Corrigé 13

Exercice 1.

a) On a

$$\sum_{j=1}^{n} w_j = \sum_{j=1}^{n} \frac{x_j - \bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left(\sum_{j=1}^{n} x_j - n\bar{x}\right) = 0.$$

On en déduit que

$$\sum_{j=1}^{n} v_j = \sum_{j=1}^{n}\left(\frac{1}{n} - \bar{x}w_j\right) = n\frac{1}{n} - \bar{x}\sum_{j=1}^{n} w_j = 1 - \bar{x}\times 0 = 1.$$

b) On a

$$\begin{aligned}\sum_{j=1}^{n} w_j^2 &= \sum_{j=1}^{n} \frac{(x_j-\bar{x})^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} = \frac{1}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}\sum_{j=1}^{n}(x_j-\bar{x})^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} \\ &= \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.\end{aligned}$$

On en déduit que

$$\sum_{j=1}^{n} v_j^2 = \sum_{j=1}^{n}\left(\frac{1}{n^2} - 2\frac{\bar{x}w_j}{n} + \bar{x}^2 w_j^2\right) = \frac{n}{n^2} - 2\frac{\bar{x}}{n}\sum_{j=1}^{n} w_j + \bar{x}^2\sum_{j=1}^{n} w_j^2 = \frac{1}{n} + \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\bar{x}^2.$$

c) En remarquant que $n\bar{x}^2 = \sum_{j=1}^{n} x_j\bar{x}$, on a $\sum_{j=1}^{n}\left(\bar{x}^2 - x_j\bar{x}\right) = 0$, et donc $\sum_{j=1}^{n}(x_j - \bar{x})x_j = \sum_{j=1}^{n}(x_j-\bar{x})^2$. Ainsi,

$$\begin{aligned}\sum_{j=1}^{n} w_j x_j &= \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{j=1}^{n}(x_j-\bar{x})\,x_j = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{j=1}^{n}(x_j-\bar{x})^2 \\ &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = 1.\end{aligned}$$

On a enfin

$$\sum_{j=1}^{n} v_j x_j = \sum_{j=1}^{n}\left(\frac{1}{n}x_j - \bar{x}w_j x_j\right) = \frac{1}{n}\sum_{j=1}^{n} x_j - \bar{x}\sum_{j=1}^{n} w_j x_j = \bar{x} - \bar{x} = 0.$$

Exercice 2. La corrélation entre une variable aléatoire X et une variable aléatoire Y est définie par $r = \dfrac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \times \text{Var}(Y)}}$. 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

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

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

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

La combinaison de (1) et (2) donne

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

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

Exercice 3.

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

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

Par ailleurs on a

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

On obtient donc immédiatement que le minimum est atteint pour

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

la réalisation de l'estimateur

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - x_0)(Y_i - y_0)}{\sum_{i=1}^{n}(x_i - x_0)^2}. \tag{3}$$

Il convient de faire attention au fait qu'un estimateur et sa réalisation sont généralement notés de la même façon mais il faut bien garder à l'esprit qu'il s'agit de deux objets différents (une variable aléatoire et sa réalisation). Par ailleurs, nous avons utilisé la notation x_i et non X_i en (3) car, dans le cours, la covariable est considérée comme non aléatoire (contrairement à la variable de réponse).

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

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

Ainsi, on obtient les estimations $\hat{b} = 0.94$ 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 $H_0 : b = 0$ contre $H_1 : b \neq 0$ au niveau de signification de 1%. On sait d'après le cours que

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

où $\hat{b}$ désigne ici l'estimateur et non sa réalisation. Ainsi on rejette H_0 si $\left|\frac{\hat{b}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}}\right| > s_1(t_8) = 3.355$. Nous rappelons que, pour $a \in \mathbb{R}$ et $n \in \mathbb{N}$, $s_a(t_n)$ est défini par $\mathbb{P}(|T| > s_a(t_n)) = a/100$, où $T \sim t_n$. On a $\left|\frac{\hat{b}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}}\right| = 3.64$. On rejette donc H_0 en faveur de $H_1 : b \neq 0$.

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

$$\hat{\beta} \pm \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} s_5(t_{n-2}),$$

où

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

On a $\hat{\sigma} \approx 0.04$, $\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \approx 1.05$ et $s_5(t_{n-2}) = 2.776$. L'intervalle de confiance recherché est donc approximativement $[-1.51, -1.29]$.