# TD III : Régression linéaire

**Exercice 1.** *1. Que vaut la somme des résidus $\hat{\varepsilon}$ d'une régression linéaire?*

*2. Le vecteur $\hat{Y}$ est-il orthogonal au vecteur des résidus $\hat{\varepsilon}$? est-il indépendant?*

*3. On note SCR1 la somme des carrés des résidus d'une régression. Après l'ajout d'une nouvelle variable explicative, on note SCR2 la somme correspondant à la nouvelle régression. Peut-on comparer SCR1 et SCR2? Que cela implique-t-il pour le R2?*

*4. Même question avec $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$ à la place de SCR1 et SCR2.*

**Exercice 2.** *On considère un modèle de régression linéaire simple $y_i = ax_i + b + \varepsilon_i$.*

*1. Rappeler l'expression des estimateurs MCO dans le cas de la régression linéaire simple.*

*2. Montrer à la main que ces estimateurs minimisent la fonction $(a, b) \to \sum_i (y_i - ax_i - b)^2$.*

**Exercice 3.** *On considère un modèle de régression linéaire simple $y_i = ax_i + b + \varepsilon_i$. On veut montrer que les estimateurs MCO ont une variance minimale parmi les estimateurs* ***linéaires sans biais*** *de $a$ et $b$. On ne considérera que le cas de $\hat{a}$. Soit $\tilde{a}$ un estimateur linéaire sans biais de $a$, qui s'écrit $\tilde{a} = \sum_i \lambda_i x_i$. On veut montrer que $\mathrm{V}(\tilde{a}) \geq \mathrm{V}(\hat{a})$.*

*1. Montrer que nécessairement $\sum_i \lambda_i = 0$ et $\sum_i \lambda_i x_i = 1$.*

*2. Calculer $\mathrm{V}(\hat{a})$ et $\mathrm{Cov}(\hat{a}, \tilde{a})$. En déduire que $\mathrm{Cov}(\hat{a} - \tilde{a}, \hat{a}) = 0$.*

*3. Conclure.*

**Exercice 4.** *Soient $m$ échantillons Gaussiens indépendants $(Y_j^{(i)},\ j \leq n_i)_{1 \leq i \leq m}$, où pour chaque $i$, l'échantillon $(Y_j^{(i)},\ j \leq n_i)$ est une collection de variables i.i.d de moyenne $\mu_i$ dépendant de $i$ et de variance $\sigma^2$ ne dépendant pas de $i$.*

*1. Montrer que ce modèle peut s'écrire comme un modèle ANOVA à un facteur.*

2. *En déduire un test de* $(H_0) : \mu_i = \mu_j \, \forall \, i, j$ *contre* $(H_1) : \exists \, i, j, \, \mu_i \neq \mu_j$.

**Exercice 5.** *Des données pour 66 communes de Seine et Marne ont été extraites de la plateforme* **`datagouv.fr`** *(données 2006). On dispose du nom de la commune, du revenu imposable brut par ménage (variable Revenu) et du Prix moyen par logement (variable PrixLogement). Des extraits de sorties R sont fournis, voir les figures ci-dessous.*

1. *Pour quelles variables $x$ et $y$ le modèle de régression linéaire simple $y = ax + b + \varepsilon$ a-t-il été demandé Ã  R? D'après les sorties R, ce modèle vous semble-t-il pertinent?*

2. *Donner les valeurs des estimateurs MCO pour le problème étudié ici en utilisant les sorties R.*

3. *Sur la légende du graphique, indiquer Ã  quoi correspond chaque numéro 1,2 et 3 parmi : droite de régression, intervalle de confiance, intervalle de prédiction.*

4. *Donner la formule permettant de calculer la sortie "Residual standard error".*

5. *La sortie* `Multiple R-squared` *est le R2 de la régression: $R2 = \frac{SCE}{SCT}$ où $SCE$ est la somme des carrés expliquée ($\sum_i (\hat{y}_i - \bar{y})^2$) et $SCT$ est la somme des carrés totale ($\sum_i (y_i - \bar{y})^2$). Retrouver la sortie* `F-statistic` *à partir du R2.*

6. *Tester au risque 5% l'effet de la variable Revenu sur la variable PrixLogement. On posera les hypothèses de test, on donnera la statistique de test et la règle de décision proprement. On donnera la réponse finale en utilisant les sorties R.*

7. *Retrouver l'expression générale d'un intervalle de confiance au niveau $1 - \alpha$ pour l'espérance de PrixLogement pour une commune dont la variable Revenu vaut une valeur $r$ donnée. En utilisant les sorties R, donner (la réalisation de) l'intervalle de confiance au niveau 95% pour l'espérance de PrixLogement pour une commune dont le revenu imposable brut par ménage est de 28 185 euros (On posera juste le calcul, ce n'est pas la peine d'effectuer le calcul).*

8. *On s'intéresse à la validation du modèle. En s'aidant des graphiques, répondre aux questions suivantes:*

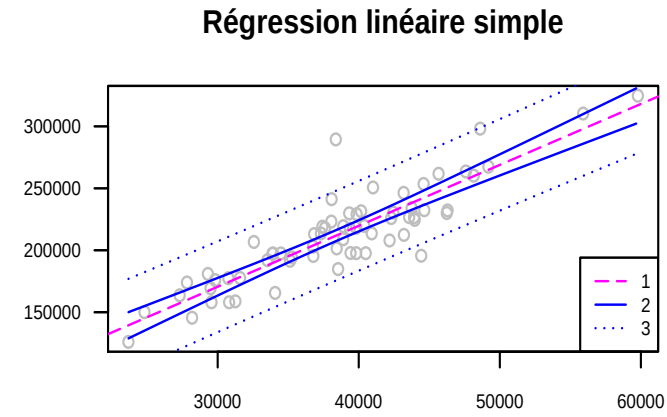
   (a) *L'hypothèse de normalité vous semble-t-elle vérifiée?*

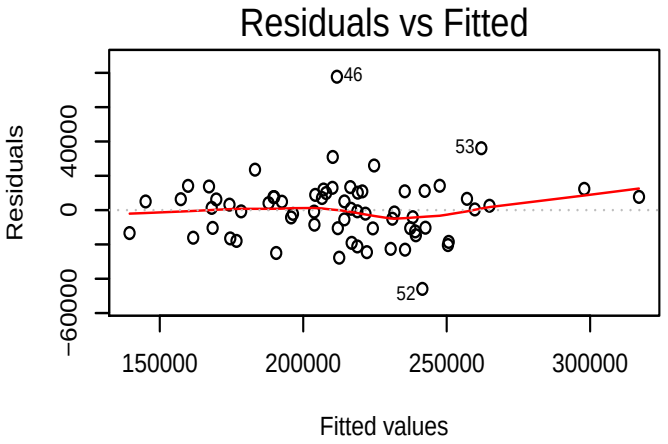
   (b) *L'hypothèse d'indépendance vous semble-t-elle vérifiée?*

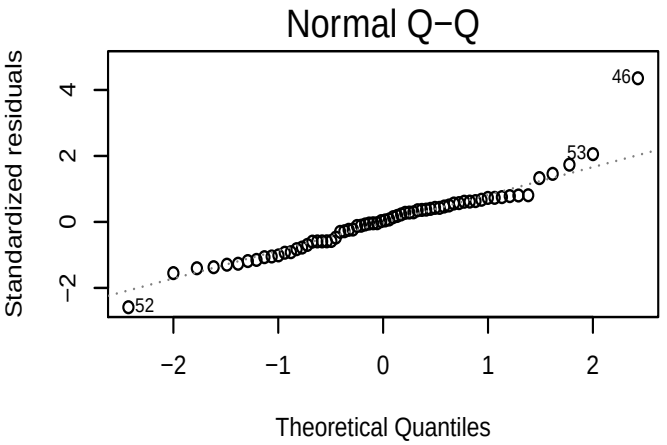
   (c) *Y a-t-il des valeurs qui vous semblent aberrantes? Si oui, sont-elles gênantes?*

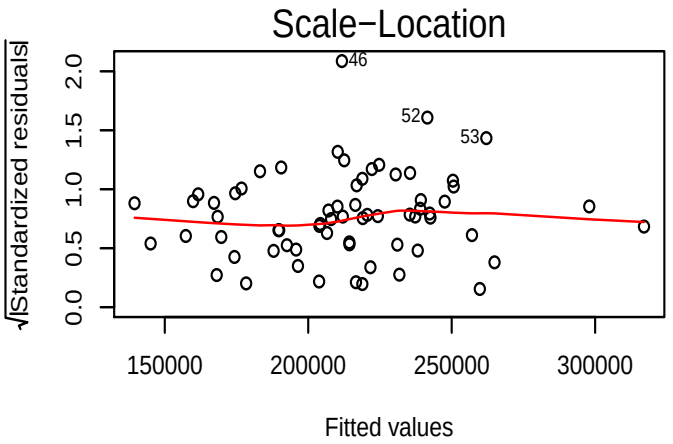
   (d) *Y a-t-il des points leviers? Si oui, sont-ils gênants?*

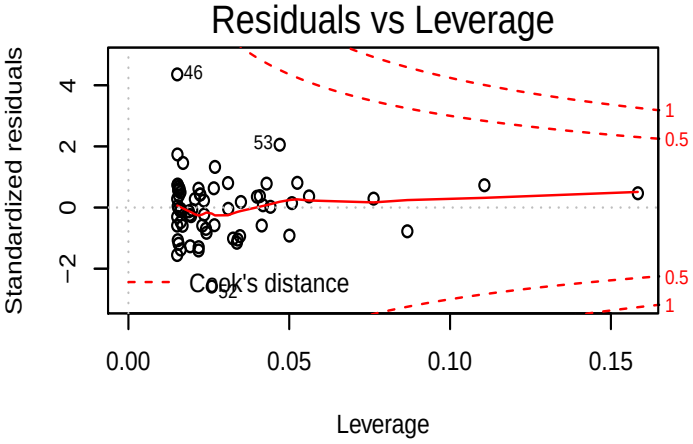
```
Call:
lm(formula = PrixLogement ~ Revenu)

Residuals:
   Min     1Q Median     3Q    Max
-45900 -10618    611   9698  77761

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.306e+04  1.263e+04   1.826   0.0725 .
Revenu      4.917e+00  3.218e-01  15.280   <2e-16 ***
---

Signif. codes 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17990 on 64 degrees of freedom
Multiple R-squared:  0.7849,        Adjusted R-squared:  0.7815
F-statistic: 233.5 on 1 and 64 DF,  p-value: < 2.2e-16
```


Régression linéaire simple
300000
250000
200000
150000
30000
40000
50000
60000
1
2
3


Residuals vs Fitted
Residuals
46
53
52
−60000
0
40000
150000
200000
250000
300000
Fitted values


Normal Q−Q
Standardized residuals
46
53
52
−2
0
2
4
−2
−1
0
1
2
Theoretical Quantiles


Scale−Location
√|Standardized residuals|
46
52
53
0.0
0.5
1.0
1.5
2.0
150000
200000
250000
300000
Fitted values


Residuals vs Leverage
Standardized residuals
46
53
52
Cook's distance
1
0.5
0.5
1
−2
0
2
4
0.00
0.05
0.10
0.15
Leverage