TD 7 : modèle linéaire, suite

Les questions marquées d'un astérisque **(*)** *sont facultatives.*

Exercice 1. (Tout sur la régression linéaire) On dispose d'un jeu de données composé de 90 couples (X_i, Y_i) représenté dans la Figure 1. On utilise la commande `lm` de R pour estimer les paramètres du modèle linéaire gaussien homoscédastique $Y = \beta_0 + X\beta_1 + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{90})$ à partir de ce jeu de données.

L'enjeu de cet exercice est de comprendre et vérifier le résultat de la commande R suivant :

```
Residuals:
     Min       1Q  Median      3Q      Max
 -97.434  -34.711   0.982  24.288  112.944

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)    77.563      41.361    1.875    0.0641  .
           X    -2.929       1.237   -2.368    0.0201  *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.71 on 88 degrees of freedom
Multiple R-squared: 0.05989, Adjusted R-squared: 0.0492
```

On note $\hat{Y} = \hat{\beta}_0 + X\hat{\beta}_1$ les valeurs prédites de Y, $e = Y - \hat{Y}$ les résidus et $\hat{\sigma}$ l'estimateur non biaisé de la variance des ε. On note $\overline{X}$ la moyenne empirique des X_i, et de même pour les autres quantités. On a :

$$\overline{X} = 33.176, \quad \overline{Y} = -19.607 = \overline{\hat{Y}},$$
$$\overline{X^2} = 1117.845, \quad \overline{Y^2} = 2852.297,$$
$$\overline{XY} = -700.9204, \quad \overline{Y\hat{Y}} = 532.212 = \overline{\hat{Y}^2}.$$

1. (a) Notons $\mathbb{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_{90} \end{pmatrix}$. Calculer $\mathbb{X}^\top Y$.
   (b) Démontrer l'égalité $\overline{Y} = \overline{\hat{Y}}$.
   (c) Démontrer l'égalité $\overline{Y\hat{Y}} = \overline{\hat{Y}^2}$.
2. Le premier tableau de sortie est titré « `Residuals` ». Que contient ce tableau ? Confirmez-vous les valeurs fournies ?
3. On regarde l'avant-dernière ligne de la sortie (« `Residual standard error` » etc. pour « écart-type des résidus »)
   (a) Rappelez la formule de l'estimateur non biaisé de la variance σ^2 des ε_i et sa loi.
   (b) À quoi font référence les mots « `on 88 degrees of freedom` » ?
   (c) Confirmez-vous la valeur fournie dans la ligne « `Residual standard error` » ?
4. Le second tableau, titré « `Coefficients` », contient les informations relatives à l'estimateur des moindres carrés $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$.
   (a) **(*)** Vérifiez les valeurs estimées pour $\hat{\beta}_0$ et $\hat{\beta}_1$ (colonne « `Estimate` »).
   (b) Rappeler la loi de $(\hat{\beta}_0, \hat{\beta}_1)$.

(c) Que contient la colonne « `Std. Error` » ? Confirmez-vous les valeurs fournies ?

Les deux dernières colonnes (« `t value` » et « `Pr(>|t|)` ») servent à tester si le coefficient β_0 (ou β_1, selon la ligne) est nul.

(d) **(*)** Que contient la colonne « `Pr(>|t|)` » ? *(Vous pouvez commencez par répondre aux deux questions suivantes avant de revenir sur celle-ci.)*

(e) Effectuer un test de niveau 5% de H0 : $\beta_0 = 0$ contre H1 : $\beta_0 \neq 0$. On fera bien attention à préciser la formule de la statistique de test, sa loi sous H0, et la valeur observée. Les valeurs et la conclusion obtenues sont-elles conformes aux données du tableau ? (Le quantile d'ordre 97.5% de la loi de Student à 88 degrés de liberté est 1.987.)

(f) Même question pour β_1.

5. (a) Rappeler la définition du coefficient de détermination R^2. Confirmez-vous la valeur fournie dans la ligne « Multiple R-squared » ?

   (b) Le coefficient R^2 ajusté associé à un modèle à n observations et p variables explicatives (terme constant exclus, donc ici $n = 90$ et $p = 1$) est défini par $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$. Cette quantité est-elle croissante ou décroissante en p ? Confirmez-vous la valeur fournie ?

   Interprétation : plus R^2 est proche de 1, plus les valeurs prédites pour Y collent aux valeurs observées. Ajouter plus de variables explicatives fait toujours augmenter R^2, ce qui peut amener à sur-apprendre. Le coefficient ajusté vise à compenser ce sur-apprentissage en pénalisant les grandes valeurs de p.

6. Etant donné une nouvelle valeur $X_{91} = 50$, proposer un intervalle de confiance pour la valeur débruitée $\beta_0 + X_{91}\beta_1$ de l'observation Y_{91} associée.

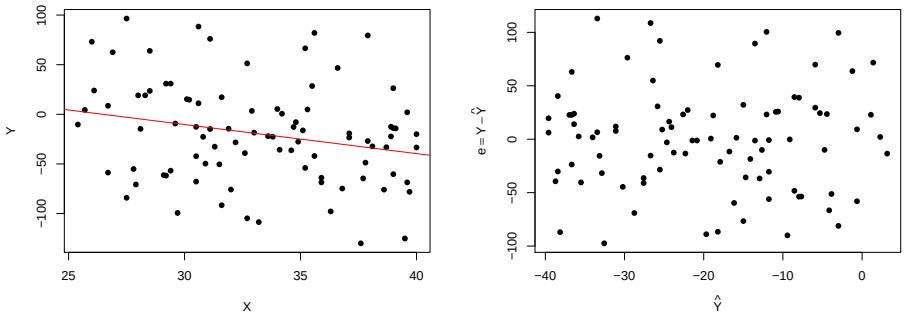
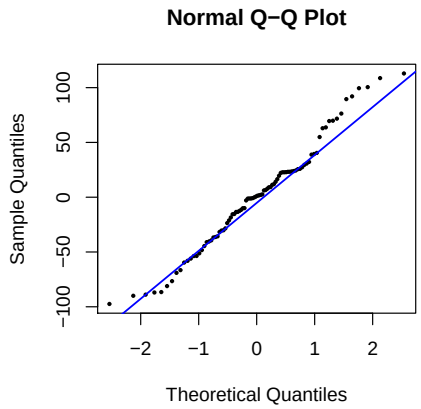


FIGURE 1 – Jeu de données et droite de régression (gauche) et résidus (droite).



indice	1	...	23	24	...	45	46
valeur	-97.434	...	-35.701	-31.741	...	0.592	1.371

indice	...	67	68	...	90
valeur	...	23.873	24.426	...	112.944

FIGURE 2 – Graphe quantile-quantile des résidus comparé à une loi normale (gauche), table des résidus triés par ordre croissant (droite).

Exercice 2. On dispose d'un jeu de données composé de 75 couples (X_i, Y_i) représentés dans les Figures 3 et 4. On commence par essayer le modèle linéaire gaussien homoscédastique $Y = \beta_0 + X\beta_1 + \varepsilon$. La commande R associée nous fournit les informations suivantes.

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
 -5.665  -3.240  -1.286   2.309  13.548

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)    6.3618      0.5274    12.06    <2e-16  ***
           X   -3.4584      0.2408   -14.36    <2e-16  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.389 on 73 degrees of freedom
Multiple R-squared: 0.7386, Adjusted R-squared: 0.7351
```

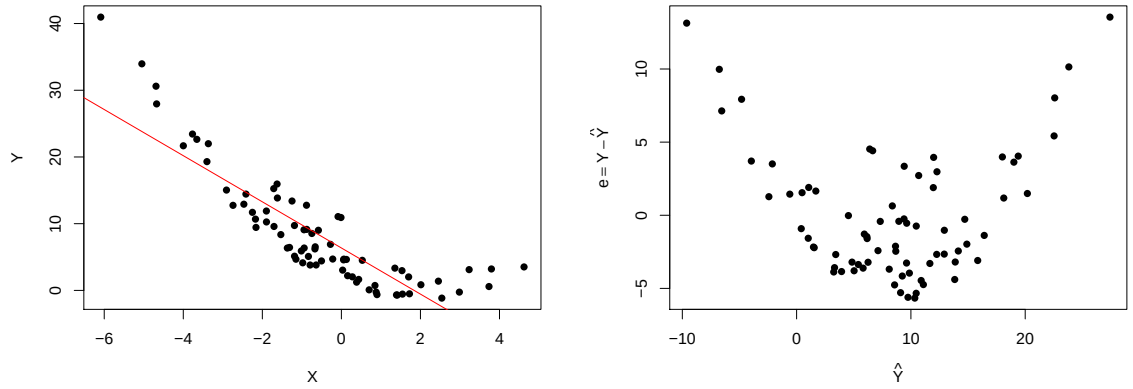


FIGURE 3 – Jeu de données et droite de régression (gauche) et résidus (droite) pour le modèle linéaire simple $Y = \beta_0 + X\beta_1 + \varepsilon$.

1. Commentez : le modèle vous semble-t-il approprié ?

On décide d'ajouter un terme quadratique au modèle pour obtenir le modèle gaussien homoscédastique $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$. La commande R associée nous fournit les informations suivantes.

```
Call:
lm(formula = Y ~ X + X2)

Residuals:
     Min       1Q   Median       3Q      Max
 -3.3736  -1.7700  -0.8768   1.5919   6.7664

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   4.13001     0.34542    11.96    <2e-16  ***
           X  -2.77213     0.14594   -19.00    <2e-16  ***
          X2   0.55188     0.04379    12.60    <2e-16  ***
```

```
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.468 on 72 degrees of freedom
Multiple R-squared: 0.9185, Adjusted R-squared: 0.9162
```

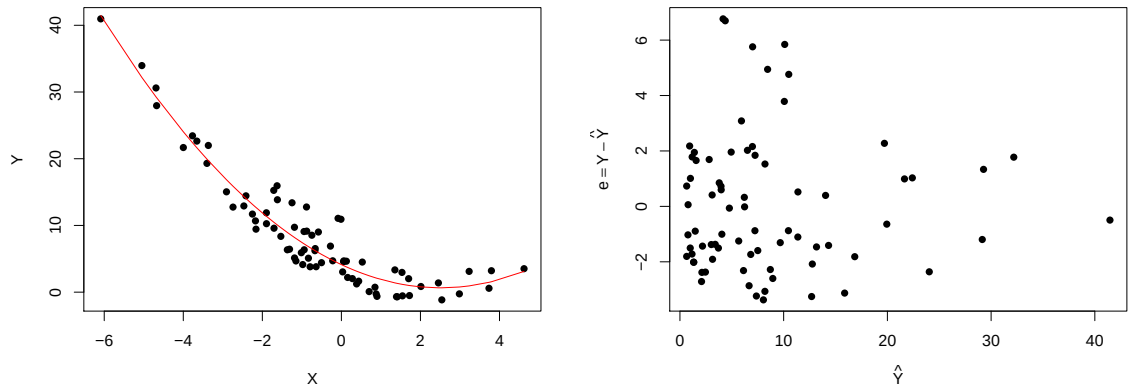


FIGURE 4 – Jeu de données et courbe de régression (gauche) et résidus (droite) pour le modèle $Y = \beta_0 + X\beta_1 + X^2\beta_2 + \varepsilon$.

2. Comparer ce modèle au modèle précédent. Vous semble-t-il mieux adapté aux données ?

On décide de voir ce qui se passe en allant encore plus loin dans les puissances de X, et on considère le modèle gaussien homoscédastique $Y = \beta_0 + X\beta_1 + X^2\beta_2 + X^3\beta_3 + \varepsilon$. La commande R associée nous fournit les informations suivantes.

```
Call:
lm(formula = Y ~ X + X2 + X3)

Residuals:
     Min       1Q   Median       3Q      Max
 -3.4006  -1.7515  -0.8935   1.5222  6.8025

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)  4.093358    0.374552   10.929   < 2e-16  ***
           X -2.823281    0.243706  -11.585   < 2e-16  ***
          X2  0.560178    0.054190   10.337  8.39e-16  ***
          X3  0.004057    0.015423    0.263     0.793
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.484 on 71 degrees of freedom
Multiple R-squared: 0.9186, Adjusted R-squared: 0.9151
```

3. Commentez ce résultat. Ce modèle est-il plus approprié que les précédents ?