur de `O3` est influencée par `Vx` ou `T12` ?

4. Formuler les différentes questions comme des tests d'hypothèses.
5. Construire des procédures de tests pour ces trois hypothèses.
6. Conclure.