CHAPITRE 2

# Réponses aux exercices du Chapitre 2

MTH3302 - Méthodes probabilistes et statistiques pour l'I.A.

Jonathan Jalbert

1. a)

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(X^\top X\right)^{-1} X^\top \boldsymbol{Y} \\
&= \left\{(QR)^\top (QR)\right\}^{-1} (QR)^\top \boldsymbol{Y} \\
&= \left\{(R^\top \underbrace{Q^\top Q}_{I} R)\right\}^{-1} (QR)^\top \boldsymbol{Y} \\
&= \left\{(R^\top R)\right\}^{-1} (QR)^\top \boldsymbol{Y} \\
&= \left(R^\top R\right)^{-1} R^\top Q^\top \boldsymbol{Y} \\
&= R^{-1}(R^\top)^{-1} R^\top Q^\top \boldsymbol{Y} \\
&= R^{-1} Q^\top \boldsymbol{Y}.
\end{aligned}$$

b)

$$\begin{aligned} H &= X\left(X^\top X\right)^{-1} X^\top \\ &= (QR)\left\{(QR)^\top (QR)\right\}^{-1} (QR)^\top \\ &= QR\left\{(R^\top R\right\}^{-1} R^\top Q^\top \\ &= QRR^{-1}\left(R^\top\right)^{-1} R^\top Q^\top \\ &= QQ^\top. \end{aligned}$$

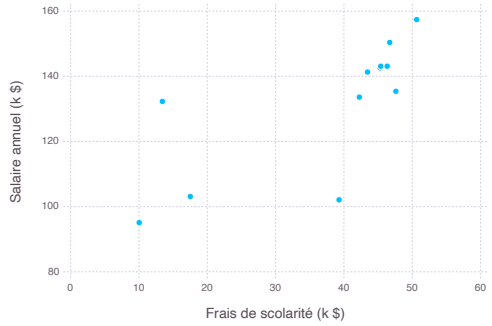
2. On a que

$$SS_T = SS_R + SS_E \quad \text{et } R^2 = \frac{SS_R}{SS_T}$$

Dans la première équation, remplaçons $SS_R$ par $R^2 \times SS_T$. On trouve que

$$\begin{aligned} SS_T &= R^2 \times SS_T + SS_E \\ R^2 \times SS_T &= SS_T - SS_E \\ R^2 &= 1 - \frac{SS_E}{SS_T}. \end{aligned}$$

3. a) Le nuage de points est le suivant :




La relation linéaire semble appropriée quoiqu'il y a présence de deux points extrêmes.

b) Si on transforme les données en milliers de dollar, on obtient $\hat{\beta}_0 = 93.6$ et $\hat{\beta}_1 = 1.02$.

c) L'interprétation de la pente ($\hat{\beta}_1 = 1.02$) est que le salaire annuel des diplômés augmente en moyenne de 1020 $ pour chaque millier de dollar d'augmentation des frais de scolarité.

d) L'interprétation de l'ordonnée à l'origine (valeur = 93.6) est que s'il y avait des universités avec des frais de scolarité nuls, la moyenne prévue pour le salaire des diplomés serait de 93 600$. Or, puisqu'il n'y a pas de telles universités, cette interprétation ne s'applique pas.

e) Le coefficient de détermination est de 0.54. Cette valeur peut être interprétée de la façon suivante : le montant des frais de scolarité expliques 54% de la variation observée des salaires des diplomés.

4. a) $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 117.1 & 4.33 & -2.86 & -2.19 \end{bmatrix}^\top$

b) $\hat{\sigma}^2 = 6.15$

c) $F = 21.5$
valeur-p $= 7.34 \times 10^{-6}$
Comme la valeur-p est plus petite que 5%, on rejette l'hypothèse nulle stipulant qu'aucune des variables explicatives possède un pouvoir prédictif significatif.

d)

$$\begin{aligned} \beta_0 &\in [-94.4\ ,\ 328.6] \\ \beta_1 &\in [-2.06\ ,\ 10.73] \\ \beta_2 &\in [-8.33\ ,\ 2.62] \\ \beta_3 &\in [-5.57\ ,\ 1.20] \end{aligned}$$

La valeur 0 est incluse dans tous ces intervalles. Cela suggère que toutes les variables explicatives ne sont significatives pas significatives.

**Remarque.** *Les intervalles de confiance semblent très larges. Ceci nous permet de suspecter la présence de multicolinéarité que nous verrons à la section 2.9.*

e) $Y_0 \in [15.0\ ,\ 27.9]$ avec une probabilité de 95%.

5. a) $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 13.7 & 0.32 & 0.61 & -2.68 \end{bmatrix}^\top$

b)

$$\begin{aligned}\beta_0 &\in [9.07\ ,\ 18.33]\\ \beta_1 &\in [0.005\ ,\ 0.64]\\ \beta_2 &\in [0.35\ ,\ 0.88]\\ \beta_3 &\in [-4.71\ ,\ -0.66]\end{aligned}$$

c) Oui car 0 n'est pas inclus dans l'intervalle de confiance.

d) La prédiction de sa note au final est de 24.4.

6. a) Non. Le nuage de points illustrant la viscosité en fonction de la température ne suggère pas une relation linéaire.

   b) $\hat{\boldsymbol{\beta}} = \begin{bmatrix}-148\,395 & 2183 & -7.739\end{bmatrix}^\top$
   $R^2_{aj} = 0.86$

   c) $\hat{\boldsymbol{\beta}} = \begin{bmatrix}-4.26 \times 10^6 & 8.61 \times 10^4 & -5.77 \times 10^2 & 1.29\end{bmatrix}^\top$
   $R^2_{aj} = 0.98$

   d) Le modèle cubique car son coefficient de détermination ajusté est supérieur.

7. Le vecteur des facteurs d'inflation de la variance est : $\text{VIF} = \begin{bmatrix}709 & 564 & 105\end{bmatrix}^\top$. Puisqu'au moins un VIF est supérieur à 10, la multicolinéarité est préoccupante.

8. a) Le vecteur des facteurs d'inflation de la variance est :

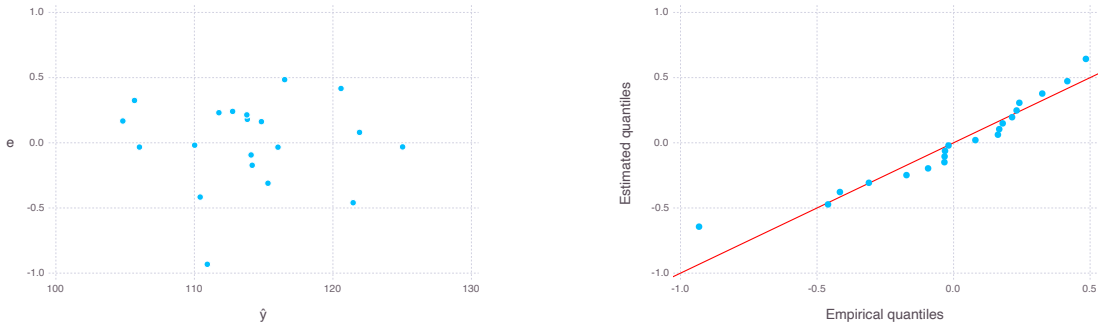
   $$\text{VIF} = \begin{bmatrix}1.76 & 8.42 & 5.33 & 1.24 & 4.41 & 1.83\end{bmatrix}^\top.$$

   Puisque tous les VIF sont inférieurs à 10, la multicolinéarité n'est pas préoccupante.

   b) La figure 2.1a illustre les résidus en fonction des estimations. On constate que qu'ils sont centrés autour de 0 et que leur dispersion semble constante. Alors les hypothèses 1 et 2 semblent satisfaites.

   La figure 2.1b illustre la droite de Henry des résidus. On constate que la loi normale est raisonnable pour les résidus. L'hypothèse 4 est donc satisfaite.

   Avec les données seulement, il est impossible de vérifier si l'hypothèse 3 d'indépendance est satisfaite. Dans ce cas, on espère que c'est le cas en postulant l'indépendance.




(a) (b)

FIGURE 2.1 – Résidus en fonction des estimations (a) et droite de Henry des résidus (b).

c) Il y a 63 modèles possibles si on exclut le modèle trivial avec aucune variable explicative. Le pire modèle ($R^2_{aj} = 0.0269$) est obtenu en n'utilisant que la variable *stress*. Le meilleur modèle ($R^2_{aj} = .9948$) est obtenu en utilisant les 6 variables explicatives. Alors les 6 variables explicatives possèdent un pouvoir prédictif significative sur la tension artérielle.

9. a) Vrai
   b) Faux
   c) Faux
   d) Vrai
   e) Faux
   f) Faux