Modéle de régression linéaire - Feuille 4

Projections

Exercice 1 Soit X de loi $\mathcal{N}(0, I_n)$ et δ un vecteur fixé de $\mathbb{R}^n$. On définit la variable aléatoire réelle Y par

$$Y = || X + \delta ||^2 = \sum_{i=1}^{n} (X_i + \delta_i)^2.$$

On dit que Y suit un $\chi^2(n, || \delta ||^2)$. Supposons $\delta \neq 0$. Calculer l'espérance et la variance de Y.

Exercice 2 Soit X de loi $\mathcal{N}(0, I_n)$. Soit P une projection orthogonale de rang r. Montrer que $X'PX$ suit un χ^2_r.

Exercice 3 Soit Z une matrice de taille $n \times q$ de rang q et soit X une matrice $n \times p$ de rang p composée des q vecteurs colonne de Z et de $p - q$ autres vecteurs linéairement indépendants. Nous considérons les deux modèles suivants

$$Y = Z\beta + \varepsilon$$
$$Y = X\beta^* + \eta$$

Supposons pour simplifier que la constante ne fait partie d'aucun modèle. Notons respectivement P_X et P_Z les projections orthogonales sur les sous-espaces $\mathcal{M}(X)$ et $\mathcal{M}(Z)$ engendrés par les p colonnes de X et les q colonnes de Z. Notons enfin $P_{X \cap Z^\perp}$ la projection orthogonale sur $\mathcal{M}(X) \cap \mathcal{M}(Z)^\perp$, orthogonal de $\mathcal{M}(Z)$ dans $\mathcal{M}(X)$, soit

$$\mathbb{R}^n = \mathcal{M}(X) \overset{\perp}{\oplus} \mathcal{M}(X)^\perp = \Big(\mathcal{M}(Z) \overset{\perp}{\oplus} (\mathcal{M}(X) \cap \mathcal{M}(Z)^\perp)\Big) \overset{\perp}{\oplus} \mathcal{M}(X)^\perp.$$

1. Exprimer $\|P_X Y\|^2$ en fonction de $\|P_Z Y\|^2$ et $\|P_{X \cap Z^\perp} Y\|^2$.
2. Comparer alors les coefficients de détermination des deux modèles, soit R^2_Z et R^2_X.
3. De facon générale, qu'en déduire quant à l'utilisation du R^2 pour le choix de variables ?

Exercice 4 Deux variables explicatives

On examine l'évolution d'une variable réponse y_i en fonction de deux variables explicatives x_i et z_i. Soit $X = (\mathbf{1}\ x\ z)$ la matrice $n \times 3$ du plan d'expérience.

1. Nous avons obtenu :

$$X'X = \begin{bmatrix} 25 & 0 & 0 \\ ? & 9.3 & 5.4 \\ ? & ? & 12.7 \end{bmatrix}, \quad (X'X)^{-1} = \begin{bmatrix} 0.04 & 0 & 0 \\ 0 & 0.1428 & -0.0607 \\ 0 & -0.0607 & 0.1046 \end{bmatrix}$$

   (a) Donner les valeurs manquantes.

   (b) Que vaut n ?

   (c) Calculer le coefficient de corrélation empirique entre x et z.

2. La régression linéaire de Y sur $(\mathbf{1}, x, z)$ donne :

$$Y = -1.61\mathbf{1} + 0.61x + 0.46z + \widehat{\varepsilon}, \quad SCR = \|\widehat{\varepsilon}\|^2 = 0.3$$

   Déterminer la moyenne empirique $\overline{y}$. Calculer la somme des carrés expliquée et totalement, ainsi que R^2, R^2_a.

Exercice 5 (Régression sur variables centrées) Nous considérons le modèle de régression linéaire

$$Y = X\beta + \varepsilon, \tag{1}$$

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

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

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

1. Donner $P_{\mathbb{1}}$, matrice de projection orthogonale sur le sous-espace engendré par le vecteur $\mathbb{1}$.
2. En déduire la matrice de projection orthogonale $P_{\mathbb{1}^\perp}$ sur le sous-espace $\mathbb{1}^\perp$ orthogonal au vecteur $\mathbb{1}$.
3. Calculer $P_{\mathbb{1}^\perp} Z$.
4. 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 :

$$\tilde{Y} = \tilde{Z}\beta_{(1)} + \eta, \tag{2}$$

   où $\tilde{Y} = P_{\mathbb{1}^\perp} Y$ et $\tilde{Z} = P_{\mathbb{1}^\perp} Z$.
5. Ecrire la SCR estimée dans le modèle (2) en fonction des variables du modèle (2). Vérifier que la SCR du modèle (2) est identique à celle qui serait obtenue par l'estimation du modèle (1).

Exercice 6 Soit $(Y_i)_{1 \le i \le n}$ une famille de variables aléatoires définie par :

$$Y_i = \theta_0 + \sum_{k=1}^{p} \theta_k \, Z_i^{(k)} + \varepsilon_i \qquad \text{pour tout } i \in \{1, \ldots, n\}, \qquad \text{où :} \tag{3}$$

— $\theta = {}^t(\theta_0, \theta_1, \ldots, \theta_p)$ est un vecteur composé de $p+1$ réels inconnus.

— pour $1 \le j \le p$, les $(Z_i^{(j)})_{1 \le i \le n}$ sont p familles de réels connues.

On note $X = \begin{pmatrix} 1 & Z_1^{(1)} & \cdots & Z_1^{(p)} \\ \vdots & \vdots & & \vdots \\ 1 & Z_n^{(1)} & \cdots & Z_n^{(p)} \end{pmatrix}$ et on suppose que son rang est $p+1$ avec $p+1 \le n+1$.

— la suite $(\varepsilon_i)_i$ est une suite de v.a.i.i.d. de loi **gaussienne** centrée de variance $\sigma^2 > 0$.

1. On note $Y = (Y_i)_{1 \le i \le n}$ et $\varepsilon = (\varepsilon_i)_{1 \le i \le n}$. Ecrire le modèle (3) sous une forme matricielle, en précisant la loi du vecteur d'erreur ε.
2. Rappeler l'expression de l'estimateur $\widehat{\theta}$ de θ par moindres carrés en fonction de X et Y. On note $\widehat{Y} = X\,\widehat{\theta}$. On mesure la qualité de la prédiction par cet estimateur avec le risque quadratique $R(\widehat{Y}) = \mathbb{E}\big(\|\widehat{Y} - X\,\theta\|^2\big)$, où $\|\cdot\|$ désigne la norme euclidienne classique. Déterminer $R(\widehat{Y})$ (en justifiant).
3. A partir du modèle (3), on veut tester l'hypothèse $H_0 : \theta_i = 0$ pour tout $i = p - p_0, \cdots, p$, où $p_0 \in \mathbb{N}^*$, contre l'hypothèse H_1, son complément. On note $\widehat{\sigma}^2 = \frac{1}{n-(p+1)} \|Y - \widehat{Y}\|^2$. Déterminer sous H_0 la loi de $\widehat{\sigma}^2$.
4. On note X^0 la matrice extraite de X contenant uniquement ses $p - p_0 + 1$ premières colonnes et $\widehat{Y}^0 = X^0\,\widehat{\theta}^0$, où $\widehat{\theta}^0$ est obtenu par régression par moindres carrés sur les $p - p_0$ premières variables. On définit :

$$\widehat{F} = \frac{\frac{1}{p_0}\|\widehat{Y} - \widehat{Y}^0\|^2}{\widehat{\sigma}^2}.$$

Montrer que sous H_0, $\|\widehat{Y} - \widehat{Y}^0\|^2 = \|P_A\varepsilon\|^2$ où A est un sous-espace vectoriel de $\mathbb{N}^n$ de dimension p_0 que l'on précisera et P_A est la matrice de la projection orthogonale sur A.

En déduire la loi du numérateur de $\widehat{F}$.

Montrer que $\widehat{F}$ suit une loi de Fisher à $(p_0, n-p-1)$ degrés de liberté. Quelle règle de décision s'en déduit pour décider de H_0 avec un risque de première espèce $\alpha \in]0,1[$?

5. On suppose jusqu'à la fin du problème que $\theta_i = 0$ pour tout $i = p - p_0, \cdots, p$. Déterminer alors $R(\widehat{Y})$ et $R(\widehat{Y}^0)$. Quel estimateur vaut-il mieux choisir entre $\widehat{\theta}$ et $\widehat{\theta}^0$?
6. Pour estimer σ^2, on utilise les estimateurs par moindres carrés non biaisés $\widehat{\sigma}^2$ et $\widehat{\sigma}_0^2$ construits respectivement à partir de $\widehat{\theta}$ et $\widehat{\theta}^0$.

   Déterminer en justifiant la loi de $\widehat{\sigma}_0^2$.

   Montrer que pour Z une variable de loi $\mathcal{N}(0,1)$, $\mathrm{Var}(Z) = 2$.

   Déterminer alors les risques quadratiques de $\widehat{\sigma}^2$ et $\widehat{\sigma}_0^2$, soit $\mathbb{E}\big[(\widehat{\sigma}^2 - \sigma^2)^2\big]$ et $\mathbb{E}\big[(\widehat{\sigma}_0^2 - \sigma^2)^2\big]$.

   Quel estimateur de σ^2 vaut-il mieux choisir entre les deux ?
7. Pour A et B deux sous-espaces vectoriels de $\mathbb{N}^n$ tels que $A \subset B$, montrer que pour tout $x \in \mathbb{N}^n$, $\|P_A x\|^2 \leq \|P_B x\|^2$. On note $\widehat{R}^2$ et $\widehat{R}_0^2$ les coefficients de détermination R^2 respectifs pour les modèles avec $\widehat{\theta}$ et avec $\widehat{\theta}^0$. Montrer que $\widehat{R}^2 \geq \widehat{R}_0^2$ presque sûrement. Par rapport à ce critère, quel estimateur choisiriez-vous ?