Modéle de régression linéaire - Feuille 2

Régression linéaire simple

Exercice 1 Soit $(z_1, y_1), \cdots, (z_n, y_n)$ des couples de réels.

1. Déterminer le réel $\widehat{a}$ qui minimise la somme des carrés résiduelle $SCR(a) = \sum_{i=1}^{n} |y_i - a \cdot z_i|^2$.
2. Que représente $\widehat{a}$ par rapport à $y_1, \cdots, y_n$? Comparer avec la valeur $\widehat{\beta}$ obtenue dans le cadre de la régression linéaire simple classique.

Exercice 2 Rappeler la formule définissant le coefficient de détermination R^2 et la développer pour montrer qu'il est égal au carré du coefficient de corrélation empiriqueentre x et y, noté $\rho_{x,y}$, c'est-à-dire :

$$R^2 = \rho_{x,y}^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2 .$$

Exercice 3 On considère le modèle de régression linéaire simple $y = \beta_1 + \beta_2 x + \epsilon$. Soit un échantillon $(x_i, y_i)_{1 \leq i \leq 100}$ de statistiques résumées

$$\sum_{i=1}^{100} x_i = 0 \qquad \sum_{i=1}^{100} x_i^2 = 400 \qquad \sum_{i=1}^{100} x_i y_i = 100 \qquad \sum_{i=1}^{100} y_i = 100 \qquad \hat{\sigma}^2 = 1$$

1. Exprimer les intervalles de confiance à 95% pour β_1 et β_2.
2. Donner l'équation de la région de confiance à 95% de (β_1, β_2). Rappel : $\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} \leq 1$ est l'équation de l'intérieur de l'ellipse centré en (x_0, y_0), dont les axes sont parallèles à ceux des abscisses et des ordonnées, et de sommets $(x_0 \pm a, y_0)$ et $(x_0, y_0 \pm b)$.
3. Représenter sur un même graphique les résultats obtenus.

Exercice 4 On appelle "fréquence seuil" d'un sportif amateur sa fréquence cardiaque obtenue aprés trois quarts d'heure d'un effort soutenu de course à pied. Celle-ci est mesurée à l'aide d'un cardio-fréquence-mètre. On cherche à savoir si l'âge d'un sportif a une influence sur sa fréquence seuil. On dispose pour cela de 20 valeurs du couple (x_i, y_i), où x_i est l'âge et y_i la fréquence seuil du sportif. On a obtenu $(\bar{x}, \bar{y}) = (35,6; 170,2)$ et :

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 1991 \qquad \sum_{i=1}^{n} (y_i - \bar{y})^2 = 189,2 \qquad \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = -195,4$$

1. Calculer les estimateurs des moindres carrés pour le modèle $y = \beta_1 + \beta_2 x + \varepsilon$.
2. Question de cours : montrer que

$$\hat{\beta}_2 = \beta_2 + \frac{\sum (x_i - \overline{x}) \varepsilon_i}{\sum (x_i - \overline{x})^2}.$$

3. Calculer le coefficient de détermination R^2. Commenter la qualité de l'ajustement des données au modèle.
4. Avec ces estimateurs, la somme des carrés des résidus vaut $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = 170$. On suppose les perturbations ε_i gaussiennes, indépendantes et de même variance σ^2. Déduire un estimateur non biaisé $\hat{\sigma}^2$ de σ^2
5. Déduire un estimateur $\hat{\sigma}_2^2$ de la variance de $\hat{\beta}_2$.
6. Calculer l'intervalle de confiance au risque 5% de β_2.
7. Tester l'hypothèse $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$ pour un risque de 5%. Conclure sur la question de l'influence de l'âge sur la fréquence seuil.

Exercice 5 On souhaite expliquer la hauteur y (en mètres) d'un arbre en fonction de sa circonférence x (en centimètres) à 1m30 du sol. On a relevé $n = 1429$ couples (x_i, y_i). On a obtenu $(\bar{x}, \bar{y}) = (47,3; 21,2)$ et :

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 102924 \qquad \sum_{i=1}^n (y_i - \bar{y})^2 = 8857 \qquad \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 26466$$

1. Calculer la droite des moindres carrés pour le modèle $y = \beta_1 + \beta_2 x + \epsilon$.
2. Calculer le coefficient de détermination R^2. Commenter la qualité de l'ajustement des données au modèle.
3. Avec ces estimateurs, la somme des carrés des résidus vaut alors $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = 2052$. Si on suppose les perturbations ϵ_i gaussiennes, centrées, indépendantes et de même variance σ^2 , en déduire un estimateur non biaisé $\hat{\sigma}^2$ de σ^2.
4. Donner un estimateur $\hat{\sigma}_1^2$ de la variance de $\hat{\beta}_1$.
5. Tester l'hypothèse $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$.

Exercice 6 Considérons le modèle statistique suivant :

$$y_i = \beta x_i + \varepsilon_i, \qquad i = 1, \cdots, n,$$

où nous supposons que les perturbations ε_i sont telles que $\mathbb{E}[\varepsilon_i] = 0$ et $\mathrm{Cov}(\varepsilon_i, \varepsilon_i) = \sigma^2 \delta_{i,j}$.

1. En revenant à la définition des moindres carrés, montrer que l'estimateur des moindres carrés de β vaut $\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$.
2. Montrer que la droite passant par l'origine et le centre de gravité du nuage de points est $y = \beta^* x$, avec $\beta^* = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$.
3. Montrer que $\hat{\beta}$ et β^* sont des estimateurs sans biais de β.
4. Montrer que $V(\beta^*) > V(\hat{\beta})$ sauf dans le cas où tous les x_i sont égaux (penser à l'inégalité de Cauchy-Schwarz). Ce résultat était-il prévisible ?

Exercice 7 Au $17^{\text{ème}}$ siècle, Huygens s'est intéressé aux forces de résistance d'un objet en mouvement dans un fluide (eau, air, etc.). Il a d'abord émis l'hypothèse selon laquelle les forces de frottement étaient proportionnelles à la vitesse de l'objet, puis, après expérimentation, selon laquelle elles étaient proportionnelles au carré de la vitesse. On réalise une expérience dans laquelle on fait varier la vitesse x d'un objet et on mesure les forces de frottement y.

1. Quel(s) modèle(s) testeriez-vous ?
2. Comment feriez-vous pour déterminer le modèle adapté ?

Exercice 8 Comparaison entre EMCO et EMV.

On rappelle que, dans le cadre d'un modèle statistique paramétrique (c'est-à-dire que la loi du modèle ne dépend que d'un nombre fini de paramètres), la densité des observations $Y_1, \dots, Y_n$ vue comme une fonction des paramètres est appelé la vraisemblance. L'estimateur du maximum de vraisemblance de ces paramètres est la valeur des paramètres qui maximise la vraisemblance. Considérons le modèle gaussien

$$Y_i = \mu + \beta\, X_i + \varepsilon_i, \quad i = 1, \dots, n \tag{1}$$

et ses paramètres (μ, β, σ^2), avec les ε_i des v.a.i.i.d. gaussiennes de loi $\mathcal{N}(0, \sigma^2)$.

1. Montrer que $-2\times$ log-vraisemblance vaut

$$L(\mu, \beta, \sigma^2) = n \cdot log(2\pi) + n \cdot log(z) + \frac{1}{z} \cdot \sum_{i=1}^{n} (Y_i - \mu - \beta X_i)^2,$$

   en notant $z = \sigma^2$ pour dériver plus facilement.
2. Comparer les estimateurs du maximum de vraisemblance avec ceux des moindres carrés de μ et β.
3. Comparer (au sens de la vitesse de convergence) l'estimateur du maximum de vraisemblance de σ^2 avec $\widehat{\sigma}^2$ défini dans le corps du chapitre.

Exercice 9 Octopus's Garden (et lien régression simple et multiple)

On cherche à mettre en oeuvre une stratégie de prédiction du poids utile du poulpe, c'est-à-dire son poids éviscéré, à partir de son poids non éviscéré. C'est en effet le poulpe éviscéré qui est commercialisé. Pour cela, un échantillon de poulpes a été collecté en 2003 lors des opérations de pêche dans les eaux mauritaniennes. Vu l'importante différence de poids entre les poulpes mâles et les poulpes femelles, on étudie ici uniquement les données concernant 240 poulpes femelles.

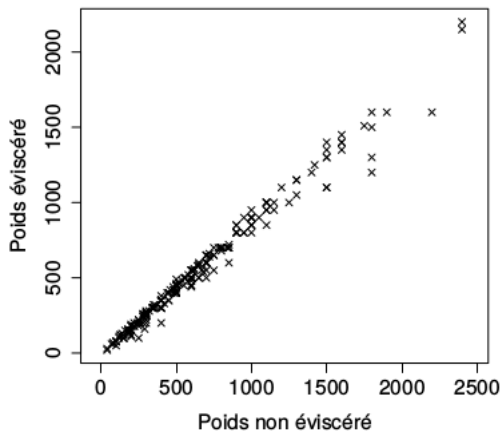

Figure 1 – Poids de poulpe éviscéré en fonction du poids non éviscéré (en grammes).

```
                        Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)            -2.312146     5.959670    -0.388    0.698
Poids non éviscéré      0.853169     0.007649   111.545   <2e-16

Residual standard error: 52.73 on 238 degrees of freedom
Multiple R-Squared: 0.9812,     Adjusted R-squared: 0.9812
F-statistic: 1.244e+04 on 1 and 238 DF,  p-value: < 2.2e-16
```

TABLE A.2 – Poids de poulpes éviscérés et non éviscérés : résultats de la régression linéaire simple (sortie R).

```
                       Estimate Std. Error t value Pr(>|t|)
Poids non éviscéré      0.85073    0.00436   195.1   <2e-16

Residual standard error: 52.63 on 239 degrees of freedom
Multiple R-Squared: 0.9938,     Adjusted R-squared: 0.9937
F-statistic: 3.807e+04 on 1 and 239 DF,  p-value: < 2.2e-16
```

TABLE A.3 – Poids de poulpes éviscérés et non éviscérés : résultats de la régression linéaire simple avec le modèle simplifié (sortie R).

1. Les données sont représentées Figure 1.
   (a) Proposer un modèle reliant le poids éviscéré et le poids non éviscéré dun poulpe.
   (b) Rappeler les formules des estimateurs des paramètres du modèle.
   (c) A partir de la table A.2 donner des estimations umériques des paramètres du modèle.
   (d) Que représente la valeur 0.698 ? Au vu de cette valeur proposer un autre modéle reliant les poids éviscéré et non éviscéré.
2. Plus généralement, considérons un échantillon de n couples de réels (z_i, y_i) suivant le modèle : $y_i = \beta z_i + \varepsilon_i$ où les erreurs ε_i sont supposées gaussiennes indépendantes centrées et de même variance σ^2.
   (a) Déterminer l'estimateur $\widetilde{\beta}$ de β minimisant la somme des carrés des écarts au modèle.
   (b) Retrouver le résultat précédent à partir de la formule générale de l'estimateur de régression linéaire multiple.
   (c) En déduire la variance de $\widetilde{\beta}$. Proposer un estimateur sans biais $\widetilde{\sigma}^2$ de σ^2.
   (d) Les résultats de l'analyse de ce nouveau modèle sont donnés table A.3. Donner $\widetilde{\beta}$ et $\widetilde{\sigma}^2$.