Corrigé 13

Exercice 1. (i). La droite est $y(x) = \widehat{a}_n + \widehat{\beta}_n x$. En mettant $x = \overline{x}_n$ on trouve

$$y(\overline{x}_n) = \widehat{a}_n + \widehat{\beta}_n \overline{x}_n = (\overline{y}_n - \widehat{\beta}_n \overline{x}_n) + \widehat{\beta}_n \overline{x}_n = \overline{y}_n.$$

(ii). On a

$$\sum_{j=1}^n r_j = \sum_{j=1}^n (y_j - \widehat{y}_j) = n\overline{y}_n - n\widehat{a}_n - \widehat{\beta}_n \sum_{j=1}^n x_j = n\overline{y}_n - n(\overline{y}_n - \widehat{\beta}_n \overline{x}_n) - \widehat{\beta}_n n\overline{x}_n = 0.$$

(iii). On a

$$\sum_{j=1}^n x_j r_j = \sum_{j=1}^n x_j (y_j - \widehat{a}_n - \widehat{\beta}_n x_j) = \sum_{j=1}^n x_j y_j - n\overline{x}_n(\overline{y}_n - \widehat{\beta}_n \overline{x}_n) - \widehat{\beta}_n \sum x_j^2$$

$$= \sum_{j=1}^n x_j y_j - n\overline{x}_n \overline{y}_n - \widehat{\beta}_n (\sum_{j=1}^n x_j^2 - n\overline{x}_n^2) = 0,$$

par la formule de $\widehat{\beta}_n$.

(iv). On a

$$\sum_{j=1}^n \overline{y}_j r_j = \sum_{j=1}^n r_j (\widehat{a}_n + \widehat{\beta}_n x_j) = \widehat{a}_n \sum_{j=1}^n r_j + \widehat{\beta}_n \sum_{j=1}^n r_j x_j = 0 + 0,$$

d'après les parties 2. et 3.

Exercice 2. La corrélation entre une variable aléatoire X et une variable aléatoire Y est définie par $r = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X) \times \mathrm{Var}(Y)}}$. Son équivalent empirique (ou encore la réalisation de son estimateur) est donc

$$\hat{r} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \tag{1}$$

D'après le cours, on sait que la réalisation de l'estimateur de la pente de la droite de régression s'écrit

$$\hat{\beta}_n = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Ainsi, en utilisant le fait que $\sum_{i=1}^n (x_i - \bar{x}) = 0$, on obtient

$$\hat{\beta}_n = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{2}$$

La combinaison de (1) et (2) donne

$$\hat{\beta}_n = \hat{r} \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

On trouve ainsi que $\hat{\beta}_n = 0.6\sqrt{20/10} \approx 0.85$. Le cours nous donne également que la réalisation de l'estimateur de l'ordonnée à l'origine de la droite de régression est $\hat{a}_n = \bar{y} - \hat{\beta}_n \bar{x}$. On a donc $\hat{a}_n = -4.5$.

Exercice 3. a) D'après l'énoncé, on doit minimiser $\sum_{i=1}^n [y_i - y_0 - \beta(x_i - x_0)]^2$. On a donc

$$\begin{aligned}
&\frac{\partial \sum_{i=1}^n [y_i - y_0 - \beta(x_i - x_0)]^2}{\partial \beta} = 0 \\
\Leftrightarrow &-2\sum_{i=1}^n [y_i - y_0 - \beta(x_i - x_0)](x_i - x_0) = 0 \\
\Leftrightarrow &\beta = \frac{\sum_{i=1}^n (x_i - x_0)(y_i - y_0)}{\sum_{i=1}^n (x_i - x_0)^2}.
\end{aligned}$$

Par ailleurs on a

$$\frac{\partial^2 \sum_{i=1}^n [y_i - y_0 - \beta(x_i - x_0)]^2}{\partial \beta^2} = 2\sum_{i=1}^n (x_i - x_0)^2 > 0.$$

On obtient donc que le minimum est atteint pour

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - x_0)(y_i - y_0)}{\sum_{i=1}^n (x_i - x_0)^2}.$$

b) Si l'on pose $(x_0, y_0) = (\bar{x}, \bar{y})$, on obtient

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

On retombe alors sur l'estimateur de la pente de la droite de régression classique. En d'autres termes, l'estimateur de la pente de régression classique correspond à la pente qui minimise l'erreur de la droite de régression forcée à passer par la moyenne $(\bar{x}, \bar{y})$.

c) La réalisation de l'estimateur de la pente de la droite de régression calculée sur notre jeu de données est $\hat{\beta} = \dfrac{24.75}{35} = 0.71$. La droite de régression est donc $y = \hat{\beta}x - \hat{\beta}x_0 + y_0 = 3.87 + 0.71x$.

Exercice 4. a) On a

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - 2\bar{x}\sum_{i=1}^n x_i + n\bar{x}^2.$$

Ainsi, en utilisant les données de l'énoncé, on obtient $\sum_{i=1}^n (x_i - \bar{x})^2 = 76.9$. De même, on a $\sum_{i=1}^n (y_i - \bar{y})^2 = 108.76$. Enfin, on obtient

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \bar{y}\sum_{i=1}^n x_i - \bar{x}\sum_{i=1}^n y_i + n\bar{x}\bar{y} = 72.17.$$

Maintenant, on rappelle que

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ et } \hat{a} = \bar{y} - \hat{\beta}\bar{x}.$$

Ainsi, on obtient les estimations $\hat{\beta} = 0.94$ et $\hat{a} = -4.56$. Finalement, on sait que l'estimateur de la variance du bruit Gaussien η est

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}x_i)^2.$$

Sa réalisation est donc $\hat{\sigma}^2 = 5.13$.

b) On teste l'hypothèse $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$ au niveau de signification de 1%. On sait d'après le cours que sous H_0

$$\frac{\hat{\beta}}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim t_{n-2} = t_8,$$

Ainsi on rejette H_0 si $\left|\frac{\hat{\beta}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}\right| > t_{8,0.995} = 3.355$. On a $\left|\frac{\hat{\beta}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}\right| = 3.64$. On rejette donc H_0 en faveur de $H_1 : \beta \neq 0$.

Exercice 5. a) Puisque $P\,V^\gamma = C$, on a

$$\log(P) + \gamma\log(V) = \log(C) \quad \text{et donc} \quad \log(P) = \log(C) - \gamma\log(V).$$

En posant $X = \log(V)$ et $Y = \log(P)$, l'équation de la droite du modèle linéaire s'écrit

$$Y = \alpha + \beta X\,,$$

où $\alpha = \log(C)$ et $\beta = -\gamma$. Nous souhaitons estimer les paramètres α et β.

b) On sait d'après le cours que les estimateurs des paramètres de la droite de régression sont donnés par

$$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{et} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

où $x_i = \log(v_i)$ et $Y_i = \log(P_i), i = 1, \ldots, 6$. On trouve $\hat{\beta} = -1.4$ et $\hat{\alpha} = 9.66$. Ainsi, on a $\hat{C} = \exp(\hat{\alpha}) = 15677.78$ et $\hat{\gamma} = -\hat{\beta} = 1.4$.

c) On a $\hat{y} = \log(\hat{p}) = \hat{\alpha} + \hat{\beta}\log(v)$. Ainsi, pour $v = 100$, on a $\hat{p} = \exp(\hat{y}) = 24.85$ kg/cm^2.

d) Soit

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}x_i)^2.$$

On sait d'après le cours que

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim t_{n-2}.$$

L'intervalle de confiance à 95% est donc donné par les bornes

$$\hat{\beta} \pm \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} t_{n-2,0.975},$$

où

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}.$$

On a $\hat{\sigma} \approx 0.04$, $\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \approx 1.05$ et $t_{n-2,0.975} = 2.776$. L'intervalle de confiance recherché est donc approximativement $[-1.51, -1.29]$.

Exercice 6. (i). Notons d'abord que $\sum_{i=1}^{n}(x_i - \overline{x}_n)\overline{y}_n = 0$ et donc

$$\widehat{\beta}_n = \frac{\sum_{i=1}^{20}(x_i - \overline{x}_n)y_i}{\sum_{i=1}^{20}(x_i - \overline{x}_n)^2} = \frac{\sum_{i=1}^{20}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^{20}(x_i - \overline{x}_n)^2} = \frac{20 \times 6.26}{20 \times 28.29} \approx 0.22,$$
$$\widehat{a}_n = \overline{y}_n - \widehat{\beta}_n\overline{x}_n \approx 18.34 - 0.22 \times 34.9 \approx 10.662.$$

(ii). Nous nous servons du fait que $\hat{y}_i = \widehat{a}_n + \widehat{\beta}_n x_i$ et $\overline{y}_n = \widehat{a}_n + \widehat{\beta}_n\overline{x}_n$ pour obtenir

$$R^2 = \frac{\sum_{i=1}^{20}(\hat{y}_i - \overline{y}_n)^2}{\sum_{i=1}^{20}(y_i - \overline{y}_n)^2} = \widehat{\beta}_n^2\frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}{\sum_{i=1}^{20}(y_i - \overline{y}_n)^2} \approx 0.22^2 \times \frac{28.29}{2.85} \approx 0.48.$$

La régression explique donc un peu moins de la moitié de la variabilité des $(y_1, \dots, y_n)$.

(iii). Les statistiques de test sont $\widehat{\beta}_n/\sqrt{\widehat{\mathrm{var}}(\widehat{\beta}_n)}$ et $\widehat{a}_n/\sqrt{\widehat{\mathrm{var}}(\widehat{a}_n)}$ et valent respectivement 4.4 et 5.65. Pour un test bilatéral au niveau 0.05, elles sont à comparer au $(1-\alpha/2)$-quantile de la loi de Student avec $n-2 = 18$ degrés de liberté, soit $t_{18,0.975} \approx 2.1$. Nous rejetons dans les deux cas l'hypothèsse de nullité du coefficient.

Nous avons modélisé la hauteur par une fonction affine de la circonférence, il semblerait évident que la droite passe par l'origine (un arbre admettant un diamétre proche de zéro doit être petit). Or, l'hypothése nulle $a = 0$ est rejetée. Les données mesurées indiquent des arbres dont la circonférence varie entre 20 et 50 cm, les estimations des paramétres du modéle sont valides pour des données dans l'intervalle $[20, 50]$cm. La relation en dehors de cet intervalle pourrait être non-linéaire.

Exercice 7. (i). D'aprés les tableaux, la valeur estimée de a est $\widehat{a}_n \approx -585.49$, qui représenterait le niveau de la mer en l'an 0. Son interprétation n'est donc pas très intuitive (c'est très loin dans le passé).

La valeur estimée de β est $\widehat{\beta}_n \approx 0.359$, et elle indique la hauteur en cm qui s'ajoute chaque année.

(ii). En utilisant 1900 comme année de référence, on a

$$\widehat{\beta}_{1n} = \frac{\sum_{i=1}^{20}(x_i - 1900 - (\overline{x}_n - 1900))(y_i - \overline{y}_n)}{\sum_{i=1}^{20}(x_i - \overline{x}_n)^2} = \frac{\sum_{i=1}^{20}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^{20}(x_i - \overline{x}_n)^2} = \widehat{\beta}_n$$

et

$$\widehat{\beta}_{0n} = \overline{y}_n - \widehat{\beta}_{1n}(\overline{x}_n - 1900) = \overline{y}_n - \widehat{\beta}_n\overline{x}_n + 1900\widehat{\beta}_n \approx 96.45.$$

Au niveau du modèle on a la même relation :

$$\alpha + \beta x = y - \epsilon = \beta_0 + \beta_1(x - 1900),$$

et donc $\beta_1 = \beta$ et $\beta_0 = \alpha + 1900\beta$. Ainsi $\beta_1 = \beta$ a la même interprétation alors que β_0 indique le niveau en 1900 et non pas à l'an zéro, ce qui semble plus utile.

(iii). Un intervalle de confiance à un niveau de 95% pour β a pour bornes

$$0.359 \pm t_{130,0.975} \times 0.0385 \approx 0.359 \pm 0.076 \approx (0.283, 0.435) \text{ cm/an},$$

et, forcément l'IC pour β_1 est le même.
Cet intervalle ne contient pas 0, on peut donc rejeter l'hypothése que le niveau de l'eau reste stable au cours des années avec un niveau de significativité 5% (ce qui est évident !)