Série 13

Exercice 1. Dans le cadre de la régression linéaire, montrer les propriétés énoncées en cours :

(i). La droite de moindres carrés passe par $(\overline{x}_n, \overline{y}_n)$.

(ii). $\sum_{j=1}^n r_j = 0$.

(iii). $\sum_{j=1}^n x_j\, r_j = \sum_{j=1}^n x_j(y_j - \widehat{y}_j) = 0$.

(iv). $\sum_{j=1}^n \widehat{y}_j\, r_j = 0$.

Exercice 2. On cherche à comprendre la relation entre la moyenne $\bar{y}$ d'un examen et celle $\bar{x}$ d'un test bonus. On sait que leur coefficient de corrélation empirique vaut $\hat{r} = 0.6$. En utilisant les données suivantes :

$$\bar{y} \;=\; 55, \quad \sum_{i=1}^n (y_i - \bar{y})^2 = 20, \quad \bar{x} = 70, \quad \sum_{i=1}^n (x_i - \bar{x})^2 = 10,$$

trouver l'équation de la droite de régression $x \to \hat{a}_n + \hat{\beta}_n \bar{x}$ de la moyenne à l'examen en fonction de celle au test bonus. Quel est le lien entre $\widehat{\beta}_n$ et $\widehat{r}$?

Exercice 3. Au problème de la régression simple classique peut s'ajouter l'obligation de devoir passer par un point (x_0, y_0) spécifique. Ce type de régression est dite régression forcée. On considère un ensemble de points (x_i, y_i) pour $i = 1, \ldots, n$, et le modèle $y_i = \beta(x_i - x_0) + y_0 + \eta_i$ où les η_i sont des Gaussiennes indépendantes de moyenne 0 et de variance 1.

a) Trouver la valeur $\hat{\beta}$ de β qui minimise l'erreur de prédiction $\sum_i \{y_i - y_0 - \beta(x_i - x_0)\}^2$. On va utiliser cette valeur comme estimation (réalisation de l'estimateur) de β, la pente de la droite de régression forcée par un point (x_0, y_0).

b) Appliquer le résultat de la question a) au cas $(x_0, y_0) = (\bar{x}, \bar{y})$ et comparer avec l'estimateur de la pente de la droite de régression simple classique.

c) Déterminer le modèle de régression forcée par le point de coordonnées $(3, 6)$ pour les données ci-dessous :

x_i	1	2	3	4	5	6	7
y_i	2.5	5.0	3.5	7	6.5	8.25	8

Exercice 4. Un échantillon aléatoire de 10 hommes âgés de 30 ans a été choisi. On a relevé les données leur salaire annuel (Y) et le nombre d'années d'école qu'ils ont suivies (X). On a trouvé les valeurs suivantes : $\Sigma\, x = 151 \quad \Sigma\, y = 96.3 \quad \Sigma\, xy = 1526.30 \quad \Sigma\, x^2 = 2357 \quad \Sigma\, y^2 = 1036.13$.

En supposant que la moyenne de la loi des salaires est une fonction linéaire du nombre d'années d'école :

a) Déterminer les estimateurs des paramètres a et β du modèle $y = a + \beta\, X + \epsilon$, ainsi que l'estimateur de σ^2 (variance de ϵ).

b) Tester l'hypothèse $H_0 : \beta = 0$ contre l'alternative $H_1 : \beta \neq 0$ au seuil de signification $\alpha = 0.01$.

Exercice 5. Transformations des données. Pour certains jeux de données, il peut être judicieux de calculer une transformation des données avant d'effectuer une régression linéaire.

Les données suivantes montrent les valeurs expérimentales de la pression P d'une masse donnée de gaz pour différentes valeurs du volume V.

Volume V (cm^3)	54.3	61.8	72.4	88.7	118.6	194.0
Pression P (kg/cm^2)	61.2	49.5	37.6	28.4	19.2	10.1

D'après les principes de la thermodynamique, on a la relation $PV^\gamma = C$, où γ et C sont des constantes dépendant des conditions d'expérience.

a) Transformer $PV^\gamma = C$ en modèle linéaire avec $y = \log(P)$. Quels sont les paramètres dont on cherche la valeur ?

b) En utilisant a), trouver les estimateurs de γ et $\log(C)$.

c) Estimer P quand $V = 100$ cm^3.

d) Calculer l'intervalle de confiance à 95% pour la pente de la droite du modèle de régression trouvé en a).

Exercice 6. Nous souhaitons exprimer la hauteur y d'un arbre en fonction de son diamètre x à 1m30 du sol. Pour ce faire, nous avons mesuré 20 couples "diamètre-hauteur", avec

$$\overline{x}_n = 34.9, \quad \overline{y}_n = 18.34, \quad \frac{1}{20}\sum_{i=1}^{20}(x_i - \overline{x}_n)^2 = 28.29, \quad \frac{1}{20}\sum_{i=1}^{20}(y_i - \overline{y}_n)^2 = 2.85, \quad \frac{1}{20}\sum_{i=1}^{20}(x_i - \overline{x}_n)(y_i - \overline{y}_n) = 6.26,$$

$$\min(x_1, \dots, x_n) = 20, \qquad \max(x_1, \dots, x_n) = 50.$$

(i). Soit $\hat{y} = \hat{a}_n + \widehat{\beta}_n x$, la droite des moindres carrés. Donner l'expression de $\widehat{\beta}_n$ en fonction des statistiques élémentaires ci-dessus. Calculer $\hat{a}_n$ et $\widehat{\beta}_n$.

(ii). Donner une mesure de la qualité de l'ajustement des données au modèle, et commenter la valeur obtenue.

(iii). Les estimations des variances de $\widehat{a}_n$ et de $\widehat{\beta}_n$ sont

$$\widehat{\mathrm{var}}(\widehat{a}_n) = 1.89^2, \quad \widehat{\mathrm{var}}(\widehat{\beta}_n) = 0.05^2.$$

Tester les deux hypothèses $H_0 : \beta = 0$ et $H_0 : a = 0$ contre $H_1 : \beta \neq 0$ et $H_1 : a \neq 0$ respectivement. Pourquoi ces tests sont-ils intéressants dans notre contexte ? Que pensez-vous du résultat ?

Exercice 7. Les données suivantes représentent le niveau maximal de la mer (en cm) à Venise mesuré chaque année entre 1887 et 2019 (on a donc $n = 132$ observations) :

année	année −1900	niveau (cm)
1887	−13	94
1888	−12	90
1889	−11	97
1890	−10	107
1891	−9	87
1892	−8	86
⋮	⋮	⋮

On va ajuster un modèle de régression,

$$y = a + \beta x + \epsilon, \tag{1}$$

où x représente l'année et ϵ est une variable aléatoire indépendante et identiquement distribuée selon une loi normale de moyenne 0 et de variance σ^2. Pour cela, on exécute le code `R` suivant

```
x=venice$year
y=venice$y

plot(x,y,xlab='Ann\'ee',ylab='niveau de la mer (cm)')
M=lm(y~x)
summary(M)
```

qui nous permet de créer la figure 1 et d'obtenir les estimations présentées dans le tableau 1. Nous considérons aussi un modèle modifié

$$y = \beta_0 + \beta_1(x - 1900) \tag{2}$$

dont les estimations associées sont représentées dans le tableau 2.

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	-585.4931	75.2428	-7.78	≈ 0
year	0.3589	0.0385	9.32	≈ 0

Table 1 – Résultats associés au modéle 1

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	96.4510	2.5209	38.26	≈ 0
year-1900	0.3589	0.0385	9.32	≈ 0

Table 2 – Résultats associés au modéle 2

(i). Quelles sont les estimations de a et β ? Quelle est l'interprétation de a ?

(ii). Quelles sont les estimations de β_0 et β_1 selon le modèle 2 ? Quelle est leur interprétation ? Quel est le lien entre (a, β) et (β_0, β_1) et quel est le lien entre leurs estimateurs ? Quelle représentation est préférable ?

(iii). Donner des intervalles de confiance à 95% pour β et pour β_1. Que peut-on dire sur l'effet du temps sur le niveau de la mer à Venise ?

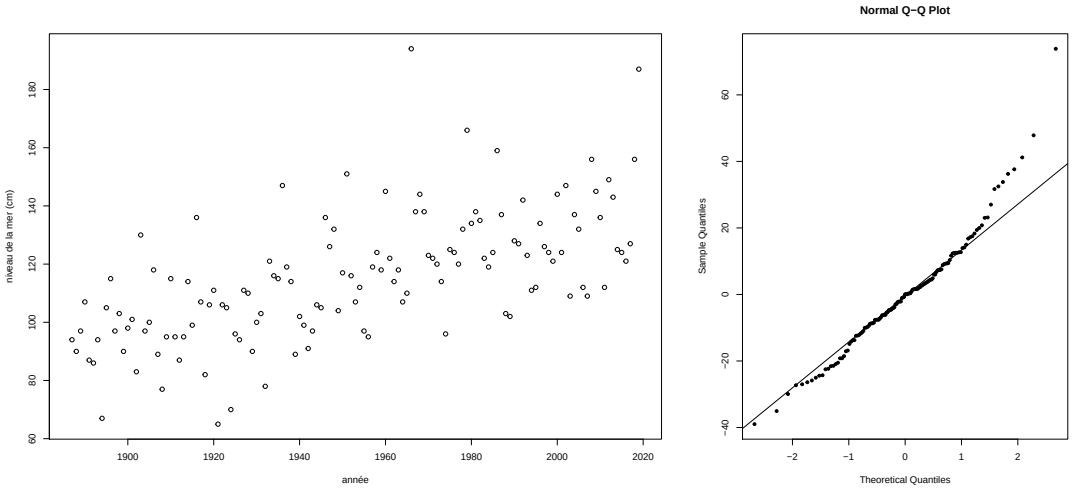

Figure 1 – Données sur les marées maximales annuelles de 1881 à 2019 à la station hydrographique de Santa Maria della Salute, à Venise : (à gauche) les données ; (à droite) qq-plot des résidus suite à une régression linéaire.