MAT3775

Patrick Boily, Gilles Lamothe

2022-12-27

Exemples

Un grand nombre d'exemples utiliseront l'ensemble de données suivant.

```
library(tidyverse)
gapminder = as.data.frame(unclass(data.frame(read.csv("Data/gapminder_SS.csv"))),
                          stringsAsFactors=TRUE)
gapminder <- gapminder[,c("country","year","continent",
                          "population","infant_mortality","fertility","gdp",
                          "life_expectancy")]

gapminder = gapminder |> mutate(lgdppc=log(gdp/population),gdppc=gdp/population)
str(gapminder)
```

```
## 'data.frame':    10545 obs. of  10 variables:
##  $ country         : Factor w/ 185 levels "Albania","Algeria",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ year            : int  1960 1960 1960 1960 1960 1960 1960 1960 1960 1960 ...
##  $ continent       : Factor w/ 5 levels "Africa","Americas",..: 4 1 1 2 2 3 2 5 4 3 ...
##  $ population      : int  1636054 11124892 5270844 54681 20619075 1867396 54208 10292328 7065525 389
##  $ infant_mortality: num  115.4 148.2 208 NA 59.9 ...
##  $ fertility       : num  6.19 7.65 7.32 4.43 3.11 4.55 4.82 3.45 2.7 5.57 ...
##  $ gdp             : num  NA 1.38e+10 NA NA 1.08e+11 ...
##  $ life_expectancy : num  62.9 47.5 36 63 65.4 ...
##  $ lgdppc          : num  NA 7.13 NA NA 8.57 ...
##  $ gdppc           : num  NA 1243 NA NA 5254 ...
```

1. Ajustement d'un modèle linéaire avec `lm()`

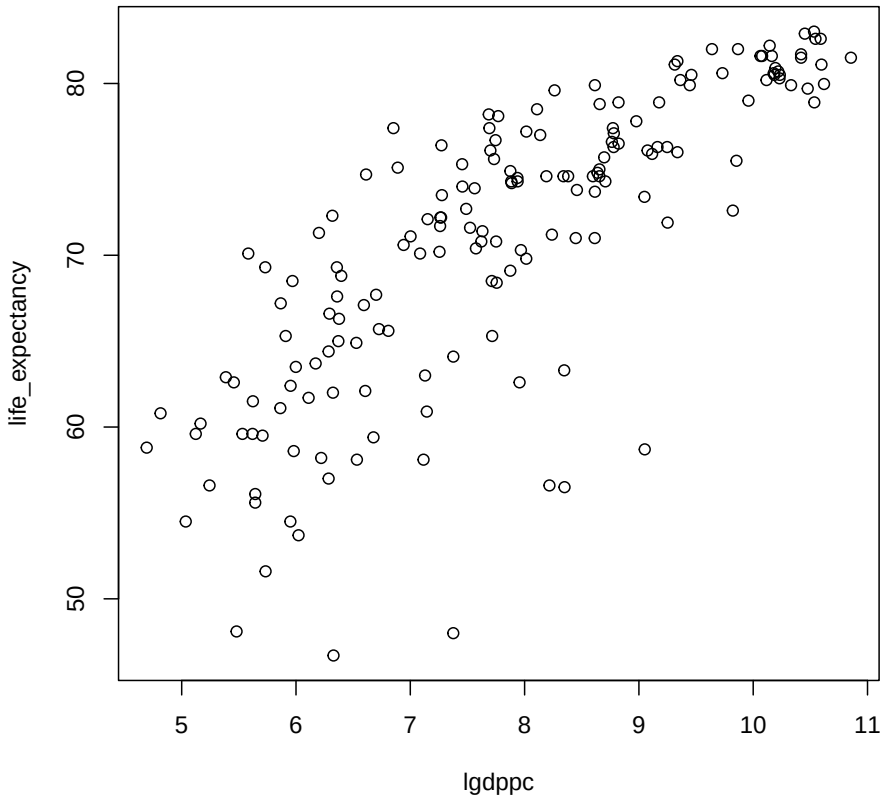
On s'intéresse pour l'instant aux observations datant de 2011 et portant sur l'espérance de vie et sur le logarithme du revenu par habitant.

```
gapminder.elr <- gapminder |>
  filter(year==2011) |>
  select(lgdppc,life_expectancy)
str(gapminder.elr)
head(gapminder.elr)
```

```
## 'data.frame':    185 obs. of  2 variables:
##  $ lgdppc         : num  7.69 7.7 7.12 9.12 9.34 ...
##  $ life_expectancy: num  77.4 76.1 58.1 75.9 76 ...
```

lgdppc	life_expectancy
7.691867	77.4
7.700730	76.1
7.115692	58.1
9.115537	75.9
9.337278	76.0
7.276390	73.5

```
plot(gapminder.elr)
```



On peut utiliser la fonction `lm()` afin d'ajuster le modèle linéaire qui décrit l'espérance de vie (variable réponse Y) en fonction du logarithme du revenu par habitant (variable prédicteur X) en 2011.

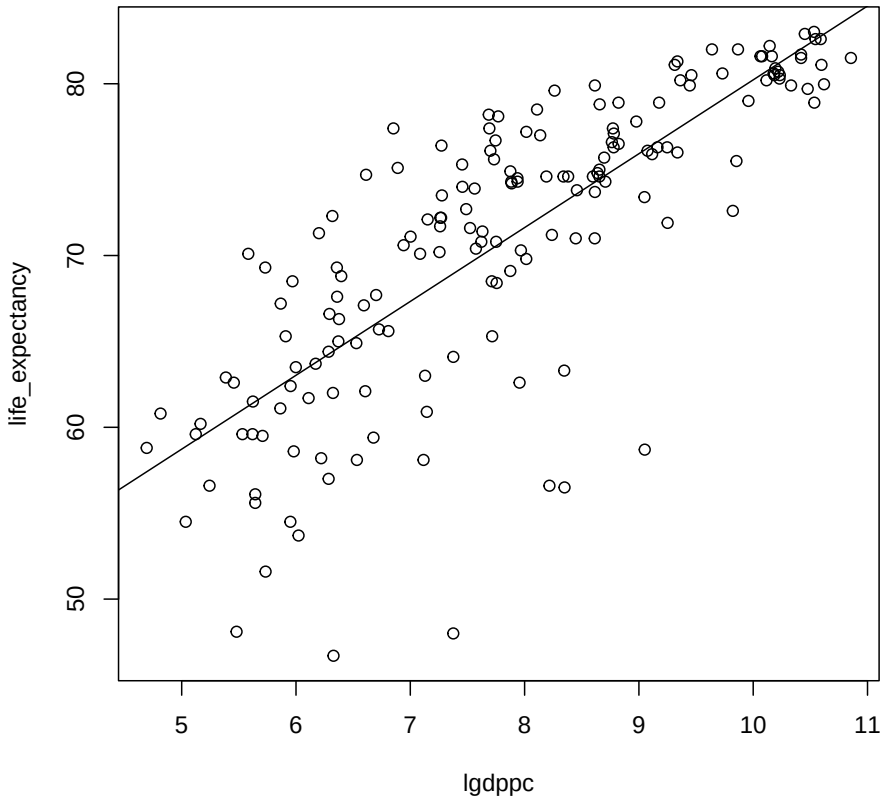
```
mod <- lm(life_expectancy ~ lgdppc, data=gapminder.elr)
mod
```

```
##
## Call:
## lm(formula = life_expectancy ~ lgdppc, data = gapminder.elr)
##
## Coefficients:
## (Intercept)       lgdppc
##       37.23         4.30
```

Le modèle linéaire estimé est ainsi

$$\widehat{\text{Life Expectancy}}_{2011} = 37.23 + 4.30 \cdot \log \text{GDP per capita}_{2011}.$$

```
plot(gapminder.elr)
abline(mod)
```



Commentaires

- Nous avons construit un objet de type `lm` nommé `mod`. Cet objet contient plusieurs composantes que l'on peut afficher l'aide de la fonction `names()`.

```
names(mod)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "na.action"     "xlevels"       "call"          "terms"
## [13] "model"
```

  La première composantes est le vecteur des estimations des coefficients $\vec{\beta}$.

```
mod$coefficients
```

```
## (Intercept)      lgdppc
##   37.229550    4.299686
```

Ainsi, $b_0 = 37.229550$ et $b_1 = 4.299686$.

La deuxième composante est le vecteur des résidus et la huitième composante est le nombre de degrés de liberté des résidus. On peut les utilisés pour calculer l'estimation de la variance de l'erreur, c'est-à-dire

$$\text{MSE} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

L'écart type de cette variance est l'écart type des résiduels

$$\text{se} = \sqrt{\text{MSE}}$$

qui décrit l'écart type de la droite d'ajustement.

```
MSE<-sum(mod$residuals^2)/mod$df.residual
MSE
```

```
## [1] 26.98027
```

```
sqrt(MSE)
```

```
## [1] 5.194254
```

- Nous pouvons utiliser plusieurs fonctions avec un object de type `lm`:

```
methods(class=lm)
```

```
##  [1] add1           alias          anova          case.names     coerce
##  [6] confint        cooks.distance deviance       dfbeta         dfbetas
## [11] drop1          dummy.coef     effects        extractAIC     family
## [16] formula        fortify        hatvalues      influence      initialize
## [21] kappa          labels         logLik         model.frame    model.matrix
## [26] nobs           plot           predict        print          proj
## [31] qr             residuals      rstandard      rstudent       show
## [36] simulate       slotsFromS3    summary        variable.names vcov
## see '?methods' for accessing help and source code
```

Par exemple, nous savons que la droite ajustée est celle qui maximise le logarithme de la fonction de vraisemblance

$$\log L = -\frac{n}{2}\left[\log\left(2\pi \cdot \frac{\text{SSE}}{n}\right) + 1\right].$$

Nous avons 3 degrés de liberté, puisqu'il y a 3 paramètres à estimer: β_0, β_1, et σ^2.

```
logLik(mod)
```

```
## 'log Lik.' -514.1646 (df=3)
```

Voici comment on pourrait vérifier que `R` utilise bien la formule précédente afib de calculer le maximum du logarithme de la fonction de vraisemblance:

```
p <- length(mod$coefficients)
n<-mod$df.residual+p
SSE<-sum(mod$residuals^2)
-n/2*(log(2*pi*SSE/n)+1)
```

```
## [1] -514.1646
```

- Nous pouvons également obtenir des intervalles de confiance pour les estimations des paramètres:

```
confint(mod)
```

```
##                 2.5 %    97.5 %
## (Intercept) 33.193051 41.266050
## lgdppc       3.794986  4.804387
```

Le niveau de confiance est 95% par défaut, mais il peut être modifié avec l'argument `level`:

```
confint(mod,level=0.98)
```

```
##                   1 %      99 %
## (Intercept) 32.427065 42.032036
## lgdppc       3.699212  4.900161
```

Il est aussi possible de spécifier les paramètres qui nous intéressent:

```
confint(mod,parm=c("lgdppc"))
```

```
##           2.5 %   97.5 %
## lgdppc 3.794986 4.804387
```

- Il est aussi possible d'obtenir un sommaire de l'ajustement avec la fonction `summary()`:

```
summary(mod)
```

```
##
## Call:
## lm(formula = life_expectancy ~ lgdppc, data = gapminder.elr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.9439  -1.8240   0.4922   3.0810  10.7078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2296     2.0445   18.21   <2e-16 ***
## lgdppc        4.2997     0.2556   16.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.194 on 166 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.6302, Adjusted R-squared:  0.628
## F-statistic: 282.9 on 1 and 166 DF,  p-value: < 2.2e-16
```

Il n'est pas nécessaire d'afficher tout le sommaire,un object qui a lui-même des composantes:

```
names(summary(mod))
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"  "na.action"
```

Par exemple, voici comment extraire l'estimation des paramètres et le test de la signification correspondant:

```
summary(mod)$coefficients
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 37.229550  2.0444622 18.20995 2.022181e-41
## lgdppc       4.299686  0.2556277 16.82011 1.069638e-37
```

On peut afficher le coefficient de détermination R^2 comme suit:

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

```
## [1] 0.6302205
```

ou encore l'écart type des résiduels $\sqrt{\text{MSE}}$:

```
summary(mod)$sigma
```

```
## [1] 5.194254
```

2. Analyse de variance avec `lm()`

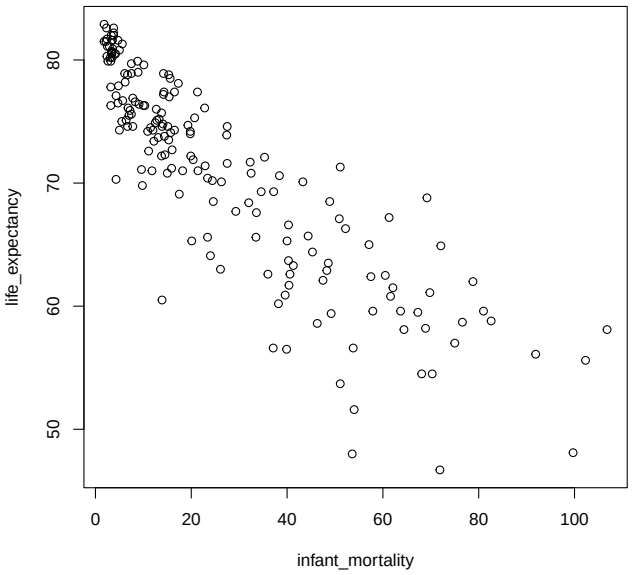
Nous allons utiliser un modèle linéaire simple pour décrire la l'ésperance de vie (Y) en fonction du taux de mortalité infantile (X) en 2011 (on a $n = 178$).

```
gapminder.em <- gapminder |>
  filter(year==2011) |>
  select(infant_mortality,life_expectancy) |>
  drop_na()
str(gapminder.em)
head(gapminder)
```

```
## 'data.frame':    178 obs. of  2 variables:
##  $ infant_mortality: num  14.3 22.8 106.8 7.2 12.7 ...
##  $ life_expectancy : num  77.4 76.1 58.1 75.9 76 73.5 82.2 80.7 70.8 72.6 ...
```

infant_mortality	life_expectancy
14.3	77.4
22.8	76.1
106.8	58.1
7.2	75.9
12.7	76.0
15.3	73.5

```
plot(gapminder.em)
```



Visuellement, il n'est pas irraisonable de s'attendre à ce que la fonction de réponse moyenne soit donnée par

$$E[Y \mid X = x] = \beta_0 + \beta_1 x.$$

Comment peut-on vérifier la signification du taux de mortalité infantile? On confronte tout simplement

$$H_0 : \beta_1 = 0 \qquad \text{à} \qquad H_1 : \beta_1 \neq 0.$$

Puisque nous avons un modèle simple (un seul prédicteur, c'est aussi un test de la signification de la régression: on pourrait se servir de la statistiqur

$$t^* = \frac{b_1}{s\{b_1\}}.$$

Mais il y s une autre approche, celle de l' analyse de variance (ANOVA). Le tableau d'ANOVA s'obtient en mettant un objet de type `lm` dans la fonction `anova()`.

```
mod <- lm(life_expectancy ~ infant_mortality, data=gapminder.em)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: life_expectancy
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## infant_mortality   1 9297.4  9297.4  528.73 < 2.2e-16 ***
## Residuals        176 3094.8    17.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Commentaires:

- La statistique du test est $F^* = 528.73$. Ceci veut dire que l'estimation de la variance de l'erreur basée sur la somme des carrés de la régression est 528.73 fois aussi grande que l'estimation de la variance de l'erreur basée sur la somme des carrés résiduels. Il est fort probable que MSR n'estime pas σ^2, mais plutˆot une quantité plus élevée. Comme nous savons que

$$E(\text{MSR}) = E\left(\frac{\text{SSR}}{1}\right) = \sigma^2(1 + \beta_1^2 s_{xx}),$$

  cela indique fortement que $\beta_1 \neq 0$.

- Pour avoir une mesure de la signification des preuves ayant à l'encontre de $H_0 : \beta_1 = 0$, nous devons calculer les chances d'avoir observé une statistique F^* aussi élevée que 528.73 en supposant que H_0 était valide: si c'est le cas, $F^* \sim F(1, n-2)$. Puisque la valeur P avec ce modèle est

$$P(F(1, 176) > 528.73) < 0.001,$$

  il y a une forte évidence en faveur de H_1.

- On peut aussi s'y prendre à l'aide d'un test F qui compare deux modèles. Dans ce cas, on confronte

$$H_0 : E[Y \mid X = x] = \beta_0 \qquad \text{à} \qquad H_1 : E[Y \mid X = x] = \beta_0 + \beta_1 x.$$

  Pour évaluer l'évidence contre H_0 et en faveur de H_1, il suffit de comparer l'ajustement des deux modèles selon la somme des carrés des résiduels, ce qui se fait de nouveau avec la fonction `anova()`.

```
mod.0 <- lm(life_expectancy ~ 1, data=gapminder.em)
anova(mod.0,mod)
```

```
## Analysis of Variance Table
##
## Model 1: life_expectancy ~ 1
## Model 2: life_expectancy ~ infant_mortality
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    177 12392.2
## 2    176  3094.8  1    9297.4 528.73 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

  Nous comparons la somme des carrés du modèle complet SSE = 3094.8 à la somme des carrés du modèle réduit SSE(R) = 12392.2, en calculant la somme des carrés supplémentaires

$$\text{ExtraSS} = \text{SSE(R)} - \text{SSE} = 9297.4.$$

Plus cette différence est élevée, plus le modèle réduit est considéré comme étant mal ajusté en comparaison avec le modèle complet. Si l'évidence est forte (ici, une réduction de près de 75% en passant du modèle réduit au modèle complet), cela suggère que l'on devrait rejeter H_0 en faveur de l'hypothèse H_1 que la pente est non-nulle.

3. Variable explicative binaire

3.1 Variable explicative binaire II

On mène une étude sur le développement de l'ectomycorhizue, une relation symbiotique entre les racines des arbres et un champignon, dans laquelle les minéraux sont transférés du champignon aux arbres et du sucre des arbres au champignon. 20 chênes rouges du nord exposés au champignon *pisolithus tinctorus* ont été cultivés dans une serre; tous les chênes ont été plantés dans le même type de terre et ont reçu la même quantité de soleil et d'eau. La moitié des spécimens (choisis au hasard) ont reçu un traitement de 368 ppm d'azote sous la forme NaNO3; les autres, non (X). On mesure la masse de la tige, en grammes, après 140 jours (Y).

Les données sont les suivantes:

```
azote = as.data.frame(unclass(data.frame(read.csv("Data/Azote.csv"))),
                              stringsAsFactors=TRUE)
azote
```

Masse	Azote
0.59	non
0.47	non
0.25	non
0.36	non
0.42	non
0.19	non
0.38	non
0.39	non
0.45	non
0.48	non
0.35	oui
0.50	oui
0.83	oui
0.77	oui
0.54	oui
0.64	oui
0.62	oui
0.64	oui
0.64	oui
0.65	oui

`R` utilise la première catégorie rencontrée en tant que catégorie de référence: les chênes n'ayant pas reçu d'azote forment ainsi le groupe de référence (il est possible de changer l'ordre des catégories afin que le groupe de traitement d'azote devienne le groupe de référence, à l'aide de: `azote$Azote <- factor(azote$Azote, levels=c("oui","non"))`, par exemple.) En général, c'est souvent le groupe de contrôle qui est utilisé comme groupe de référence, ce qui est déjà le cas ici.

Voici quelques statistiques descriptives pour la masse de la tige dans chacun des groupes:

```
library(dplyr)
azote.s <- azote |> group_by(Azote) |>
  summarise(mean = mean(Masse),
            var = var(Masse),
            n = n()) |>
  as.data.frame()
```

Azote	mean	var	n
non	0.398	0.0132178	10
oui	0.618	0.0180400	10

Si on suppose que les variances des deux populations sont égales, alors on peut l'approximer par la variance pondérée

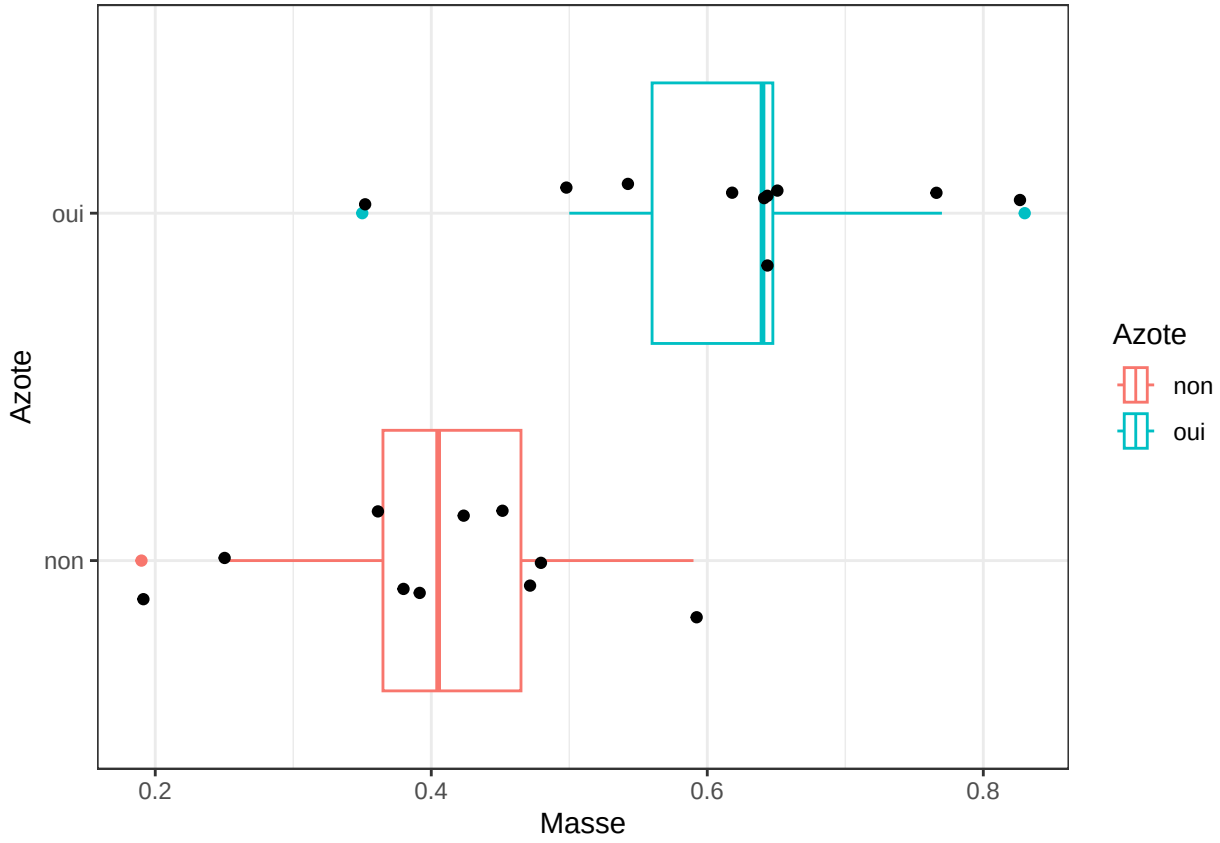
$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = 0.0156289.$$

```
s2p = sum((azote.s$n-1)*azote.s$var)/(sum(azote.s$n)-2)
s2p
```

```
## [1] 0.01562889
```

Pour comparer les deux groupes visuellement, on peut utiliser des diagrammes à boîte comparatifs avec une superposition de points.

```
azote |> ggplot(aes(x=Masse,y=Azote,color=Azote)) +
  geom_boxplot() +
  geom_jitter(color="black", height=0.2) +
  theme_bw()
```



On peut utiliser un test T de Student afin de comparer les deux moyennes. La valeur observée de la statistique du test est

$$t^* = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = -3.935.$$

La valeur p correspondante est

$$2P(T(18) > |-3.935|) = 0.0009707.$$

```
t.test(Masse~Azote,data=azote,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Masse by Azote
## t = -3.935, df = 18, p-value = 0.0009707
## alternative hypothesis: true difference in means between group non and group oui is not equal to 0
## 95 percent confidence interval:
##  -0.3374597 -0.1025403
## sample estimates:
## mean in group non mean in group oui
##             0.398             0.618
```

Le traitement semble certainement avoir un effet. On peut aussi utiliser une approche de régression pour effectuer ce test (ceci va nous aider à généraliser le test à la comparaison de plus de 2 groupes).

Pour identifier les groupes, nous utilisons la variable muette

$$x_i = \begin{cases} 1 & \text{observation } i \text{ dans le groupe de traitement} \\ 0 & \text{autrement} \end{cases}$$

Considérons le modèle de régression linéaire simple: $Y_1, Y_2, ..., Y_n$ sont des variables aléatoires normales, indépendantes, et telles tel que

$$E[Y_i \mid X = x_i] = \beta_0 + \beta_1 x_i = \begin{cases} \beta_0 = \mu_1, & x_i = 0 \\ \beta_0 + \beta_1 = \mu_2, & x_i = 1 \end{cases}$$

et $V[Y_i] = \sigma^2$ for $i = 1, \ldots, n$. Dans notre exemple, nous avons ainsi deux populations normales indépendantes, de variances égales.

Lorsque la variable explicative est catégorique (un `factor` dans la terminologie de `R`), `R` code automatiquement des variables muettes. On affiche ces variables muettes, à l'aide de la fonction `contrasts()`.

```
contrasts(azote$Azote)
```

```
##     oui
## non   0
## oui   1
```

On interprète cette variable de la façon suivante: elle prend la valeur 0 si c'est une observation sans azote; 1 si c'est une observation avec azote.

On peut maintenant ajuster le modèle linéaire pour décrire la masse en fonction du groupe de traitement.

```
mod<-lm(Masse~Azote,data=azote)
mod$coefficients
```

```
## (Intercept)    Azoteoui
##       0.398       0.220
```

La masse moyenne est ainsi:

$$\hat{\mu}_{Y|X=x} = b_0 + b_1 x = \begin{cases} b_0 = 0.398, & x_i = 0 \\ b_0 + b_1 = 0.398 + 0.220 = 0.618, & x_i = 1 \end{cases}$$

L'estimation de la variance σ^2 est MSE $= 0.01562889$.

```
summary(mod)$sigma^2
```

```
## [1] 0.01562889
```

Ce sont les valeurs obtenues de $\overline{y}_1$, $\overline{y}_2$, et s_p^2, respectivement.

Remarquons que $\mu_1 = \mu_2$ si et seulement si $\beta_1 = 0$; lorsque l'on test pour la signification de la variable explicative qui identifie le traitement, nous pouvons aussi l'interpréter comme un test d'égalité des moyennes.

```
summary(mod)$coefficients
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)    0.398 0.03953339 10.06744 8.052524e-09
## Azoteoui       0.220 0.05590866  3.93499 9.707043e-04
```

La valeur observée de la statistique du test est

$$t^* = \frac{b_1}{s\{b_1\}} = 3.935$$

et la valeur p du test est

$$2P(T(18) > |3.935|) = 0.0009707.$$

3.2 Variable explicative binaire II

Nous avons deux variables explicative `x1` (qui est catégorique) et `x2` (qui est quantitative). Nous affichons le codage des variables muettes pour `x1`.

```
> contrasts(x1)
         groupe 2 groupe 3
groupe 1        0        0
groupe 2        1        0
groupe 3        0        1
```

Voici un sommaire de l'ajustement du modèle.

```
> summary(mod)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-25.702 -11.913   0.602   7.663  35.245

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3141    15.1880   0.350 0.729244
x1groupe 2   -5.9883     6.9386  -0.863 0.396005
x1groupe 3   -6.6344     6.9633  -0.953 0.349483
x2            1.1677     0.3003   3.889 0.000624 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.51 on 26 degrees of freedom
Multiple R-squared:  0.3788,    Adjusted R-squared:  0.3072
F-statistic: 5.286 on 3 and 26 DF,  p-value: 0.005573
```

(a) Quelle est la taille n de cette étude.

(b) Donner la fonction de la moyenne estimée pour chacun des 3 niveaux de la variable catégorique `x1`.

(c) Tester pour la signification de la régression. Formuler les hypothèses, donner la statistique du test et la valeur p. Donner la conclusion à $\alpha = 5\%$.

(d) Nous avons ajusté un modèle réduit. Nous allons comparer ce modèle réduit au modèle complet ci-haut. Cette comparaison va nous permettre de tester quelles hypothèses? Formuler les hypothèses, donner la statistique du test, la valeur p et la conclusion du test à $\alpha = 5\%$.

```
> mod0<-lm(y~x2)
> summary(mod0)$sigma
[1] 28.46512
```

Réponses:

(a) On a $n - p = 26$ et $p = 4$, alors, $n = 26 + 4 = 30$.

(b) La fonction de la moyenne estimée est

$$\begin{aligned} E\{Y\} &= 5{,}3141 - 5{,}9883\, I\{x_1 = \text{Groupe 2}\} - 6{,}6344\, I\{x_1 = \text{Groupe 3}\} + 1{,}1677\, x_2 \\ &= \begin{cases} 5{,}3141 + 1{,}1677\, x_2, & \text{si } x_1 = \text{Groupe 1} \\ -0{,}6742 + 1{,}1677\, x_2, & \text{si } x_1 = \text{Groupe 2} \\ -1{,}3203 + 1{,}1677\, x_2, & \text{si } x_1 = \text{Groupe 3} \end{cases} \end{aligned}$$

(c) On veut tester $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ contre H_a: au moins un des ces β est non nul. La régression est significative ($F(3, 26) = 5{,}286; p = 0{,}0056$).

(d) Nous allons tester

$$H_0 : \ E\{Y\} = \beta_0 + \beta_3 \, x_2$$

contre

$$H_a : \ E\{Y\} = \beta_0 + \beta_1 \, I\{x_1 = \text{Groupe 1}\} + \beta_2 \, I\{x_1 = \text{Groupe 2}\} + \beta_3 \, x_2.$$

Ceci est un test pour la signification du prédicteur catégorique `x1`. Il nous faudra la somme de carrés résiduelle pour chacun des modèles.

Pour le modèle complet : On a $15{,}51 = \sqrt{\text{MSE}} = \sqrt{\text{SSE}/(n-p)} = \sqrt{\text{SSE}/26}$, alors $\text{SSE} = (15{,}51)^2(26) = 6254{,}563$.

Pour le modèle réduit : On a $28{,}46512 = \sqrt{\text{MSE(R)}} = \sqrt{\text{SSE(R)}/(n-q)} = \sqrt{\text{SSE}/28}$, alors $\text{SSE} = (28.46512)^2(28) = 22\,687{,}37$.

ExtraSS : La différence des sommes de carrés résiduelle est $\text{ExtraSS} = \text{SSE(R)} - \text{SSE} = 22\,687{,}37 - 6254{,}563 = 16\,432{,}81$.

La statistique du test est

$$F_0 = \frac{\text{ExtraSS}/(p-q)}{\text{SSE}/(n-p)} = \frac{16\,432{,}81/(4-2)}{6254{,}563/26} = 34{,}15531.$$

La valeur p est $P(F(2, 26) > 34{,}15531) = 5{,}31 \times 10^{-8}$. Le prédicteur `x_1` est significatif.

```
> 1-pf(34.15531,2,26)
[1] 5.313317e-08
```

4. Diagnostiques et mesures correctives

Lors de l'évaluation de la pertinence du modèle linéaire, nous devons suivre l'ordre suivant

1. Identifier les valeurs aberrantes et les observations influentes;
2. Test des erreurs dans la spécification de la fonction de la moyenne;
3. Test d'hétéroscédasticité;
4. Testez la normalité des erreurs aléatoires.

4.1 Diagnostique pour la spécification de la fonction de la moyenne

Nous pouvons utiliser le diagramme des résidus contre les valeurs ajustées, et nous pouvons effectuer le test RESET de Ramsey (Regression Equation Specification Error Test).

Considérons le modèle linéaire:

$$E\{Y\} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}.$$

Nous ajustons le modèle pour obtenir les valeurs ajustées

$$\hat{y}_i = b_0 + b_1 x_{1,i} + \cdots + b_{p-1} x_{p-1,i}$$

pour $i = 1, \ldots, n$. Ensuite, on ajoute $\hat{y}^2$ et $\hat{y}^3$ comme prédicteurs dans le modèle:

$$E\{Y\} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3.$$

Le test RESET de Ramsey consiste à tester

$$H_0 : \gamma_1 = \gamma_2 = 0 \quad \text{vs.} \quad H_1 : \gamma_1 \neq 0 \text{ or } \gamma_2 \neq 0.$$

Si le test RESET de Ramsey est significatif, alors nous avons des preuves d'effets d'ordre supérieur manquants dans le modèle. Cela signifie que nous avons des preuves significatives qu'il y a une erreur dans la spécification de la fonction de la moyenne.

Le test de Ramsey est général dans le sens où il peut identifier des erreurs dans la spécification du modèle. Cependant, il ne peut pas nous dire quelle est la cause de l'erreur. Cela signifie qu'il y a des effets non linéaires. Nous devons étudier la relation partielle entre Y et les prédicteurs et essayer de trouver une relation non linéaire. Alternativement, la non-linéarité pourrait être causée par des interactions entre les prédicteurs.

5. Test d'inadéquation de l'ajustement linéaire

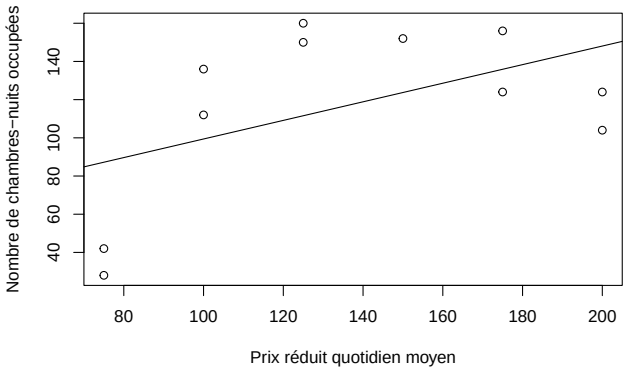
Une chaîne d'hôtels offre une promotion durant le mois de février: dans chacune de 11 succursales, la direction réduit le prix quotidien moyen, qui varie d'une location à l'autre, et prend note du nombre de chambres-nuits supplémentaires qui sont occupé durant le mois.

```
prix.reduit.moyen <- c(125,100,200,75,150,175,75,175,125,200,100)
n.chambres.supp   <- c(160,112,124,28,152,156,42,124,150,104,136)
hotels            <- data.frame(prix.reduit.moyen,n.chambres.supp)
str(hotels)
```

```
## 'data.frame':    11 obs. of  2 variables:
##  $ prix.reduit.moyen: num  125 100 200 75 150 175 75 175 125 200 ...
##  $ n.chambres.supp  : num  160 112 124 28 152 156 42 124 150 104 ...
```

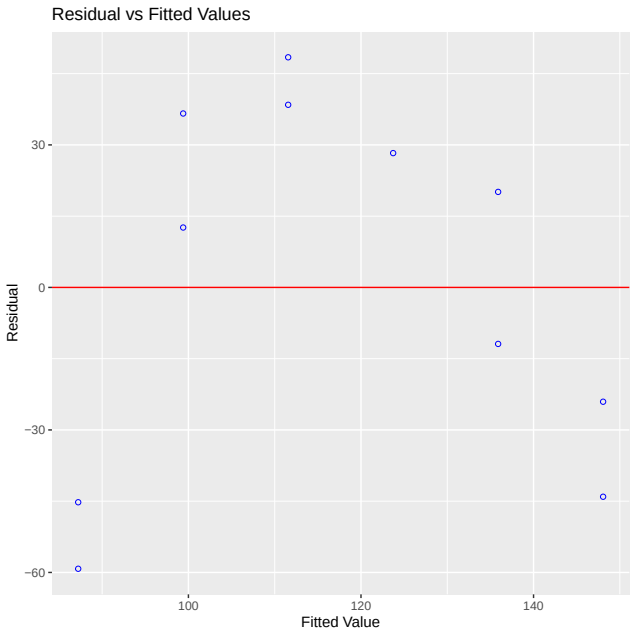
Visuellement, il semble évident que l'association entre les deux variables n'est pas linéaire.

```
plot(hotels$prix.reduit.moyen,hotels$n.chambres.supp,
     xlab="Prix réduit quotidien moyen",
     ylab="Nombre de chambres-nuits occupées")
mod <- lm(n.chambres.supp ~ prix.reduit.moyen, data=hotels)
abline(mod)
```

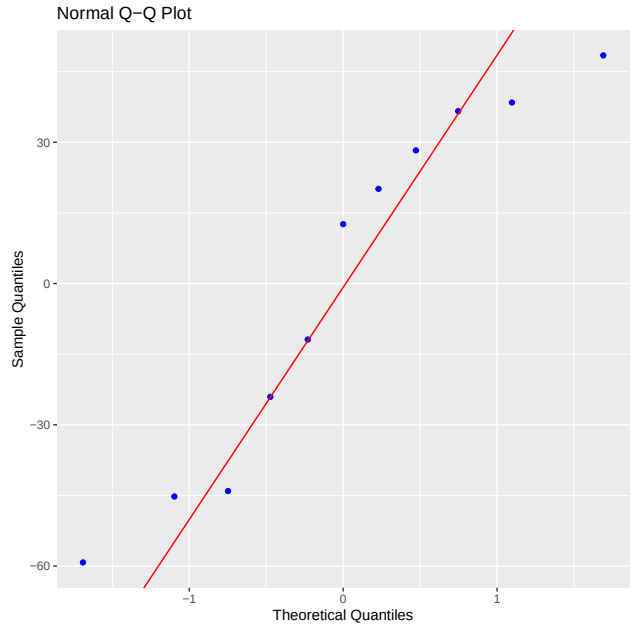


Nous pouvons obtenir les diagrammes de résidus comme suit:

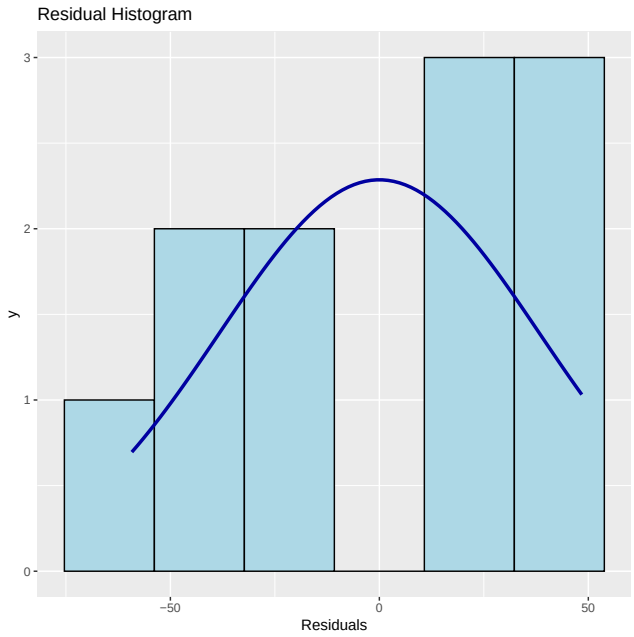
```
# produce residual vs. fitted plot
olsrr::ols_plot_resid_fit(mod)
```




```
# create Q-Q plot for residuals
olsrr::ols_plot_resid_qq(mod)
```




```
# histogram of residuals
olsrr::ols_plot_resid_hist(mod)
```



Nous cherchons à tester

$$H_0 : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad \text{vs.} \quad H_1 : E\{Y\} \neq \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

Pour confronter ces deux hypothèses, nous emboîtons le modèle de régression linéaire simple dans un modèle plus général. On considère une stratification des unités selon la valeur de x, c'est-à-dire que les unités dans le même groupe prennent la même valeur de x.

On obtient un tableau de fréquences pour x =prix réduit quotidien moyen. On observe qu'il y a $c = 6$ groupes et que chaque groupe contient 2 unités sauf pour le groupe $x = 150$ (ce dernier groupe contient une seule unité).

```
table(hotels$prix.reduit.moyen)
```

```
##
##  75 100 125 150 175 200
##   2   2   2   1   2   2
```

Si chaque groupe a sa propre moyenne, on peut considérer que x est une variable catégorielle avec $c = 6$ catégories (ce que l'on peut implémenter dans `R` à l'aide de la fonction `factor()`). Nous allons ajouter une variable catégorielle au *data frame* `hotels` ; on affiche aussi les niveaux de cette variable. Le modèle d'ANOVA correspondant est le modèle le plus complexe possible puisque nous n'imposons aucune structure à $E\{Y \mid X = x\}$.

```
hotels$prix.reduit.moyen.cat <- factor(hotels$prix.reduit.moyen)
levels(hotels$prix.reduit.moyen.cat)
```

```
## [1] "75"  "100" "125" "150" "175" "200"
```

Le modèle linéaire (complet) est

$$Y_i = \beta_0 + \beta_1 x_{i,2} + \cdots + \beta_5 x_{i,6} + \varepsilon_i = \begin{cases} \beta_0 = \mu_1 & \text{si la } i\text{ème unité provient du groupe 1} \\ \beta_0 + \beta_1 = \mu_2 & \text{si la } i\text{ème unité provient du groupe 2} \\ \vdots & \vdots \\ \beta_0 + \beta_5 = \mu_6 & \text{si la } i\text{ème unité provient du groupe 6} \end{cases}$$

où $\varepsilon_1, \ldots, \varepsilon_n$ sont des variables aléatoires i.i.d. normales $\mathcal{N}(0, \sigma^2)$. Ce modèle est parfois appellé un modèle d'ANOVA; le paramètre $\beta_{j-1} = \mu_j - \mu_1$ est l'**effet de groupe** j (en comparaison avec le groupe de référence 1).

Voici un sommaire de l'ajustement du modèle d'ANOVA.

```
mod.ANOVA <- lm(n.chambres.supp ~ prix.reduit.moyen.cat, data=hotels)
summary(mod.ANOVA)
```

```
##
## Call:
## lm(formula = n.chambres.supp ~ prix.reduit.moyen.cat, data = hotels)
##
## Residuals:
##          1          2          3          4          5          6          7
##  5.000e+00 -1.200e+01  1.000e+01 -7.000e+00 -3.331e-15  1.600e+01  7.000e+00
##          8          9         10         11
## -1.600e+01 -5.000e+00 -1.000e+01  1.200e+01
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   35.00      10.71   3.267 0.022282 *
## prix.reduit.moyen.cat100      89.00      15.15   5.874 0.002030 **
## prix.reduit.moyen.cat125     120.00      15.15   7.919 0.000517 ***
## prix.reduit.moyen.cat150     117.00      18.56   6.305 0.001478 **
## prix.reduit.moyen.cat175     105.00      15.15   6.930 0.000960 ***
## prix.reduit.moyen.cat200      79.00      15.15   5.214 0.003428 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.15 on 5 degrees of freedom
## Multiple R-squared:  0.9423, Adjusted R-squared:  0.8845
## F-statistic: 16.32 on 5 and 5 DF,  p-value: 0.004085
```

Le modèle estimé est ainsi

$$\hat{E}\{Y \mid X = x\} = \begin{cases} \overline{y}_1 = 35 & \text{si } x = 75 \\ \overline{y}_2 = 35 + 89 = 124 & \text{si } x = 100 \\ \overline{y}_3 = 35 + 120 = 155 & \text{si } x = 125 \\ \overline{y}_4 = 35 + 117 = 152 & \text{si } x = 150 \\ \overline{y}_5 = 35 + 105 = 140 & \text{si } x = 175 \\ \overline{y}_6 = 35 + 79 = 114 & \text{si } x = 200 \end{cases}$$

L'estimation de la variance (de l'erreur) est $s^2_\varepsilon = \text{MSE} = (15.15)^2 = 229.52$; le coefficient de détermination du modèle d'ANOVA est $R^2 = 0.9423$. Le coefficient de détermination du modèle réduit (le modèle de régression linéaire simple obtenu un peu plus tôt), cependant, est $R^2 = 0.2586$.

```
mod <- lm(n.chambres.supp ~ prix.reduit.moyen, data=hotels)
summary(mod)$r.squared
```

```
## [1] 0.2585808
```

La différence dans l'ajustement des deux modèles semble démontrer que la supposition de linéarité du modèle réduit n'est pas justifiée. L'évidence est-elle significative?

Nous utilisons le test linéaire général afin de comparer les deux modèles. En général, la statistique de test est

$$F = \frac{\text{ExtraSS}/(p-q)}{\text{MSE}} = \frac{(\text{SSE}(R) - \text{SSE})/(p-q)}{\text{MSE}}$$

où p est le nombre de paramètres du modèle complet (ANOVA) et SSE sa somme de carrés des résidus (erreurs), q le nombre de paramètres du modèle réduit (linéaire) et SSE(R) sa somme de carrés des résidus, et MSE est l'écart-type des résidus du modèle complet ; si H_0 est valide, nous avons $F \sim F(p-q, n-p)$.

Dans le modèle complet, il y a $p = c = 6$ paramètres ; dans le modèle réduit, il n'y en a $q = 2$ (puisque $p - q = c - 2 > 0$, on doit avoir au moins $c = 3$ valeurs de x). La valeur observée de la statistique de test est calculée en utilisant la fonction `anova()`.

```
anova(mod,mod.ANOVA)
```

```
## Analysis of Variance Table
##
## Model 1: n.chambres.supp ~ prix.reduit.moyen
## Model 2: n.chambres.supp ~ prix.reduit.moyen.cat
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1      9 14742
## 2      5  1148  4     13594 14.801 0.005594 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nous avons ainsi

$$F_0 = \frac{(\text{SSE} - \text{SSE}(R))/(c-2)}{\text{MSE}} = \frac{(14742 - 1148)/(6-2)}{1148/5} = 14.801;$$

la valeur$-p$ du test est $P(F(4,5) > 14.801) = 0.0056$, d'ou l'on rejete l'hypothèse nulle de linéarité de l'ajustement.

6. Corrélations

6.1 Nuages de points I

Considérons les nuages de points ci-dessous.

Faire une correspondance avec les corrélations suivantes $-0{,}977$, $-0{,}021$, $0{,}736$, et $0{,}951$, et les nuages de points ci-haut.

Réponses: (a) $-0{,}977$ (b) $0{,}736$ (c) $0{,}951$ (d) $-0{,}021$

6.2 Nuages de points II

La puissance produite par un moteur est mesurée en chevaux-vapeur (ou chevaux, plus simplement). Elle correspond à la puissance nécessaire pour lever, sur une hauteur d'un pied, un poids de 550 livres en une seconde ou de 33 000 livres en une minute. Elle est mesurée en fonction de la vitesse à laquelle le travail est effectué.

Dans le fichier `Fuel_economy_2007.csv`, on a les cotes de puissance (en chevaux) annoncées et la consommation d'essence prévue (en mpg) pour plusieurs véhicules en 2007.

Nous importons les données avec R et nous affichons quelques rangés du jeu de données.

```
voitures<-read.csv("Data/Fuel_economy_2007.csv")
head(voitures)
```

```
##             Vehicle Horsepower Highway.Gas.Mileage..mpg.
## 1           Audi A4        200                         32
## 2           BMW 328        230                         30
## 3    Buick LaCrosse        200                         30
## 4      Chevy Cobalt        148                         32
## 5 Chevy TrailBlazer        291                         22
## 6   Ford Expedition        300                         20
```

(a) Donner un nuage de points de la puissance contre la consommation d'essence et superimposer une courbe lisse. Décrire l'orientation et la forme de l'association.

(b) Voici quelques statistiques obtenues avec `R`.

```
x <- voitures$Horsepower
y <- voitures$Highway.Gas.Mileage..mpg.
sum((x-mean(x))^2)
```

```
## [1] 61503.6
```

```
sum((y-mean(y))^2)
```

```
## [1] 572.4
```

```
sum((x-mean(x))*(y-mean(y)))
```

```
## [1] -5154.2
```

Remarques:

- `y-mean(y)` est le vecteur $(y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})$.
- `sum` calcul la somme des composantes du vecteur.
- `mean` calcul la moyenne des composantes du vecteur.

En utilisant ces statistiques, calculer la corrélation de Pearson entre la puissance contre la consommation d'essence.

(c) On supposant que `x` et `y` sont des vecteurs numériques de même taille, alors la commande `cor(x,y)` calcul la corrélation de ¨Pearson entre `x` et `y`. Utiliser la fonction `cor` pour calculer la corrélation entre la puissance et la consommation d'essence.

(d) Au Canada, on décrit la consommation d'essence en L/100km. Nos données sont mesurées en mpg. Soit w la consommation en L/100km et y la consommation en mpg. Voici une formule de conversion:

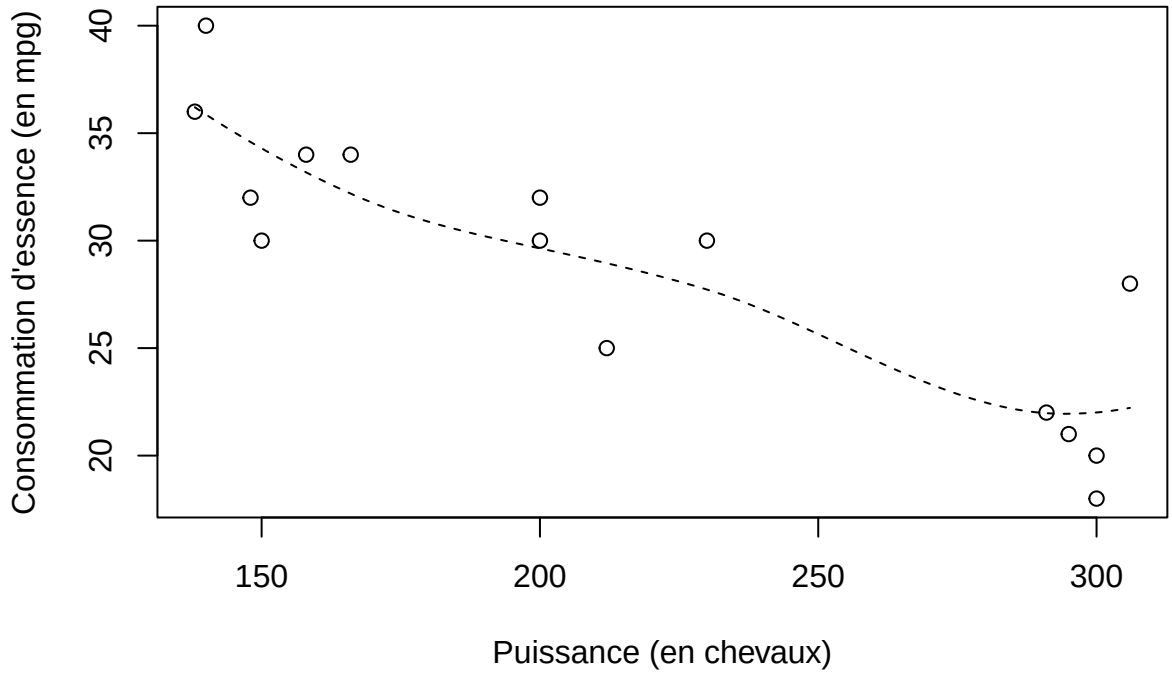
$$w = \frac{235{,}215}{y}.$$

Si on mesure la consommation d'essence en mpg, alors la corrélation entre la puissance et la consommation d'essence est $r = -0{,}869$. Si nous mesurons la consommation en L/100km, est que la corrélation entre la puissance et la consommation d'essence demeure égal à $r = -0{,}869$? Sinon, calculer la correcte valeur de la corrélation.

Réponses:

(a) L'association entre la puissance et la consommation d'essence (en mpg) est approximativement linéaire et négative.

```
with(voitures,
plot(x=Horsepower,y=Highway.Gas.Mileage..mpg.,
xlab="Puissance (en chevaux)",
ylab="Consommation d'essence (en mpg)"))
## ajuster une courbe lisse (loess=lowess=locally weighted scatterplot smoothing)
mod.loess<-loess(Highway.Gas.Mileage..mpg.~Horsepower,
data=voitures)
## obtenir l'étendue pour x
xlim<-range(voitures$Horsepower)
## construire un nouveau jeu de données
xnew<-seq(xlim[1],xlim[2],length.out=100)
ynew<-predict(mod.loess,data.frame(Horsepower=xnew))
## add Lowess Smooth to the plot
lines(x=xnew,y=ynew,lty=2)
```



(b) On a $s_{xy} = -5154{,}2$, $s_{xx} = 61503{,}6$, et $s_{yy} = 572{,}4$. Alors la corrélation de Pearson entre x et y est

$$r = \frac{s_{xy}}{\sqrt{s_{xx}\, s_{yy}}} = -0{,}869.$$

(c) La corrélation entre la puissance et la consommation d'essence est $r = -0{,}867$.

```
with(voitures,cor(Horsepower,Highway.Gas.Mileage..mpg.))
```

```
## [1] -0.8686827
```

(d) La formule de conversion n'est pas linéaire. Alors, c'est possible que la corrélation puisse changer de valeur si nous mesurons la consommation d'essence en L/100km. Nous utilisons R ci-bas pour convertir la consommation en L/100km, et ensuite calculer la corrélation entre la puissance et la consommation d'essence (enL/100km). Cette corrélation est $r = 0{,}851$.

```
w<-235.215/voitures$Highway.Gas.Mileage..mpg.
cor(w,voitures$Horsepower)
```

```
## [1] 0.8511043
```

6.3 Nuages de points III

1. La masse musculaire d'une personne devrait diminuer avec l'âge. Pour explorer cette association chez les femmes, un nutritionniste a sélectionné au hasard 15 femmes parmi chacun des tranches d'âges de 10 ans, commençant à 40 ans et se terminant à 79 ans. Les données sont dans le fichier `Masse.csv`. Il y a deux variables dans ce jeu de données: x = l'âge de la participante, et y = la masse musculaire de la partipante.

Nous importons les données avec R, et nous affichons quelques rangés du jeu de données.

```
masse<-read.csv("Data/Masse.csv")
head(masse)
```

```
##   Masse Age
## 1   106  43
## 2   106  41
## 3    97  47
## 4   113  46
## 5    96  45
## 6   119  41
```

(N.B. Avec R markdown, on doit sauvegarder les fichier CSV dans le même dossier que le fichier .rmd.)

Voici la structure du jeu de données.

```
str(masse)
```

```
## 'data.frame':    60 obs. of  2 variables:
##  $ Masse: int  106 106 97 113 96 119 92 112 92 102 ...
##  $ Age  : int  43 41 47 46 45 41 47 41 48 48 ...
```

(a) Il y a combien d'observations dans ce jeu de données.

(b) On calcul quelques sommes pour résumer les données.

```
x<-masse$Age
y<-masse$Masse
c(sum(x), sum(y), sum(x^2), sum(y^2), sum(x*y))
```

```
## [1]   3599   5098 224091 448662 296024
```

```
rx<-rank(masse$Age)
ry<-rank(masse$Masse)
c(sum(rx), sum(ry), sum(rx^2), sum(ry^2), sum(rx*ry))
```

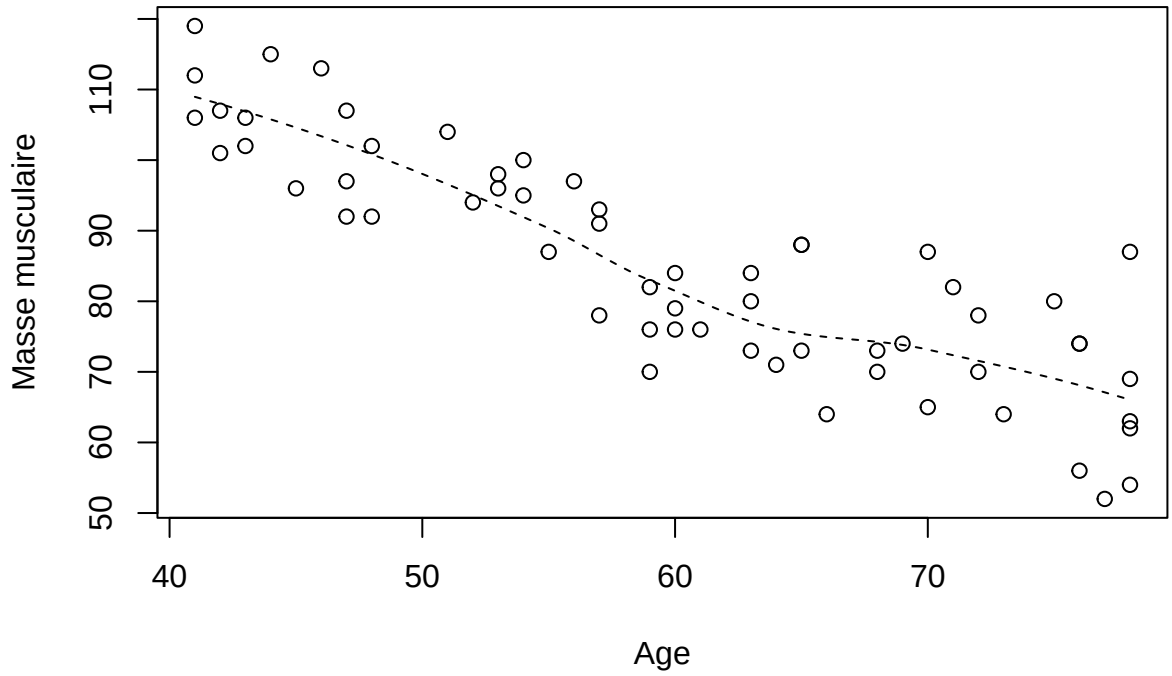
```
## [1]  1830.00  1830.00 73780.00 73794.00 40256.25
```

- (i) En se basant sur ces sommes, calculer la covariance entre l'âge et la masse et aussi calculer la corrélation (de Pearson) entre l'âge et la masse.
- (ii) En se basant sur ces sommes, calculer la corrélation de Spearman entre l'âge et la masse.

(c) Voici un nuage de points de la masse musculaire contre l'âge avec une superposition d'une courve lisse. Décrire l'association entre l'âge et la masse musculaire (orientation, forme, et intensité).

```
with(masse,
plot(x=Age,y=Masse,
xlab="Age",
```

```
ylab="Masse musculaire"))
## ajuster une courbe lisse
mod.loess<-loess(Masse~Age,
data=masse)
## obtenir l'étendue pour x
xlim<-range(masse$Age)
## construire un nouveau jeu de données
xnew<-seq(xlim[1],xlim[2],length.out=100)
ynew<-predict(mod.loess,data.frame(Age=xnew))
## add Lowess Smooth to the plot
lines(x=xnew,y=ynew,lty=2)
```



Réponses:

(a) Il y a $n = 60$ observations.

(b) On calcul quelques sommes pour résumer les données.

```
x<-masse$Age
y<-masse$Masse
c(sum(x), sum(y), sum(x^2), sum(y^2), sum(x*y))
```

```
## [1]   3599   5098 224091 448662 296024
```

```
rx<-rank(masse$Age)
ry<-rank(masse$Masse)
c(sum(rx), sum(ry), sum(rx^2), sum(ry^2), sum(rx*ry))
```

```
## [1]  1830.00  1830.00 73780.00 73794.00 40256.25
```

- (i) On a

$$s_{xy} = \left(\sum_{i=1}^{n} x_i\, y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \Big/ n = 296\,024 - (3599)(5098)/60 = -9771{,}033,$$

alors la covariance entre l'âge et la masse est

$$\hat{\sigma}_{X,Y} = \frac{s_{xy}}{n-1} = \frac{-9771{,}033}{60-1} = -165{,}6107.$$

En outre, on a

$$s_{xx} = \left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n = 224\,091 - (3599)^2/60 = 8\,210{,}983,$$

$$s_{yy} = \left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n = 448\,662 - 5098^2/60 = 15\,501{,}93,$$

alors la corrélation de Pearson entre l'âge et la masse est

$$r = \frac{s_{xy}}{\sqrt{s_{xx}\, s_{yy}}} = \frac{-9\,771{,}033}{\sqrt{(8\,210{,}983)(15\,501{,}93)}} = -0{,}866.$$

- (ii) On a

$$\begin{aligned} s_{R_x,R_y} &= \left(\sum_{i=1}^{n} R_{x,i}\, R_{y,i}\right) - \left(\sum_{i=1}^{n} R_{x,i}\right)\left(\sum_{i=1}^{n} R_{y,i}\right) \Big/ n \\ &= 40\,256{,}25 - (1830)(1830)/60 = -15\,558{,}75. \end{aligned}$$

En outre, on a

$$s_{R_xR_x} = \left(\sum_{i=1}^{n} R_{x,i}^2\right) - \left(\sum_{i=1}^{n} R_{x,i}\right)^2 \Big/ n = 73\,780 - (1830)^2/60 = 17\,965,$$

$$s_{R_yR_y} = \left(\sum_{i=1}^{n} R_{y,i}^2\right) - \left(\sum_{i=1}^{n} R_{y,i}\right)^2 \Big/ n = 73\,794 - 1830^2/60 = 17\,979,$$

alors la corrélation de Spearman entre l'âge et la masse est

$$r_S = \frac{s_{R_xR_y}}{\sqrt{s_{R_xR_x}\, s_{R_yR_y}}} = \frac{-15\,558{,}75}{\sqrt{(17\,965)(17\,979)}} = -0{,}8657.$$

(c) L'association entre l'âge et la masse est approximativement linéaire, et négative, avec une corrélation de Pearson de $r = -0{,}866$.

6.4 Inférence concernant une corrélation

Nous avons des données provenant d'une étude avec des volontaires en bonne santé. Un stimulus est appliqué aux doigts du sujet et on mesure la vitesse de conduction de la moelle épinière (VC). On veut décrire l'association entre la taille de l'indivu (en cm) et la vitesse de conduction de la moelle épinière pour les individus en bonne santé.

On importe les données avec R et on affiche quelques rangés.

```
VC <- read.csv("Data/VC.csv")
head(VC)
```

```
##   Taille.en.cm   VC
## 1          149 14.4
## 2          149 13.4
## 3          155 13.5
## 4          155 13.5
## 5          156 13.0
## 6          156 13.6
```

Testons $H_0 : \rho = 0$ (o'u ρ est la corrélation (de Pearson) entre la vitesse de conduction et la taille de l'individu) vs $H_1 : \rho \neq 0$.

Réponse:

La valeur de la corrélation dans l'ensemble de données est :

```
cor(VC$Taille.en.cm,VC$VC)
```

```
## [1] 0.8478829
```

Cette valeur n'est pas très proche de 0. On peut utiliser la fonction `cor.test()` afin d'obtenir un intervalle de confiance de la corrélation ρ:

```
with(VC,cor.test(Taille.en.cm,VC))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Taille.en.cm and VC
## t = 19.781, df = 153, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7967316 0.8869721
## sample estimates:
##       cor
## 0.8478829
```

La valeur observée de la statistique du test est

$$t^* = r\sqrt{\frac{n-2}{1-r^2}} = 19.781$$

La valeur$-p$ du test est $2P(T(153) > |19.781|) < 0.00001$; l'évidence suggère que la correeee2élation est non-nulle a p

7. Probabilités et statistiques

7.1 Probabilities I

Calculer les probabilités suivante tel que T suit une loi $t(15)$.

(a) $P(T > 2{,}45)$;

(b) $P(T < 2{,}45)$;

(c) $2P(T > 4{,}34)$.

Réponses:

(a) $P(T > 2{,}45) = 0{,}0135$;

```
1-pt(2.45,15)
```

```
## [1] 0.01352069
```

(b) $P(T < 2{,}45) = 0{,}9865$;

```
pt(2.45,15)
```

```
## [1] 0.9864793
```

(c) $2P(T > 4{,}34) = 0{,}00058$.

```
2*(1-pt(4.34,15))
```

```
## [1] 0.0005829995
```

7.2 Probabilities II

Obtenir les quantiles suivants.

(a) le 95$^{\text{e}}$ centile de la loi $t(34)$;

(b) le 97,5$^{\text{e}}$ centile de la loi $t(44)$.

Réponses:

(a) Le 95$^{\text{e}}$ centile de la loi $t(34)$ est $t(0{,}95; 34) = 1{,}6909$.

```
qt(0.95,34)
```

```
## [1] 1.690924
```

(b) le 97,5$^{\text{e}}$ centile de la loi $t(44)$ est $t(0{,}975; 44) = 2{,}0154$.

```
qt(0.975,44)
```

```
## [1] 2.015368
```

7.3 Probabilities III

Soit $Y \sim N(\mu = 125, \sigma^2 = 25)$, et soit $(1/\sigma^2)\, V \sim \chi^2(10)$. Supponsons que Y et V sont indépendantes.

(a) Calculer

$$P\left(\frac{Y - 125}{\sqrt{V/10}} > 2{,}75\right).$$

(b) Calculer

$$P\left(\frac{(Y-125)^2}{V/10} > 7{,}12\right).$$

Réponses:

On a

$$Z = \frac{Y-125}{\sigma} \sim N(0,1),$$

et Z et indépendant de $U = (1/\sigma^2)\,V \sim \chi^2(10)$, alors

$$T = \frac{Y-125}{\sqrt{V/10}} = \frac{(Y-125)/\sigma}{\sqrt{(1/\sigma^2)V/10}} = \frac{Z}{\sqrt{U/10}} \sim t(10).$$

En outre,

$$\frac{(Y-125)^2}{V/10} = T^2 \sim F(1,10).$$

(a) On

$$P\left(\frac{Y-125}{\sqrt{V/10}} > 2{,}75\right) = P(t(10) > 2{,}75) = 0{,}0102.$$

```
1-pt(2.75,10)
```

```
## [1] 0.01023912
```

(b) On veut

$$P\left(\frac{(Y-125)^2}{V/10} > 7{,}12\right) = P(F(1,10) > 7{,}12) = 0{,}02356.$$

```
1-pf(7.12,1,10)
```

```
## [1] 0.02355988
```

7.4 Statistiques I

Supposons que $\hat{\theta}$ est un estimateur d'un paramètre inconnu θ, tel que

$$T = \frac{\hat{\theta}-\theta}{\text{se}(\hat{\theta})} \sim t(15).$$

D'un échantillon aléatoire, on observe $\hat{\theta} = -3{,}2$ et $s\{\hat{\theta}\} = 4{,}5$.

a. Tester $H_0:\ \theta = 0$ contre $H_a:\ \theta \neq 0$ 'a $\alpha = 5\%$. Donner la valeur observée de la statistique t du test et la conclusion du test.

b. Donner un intervalle de confiance à 95% pour θ.

Réponses:

a. La valeur observée de la statistique t du test est

$$t_0 = \frac{\hat{\theta}-0}{s\{\hat{\theta}\}} = \frac{-3{,}2-0}{4{,}5} = -0{,}71111.$$

Puisque $|t_0|=0,!71111<2,!13145=t(0,!975;10)$, alors les preuves contre H_0 ne sont pas significatives à $\alpha = 5\%$.

b. Un intervalle de confiance à 95% pour θ est

$$\hat{\theta} \pm t(0{,}975;15)\, s\{\hat{\theta}\} = -3.2 \pm 2.13145\,(4.5) =]-12.792; 6.392[.$$

7.5 Statistiques II

Supposons que $\hat{\theta}$ est un estimateur d'un param'etre inconnu θ, et que

$$T = \frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} \sim t(20).$$

D'un échantillon aléatoire, on observe $\hat{\theta} = 2{,}5$ et $s\{\hat{\theta}\} = 0{,}75$.

a. Tester $H_0 : \ \theta = 0$ contre $H_a : \ \theta \neq 0$ 'a $\alpha = 5\%$. Donner la valeur observée de la statistique t du test et la conclusion du test.

b. Donner un intervalle de confiance à 95% pour θ.

Réponses:

a. La valeur observée de la statistique t du test est

$$t_0 = \frac{\hat{\theta} - 0}{s\{\hat{\theta}\}} = \frac{2{,}5 - 0}{0{,}75} = 3{,}3333.$$

Puisque $|t_0| = 3{,}3333 \geq 2{,}08596 = t(0{,}975; 20)$, alors les preuves contre H_0 sont significatives à $\alpha = 5\%$.

```
qt(0.975,20)
```

```
## [1] 2.085963
```

b. Un intervalle de confiance 'a 95% pour θ est

$$\hat{\theta} \pm t(0.975; 20)\, s\{\hat{\theta}\} = 2{,}5 \pm 2{,}08596\,(0{,}75) =]0{,}936; 4{,}064[.$$

7.6 Statistiques III

Supposons que $\hat{\theta}$ est un estimateur d'un paramètre inconnu θ, tel que

$$T = \frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} \sim t(28).$$

D'un échantillon aléatoire, on observe $\hat{\theta} = -0{,}211$ et $s\{\hat{\theta}\} = 3{,}235$.

a) Tester $H_0 : \ \theta = 0$ contre $H_a : \ \theta \neq 0$. Donner la valeur observée de la statistique t du test, et la valeur p du test.

b) Donner la conclusion du test de la partie a) à $\alpha = 5\%$.

Réponses:

a. La valeur observée de la statistique t du test est

$$t_0 = \frac{\hat{\theta} - 0}{s\{\hat{\theta}\}} = \frac{-0{,}211 - 0}{3{,}235} = -0{,}06522.$$

La valeur p est $2\,P(t(28) \geq |-0{,}06522|) = 0{,}948$.

```
2*(1-pt(.06522,28))
```

```
## [1] 0.9484623
```

b. Les preuves contre $\theta = 0$ en faveur de $\theta \neq 0$ ne sont pas significatives à $\alpha = 5\%$ $(t(28) = -0{,}06522; p = 0{,}948)$.

7.7 Statistiques IV

Supposons que $\hat{\theta}$ est un estimateur d'un param'etre inconnu θ, et que

$$T = \frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} \sim t(28).$$

D'un échantillon aléatoire, on observe $\hat{\theta} = 2{,}5$ et $s\{\hat{\theta}\} = 0{,}75$.

(a) Tester $H_0 : \ \theta = 0$ contre $H_a : \ \theta \neq 0$. Donner la valeur observée de la statistique t du test et la valeur p.

(b) Donner la conclusion du test de la partie a) à $\alpha = 5\%$.

Réponses:

(a) La valeur observée de la statistique t du test est

$$t_0 = \frac{\hat{\theta} - 0}{s\{\hat{\theta}\}} = \frac{2{,}5 - 0}{0{,}75} = 3{,}3333.$$

La valeur p est $2\,P(t(28) \geq |3{,}3333|) = 0{,}0024$.

```
2*(1-pt(3.3333,28))
```

```
## [1] 0.0024247
```

(b) Les preuves contre $\theta = 0$ en faveur de $\theta \neq 0$ sont significatives à $\alpha = 5\%$ $(t(28) = 3{,}3333; p = 0{,}0024)$.

7.8 Probabilités IV

Supponsons que Y_1, Y_2, Y_3 sont des variables aléatoires indépendantes et normales tel que

$$\mu_1 = E\{Y_1\} = 23; \mu_2 = E\{Y_2\} = 15; \mu_3 = E\{Y_3\} = 10$$

et

$$\sigma_1^2 = V[Y_1] = 2; \sigma_2^2 = V[Y_2] = 3; \sigma_3^2 = V[Y_3] = 1.$$

Quelle est la loi de probabilité à $W = 2\,Y_1 + 3\,Y_2 - Y_3$?

Réponse:

On a $W \sim N(E\{W\}; V[W])$ où

$$E\{W\} = 2\,E\{Y_1\} + 3\,E\{Y_2\} - E\{Y_3\} = 2\,(23) + 3\,(15) - 10 = 81$$

et

$$V[W] = 2^2\,V[Y_1] + 3^2\,V[Y_2] + (-1)^2\,V[Y_3] = 2^2\,(2) + 3^2\,(3) + (-1)^2(1) = 36.$$

8. Ajustement d'un modèle linéaire

8.1 Régression linéaire I

Supposons que $\hat{y} = b_0 + b_1\, x$ est un modèle linéaire estimé par la méthode des moindres carrés. Trouver les valeurs maquantes dans le tableau ci-bas.

Notation:

- $\bar{x}$ et s_x sont la moyenne et l'écart type de l'échantillon pour la variable X.
- $\bar{y}$ et s_y sont la moyenne et l'écart type de l'échantillon pour la variable Y.
- r est la corrélation de Pearson (de l'échantillon) entre X et Y.

	$\overline{x}$	s_x	$\overline{y}$	s_y	r	$\hat{y} = b_0 + b_1\, x$
i)	30	4	18	6	-0,2	
ii)	100	18	60	10	0,9	
iii)		0,8	50	15		$\hat{y} = -10 + 15\, x$
iv)			18	4	-0,6	$\hat{y} = 30 - 2\, x$

Rappel de vos cours d'intro à la statistique: La variance et l'écart type de l'échantillon pour X sont respectivement

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{s_{xx}}{n-1} \quad \text{et} \quad s = \sqrt{s^2} = \sqrt{s_{xx}/(n-1)}.$$

Pour Y, c'est semblable. L'écart type de l'échantillon est $s_y = \sqrt{s_{yy}/(n-1)}$.

Réponses:

1. La droite estimée est $\hat{y} = b_0 + b_1\, x$, où

$$b_1 = \frac{r\,\sqrt{s_{yy}}}{\sqrt{s_{xx}}} = \frac{r\,\sqrt{s_{yy}/(n-1)}}{\sqrt{s_{xx}/(n-1)}} = r\,\frac{s_y}{s_x} \quad \text{et} \quad b_0 = \overline{y} - b_1\,\overline{x}.$$

Voici le tableau avec les valeurs manquantes. Les calculs sont sous le tableau.

	$\overline{x}$	s_x	$\overline{y}$	s_y	r	$\hat{y} = b_0 + b_1\, x$
i)	30	4	18	6	-0,2	$\hat{y} = 27 - 0{,}3\, x$
ii)	100	18	60	10	0,9	$\hat{y} = 10 + 0{,}5\, x$
iii)	4	0,8	50	15	0,8	$\hat{y} = -10 + 15\, x$
iv)	6	1,2	18	4	-0,6	$\hat{y} = 30 - 2\, x$

(i) On a

$$b_1 = \frac{r\, s_y}{s_x} = \frac{(-0{,}2)(6)}{4} = -0{,}3 \quad \text{et} \quad b_0 = \overline{y} - b_1\,\overline{x} = 18 - (-0{,}3)(30) = 27.$$

(ii) On a

$$b_1 = \frac{r\, s_y}{s_x} = \frac{(0{,}9)(10)}{18} = 0{,}5 \quad \text{et} \quad b_0 = \overline{y} - b_1\,\overline{x} = 60 - (0{,}5)(100) = 10.$$

(iii) On a

$$15 = b_1 = \frac{r\, s_y}{s_x} = \frac{r\,(15)}{0{,}8} \quad \text{et} \quad -10 = b_0 = \overline{y} - b_1\,\overline{x} = 50 - (15)\,\overline{x}.$$

Thus, $r = 15\,(0.8)/15 = 0{,}8$ et $\overline{x} = (50 - (-10))/15 = 4$. \\ (iv) On a

$$-2 = b_1 = \frac{r\, s_y}{s_x} = \frac{(-0.6)\,(4)}{s_x} \quad \text{et} \quad 30 = b_0 = \overline{y} -_1\ \overline{x} = 18 - (-2)\,\overline{x}.$$

Donc, $s_x = (-0{,}6)\,(4)/(-2) = 1{,}2$ et $\overline{x} = (30 - 18)/2 = 6$.

8.2 Régression linéaire II

Voici un nuage de points et de données.

Une enquête sur le coût de la vie a déterminé le coût de la vie dans les 25 villes les plus coûts de la vie du monde. Ce classement considère la ville de New York comme 100 et exprime les autres villes en pourcentage du coût de la vie à New York. Par exemple, le coût de la vie à Tokyo en 2007 est de 122,1, donc le coût de la vie à Tokyo était de 22,1% plus élevé qu'à New York en 2007. L'écart type du coût de la vie en 2007 est de 11,9147; alors que c'est de 10,8517 pour 2008.

(i) La droite par les moindres carrés pour prévoir le coût de la vie en 2008 en fonction du coût de la vie en 2007 est

$$\widehat{\text{coût08}} = 21{,}75 + 0{,}84\,(\text{coût07}).$$

Calculez la corrélation (de Pearson) entre le coût de la vie en 2007 et 2008.

(ii) Décrivez l'association entre le coût de la vie en 2007 et 2008.

(iii) Calculer le coefficient de détermination R^2 et interpréter sa valeur dans le contexte de cette étude.

(iv) Utilisez la droite des moindres carrés de (i) pour calculer le résidu pour Oslo.

(v) Que nous dit le résidu calculé en (iv) sur Oslo?

Réponses:

(i) On a

$$0{,}84 = b_1 = r\,s_y/s_x = r\,(10{,}8517/11{,}9147) \quad \Rightarrow \quad r = (11{,}9147)(0{,}84)/10{,}8517 = 0{,}922.$$

(ii) L'association entre le coût de la vie en 2007 et 2008 est positive et linéaire avec une corrélation de 0,922.

(iii) $R^2 = r^2 = (0{,}922)^2 = 0{,}8501$. Alors, 85% de la variabilité dans les coûts de la vie en 2008 est expliquée par le modèle linéaire du coût de la vie en 2007.

(iv) On a observé coût07 = 105,8 et coût08 = 118,3 pour Oslo. La valeur ajustée pour Oslo est

$$\widehat{\text{coût08}} = 21{,}75 + 0{,}84\,(105{,}8) = 110{,}622.$$

Alors, le résidu pour Oslo est $e = \text{coût08} - \widehat{\text{coût08}} = 118{,}3 - 110{,}622 = 7{,}678$.

(v) Le résidu est positif, donc le coût de la vie prévu pour 2008 est inférieur au coût de la vie observé pour 2008 de 7,678 unités.

8.3 Régression linéaire III

Supposons un modèle de régression où nous supposons que $Y_1, \ldots, Y_{50}$ sont des variable indépendantes normales avec une variance commune σ^2. Nous avons deux modèles pour la fonction de la moyenne. Nous utilisons la méthode des moindres carrés pour estimer le paramètres de la fonction de la moyenne pour les deux modèles. Ensuite, nous calculons la somme de carrés résiduelle pour les deux modèles. On obtient

$$\text{pour le modèle 1: } SSE = 1222; \quad \text{pour le modèle 2: } SSE = 995.$$

(a) Selon la somme de carrés résiduelle, lequel des modèles est le meilleur.

(b) Selon le max du log de vraisemblance, lequel des modèles est le meilleur. Cette réponse vous surprend-elle?

Réponses:

(a) La modèle 2 a la plus petite somme de carrés résiduelle. Alors selon SSE, le modèle 2 est le mieux ajusté aux données.

(b) Pour le modèle 1, on a

$$\ell = -(n/2)\,\ln(2\,\pi\,\text{SSE}/n) - n/2 = -(50/2)\,\ln(2\,\pi\,(1222/50)) - 50/2 = -150{,}8525.$$

Pour le modèle 2, on a

$$\ell = -(n/2)\,\ln(2\,\pi\,\text{SSE}/n) - n/2 = -(50/2)\,\ln(2\,\pi\,(995/50)) - 50/2 = -145{,}7149.$$

La plus grande de ces deux statistiques est pour le modèle 2. Ainsi, selon la statistique de la log-vraisemblance maximale, le modèle 2 est meilleur. Ce résultat ne devrait pas être surprenant puisque SSE et ℓ sont équivalentes dans le sense qu'elle prèfère toujours le même modèle.

8.4 Régression linéaire IV

Les données concernant la résistance (x) (en ohms) et le temps de défaillance (y) (en minutes) de certaines résisteurs surchargées sont dans le fichier `defaillance.csv`.

Nous importons les données. Nous affichons les noms des colonnes, les écarts types des colonnes, les moyennes des colonnes et la dimension du jeu de données.

```
defaillance <- read.csv("Data/defaillance.csv")
names(defaillance)
```

```
## [1] "resistance"          "temps.de.defaillance"
```

```
sapply(defaillance,sd)
```

```
##           resistance temps.de.defaillance
##             6.781128             8.544852
```

```
sapply(defaillance,mean)
```

```
##           resistance temps.de.defaillance
##             38.62500             33.83333
```

```
dim(defaillance)
```

```
## [1] 24  2
```

Remarque: La fonction `sapply` nous permet d'appliquer une fonction sur toute les colonnes d'un jeu de données (un dataframe). La commande `sapply(defaillance,sd)` applique la fonction `sd` sur toutes les colonnes du jeu de données `defaillance`.

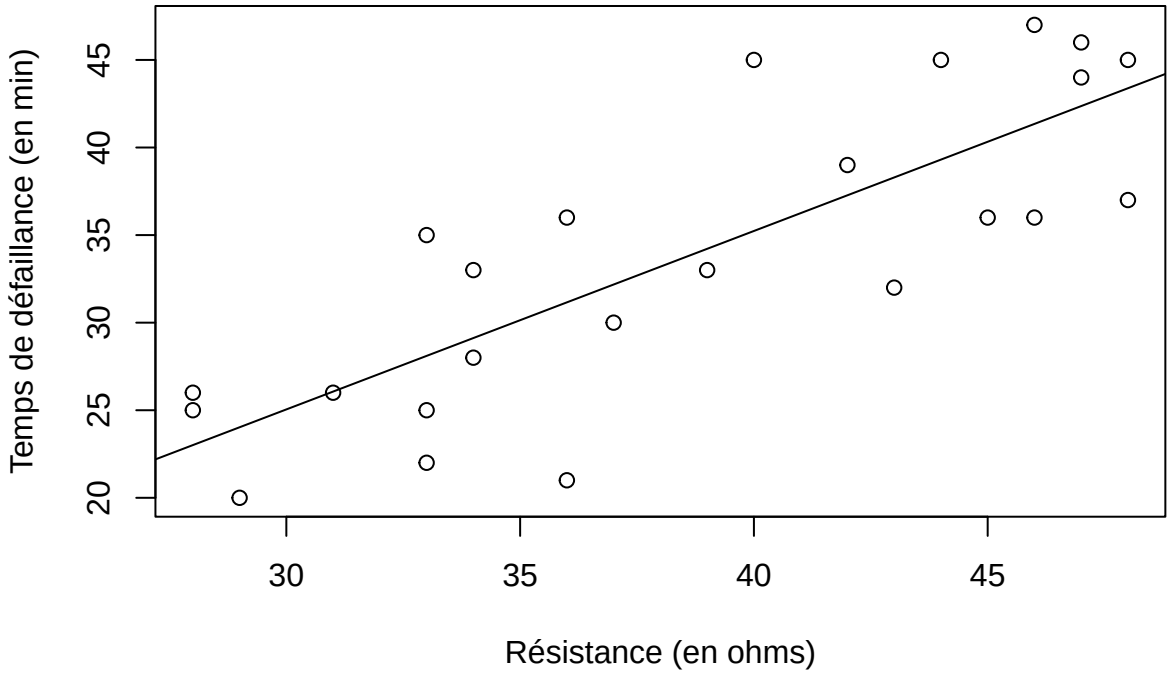
Voici la corrélation (de Pearson) entre le temps de défaillance et la résistance.

```
with(defaillance,cor(temps.de.defaillance,resistance))
```

```
## [1] 0.8085055
```

Voici un nuage de points pour le temps de défaillance contre la résistance avec la superposition de la droite de régression estimée.

```
with(defaillance,plot(resistance,temps.de.defaillance,
  xlab="Résistance (en ohms)",ylab="Temps de défaillance (en min)"))
mod<-lm(temps.de.defaillance~resistance,data=defaillance)
abline(mod)
```



(a) Donner la droite des moindes carrées qui exprime le temps de défaillance en fonction de la résistance.

(b) Donner la valeur de R^2 (le coefficient de détermination) et interpréter dans le contexte de cette question.

(c) Donner une estimation de la variance de l'erreur σ^2.

Réponses:

Dans la sortie de la question, on nous a donné les statistiques suivantes : $s_x = 6{,}78113; s_y = 8{,}54485; \bar{x} = 38{,}62500; \bar{y} = 33{,}83333$; $r = 0{,}80851$; et $n = 24$.

(a) La pente estim'ee est

$$b_1 = r\,\frac{s_y}{s_x} = 0{,}80851\,\left(\frac{8{,}54485}{6{,}78113}\right) = 1{,}01880.$$

L'ordonnée à l'origine estimée est

$$b_0 = \bar{y} - b_1\,\bar{x} = 33{,}83333 - (1{,}01880)(38{,}62500) = -5{,}51782.$$

Alors, la droite estimée est

$$\hat{y} = -5{,}51782 + 1{,}01880\,x.$$

(b) On a $R^2 = r^2 = (0{,}80851)^2 = 0{,}653$. Alors, 65,3% de la variabilité dans le temps de défaillance est expliquée par le modèle linéaire.

(c) On a

$$\begin{aligned} \text{SSE} &= s_{yy} - \text{SSR} = s_{yy} - b_1^2\, s_{xx} = (n-1)s_y^2 - b_1^2\,(n-1)s_{xx}^2 \\ &= (24-1)(8{,}54485)^2 - 1{,}01880^2\,(24-1)(6{,}78113)^2 = 581{,}5664. \end{aligned}$$

L'estimation de σ^2 est

$$\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{581{,}5664}{24-2} = 26{,}43484.$$

8.5 Régression linéaire V

Considérons un modèle de régression linéaire simple. L'estimation de l'ordonnée à l'origine β_0 est $b_0 = \bar{y} - b_1\,\bar{x}$.

(a) Démontrer que $E\{\bar{Y}\} = \beta_0 + \beta_1\,\bar{x}$ où $\bar{Y} = \sum_{i=1}^n Y_i/n$.

(b) Nous avons démontrer durant une leçon que l'estimation de la pente $b_1 = s_{xY}/s_{xx}$ est un estimateur non-biasé de la pente β_1, c'est-à-dire $E\{b_1\} = \beta_1$. Là démontrer que b_0 est un estimateur non-biasé de β_0, c'est-à-dire $E\{b_0\} = \beta_0$.

(c) Démontrer que b_0 peut être écrit sous la forme suivante

$$b_0 = \sum_{i=1}^n \left[\frac{1}{n} - \frac{(x_i - \bar{x})}{s_{xx}}\right] Y_i.$$

(d) Est-ce que b_0 est une variable aléatoire normale? (Pourquoi?)

(e) Démontrer que

$$V[b_0] = \sigma^2\,\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right).$$

Réponses:

(a) D'après le modèle linéaire simple, on a $E\{Y_i\} = \beta_0 + \beta_1\,x_i$, alors

$$E\{\bar{Y}\} = \sum_{i=1}^n (1/n)E\{Y_i\} = \sum_{i=1}^n (1/n)(\beta_0 + \beta_1\,x_i) = \beta_0 + \beta_1 \sum_{i=1}^n x_i/n = \beta_0 + \beta_1\,\bar{x}.$$

(b) En utilisant le résultat de la partie (a) et que $E\{b_1\} = \beta_1$, on a

$$E\{b_0\} = E\{\bar{Y}\} - \bar{x}\,E\{b_1\} = (\beta_0 + \beta_1\,\bar{x}) - \bar{x}\beta_1 = \beta_0.$$

Alors, b_0 est un estimateur non-biasé de β_0.

(c) Rappelons-nous que $s_{xY} = \sum_{i=1}^n (x_i - \bar{x})\,Y_i$. Alors,

$$\begin{aligned} b_0 &= \bar{Y} - b_1\,\bar{x} = \bar{Y} - \frac{s_{xY}}{s_{xx}}\,\bar{x} \\ &= \left(\sum_{i=1}^n (1/n)Y_i\right) - \left(\frac{\sum_{i=1}^n (x_i - \bar{x})\,Y_i}{s_{xx}}\right)\,\bar{x} \\ &= \left(\sum_{i=1}^n (1/n)Y_i\right) - \left(\sum_{i=1}^n [(x_i - \bar{x})\bar{x}/s_{xx}]\,Y_i\right) \\ &= \sum_{i=1}^n \left[\frac{1}{n} - \frac{(x_i - \bar{x})\,\bar{x}}{s_{xx}}\right] Y_i. \end{aligned}$$

(d) Selon la partie (d), b_0 est une combinaison linéaire de $Y_1, Y_2, \ldots, Y_n$ qui sont des variables aléatoires indépendantes et normales, alors b_0 est une variable aléatoire normale.

(e) De la partie (d), on a

$$b_0 = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\,\bar{x}}{s_{xx}} \right] Y_i.$$

Puisque $Y_1, \ldots, Y_n$ sont indépendantes et $V[Y_i] = \sigma^2$ pour $i = 1, 2, \ldots, n$, alors

$$V[b_0] = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\,\bar{x}}{s_{xx}} \right]^2 V[Y_i] = \sigma^2 \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\,\bar{x}}{s_{xx}} \right]^2.$$

Mais,

$$\begin{aligned} & \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\,\bar{x}}{s_{xx}} \right]^2 \\ = & \sum_{i=1}^{n} \left[\frac{1}{n^2} - 2\,\frac{(x_i - \bar{x})\,\bar{x}}{n\,s_{xx}} + \frac{(x_i - \bar{x})^2\,\bar{x}^2}{s_{xx}^2} \right] \\ = & \frac{n}{n^2} - 2\,\frac{\bar{x}}{n\,s_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) + \frac{\bar{x}^2}{s_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2. \end{aligned}$$

Mais, $s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ et $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$, alors

$$\sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\,\bar{x}}{s_{xx}} \right]^2 = \frac{n}{n^2} - 2\,\frac{\bar{x}}{n\,s_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) + \frac{\bar{x}^2}{s_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}.$$

Alors,

$$V[b_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right).$$

9. Régression linéaire multiple

9.1 Régression linéaire multiple I

Supposons que nous avons un modèle linéaire avec $p - 1 = 9$ prédicteurs et $n = 125$ observations. Nous avons ajusté le modèle et nous avons calculer la somme de carré résiduelle $\sum_{i=1}^{125} e_i^2 = 356$. La variance de l'échantillon pour la variance dépendantes est $s_y^2 = 34$.

(a) Donner une estimation de la variance de l'erreur.

(b) Donner l'écart type résiduel.

(c) Calculer le coefficient de détermination R^2.

Réponses:

(a) Une estimation de σ^2 est $\text{MSE} = \sqrt{\text{SSE}/(n-p)} = 356/(125 - 10) = 3{,}09565$.

(b) L'écart type résiduel est $s_e = \sqrt{\text{MSE}} = \sqrt{3{,}09565}$.

(c) On a $s_{yy} = (n-1)\,s_y^2 = 126\,(34) = 4\,284$ et $\text{SSR} = s_{yy} - \text{SSE} = 4\,284 - 356 = 3928$. Alors, le coefficient de détermination est

$$R^2 = \frac{\text{SSR}}{s_{yy}} = \frac{3\,928}{4\,284} = 0{,}9169.$$

9.2 Régression linéaire multiple II

Nous avons ajusté un modèle linéaire avec la fonction de la moyenne suivante

$$E\{Y\} = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_3.$$

Nous avons aussi extrait la matrice du plan X avec R.

```
> mod<-lm(y~x1+x2+x3)
> X<-model.matrix(mod)
```

Là nous calculons l'inverse de la matrice $X'X$, c'est-à-dire nous obtenons $(X'X)^{-1}$.

```
> ## inverse de (X'X)
> solve(t(X) %*% X)
            (Intercept)            x1            x2            x3
(Intercept)  1.30376082 -4.873540e-03 -1.600293e-02 -8.779750e-03
x1          -0.00487354  2.706184e-04 -1.285093e-04  3.860415e-05
x2          -0.01600293 -1.285093e-04  4.201381e-04  1.267343e-05
x3          -0.00877975  3.860415e-05  1.267343e-05  1.729348e-04
```

Nous affichons aussi l'estimation des paramètres de la fonction de la moyenne, l'écart type résiduel, et le nombre de degré de liberté de l'erreur.

```
> mod$coefficients
(Intercept)          x1          x2          x3
 17.8425757  19.9823858 -18.9075439   0.9351907
 > summary(mod)$sigma
[1] 10.0958
> mod$df.residual
[1] 36
```

N.B. Le symbole pour la mutiplication matricielle avec R est `%*%`. En outre, si `A` est une matrice inversible, alors `solve(A)` donne l'inverse de la matrice.

(a) Tester $H_0 : \beta_1 = 0$ contre $H_a : \beta_1 \neq 0$. Utiliser un niveau de signification de $\alpha = 5\%$.

(b) Donner un intervalle de confiance à 95% pour β_1.

Réponses:

L'écart type résiduel est $s_e = \sqrt{\text{MSE}} = 10{,}0958$. L'estimation de β_1 est $b_1 = 19{,}9823858$ et l'erreur type estimée de l'estimation est

$$s\{b_1\} = \sqrt{\text{MSE}(X'X)^{-1}_{11}} = s_e \sqrt{(X'X)^{-1}_{11}} = 10{,}0958\, \sqrt{2.706184 \times 10^{-4}} = 0{,}16608.$$

N.B. L'indice pour les coefficients prend les valeurs $j = 0, 1, 2, \ldots, p-1$. Alors, $(X'X)^{-1}_{11}$ est la deuxième valeur dans la diagonale principale de $X'X$.

La valeur observée de la statistique du test est

$$t_0 = \frac{b_1}{s\{b_1\}} = \frac{19{,}9823858}{0{,}16608} = 120{,}3178.$$

La valeur p est $2\, P(t(36) \geq 120,3178) < 0{,}0001$.

```
2*(1-pt(120.3178,36))
```

```
## [1] 0
```

À un niveau de signification de $\alpha = 5\%$, le predicteur x_1 est significatif.

(b) Un intervalle de confiance à 95% pour β_1 est

$$b_1 \pm t(0{,}975; 36)s\{b_1\} =]19.65; 20{,}32[.$$

où $t(0{,}975; 36) = 2{,}02809$, $b_1 = 19{,}9823858$ et $s\{b_1\} = 0{,}16608$.

```
qt(0.975,36)
```

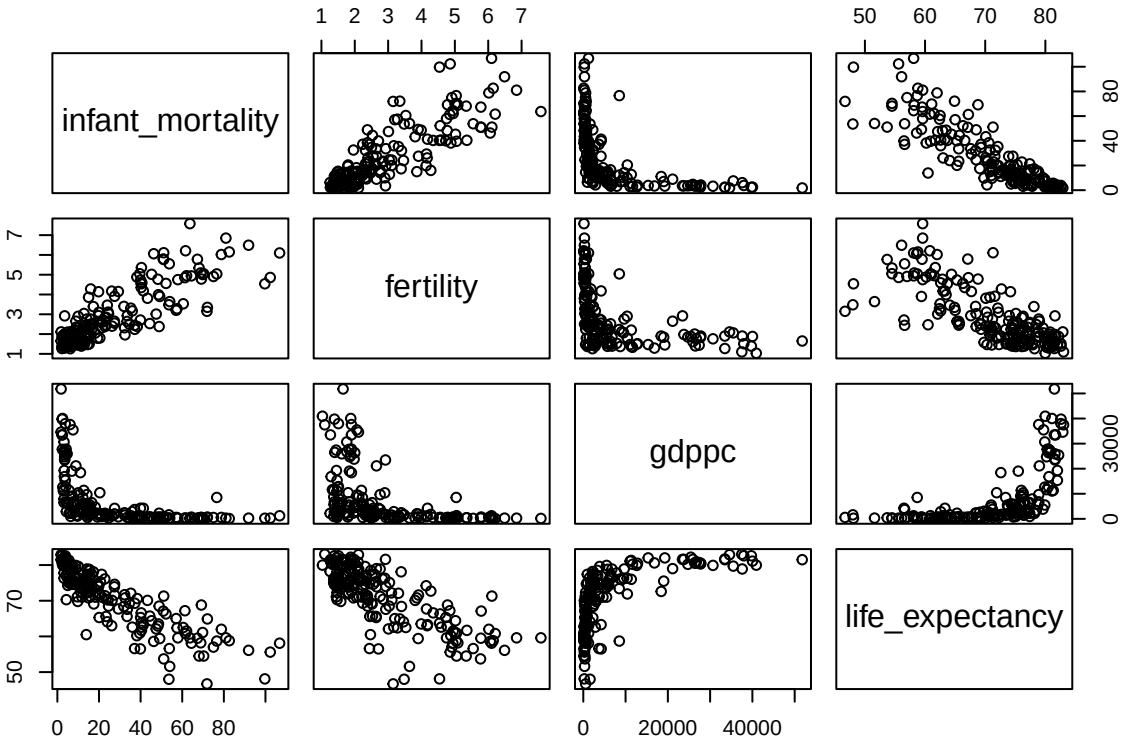
```
## [1] 2.028094
```

9.3 Régression linéaire multiple

```
library(dplyr)
gapminder.rlm <- gapminder |>
  filter(year==2011) |>
  select(infant_mortality, fertility, gdppc,life_expectancy)
str(gapminder.rlm)
```

```
## 'data.frame':    185 obs. of  4 variables:
##  $ infant_mortality: num  14.3 22.8 106.8 7.2 12.7 ...
##  $ fertility       : num  1.75 2.83 6.1 2.12 2.2 1.5 1.69 1.88 1.44 1.96 ...
##  $ gdppc           : num  2190 2210 1231 9096 11353 ...
##  $ life_expectancy : num  77.4 76.1 58.1 75.9 76 ...
```

```
plot(gapminder.rlm)
```




```
head(gapminder.rlm)
```

```
##   infant_mortality fertility     gdppc life_expectancy
## 1             14.3      1.75  2190.460            77.4
## 2             22.8      2.83  2209.961            76.1
## 3            106.8      6.10  1231.135            58.1
## 4              7.2      2.12  9095.516            75.9
```

```
## 5             12.7      2.20 11353.457            76.0
## 6             15.3      1.50  1445.759            73.5
```

```
summary(gapminder.rlm)
```

```
##  infant_mortality    fertility         gdppc          life_expectancy
##  Min.   :  1.800   Min.   :1.030   Min.   :  109.3   Min.   :46.70
##  1st Qu.:  7.275   1st Qu.:1.790   1st Qu.:  662.9   1st Qu.:65.30
##  Median : 16.250   Median :2.350   Median : 2329.2   Median :73.70
##  Mean   : 26.699   Mean   :2.854   Mean   : 7486.3   Mean   :71.18
##  3rd Qu.: 40.375   3rd Qu.:3.640   3rd Qu.: 8511.8   3rd Qu.:77.40
##  Max.   :106.800   Max.   :7.580   Max.   :51787.6   Max.   :83.02
##  NA's   :7                         NA's   :17
```

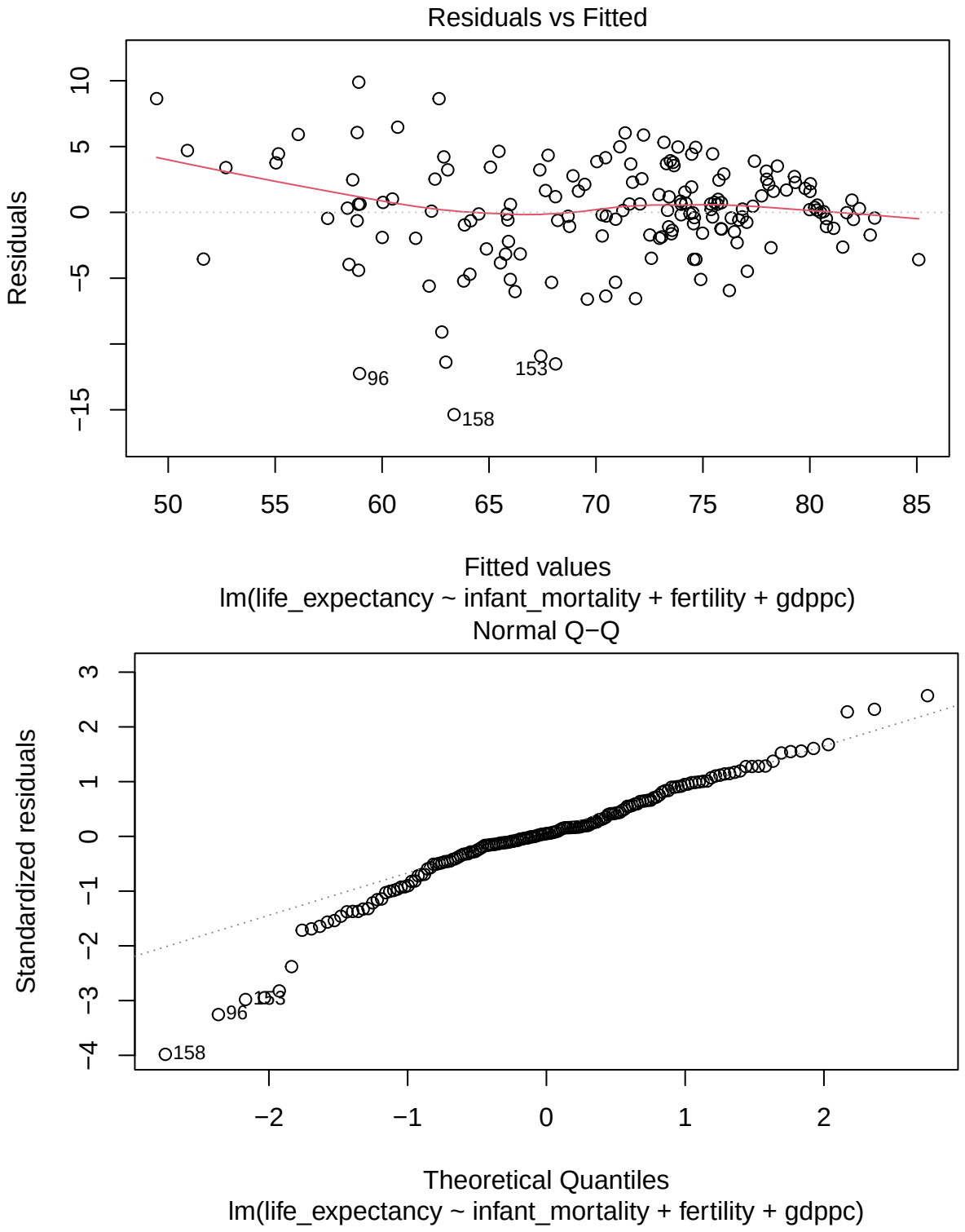
```
attributes(summary(gapminder.rlm))
```

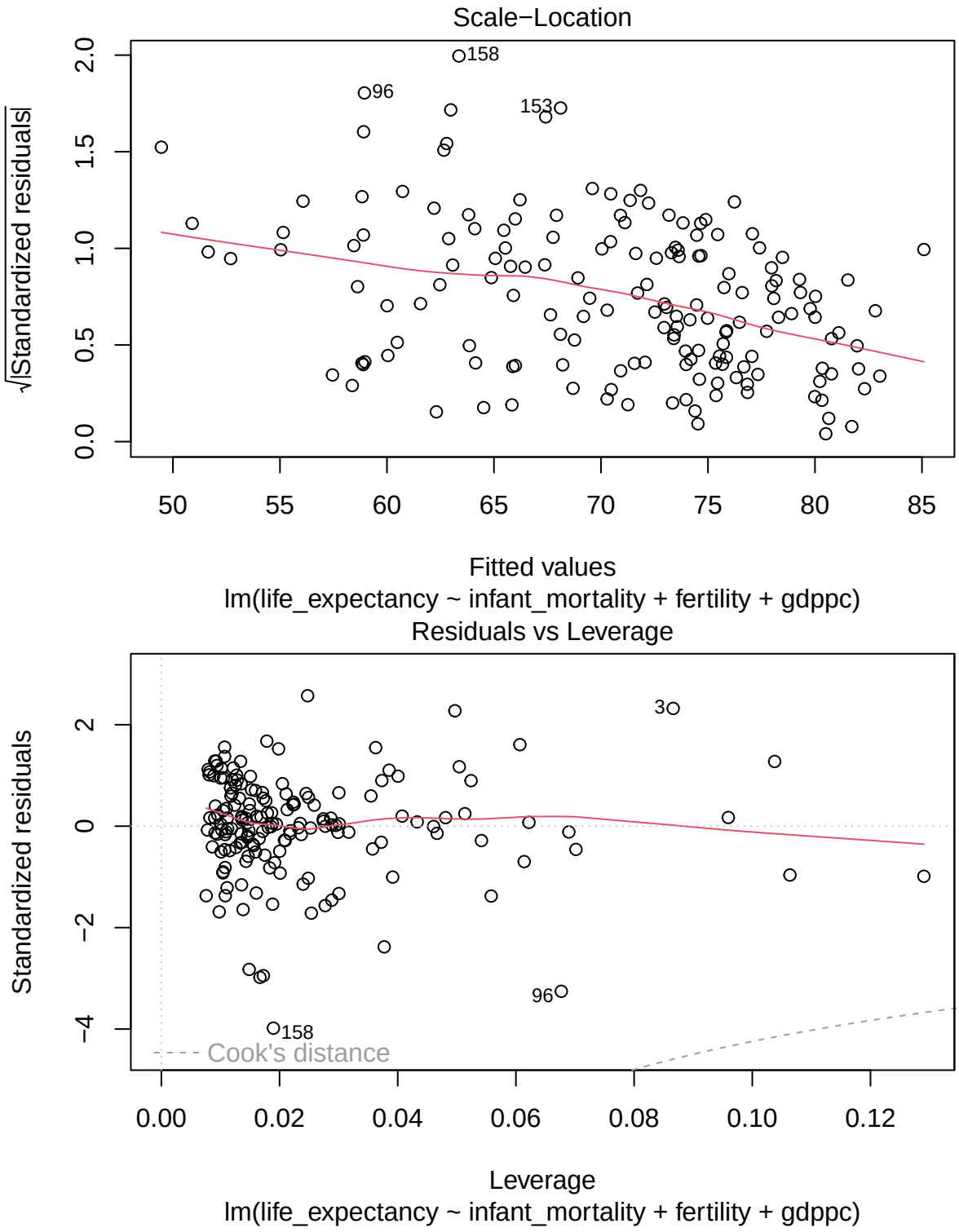
```
## $dim
## [1] 7 4
##
## $dimnames
## $dimnames[[1]]
## [1] "" "" "" "" "" "" ""
##
## $dimnames[[2]]
## [1] "infant_mortality" "  fertility"      "    gdppc"        "life_expectancy"
##
##
## $class
## [1] "table"
```

```
mod.rlm.1 <- lm(life_expectancy ~ infant_mortality + fertility + gdppc, data=gapminder.rlm)
summary(mod.rlm.1)
```

```
##
## Call:
## lm(formula = life_expectancy ~ infant_mortality + fertility +
##     gdppc, data = gapminder.rlm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.365  -1.615   0.192   2.409   9.890
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      77.3967309  0.8739902  88.556  < 2e-16 ***
## infant_mortality -0.2393955  0.0243058  -9.849  < 2e-16 ***
## fertility        -0.4231811  0.3857345  -1.097    0.274
## gdppc             0.0001704  0.0000341   4.997  1.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.896 on 162 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.7931, Adjusted R-squared:  0.7892
## F-statistic:   207 on 3 and 162 DF,  p-value: < 2.2e-16
```

```
plot(mod.rlm.1)
```






```
mod.1 <- lm(infant_mortality ~ fertility + gdppc, data=gapminder.rlm)
mod.2 <- lm(fertility ~ infant_mortality + gdppc, data=gapminder.rlm)
mod.3 <- lm(gdppc ~ infant_mortality + fertility, data=gapminder.rlm)

summary(mod.1)$r.squared
```

```
## [1] 0.7456988
```

```
summary(mod.2)$r.squared
```

```
## [1] 0.7199269
```

```
summary(mod.3)$r.squared
```

```
## [1] 0.2772157
```

```
intercept_only <- lm(life_expectancy ~ 1, data=gapminder.rlm[complete.cases(gapminder.rlm),])

#define model with all predictors
all <- lm(life_expectancy ~ infant_mortality + fertility + gdppc, data=gapminder.rlm[complete.cases(gapr

#perform forward stepwise regression
forward <- step(intercept_only, direction='forward', scope=formula(all))
```

```
## Start:  AIC=710.95
## life_expectancy ~ 1
##
##                     Df Sum of Sq     RSS    AIC
## + infant_mortality  1    9024.4  2857.8 476.41
## + fertility         1    6884.8  4997.3 569.18
## + gdppc             1    4435.6  7446.5 635.38
## <none>                          11882.2 710.95
##
## Step:  AIC=476.41
## life_expectancy ~ infant_mortality
##
##             Df Sum of Sq    RSS    AIC
## + gdppc      1    380.80 2477.0 454.67
## <none>                   2857.8 476.41
## + fertility  1     20.11 2837.7 477.23
##
## Step:  AIC=454.67
## life_expectancy ~ infant_mortality + gdppc
##
##             Df Sum of Sq    RSS    AIC
## <none>                   2477.0 454.67
## + fertility  1    18.267 2458.7 455.44
```

```
#view results of forward stepwise regression
forward$anova
```

```
##                  Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                    NA        NA       165  11882.178 710.9540
## 2 + infant_mortality -1 9024.3761       164   2857.802 476.4062
## 3            + gdppc -1  380.8033       163   2476.999 454.6673
```

```
#view final model
forward$coefficients
```

```
##      (Intercept) infant_mortality            gdppc
##    76.7403353049    -0.2608641520     0.0001708017
```

9.4 Régression linéaire multiple

Considérons les données provenant d'une étude observationelle pour décrire le temps de rétablissement (en mois) selon la forme pré-chirugie du patient. La forme du patient est un prédicteur catégorique avec 3 niveaux: 1=inférieure à la moyenne; 2=moyenne; 3=supérieure à la moyenne.

On importe les données et on affiche la structure du jeu de données.

```
#genou <- read.csv("Data/genou.csv")
#str(genou)
#plot(genou)
```

On remarque que la variable `Groupe` (la forme) est une variable numérique; en réalité, c'est une variable catégorique. Nous allons la transformer en un facteur (un type de variable catégorique dans `R`).

```
#genou$Groupe<-factor(genou$Groupe)
#str(genou)
```

Avec un facteur, on peut afficher les niveaux, un tableau de fréquence, et le codage du facteur (pour la modèlisation).

On affiche les niveaux; on remarque qu'il y en a trois.

```
#levels(genou$Groupe)
```

Voici le tableau de fréquence pour la variable `Groupe`. On voit que ce n'est pas une étude équilibrée; le nombre d'observations n'est pas constant dans chaque groupe.

```
#table(genou$Groupe)
```

Il y a deux **variables muettes**, une pour le groupe 2 et l'autre pour le groupe 3. Dans ce qui suit, chaque colonne est une variable muette. La variable muette 2 prend la valeur 1 seulement si l'observation est dans le groupe 2, sinon c'est 0; la variable muette 3 prend la valeur 1 seulement si l'observation est dans le groupe 3, sinon c'est 0.

```
#contrasts(genou$Groupe)
```

On ajuste le modèle d'ANOVA et on affiche un sommaire de l'ajustement.

```
#mod<-lm(Temps ~ Groupe, data=genou)
#summary(mod)
```

Le modèle d'ANOVA est significatif ($F(2, 21) = 16.96$; $p < 0,0001$). On appelle parfois ce test une analyse de variance (ANOVA). C'est un test pour l'égalité des moyennes.

Ainsi, on peut conclure qu'il y a des preuves significatives que le temps moyen de rétablissement varie selon le groupe du patient. Puisqu'il n'y a qu'un prédicteur dans le modèle,on peut aussi afficher le tableau de l'ANOVA pour le test de la signification de la régression.

```
#anova(mod)
```

Ainsi, on estime que le temps de rétablissement pour un patient avec un forme **inférieure** à la moyenne est 38 mois; le temps moyen de rétablissement pour un patient avec une forme **moyenne** est $38 - 6 = 32$ mois, et celui pour un patient avec une forme supérieure à la moyenne est $38 - 14 = 24$ mois. De plus, l'écart type résiduel est 4.451 mois.

Dans l'exemple de la chirurgie des genoux, nous avons un **facteur observationnel**. Les chercheurs n'ont pas assigné la forme pré-chirurgie au patient; ils viennent avec une certaine forme. Il est possible que les différences observées peuvent être expliquer par une variable de confusion.

Par exemple, les chercheurs peuvent être malchanceux il se pourrait que le groupe 1 soit peuplé de patients plus agés, et c'est peut-être plutôt ceci qui a été observé. Pour des études observationnelles, il est important

d'utiliser les connaissances dans le domaine d'application et d'essayer de proposer quelques variables de confusion.

Dans les applications médicales, l'âge et le sexe sont souvent utilisées comme variables de confusion. Ici, on contrôle pour l'âge du participant. Le sexe est souvent une variable explicative importante en médécine; la pratique commune est de séparer les sexes. Les études utilisent souvent seulement des hommes ou seulement des femmes: dans cette étude, il n'y a que des hommes.

Nous allons décrire le temps de rétablissement selon la forme du patient et son âge. Le modèle linéaire général possède un prédicteur catégorique et un prédicteur quantitatif. La fonction systématique du temps de réponse moyenne est

$$E\{Y\} = \beta_0 + \beta_1 I\{\texttt{Groupe} = 2\} + \beta_2 I\{\texttt{Groupe} = 3\} + \beta_3 \times \texttt{Age}.$$

β_0 est le temps de réadaptation moyen d'un patient avec un forme inférieure, d'âge 0. Mais cela ne veut rien dire puisqu'on ne peut pas faire affaire à un patient d'âge zéro.

On donne un sens à l'ordonnée à l'origine, nous allons centrer le prédicteur quantitatif autour de la moyenne $x = 23.575$. On pourrait aussi utiliser une valeur proche de la moyenne, comme 24.

```
#mean(genou$Age)
```

Le modèle devient:

$$E\{Y\} = \beta_0 + \beta_1 I\{\texttt{Groupe} = 2\} + \beta_2 I\{\texttt{Groupe} = 3\} + \beta_3 \times (\texttt{Age} - 24).$$

Interprétation des paramètres:

- β_0 est le temps moyen de rétablissement d'un patient de 24 ans avec une forme inférieure à la moyenne;
- quelque soit la forme du patient, le taux de variation de $E\{Y\}$ par rapport à l'âge est β_3;
- pour deux patients du même âge, le temps de rétablissement moyen entre un patient de forme moyenne et de forme inférieure à la moyenne est β_1 (effet du groupe 2);
- pour deux patients du même âge, le temps de rétablissement moyen entre un patient de forme supérieure à la moyenne et de forme inférieure à la moyenne est β_2 (effet du groupe 3).

```
#genou$Age.c <- genou$Age-24
```

On ajuste un modèle linéaire général pour décrire le temps de rétablissement moyen selon la forme pré-chirurgique et l'âge du patient. Le modèle est significatif ($F(3, 20) = 1170$; $p < 0.0001$) et $R^2 = 0.9943$.

```
#mod.1 <- lm(Temps ~ Groupe + Age.c, data=genou)
#summary(mod.1)
```

Le modèle est significatif, alors on peut conclure qu'il y ait au moins un prédicteur utile pour décrire la distribution du temps de rétablissement. Est-ce que la forme pré-chirurgique est significative? Est-ce que l'âge est significatif?

On cherche à tester

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{envers} \quad H_1 : \beta_1 \neq= 0 \text{ or } \beta_2 \neq 0.$$

Autrement dit, on veut savoir si nous pouvons éliminer le prédicteur `Groupe` du modèle. On peut utiliser un test linéaire général en comparant le modèle complet au modèle réduit. La forme pré-chirurgique est un prédicteur significatif ($F(2, 20) = 300.11$; $p < 0.0001$).

```
#mod.0 <- lm(Temps ~ Age.c,data=genou)
#anova(mod.0,mod.1)
```

La régression est significative: on rejette H_0 en faveur de H_1.

Pour tester

$$H_0 : \beta_3 = 0 \quad \text{envers} \quad H_1 : \beta_3 \neq 0;$$

on invoque une hypothèse qui ne contient qu'un seul paramètre, on peut utiliser un test t ou un test F. Les deux tests sont équivalents, avec $t^2 = F$.

Voici le tableau des tests t liés aux coefficients. L'âge est un prédicteur significatif ($t(20) = 36.4608$; $p < 0.0001$).

```
#summary(mod.1)$coefficients
#mod.1$df.residual
```

Si on utilise plutôt le test linéaire général pour la signification de l'âge, on obtient que c'est un prédicteur significatif ($F(1, 20) = 1329.4$; $p < 0.0001$).

```
#mod.1 <- lm(Temps ~ Groupe + Age.c,data=genou)
#mod.2 <- lm(Temps ~ Groupe, data=genou)
#anova(mod.1,mod.2)
```

Voici l'estimation des paramètres du modèle:

```
#mod <- lm(Temps ~ Groupe + Age.c,data=genou)
#mod$coefficients
#summary(mod)$sigma
```

On estime qu'un patient de 24 ans avec une forme inférieure à la moyenne aura un temps moyen de rétablissement de 35.4 mois. On estime que le temps moyen de rétablissement augmentera de 1.16 mois par année ajoutée à l'âge du patient. Si un patient a une forme moyenne au lieu d'une forme inférieure à la moyenne, alors le temps moyen de rétablissement est réduit de 1.85 mois par année ajoutée à l'âge du patient.

La réduction est 8.72 mois par année ajoutée à l'âge du patient pour un patient avec une forme supérieure à la moyenne en comparaison à un patient avec une forme inférieure à la moyenne. L'écart type résiduel est 0.56 mois.

10. Formes quadratiques

10.1 Formes quadratiques I

Pour chacun des cas ci-dessous, la matrice A est la matrice de forme quadratique Q des variables aléatoires non corrélées $Y_1,Y_2,Y_3. $ En outre, supposons que $E\{Y_i\} = 0$ and $V[Y_i] = \sigma^2 = 3$, pour $i = 1, 2, 3$. Pour chacune des formes quadratiques, calculer $E\{Q\}$.

$$(i)\ \ A = \begin{bmatrix} 1 & 4 & 6 \\ 4 & 0 & 6 \\ 6 & 6 & 5 \end{bmatrix}; \quad (ii)\ \ A = \begin{bmatrix} 1 & 4 & 6 \\ 3 & 0 & 6 \\ 6 & 4 & 10 \end{bmatrix}; \quad (iii)\ \ A = \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix}.$$

Réponses:

Nous utilisons le résultat suivant:

$$E\{Q\} = E\{Y\, A\, Y'\} = \sigma^2 \operatorname{tr}(A) + E\{Y\}\, A\, E\{Y'\}.$$

Mais, $E\{Y\} = 0$, alors, $E\{Q\} = \sigma^2 \operatorname{tr}(A) = 3 \operatorname{tr}(A)$.

(i) $E\{Q\} = 3 \operatorname{tr}(A) = 3\,(1 + 0 + 5) = 18$

(ii) $E\{Q\} = 3 \operatorname{tr}(A) = 3\,(1 + 0 + 10) = 33$

(iii) $E\{Q\} = 3 \operatorname{tr}(A) = 3\,(2/3 + 2/3 + 2/3) = 6$

10.2 Formes quadratiques II

Voici des formes quadratiques de Y_1, Y_2, Y_3. Dans chaque cas, donnez la matrice de la forme quadratique, et calculez la trace de la matrice.

(a) $Q = Y_1^2 - 5\,Y_2^2 + 10\,Y_3^2 - 2\,Y_1\,Y_2 - 10\,Y_1\,Y_3 + 6\,Y_2\,Y_3$.

(b) $Q = 5\,Y_1^2 + 3\,Y_2^2 + 2\,Y_3^2 - 10\,Y_1\,Y_2 - 8\,Y_1\,Y_3 + 5\,Y_2\,Y_3$.

Réponses:

(a) La matrice de la forme quadratique est

$$A = \begin{bmatrix} 1 & -1 & -5 \\ -1 & -5 & 3 \\ -5 & 3 & 10 \end{bmatrix}.$$

Sa trace est $\mathrm{tr}(A) = 1 + (-5) + 10 = 6$.

(b) La matrice de la forme quadratique est

$$A = \begin{bmatrix} 5 & -5 & -4 \\ -5 & 3 & 2.5 \\ -4 & 2.5 & 2 \end{bmatrix}.$$

Sa trace est $\mathrm{tr}(A) = 5 + 3 + 2 = 10$.

11. Bonferroni

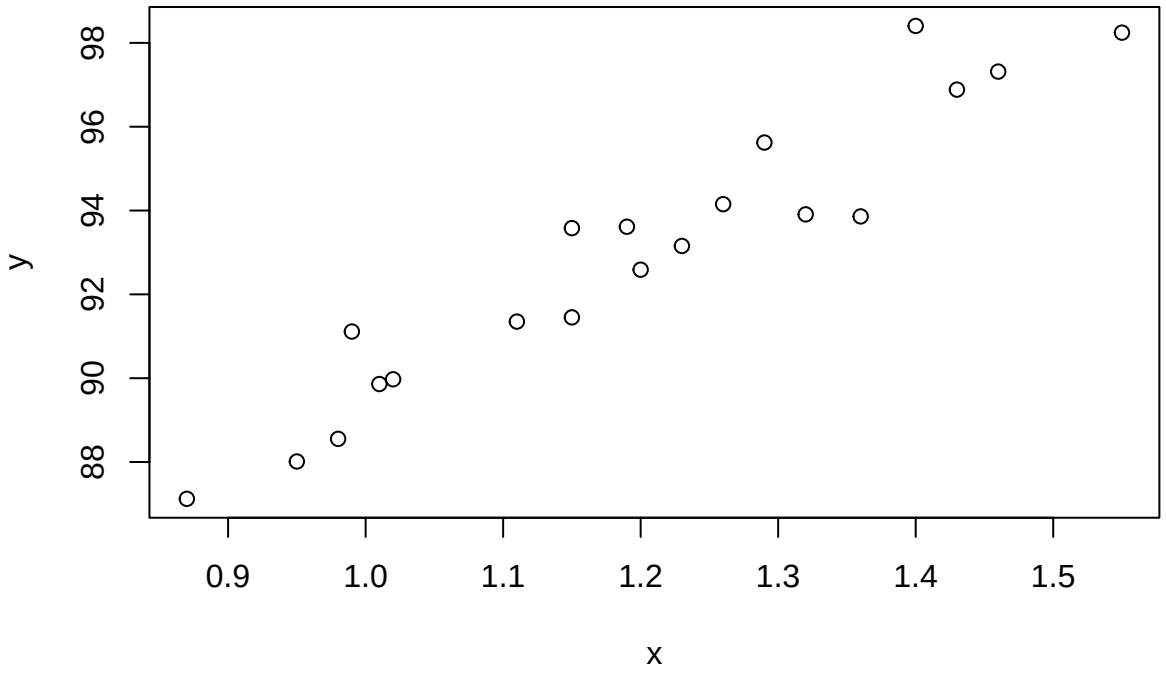
Imaginons que la relation linéaire réelle liant Y et X soit $y = 75 + 15x + \varepsilon$, où $\varepsilon \sim N(0, 1)$.

Nous prélevons des échantillons (de taille n) de la réponse pour les prédicteurs suivants:

```
n=20
x = c(0.99,1.02,1.15,1.29,1.46,1.36,0.87,1.23,1.55,1.40,1.19,1.15,0.98,1.01,1.11,1.20,1.26,1.32,1.43,0.9
somme.X = sum(x)
somme.X.2 = sum(x*x)
```

Pour le premier échantillon, les réponses observées sont:

```
set.seed(0)
beta0 = 75
beta1 = 15
y = beta0 + beta1*x + rnorm(n)
plot(x,y)
```

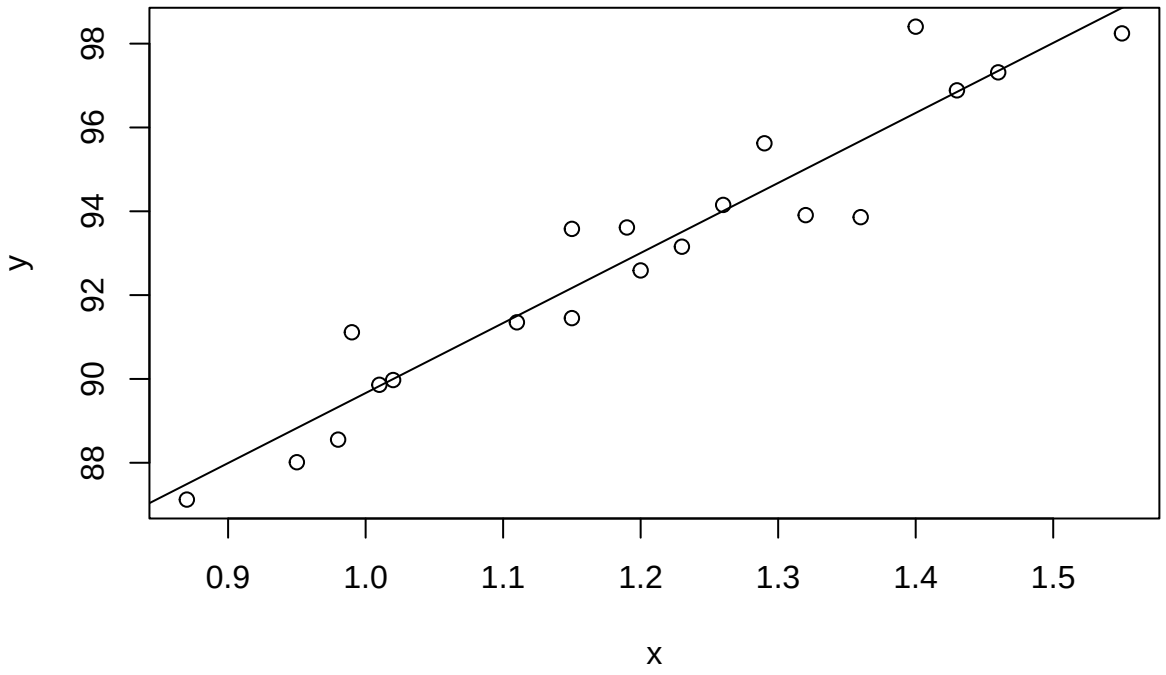


L'équation de la droite de meilleur ajustement, dans ce cas, est:

```
(mod = lm(y~x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##       72.95        16.71
```

```
plot(x,y)
abline(mod)
```



On peut calculer

```
somme.Y = sum(y)
somme.X.Y = sum(x*y)
somme.Y.2 = sum(y*y)
b1 = (somme.X.Y-n*mean(x)*mean(y))/(somme.X.2-n*(mean(x))^2)
b0 = mean(y) - b1*mean(x)
SSE = somme.Y.2 -n*(mean(y))^2 - b1^2*(somme.X.2-n*(mean(x))^2)
sigma.2.hat = SSE/(n-2)
s.b1 = sqrt(sigma.2.hat/(somme.X.2-n*(mean(x))^2))
s.b0 = sqrt(sigma.2.hat*(1/n+(mean(x))^2/(somme.X.2-n*(mean(x))^2)))
```

À un niveau de confiance de 95%, l'intervalle de confiance pour l'ordonnée à l'origine β_0 est ainsi:

```
alpha=0.05
c(b0-qt(1-alpha/2,n-2)*s.b0,b0+qt(1-alpha/2,n-2)*s.b0)
```

```
## [1] 69.88696 76.02121
```

La valeur réelle de β_0 se retrouve bien dans l'I.C.:

```
(beta0 > b0-qt(1-alpha/2,n-2)*s.b0) & (beta0 < b0+qt(1-alpha/2,n-2)*s.b0)
```

```
## [1] TRUE
```

Celui de la pente β_1 est:

```
c(b1-qt(1-alpha/2,n-2)*s.b1,b1+qt(1-alpha/2,n-2)*s.b1)
```

```
## [1] 14.17464 19.24365
```

La valeur réelle de β_0 se retrouve bien dans l'I.C.:

```
(beta1 > b1-qt(1-alpha/2,n-2)*s.b1) & (beta1 < b1+qt(1-alpha/2,n-2)*s.b1)
```

```
## [1] TRUE
```

Simultanément, (β_0, β_1) se retrouvent dans leurs I.C. respectifs:

```
(beta0 > b0-qt(1-alpha/2,n-2)*s.b0) & (beta0 < b0+qt(1-alpha/2,n-2)*s.b0) & (beta1 > b1-qt(1-alpha/2,n-
```

```
## [1] TRUE
```

Répétons l'expérience à $m = 10,000$ reprises:

```
m = 10000
g=1
set.seed(0)
ICb0 = c()
ICb1 = c()
ICb0b1 = c()
for(j in 1:m){
  y = beta0 + beta1*x + rnorm(n, sd=10)
  somme.Y = sum(y)
  somme.X.Y = sum(x*y)
  somme.Y.2 = sum(y*y)
  b1 = (somme.X.Y-n*mean(x)*mean(y))/(somme.X.2-n*(mean(x))^2)
  b0 = mean(y) - b1*mean(x)
  SSE = somme.Y.2 -n*(mean(y))^2 - b1^2*(somme.X.2-n*(mean(x))^2)
  sigma.2.hat = SSE/(n-2)
  s.b1 = sqrt(sigma.2.hat/(somme.X.2-n*(mean(x))^2))
  s.b0 = sqrt(sigma.2.hat*(1/n+(mean(x))^2/(somme.X.2-n*(mean(x))^2)))

  ICb0[j] = (beta0 > b0-qt(1-(alpha/g)/2,n-2)*s.b0) & (beta0 < b0+qt(1-(alpha/g)/2,n-2)*s.b0)
  ICb1[j] = (beta1 > b1-qt(1-(alpha/g)/2,n-2)*s.b1) & (beta1 < b1+qt(1-(alpha/g)/2,n-2)*s.b1)
  ICb0b1[j] = (beta0 > b0-qt(1-(alpha/g)/2,n-2)*s.b0) & (beta0 < b0+qt(1-(alpha/g)/2,n-2)*s.b0) & (beta
}
```

Individuellement, nous avons:

```
sum(ICb0)/m
```

```
## [1] 0.9523
```

```
sum(ICb1)/m
```

```
## [1] 0.9518
```

Simultanément:

```
sum(ICb0b1)/m
```

```
## [1] 0.9459
```

Nous n'atteignons pas le cap des 95%!!

Si l'on utilise la procédure de Bonferroni, au contraire:

```
m = 10000
g=2
set.seed(0)
ICb0 = c()
ICb1 = c()
ICb0b1 = c()
for(j in 1:m){
  y = beta0 + beta1*x + rnorm(n, mean=0, sd = 400)
  somme.Y = sum(y)
  somme.X.Y = sum(x*y)
  somme.Y.2 = sum(y*y)
```

```
  b1 = (somme.X.Y-n*mean(x)*mean(y))/(somme.X.2-n*(mean(x))^2)
  b0 = mean(y) - b1*mean(x)
  SSE = somme.Y.2 -n*(mean(y))^2 - b1^2*(somme.X.2-n*(mean(x))^2)
  sigma.2.hat = SSE/(n-2)
  s.b1 = sqrt(sigma.2.hat/(somme.X.2-n*(mean(x))^2))
  s.b0 = sqrt(sigma.2.hat*(1/n+(mean(x))^2/(somme.X.2-n*(mean(x))^2)))

  ICb0[j] = (beta0 > b0-qt(1-(alpha/g)/2,n-2)*s.b0) & (beta0 < b0+qt(1-(alpha/g)/2,n-2)*s.b0)
  ICb1[j] = (beta1 > b1-qt(1-(alpha/g)/2,n-2)*s.b1) & (beta1 < b1+qt(1-(alpha/g)/2,n-2)*s.b1)
  ICb0b1[j] = (beta0 > b0-qt(1-(alpha/g)/2,n-2)*s.b0) & (beta0 < b0+qt(1-(alpha/g)/2,n-2)*s.b0) & (beta
}
sum(ICb0)/m
```

```
## [1] 0.9759
```

```
sum(ICb1)/m
```

```
## [1] 0.9755
```

```
sum(ICb0b1)/m
```

```
## [1] 0.9727
```