Régression linéaire simple

Semaine 8

Table des matières

Objectifs

Les objectifs de ce CTD sont, dans un premier temps, d'établir quelques propriétés théoriques du modèle de régression linéaire simple. Dans un second temps, ces propriétés seront appliquées à l'étude de la relation entre deux variables simulées, pour lesquelles l'ensemble des hypothèses du modèle peuvent être vérifiées.

Le modèle de régression linéaire simple implique une seule variable indépendante (ou prédicteur), notée X. La relation peut être exprimée comme suit :

$Y = b_0 + b_1 X + \epsilon$, où

- Y est la variable dépendante,
- X est le prédicteur,
- b_0 est l'ordonnée à l'origine ou intercept,
- b_1 est le coefficient de la pente ou taille de l'effet de X sur Y.
- ϵ est l'erreur résiduelle représentant l'écart entre les valeurs observées et celles prédites (ou ajustées). L'erreur ϵ est une variable aléatoire gaussienne de moyenne nulle et de variance σ^2, indépendante de X.

Exercice 1. Quelques propriétés mathématiques du modèle de régression linéaire.

On suppose que les variables X et Y sont reliées selon le modèle de régression linéaire simple décrit ci-dessus.

Question 1

Montrer que l'espérance conditionnelle de Y sachant $X = x$ est égale à

$$\mathbb{E}[Y|X = x] = b_0 + b_1 x\,.$$

Solution.

$\mathbb{E}[Y|X = x] = \mathbb{E}[b_0 + b_1 X + \epsilon | X = x] = \mathbb{E}[b_0 + b_1 x + \epsilon] = \mathbb{E}[b_0] + \mathbb{E}[b_1 x] + \mathbb{E}[\epsilon]$ car l'espérance est linéaire

$\Rightarrow \mathbb{E}[Y|X = x] = b_0 + b_1 x + \mathbb{E}[\epsilon]$ car b_0, b_1 sont des constantes et x une valeur fixe.

Or ϵ est une variable aléatoire gaussienne de moyenne nulle $\Rightarrow \mathbb{E}[\epsilon] = 0$

$\Rightarrow \mathbb{E}[Y|X = x] = b_0 + b_1 x$

En admettant que le terme "régression" est un synonyme du terme "espérance conditionnelle", la droite de régression est l'espérance de Y en fonction de X.

Question 2

Montrer que

$$b_1 = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

et que

$$b_0 = \mathbb{E}[Y] - b_1\mathbb{E}[X]\,.$$

Solution.

$Cov(X,Y) = Cov(X, b_0 + b_1X + \epsilon) = Cov(X, b_0) + Cov(X, b_1X) + Cov(X, \epsilon)$ car la covariance est bilinéaire

Or, $Cov(X, b_0) = 0$ car b_0 est une constante et $Cov(X, \epsilon) = 0$ car X et ϵ sont indépendants.

$\Rightarrow Cov(X,Y) = Cov(X, b_1X) = b_1Cov(X,X) = b_1(\mathbb{E}[X^2] - \mathbb{E}[X]^2) = b_1 \times Var(X)$

$\Rightarrow b_1 = \frac{Cov(X,Y)}{Var(X)}$

$\mathbb{E}[Y] = \mathbb{E}[b_0 + b_1X + \epsilon] = \mathbb{E}[b_0] + \mathbb{E}[b_1X] + \mathbb{E}[\epsilon]$ car l'espérance est linéaire

Or ϵ est une variable aléatoire gaussienne de moyenne nulle $\Rightarrow \mathbb{E}[\epsilon] = 0$

$\Rightarrow \mathbb{E}[Y] = b_0 + b_1\mathbb{E}[X] \Rightarrow b_0 = \mathbb{E}[Y] - b_1\mathbb{E}[X]$

Question 3

Proposer des estimateurs des paramètres de régression b_0 et b_1.

Solution.

Soient :

$$\bar{y} = \frac{1}{n}\sum_{i=0}^{n} y_i$$

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

$$\hat{\mathrm{Var}}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\hat{\mathrm{Cov}}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

D'après la question 2, on prend alors comme estimateurs des paramètres de régression b_0 et b_1 :

$$\hat{b_1} = \frac{\sum_{i=0}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=0}^{n}(x_i - \bar{x})^2}$$

$$\hat{b_0} = \bar{y} - \hat{b_1}\bar{x}$$

Question 4

Que devient la valeur de b_1 lorsque les variables X et Y sont respectivement divisées par $\sqrt{\mathrm{Var}(X)}$ et $\sqrt{\mathrm{Var}(Y)}$? Comment interpréter la pente de la droite de régression lorsque les variables sont normalisées (variances égales à 1) ?

Solution.

$b_1 = \frac{Cov(\frac{X}{\sqrt{Var(X)}}, \frac{Y}{\sqrt{Var(Y)}})}{Var(\frac{X}{\sqrt{Var(X)}})} = \frac{Var(X)}{\sqrt{Var(X)\times Var(Y)}} \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(X,Y)}{\sqrt{Var(X)\times Var(Y)}}$

Lorsque les variables sont normalisées $\Rightarrow$ la pente de la droite de régression est b_1 et sa valeur est comprise entre -1 et 1 ($\Rightarrow b_1 = \rho$).

Question 5

On note ρ le coefficient de corrélation linéaire entre X et Y

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}\,.$$

On rappelle que la variance de ϵ est égale à σ^2. Démontrer que

$$\rho^2 = 1 - \frac{\sigma^2}{\mathrm{Var}(Y)}\,.$$

Solution.

$Var(Y) = Var(b_0 + b_1 X + \epsilon) = b_1^2 Var(X) + Var(\epsilon) = (\frac{Cov(X,Y)}{Var(X)})^2 \times Var(X) + \sigma^2 = \frac{Cov(X,Y)^2}{Var(X)} + \sigma^2$

$\Rightarrow \frac{Cov(X,Y)^2}{Var(X)} = Var(Y) - \sigma^2$

Or d'après la question 2, $\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \Rightarrow \rho^2 = \frac{Cov(X,Y)^2}{Var(X)\times Var(Y)} = \frac{Var(Y)-\sigma^2}{Var(Y)}$

$\Rightarrow \rho^2 = 1 - \frac{\sigma^2}{Var(Y)}$

Question 6

Montrer que la variance de la réponse Y peut être partitionnée en une *part expliquée* par le prédicteur X et une part *expliquée* par le résidu ϵ. En déduire que la part de variance expliquée par le prédicteur est égale à ρ^2. La part de variance expliquée par le résidu est égale à $1 - \rho^2$.

Solution.

D'après la formule d'Eve (formule de la variance totale) : $Var(Y) = Var(\mathbb{E}[Y|X]) + \mathbb{E}(Var(Y|X))$

Avec :

$Var(\mathbb{E}[Y|X]) = Var(b_0 + b_1 X)$ d'après la question 1

$\Rightarrow Var(\mathbb{E}[Y|X]) = b_1^2 Var(X)$

Et :

$\mathbb{E}(Var(Y|X)) = \mathbb{E}(Var(b_0 + b_1 X + \epsilon|X)) = \mathbb{E}(Var(\epsilon|X)) = \mathbb{E}(\sigma^2) = \sigma^2$

$\Rightarrow Var(Y) = b_1^2 Var(X) + \sigma^2$

Donc la variance de la réponse Y peut être partitionnée en une partie expliquée par le prédicteur X,

$$b_1^2 Var(X)$$

, et une partie expliquée par le résidu ϵ, σ^2.

La part de variance expliquée par X est définie par :

$r^2 = \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)}$

$\Rightarrow r^2 = \frac{b_1^2 Var(X)}{Var(Y)} = \frac{Var(Y)-\sigma^2}{Var(Y)} = 1 - \frac{\sigma^2}{Var(Y)}$

On a bien $r^2 = \rho^2$ donc la part de variance expliquée par le prédicteur est bien égale à ρ^2.

Exercice 2. Estimateurs des moindres carrés (ordinaires)

On dispose désormais de n observations indépendantes du modèle de régression linéaire simple

$$y_i = b_0 + b_1 x_i + \epsilon_i \,, \quad i = 1, \dots, n \,.$$

Les résidus ϵ_i sont supposés représenter n tirages indépendants de la loi $N(0, \sigma^2)$. On considère la fonction *de perte* suivante

$$\mathrm{RSS}(b_0, b_1) = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right)^2 \,.$$

L'acronyme RSS signifie *Residual Sum of Squares*, c'est à dire, la somme des carrés des résidus.

Question 1

On note $\bar{x}$ et $\bar{y}$ les moyennes empiriques des échantillons $\mathbf{x}$ et $\mathbf{y}$. Montrer que les valeurs $\hat{b}_0$ et $\hat{b}_1$ minimisant la fonction de perte $\mathrm{RSS}(b_0, b_1)$ sont égales à

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} x_i y_i / n - \bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 / n - \bar{x}^2} \,,$$

et

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} \,.$$

Comparer ces résultats aux valeurs théoriques de l'exercice précédent.

Solution.

Pour minimiser la fonction de perte $RSS(b_0, b_1)$ on la dérive par rapport aux paramètres (b_0, b_1) et on admet pour l'instant que le point où le gradient s'annule est unique et correspond à un minimum.

$$\frac{\partial RSS(b_0, b_1)}{\partial b_0} = 0 \Leftrightarrow -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0 \Leftrightarrow b_0 \times n = \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i \Leftrightarrow b_0 = \bar{y} - b_1 \bar{x}$$

Et,

$$\frac{\partial RSS(b_0,b_1)}{\partial b_1} = 0 \Leftrightarrow -2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0 \Leftrightarrow \sum_{i=1}^{n} x_i y_i - b_0 \times n\bar{x} - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

Or $b_0 = \bar{y} - b_1\bar{x}$,

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - n(\bar{y} - b_1\bar{x})\bar{x} - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} + b_1(n\bar{x}^2 - \sum_{i=1}^{n} x_i^2) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{x_i y_i}{n} - \bar{y}\bar{x} + b_1(\bar{x}^2 - \sum_{i=1}^{n} \frac{x_i^2}{n}) = 0 \; car \; n \neq 0$$

$$\Leftrightarrow b_1 = \frac{\sum_{i=1}^{n} \frac{x_i y_i}{n} - \bar{y}\bar{x}}{\sum_{i=1}^{n} \frac{x_i^2}{n} - \bar{x}^2}$$

Les valeurs minimisant la fonction de perte $\mathrm{RSS}(b_0, b_1)$ sont donc bien :

$$\Leftrightarrow \hat{b_1} = \frac{\sum_{i=1}^{n} \frac{x_i y_i}{n} - \bar{y}\bar{x}}{\sum_{i=1}^{n} \frac{x_i^2}{n} - \bar{x}^2}$$

Et :

$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x}\,.$$

On obtient bien la même chose que dans l'exercice 1.

Justification que le point où le gradient s'annule est unique et correspond à un minimum :

$$\frac{\partial^2 \mathrm{RSS}}{\partial b_0^2} = 2n, \qquad \frac{\partial^2 \mathrm{RSS}}{\partial b_0 \partial b_1} = 2\sum_{i=1}^{n} x_i, \qquad \frac{\partial^2 \mathrm{RSS}}{\partial b_1^2} = 2\sum_{i=1}^{n} x_i^2.$$

La matrice Hessienne est donc :

$$H = 2\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}.$$

$$\det\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 = n^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right) = n^2\hat{Var}(x).$$

Ainsi,

$$\Rightarrow \det(H) = 4n^2\hat{Var}(x).$$

On suppose que $\forall\ i = 1; 2; ...; n$, les x_i ne sont pas tous égaux $\Rightarrow \hat{Var}(x) > 0$

$\hat{Var}(x) > 0 \Rightarrow det(H) > 0 \Rightarrow$ H est définie positive $\Rightarrow$ la fonction RSS est strictement convexe sur (b_0, b_1) ie le point où le gradient s'annule en $(\hat{b}_0, \hat{b}_1)$ est unique et correspond bien au minimum de la fonction RSS.

Question 2

Pour un échantillon de taille n, on définit la **somme totale des carrés** (TSS) de la manière suivante

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \ ,$$

et le **coefficient de détermination** de la manière suivante

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \, .$$

Quels paramètres sont estimés par ces statistiques ? Justifier votre réponse. Pourquoi le coefficient de détermination est-il parfois appelé "pourcentage de variance expliquée" ?

Solution.

Dans l'exercice 1, on a fini par trouver la formule suivante :

$$\rho^2 = 1 - \frac{\sigma^2}{\mathrm{Var}(Y)}$$

Cette formule est très proche de l'expression du coefficient de détermination :

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \, .$$

On en déduit que R^2, RSS *et* TSS estiment respectivement la proportion de variance expliquée par le modèle, une estimation de $n \times \sigma^2$ et une estimation de $n \times Var(Y)$.

Le coefficient de détermination est parfois appelé "pourcentage de variance expliquée" car il exprime la fraction de la variabilité totale de Y qui est expliquée par la régression :

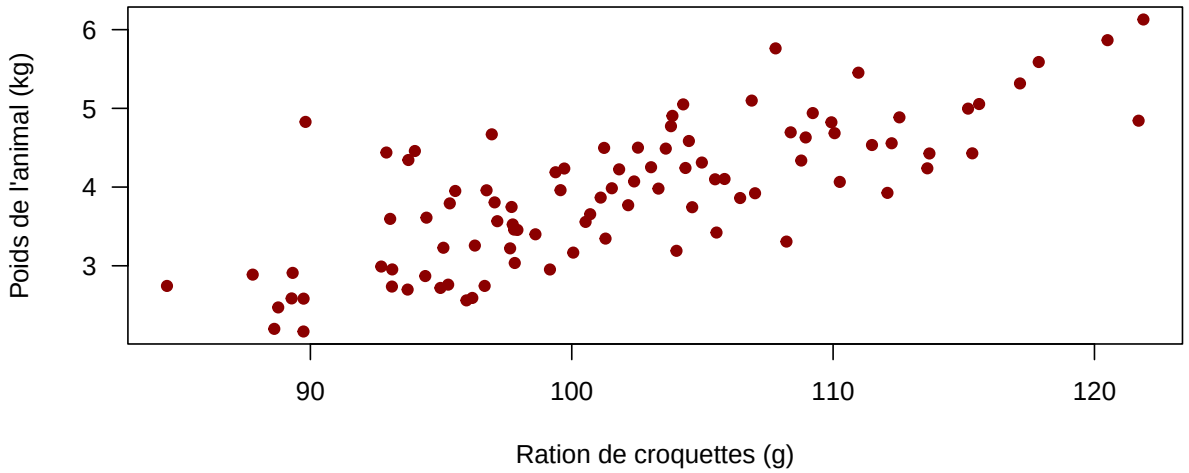
$R^2 = \frac{variance\ expliquee}{variance\ totale}$

Exercice 3.

Le poids d'un chat domestique dans une population est représenté par une variable y, exprimée en kg. On considère la variable x représentant la dose alimentaire journalière (croquettes), exprimée en g, donnée à un animal adulte. Nous cherchons à décrire l'effet de la dose alimentaire journalière sur le poids d'un animal. Un échantillon de taille $n = 95$ est fidèlement reproduit dans l'expérience suivante.

```r
# simulation : est-elle seulement réaliste...
set.seed(123)
nn = 100

# ration de croquettes king-size
x = rnorm(nn, 100, sd = 10)

# forcement gros-minet est en surcharge pondérale
epsilon = rnorm(nn, 0, sd = 0.6)
y = -5.2 + 0.09 * x  + epsilon

# y a pas de chatons (moins de 2 kg)
```

```r
x <- x[y > 2]
y <- y[y > 2]

# et voici nos données
plot(x, y,
     cex = .9, col = "red4", pch = 19, las = 1,
     xlab = "Ration de croquettes (g)",
     ylab = "Poids de l'animal (kg)")
```



Question 1

À l'aide de la commande `lm`, ajuster un modèle de régression linéaire aux données : y est la réponse, x le prédicteur. Afficher le résumé des estimations. Reprendre le graphique ci-dessus en représentant la droite de régression linéaire (`abline`). Inclure une légende en haut à gauche du graphique indiquant la signification du tracé et donnant le coefficient de détermination R^2.

Solution. Compléter le code suivant. Remarquer que la commande `summary` renvoie un objet de classe "summary.lm". Extraire de cet objet l'argument correspondant au calcul de R^2.

```r
## on ajuste un modèle de régression aux données avec y la réponse et x le prédicteur
mod = lm(y ~x)

## on cherche à obtenir un résumé statistique du modèle de régression précédent
resume = summary(mod)
resume
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.13063 -0.42763 -0.01082  0.37555  1.86798
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.254769   0.730255  -5.826 8.08e-08 ***
## x            0.080329   0.007147  11.240  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5754 on 93 degrees of freedom
## Multiple R-squared:  0.576,  Adjusted R-squared:  0.5714
## F-statistic: 126.3 on 1 and 93 DF,  p-value: < 2.2e-16
```

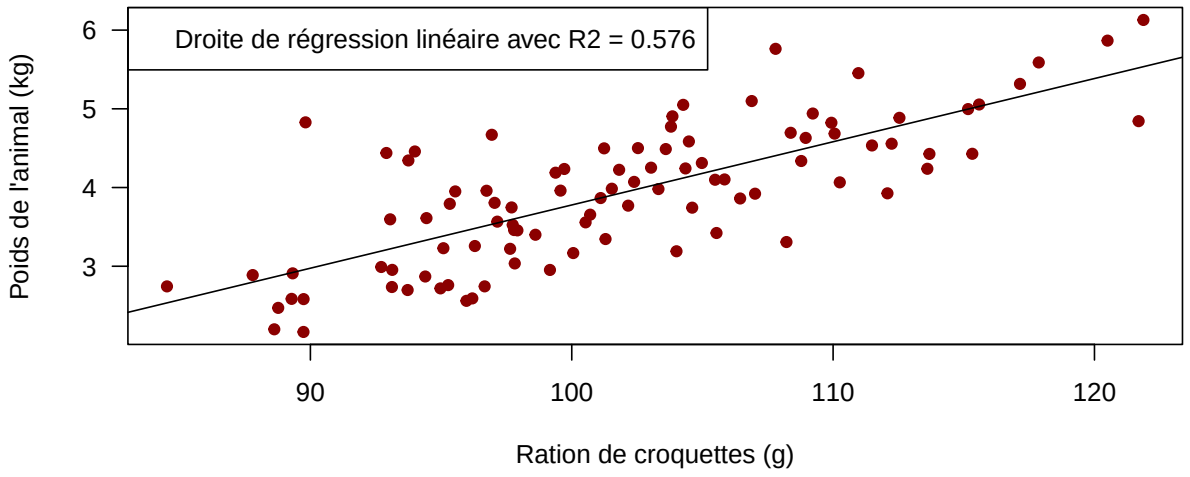
```r
## extraire le coefficient de détermination R2
R2 = resume$r.squared
```

p-valeur $< 0{,}05$ donc il y a bien une relation linéaire entre la dose alimentaire et le poids de l'animal.

Un chat ayant une ration de x g a un poids moyen estimé de -4,25 + 0,0803x.

$R^2 \approx 0,576 \Rightarrow 57{,}6\%$ de la variabilité du poids des chats est expliquée par la ration alimentaire.

```r
# revoici nos données
plot(x, y,
     cex = .9, col = "red4", pch = 19, las = 1,
     xlab = "Ration de croquettes (g)",
     ylab = "Poids de l'animal (kg)")

# ajouter une légende
legend(x = "topleft", legend = paste("Droite de régression linéaire avec R2 =",round(R2,3) ))

# tracer la droite de régression
abline(mod)
```



Question 2

Quel est l'effet prédit d'une augmentation (respectivement d'une diminution) de 10 g de la ration journalière sur le poids des chats ? Calculez ces effets en grammes sans utiliser la commande `lm`, puis vérifier votre réponse à l'aide du résultat de la commande.

Solution.

On pose f(x) = a +bx avec a l'ordonnée à l'origine et b le coefficient directeur de la droite de régression linéaire.

f(x+10) - f(x) = a + b(x+10) - (a + bx) = 10*b

f(x-10) - f(x) = a + b(x-10) - (a + bx) = -10*b

```
b = cov(x,y) / var(x)
cat("Effet d'une augmentation de 10 g de la ration journalière :", round(10*b*1000, 0),"g", "\n")
```

```
## Effet d'une augmentation de 10 g de la ration journalière : 803 g
```

```
cat("Effet d'une augmentation de 10 g de la ration journalière :", round(-10*b*1000, 0),"g")
```

```
## Effet d'une augmentation de 10 g de la ration journalière : -803 g
```

```
#effet prédit avec la commande lm
c = coef(mod)[2]
cat("Effet d'une augmentation de 10 g de la ration journalière :", round(10*c*1000, 0),"g", "\n")
```

```
## Effet d'une augmentation de 10 g de la ration journalière : 803 g
```

```
cat("Effet d'une augmentation de 10 g de la ration journalière :", round(-10*c*1000, 0),"g")
```

```
## Effet d'une augmentation de 10 g de la ration journalière : -803 g
```

Les résultats sont identiques.

Question 3

Comment interprétez-vous les effets prédits ? Les prédictions sont-elles valides pour un animal en particulier ?

Solution.

Les animaux qui reçoivent une ration journalière supérieur de 10 g à ceux d'une population donnée ont, en moyenne, un poids supérieur d'environ 803 g.

Il s'agit d'une interprétation statistique à l'échelle de la population pas d'une prédiction exacte pour un animal particulier : chaque chat a sa variabilité individuelle.

Question 4

Estimer l'espérance conditionnelle du poids d'un animal sachant que sa ration journalière de croquettes est 112 g. Interpréter le résultat.

Solution.

```r
# d'après la question 1 del'exercice espérance conditionnelle est la 'prédiction' de la variable répons
a = coef(mod)[1]
b = coef(mod)[2]
E = a + b*112

cat("L'espérance conditionnelle du poids d'un animal sachant que sa ration journalière de corquettes es
```

```
## L'espérance conditionnelle du poids d'un animal sachant que sa ration journalière de corquettes est
```

```r
round(predict(mod, newdata = data.frame(x = 112)), 2)
```

```
##    1
## 4.74
```

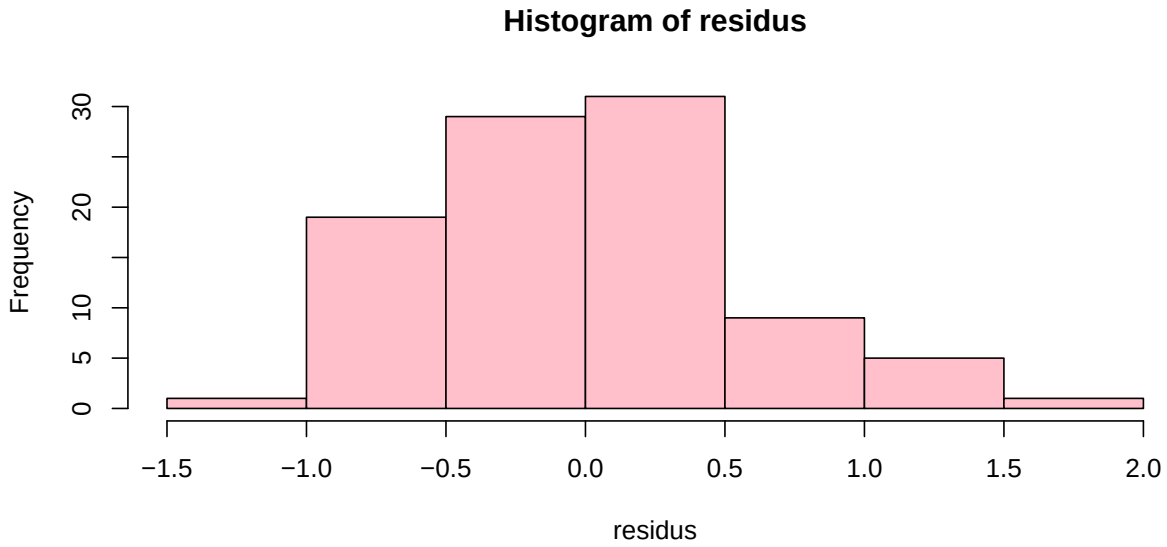
Un chat recevant 112 g de croquettes par jour a un poids moyen attendu d'environ 4,74 kg.

Question 5

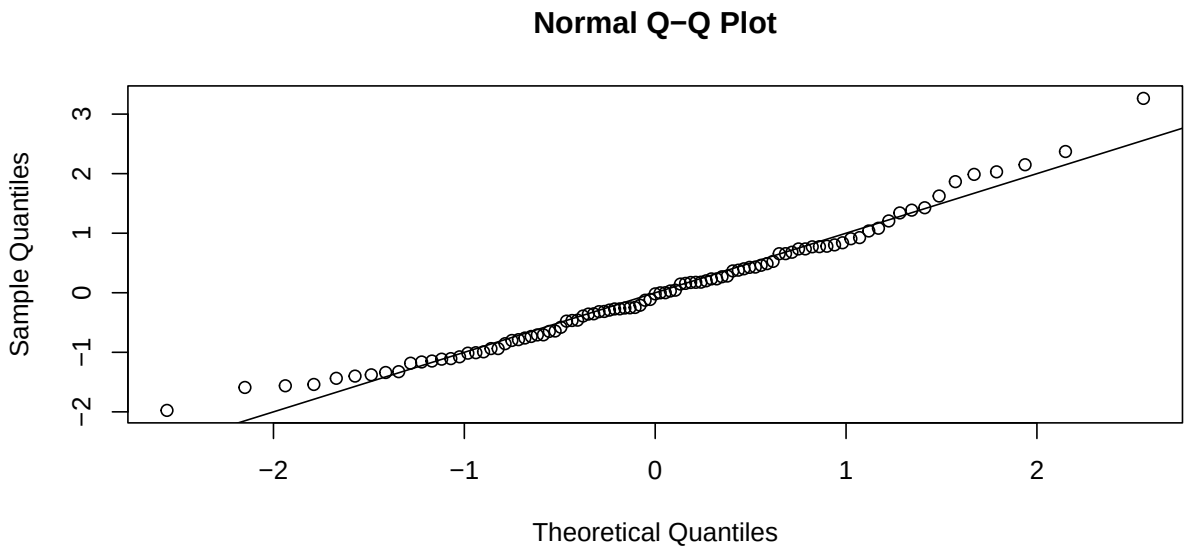
Afficher l'histogramme des résidus, puis un graphe quantile-quantile pour vérifier l'adéquation de la loi normale. Commenter le résultat.

Solution. Compléter le code suivant.

```r
#residus = mod$residuals
residus <- mod$residuals
hist(residus, col = "pink")
```




```r
# Vérification graphique de l'adéquation de la loi normale
qqnorm((residus - mean(residus))/sd(residus)); abline(0,1)
```



L'histogramme et le graphique Q–Q ne révèlent pas d'écarts importants à la normalité. Les résidus semblent donc approximativement normaux, ce qui valide l'hypothèse de normalité du modèle de régression linéaire.

Question 6

Calculer la somme des carrés des résidus (RSS), puis le coefficient de détermination de la régression. Vérifier le résultat à partir d'un calcul direct, puis du résumé du modèle ajusté.

Solution. Compléter le code suivant et commenter les résultats.

```
## Calcul de RSS
RSS <- sum(residuals(mod)^2)

## TSS
TSS <- sum((y - mean(y))^2)

## verification : R2 est le coef de correlation
R2 <- 1 - RSS / TSS
cat("vérification, R2 =", R2)
```

```
## vérification, R2 = 0.5759938
```

```
## et dans summary(mod)
summary(mod)$r.squared
```

```
## [1] 0.5759938
```

Question 7

On admet que l'estimateur

$$\hat{\sigma}^2 = \mathrm{RSS}/(n-2)$$

est un estimateur sans biais de σ^2, la variance de l'erreur résiduelle. Calculer l'estimation de l'écart-type résiduel. Identifier les estimations effectuées précedemment dans le résumé du modèle.

Solution. Utiliser la commande `summary`. Il y a quatre nombres à identifier : les deux coefficients de régression, le coefficient de détermination et l'écart-type de l'erreur résiduelle.

```
n <- length(y)
RSS <- sum(residuals(mod)^2)
sigma <- sqrt(RSS / (n - 2))
s <- summary(mod)

b0 <- s$coefficients[1, "Estimate"]
b1 <- s$coefficients[2, "Estimate"]
R2 <- s$r.squared
sigma_summary <- s$sigma

cat("b0 :", b0, "\n")
```

```
## b0 : -4.254769
```

```
cat("b1 :", b1, "\n")
```

```
## b1 : 0.08032909
```

```
cat("R^2:", R2, "\n")
```

```
## R^2: 0.5759938
```

```
cat("Écart-type résiduel :",sigma, "\n")
```

```
## Écart-type résiduel : 0.5754129
```

```
cat("summary(mod)$sigma =", sigma_summary, "\n")
```

```
## summary(mod)$sigma = 0.5754129
```

On observe que écart-type de l'erreur résiduelle représente la dispersion moyenne des points observés autour de la droite de régression : plus il est petit, plus l'ajustement est précis.

Exercice 4. Le matou revient pour les tests statistiques

On poursuit l'exemple concernant le poids d'un chat domestique, y, exprimé en kg, expliqué par une variable x représentant la dose journalière de croquettes donnée à un animal, exprimée en g.

Question 1

Rappeler l'estimation de la taille d'effet (ou coefficient de régression) b_1.

Solution. Utiliser par exemple la commande `coefficients`.

```r
b1 <- coefficients(mod)[2]
b1
```

```
##          x
## 0.08032909
```

Question 2

Calculer la statistique de Student t_{n-2} (t-score) associée à l'estimation $\hat{b}_1$ en supposant $b_1 = 0$. Vérifier le résultat à l'aide du résumé du modèle (`summary`).

Solution. D'après un résultat de cours (que l'on peut retrouver dans l'exercice 5), la variance de $\hat{b}_1$ est σ^2/ns_x^2. On définit la valeur t en normalisant l'estimateur (standardisation). On retrouve la valeur t dans la colonne "t value" du tableau des coefficients du modèle ajusté.

```r
cat("Le t-score est égal à : ",
    round(coef(summary(mod))[6], 4),
    "\n")
```

```
## Le t-score est égal à :  11.2399
```

Question 3

Pour la pente de la régression b_1, donner un intervalle de confiance bilatéral au seuil $1 - \alpha = 95\%$. Calculer cet intervalle de confiance à l'aide des quantiles de la loi de Student (`qt`). Vérifier le résultat avec la fonction `confint`.

Solution. Pour construire l'intervalle de confiance, on admettra que la variable aléatoire

$$T_{n-2} = \sqrt{ns_x^2}\frac{\hat{b}_1 - b_1}{\hat{\sigma}}$$

suit une loi de Student à $n-2$ degrés de liberté.

On a :

$$\mathrm{P}(\Phi^{-1}(0.05/2) \leq T_{n-2} \leq \Phi^{-1}(1 - 0.05/2)) = 0.95$$

Ce qui est équivalent à :

$$\mathrm{P}(\hat{b}_1 - \frac{\hat{\sigma}}{\sqrt{ns_x^2}}\Phi^{-1}(1 - 0.05/2) \leq b_1 \leq \hat{b}_1 - \frac{\hat{\sigma}}{\sqrt{ns_x^2}}\Phi^{-1}(0.05/2)) = 0.95$$

```r
# calcul des quantiles de la loi tn-2
# alpha = 0.05
t_lower = qt(0.025, n - 2)
t_upper = qt(0.975, n - 2)

# bornes de l'intervalle de confiance
b_1_chap = coef(summary(mod))[2]
sigma = summary(mod)$sigma
s_x_2 = (n-1)/n*var(x)
lower =  b_1_chap - sigma / sqrt(n*s_x_2) * t_upper
upper = b_1_chap - sigma / sqrt(n*s_x_2) * t_lower
```

```r
conf_int = c(lower, upper)
names(conf_int) = c("2.5%", "97.5%")

# Affichage
cat("valeur de b_1_chap : \n")
```

```
## valeur de b_1_chap :
```

```r
b_1_chap
```

```
## [1] 0.08032909
```

```r
cat("Intervalle de confiance : \n")
```

```
## Intervalle de confiance :
```

```r
conf_int
```

```
##       2.5%      97.5%
## 0.06613706 0.09452112
```

On peut vérifier ce résultat directement avec la fonction `confint`

```r
confint(mod)["x",]
```

```
##      2.5 %     97.5 %
## 0.06613706 0.09452112
```

Question 4

Calculer la p-valeur du test de l'hypothèse nulle "$H_0 : b_1 = 0$" en utilisant la commande `pt`. Vérifier le résultat à l'aide du résumé du modèle. L'augmentation de la dose de croquettes d'un écart-type a-t-elle un effet significatif sur le poids de l'animal (au seuil 5%) ?

Solution. La p-valeur du test bilatéral correspond à

$$p = \mathbb{P}(|T_{n-2}| > |t|)$$

où t est le t-score calculé précédemment (et $b_1 = 0$).

On a :

$$p = \mathbb{P}(|T_{n-2}| > |t|) = \mathbb{P}(T_{n-2} > |t[) + \mathbb{P}(T_{n-2} < -|t[)$$

(évênements incompatibles) Donc :

$$p = 2\mathbb{P}(T_{n-2} > |t[)$$

(loi de Student symétrique autour de 0)

```r
# Calcul avec pt()
t = coef(summary(mod))[6]
cat("Calcul avec pt() :", 2*pt(abs(t),n-2, lower = FALSE), "\n")
```

```
## Calcul avec pt() : 5.079564e-19
```

```r
# Affichage avec le résumé
cat("Dans le résumé :", summary(mod)$coefficients[8])
```

```
## Dans le résumé : 5.079564e-19
```

On a $p < 0.05$, donc b_1 est significativement différent de 0. Ainsi, l'augmentation de la dose de croquette a un effet significatif sur le poid du chat.

Question 5

Test de pertinence de la régression. Calculer la p-valeur pour le test de l'hypothèse nulle "$H_0 : \rho^2 = 0$" en utilisant la commande `pf`. Vérifier le résultat à l'aide du résumé du modèle.

Solution. D'après le cours, la p-valeur du test correspond à

$$p = \mathbb{P}(F_{1,n-2} > z)$$

où $F_{1,n-2}$ est de loi de Fisher $F(1, n-2)$ et

$$z = (n-2)\frac{R^2}{1-R^2} = (n-2)\frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{RSS}}.$$

Faire les calculs.

```r
# Calcul
z = (n-2) * (TSS - RSS) / RSS
cat("Le calcul donne :", pf(z, 1, n-2, lower = FALSE), "\n")
```

```
## Le calcul donne : 5.079564e-19
```

```r
# D'après le résumé
f <- summary(mod)$fstatistic
cat("Le résumé donne :", pf(f[1], f[2], f[3], lower.tail = FALSE))
```

```
## Le résumé donne : 5.079564e-19
```

On a $p < 0.05$, donc ρ^2 est significativement différent de 0. On obtient alors la même conclusion que précédemment.

Question 6

Vérifier que la p-valeur obtenue est exactement égale à celle du test de significativité de la pente b_1. Pourquoi est-ce vrai ?

Solution. Vérification : Il s'agit de la dernière ligne du résumé du modèle.

```r
# extraire t, p-value du test de la pente
t_val <- coef(summary(mod))["x", "t value"]
p_t   <- coef(summary(mod))["x", "Pr(>|t|)"]

# extraire la statistique F et sa p-value depuis summary(mod)
fstat <- summary(mod)$fstatistic
f_val <- fstat[1]
df1   <- fstat[2]
df2   <- fstat[3]
p_f   <- pf(f_val, df1, df2, lower.tail = FALSE)

# vérifier l'égalité numérique et relation F = t^2
cat("t-value :", round(t_val, 6), "\n")
```

```
## t-value : 11.23995
```

```r
cat("t-test p-value :", signif(p_t, 6), "\n")
```

```
## t-test p-value : 5.07956e-19
```

```r
cat("F-stat :", round(f_val, 6), "  (df1 =", df1, ", df2 =", df2, ")\n")
```

```
## F-stat : 126.3364   (df1 = 1 , df2 = 93 )
```

```r
cat("F-test p-value :", signif(p_f, 6), "\n\n")
```

```
## F-test p-value : 5.07956e-19
```

La p-valeur obtenue est exactement égale à celle du test de significativité de la pente b_1.

Le test de Student teste $H_0 : b_1 = 0$

Le test de Fisher teste $H_0 : \rho^2 = 0$

D'après l'exercice 3 :

$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$
$\hat{b}_0 = \bar{y}_i - \hat{b}_1 \bar{x}_i$

Donc :

$\hat{b}_1 = 0 \Leftrightarrow \hat{y}_i = \bar{y} \Leftrightarrow RSS = TSS \Leftrightarrow R^2 = 0$

De plus :

$b_1 = 0 \Leftrightarrow \frac{Cov(X,Y)}{Var(X)} = 0 \Leftrightarrow Cov(X,Y) = 0$ d'après la question 2 de l'exercice 1.

Or d'après la question 5 de l'exercice 1 : $\rho^2 = (\frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}})^2$

Donc $b_1 = 0 \Leftrightarrow \rho^2 = 0$

Donc dans ce cas le test de Fisher équivaut au test de Student donc le test t et le test F donnent la même p-valeur.

Exercice pour les pas fatigués (facultatif).

On dispose de n observations indépendantes du modèle de régression linéaire

$$y_i = b_0 + b_1 x_i + \epsilon_i\,, \quad i = 1, \dots, n\,,$$

où les résidus ϵ_i sont issus de tirages indépendants de la loi $N(0, \sigma^2)$.

Dans ce qui suit, on considère que le vecteur $\mathbf{x} = (x_1, \dots, x_n)$ contient des valeurs fixes. Les calculs s'effectueront conditionnellement à ces valeurs. Seules les observations y_i et ϵ_i sont donc issues de tirages aléatoires. On notera Y_i la réponse aléatoire correspondant à l'observation y_i.

Question 1

Montrer que les estimateurs $\widehat{b}_1$ et $\widehat{b}_0$ sont des estimateurs sans biais de b_1 et b_0 conditionnellement à $\mathbf{x}$.

Solution. On note $s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$, la variance empirique (non-corrigée) calculée à partir de $\mathbf{x}$. Utiliser la définition LUE de l'estimateur $\widehat{b}_1$ (cf classe inversée)

$$\widehat{b}_1 = \frac{1}{n s_x^2} \sum_{i=1}^n (x_i - \bar{x}) Y_i\,,$$

afin de calculer l'espérance conditionnelle demandée : $\mathbb{E}[\widehat{b}_1 \mid \mathbf{x}]$.

Question 2

Montrer que la variance de $\widehat{b}_1$ est égale à $\sigma^2/n s_x^2$ conditionnellement à $\mathbf{x}$, où s_x^2 est la variance empirique de l'échantillon $\mathbf{x}$. Conclure que l'estimateur est convergent.

Solution. Utiliser à nouveau le fait que $\widehat{b}_1$ s'exprime comme combinaison linéaire des variables Y_i. Calculer $\mathrm{Var}(\widehat{b}_1|\mathbf{x})$.