École d'actuariat
Université Laval

ACT-2003 Modèles linéaires en actuariat
Exercice Supplémentaire Chapitre 3

Marie-Pier Côté
Automne 2018

Question 1. Le modèle de régression linéaire multiple

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \text{ pour } i = 1, ..., n$$

a été ajusté à des données avec la méthode des moindres carrés.

(i) La figure 1 montre le QQ-plot des résidus studentisés. À la lumière de ce graphique, y a-t-il un postulat du modèle qui n'est pas vérifié ? Si oui, lequel et pourquoi ? S'il y a lieu, expliquer l'impact de la violation de ce postulat.

(ii) La figure 2 montre les résidus studentisés en fonction de chacune des variables exogènes et en fonction des valeurs prédites. Utiliser ces graphiques pour commenter sur la validité des postulats du modèle. Y en a-t-il qui ne sont pas respectés ? S'il y a lieu, expliquer l'impact de la violation de ce ou ces postulats.

FIGURE 1: QQ-Plot des résidus studentisés

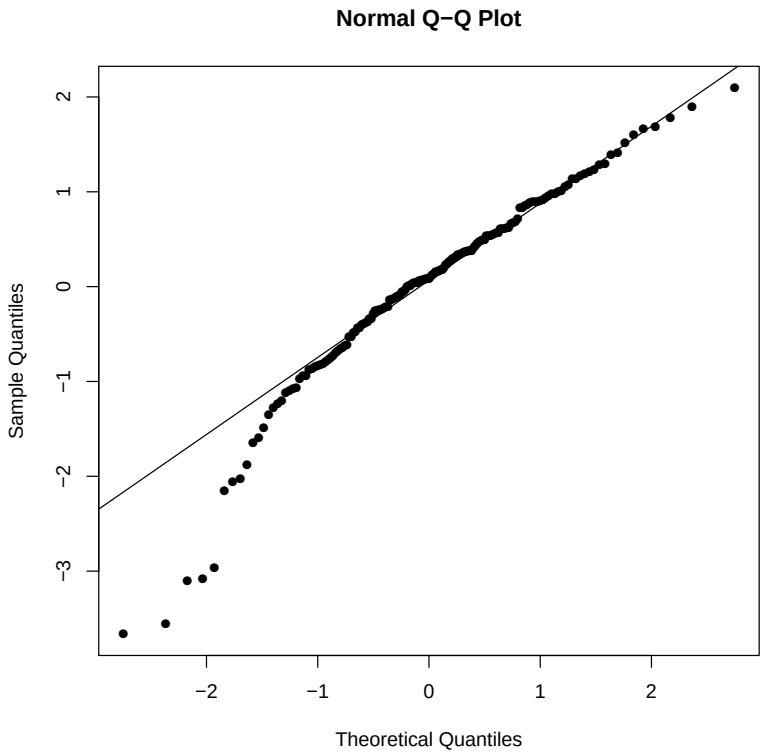
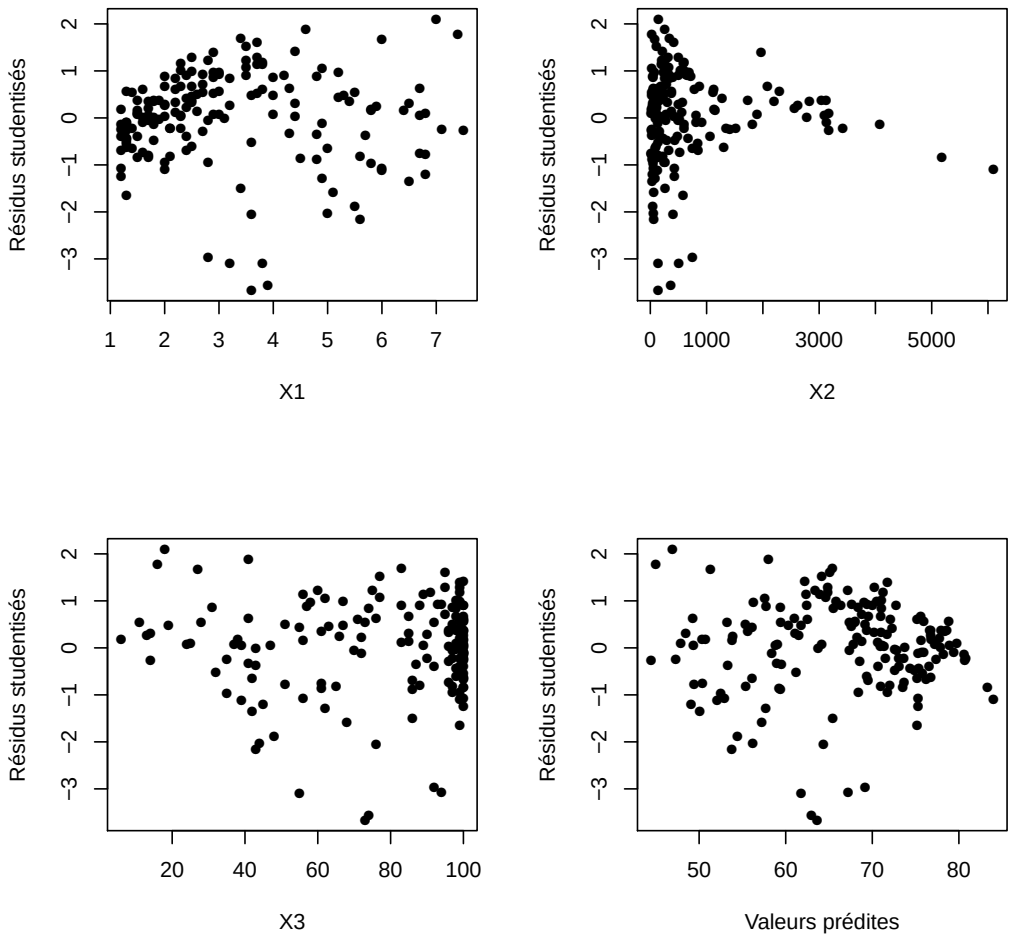


FIGURE 2: Nuage de points des résidus studentisés en fonction de chacune des variables exogènes et en fonction de la variable prédite



Question 2. La base de données `OutlierExample.csv` disponible sur le site du cours contient 19 observations de base, et trois observations supplémentaires, notées par les `CODES` 1, 2 et 3, qui sont aberrantes ou influentes.

- Importez la base de données et tracez un nuage de points de `Y` en fonction de `X`.
- Roulez les lignes de code suivantes pour observer le graphique avec les 3 points ajoutés

```
library(ggplot2)
ggplot(dat, aes(x= X, y= Y, label=CODES))+
  geom_point() +
  geom_text(aes(label=ifelse(CODES>0,CODES,'')),hjust=0,vjust=0)
```

- Ajustez un modèle linéaire en incluant seulement les 19 points dont le code est 0. Regardez l'ajustement et commentez.
- Ajustez un modèle linéaire en incluant les 19 points dont le code est 0 et le point de code 1. Quel est l'impact de l'inclusion de ce point sur le R^2 et sur les estimations des paramètres ? Étudiez le résultat de la fonction `influence.measures()`.
- Ajustez un modèle linéaire en incluant les 19 points dont le code est 0 et le point de code 2. Quel est l'impact de l'inclusion de ce point sur le R^2 et sur les estimations des paramètres ? Étudiez le résultat de la fonction `influence.measures()`.
- Ajustez un modèle linéaire en incluant les 19 points dont le code est 0 et le point de code 3. Quel est l'impact de l'inclusion de ce point sur le R^2 et sur les estimations des paramètres ? Étudiez le résultat de la fonction `influence.measures()`.

SOLUTIONS

Question 1.

(i) Le postulat de normalité semble violé.
La distribution des résidus a une queue inférieure plus épaisse que la loi normale, ce que l'on voit à gauche du Q-Q plot, puisque les poits ne sont pas alignés.
Le postulat de normalité n'est pas critique, parce que les estimateurs des moindres carrés ont un sens quand même. Toutefois, les tests d'hypothèses et les intervalles de confiance ne sont pas valides.

(ii) Le graphique des résidus en fonction de x_2 montre que le postulat de linéarité semble violé. Cela implique que le modèle n'est pas valide.
On observe de l'hétéroscédasticité (par exemple, dans les graphiques 1, 3 ou 4) puisque les résidus ne semblent pas avoir une variance constante.
Cela signifie que les variances des paramètres ne sont pas calculées de façon appropriée OU il faudrait effectuer une transformation sur les variables pour régler ces problèmes.

Question 2. On pourrait croire qu'un point sur 20, ça ne change rien, mais ce n'est pas le cas ! Le point 1 a un impact sur la pente et la qualité de l'ajustement. Le point 2 a un grand levier mais n'affecte pas beaucoup les estimations, le point 3 a un grand levier et un gros impact.

```
dat <- read.csv("OutlierExample.csv")

dim(dat)

summary(dat)

library(ggplot2)

ggplot(dat, aes(x= X, y= Y, label=CODES))+
  geom_point() +
  geom_text(aes(label=ifelse(CODES>0,CODES,'')),hjust=0,vjust=0)

fit0 <- lm(Y~X,dat,subset=(CODES==0))
summary(fit0)
plot(dat[,1:2],pch=16)
points(dat[match(1:3,dat$CODES),1:2],col=2:4,pch=16:18,cex=1.2)
abline(fit0)

fit1 <- lm(Y~X,dat,subset=(CODES<=1))
summary(fit1)
```

```
abline(fit1,col=2,lty=2)

fit2 <- lm(Y~X,dat,subset=(CODES%in%c(0,2)))
summary(fit2)
abline(fit2,col=3,lty=3)

fit3 <- lm(Y~X,dat,subset=(CODES%in%c(0,3)))
summary(fit3)
abline(fit3,col=4,lty=4)

influence.measures(fit0)
influence.measures(fit1)
influence.measures(fit2)
influence.measures(fit3)
```