Ecole Polytechnique Universitaire de Lille Lille, le 14 Décembre 2017
Université des Sciences et Technologies de Lille

Examen de Régression Linéaire GIS2A-4

(cours de C. Preda)

Temps de travail : 2h
Tous documents autorisés.

Exercice 1. (9p)

Un modèle de régression linéaire multiple de la forme suivante a été ajusté sur un jeu de données :

$$Y_i = \beta_0 + \beta_1 \cdot x_{i,1} + \beta_2 \cdot x_{i,2} + \beta_3 \cdot x_{i,3} + \beta_4 \cdot x_{i,4} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

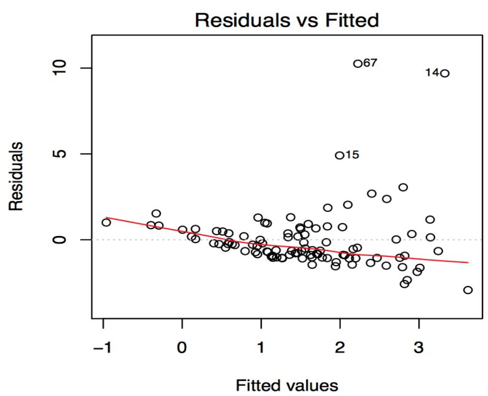
Avec le logiciel R on obtient la sortie suivante :

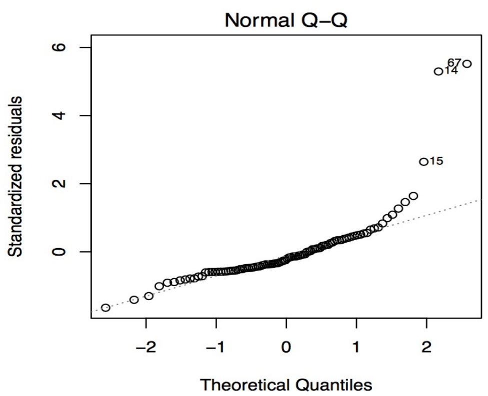
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)      ???     0.1960   8.438 3.57e-13
x1            5.3036     2.5316     ??? 0.038834
x2            4.0336     2.4796   1.627 0.107111
x3           -9.3153     2.4657  -3.778 0.000276
x4            0.5884     2.2852   0.257 0.797373

Residual standard error: 1.892 on 95 degrees of freedom
Multiple R-squared: 0.1948,Adjusted R-squared: ???
F-statistic: 5.745 on 4 and 95 DF,  p-value: 0.0003483
```

On demande :

1) Combien d'observations ont été utilisées pour estimer le modèle ?
2) Quelle est la valeur de l'estimation de β_0 et de la statistique de test pour vérifier l'hypothèse $\beta_1 = 0$.
3) Avec un risque d'erreur de 5%, peut-on rejeter l'hypothèse selon laquelle la variable X3 n'influence pas Y dans ce modèle ?
4) Quelle est l'estimation de la variance de ε ?
5) Quelle est votre interprétation pour le coefficient de la variable X3 ?
6) Que représente R2 ici ? Calculer aussi R2-ajusté et expliquer son utilité.
7) Quelle hypothèse nulle est associeé à la statistique de test de Fisher *F-statistic ?*
8) Quelle est la prédiction faite par le modèle pour $x_1 = 2$, $x_2 = 1$, $x_3 = 4$ et $x_4 = 0$? Que représente cette prédiction ?

9) Voici le graphique des résidus :





Le test de normalité de Shapiro Wilks nous donne une p-valeur égale à 0.002 et le test de homoscédasticité de Pagan p-valeur < 0.001.
Que pouvez vous commenter quant aux conditions d'applications du modèle ?

Exercice 2 (6p).
Soit X_1 et X_2 deux variables quantitatives qui représentent les observations d'une même variable aléatoire X par deux observateurs différents. On utilise le modèle de régression

$$X_2 = \beta X_1 + \varepsilon \text{ , } \varepsilon \ \gamma \ \Omega(0,\sigma^2) \text{ indépendant de } X_1$$

notamment pour vérifier l'hypothèse selon laquelle les deux observateurs mesurent la même chose (mesures concordantes).

On demande :

1) Quelle est l'estimateur de moindres carrés de β ?
2) Proposez un estimateur pour σ^2.
3) Quelle est la statistique de test utilisée pour vérifier l'hypothèse selon laquelle les deux observateurs concordent ?
4) Pour $\beta = 1$ tracez la droite de régression.

Exercice 3 (5p). Dans un échantillon aléatoire de 100 sujets choisis dans une population en surcharge pondérale, on étudie la relation entre le tour de taille en centimètres et la glycémie à jeun.

On obtient les valeurs suivantes :
Tour de Taille : TT = 91 ± 14 (Moyenne ± écart-type)
Glycémie : G = 1.10 ± 0,15 (Moyenne ± écart-type)
La covariance entre TT et G a été calculée : Cov(TT,G) = 0.84.

a) Donner un intervalle de confiance de niveau 95% pour la moyenne du tour de taille dans la population étudiée.
b) Tracer sur un graphe la droite de régression linéaire expliquant le TT en fonction de G.
c) Quelle est la qualité du modèle linéaire ?
d) On observe un individu dont la glycémie à jeun vaut 1.1. Quelle est la valeur prédite par le modèle linéaire pour le tour de taille de cet individu ?