L2 MIASHS
Année 2021-22

Statistiques Descriptives 2

TD 1 de Statistiques Descriptives 2

Exercice 1 Un père a deux garçons, et s'inquiète de la croissance de son cadet qu'il trouve petit. Il décide de faire un modèle familial à partir des mesures de taille en fonction de l'âge de l'aîné :

age	3	4	5	6	7	8	9	10	11	12
taille	96	104.8	110.3	115.3	121.9	127.4	130.8	136	139.7	144.5

i. Représenter les données sur un graphique et justifier visuellement l'utilisation d'un modèle de régression linéaire simple. Discuter les hypothèses nécessaires dans le cas où on souhaite modéliser par un modèle linéaire Gaussien.

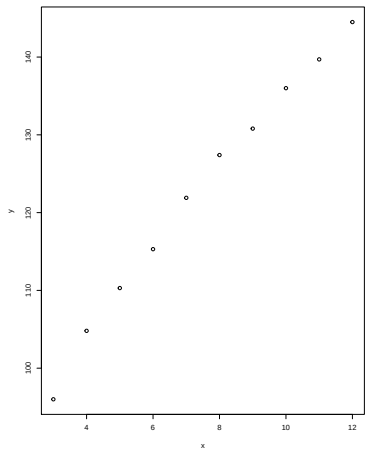

Figure 1 – Représentation des données

Le modèle linéaire gaussien demande que les données $(x_1, Y_1), \ldots, (x_n, Y_n)$ soient indépendantes et $Y_i \sim \mathcal{N}(ax_i + b, \sigma^2)$.

ii. Estimer les coefficients de la régression et tracez sur le graphique la droite de régression estimée.

Réponse : En utilisant les formules du cours,

$$\widehat{a} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

et

$$\widehat{b} = \bar{y}_n - \widehat{a}\bar{x}_n,$$

on trouve que le coefficient directeur est

$$\hat{a} = 5.22$$

et l'ordonnée à l'origine

$$\hat{b} = 83.52.$$

Les commandes sous R sont les suivantes :

```
install.packages("ggplot2")
library("ggplot2")
x = c(3,4,5,6,7,8,9,10,11,12)
y = c(96,104.8,110.3,115.3,121.9,127.4,130.8,136,139.7,144.5)
res = lm(y ~ x)
res$coefficients
plot(x,y)
abline(res$coefficients)
```

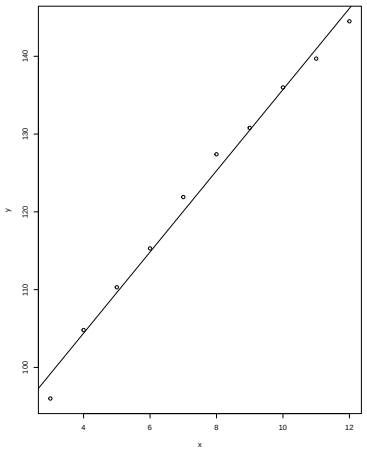

FIGURE 2 – Représentation des données superposées à la droite de régression empirique

iii. Représenter les résidus. La régression semble-t-elle valable ?

Réponse : Les résidus sont

$$-3.18 \ 0.40 \ 0.68 \ 0.46 \ 1.84 \ 2.12 \ 0.30 \ 0.28 \ -1.24 \ -1.66$$

et la représentation des résidus donne

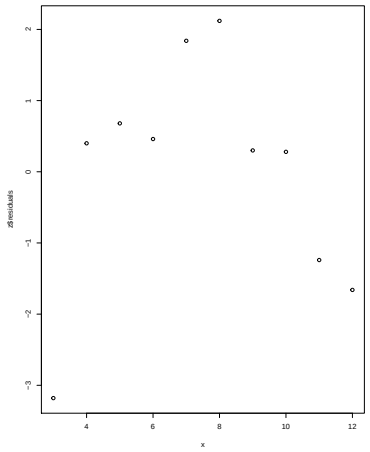

FIGURE 3 – Représentation des résidus

A l'affichage des résidus on remarque qu'ils sont majoritairement positifs et les signes n'alternent pas, ce qui contredit intuitivement l'indépendance de ces résidus. De plus, les résidus semblent croître jusqu'a une valeur maximale puis redescendre dans les valeurs négatives, ce qui renforce notre hypothèse que les résidus sont trop structurés pour être du "bruit". On peut peut etre tenter un modèle quadratique pour ces données.

Pour information, les données proviennent des études auxologique du Docteur Sempé dont une partie a été publiée par Abidi et al (1996). Ces données mesurées sur des milliers d'enfants (de 1 mois à 19 ans) ont permis d'établir un modèle de croissance humaine qui fournit les prédictions du carnet de santé. Il s'écrit de la manière suivante :

$$Y = \theta_1 \left[1 - \frac{1}{1 + \left((x + \theta_8) / \theta_2 \right)^{\theta_3} + \left((x + \theta_8) / \theta_4 \right)^{\theta_5} + \left((x + \theta_8) / \theta_6 \right)^{\theta_7}} \right]$$

où θ_1 représente la taille adulte, θ_x le temps de grossesse, et les couples $(\theta_2, \theta_3), (\theta_4, \theta_5)$ et (θ_6, θ_7) permettent de modéliser respectivement la phase de croissance initiale (juste après la naissance), la phase de croissance centrale (pré-adolescente) et la phase finale.

Exercice 2 La tableau suivant contient la liste de 14 pays d'Amérique du Nord et d'Amérique Centrale, dont la population dépassait le million d'habitants en 1985.

Observations	pays	taux d'urbanisation	taux de natalité
1	Canada	55.0	16.2
2	costa-Rica	27.3	30.5
3	Cuba	33.3	16.9
4	USA	56.5	16.0
5	El Salvador	11.5	40.2
6	Guatemala	14.2	38.4
7	Haiti	13.9	41.3
8	Honduras	19.0	43.9
9	Jamaique	33.1	28.3
10	Mexique	43.2	33.9
11	Nicaragua	28.5	44.2
12	Trinidade/Tobago	6.8	24.6
13	Panama	37.7	28.0
14	Rep. Dominicaine	37.1	33.1

Pour chaque pays, on mesure le taux de natalité y_i (nombre de naissances annuel pour 1000 habitants) ainsi que le taux d'urbanisation x_i (pourcentage de la population vivant dans des villes de plus de 100000 habitants). On fait l'hypothèse d'un modèle de régression linéaire simple du type $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, c'est-d-dire que le taux de natalité dépend linéairement du taux d'urbanisation.

1. Représenter graphiquement les données.

   Voila les commandes R pour récupérer les données :

```
data2 = cbind(c(55.0,16.2),
c(27.3,30.5),
c( 33.3 , 16.9),
c( 56.5 , 16.0),
c( 11.5 , 40.2),
c( 14.2 , 38.4),
c( 13.9 , 41.3),
c( 19.0 , 43.9),
c( 33.1 , 28.3),
c( 43.2 , 33.9),
c( 28.5 , 44.2),
c( 6.8 , 24.6),
c( 37.7 , 28.0),
c( 37.1 , 33.1))

y = data2[2,]
x = data2[1,]
res = lm(y   x)
res$coefficients
plot(x,y)
abline(res$coefficients)

colnames(data2) <- c("observations", "pays", "urbanisation", "natalite")
```

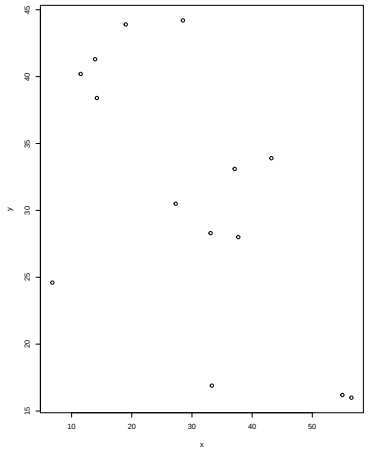

FIGURE 4 – Représentation des données

2. Estimer les paramètres β_0 et β_1 du modèle et tracer la droite de régression correspondante.

   Réponse : les coefficients sont donnés par

$$\hat{\beta}_1 = -0.3988675$$

$$\hat{\beta}_0 = 42.9905457$$

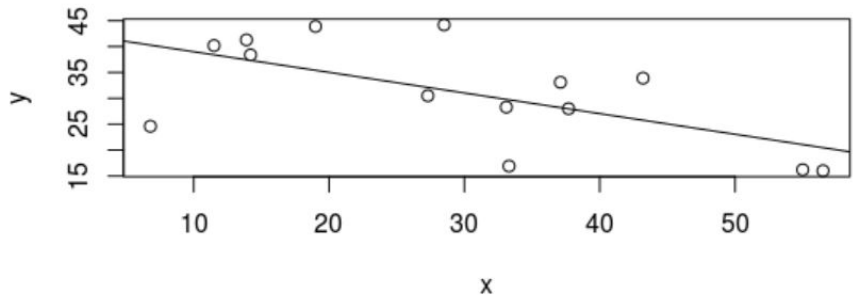

FIGURE 5 – Représentation de la droite de régression et des données

3. Calculer la somme des carrés des résidus.

   Réponse : les résidus sont donnés par

$$-4.85 \ -1.60 \ -12.80 \ -4.45 \ 1.79 \ 1.07 \ 3.85 \ 8.49 \ -1.49 \ 8.14 \ 12.58 \ -15.68 \ 0.047 \ 4.91$$

   La somme des carrés des résidus est égale à 797.8449.

   On affiche les résidus dans la Figure 6.

— En résumé, on peut utiliser la fonction

   > summary(res)

   qui donne :

   Call : lm(formula = y x)
   Residuals : Min 1Q Median 3Q Max -15.6782 -3.7413 0.5601 4.6440 12.5772
   Coefficients : Estimate Std. Error t value Pr(>|t|)

(Intercept)	42.9905	4.8454	8.872	1.28e-06 ***
x	-0.3989	0.1453	-2.746	0.0177 *

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

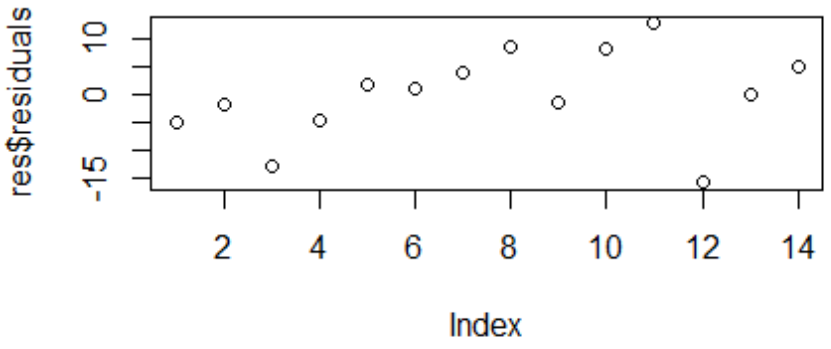

FIGURE 6 – Caption

Residual standard error : 8.154 on 12 degrees of freedom Multiple R-squared : 0.3859, Adjusted R-squared : 0.3347 F-statistic : 7.54 on 1 and 12 DF, p-value : 0.01774

Cela nous informe en particulier l'importance des coefficients au moyen des $p-$values : nous verrons en cours de statistiques inférentielles en L3 que plus la p-value est forte, plus on peut penser que le coefficient auquel elle est associée est négligeable. Lorsque β_1 est négligeable, on peut en déduire que "x n'a pas d'influence sur y".