Devoir nº 2*

Julien HÉBERT-DOUTRELOUX Alexandre PACHOT

17 février 2020

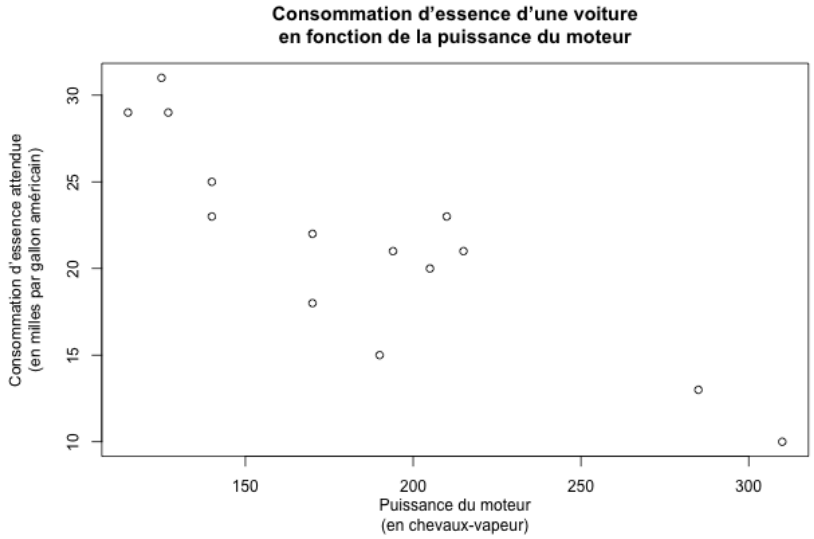
Table des matières

1 Consommation d'essence

a) Graphique du nuage de points

En observant le nuage de points de la figure 1, on remarque une corrélation négative entre la consommation d'essence et la puissance du moteur : plus un moteur est puissant, moins la voiture peut parcourir de kilomètres avec une même quantité d'essence. La relation entre ces deux variables semble linéaire. On ne peut pas considérer la relation entre ces deux variables comme forte, notamment pour la fourchette allant de 180 à 225 chevaux. D'autre part, on peut remarquer qu'il y a plus de données pour les voitures ayant un moteur entre 100 et 225 chevaux-vapeur (12 données) que pour les voitures ayant entre 280 et 320 chevaux (2 données). On remarque également qu'il n'y a pas données pour les voitures ayant entre 225 et 280 chevaux.

FIGURE 1 – Graphique du nuage de points



Code R :

```
efficacite = read.csv("~/donnees/efficacite.txt", sep="")
png("~/images/d2_essencePlot.png", width = 645, height = 425)
par(mar=c(5, 6, 4, 2) + 0.1)
plot(
  efficacite,
  main = "Consommation d'essence d'une voiture\n
    en fonction de la puissance du moteur",
```

*STT 1700 - Introduction à la statistique - Université de Montréal - Hiver 2020 - Christian LÉGER

```
  xlab = "Puissance du moteur\n(en chevaux-vapeur)",
  ylab = "Consommation d'essence attendue\n
    (en milles par gallon américain)")
dev.off()
```

b) Corrélation

La corrélation entre la puissance du moteur et la consommation d'essence est de $-0,8779545$.

Code R :

```
> cor(efficacite$chevaux.vapeur, efficacite$mpg)
[1] -0.8779545
```

c) Transformations linéaires

Lorsqu'une voiture fait x miles avec 1 gallon d'essence, elle fait $x \times 1,609$ km avec 3,78541 litres. Si $f(x)$ est la consommation d'essence exprimée en kilomètres par litres, alors :

$$f(x) = \frac{1,609}{3,78541}x$$

Lorsqu'une voiture fait x kilomètres avec 1 litre d'essence, il lui faut $1/x$ litre pour un kilomètre, et $100/x$ litres pour 100 kilomètres. Si $g(x)$ est la consommation d'essence exprimée en litres par 100 kilomètres, alors :

$$g(x) = \frac{100}{f(x)} = \frac{378,541}{1,609x}$$

$f(x)$ est une transformation linéaire car :

$$f(kx) = k.f(x)$$

$$f(x+y) = f(x) + f(y)$$

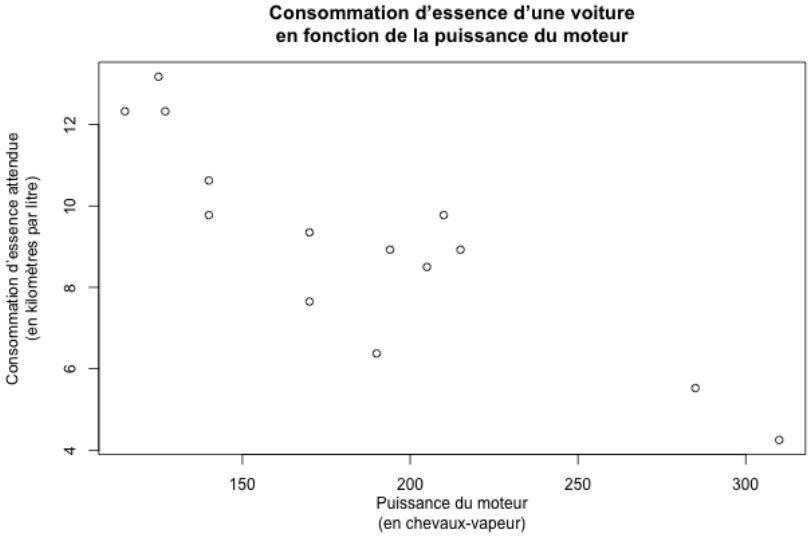
Par contre $g(x)$ n'est pas une transformation linéaire. En effet :

$$g(x+y) \neq g(x) + g(y)$$

d) Consommation d'essence en kilomètres par litre

Le nuage de points de la consommation d'essence en kilomètres par litre est représenté à la figure 2. Il est exactement identique à celui de la figure précédente, ce qui est normal, car on a seulement multiplié les données précédentes par une constante ($1,609/3,78541 \approx 0,43$), ce qui fait qu'on multiplié l'échelle des ordonnées par cette constante.

FIGURE 2 – Consommation d'essence en kilomètres par litre



La corrélation est égale -0,8779545. Elle est exactement la même que pour le nuage de points précédents, ce qui est normal, car la seule transformation qui a été faite a été la multiplication par une constante, ce qui ne modifie en rien le lien qu'il y a entre deux variables.

Code R :

```
> efficacite$kmpl = 1.609 / 3.78541 * efficacite$mpg
> plot(
    x = efficacite$chevaux.vapeur,
    y = efficacite$kmpl,
```

```
    main = "Consommation d'essence d'une voiture\n
      en fonction de la puissance du moteur",
    xlab = "Puissance du moteur\n(en chevaux-vapeur)",
    ylab = "Consommation d'essence attendue\n
      (en kilomètres par litre)")
> cor(efficacite$chevaux.vapeur, efficacite$kmpl)
[1] -0.8779545
```

e) Consommation d'essence en litre par 100 kilomètres

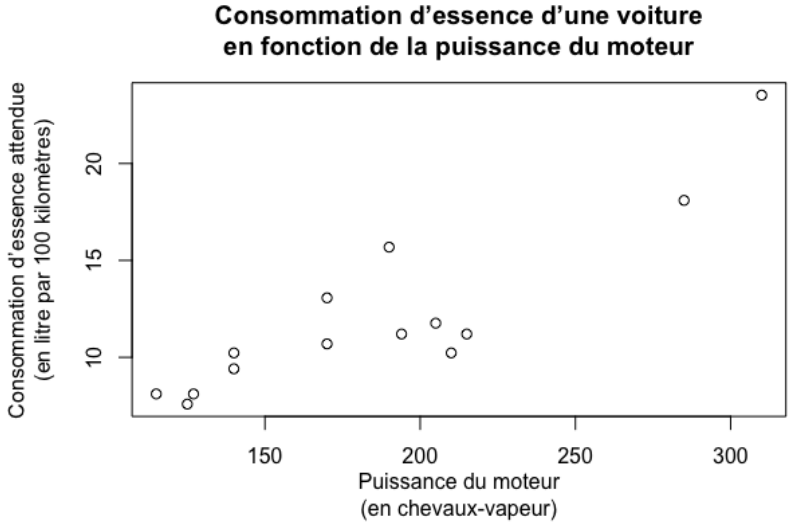
La direction du nuage de points a changé (fig. 3), par contre sa forme et sa force sont restées approximativement les mêmes. Étant donné qu'il s'agit d'une transformation non linéaire, il est normal que la forme du nuage ne soit pas identique. La valeur de la corrélation est de 0,8887396, ce qui en valeur absolue est similaire à -0,8779545. Le changement de signe de la corrélation reflète le changement de direction, la non-égalité de deux corrélations est la conséquence de la transformation non linéaire. Le fait que la forme et la force sont approximativement les mêmes, cela est représenté par les deux corrélations proches en valeur absolue.

On peut donc conclure qu'une transformation non linéaire modifie le nuage de points, et poser comme hypothèse qu'une transformation linéaire ne modifie pas le nuage de points. On démontre assez facilement que si l'on multiplie tout les y_i par un facteur k, on multiplie la moyenne et l'écart-type par ce même facteur et que cela n'a aucune incidence sur la corrélation.

Code R :

```
> efficacite$lp100km = 378.541 / (1.609 * efficacite$mpg)
> plot(
    x = efficacite$chevaux.vapeur,
    y = efficacite$lp100km,
    main = "Consommation d'essence d'une voiture\n
      en fonction de la puissance du moteur",
    xlab = "Puissance du moteur\n(en chevaux-vapeur)",
    ylab = "Consommation d'essence attendue\n
      (en litre par 100 kilomètres)")
> cor(efficacite$chevaux.vapeur, efficacite$lp100km)
[1] 0.8887396
```

FIGURE 3 – Consommation d'essence en litre par 100 kilomètres



2 ExamensA19

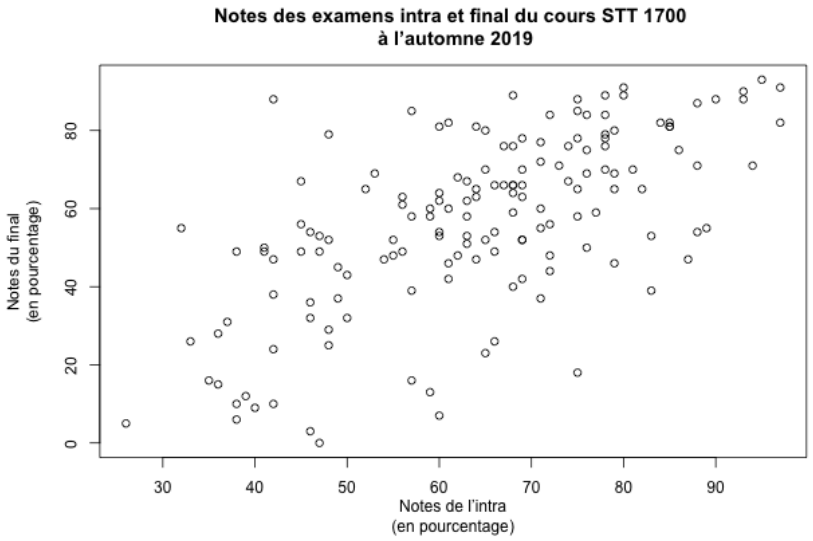
a) Graphique du nuage de points

Lorsqu'on observe le nuage de points (fig. 4), il semble plutôt avoir une corrélation positive entre les deux notes. En effet, les notes s'alignent plutôt par rapport à la première bissectrice. Il n'y a aucun étudiant qui a eu une note inférieure à 40 à l'intra et supérieur à 60 au final. De même, il n'y a aucun étudiant qui a eu une note supérieure à 80 à l'intra et une note inférieure à 40 au final. On peut également noter que les notes de l'intra se répartissent majoritairement entre 30 et 100 %, alors que les notes du final sont plus entre 0 et 90 %.

Code R :

```
examensA19 = read.delim("~/donnees/examensA19.txt")
png("~/images/d2_examPlot.png", width = 645, height = 425)
par(mar=c(5, 6, 4, 2) + 0.1)
plot(
```

FIGURE 4 – Graphique du nuage de points




```
  examensA19,
  main = "Notes des examens intra et final du cours STT 1700
    \n à l'automne 2019",
  xlab = "Notes de l'intra\n(en pourcentage)",
  ylab = "Notes du final\n(en pourcentage)")
dev.off()
```

b) Équation de la droite des moindres carrés

L'équation de la droite des moindres carrés de la note du final par rapport à celle de l'intra est $y = 0,9043x - 1,2567$. C'est-à-dire que quelqu'un qui a eu 100 % à l'intra, devrait avoir 89 % au final.

Code R :

```
> lm(Final~Intra, data=examensA19)
(Intercept)        Intra
    -1.2567       0.9043
```

c) Variation expliquée par la corrélation

Le carré du coefficient de corrélation r^2 est la fraction de y *expliquée* par la régression des moindres carrés de y sur x.

Le coefficient de corrélation entre les deux variables est 0,6374878. La proportion de la variation totale des notes du final expliquée par la régression du final par l'intra est de 40 %.

Code R :

```
> cor(examensA19$Intra,examensA19$Final)
[1] 0.6374878
> 0.6374878 * 0.6374878
[1] 0.4063907
```

d) Prévision

La droite de régression permet de faire des prévisions pour des valeurs de x observées. Soit x_0 la valeur observée, la prévision est $y_0 = \hat{b}x_0 + \hat{a}$. Où $\hat{a}$ et $\hat{b}$ sont les coefficients de la droite des moindres carrés.

Ainsi, un étudiant qui a 50 % à l'intra, devrait avoir 44 % $(0,9043 \times 50 + (-1.2567))$ au final. La prévision pour le final monte à 53 % lorsque la note obtenue à l'intra est de 60 %.

Code R (vérification) :

```
> examensA19.lm = lm(Final~Intra, data=examensA19)
> predict(examensA19.lm, data.frame(Intra=c(50,60)))
       1        2
43.95855 53.00160
```

3 Hypertension

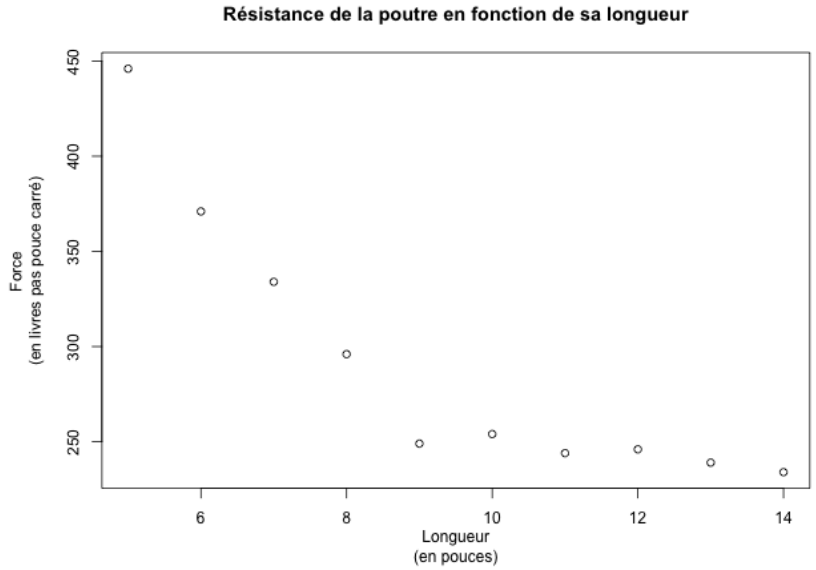
Les personnes qui ont un problème d'hypertension ont une pression artérielle beaucoup plus élevée que la normale. Afin de réduire ce problème de pression, ils utilisent des médicaments contre l'hypertension. Cela réduit leur pression, néanmoins, elle reste plus élevée que la normale. Ce n'est pas parce que leur pression est plus élevée que cela est causé par les médicaments contre l'hypertension. Sans eux, leur pression serait largement plus élevée !

4 Résistance du bois

a) Graphique du nuage de points

Le graphique du nuage de points pour étudier comment la longueur de la poutre affecte sa résistance à une force donnée est représenté à la figure 5.

FIGURE 5 – Graphique du nuage de points



Code R :

```
bois <- read.csv("~/donnees/bois.txt", sep="")
png("~/images/d2_boisPlot.png", width = 650, height = 460)
par(mar=c(5, 6, 4, 2) + 0.1)
plot(
  bois,
  main = "Résistance de la poutre en fonction de sa longueur",
  xlab = "Longueur\n(en pouces)",
  ylab = "Force\n(en livres pas pouce carré)")
dev.off()
```

b) Allure générale du graphique

La courbe à la forme d'une exponentielle négative (e^{-x}) : plus la poutre est longue, moins elle est résistante. Bien qu'une poutre de 10 pouces soit plus résistante qu'une poutre de 9 pouces et qu'une poutre de 12 pouces soit plus résistante qu'une poutre de 11 pouces, il n'y a pas de données qui semblent aberrantes. Le nuage de points représente la courbe de flambage des poutres en copeaux de bois.

c) Régression des moindres carrés

c.1) Calcul de la moyenne

$$\begin{aligned}\bar{x} &= (1/n)\sum_{i=1}^{n} x_i \\ &= (1/10)(5+\ldots) \\ &= 9,5 \\ \bar{y} &= 291,3\end{aligned}$$

Code R (vérification) :

```
> mean(bois$longueur)
[1] 9.5
> mean(bois$force)
[1] 291.3
```

c.2) Calcul de l'écart-type (de l'échantillon)

Formule classique

$$\begin{aligned}s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= \frac{1}{9}\{(5-9,5)^2+\ldots\} \\ s_x &= 3,02765 \\ s_y &= 71,19465\end{aligned}$$

Méthode « de la calculatrice »

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right\} \\ &= \frac{1}{9}\left\{985 - \frac{1}{10}(95)^2\right\} \\ s_x &= 3,02765 \\ s_y &= 71,19465 \end{aligned}$$

Code R (vérification) :

```
> sd(bois$longueur)
[1] 3.02765
> sd(bois$force)
[1] 71.19465
```

c.3) Calcul de la corrélation

Formule classique

$$\begin{aligned} r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \\ &= \frac{1}{9}\left\{\left(\frac{5-9,5}{3,02765}\right)\left(\frac{446-291,3}{71,19465}\right)+\ldots\right\} \\ &= -0,882229 \end{aligned}$$

Méthode « de la calculatrice »

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n} x_i y_i - (1/n)\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{(n-1)s_x s_y} \\ &= \frac{25\,962 - (1/10)\times 95 \times 2\,913}{9 \times 3,02765 \times 71,19465} \\ &= -0,882229 \end{aligned}$$

Code R (vérification) :

```
> cor(bois$longueur,bois$force)
[1] -0.882229
```

c.4) Calcul du coefficient directeur

Formule classique

$$\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ &= \frac{25\,962 - 10 \times 9,5 \times 291,3}{985 - 10 \times 9,5^2} \\ &= -20,75 \end{aligned}$$

Méthode « de la calculatrice »

$$\begin{aligned} \hat{b} &= r\frac{s_y}{s_x} \\ &= -0,882229 \times \frac{71,19465}{3,02765} \\ &= -20,75 \end{aligned}$$

c.5) Calcul de l'ordonnée à l'origine

$$\begin{aligned} \hat{a} &= \bar{y} - \hat{b}\bar{x} \\ &= 291,3 - (-20,75) \times 9,5 \\ &= 488,38 \end{aligned}$$

Code R (vérification) :

```
> lm(bois$force ~ bois$longueur)
Coefficients:
  (Intercept)  bois$longueur
       488.38         -20.75
```

c.6) Droite de régression des moindres carrés

Lorsqu'on ajoute la droite de régression des moindres carrés, $y = -20,75x + 488,38$ on obtient la figure 6. On remarque que les résidus sont tous positifs, puis tous négatifs et finalement tous positifs, comme si notre distribution était quadratique. Cette distribution non aléatoire des résidus signifie que la droite n'est pas appropriée.

FIGURE 6 – Droite de régression des moindres carrés

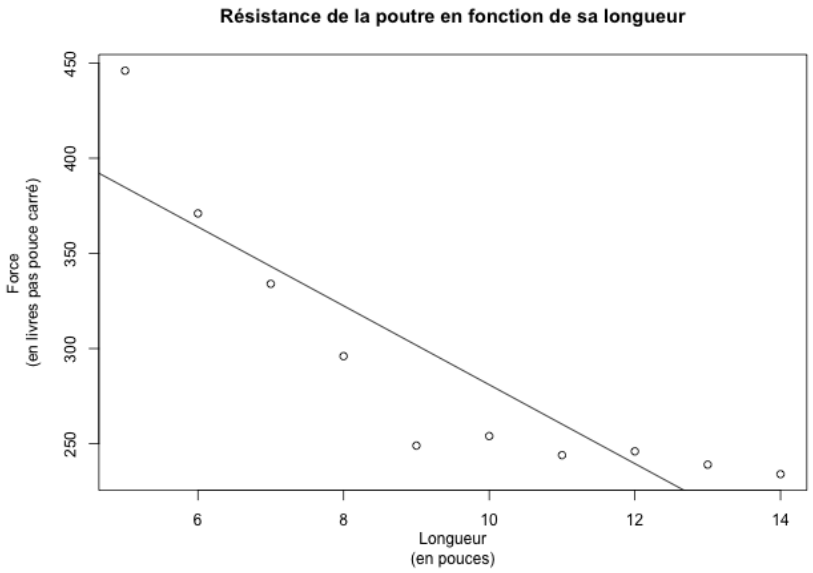
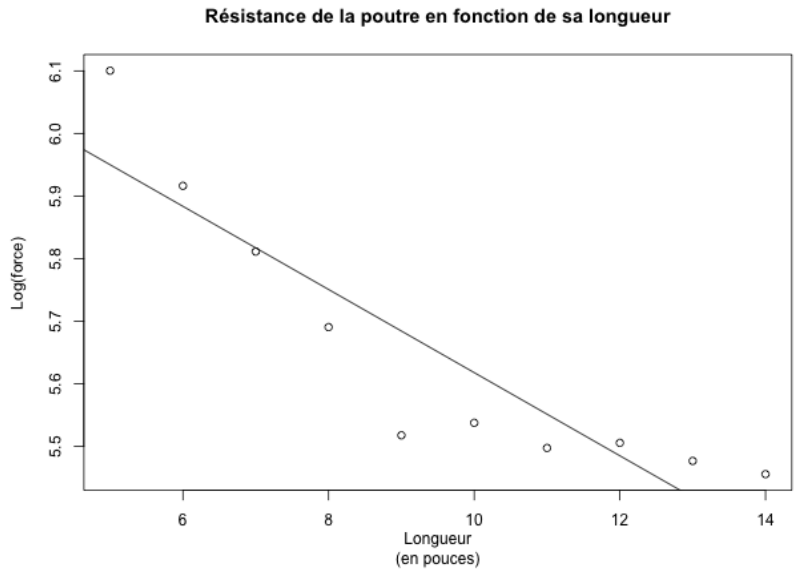


FIGURE 7 – Log(force)



Code R :

```
abline(lm(bois$force ~ bois$longueur))
```

Tout au début, nous avons dit que la courbe avait la forme d'une exponentielle. Qu'en est-il si l'on considère le logarithme de la force ? Dans ce cas-là, on obtient le graphique de la figure 7. Ce qui ne change en rien la distribution des résidus. Essayons plutôt de modaliser le nuage de points par deux segments de droites.

Code R :

```
plot(
  x = bois$longueur,
  y = log(bois$force),
  main = "Résistance de la poutre en fonction de sa longueur",
  xlab = "Longueur\n(en pouces)",
  ylab = "Log(force)")
abline(lm(log(bois$force) ~ bois$longueur))
```

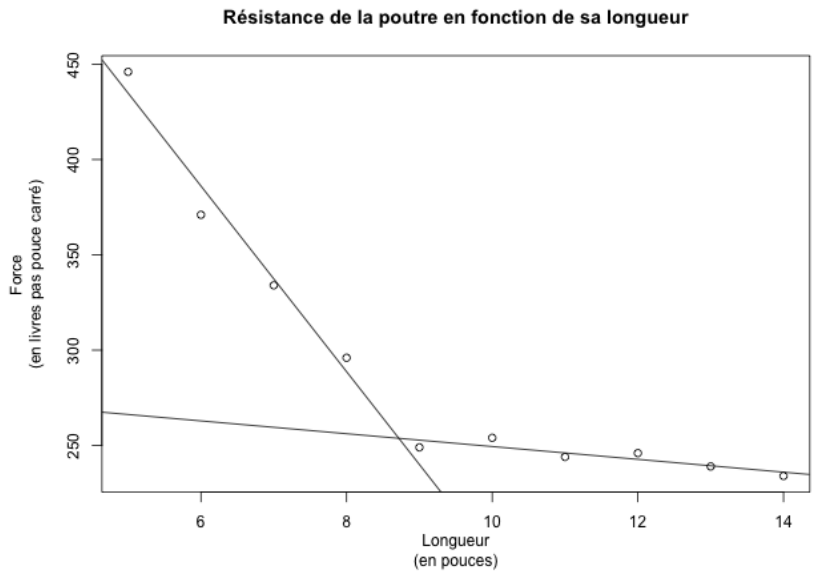
d) Deux droites

La droite de régression des moindres carrés pour les quatre premières valeurs a pour équation $y = -48,7x + 678,3$, et pour les six dernières valeurs, l'équation de cette droite est $y = -3,371x + 283,105$. Lorsqu'on observe la répartition des résidus (fig. 8), elle semble bien aléatoire pour la deuxième partie, et un peu moins pour la première droite où on a un résidu positif, deux résidus négatifs, puis un résidu positif. Il faudrait plus de données afin de juger la pertinence d'une approximation linéaire. Néanmoins, l'approximation est visuellement largement plus satisfaisante que tout ce que nous avons obtenu jusqu'à maintenant.

Nonobstant la question sur la linéarité des quatre premières observations, on peut considérer que ces deux droites décrivent adéquatement les données. On peut se poser la question du rattachement de l'observation 5 (longueur = 9) à un nuage de données plutôt qu'à un autre. Il serait intéressant de demander aux experts du bois s'il y a une modification dans le processus de fabrication des poutres selon la longueur de celles-ci.

De même, il serait intéressant de comparer la résistance des poutres en co-

FIGURE 8 – Deux droites



peaux de bois par rapport aux poutres en bois massif, et cela pour deux raisons. D'une part, afin de savoir qu'elle est la perte de résistance avec des poutres de qualité moindre, et d'autre part afin de vérifier si l'on obtient un phénomène équivalent lors de la conduite des tests de résistance.

Code R :

```
> lm(force~longueur, data=bois, subset=0:4)
  (Intercept)     longueur
      678.3        -48.7
> lm(force~longueur, data=bois[c(5:10),])
  (Intercept)     longueur
    283.105       -3.371
> abline(lm(force~longueur, data=bois, subset=0:4))
> abline(lm(force~longueur, data=bois[c(5:10),]))
```

5 Régression à la moyenne

Quelqu'un qui à l'intra a eu la moyenne du groupe, $\bar{x}$, aura au final la note $0,88\bar{x}+1,09$, c'est-à-dire qu'il aura une note inférieure au final, à moins d'avoir eu une note inférieure à $1,09/(1-0,88)=9,08$ à l'intra. Dans ce cas-là, la note de l'examen final devrait être meilleure que celle de l'intra.

Quelqu'un qui a une note de $\bar{x}+10$ à l'intra, devrait avoir la note $0,88(\bar{x}+10)+1,09$ au final. C'est-à-dire qu'il devrait être à 8,8 points au-dessus de la moyenne de l'examen final.