Problème – Minimisation de l'erreur de prédiction

Dans ce problème, on s'intéresse à l'erreur de prédiction dans le modèle linéaire. On travaille ici avec un modèle linéaire, non supposé gaussien, et on note σ^2 la variance (constante) du bruit.

Première partie : cas à une variable explicative. On cherche à apprendre le modèle linéaire suivant pour prédire le réel Y_i en fonction d'une variable explicative réelle z_i :

$$\forall 1 \leq i \leq n, \quad Y_i = \beta_0 + \beta_1 z_i + \varepsilon_i \,.$$

On notera $\overline{y}$ et $\overline{z}$ les moyennes empiriques de $z = (z_1, \ldots, z_n)^T$ et de $Y = (Y_1, \ldots, Y_n)^T$ (attention, elles dépendent de n). On suppose que le modèle est identifiable et on note $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ l'estimateur des moindres carrés de β.

Sous le même modèle, on observe une nouvelle valeur z_{n+1} de la variable explicative et on cherche à prédire la variable réponse Y_{n+1} avec l'estimateur

$$\hat{y}_{n+1} := \hat{\beta}_0 + \hat{\beta}_1 z_{n+1} = x_{n+1}^T \hat{\beta}, \quad \text{où } x_{n+1} := \begin{bmatrix} 1 \\ z_{n+1} \end{bmatrix} .$$

L'erreur de prédiction est définie par :

$$\mathrm{err}(z_{n+1}) := \mathbb{E}\left[(Y_{n+1} - \hat{y}_{n+1})^2\right] .$$

Notons que dans cette espérance, l'aléa vient du bruit ε_{n+1} dans Y_{n+1} ainsi que de l'aléa dans le $\hat{\beta}$.

1. Redémontrer que si S est un vecteur aléatoire de $\mathbb{R}^m$ de matrice de covariance (finie) C, alors pour tout vecteur $u \in \mathbb{R}^m$, $\mathrm{Var}(u^T S) = u^T C u$. *Solution. On peut développer pour voir que*

$$\begin{aligned} Var(u^T S) &= Var\left(\sum_{i=1}^m u_i S_i\right) \\ &= \sum_{1 \leq i,j \leq m} Cov(u_i S_i, u_j S_j) \\ &= \sum_{1 \leq i,j \leq m} u_i u_i C_{i,j} = u^T C u \,. \end{aligned}$$

2. Que vaut $\mathbb{E}\left[Y_{n+1} - \hat{y}_{n+1}\right]$? *Solution. Elle est égale à $\mathbb{E}\left[x_{n+1}^T(\beta - \hat{\beta}) + \varepsilon_{n+1}\right] = x_{n+1}^T 0 + 0 = 0$ par nullité du biais de $\hat{\beta}$ et d'après les hypothèses sur les ε_i.*

3. Ecrire la matrice X du plan d'expérience dans ce modèle (on mettra l'intercept dans la première colonne), et montrer que

$$(X^T X)^{-1} = \frac{1}{n v(z)} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n z_i^2 & -\overline{z} \\ -\overline{z} & 1 \end{bmatrix},$$

avec $v(z)$ la variance empirique de z. *Solution. Dans notre cas on a $X = \left[\begin{array}{c|c} \mathbf{1}_n & z \end{array}\right] \in \mathbb{R}^{n \times 2}$, et $X^T X = n \begin{bmatrix} 1 & \overline{z} \\ \overline{z} & \frac{1}{n}\sum_{i=1}^n z_i^2 \end{bmatrix}$ de déterminant $n v(z)$. Il vient donc $(X^T X)^{-1} = \frac{1}{n v(z)} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n z_i^2 & -\overline{z} \\ -\overline{z} & 1 \end{bmatrix}$.*

4. En utilisant les résultats des questions 1, 2 et 3, montrer que

$$\mathrm{err}(z_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(z_{n+1} - \overline{z})^2}{\sum_{i=1}^n (z_i - \overline{z})^2}\right).$$

Solution. Comme la variable dont on prend le moment d'ordre deux est centré d'après la question 2, on calcule en fait une variance.

$$\begin{aligned}\mathrm{err}(z_{n+1}) &= \mathrm{Var}(\varepsilon_{n+1} + x_{n+1}^T(\beta - \hat{\beta})) \\ &= \sigma^2 + x_{n+1}^T(\sigma^2 (X^T X)^{-1}) x_{n+1}\,.\end{aligned}$$

D'après la question 3 on a $(X^T X)^{-1} = \frac{1}{nv(z)} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n z_i^2 & -\overline{z} \\ -\overline{z} & 1 \end{bmatrix}$ *avec* $v(z)$ *variance empirique de* z*. Cela donne*

$$\begin{aligned}\mathrm{err}(z_{n+1}) &= \sigma^2 + x_{n+1}^T(\sigma^2 (X^T X)^{-1}) x_{n+1} \\ &= \sigma^2 + \frac{\sigma^2}{nv(z)} \left(\frac{1}{n}\sum_{i=1}^n z_i^2 - 2\overline{z} z_{n+1} + z_{n+1}^2\right) \\ &= \sigma^2 + \frac{\sigma^2}{nv(z)} \left(v(z) + \overline{z}^2 - 2\overline{z} z_{n+1} + z_{n+1}^2\right) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(z_{n+1} - \overline{z})^2}{nv(z)}\right).\end{aligned}$$

5. Pour quelle(s) valeur(s) de z_{n+1} l'erreur de prédiction est-elle minimale ? Interpréter ce résultat. *Solution. L'erreur est minimale en* $z_{n+1} = \overline{z}$ *et vaut* $\sigma^2 + \sigma^2/n$*. On est toujours meilleur pour prédire lorsqu'on observe le barycentre de ce qu'on a déjà observé. En effet, intuitivement, c'est le point le plus typique.*

6. Quelle est la limite de l'erreur de prédiction minimale lorsque $n \to \infty$? Interpréter cette valeur limite. *Solution. Cette erreur tend vers* σ^2 *lorsque* n *est grand, c'est l'erreur minimale de prédiction qu'on puisse faire, car la variable* Y_{n+1} *est bruitée avec un bruit indépendant et de variance* σ^2 *: on ne peut pas faire mieux.*

Deuxième partie : cas général. Nous considérons cette fois-ci le cas général à plusieurs covaraibles. Ici, pour tout $1 \leq i \leq n$,

$$Y_i = \beta_0 + \beta_1 z_{1,i} + \ldots + \beta_p z_{p,i} + \varepsilon_i\,.$$

On note

$$Z := \left[\begin{array}{c|c|c|c} & & & \\ z_1 & \ldots & z_p & \\ & & & \end{array}\right] \in \mathbb{R}^{n \times p},$$

avec pour tout $1 \leq \ell \leq p$, $z_\ell := \begin{bmatrix} z_{\ell,1} & \ldots & z_{\ell,n} \end{bmatrix}^T \in \mathbb{R}^n$. On note $\overline{z}$ le vecteur des moyennes empiriques $\overline{z} := \begin{bmatrix} \overline{z_1} & \ldots & \overline{z_p} \end{bmatrix}^T \in \mathbb{R}^p$ (dont les entrées dépendent de n).

7. En notant $\beta = \begin{bmatrix} \beta_0 & \beta_1 & \ldots & \beta_p \end{bmatrix}^T$, écrire le modèle matriciellement sous la forme $Y = X\beta + \varepsilon$ en précisant X. *Solution. On l'a déjà fait,* X *s'écrit*

$$X = \left[\begin{array}{c|c|c|c} & & & \\ \mathbf{1}_n & z_1 & \ldots & z_p \\ & & & \end{array}\right] \in \mathbb{R}^{n \times (p+1)},$$

8. Ecrire la matrice X^TX sous forme de 4 blocs faisant intervenir Z, $\overline{z}$ et n. *Solution. Déjà fait en séance d'exercices. On a que* $X^TX = \begin{bmatrix} n & n\overline{z}^T \\ n\overline{z} & Z^TZ \end{bmatrix}$

Dans toute la suite, on suppose que le modèle est identifiable.

9. On donne la formule d'inversion matricielle par blocs : soit A une matrice inversible s'écrivant par blocs $A = \begin{bmatrix} T & U \\ V & W \end{bmatrix}$ avec T inversible. Alors $Q = W - VT^{-1}U$ est inversible et l'inverse de A est :

$$A^{-1} = \begin{bmatrix} T^{-1} + T^{-1}UQ^{-1}VT^{-1} & -T^{-1}UQ^{-1} \\ -Q^{-1}VT^{-1} & Q^{-1} \end{bmatrix}.$$

On note $\Gamma := \frac{1}{n}Z^TZ - \overline{z}\,\overline{z}^T$. Ecrire la matrice $(X^TX)^{-1}$ sous la forme d'une matrice par blocs en fonction de $n, \overline{z}$ et Γ^{-1}. *Solution. La formule donne directement* $Q = n\Gamma$ *et*

$$(X^TX)^{-1} = \frac{1}{n}\begin{bmatrix} 1 + \overline{z}^T\Gamma^{-1}\overline{z} & -(\Gamma^{-1}\overline{z})^T \\ -\Gamma^{-1}\overline{z} & \Gamma^{-1} \end{bmatrix}.$$

10. Comme dans la première partie, on observe désormais un vecteur

$$x_{n+1} = (1, z_{n+1}) = (1, z_{1,n+1}, \dots z_{\ell,n+1})$$

et l'on cherche à prédire la variable réponse Y_{n+1} avec l'estimateur $\hat{y}_{n+1} := x_{n+1}^T\hat{\beta}$. L'erreur de prédiction est définie comme précédemment.

Exprimer $\mathrm{err}(z_{n+1})$ en fonction de n, de σ^2, du vecteur $(z_{n+1} - \overline{z})$ et de la matrice Γ^{-1}. *Solution. On reprend les calculs de la question 3., et on a encore*

$$\begin{aligned} \mathrm{err}(z_{n+1}) &= \sigma^2 + x_{n+1}^T(\sigma^2(X^TX)^{-1})x_{n+1} \\ &= \sigma^2 + \frac{\sigma^2}{n}\left(1 + z^T\Gamma^{-1}z - 2z_{n+1}^T\Gamma^{-1}\overline{z} + z_{n+1}^T\Gamma^{-1}z_{n+1}\right) \\ &= \sigma^2\left(1 + \frac{1}{n} + (z_{n+1} - \overline{z})^T\Gamma^{-1}(z_{n+1} - \overline{z})\right). \end{aligned}$$

11. On admet dans cette question que Γ est symétrique définie positive. Pour quelle(s) valeur(s) de z_{n+1} l'erreur de prédiction est-elle minimale ? *Solution. Ben du coup c'est évident, il faut dire que* Γ^{-1} *est encore symétrique définie positive, et le minimum est donc atteint en* $z_{n+1} = \overline{z}$. *Le reste du temps, c'est au-dessus !*

12. ($\star$) Montrer que $\Gamma = \frac{1}{n}Z^TZ - \overline{z}\,\overline{z}^T$ est bien symétrique définie positive. *On pourra chercher à l'écrire sous la forme* $\frac{1}{n}WW^T$ *avec* W *une matrice bien choisie. Solution. Symétrie évidente. Ensuite, l'idée c'est de voir qu'elle ressemble fortement à une matrice de covariance. C'est même carrément une matrice de covariance empirique. Si on note* $\tilde{Z}$ *la matrice*

$$\tilde{Z} := \left[\begin{array}{c|c|c} & & \\ z_1 - \overline{z_1}\mathbf{1}_n & \dots & z_p - \overline{z_p}\mathbf{1}_n \\ & & \end{array}\right] \in \mathbb{R}^{n\times p},$$

on a que $\Gamma = \frac{1}{n}\tilde{Z}^T\tilde{Z}$. *On a donc la positivité qui en découle. On peut ensuite évoquer le résultat d'inversion par blocs qui implque que* Γ *est inversible, ou remarquer que* $\tilde{Z}$ *est*

injective. En effet, si $u \in \mathrm{Ker}\tilde{Z}$*, alors* $u_1 z_1 + \ldots + u_p z_p = \left(\sum_{\ell=1}^{p} u_\ell \overline{z_\ell}\right) \mathbf{1}_n$*, et comme* X *est de rang plein par hypothèse,* $u = 0$*.*