TD 2 : Régression linéaire multiple

Exercice 1. Rappels de cours.

1. Rappeler le principe d'une régression linéaire multiple. Préciser les hypothèses.
2. Faire un schéma pour donner une interprétation géométrique à la régression linéaire multiple. Retrouvez l'expression de l'estimateur des moindres carrés $\hat{\beta}$.
3. Donner l'expression de la matrice de projection $\mathbf{P}^{\mathbf{X}}$ et de l'estimateur $\hat{\beta}$. Vérifier que $\mathbf{P}^{\mathbf{X}}$ est bien une matrice de projection.
4. Quelles sont les hypothèses supplémentaires dans le cas gaussien ?
5. Dans le cas Gaussien, retrouvez l'expression des estimateurs du maximum de vraisemblance pour β et σ^2 en anulant le gradient de la fonction à maximiser.
6. Dans le cas Gaussien, retrouvez la loi de $\hat{\beta}$ et $\hat{\sigma}^2$, à variance connue ou inconnue.

On conseille de toujours faire attention à la dimension des objets (matrices et vecteurs) qu'on manipule.

Exercice 2. Régression simple vs régression multiple.

1. Rappeler les expressions de $\hat{\beta}_0$ et $\hat{\beta}_1$ dans le cas d'une régression simple.
2. Rappeler l'expression de $\hat{\beta}$ dans le cas d'une régression multiple.
3. Retrouver le résultat de la question 1 à partir de celui de la question 2.
4. Rappeler les expressions des variances et covariance de $\hat{\beta}_0$ et $\hat{\beta}_1$ pour une régression simple.
5. Rappeler l'expression de la matrice de variance-covariance de $\hat{\beta}$ pour une régression multiple.
6. Retrouver le résultat de la question 4 à partir de celui de la question 5.

Exercice 3. Régression à deux variables. On étudie l'évolution d'une variable y en fonction de deux variables x et z. On dispose de n observations de ces variables. On note $X = \begin{pmatrix}\mathbb{1} & x & z\end{pmatrix}$, où $\mathbb{1}$ est le vecteur constant, et x et z sont les vecteurs des variables explicatives. Nous avons obtenu les résultats suivants:

$$X^T X = \begin{pmatrix} 30 & 0 & 0 \\ ? & 10 & 7 \\ ? & ? & 15 \end{pmatrix} \quad , \quad \|\hat{\varepsilon}\|^2 = 12 \quad , \quad \hat{\beta} = \begin{pmatrix} -2 \\ 1 \\ 2 \end{pmatrix} .$$

1. (a) Donner les valeurs manquantes. Que vaut n ?
   (b) Calculer le coefficient de corrélation empirique entre x et z.
2. (a) Calculer $\sum_{i=1}^n \hat{\varepsilon}_i$, puis en déduire la valeur de la moyenne arithmétique $\bar{y}$.
   (b) Calculer la somme des carrés résiduels (SCR), la somme des carrés expliquée (SCE), la somme des carrés totale (SCT) et le coefficient de détermination R^2.
3. (a) Calculer $X^T y$ en utilisant la valeur de $\hat{\beta}$, puis en déduire $\sum x_i y_i$ et $\sum z_i y_i$.
   (b) Calculer les coefficients de corrélation $\rho_{x,y}$ et $\rho_{z,y}$. En déduire la valeur du R^2 pour le modèle de régression de y par $\mathbb{1}$ et x, puis de y par $\mathbb{1}$ et z.
4. (a) Sous l'hypothèse gaussienne, donnez la loi de $\hat{\beta}_2$ le coefficient associé à x en fonction de β_2 et σ^2.
   (b) Calculer $(X^T X)^{-1}$.
   (c) Calculez un estimateur sans biais de la variance σ^2.
   (d) Proposez un intervalle de confiance à 95% pour β_2. Que peut-on conclure quant à la nullité de ce coefficient ? On donne le quantile à 97.5% de la loi de Student à 27 degrés de libertés : $t_{27}(0.975) = 2.05$.

Exercice 4. Interprétation géométrique.

1. Nous avons une variable y à expliquer par une variable x. Nous avons effectué $n = 2$ mesures et trouvé
$$(x_1, y_1) = (4, 5) \text{ et } (x_2, y_2) = (1, 5)$$
Représenter les variables, estimer β dans le modèle $y_i = \beta x_i + \varepsilon_i$ et représenter $\hat{y}$.
2. Nous avons maintenant une variable y à expliquer par deux variables x_1 et x_2. Nous avons effectué $n = 3$ mesures et trouvé
$$(x_{1,1}, x_{1,2}, y_1) = (3, 2, 0), \quad (x_{2,1}, x_{2,2}, y_2) = (3, 3, 5), \quad (x_{3,1}, x_{3,2}, y_3) = (0, 0, 3)\,.$$
Représenter les variables, estimer β dans le modèle $y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$ et représenter $\hat{y}$.

Exercice 5. Croissance de R^2. Soit X une matrice de taille $n \times p$ composée de p vecteurs indépendants de $\mathbb{R}^n$. Nous notons X_q la matrice composée des $q < p$ premiers vecteurs de X. On suppose que la première colone de X est égale à $\mathbb{1}$, i.e. que l'intercept est inclu dans les deux modèles. Nous avons les deux modèles suivants :
$$(1) \quad Y = X\beta + \varepsilon \quad \text{et} \quad (2) \quad Y = X_q\beta_q + \varepsilon_q\,.$$
Comparer les R^2 dans les deux modèles.

Exercice 6. Régression sur données agrégées par groupes On suppose le modèle de régression
$$Y = X\beta + \varepsilon\,, \quad \text{avec} \quad \mathbb{E}[\varepsilon] = 0 \quad \text{et} \quad \mathrm{Var}(\varepsilon) = \sigma^2 I_n\,.$$
Les données individuelles $(x_{i1}, \ldots, x_{ip}, y_i)$ ne sont cependant pas disponibles. On observe seulement les moyennes sur I groupes, notés $C_1, \ldots, C_I$, d'effectifs $n_1, \ldots, n_I$:
$$\bar{y}_k = \frac{1}{n_k}\sum_{i \in C_k} y_i \quad \text{et} \quad \bar{x}_{kj} = \frac{1}{n_k}\sum_{i \in C_k} x_{ij}\,.$$
En notant $\bar{\varepsilon}_k = \frac{1}{n_k}\sum_{i \in C_k} \varepsilon_i$, on a alors $\bar{Y} = \bar{X}\beta + \bar{\varepsilon}$.

1. Calculer $\mathbb{E}[\bar{\varepsilon}]$ et $\mathrm{Var}[\bar{\varepsilon}]$.
2. On pose
$$M = \mathrm{diag}(\sqrt{n_1}, \ldots, \sqrt{n_I})\,, \quad Y^* = M\bar{Y}\,, \quad X^* = M\bar{X}\,, \quad \varepsilon^* = M\bar{\varepsilon}\,.$$
Quelle est la relation entre Y^*, X^* et ε^* ? Calculer $\mathbb{E}[\varepsilon^*]$ et $\mathrm{Var}(\varepsilon^*)$.
3. En déduire un estimateur de β.
4. Application numérique : $I = 3$ avec $n_1 = 1$ et $n_2 = n_3 = 2$. $\bar{X}_1^T = (1, 1, 1), \bar{X}_2^T = (7, 12, 5)$ et $\bar{Y}^T = (15, 25, 10)$.

Exercice 7. Estimateurs linéaires Soit θ_1 et θ_2 deux paramètres réels inconnus et soit :
- Y_1 un estimateur sans biais de $\theta_1 + \theta_2$ et de variance σ^2
- Y_2 un estimateur sans biais de $2\theta_1 - \theta_2$ et de variance $4\sigma^2$
- Y_3 un estimateur sans biais de $6\theta_1 + 3\theta_2$ et de variance $9\sigma^2$

Les estimateurs Y_1, Y_2 et Y_3 étant indépendants, nous cherchons les estimateurs sans biais de θ_1 et θ_2, linéaires en Y_1, Y_2 et Y_3, et de variance minimale.

1. Notons $\tilde{\theta} = \alpha Y_1 + \beta Y_2 + \gamma Y_3$.
   (a) Quelles sont les équations à satisfaire pour que $\tilde{\theta}$ soit un estimateur sans biais de θ_1?

(b) Dans ce cas-là, exprimer la variance de $\tilde{\theta}$ et la minimiser.
(c) Idem pour θ_2.

2. On pose $Z_1 = Y_1$, $Z_2 = Y_2/2$, et $Z_3 = Y_3/3$, et on note $Z = (Z_1, Z_2, Z_3)^T$ et $\theta = (\theta_1, \theta_2)^T$.
   (a) Trouver la matrice X telle que $\mathbb{E}[Z] = X\theta$.
   (b) Que vaut la matrice de variance-covariance de Z ?
   (c) On peut alors écrire $Z = X\theta + \varepsilon$. Retrouver les estimateurs de θ_1 et θ_2 calculés question 1.

Exercice 8. Théorème de Gauss Markov L'objectif de cet exercice est de démontrer le théorème de Gauss-Markov. On se place dans le cadre du modèle de régression multivarié classique (sans hypothèse gaussienne). Soit $\hat{\boldsymbol{\beta}}$ l'estimateur des moindres carrés du vecteur de coefficient $\boldsymbol{\beta}$.

1. Rappellez les hypothèses du modèle de régression multiple, ainsi que la définition de l'estimateur des moindres carrés, et donnez son expression.
2. Montrez que l'estimateur des moindres carrés $\hat{\beta}$ est linéaire et sans biais. Retrouvez l'expression de sa matrice de variance-covariance.
3. Soit $\tilde{\boldsymbol{\beta}}$ un autre estimateur linéaire sans biais de $\boldsymbol{\beta}$.
   (a) Montrez qu'il existe une matrice $\mathbf{B}$ déterministe telle que $\tilde{\boldsymbol{\beta}} = \mathbf{By}$. Précisez les dimensions de $\mathbf{B}$.
   (b) Montrez que, pour tout vecteur $\boldsymbol{\beta}$ de coefficient, $\mathbf{BX}\boldsymbol{\beta} = \boldsymbol{\beta}$. En déduire que $\mathbf{BX} = \mathbf{I}_p$ où $\mathbf{I}_p$ est la matrice identité de taille p.
4. Soient $\mathbf{S}_1$ et $\mathbf{S}_2$ deux matrices symétriques réelles de taille $p \times p$. On dit que $\mathbf{S}_1 \leq \mathbf{S}_2$ si la matrice $\mathbf{S}_2 - \mathbf{S}_1$ est une matrice symétrique positive, i.e. si elle est symétrique, et, pour tout $\mathbf{u} \in \mathbb{R}^p$, $\mathbf{u}^T(\mathbf{S}_2 - \mathbf{S}_1)\mathbf{u} \geq 0$. On cherche à montrer que $\mathbb{V}[\hat{\boldsymbol{\beta}}] \leq \mathbb{V}[\tilde{\boldsymbol{\beta}}]$.
   (a) Montrez: $\mathbb{V}[\tilde{\boldsymbol{\beta}}] - \mathbb{V}[\hat{\boldsymbol{\beta}}] = \mathbb{V}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}] + \mathbb{C}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}] + \mathbb{C}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}]^T$.
   (b) Montrez: $\mathbb{C}[\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.
   (c) Montrez: $\mathbb{C}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}] = 0$.
   (d) En déduire que $\mathbb{V}[\tilde{\boldsymbol{\beta}}] - \mathbb{V}[\hat{\boldsymbol{\beta}}]$ est une matrice symétrique positive.
5. Conclure la démontration du théorème de Gauss-Markov.