Examen Final

Durée 2h. Les documents, la calculatrice, les téléphones portables, tablettes, ordinateurs ne sont pas autorisés. Les exercices sont indépendants. La qualité de la rédaction sera prise en compte.

Exercice 1. Interprétation géométrique en régression linéaire simple

On considère un modèle de régression linéaire simple avec constante :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

où $\mathbf{Y}$ est un vecteur aléatoire à valeurs dans $\mathbb{R}^3$, $\boldsymbol{\beta} \in \mathbb{R}^2$, $\boldsymbol{\epsilon}$ est un vecteur aléatoire vérifiant les conditions classiques d'un modèle de régression linéaire, et $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix}$.

On note $\mathbb{1} = (1,1,1)^T$, $\mathbf{x} = (x_1, x_2, x_3)^T$, $\langle \cdot, \cdot \rangle$ le produit scalaire sur $\mathbb{R}^3$ et $\mathbf{P_X}$ la matrice de projection sur l'espace vectoriel $\mathcal{M}(\mathbf{X})$ engendré par $\mathbb{1}$ et $\mathbf{x}$.

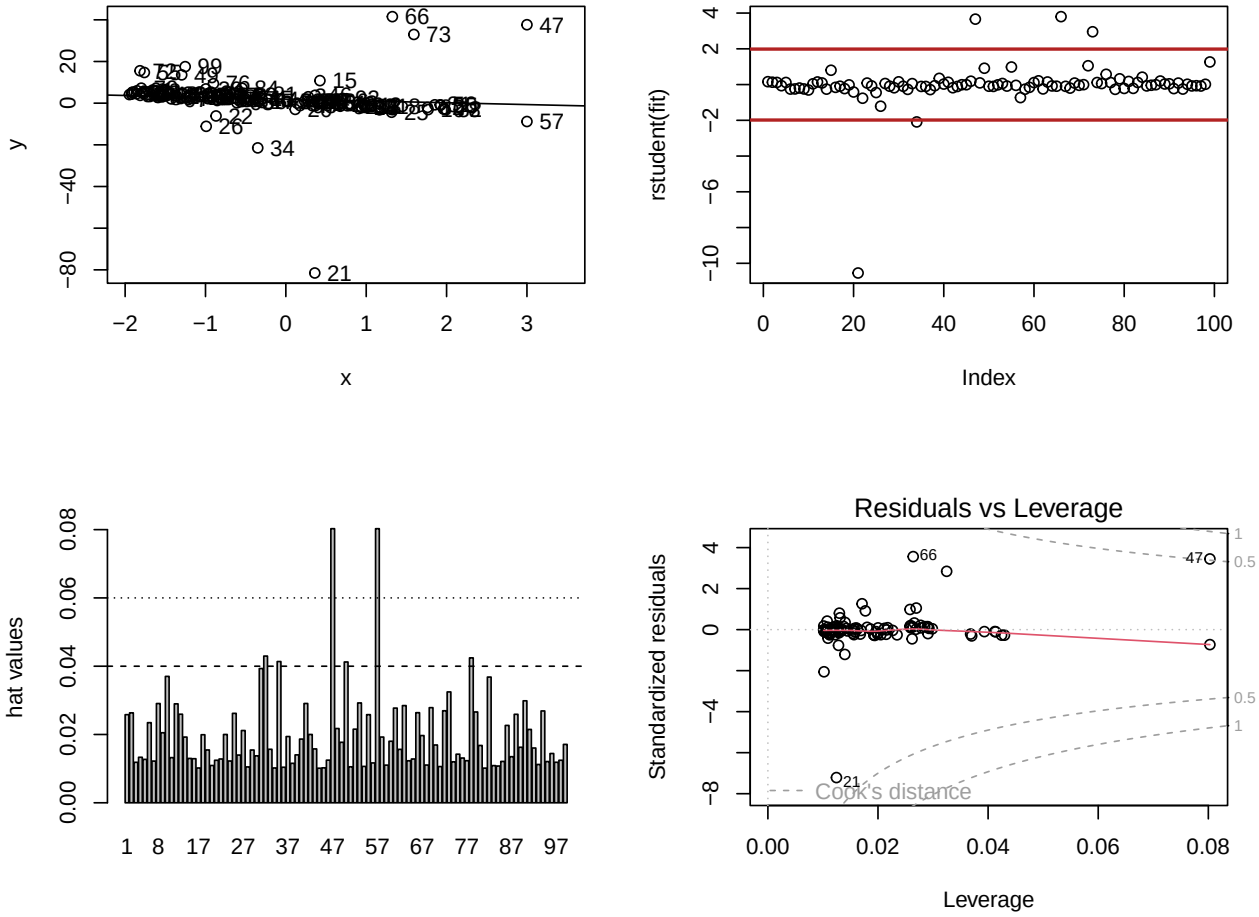
1. Donnez les expressions de $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{Y}}$, $\hat{\boldsymbol{\epsilon}}$ et $\mathbf{P_X}$ en fonction des données $\mathbf{Y}$ et $\mathbf{X}$.
2. Représentez graphiquement : le vecteur $\mathbb{1}$, le vecteur $\mathbf{x}$, $\mathcal{M}(\mathbf{X})$, $\mathbf{Y}$, $\hat{\mathbf{Y}}$, $\bar{\mathbf{Y}}\mathbb{1}$, $\hat{\boldsymbol{\epsilon}}$. Représentez tous les angles droits visibles sur le graphe, ainsi que l'angle θ permettant de définir le critère du R^2. Quelle est l'interprétation géométrique de l'équation de la variance totale (TSS = ESS + RSS)?
3. Donner une expression simplifiée de :
   - $\langle \mathbf{Y} - \hat{\mathbf{Y}}, \mathbf{Y} \rangle$,
   - $\langle \mathbf{Y}, \mathbb{1} \rangle$,
   - $\left\| \mathbf{Y} - \frac{1}{n}\langle \mathbf{Y}, \mathbb{1} \rangle \mathbb{1} \right\|^2 - \left\| \mathbf{P_X}(\mathbf{Y} - \bar{\mathbf{Y}}\mathbb{1}) \right\|^2$,
   - $\|\mathbf{Y}\|^2 - \|\hat{\boldsymbol{\epsilon}}\|^2$,
   - $\mathbf{P_X}(\mathbf{Y} - \bar{\mathbf{Y}}\mathbb{1}) - (\mathbf{Y} - \bar{\mathbf{Y}}\mathbb{1})$,
   - $\mathbf{P_X}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$.

Exercice 2. Validation de Modèle

On mesures deux variables $\mathbf{y}$ et $\mathbf{x}$, et l'on procède à l'analyse ci-dessous.

```
## Fit
fit <- lm(y ~ x)

par(mfrow = c(2, 2))
## Plot Data
plot(y ~ x, xlim = c(-2, 3.5)); abline(fit)
text(y ~ x, adj = -0.5)
## Plot studentized residuals
plot(rstudent(fit))
alp <- 0.05
qs <- qt(1 - alp/2, n-2)
abline(h = c(qs, -qs), col = "firebrick", lwd = 2)
## Plot leverage
barplot(hatvalues(fit), ylab = "hat values")
abline(h = c(2 * 2 / n, 3 * 2 / n), lty = c(2, 3))
## Plot Cook distance
plot(fit, 5)
```



1. Quel modèle a-t-on utilisé ? Rappelez les hypothèses sous-jacentes.
2. On rappelle la définition du *résidu studentisé* au point i:

$$t_i^* = \frac{y_i - \hat{y}_i^P}{\sqrt{\hat{\sigma}^2_{(-i)}\left(1 + \mathbf{x}^i(\mathbf{X}^T_{(-i)}\mathbf{X}_{(-i)})^{-1}(\mathbf{x}^i)^T\right)}}$$

   Explicitez les termes impliqués dans la définition. Que représentent $\hat{\sigma}^2_{(-i)}$, $\hat{y}_i^P$, $\mathbf{x}^i$, $\mathbf{X}_{(-i)}$? Quelles propriétés ont ces résidus studentisés ?
3. Sur quelle métrique se base-t-on pour détecter des points levier (en anglais, *leverage*) ?
4. On rappelle la définition de la *distance de Cook* pour un point i:

$$C_i = \frac{1}{p\hat{\sigma}^2}\left(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}}\right)^T (\mathbf{X}^T\mathbf{X}) \left(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}}\right)$$

   Explicitez les termes de l'équation. Que représente $\hat{\boldsymbol{\beta}}_{(-i)}$?
   Cette quantité est-elle une distance ? Quelle signification peut-on donner à des points dont la distance de Cook est élevée ?
5. Dans les données représentées ci-dessus :
   - Y-a-t-il des points aberrants ?
   - Y-a-t-il des points leviers ?

– Si l'on retirait ces potentiels points aberrants ou levier de l'analyse, obtiendrait-on des résultats très différents pour la régression linéaire ?

Justifiez vos réponses.

Exercice 3. Régression robuste

Dans cet exercice, on se place dans le cadre du modèle linéaire simple avec des erreurs qui suivent une loi de Cauchy. Plus précisément, on considère le modèle:

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i \qquad \forall 1 \leq i \leq n \tag{1}$$

avec les ϵ_i iid, qui suivent une loi de Cauchy paramètre de position 0 et de paramètre de dispersion γ.

1. On rappelle la densité f de la loi de Cauchy de paramètre de position $m \in \mathbb{R}$ et de paramètre de dispersion $\gamma > 0$:

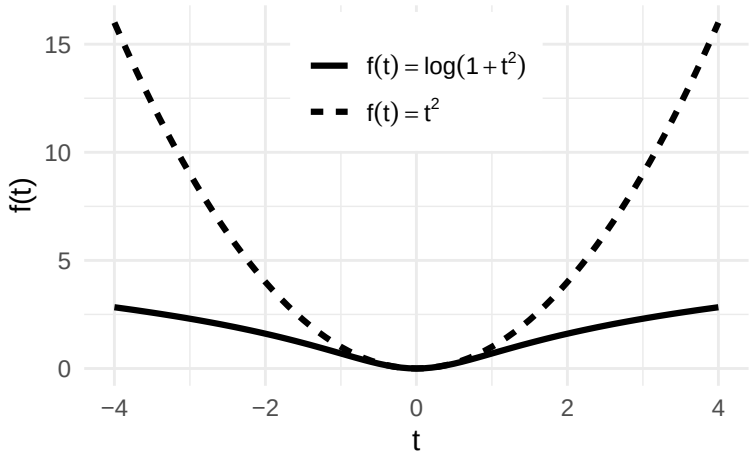
$$f(t) = \frac{1}{\pi\gamma\left(1 + \left(\frac{t-m}{\gamma}\right)^2\right)} \qquad t \in \mathbb{R}.$$

   (a) Vérifiez que la fonction f est bien une densité de probabilité.
   (b) Cette loi admet-elle une espérance ? Une variance ?
   (c) Soit Z une variable aléatoire suivant une loi de Cauchy de paramètres m et γ, et soit $a \in \mathbb{R}$ une constante. Quelle est la loi de $Z + a$? Justifiez.

2. On se place dans le cadre du modèle (1).
   (a) Ce modèle suit-il les hypothèses traditionnelles de la régression linéaire ?
   (b) Quelle est la loi d'une observation y_i suivant ce modèle ?
   (c) Donnez l'expression de la vraisemblance $L(\beta_1, \beta_2, \gamma|y)$ du vecteur $y = (y_1, \ldots, y_n)$ dans ce modèle. En déduire que les estimateurs du maximum de vraisemblance $(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\gamma})$ sont donnés par:

$$(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\gamma}) = \underset{(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\gamma}) \in \mathbb{R}^2 \times \mathbb{R}_+^*}{\operatorname{argmin}} \; n\log(\gamma) + \sum_{i=1}^{n} \log\left(1 + \left(\frac{y_i - \beta_1 + \beta_2 x_i}{\gamma}\right)^2\right). \tag{2}$$

3. En utilisant le graphique ci-dessous, justifiez empiriquement et sans faire de calculs qu'un point aberrant (*outlier*) a moins d'influence sur la minimisation de l'équation (2) que sur la minimisation des moindres carrés. (On pourra raisonner à paramètre γ fixé.) Cela vous paraît-il cohérent avec les formes des densités gaussiennes et de Cauchy ?



4. La propriété ci-dessus rend la régression avec résidus de Cauchy plus "robuste", c'est-à-dire moins sensible aux données aberrantes. On cherche dans cette question à l'illustrer sur les données de l'exercice précédent (exercice 2).
On suppose que l'on a accès à une fonction `lmCauchy` qui, étant donnée une réponse y et un prédicteur x, renvoie les estimateurs $(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\gamma})$ définis précédemment. En reprenant les données de l'exercice précédent, on exécute les lignes de code suivantes.

```
coef(lm(y ~ x))

## (Intercept)           x
##   1.9358731  -0.8730434

coef(lm(y[-47] ~ x[-47]))

## (Intercept)      x[-47]
##    1.305202   -1.778146

lmCauchy(y, x)

##      beta1      beta2      gamma
##  0.9775317 -2.0877324  0.8506250

lmCauchy(y[-47], x[-47])

##      beta1      beta2      gamma
##  0.9756348 -2.0893676  0.8343814
```

   (a) Avec la régression classique, que se passe-t-il lorsque l'on enlève le point 47 ? Cela vous surprend-il au vu des résultats de l'exercice 2 ?
   (b) Que dire de l'estimation de la pente par le modèle de régression avec résidus Cauchy? En quel estimateur avez-vous le plus confiance ?