1 Théorie

The data are derived from the observation of a statistical sample of n vectors in $\mathbb{R}^{p+1}$:

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

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

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

où

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

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

1.1 Estimation par moindres carrés du vecteur $\boldsymbol{\beta}$

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

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

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* en $\mathbf{u} \in \mathbb{R}^p$:
$$\mathbf{Z}^T\mathbf{Y} = \mathbf{Z}^T\mathbf{Z}\mathbf{u} \,.$$

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 $\theta \in \Theta$, $\mathbb{E}_\theta[\boldsymbol{\varepsilon}(\theta)] = 0$ et $\mathrm{Var}_\theta(\boldsymbol{\varepsilon}(\theta)) = \mathrm{I}_n$.

Notons que l'hypothèse (GM1) implique que la matrice de Gram $\mathbf{Z}^T\mathbf{Z}$ est inversible et l'hypothèse (GM2) entraine que pour tout $\theta \in \Theta$

$$\mathbb{E}_\theta\left[\mathbf{Y}\right] = \mathbf{Z}\boldsymbol{\beta}\ , \qquad \mathrm{Var}_\theta\left[\mathbf{Y}\right] = \sigma^2 \mathrm{I}_n\ .$$

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 :

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

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

4. Montrer que l'estimateur des moindres carrés est un estimateur sans biais de β.
5. Montrer que pour tout $\theta \in \Theta$, la matrice de covariance de cet estimateur est donnée par :

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

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

6. Montrer que l'estimateur $\tilde{\boldsymbol{\beta}}$ 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 $\theta \in \Theta$,

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

   Noter que cette quantité est la matrice de covariance $\mathrm{Cov}_\theta(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}})$.

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

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

En conclusion, nous venons d'établir que l'estimateur des moindres carrés de $\boldsymbol{\beta}$ est

$$\widehat{\boldsymbol{\beta}} := (\mathbf{Z}^T\mathbf{Z})^{-1}\,\mathbf{Z}^T\mathbf{Y}\ .$$

C'est un estimateur sans biais, de variance $\sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}$. Il est de variance minimale dans la classe des estimateurs sans biais et linéaires.

1.2 Estimation de la variance σ^2 et Coefficient de détermination

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

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

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) :

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

9. En observant que pour tout vecteur $w \in \mathbb{R}^\ell$, on a $\mathrm{Tr}(w\,w^T) = \|w\|^2$, montrer que

$$\hat{\sigma}^2 := (n-p)^{-1}\mathrm{SSE} = \frac{1}{n-p}\|\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}\|^2$$

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

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

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

10. Montrer que

$$\|\mathbf{Y}\|^2 = \mathrm{RSS} + \mathrm{SSE}. \tag{7}$$

> En conclusion, l'estimateur
>
> $$\hat{\sigma}^2 := \frac{1}{n-p}\|\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}\|^2$$
>
> est un estimateur sans biais de σ^2. De plus, *le coefficient de détermination* défini par
>
> $$R^2 := \frac{\mathrm{RSS}}{\|\mathbf{Y}\|^2}$$
>
> est un nombre dans $[0, 1]$, qui représente la part de variation de $\mathbf{Y}$ expliquée par le modèle de régression.

1.3 Cas de la régression linéaire gaussienne

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

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

11. Déterminer l'estimateur du maximum de vraisemblance du paramètre θ.
12. Pour tout $\theta \in \Theta$, déterminer la distribution de l'estimateur des moindres carrés $\hat{\boldsymbol{\beta}}$ sous $\mathbb{P}_\theta$.
13. Pour tout $\theta \in \Theta$, déterminer la distribution de $\hat{\sigma}^2$ sous $\mathbb{P}_\theta$.

14. Pour tout $\theta \in \Theta$, montrer que $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}^2$ sont indépendants sous $\mathbb{P}_\theta$.

> En conclusion, dans le cas où les erreurs de régression sont gaussiennes, l'estimateur de $\boldsymbol{\beta}$ par moindres carrés et celui par maximum de vraisemblance coïncident. L'estimateur par maximum de vraisemblance de σ^2 est un estimateur biaisé. Enfin, les deux estimateurs $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}^2$ sont indépendants et sous $\mathbb{P}_\theta$, nous avons
>
> $$\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}\right)\ , \qquad (n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

1.4 Tests statistiques, cas régression linéaire gaussienne

15. Soit $\mathbf{x} \in \mathbb{R}^p$. Montrer que sous $\mathbb{P}_\theta$,

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

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

16. Soit $\alpha \in]0,1[$. Déterminer un intervalle de confiance bilatéral de niveau de couverture $1-\alpha$ pour $\boldsymbol{\beta}^T\mathbf{x}$.
17. Soit $\alpha \in]0,1[$. 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.

18. Déterminer la p-valeur de ce test.
19. Soit A une matrice de taille $q \times p$ de rang $q \leq p$. Montrer que sous $\mathbb{P}_\theta$,

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

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

20. Déterminer une région de confiance pour le vecteur (β_1, β_2) (on a posé $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_p]^T$).

2 Pratique

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.