LMAFY1101 - Solutions - Série 8

Régression et corrélation

Exercice 1

1

Avant d'utiliser un jeu de données il est toujours utile de faire une brève analyse descriptive afin de voir à quoi elles ressemblent et d'éventuellement repérer des caractéristiques particulières.

Nous allons commencer par utiliser les fonctions `str()` et `summary()`. Cela nous permet d'obtenir quelques informations sur les variables, comme, par exemple, leur type, leurs minima et maxima, leur moyenne, etc..

```
str(logement)
```

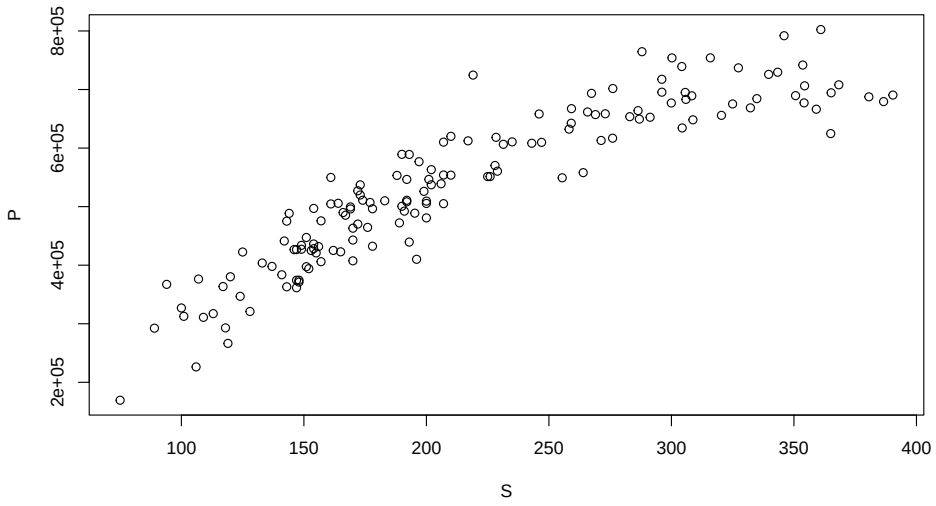
```
'data.frame':   150 obs. of  2 variables:
 $ P: num  426693 509976 397909 609786 589317 ...
 $ S: num  147 183 137 247 190 ...
```

```
summary(logement)
```

```
       P                S
 Min.   :169380   Min.   : 75
 1st Qu.:431988   1st Qu.:154
 Median :532126   Median :196
 Mean   :536278   Mean   :215
 3rd Qu.:655309   3rd Qu.:275
 Max.   :802372   Max.   :390
```

Ensuite, afin de répondre à la question sur le lien linéaire, nous allons tracer le nuage de points des paires (S_i, P_i).

```
plot(P ~ S, data = logement)
```



Nous voyons bien que les points sont “plus ou moins” alignés, cela signifie qu'une tendance linéaire se dégage et que nous pouvons envisager d'utiliser un modèle linéaire pour exprimer la relation entre nos deux variables.

2

```
r <- cor(logement$P, logement$S)
r
```

```
[1] 0.909
```

→ Il existe une forte dépendance linéaire entre la surface et le prix de vente. Le coefficient étant positif, cela signifie que plus la surface de la maison est élevée, plus son prix de vente sera également élevé.

Pour une meilleur interprétation, calculons le r^2:

```
r^2
```

```
[1] 0.827
```

→ Un modèle linéaire simple de type $P \sim S$ permettrai d'expliquer plus que 82% des variabilités observés dans le prix.

3

```
lm(P ~ S, data = logement) |>
  coef()
```

```
(Intercept)           S
     201854        1553
```

L'équation du modèle peut être écrit comme: $\hat{P} = 201854 + 1553 \times S$.

4

- Interprétation de $\hat{\beta}_0 \approx 201854$: Selon notre modèle linéaire, le prix moyen d'une maison avec une surface de 0 m^2 ($S = 0$) est de 201854 euros. Cette interprétation n'est pas pertinente car cela n'a pas de sens de parler d'un bien dont la surface est nulle. En plus, notez que la surface minimale enregistrée est de 75 m^2, ce qui signifie que 0 est une extrapolation (dangereuse) de l'équation linéaire au-delà du domaine observé.

- Interprétation de $\hat{\beta}_1 \approx 1553$: Selon notre modèle linéaire, chaque fois que la surface augmente d'un m^2, le prix augmente en moyenne de 1553 euros.

5

Selon le nouveau modèle, nous avons que

$$\begin{aligned} P &= \alpha_0 + \alpha_1 \times (S - 75) + \epsilon \\ &= (\alpha_0 - 75\alpha_1) + \alpha_1 \times S + \epsilon \end{aligned}$$

Cette dernier équation signifie que $(\hat{\alpha}_0 - 75\hat{\alpha}_1, \hat{\alpha}_1)$ ne peut être que $(\hat{\beta}_0, \hat{\beta}_1)$ l'estimateur de moindres carrés du premier modèle ($P \sim S$), càd.

$$\hat{\alpha}_1 = \hat{\beta}_1 = 1552.544 \text{ et } \hat{\alpha}_0 = \hat{\beta}_0 + 75\hat{\beta}_1 = 318294.6 \text{ et}$$

Ce que nous pouvons vérifier, en R, comme suit

```
lm(P ~ I(S - 75), data = logement) |>
  coef()
```

```
(Intercept)   I(S - 75)
     318295        1553
```

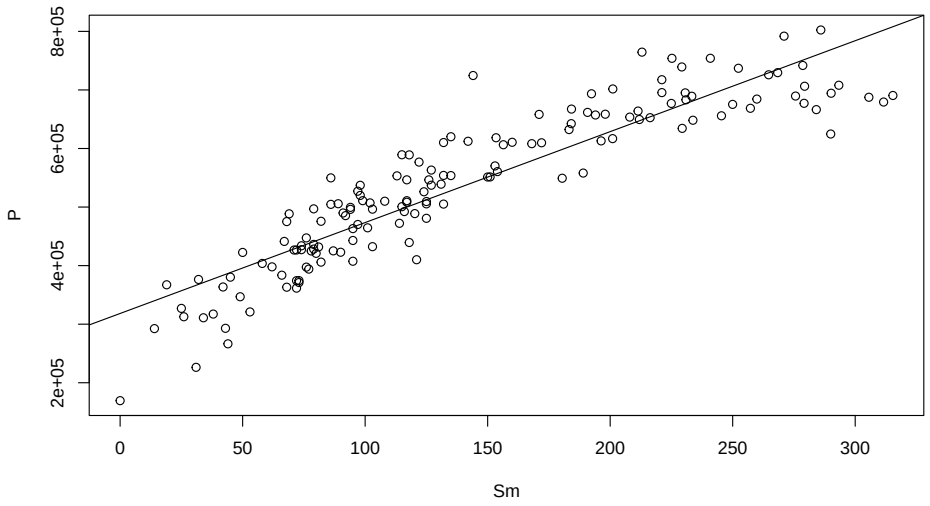
```
# ou
logement <- transform(logement, Sm = S - 75)
msm <- lm(P ~ Sm, data = logement)
msm |>
  coef()
```

```
(Intercept)          Sm
     318295        1553
```

- Interprétation de $\hat{\alpha}_0 \approx 318295$: Selon notre modèle linéaire, le prix moyen d'une maison avec une surface de 75 m^2 ($Sm = 0$) est de 31829 euros.
- L'interprétation de $\hat{\alpha}_1$ est la même que celle de $\hat{\beta}_1$; voir le point (4).

6

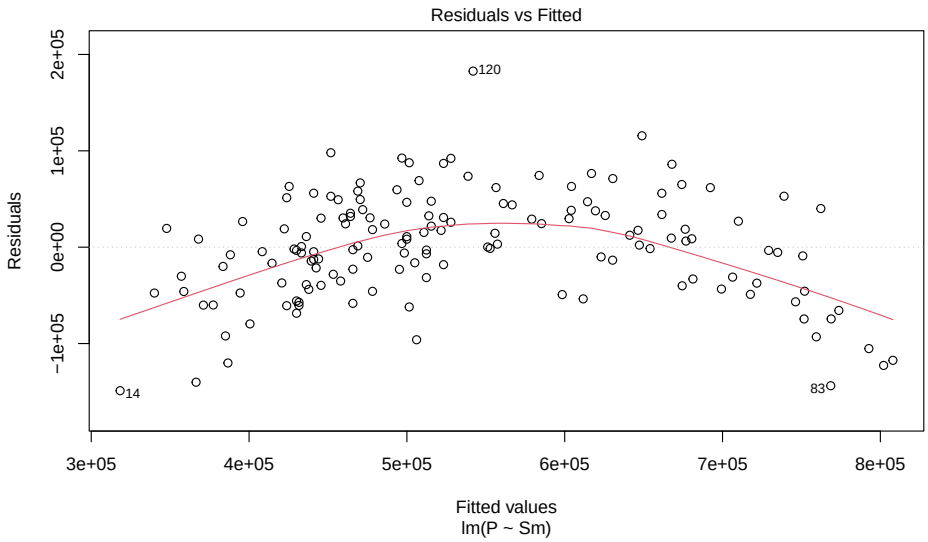
```
plot(P ~ Sm, data = logement)
abline(msm)
```



Globalement, nous pouvons dire que les points sont bien alignés sur la droite de régression. Mais la droite ajustée ne capture pas l'entièreté de la relation entre le prix et la surface. En effet, les observations présentant une petite surface (moins de 100 m^2) et celles avec une grande surface (plus de 250 m^2) semblent s'écarter de la droite.

Pour juger de la qualité d'un modèle, il est préférable d'examiner ces résidus.

```
plot(msm, 1)
```

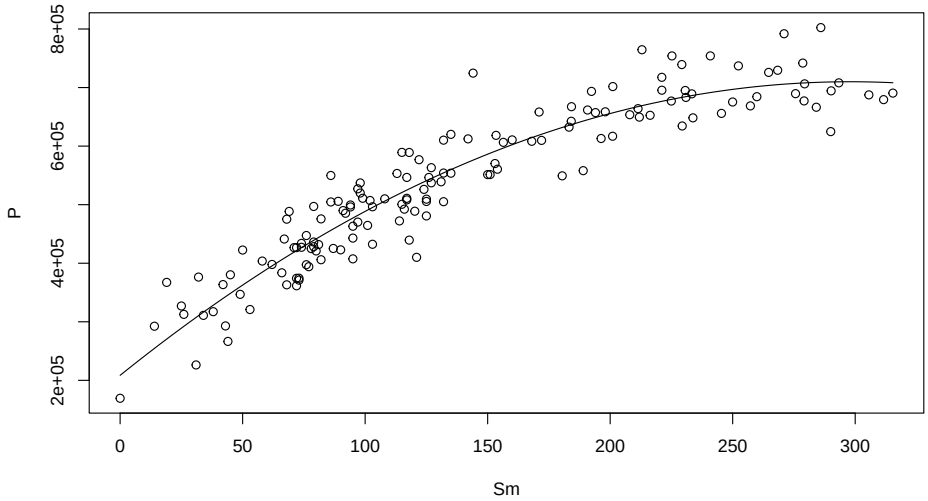


Cette figure renforce notre observation précédente. Nous constatons une structure quadratique dans les résidus ce qui indique un mauvais ajustement de la droite aux données.

7

La forme des résidus suggère l'introduction d'un terme quadratique.

```
msm2 <- lm(P ~ Sm + I(Sm^2), data = logement)
plot(P ~ Sm, data = logement)
curve(208577 + 3364 * x + -5.643 * x^2, add = TRUE)
```



L'équation de notre modèle est donnée par

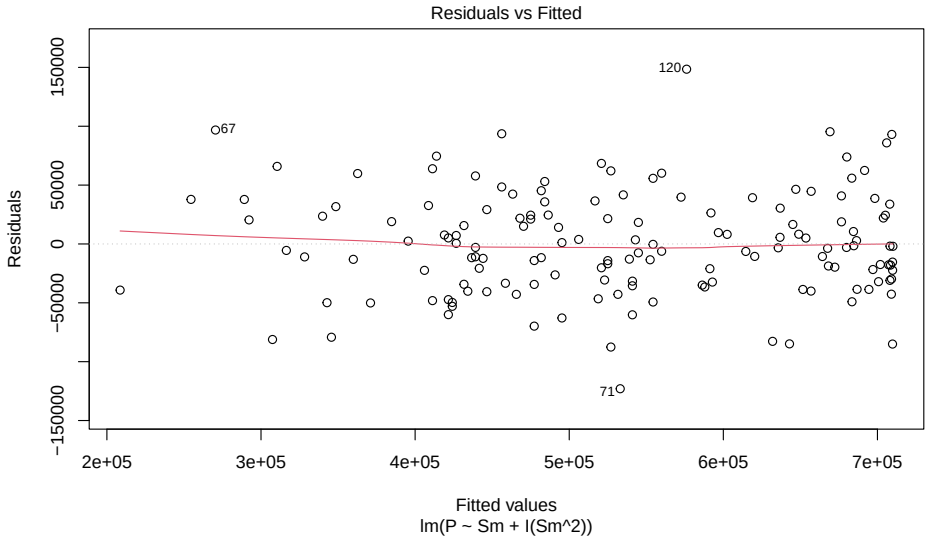
$$\hat{P} = 208577 + 3365 \times (S - 75) - 6 \times (S - 75)^2$$

Examinant la qualité de ce modèle.

```
summary(msm2)$r.squared
```

```
[1] 0.89
```

```
plot(msm2, 1)
```



Ce modèle a un plus grand coefficient de détermination et ses résidus semblent se disperser de manière aléatoire et sans forme particulière de part et d'autre de l'axe $y = 0$. Nous validons ce choix.

Exercice 2

1

```
Cars <- transform(cars, speed = speed * 1.609344, dist = 0.3048 *
  dist)
```

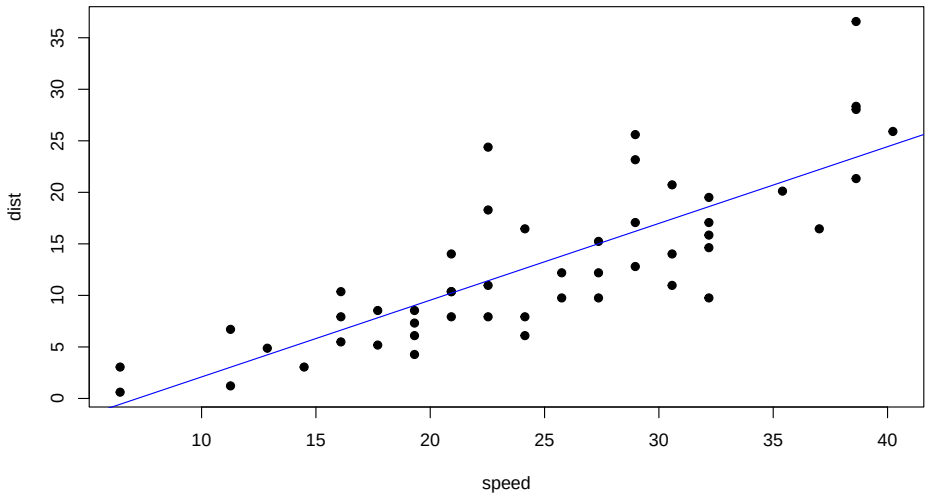
2

```
msp <- lm(dist ~ speed, data = Cars)
msp
```

```
Call:
lm(formula = dist ~ speed, data = Cars)

Coefficients:
(Intercept)        speed
     -5.358        0.745
```

```
plot(dist ~ speed, data = Cars, pch = 19)
abline(msp, col = "blue")
```



3

```
confint(msp)[2, ]
```

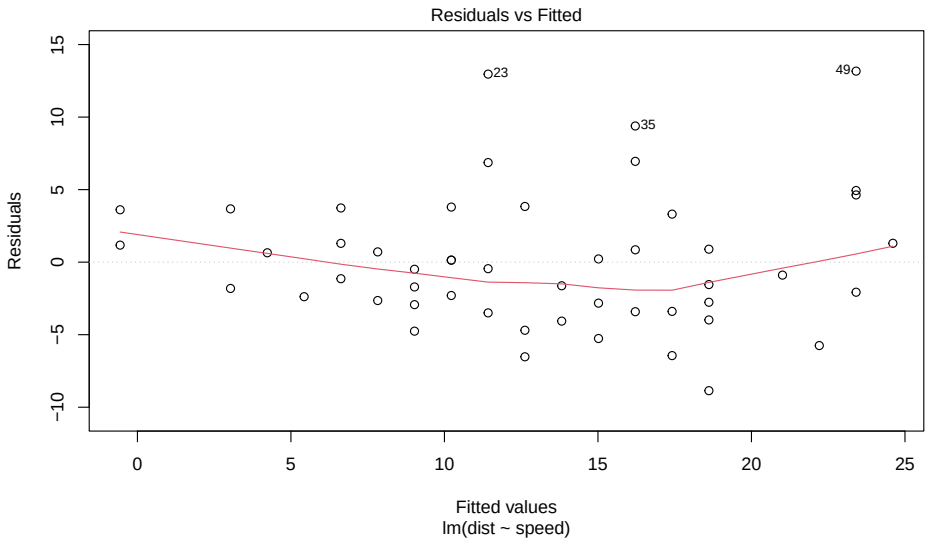
```
 2.5 % 97.5 %
 0.587  0.903
```

D'après notre modèle, nous estimons, avec une confiance de 95%, que la vraie valeur de la pente se situe quelque part entre 0.587 et 0.903. Autrement dit, chaque fois que la vitesse augmente de $1 km/h$, la distance de freinage augmentera en moyenne par $0.587\,m$ à $0.903\,m$.

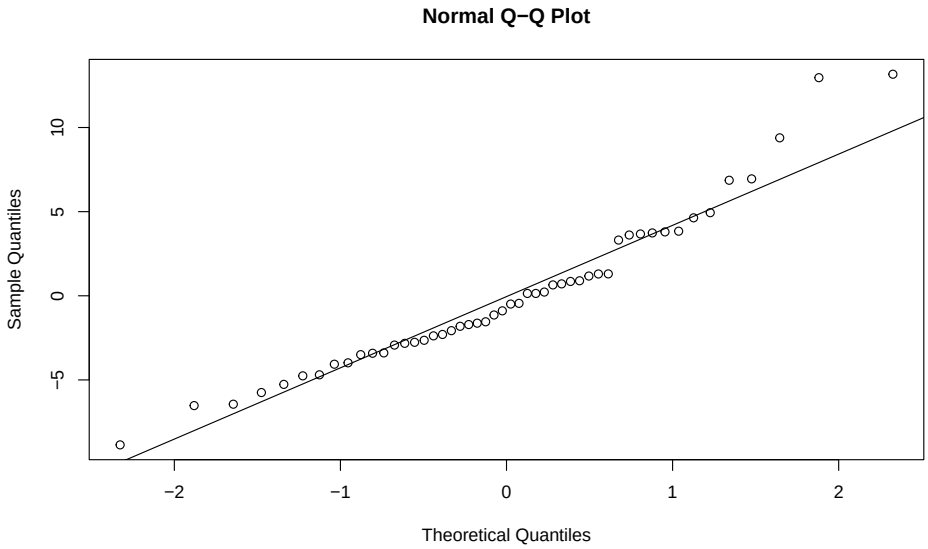
4

Les hypothèses dont nous avons besoin sont celles de la linéarité et de la normalité.

```
# Linéarité
plot(msp, 1)
```




```
# Normalité
resid(msp) |>
  qqnorm()
resid(msp) |>
  qqline()
```



Nous n'avons pas de raison de mettre en doute ces hypothèses.

5

Selon les chercheurs $dist = -5.6 + 0.8 \times speed$. Or, selon notre modèle $\widehat{dist} = -5.36 + 0.74 \times speed$. Ces deux affirmations ne coïncident pas (exactement), mais il faut se rappeler que les coefficients que nous avons calculés ne sont que des estimations (approximations) et non des valeurs réelles.

Pour répondre correctement à cette question, il faut consulter les un intervalles de confiances.

```
confint(msp)
```

```
             2.5 % 97.5 %
(Intercept) -9.500 -1.216
speed        0.587  0.903
```

Ces intervalles couvrent les valeurs suggérées par les chercheurs, donc, à 95%, nous ne pouvons donc pas rejeter leur modèle.

6

```
predict(msp, new = data.frame(speed = c(19.312, 24.784, 30.578)),
  interval = "confidence", level = 0.95)
```

```
    fit   lwr  upr
1  9.02  7.44 10.6
2 13.10 11.77 14.4
3 17.42 15.80 19.0
```

7

```
SCT <- sum((Cars$dist - mean(Cars$dist))^2)
SCR <- sum((Cars$dist - fitted(msp))^2)
SCE <- sum((fitted(msp) - mean(Cars$dist))^2)
c(SCT, SCR + SCE)
```

```
[1] 3023 3023
```

```
r2 <- SCE/SCT
r2
```

```
[1] 0.651
```

Exercice 3

Un scientifique désire étudier l'influence d'un antibiotique sur une culture bactérienne. Il répartit dans 10 tubes des volumes égaux de culture additionnée d'une quantité X (en mL) d'antibiotique, et il mesure (après incubation) la densité optique Y (grandeur sans unité). Il a noté que $\bar{x} = 0.6$, $\bar{y} = 4.15$, $s_x = 0.298$ et $s_y = 0.906$.

1

On sait que la droite de régression passe par $(\bar{x}, \bar{y}) = (0.6, 4.15)$. On donne un autre point dans la question qui est $(0, \hat{\beta}_0) = (0, 2.335)$. On peut donc évaluer la pente de la droite :

$$\hat{\beta}_1 = \frac{\Delta y}{\Delta x} = \frac{4.15 - 2.335}{0.6 - 0} = 3.025$$

L'équation de la droite de régression est donc $\hat{Y} = 2.335 + 3.025X$.

2

$$\begin{aligned} r^2 &= \frac{SCE}{SCT} = \frac{SCT - SCR}{SCT} \\ &= 1 - \frac{SCR}{SCT} = 1 - \frac{SCR}{(n-1)s_y^2} \\ &= 1 - \frac{0.0645}{(10-1) \times 0.906^2} = 0.991. \end{aligned}$$

3

Commençons par le r^2. On sait que

$$\begin{aligned} Cor(Z, Y) &= \frac{1}{n-1} \sum_i \left(\frac{z_i - \bar{z}}{s_z} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \\ &= \frac{1}{n-1} \sum_i \left(\frac{10x_i - 10\bar{x}}{10s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \\ &= \frac{1}{n-1} \sum_i \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \\ &= Cor(X, Y) \end{aligned}$$

Donc le r^2 ne change pas. Pour $\hat{\beta}_1$, on a que

$$\hat{\beta}_1^Z = \frac{s_y}{s_z} r = \frac{s_y}{10 s_x} r = \frac{1}{10} \hat{\beta}_1^X.$$

Et pour $\hat{\beta}_0$, on a que

$$\hat{\beta}_0^Z = \bar{y} - \hat{\beta}_1^Z \bar{z} = \bar{y} - \frac{1}{10} \hat{\beta}_1^X 10\bar{x} = \bar{y} - \hat{\beta}_1^X \bar{x} = \hat{\beta}_0^X.$$

4

Selon le modèle estimé, une augmentation de l'antibiotique de 1 mL entraînerait une augmentation de la densité optique de 3.025, en moyenne. Donc une augmentation en antibiotique de 10 mL entraînerait une augmentation de la densité optique de 30.25, en moyenne. En effet, Selon l'équation du modèle

$$\hat{Y}(x + 10) - \hat{Y}(x) = (2.335 + 3.025(x + 10)) - (2.335 + 3.025x) = 30.25.$$