MAT3375

Patrick Boily, Gilles Lamothe

2023-05-03

Examples

Many examples will use the following data set.

```
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. Fitting a linear model with `lm()`

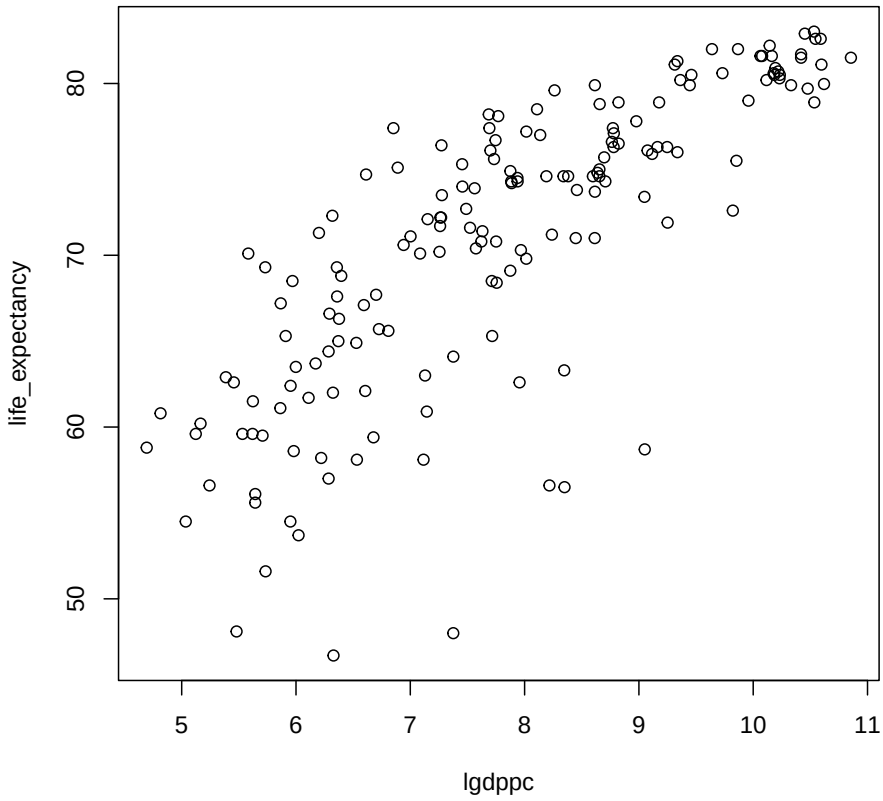
The focus for now is on observations from 2011 , in particular relating to life expectancy and log gross domestic product per capita.

```
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)
```



The function `lm()` can be used to fit the linear model that describes life expectancy (response variable Y) as a function of the logarithm of gdp per capita (predictor variable X) in 2011.

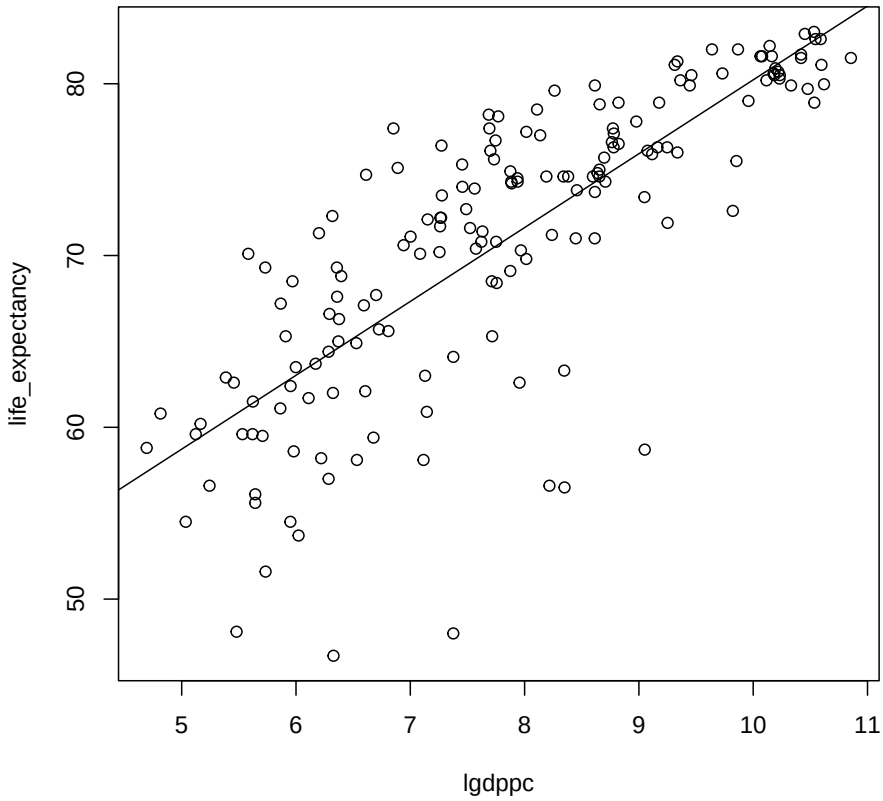
```
mod <- lm(life_expectancy ~ lgdppc, data=gapminder.elr)
mod
```

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

The estimated linear model is as follows:

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

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



Comments:

- We have created an object of type `lm` named `mod`. This object contains several components (attributes), that we can display using the `names()` function.

  ```
  names(mod)
  ```

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

  The first attribute is the vector of coefficient estimates $\vec{\beta}$.

```
mod$coefficients
```

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

Thus, $b_0 = 37.229550$ and $b_1 = 4.299686$.

The second attribute is the vector of residuals, the eighth is the number of degrees of freedom of the residuals. They can be used to calculate the estimate of the variance of the error, namely:

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

The standard deviation of this variance is the standard deviation of the residuals

$$textse = \sqrt{\text{MSE}},$$

which describes the standard deviation of the line of best fit.

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

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

```
sqrt(MSE)
```

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

- Several functions can be used with an object of 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
```

As an example, we know that the fitted line is the one that maximizes the log likelihood function

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

There are 3 parameters to estimate: β_0, β_1, and σ^2.

```
logLik(mod)
```

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

Here is how we could verify that `R` uses the previous formula to calculate the maximum of the log likelihood function:

```
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
```

- We can also obtain confidence intervals for the parameter estimates:

```
confint(mod)
```

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

The default confidence level is 95%, but it can be changed *via* the `level` argument:

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

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

We can also specify which parameters interest us:

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

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

- It is also possible to get a summary of the fit with the `summary()` function:

```
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
```

It is not necessary to display the entire summary, which is itself an object with attributes:

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

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

As an example, here is how we would extract the parameter estimates and the corresponding significance tests:

```
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
```

We can display the coefficient of determination R^2 as follows:

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

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

or the standard deviation of residuals $\sqrt{\text{MSE}}$:

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

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

2. Analysis of variance with `lm()`.

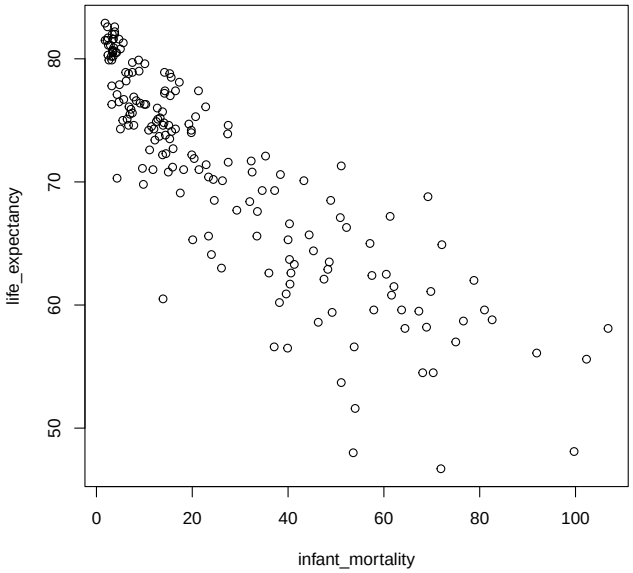
We will use a simple linear model to explain the life expectancy (Y) as a function of the infant mortality rate (X) in 2011 (we have $n = 178$).

```
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)
```



Visually, it is not unreasonable to expect the average response function to be given by

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

How can we check the significance of the infant mortality rate? We simply test

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

Since we have a simple model (a single predictor), it is also a test of the significance of the regression: we could instead use the statistic

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

But there is another approach, the analysis of variance (ANOVA). The ANOVA table is obtained by calling the `anova()` function on an object of type `lm`.

```
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
```

Comments:

- The test statistic value is $F^* = 528.73$. This means that the estimate of the variance of the error based on the sum of squares of the regression is 528.73 times as large as the estimate of the variance of the error based on the sum of the residual squares. It is quite likely that MSR is not a good estimate σ^2, but is rather a much larger quantity. Since we know that

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

  this strongly suggests that $\beta_1 \neq 0$.

- To get a measure of the significance of the evidence against $H_0 : \beta_1 = 0$, we need to calculate the odds of having observed a F^* statistic as high as 528.73 assuming that H_0 was valid: if so, $F^* \sim F(1, n-2)$. Since the P value is

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

  the evidence is strong in favour of H_1.

- This can also be done using an F-test that compares two models. In this case, we test

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

  To evaluate the evidence against H_0 and in favor of H_1, it is sufficient to compare the fit of the two models according to the sum of squares of the residuals, which is again done with the `anova()` function.

```
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
```

  We compare the sum of squares of the full model SSE = 3094.8 to the sum of squares of the reduced model SSE(R) = 12392.2, by calculating the sum of additional squares

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

  The larger this difference, the more poorly the reduced model is considered to fit compared to the full model. If the evidence is strong (in this case, a reduction of nearly 75% from the reduced model to

the full model), it suggests that one should reject H_0 in favor of the hypothesis H_1 that the slope is non-zero.

3. Binary explanatory variables

3.1 Binary explanatory variables I

A study is being conducted on the development of ectomycorrhizae, a symbiotic relationship between tree roots and a fungus in which minerals are transferred from the fungus to the trees and in return sugar goes from the trees to the fungus. 20 northern red oaks exposed to the fungus *pisolithus tinctorus* were grown in a greenhouse; all oaks were planted in the same type of soil and received the same amount of sun and water. Half of the specimens (selected at random) were treated with 368 ppm nitrogen in the form of NaNO3; the others were not (X). The mass of the stem, in grams, is measured after 140 days (Y).

The details are as follows:

```
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 uses the first category it encounters as the reference category: the oaks that did not receive nitrogen thus form the reference group (it is possible to change the order of the categories so that the nitrogen treatment group becomes the reference group, using: `nitrogen$Nitrogen <- factor(nitrogen$Nitrogen, levels=c("yes", "no"))`, for example.) In general, it is often the control group that is used as the reference group, which is already the case here.

Here are some descriptive statistics for stem mass in each of the groups:

```
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

If we assume that the variances of the two populations are equal, then we can approximate it by the weighted variance

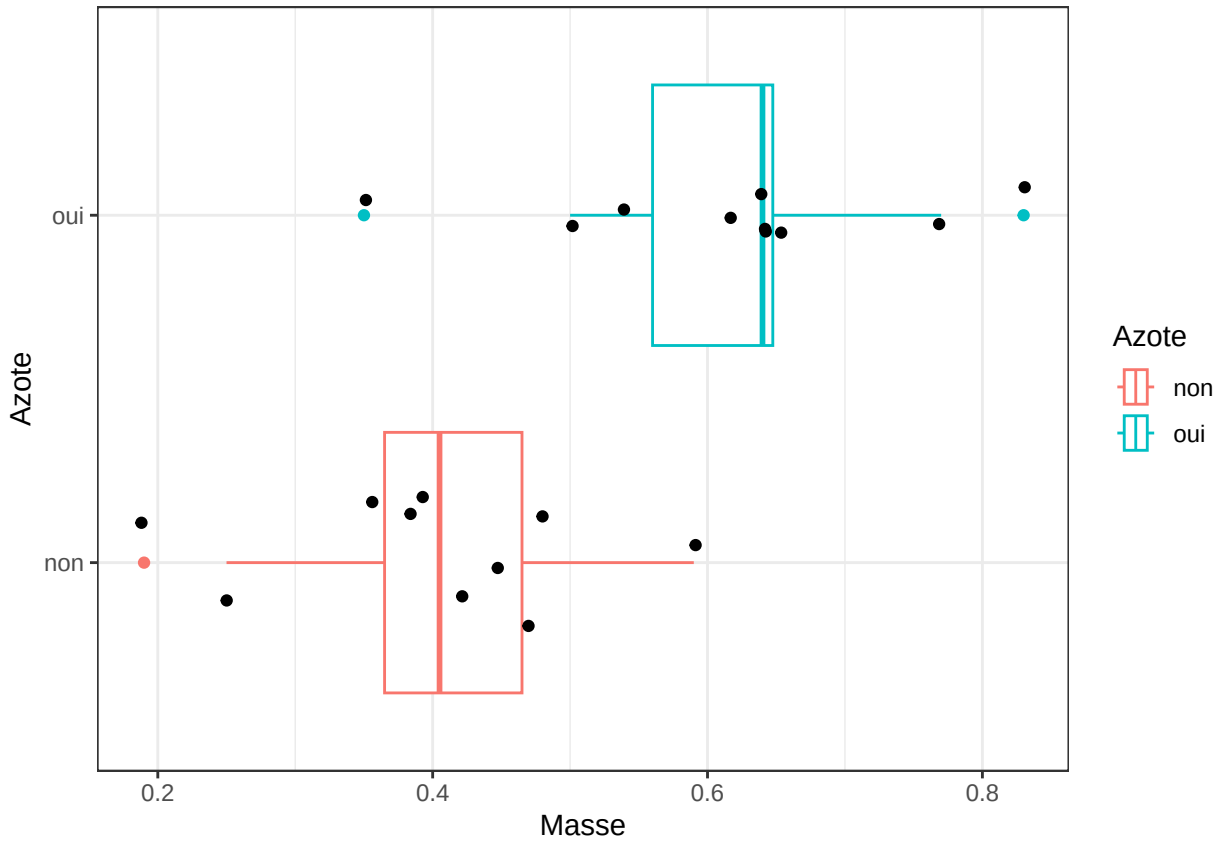
$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = 0.0156289.$$

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

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

To compare the two groups visually, we can use comparative box plots with an overlay of points, say.

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



A Student's T test can be used to compare the two means. The observed value of the test statistic is

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

The corresponding $p-$value is

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

```
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
```

The treatment certainly seems to have an effect. We can also use a regression approach to perform this test (this will help us generalize the test to the comparison of more than 2 groups).

To identify the groups, we use the dummy variable

$$x_i = \begin{cases} 1 & \text{observation } i \text{ is in the treatment group} \\ 0 & \text{otherwise} \end{cases}$$

Consider the simple linear regression model: $Y_1, Y_2, ..., Y_n$ are independent normal random variables such that

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

and $V[Y_i] = \sigma^2$ for $i = 1, \ldots, n$. In our example, we have two independent normal populations with equal variances.

When the explanatory variable is categorical (a `factor` in `R` terminology), `R` automatically encodes it as a dummy variable. We can display these dummy variables, using the `contrasts()` function.

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

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

We interpret this variable as follows: it takes the value 0 if it is an observation without nitrogen; 1 if it is an observation with nitrogen.

We can now fit a linear model to describe the mass as a function of the treatment group.

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

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

The average mass is thus:

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

The estimation of the variance σ^2 is thus MSE $= 0.01562889$.

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

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

These are the obtained values of $\overline{y}_1$, $\overline{y}_2$, and s_p^2, respectively.

Note that $\mu_1 = \mu_2$ if and only if $\beta_1 = 0$; when we test for the significance of the explanatory variable that identifies the treatment, we can also interpret it as a test of equality of means.

```
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
```

The observed value of the test statistic is

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

and the $p-$value of the test is

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

3.2 Binary explanatory variables II

We have two explanatory variables: `x1` (categorical) and `x2` (quantitative). We display the coding of the dummy variables for `x1`.

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

Here is a summary of the model fit.

```
> 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) What is the size n in this study?

(b) Give the function of the estimated mean for each of the 3 levels of the categorical variable `x1`.

(c) Test for significance of the regression. Formulate the hypotheses, give the test statistic and the corresponding $p-$value. What is the conclusion when $\alpha = 5\%$?

(d) Assume that a reduced model has been fitted. We compare this reduced model to the full model above and obtain

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

What hypotheses can we now test? Formulate them, give the test statistic, the corresponding $p-$value, and the conclusion of the test at $alpha = 5\%$.

Answers:

(a) We have $n - p = 26$ and $p = 4$, and so $n = 26 + 4 = 30$.

(b) The function of the estimated mean is

$$\begin{aligned} E\{Y\} &= 5.3141 - 5.9883\, I\{x_1 = \text{Groupe 2}\} - 6.6344\, I\{x_1 = \text{Groupe 3}\} + 1.1677\, x_2 \\ &= \begin{cases} 5.3141 + 1.1677\, x_2, & \text{si } x_1 = \text{Groupe 1} \\ -0.6742 + 1.1677\, x_2, & \text{si } x_1 = \text{Groupe 2} \\ -1.3203 + 1.1677\, x_2, & \text{si } x_1 = \text{Groupe 3} \end{cases} \end{aligned}$$

(c) We are testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs. H_a: at least one of the β is non-zero. The regression is significant since $(F(3, 26) = 5.286; p = 0.0056)$.

(d) We test

$$H_0 :\ E\{Y\} = \beta_0 + \beta_3\, x_2 \quad \text{vs.} H_a :\ E\{Y\} = \beta_0 + \beta_1\, I\{x_1 = \text{Groupe 1}\} + \beta_2\, I\{x_1 = \text{Groupe 2}\} + \beta_3\, x_2.$$

This is a test for the significance of the categorical predictor `x1`. We will need the residual sum of squares for each model.

For the complete model, we have $15.51 = \sqrt{\text{MSE}} = \sqrt{\text{SSE}/(n-p)} = \sqrt{\text{SSE}/26}$, and so $\text{SSE} = (15.51)^2(26) = 6254.563$.

For the reduced model, we have $28.46512 = \sqrt{\text{MSE(R)}} = \sqrt{\text{SSE(R)}/(n-q)} = \sqrt{\text{SSE}/28}$, and so $\text{SSE} = (28.46512)^2(28) = 22\,687.37$.

ExtraSS: The difference of the residual sums of squares is $\text{ExtraSS} = \text{SSE(R)} - \text{SSE} = 22\,687.37 - 6254.563 = 16\,432.81$.

The value of the test statistic is thus:

$$F_0 = \frac{\text{ExtraSS}/(p-q)}{\text{SSE}/(n-p)} = \frac{16\,432.81/(4-2)}{6254.563/26} = 34.15531.$$

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

The $p-$value is thus $P(F(2, 26) > 34.15531) = 5.31 \times 10^{-8}$, and we conclude that the predictor `x1` is significant.

4. Diagnostics and remedial measures

When assessing the suitability of a linear model, we should follow the following order:

1. identify outliers and influential observations;
2. test for errors in the specification of the mean function;
3. test for heteroscedasticity;
4. test for normality of random errors.

4.1 Diagnostic for the specification of the mean function

We can use the plot of the residuals against the fitted values, and/or we can perform the Ramsey RESET test (Regression Equation Specification Error Test).

Consider the linear model

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

We fit the model to obtain the fitted values

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

for $i = 1, \ldots, n$. Then we add $\hat{y}^2$ and $\hat{y}^3$ as model predictors:

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

The Ramsey RESET test pits

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

If the Ramsey RESET test is significant, then we have evidence of missing higher order effects in the model. This means that we have significant evidence that there is an error in the specification of the mean function.

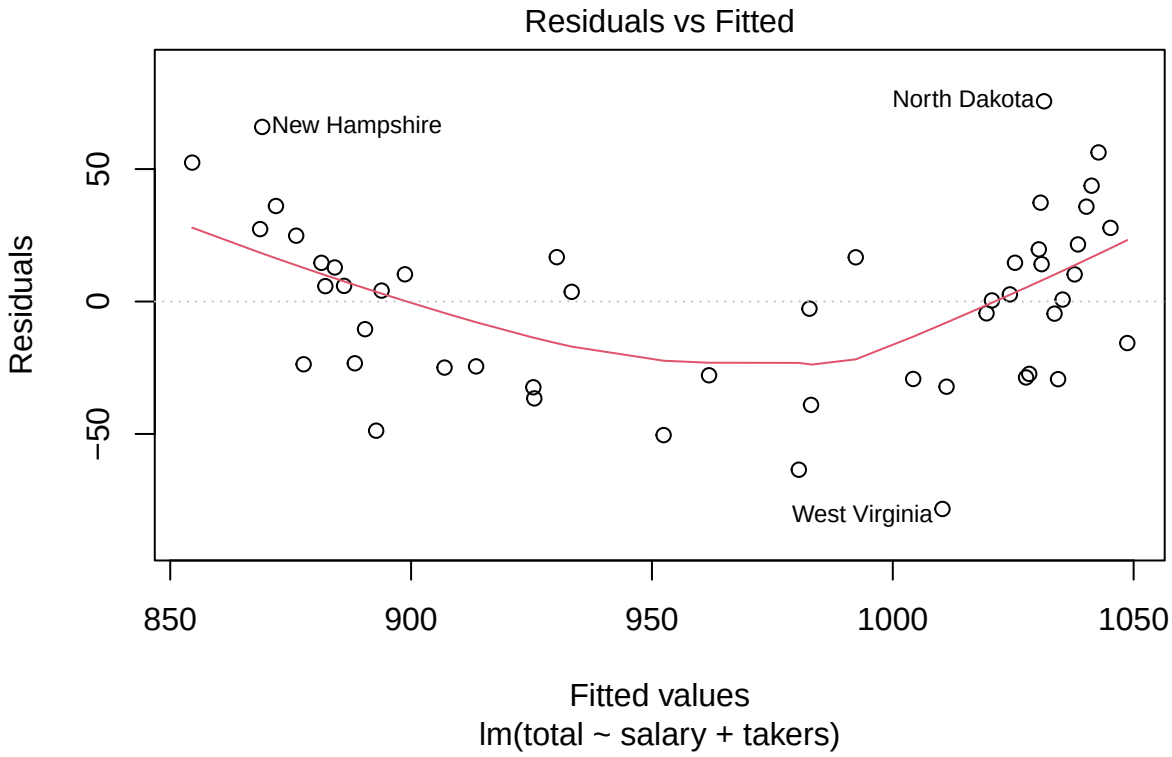
The Ramsey test is general in the sense that it can identify errors in the model specification. However, it cannot tell us what the cause of the error is, only that there are nonlinear effects. We need to study the partial relationship between Y and the predictors and try to find a nonlinear relationship. Alternatively, the nonlinearity could be caused by interactions between the predictors.

Example: we study the `sat` dataset from the `faraway` library.

```
library(faraway)
str(sat)
```

```
## 'data.frame':    50 obs. of  7 variables:
##  $ expend: num  4.41 8.96 4.78 4.46 4.99 ...
##  $ ratio : num  17.2 17.6 19.3 17.1 24 18.4 14.4 16.6 19.1 16.3 ...
##  $ salary: num  31.1 48 32.2 28.9 41.1 ...
##  $ takers: int  8 47 27 6 45 29 81 68 48 65 ...
##  $ verbal: int  491 445 448 482 417 462 431 429 420 406 ...
##  $ math  : int  538 489 496 523 485 518 477 468 469 448 ...
##  $ total : int  1029 934 944 1005 902 980 908 897 889 854 ...
```

```
mod <- lm(total ~ salary + takers, dat=sat)
plot(mod, which=1)
```



The graph of the residuals suggests that mean function is not well specified. We perform the Ramsey RESET test from the `lmtest` library.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
resettest(mod,power=c(2,3))
```

```
##
##  RESET test
##
## data:  mod
## RESET = 13.695, df1 = 2, df2 = 45, p-value = 2.262e-05
```

There is evidence that there is an error in the specification of the model ($F(1, 45) = 13.695$, $p < 0.0001$).

4.2 Homoscedasticity

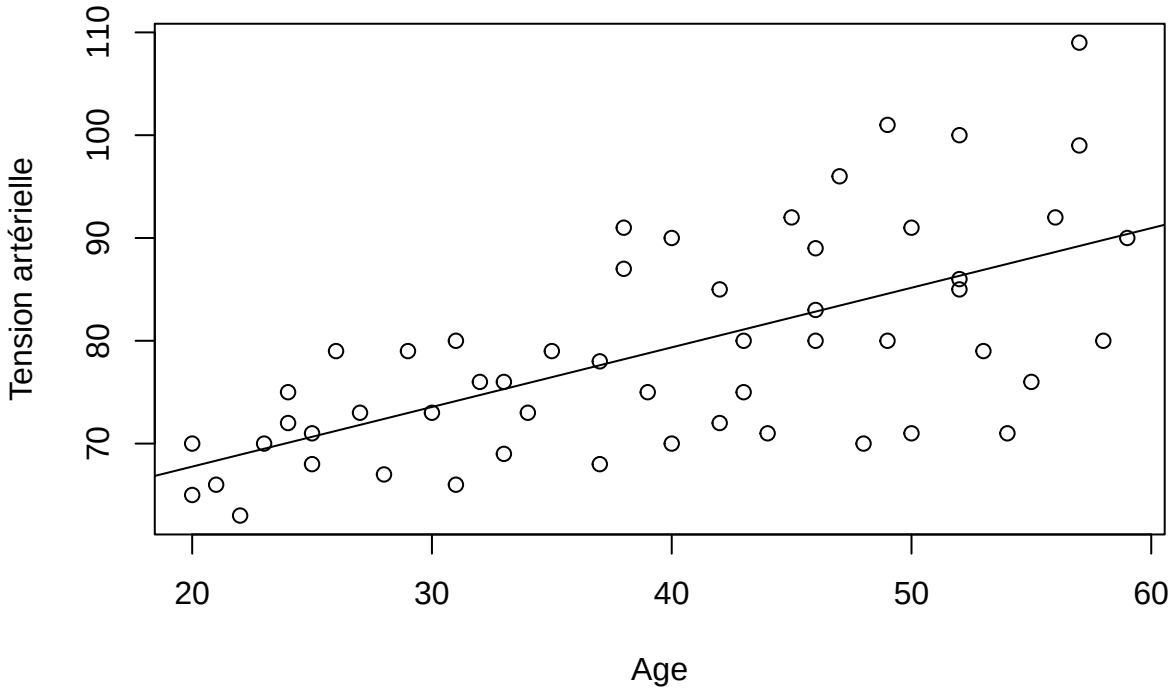
Consider the dataset `BloodPressure.csv`. This is data from a health study where the association between diastolic blood pressure and age is studies on $n = 54$ women aged 20 to 60. We import the data and display some rows.

```
BP <- read.csv("Data/BloodPressure.csv")
str(BP)
```

```
## 'data.frame':    54 obs. of  2 variables:
##  $ Age        : int  27 21 22 24 25 23 20 20 29 24 ...
##  $ DiastolicBP: int  73 66 63 75 71 70 65 70 79 72 ...
```

We fit a linear model to express diastolic blood pressure as a function of age, and we overlay the least squares line on the scatter plot of blood pressure versus age.

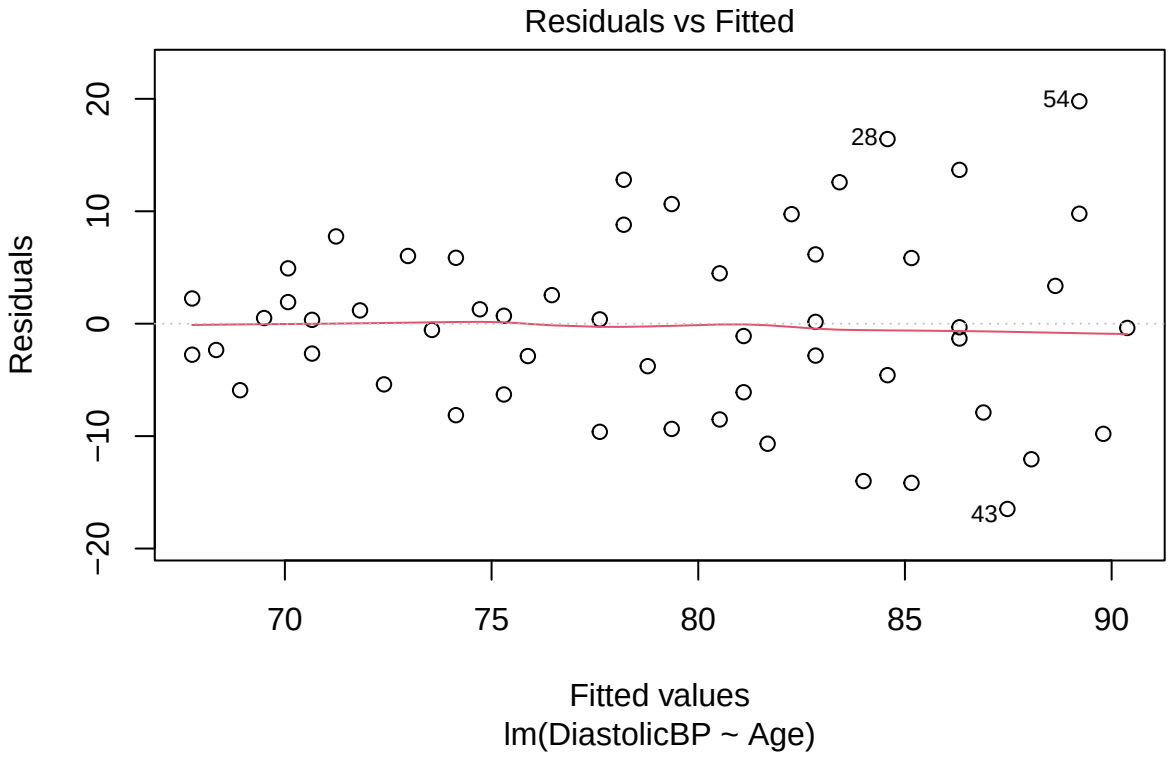
```
mod <- lm(DiastolicBP ~ Age, data=BP)
with(BP, plot(Age, DiastolicBP, ylab="Tension artérielle", xlab="Age"))
abline(mod)
```



The diagram suggests that the relationship between blood pressure and age is linear, but that the variance of the error increases with age. If we wanted to express blood pressure as a linear function of age, we would have to perform a diagnostic on the model specification.

There is no trend in the residual plot. This suggests that blood pressure expressed as a linear function of age is well specified. In addition, there is no significant evidence that the model is misspecified according to the Ramsey RESET test ($F(1, 50) = 0.077; p = 0, 926$).

```
plot(mod,which=1)
```




```
library(lmtest)
resettest(mod,powers=c(2,3))
```

```
##
##  RESET test
##
## data:  mod
## RESET = 0.07704, df1 = 2, df2 = 50, p-value = 0.926
```

After checking that the model is well specified, we can check the homoscedasticity of the model (i.e. that the variance of the error does not depend on the predictor levels/values).

Recall that $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ and that we can show that $\mathbf{e} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H})\right)$. Thus

$$V[e_i] = \sigma^2(1 - h_{ii}), \quad i = 1, 2, \ldots, n.$$

Then, even if the variance of the random error is constant, i.e., $V[\varepsilon_i] = \sigma^2$, for $i = 1, 2, \ldots, n$, the residuals (which are the observed errors) do not necessarily have the same variance. So, to evaluate the homoscedasticity status, we use the standardized residuals

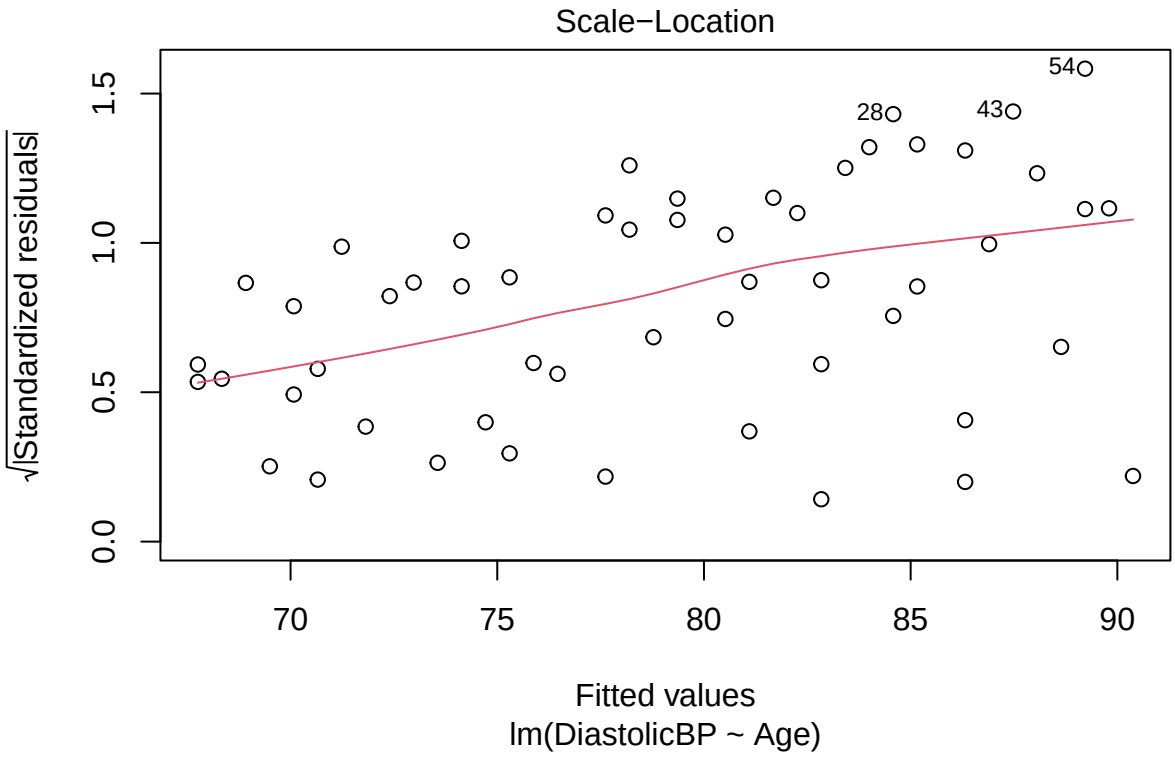
$$r_i = \frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{\text{MSE}(1 - h_{ii})}}, \quad i = 1, \ldots, n.$$

These standardized residuals will have a variance that is approximately equal to 1. We use a scaling and residual location plot as a visual tool to identify heteroscedasticity: $\sqrt{|r_i|}$ vs. $\hat{y}_i$.

If the model is homoscedastic, then we should expect a horizontal trend (close to 1) in the plot. But, if there is a pronounced trend in the plot, this suggests that the variance is not constant.

This plot is obtained with the command `plot(mod, which=3)`, assuming that `mod` is an `lm` object.

```
plot(mod, which=3)
```



There is a positive trend in the scaling and residual location plot for the blood pressure study. Thus, we conclude that the error variance increases with the value of the blood pressure estimate.

But using a visual tool is subjective; are there formal tests to identify heteroscedasticity?

Here are two:

(i) the Breusch-Pagan Studentized test, and

(ii) White's test with 2 degrees of freedom.

The Breusch-Pagan Studentized test is performed as follows:

- **Step 1:** fit the linear model and obtain the residuals;
- **Step 2:** fit a linear model that expresses the square of the residual as a function of the predictors x_1, x_{p-1}, and obtain the corresponding coefficient of determination: $R^2_{e^2|x_1,\ldots,x_{p-1}}$;
- **Step 3:** the test statistic is $X^2_{\text{BP}} = nR^2_{e^2|x_1,\ldots,x_{p-1}}$;
- **Step 4:** the $p-$value of the test is

$$p = P(\chi^2(p-1) \geq X^2_{\text{BP}}),$$

where X^2_{BP} is the observed value of the test statistic.

The Breusch-Pagan test is a score test (a Lagrange multiplier test):

- the test attempts to identify linear heteroscedasticity in $x_1, \ldots, x_{p-1}$; however, it is possible that heteroscedasticity is nonlinear in $x_1, \ldots, x_{p-1}$;
- White (in 1980) suggested fitting (e^2) to the predictors by adding quadratic and bilinear terms to the model; the number of parameters to be estimated is large and the test is often difficult to perform. Here is a simplification of White's test.
- **Step 1:** we fit the linear model and obtain the fitted values;

- **Step 2:** we fit a linear model that expresses the square of the residual as a function of the fitted value $\hat{y}$ and its square $\hat{y}^2$: $R^2_{e^2|\hat{y},\hat{y}^2}$;
- **Step 3:** the test statistic is $X^2_{\mathrm{W}} = nR^2_{e^2|\hat{y},\hat{y}^2}$;
- **Step 4:** the test's $p-$value is
$$p = P(\chi^2(2) \geq X^2_{\mathrm{W}}),$$
where X^2_{W} is the observed value of the test statistic.

Consider the data for blood pressure by age. According to the Breusch-Pagan Studentized test, there is significant evidence of heteroscedasticity ($X^2(2-1) = 12.5412$; $p = 0.000398$). It is not reasonable to assume that the variance of the error is constant.

```
mod <- lm(DiastolicBP ~ Age, data=BP)
# get the residuals
e <- mod$residuals
# fit e^2 vs predictors
mod.BP <- lm(e^2 ~ Age, data=BP)
# get R^2 and n
p <- length(mod.BP$coefficients)
n <- mod.BP$df.residual+p
R.2 <- summary(mod.BP)$r.squared
# observed value of the Breush-Pagan studentized test
n*R.2
```

```
## [1] 12.54124
```

```
# p-value
1-pchisq(n*R.2,p-1)
```

```
## [1] 0.000398068
```

The Breusch-Pagan test assumes as an alternative that the error variance is a monotonic function of the predictors. However, the variance function may be more complex. There may be nonlinear effects that the Breusch-Pagan test fails to identify.

White's 2-degree-of-freedom test for heteroscedasticity may be useful in identifying a variance function that is nonlinear from the predictors.

The Breusch-Pagan test and the White test can sometimes contradict each other:

- if the error variance is a monotonic function of the predictors, then the Breusch-Pagan test will be more powerful than the White test because the latter is more general;
- if the Breush-Pagan test is significant but the White test is not, it is possible that the sample size is too small for the White test;
- if the variance function is non-linear with respect to the predictors, then White's test is more powerful.

Thus, if White's test is significant, but the Breusch-Pagan test is not, then the variance function is most likely a nonlinear and complex function: if the function of the mean is not well specified, then we must proceed with caution in interpreting the significance of the heteroskedasticity tests. It could simply detect the error in the specification of the mean function.

We perform White's test for heteroscedasticity with 2 degrees of freedom. There is significant evidence of heteroscedasticity ($X^2(2) = 12.6517$; $p = 0.0018$).

```
mod <- lm(DiastolicBP ~ Age, data=BP)
# get residuals and fitted values
e <- mod$residuals
y.chapeau <- mod$fitted.values
```

```r
# get n
p <- length(mod$coefficients)
n <- mod$df.residual+p
# White's test
mod.White <- lm(e^2 ~ y.chapeau + I(y.chapeau^2), data=BP)
# get R^2 and n
R.2 <- summary(mod.White)$r.squared
# observed value of the White test
n*R.2
```

```
## [1] 12.65166
```

```r
# p-value
1-pchisq(n*R.2,2)
```

```
## [1] 0.00178948
```

5. Goodness-of-fit test of the linear model

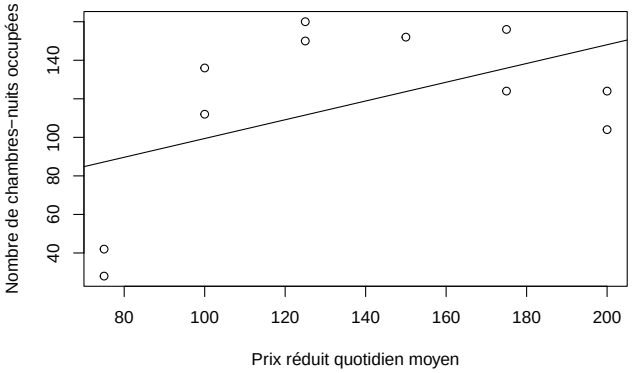
A hotel chain offers a promotion during the month of February: at each of its 11 locations, management reduces the average daily rate, which varies from one location to another, and records the number of additional room-nights that are occupied during the month.

```
avg.price.discount <- c(125,100,200,75,150,175,75,175,125,200,100)
n.add.rooms        <- c(160,112,124,28,152,156,42,124,150,104,136)
hotels             <- data.frame(avg.price.discount,n.add.rooms)
str(hotels)
```

```
## 'data.frame':    11 obs. of  2 variables:
##  $ avg.price.discount: num  125 100 200 75 150 175 75 175 125 200 ...
##  $ n.add.rooms       : num  160 112 124 28 152 156 42 124 150 104 ...
```

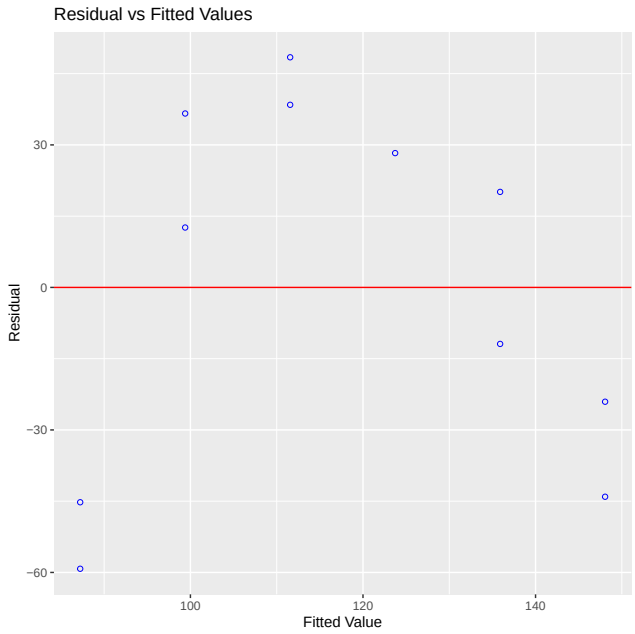
Visually, it seems obvious that the relationship between the two variables is not linear:

```
plot(hotels$avg.price.discount,hotels$n.add.rooms,
     xlab="Prix réduit quotidien moyen",
     ylab="Nombre de chambres-nuits occupées")
mod <- lm(n.add.rooms ~ avg.price.discount, data=hotels)
abline(mod)
```

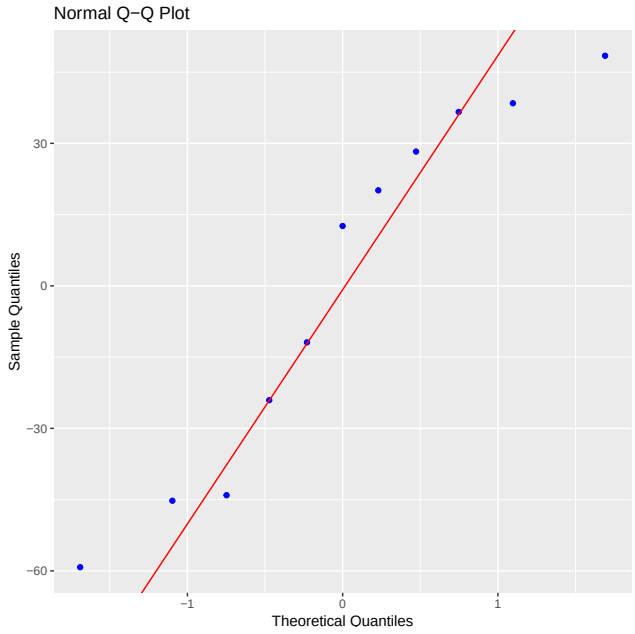


We can obtain the residual charts as follows:

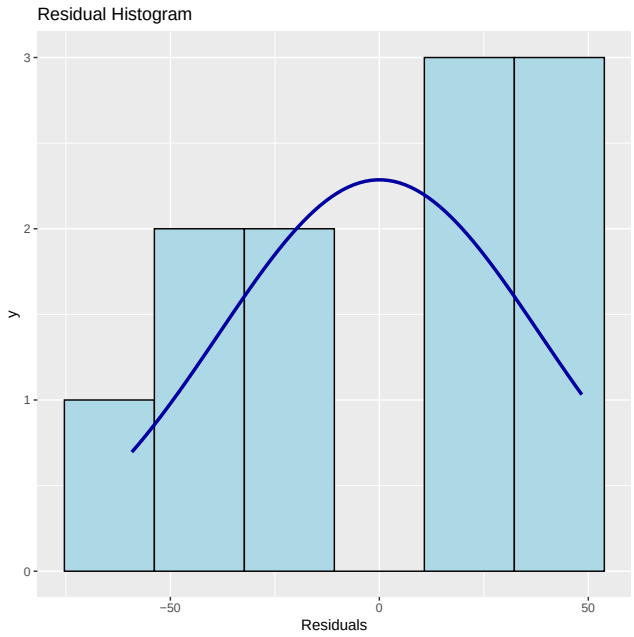
```
# produce residual vs. fitted plot
olsrr::ols_plot_resid_fit(mod)
```




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




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



We pit

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

To confront these two hypotheses, we nest the simple linear regression model in a more general model. We consider a stratification of units according to the value of x, i.e. units in the same group take the same value of x.

We obtain a frequency table for x =average daily discounted price. We observe that there are $c = 6$ groups and that each group contains 2 units except for the $x = 150$ group (which contains only one unit).

```
table(hotels$avg.price.discount)
```

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

If each group has its own average, we can consider that x is a categorical variable with $c = 6$ categories (implemented in `R` using the `factor()` function). We will add a categorical variable to the `hotels` data frame; we also display the levels of this variable. The corresponding ANOVA model is the most complex model possible since we do not impose any structure on $E\{Y \mid X = x\}$.

```
hotels$avg.price.discount.cat <- factor(hotels$avg.price.discount)
levels(hotels$avg.price.discount.cat)
```

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

The (complete) linear model is

$$Y_i = \beta_0 + \beta_1 x_{i,2} + \cdots + \beta_5 x_{i,6} + \varepsilon_i = \begin{cases} \beta_0 = \mu_1 & \text{if the } i\text{th unit is from group 1} \\ \beta_0 + \beta_1 = \mu_2 & \text{if the } i\text{th unit is from group 2} \\ \vdots & \vdots \\ \beta_0 + \beta_5 = \mu_6 & \text{if the } i\text{th unit is from group 6} \end{cases}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. normal random variables $\mathcal{N}(0, \sigma^2)$. This model is sometimes called an ANOVA model; the parameter $\beta_{j-1} = \mu_j - \mu_1$ is the **group effect** j (in comparison with reference group 1).

Here is a summary of the ANOVA model fit:

```
mod.ANOVA <- lm(n.add.rooms ~ avg.price.discount.cat, data=hotels)
summary(mod.ANOVA)
```

```
##
## Call:
## lm(formula = n.add.rooms ~ avg.price.discount.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 *
## avg.price.discount.cat100     89.00      15.15   5.874 0.002030 **
## avg.price.discount.cat125    120.00      15.15   7.919 0.000517 ***
## avg.price.discount.cat150    117.00      18.56   6.305 0.001478 **
## avg.price.discount.cat175    105.00      15.15   6.930 0.000960 ***
## avg.price.discount.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
```

The estimated model is thus

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

The estimate of the variance (of the error) is $s_\varepsilon^2 = \text{MSE} = (15.15)^2 = 229.52$; the coefficient of determination of the ANOVA model is $R^2 = 0.9423$. The coefficient of determination of the reduced model (the simple linear regression model obtained earlier), however, is $R^2 = 0.2586$.

```
mod <- lm(n.add.rooms ~ avg.price.discount, data=hotels)
summary(mod)$r.squared
```

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

The difference in the fit of the two models seems to demonstrate that the assumption of linearity of the reduced model is not justified. Is the evidence significant?

We use the general linear test to compare the two models. In general, the test statistic is

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

where p is the number of parameters of the complete model (ANOVA) and SSE its sum of squares of the residuals (errors), q is the number of parameters of the reduced (linear) model and SSE(R) its sum of squares of the residuals, and MSE is the standard deviation of the residuals of the full model; if H_0 is valid, then $F \sim F(p-q, n-p)$.

In the full model, there are $p = c = 6$ parameters; in the reduced model, there are only $q = 2$ (since $p - q = c - 2 > 0$, we must have at least $c = 3$ values of x). The observed value of the test statistic is calculated using the `anova()` function.

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

```
## Analysis of Variance Table
##
## Model 1: n.add.rooms ~ avg.price.discount
## Model 2: n.add.rooms ~ avg.price.discount.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
```

The observed value of the test statistic is thus

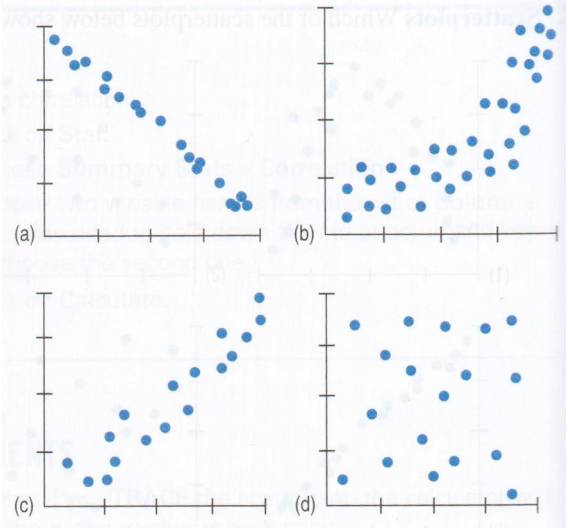
$$F_0 = \frac{(\text{SSE} - \text{SSE}(R))/(c-2)}{\text{MSE}} = \frac{(14742 - 1148)/(6-2)}{1148/5} = 14.801;$$

the $p-$value of the test is $P(F(4,5) > 14.801) = 0.0056$, and so we reject the null hypothesis of linearity of the adjustment.

6. Correlations

6.1 Scatterplot I

Considérons les nuages de points ci-dessous.



Match the following correlations -0.977, -0.021, 0.736, and 0.951, with the scatterplots above.

Answers: (a) -0.977 (b) 0.736 (c) 0.951 (d) -0.021

6.2 Scatterplots II

A horsepower is the power required to lift a weight of 550 pounds over a height of 1 foot in 1 second (or 33,000 pounds in one minute). Horsepowers are measured in terms of the speed at which the work is done.

In the file `Fuel_economy_2007.csv`, we have the claimed horsepower ratings and predicted fuel consumption (in mpg) for several vehicles in 2007.

We import the data with R and display some its rows.

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

```
##             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) Provide a scatterplot of power versus gasoline consumption and overlay a smooth curve on the plot. Describe the orientation and shape of the association.

(b) Consider the following statistics obtained with `R`.

```
x <- cars$Horsepower
y <- cars$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
```

Notes:

- `y-mean(y)` is the vector $(y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})$
- `sum` adds up the components of a vector
- `mean` computes the mean value of the components of a vector.

Using these statistics, calculate the Pearson correlation between power and gasoline consumption.

(c) Assuming that `x` and `y` are numerical vectors of the same size, then the command `cor(x,y)` calculates the Pearson correlation between `x` and `y`. Use the function `cor()` to compute the correlation between horsepower and gasoline consumption.

(d) In Canada, gasoline consumption is described in L/100km. The data, however, is measured in mpg. Let w be the consumption in L/100km and y the consumption in mpg. Here is the conversion formula:

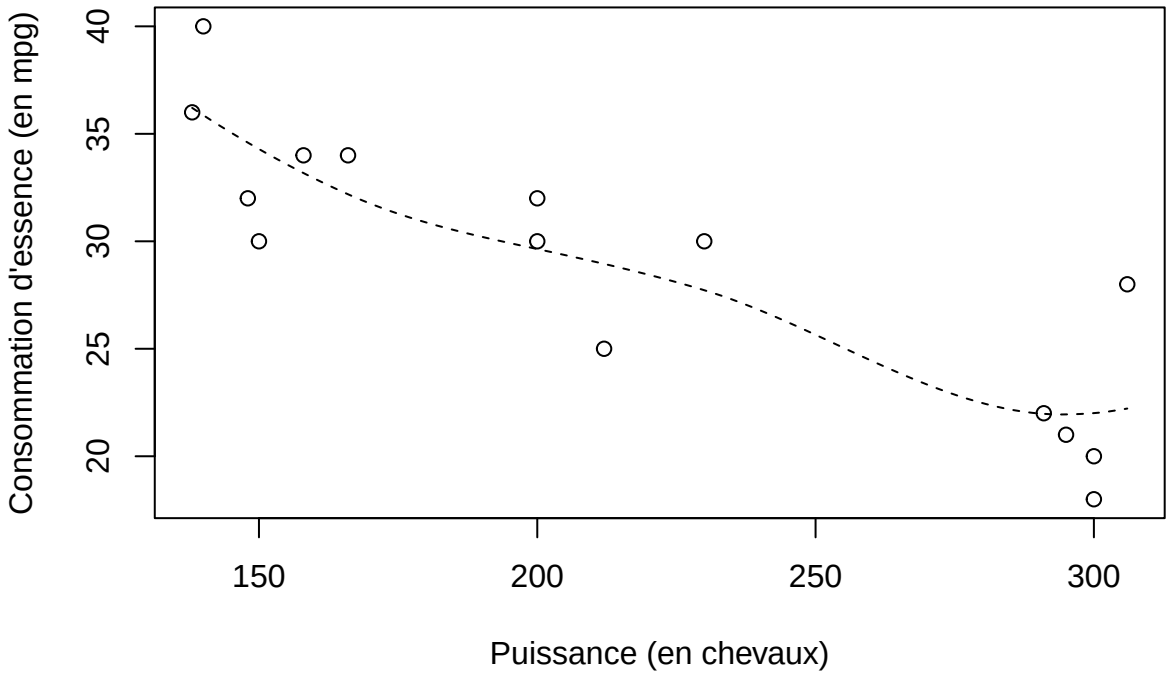
$$w = \frac{235.215}{y}.$$

If we measure gasoline consumption in mpg, then the correlation between power and gasoline consumption is $r = -0.869$. If we measure fuel economy in L/100km, does the correlation between horsepower and fuel economy remain equal to $r = -0.869$? If not, compute the value for the correlation.

Answers:

(a) The association between power and gasoline consumption (in mpg) is approximately linear and negative.

```
with(cars,
plot(x=Horsepower,y=Highway.Gas.Mileage..mpg.,
xlab="Puissance (en chevaux)",
ylab="Consommation d'essence (en mpg)"))
## fit a smooth curve (loess=lowess=locally weighted scatterplot smoothing)
mod.loess<-loess(Highway.Gas.Mileage..mpg.~Horsepower,
data=cars)
## get the range of x
xlim<-range(cars$Horsepower)
## build a new dataset
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) We have $s_{xy} = -5154.2$, $s_{xx} = 61503.6$, and $s_{yy} = 572.4$. The Pearson correlation between x and y is thus:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}\, s_{yy}}} = -0.869.$$

(c) The correlation between power and fuel consumption is $r = -0.867$:

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

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

(d) The conversion formula is not linear, so it is possible that the correlation could change if we measure gasoline consumption in L/100km. We use `R` to convert the fuel consumption to L/100km, and then calculate the correlation between power and fuel consumption (in L/100km). This correlation is $r = 0.851$:

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

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

6.3 Scatterplots III

A person's muscle mass should decrease with age. To explore this association in women, a nutritionist randomly selected 15 women from each of the 10-year age groups starting at age 40 and ending at age 79. The data are in the file `Masse.csv`. There are two variables in this dataset: x = age of the participant, and y = muscle mass of the participant.

We import the data with `R`, and display some rows of the dataset.

```
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
```

Here is the structure of the dataset:

```
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) How many observations are there in this dataset?

(b) We calculate some sums to summarize the data:

```
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))
```

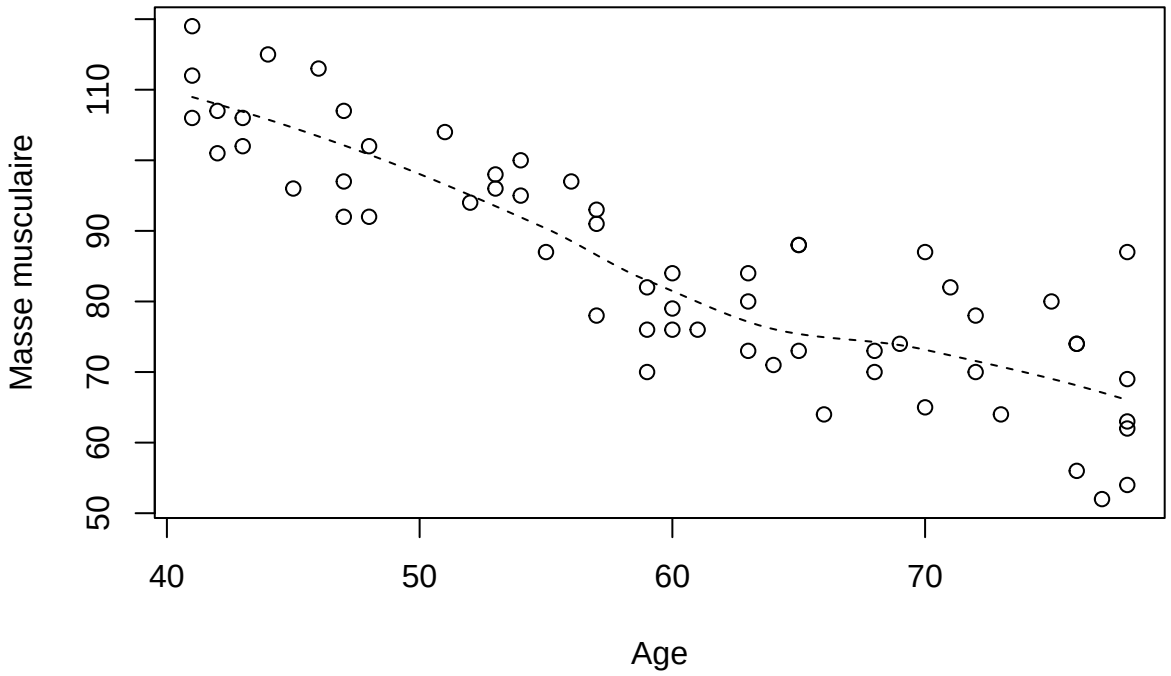
```
## [1]  1830.00  1830.00 73780.00 73794.00 40256.25
```

- (i) Based on these sums, compute the covariance between age and mass and also compute the (Pearson) correlation between age and mass.
- (ii) Based on these sums, compute the Spearman correlation between age and mass.

(c) Here is a scatter plot of muscle mass versus age with a smooth curve overlay. Describe the association between age and muscle mass (orientation, shape, and intensity).

```
with(masse,
plot(x=Age,y=Masse,
xlab="Age",
ylab="Masse musculaire"))
## fit a smooth curve
mod.loess<-loess(Masse~Age,
data=masse)
## get the range for x
xlim<-range(masse$Age)
```

```
## build a new dataset
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)
```



Answers:

(a) There are $n = 60$ observations.

(b) Let' see. . . .

- (i) We have

$$s_{xy} = \left(\sum_{i=1}^{n} x_i\, y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \Big/ n = 296\,024 - (3599)(5098)/60 = -9771.033,$$

So the covariance between age and mass is:

$$\hat{\sigma}_{X,Y} = \frac{s_{xy}}{n-1} = \frac{-9771.033}{60-1} = -165.6107.$$

In addition, we have:

$$s_{xx} = \left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n = 224\,091 - (3599)^2/60 = 8\,210.983,$$

$$s_{yy} = \left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n = 448\,662 - 5098^2/60 = 15\,501.93,$$

So the Pearson correlation between age and mass is

$$r = \frac{s_{xy}}{\sqrt{s_{xx}\, s_{yy}}} = \frac{-9\,771.033}{\sqrt{(8\,210{,}983)(15\,501.93)}} = -0.866.$$

- (ii) We have

$$\begin{aligned} s_{R_x,R_y} &= \left(\sum_{i=1}^{n} R_{x,i}\, R_{y,i}\right) - \left(\sum_{i=1}^{n} R_{x,i}\right)\left(\sum_{i=1}^{n} R_{y,i}\right) \Big/ n \\ &= 40\,256.25 - (1830)(1830)/60 = -15\,558.75. \end{aligned}$$

In addition, we have:

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

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

so the Spearman correlation between age and mass is

$$r_S = \frac{s_{R_x R_y}}{\sqrt{s_{R_x R_x}\, s_{R_y R_y}}} = \frac{-15\,558.75}{\sqrt{(17\,965)(17\,979)}} = -0.8657.$$

(c) The association between age and mass is approximately linear, and negative, with a Pearson correlation of $r = -0.866$.

6.4 Inference concerning a correlation

We have data from a study with healthy volunteers. A stimulus is applied to the subject's fingers and the spinal cord conduction velocity (VC) is measured. We want to describe the association between the height of the individual (in cm) and the spinal cord conduction velocity for healthy individuals.

We import the data with R and display some rows.

```
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
```

Test $H_0 : \rho = 0$ (o'u ρ is the (Pearson) correlation between the conduction velocity and the size of the individual) vs. $H_1 : \rho \neq 0$.

Answer:

The correlation in the dataset is

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

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

This value is not very close to 0. We can use the function `cor.test()` to obtain a confidence interval of the correlation 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
```

The observed value of the test statistic is

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

The $p-$value of the test is $2P(T(153) > |19.781|) < 0.00001$; the evidence suggests that the correlation is indeed non-zero.

7. Probability and statistics

7.1 Probability I

Compute the following probabilities assuming that T follows a $t(15)$ distribution.

(a) $P(T > 2.45)$;

(b) $P(T < 2.45)$;

(c) $2P(T > 4.34)$.

Answers:

(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 Probability II

Obtain the following quantiles:

(a) 95^{th} centile of the $t(34)$ distribution;

(b) 97.5^{th} centile of the $t(44)$ distribution.

Answers:

(a) LThe 95^{th} centile of the $t(34)$ distribution is $t(0.95; 34) = 1.6909$.

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

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

(b) the 97.5^{th} centile of the $t(44)$ distribution is $t(0.975; 44) = 2.0154$.

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

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

7.3 Probability III

Let $Y \sim N(\mu = 125, \sigma^2 = 25)$ and $(1/\sigma^2)\, V \sim \chi^2(10)$. Assume that Y and V are independent.

(a) Compute

$$P\left(\frac{Y - 125}{\sqrt{V/10}} > 2.75\right).$$

(b) Compute

$$P\left(\frac{(Y-125)^2}{V/10} > 7.12\right).$$

Answers:

We have

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

and Z is independent from $U = (1/\sigma^2)\,V \sim \chi^2(10)$, so that

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

Thus

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

(a) We have

$$P\left(\frac{Y-125}{\sqrt{V/10}} > 2.75\right) = P(t(10) > 2.75) = 0.0102.$$

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

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

(b) We want

$$P\left(\frac{(Y-125)^2}{V/10} > 7.12\right) = P(F(1,10) > 7.12) = 0.02356.$$

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

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

7.4 Statistics I

Suppose that $\hat{\theta}$ is an estimator of an unknown parameter θ, and that

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

From a random sample, we observe $\hat{\theta} = -3.2$ and $s\{\hat{\theta}\} = 4.5$.

a. Test $H_0:\ \theta = 0$ against $H_a:\ \theta \neq 0$ at $\alpha = 5\%$. Give the observed value of the test statistic t and the conclusion of the test.

b. Give a 95% confidence interval for θ.

Answers:

a. The observed value of the t test statistic is

$$t_0 = \frac{\hat{\theta} - 0}{s\{\hat{\theta}\}} = \frac{-3.2 - 0}{4.5} = -0.71111.$$

Since $|t_0| = 0.71111 < 2.13145 = t(0.975; 10)$, then the evidence against H_0 is not significant at $\alpha = 5\%$.

b. A 95% confidence interval for θ is given by:

$$\hat{\theta} \pm t(0.975; 15)\, s\{\hat{\theta}\} = -3.2 \pm 2.13145\,(4.5) =]-12.792; 6.392[.$$

7.5 Statistics II

Suppose that $\hat{\theta}$ is an estimator of an unknown parameter θ, and that

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

From a random sample, we observe $\hat{\theta} = 2.5$ and $s\{\hat{\theta}\} = 0.75$.

a. Test $H_0: \theta = 0$ against $H_a: \theta \neq 0$ at $\alpha = 5\%$. Give the observed value of the test statistic t and the conclusion of the test.

b. Give a 95% confidence interval for θ.

Answers:

a. The observed value of the t test statistic is

$$t_0 = \frac{\hat{\theta} - 0}{s\{\hat{\theta}\}} = \frac{2.5 - 0}{0.75} = 3.3333.$$

Since $|t_0| = 3.3333 \geq 2.08596 = t(0.975; 20)$, then the evidence against H_0 and in favour of H_1 is significant at $\alpha = 5\%$.

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

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

b. A 95% confidence interval for θ is given by:

$$\hat{\theta} \pm t(0.975; 15)\, s\{\hat{\theta}\} = 2.5 \pm 2.08596\,(0.75) =]0.936; 4.064[.$$

7.6 Statistiques III

Suppose that $\hat{\theta}$ is an estimator of an unknown parameter θ, and that

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

From a random sample, we observe $\hat{\theta} = -0.211$ and $s\{\hat{\theta}\} = 3.235$.

a. Test $H_0: \theta = 0$ against $H_a: \theta \neq 0$ at $\alpha = 5\%$. Give the observed value of the test statistic t and the $p-$value of the test.

b. What can you conclude at $\alpha = 5\%$.

Answers:

a. The observed value of the t test statistic is

$$t_0 = \frac{\hat{\theta} - 0}{s\{\hat{\theta}\}} = \frac{-0.211 - 0}{3.235} = -0.06522.$$

The $p-$value of the test is $2\,P(t(28) \geq |-0.06522|) = 0.948$.

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

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

b. The evidence against $\theta = 0$ in favour of $\theta \neq 0$ is not significative at $\alpha = 5\%$ ($t(28) = -0.06522; p = 0.948$).

7.7 Statistics IV

Suppose that $\hat{\theta}$ is an estimator of an unknown parameter θ, and that

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

From a random sample, we observe $\hat{\theta} = 2.5$ and $s\{\hat{\theta}\} = 0.75$.

a. Test $H_0: \ \theta = 0$ against $H_a: \ \theta \neq 0$ at $\alpha = 5\%$. Give the observed value of the test statistic t and the $p-$value of the test.

b. What can you conclude at $\alpha = 5\%$. D'un échantillon aléatoire, on observe $\hat{\theta} = 2.5$ et $s\{\hat{\theta}\} = 0.75$.

Answers:

(a) The observed value of the t test statistic is

$$t_0 = \frac{\hat{\theta} - 0}{s\{\hat{\theta}\}} = \frac{2.5 - 0}{0.75} = 3.3333.$$

The $p-$value of the test is $2\,P(t(28) \geq |3.3333|) = 0.0024$.

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

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

The evidence against $\theta = 0$ in favour of $\theta \neq 0$ is significative at $\alpha = 5\%$ ($t(28) = 3.3333; p = 0.0024$).

7.8 Probability IV

Let Y_1, Y_2, Y_3 be normal independent random variables with

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

and

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

What is the distribution of $W = 2\,Y_1 + 3\,Y_2 - Y_3$?

Answer:

We have $W \sim N(E\{W\}; V[W])$ where

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

and

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

8. Fitting a linear model

8.1 Linear regression I

Suppose that $\hat{y} = b_0 + b_1\, x$ is a linear model estimated by the least squares method. Find the missing values in the table below.

Notation:

- $\bar{x}$ and s_x are the sample mean and standard deviation for the variable X.
- $\bar{y}$ and s_y are the sample mean and standard deviation for the variable Y.
- r is the Pearson correlation (of the sample) of X and 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	$\hat{y} = 30 - 2\, x$

From introductory statistics courses, we know that the sample variance and standard deviation for X are respectively

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

For Y, it's the same idea: $s_y = \sqrt{s_{yy}/(n-1)}$.

Answers:

1. The line of best fit is $\hat{y} = b_0 + b_1\, x$, where

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

Here is the table with the missing values. The calculations are just below the table.

	$\overline{x}$	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) We have

$$b_1 = \frac{r\, s_y}{s_x} = \frac{(-0.2)(6)}{4} = -0.3 \quad \text{and} \quad b_0 = \overline{y} - b_1\,\overline{x} = 18 - (-0.3)(30) = 27.$$

(ii) We have

$$b_1 = \frac{r\, s_y}{s_x} = \frac{(0.9)(10)}{18} = 0.5 \quad \text{and} \quad b_0 = \overline{y} - b_1\,\overline{x} = 60 - (0.5)(100) = 10.$$

(iii) We have

$$15 = b_1 = \frac{r\, s_y}{s_x} = \frac{r\,(15)}{0.8} \quad \text{and} \quad -10 = b_0 = \overline{y} - b_1\,\overline{x} = 50 - (15)\,\overline{x}.$$

Thus, $r = 15\,(0.8)/15 = 0.8$ and $\overline{x} = (50 - (-10))/15 = 4$. \\ (iv) We have

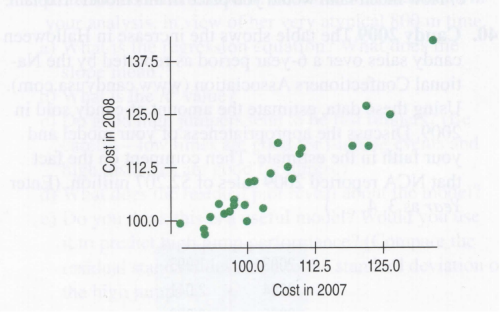
$$-2 = b_1 = \frac{r\, s_y}{s_x} = \frac{(-0.6)\,(4)}{s_x} \quad \text{and} \quad 30 = b_0 = \overline{y} -_1\, \overline{x} = 18 - (-2)\,\overline{x}.$$

Thus, $s_x = (-0.6)\,(4)/(-2) = 1.2$ and $\overline{x} = (30 - 18)/2 = 6$.

8.2 Linear regression II

Here is a scatterplot of observations:

City	2007	2008	City	2007	2008
Moscow	134.4	142.4	Tel Aviv	97.7	105.0
Tokyo	122.1	127.0	Sydney	94.9	104.1
London	126.3	125.0	Dublin	99.6	103.9
Oslo	105.8	118.3	Rome	97.6	103.9
Seoul	122.4	117.7	St. Petersburg	103.0	103.1
Hong Kong	119.4	117.6	Vienna	96.9	102.3
Copenhagen	110.2	117.2	Beijing	95.9	101.9
Geneva	109.8	115.8	Helsinki	93.3	101.1
Zurich	107.6	112.7	New York City	100.0	100.0
Milan	104.4	111.3	Istanbul	87.7	99.4
Osaka	108.4	110.0	Shanghai	92.1	98.3
Paris	101.4	109.4	Amsterdam	92.2	97.0
Singapore	100.4	109.1			



A cost of living survey determined the cost of living in the 25 most expensive cities in the world. This ranking takes New York City as 100 and expresses the other cities as a percentage of the cost of living in New York. For example, the cost of living in Tokyo in 2007 is 122.1, so the cost of living in Tokyo was 22.1% higher than in New York in 2007. The standard deviation of the cost of living in 2007 is 11.9147; while it is 10.8517 for 2008.

(i) The least squares line to predict the cost of living in 2008 as a function of the cost of living in 2007 is

$$\widehat{\text{cost08}} = 21.75 + 0.84\,(\text{cost07}).$$

Compute the (Pearson) correlation between the cost of living in 2007 and 2008.

(ii) Describe the association between the cost of living in 2007 and 2008.

(iii) Compute the coefficient of determination R^2 and interpret its value in the context of this study.

(iv) Use the least squares line from (i) to calculate the residual for Oslo.

(v) What does the residual calculated in (iv) tell us about Oslo?

Answers:

(i) We have

$$0.84 = b_1 = r\,s_y/s_x = r\,(10.8517/11.9147) \quad \Rightarrow \quad r = (11.9147)(0.84)/10.8517 = 0.922.$$

(ii) The association between the cost of living in 2007 and 2008 is positive and nearly linear with a correlation of 0.922.

(iii) $R^2 = r^2 = (0.922)^2 = 0.8501$. Thus, 85% of the variability in the cost of living in 2008 is explained by the linear cost of living model in 2007.

(iv) We have that cost07 = 105.8 and cost08 = 118.3 for Oslo. The fitted value for Oslo is thus

$$\widehat{\text{cost08}} = 21.75 + 0.84\,(105.8) = 110.622.$$

The Oslo residual is thus $e = \text{cost08} - \widehat{\text{cost08}} = 118.3 - 110.622 = 7.678$.

(v) The residual is positive, so the expected cost of living for 2008 is lower than the observed cost of living for 2008 by 7.678 units.

8.3 Linear Regression III

Assume a regression model where $Y_1, \ldots, Y_{50}$ are normal independent variables with a common variance $sigma^2$. We have two models for the mean function. We use the least squares method to estimate the parameters of the mean function for both models. Then we compute the residual sum of squares for both models. We obtain

$$\text{for model 1: SSE} = 1222; \quad \text{for model 2: SSE} = 995.$$

(a) According to the residual sum of squares, which of the models is best?

(b) According to the max of the log likelihood, which of the models is the best? Is this surprising?

Answers:

(a) Model 2 has the smallest residual sum of squares. So according to SSE, model 2 is the best fit to the data.

(b) For model 1, we have

$$\ell = -(n/2)\,\ln(2\,\pi\,\text{SSE}/n) - n/2 = -(50/2)\,\ln(2\,\pi\,(1222/50)) - 50/2 = -150.8525.$$

For model 2, we have

$$\ell = -(n/2)\,\ln(2\,\pi\,\text{SSE}/n) - n/2 = -(50/2)\,\ln(2\,\pi\,(995/50)) - 50/2 = -145.7149.$$

The larger of these two statistics is for model 2. Thus, according to the maximum log-likelihood statistic model 2 is better. This result should not be surprising since SSE and ℓ are equivalent in the sense that they always prefer the same model.

8.4 Linear regression IV

The data concerning the resistance (x) (in ohms) and the failure time (y) (in minutes) of some overloaded resistors are in the file `defaillance.csv`.

We import the data and display the column names, the column standard deviations, the column means, and the dataset size.

```
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
```

Note: the `apply()` function allows us to apply a function to all the columns of a dataframe. The command `sapply(defaillance,sd)` applies the `sd` function to all the columns of the `defaillance` dataset.

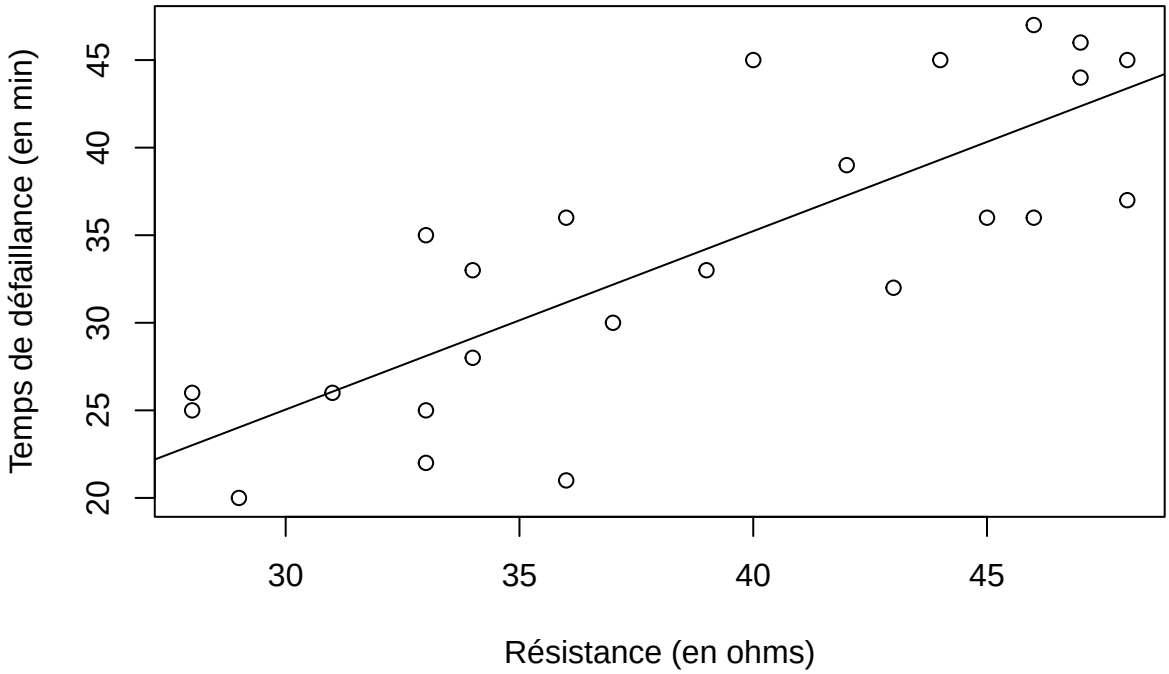
Here is the (Pearson) correlation between failure time and resistance.

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

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

Here is a scatterplot for the failure time against resistance with the estimated regression line overlay.

```
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) Give the least squares line that expresses the failure time as a function of resistance.

(b) Give the value of R^2 (the coefficient of determination) and interpret it in the context of this question.

(c) Give an estimate of the variance of the σ^2 error.

Answers:

We have:

```
x = defaillance$resistance
y = defaillance$temps.de.defaillance
(bar.x = mean(x))
```

```
## [1] 38.625
```

```
(bar.y = mean(y))
```

```
## [1] 33.83333
```

```
(s.x = sd(x))
```

```
## [1] 6.781128
```

```
(s.y = sd(y))
```

```
## [1] 8.544852
```

```
(s.xx = sum((x-mean(x))^2))
```

```
## [1] 1057.625
```

```
(s.xy = sum((x-mean(x))*(y-mean(y))))
```

```
## [1] 1077.5
```

```
(s.yy = sum((y-mean(y))^2))
```

```
## [1] 1679.333
```

```
(r = cor(x,y))
```

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

```
(n = length(x))
```

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

(a) The slope of the line of best fit is:

$$b_1 = r\,\frac{s_y}{s_x} = 0.80851\left(\frac{8.54485}{6.78113}\right) = 1.01880.$$

```
(b1 = r*s.y/s.x)
```

```
## [1] 1.018792
```

```
(b1 = s.xy/s.xx)
```

```
## [1] 1.018792
```

The intercept of the line of best fit is:

$$b_0 = \bar{y} - b_1\,\bar{x} = 33.83333 - (1.01880)(38.62500) = -5.51782.$$

```
(b0=bar.y-b1*bar.x)
```

```
## [1] -5.517512
```

The line of best fit is thus

$$\hat{y} = -5.51782 + 1.01880\,x.$$

(b) We have $R^2 = r^2 = (0.80851)^2 = 0.653$. Then, 65.3% of the variability in failure time is explained by the linear model.

```
(R.2 = r^2)
```

```
## [1] 0.6536811
```

(c) We have

$$\begin{aligned} \text{SSE} &= s_{yy} - \text{SSR} = s_{yy} - b_1^2\,s_{xx} = (n-1)s_y^2 - b_1^2\,(n-1)s_{xx}^2 \\ &= (24-1)(8.54485)^2 - 1.01880^2\,(24-1)(6.78113)^2 = 581.5664. \end{aligned}$$

The estimate of σ^2 is

$$\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{581.5664}{24-2} = 26.43484.$$

```
(MSE = (s.yy - b1^2*s.xx)/(n-2))
```

```
## [1] 26.43567
```

9. Multiple linear regression

9.1 Multiple Linear Regression I

Suppose we have a linear model with $p-1=9$ predictors and $n=125$ observations. We fit the model and calculate the residual sum of squares $\sum_{i=1}^{125} e_i^2 = 356$. The sample variance for the dependent variance is $s_y^2 = 34$.

(a) Give an estimate of the variance of the error.

(b) Give the residual standard deviation.

(c) Calculate the coefficient of determination R^2.

Answers:

(a) We provide an estimate of σ^2: $\text{MSE} = \sqrt{\text{SSE}/(n-p)} = 356/(125-10) = 3.09565$.

(b) The standard deviation of the residuals is $s_e = \sqrt{\text{MSE}} = \sqrt{3.09565}$.

(c) We have $s_{yy} = (n-1)\, s_y^2 = 126\,(34) = 4\,284$ and $\text{SSR} = s_{yy} - \text{SSE} = 4\,284 - 356 = 3928$. The coefficient of determination is thus

$$R^2 = \frac{\text{SSR}}{s_{yy}} = \frac{3\,928}{4\,284} = 0.9169.$$

9.2 Multiple Linear Regression II

We fitted a linear model with the following mean function:

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

We also extracted the design matrix $\mathbf{X}$ with `R`.

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

We compute the inverse of the matrix $X'X$, i.e. we obtain $(X'X)^{-1}$.

```
> ## inverse of (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
```

We also display the parameter estimate of the function of the mean, the residual standard deviation, and the number of degrees of freedom of the error.

```
> 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
```

Note: the symbol for matrix mutiplication in R is `%*%`; if `A` is an invertible matrix, then `solve(A)` computes its inverse.

(a) Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. Use a significance level of $\alpha = 5\%$.

(b) Give a 95% confidence interval for β_1.

Answers:

The residual standard deviation is $s_e = \sqrt{\text{MSE}} = 10.0958$. The estimate of β_1 is $b_1 = 19.9823858$ and the standard error of the estimate is

$$s\{b_1\} = \sqrt{\text{MSE}(X'X)^{-1}_{11}} = s_e\sqrt{(X'X)^{-1}_{11}} = 10.0958\sqrt{2.706184 \times 10^{-4}} = 0.16608.$$

Note that the coefficient indices range from $j = 0, 1, 2, \ldots, p-1$. Thus, $(X'X)^{-1}_{1,1}$ is the second value in the main diagonal of $X'X$.

The observed value of the test statistic is

$$t_0 = \frac{b_1}{s\{b_1\}} = \frac{19.9823858}{0.16608} = 120.3178.$$

The $p-$value is thus $2\,P(t(36) \geq 120,3178) < 0.0001$.

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

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

At a significance level of $\alpha = 5\%$, the predictor x_1 is significant.

(b) Here is a 95% confidence interval for β_1:

$$b_1 \pm t(0.975; 36)s\{b_1\} =]19.65; 20.32[.$$

where $t(0.975; 36) = 2.02809$, $b_1 = 19.9823858$ et $s\{b_1\} = 0.16608$.

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

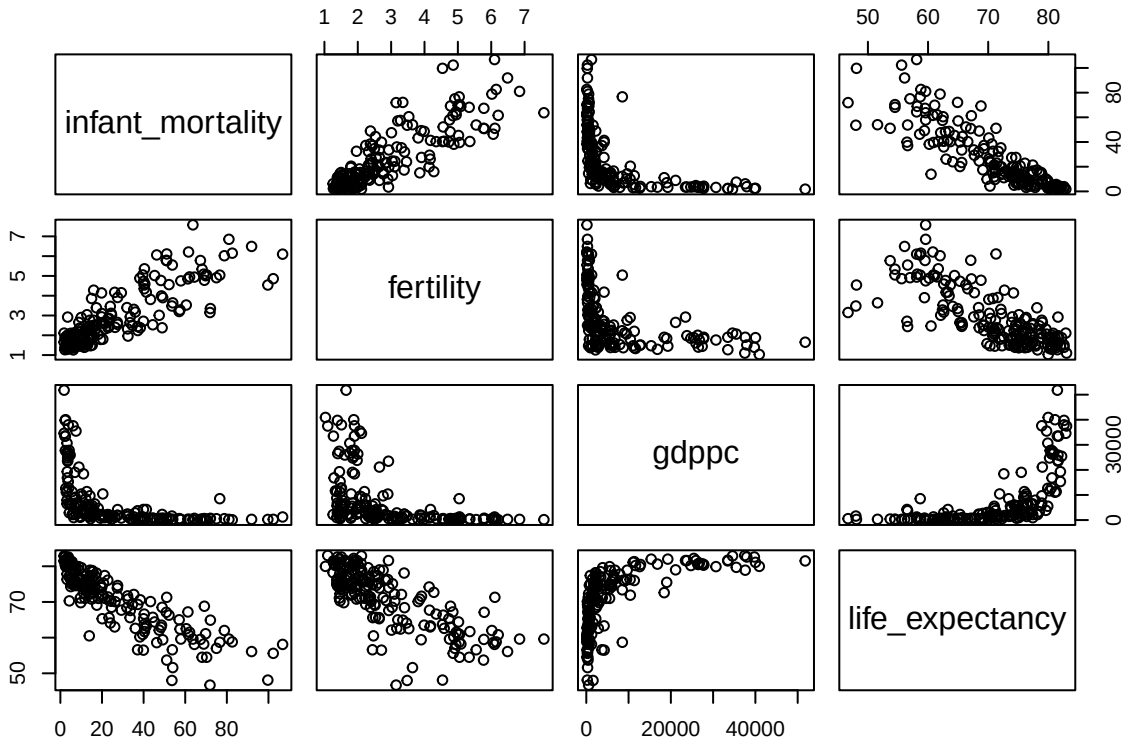
```
## [1] 2.028094
```

9.3 Multiple Linear Regression

```
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))
```

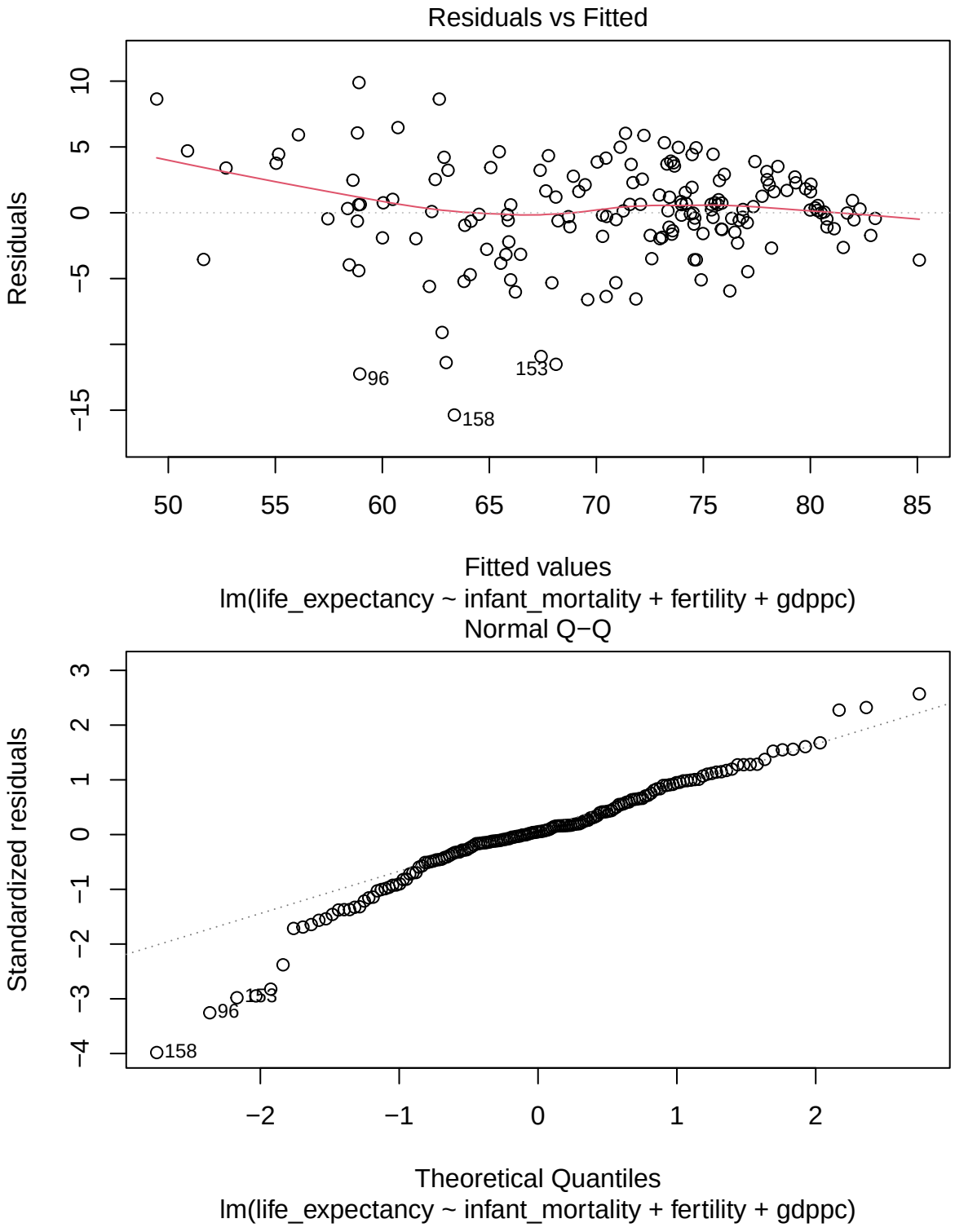
```
## $dim
## [1] 7 4
##
## $dimnames
## $dimnames[[1]]
## [1] "" "" "" "" "" "" ""
##
## $dimnames[[2]]
## [1] "infant_mortality" "  fertility"      "    gdppc"        "life_expectancy"
```

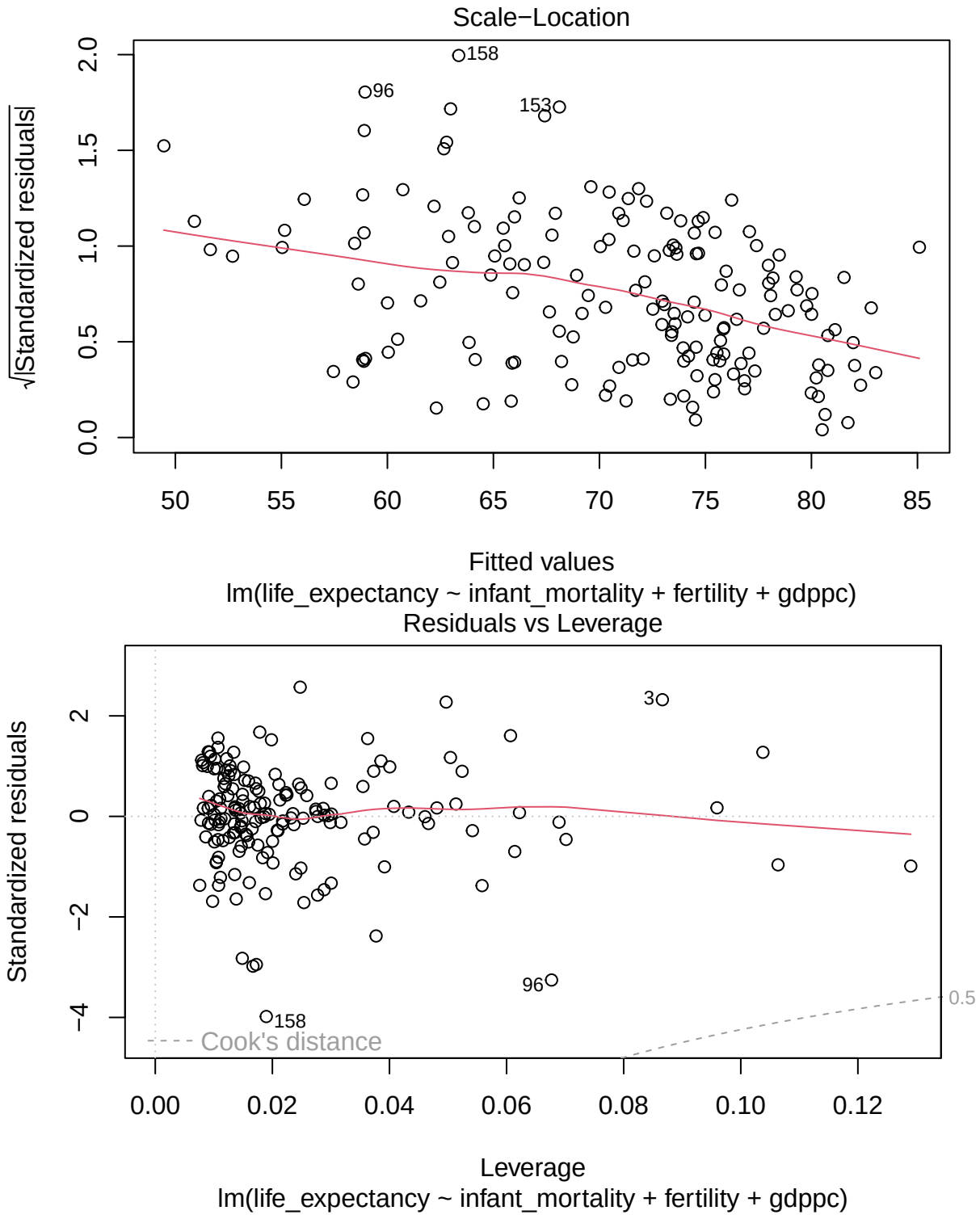
```
##
##
## $class
## [1] "table"
```

```
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)
```


Residuals vs Fitted
Residuals
10
5
0
−5
−15
96
153
158
50
55
60
65
70
75
80
85
Fitted values
lm(life_expectancy ~ infant_mortality + fertility + gdppc)
Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−4
96
153
158
−2
−1
0
1
2
Theoretical Quantiles
lm(life_expectancy ~ infant_mortality + fertility + gdppc)




```
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(gapı

#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 Multiple Linear Regression

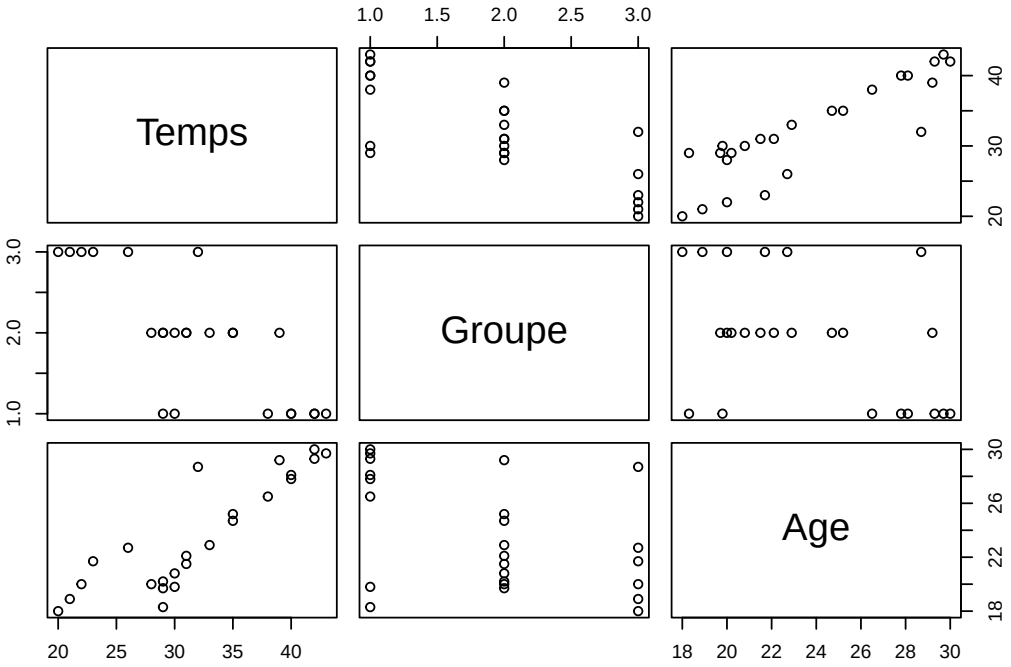
Consider data from an observational study to describe time to recovery (in months) based on a patient's pre-surgery form. Patient fitness level is a categorical predictor with 3 levels: 1=below average; 2=average; 3=above average.

We import the data and display the structure of the dataset.

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

```
## 'data.frame':    24 obs. of  3 variables:
##  $ Temps : int  29 42 38 40 43 40 30 42 30 35 ...
##  $ Groupe: int  1 1 1 1 1 1 1 1 2 2 ...
##  $ Age   : num  18.3 30 26.5 28.1 29.7 27.8 19.8 29.3 20.8 25.2 ...
```

```
plot(genou)
```



Note: genou = knee, Temps = Recovery time

The variable `Groupe` (the fitness level) is a numerical variable; but in reality it is a categorical variable. We will transform it into a factor (a type of categorical variable in `R`).

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

```
## 'data.frame':    24 obs. of  3 variables:
##  $ Temps : int  29 42 38 40 43 40 30 42 30 35 ...
##  $ Groupe: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 2 2 ...
##  $ Age   : num  18.3 30 26.5 28.1 29.7 27.8 19.8 29.3 20.8 25.2 ...
```

With a factor, you can display the levels, a frequency table, and the coding of the factor (for modelling).

We display the levels; we notice that there are three of them.

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

```
## [1] "1" "2" "3"
```

Here is the frequency table for the variable 'Group'. We can see that this is not a balanced study; the number of observations is not constant in each group.

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

```
##
##  1  2  3
##  8 10  6
```

There are two **dummy variables**, one for group 2 and one for group 3. In the following, each column is a dummy variable. Dummy 2 takes the value 1 only if the observation is in group 2, otherwise it is 0; dummy 3 takes the value 1 only if the observation is in group 3, otherwise it is 0.

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

```
##   2 3
## 1 0 0
## 2 1 0
## 3 0 1
```

The ANOVA model is fitted and a summary of the fit is displayed.

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

```
##
## Call:
## lm(formula = Temps ~ Groupe, data = genou)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -9.0   -3.0   -0.5    3.0    8.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.000      1.574  24.149  < 2e-16 ***
## Groupe2       -6.000      2.111  -2.842  0.00976 **
## Groupe3      -14.000      2.404  -5.824 8.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.451 on 21 degrees of freedom
## Multiple R-squared:  0.6176, Adjusted R-squared:  0.5812
## F-statistic: 16.96 on 2 and 21 DF,  p-value: 4.129e-05
```

The ANOVA model is significant ($F(2, 21) = 16.96$; $p < 0.0001$). This test is sometimes called an analysis of variance (ANOVA). It is a test for equality of means.

Thus, we can conclude that there is significant evidence that the mean time to recovery varies by patient group. Since there is only one predictor in the model, we can also display the ANOVA table for testing the significance of the regression.

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: Temps
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Groupe     2    672  336.00  16.962 4.129e-05 ***
```

```
## Residuals 21    416   19.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, the estimated recovery time for a patient with a **below average** fitness level is 38 months; the average recovery time for a patient with an **average** fitness level is \$38 - 6 = \$32 months, and that for a patient with an **above average** fitness level is \$38 - 14 = \$24 months. In addition, the residual standard deviation is 4.451 months.

In the knee surgery example, we have an **observational factor**. The researchers did not assign the pre-surgery form to the patient; they come up with a certain form. It is possible that the observed differences can be explained by a confounding variable.

For example, the researchers may be unlucky in that group 1 is populated by older patients, for instance. For observational studies, it is important to use the knowledge in the field of application (domain expertise) and to try to come up with potential confounding variables.

In medical applications, age and gender are often used as confounding variables. Here, we control for the age of the participant. Gender is often an important explanatory variable in medicine; the common practice is to separate the sexes. Studies often use only men or only women: in this study, there are only men.

We will describe the recovery time by patient fitness level and age. The general linear model has a categorical and a quantitative predictor. The systematic function of the mean response time is

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

β_0 is the average rehabilitation time of a patient with an inferior fitness level, of age 0. But this is meaningless since we cannot deal with a patient of age zero.

To give meaning to the intercept, we will center the quantitative predictor around the mean $x = 23.575$. We could also use a value close to the mean, like 24.

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

```
## [1] 23.575
```

The model then becomes:

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

Interpretation of the parameters:

- β_0 is the average recovery time of a 24-year-old patient with below-average form;
- whatever the fitness level of the patient, the rate of change of $E\{Y\}$ with respect to age is β_3;
- for two patients of the same age, the average recovery time between a patient of average and below-average fitness levels is β_1 (group 2 effect);
- for two patients of the same age, the mean recovery time between a patient with above-average and below-average fitness levels is β_2 (group 3 effect).

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

A general linear model is fitted to describe mean recovery time by pre-surgical fitness level and patient age. The model is significant (\$F(3,20) = \$1170; $p < 0.0001$) and $R^2 = 0.9943$.

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

```
##
## Call:
## lm(formula = Temps ~ Groupe + Age.c, data = genou)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03891 -0.36892  0.05891  0.33098  0.89991
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.44656    0.20842 170.070  < 2e-16 ***
## Groupe2     -1.84738    0.28694  -6.438  2.8e-06 ***
## Groupe3     -8.72289    0.33296 -26.198  < 2e-16 ***
## Age.c        1.16729    0.03201  36.461  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5552 on 20 degrees of freedom
## Multiple R-squared:  0.9943, Adjusted R-squared:  0.9935
## F-statistic:  1170 on 3 and 20 DF,  p-value: < 2.2e-16
```

The model is significant, so we can conclude that there is at least one useful predictor to describe the distribution of recovery time. Is the pre-surgical fitness level significant? Is age significant?

We pit

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

In other words, we want to know if we can eliminate the predictor `Groupe` from the model. We can use a general linear test by comparing the full model to the reduced model. The pre-surgical form is a significant predictor ($F(2, 20) = 300.11$; $p < 0.0001$).

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

```
## Analysis of Variance Table
##
## Model 1: Temps ~ Age.c
## Model 2: Temps ~ Groupe + Age.c
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     22 252.249
## 2     20   6.166  2    246.08 399.11 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The regression is significant: we reject H_0 in favor of H_1.

To test

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_1 : \beta_3 \neq 0,$$

we invoke a hypothesis that contains only one parameter, so that we can use a t test or an F test. Both tests are equivalent, with $t^2 = F$.

Here is the table of t tests related to the coefficients. Age is a significant predictor ($t(20) = 36.4608$; $p < 0.0001$).

```
summary(mod.1)$coefficients
```

```
##              Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 35.446561 0.20842377 170.069669 4.374171e-33
```

```
## Groupe2     -1.847379 0.28694289  -6.438141 2.802274e-06
## Groupe3     -8.722893 0.33296397 -26.197708 5.914908e-17
## Age.c        1.167286 0.03201483  36.460804 9.075467e-20
```

```
mod.1$df.residual
```

```
## [1] 20
```

If we use the general linear test for the significance of age instead, we obtain that it is a significant predictor ($F(1, 20) = 1329.4$; $p < 0.0001$).

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

```
## Analysis of Variance Table
##
## Model 1: Temps ~ Groupe + Age.c
## Model 2: Temps ~ Groupe
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     20   6.17
## 2     21 416.00 -1   -409.83 1329.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated model parameters are:

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

```
## (Intercept)     Groupe2     Groupe3       Age.c
##   35.446561   -1.847379   -8.722893    1.167286
```

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

```
## [1] 0.5552363
```

It is estimated that a 24 year old patient with a below average fitness level will have an average recovery time of 35.4 months, and that the mean recovery time will increase by 1.16 months per year added to the patient's age. If a patient has an average fitness level instead of a below average fitness level, then the average recovery time is reduced by 1.85 months per year added to the patient's age.

The reduction is 8.72 months per year added to the patient's age for a patient with an above-average fitness level compared to a patient with a below-average fitness level. The residual standard deviation is 0.56 months.

10. Quadratic forms

10.1 Quadratic forms I

For each of the cases below, the matrix A is the quadratic form matrix Q of the uncorrelated random variables \$Y_1,Y_2,Y_3. Furthermore, suppose that $E\{Y_i\} = 0$ and $V[Y_i] = \sigma^2 = 3$, for $i = 1, 2, 3$. For each of the quadratic forms, calculate $E\{Q\}$.

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

Answers:

We use the following result:

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

But $E\{Y\} = 0$, so that $E\{Q\} = \sigma^2\,\mathrm{tr}(A) = 3\,\mathrm{tr}(A)$.

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

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

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

10.2 Quadratic forms II

Here are quadratic forms of Y_1, Y_2, Y_3. In each case, give the matrix of the quadratic form, and compute the trace of the matrix.

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

Answers:

(a) The matrix of the quadratic form is

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

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

(b) The matrix of the quadratic form is

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

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

11. Bonferroni

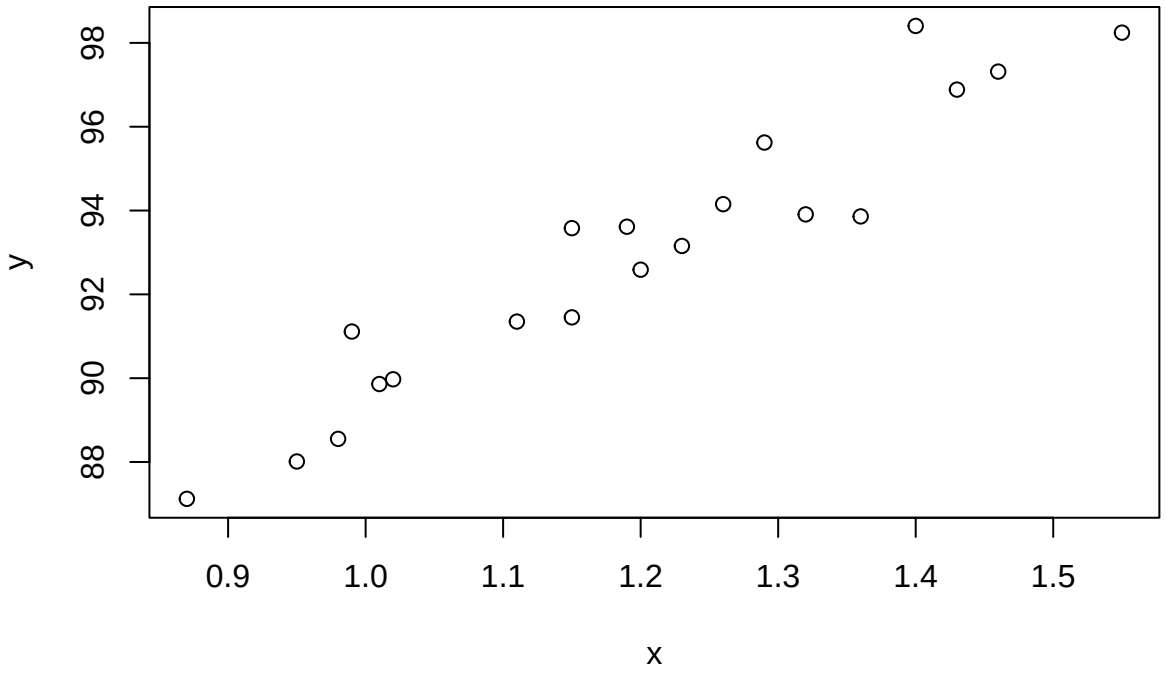
Let us imagine that the real linear relation linking Y and X is $y = 75 + 15x + \varepsilon$, where $\varepsilon \sim N(0, 1)$.

Draw samples of the response Y (of size n) for the following predictors values:

```
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
sum.X = sum(x)
sum.X.2 = sum(x*x)
```

For the first sample, the observed responses are:

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

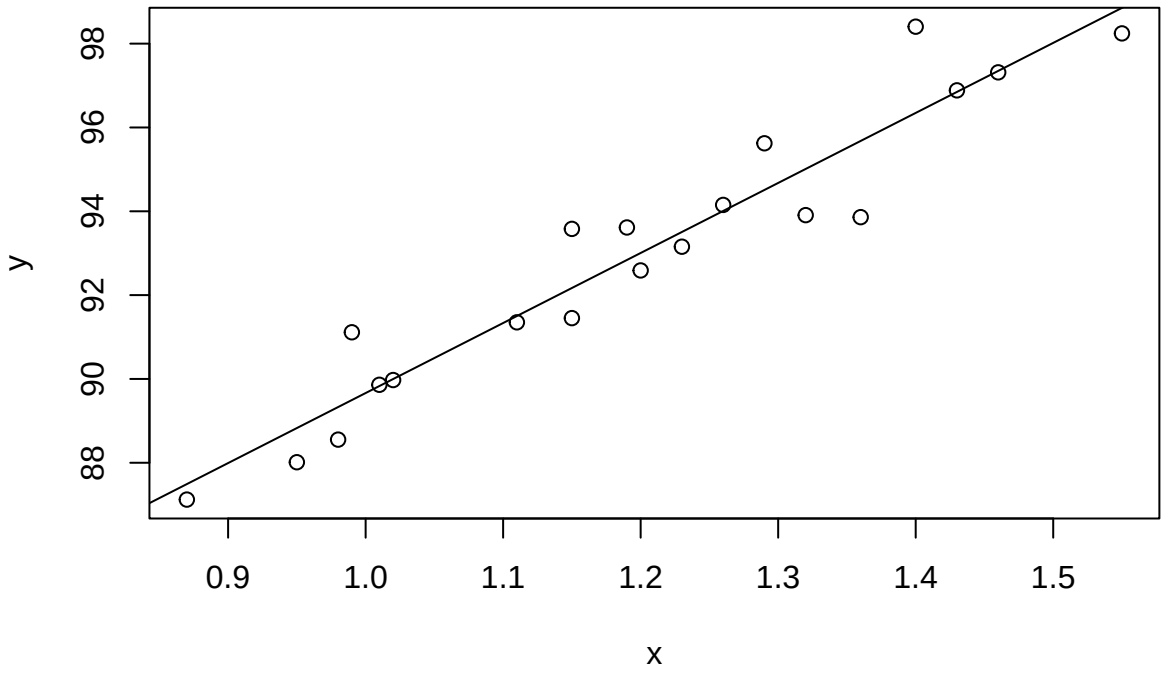


The equation of the line of best fit, in this case, is:

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

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

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



We can compute the following quantities:

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

At a confidence level of 95%, the confidence interval for the intercept β_0 is thus:

```
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
```

The real value of β_0 can indeed be found in the C.I.:

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

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

The C.I. for the slope β_1 is:

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

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

The real value of β_0 can also be found in the C.I.:

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

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

Simultaneously, (β_0, β_1) are both found in their respective C.I:

```
(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
```

Now, let's repeat the experiment $m = 10,000$ times:

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

Individually, we have:

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

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

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

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

Simultaneously, however:

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

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

We do not reach the 95% mark!

If we use the Bonferroni procedure, on the contrary:

```
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)
  sum.Y = sum(y)
  sum.X.Y = sum(x*y)
  sum.Y.2 = sum(y*y)
```

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

12. Supplementary Examples

12.1 Linearity test

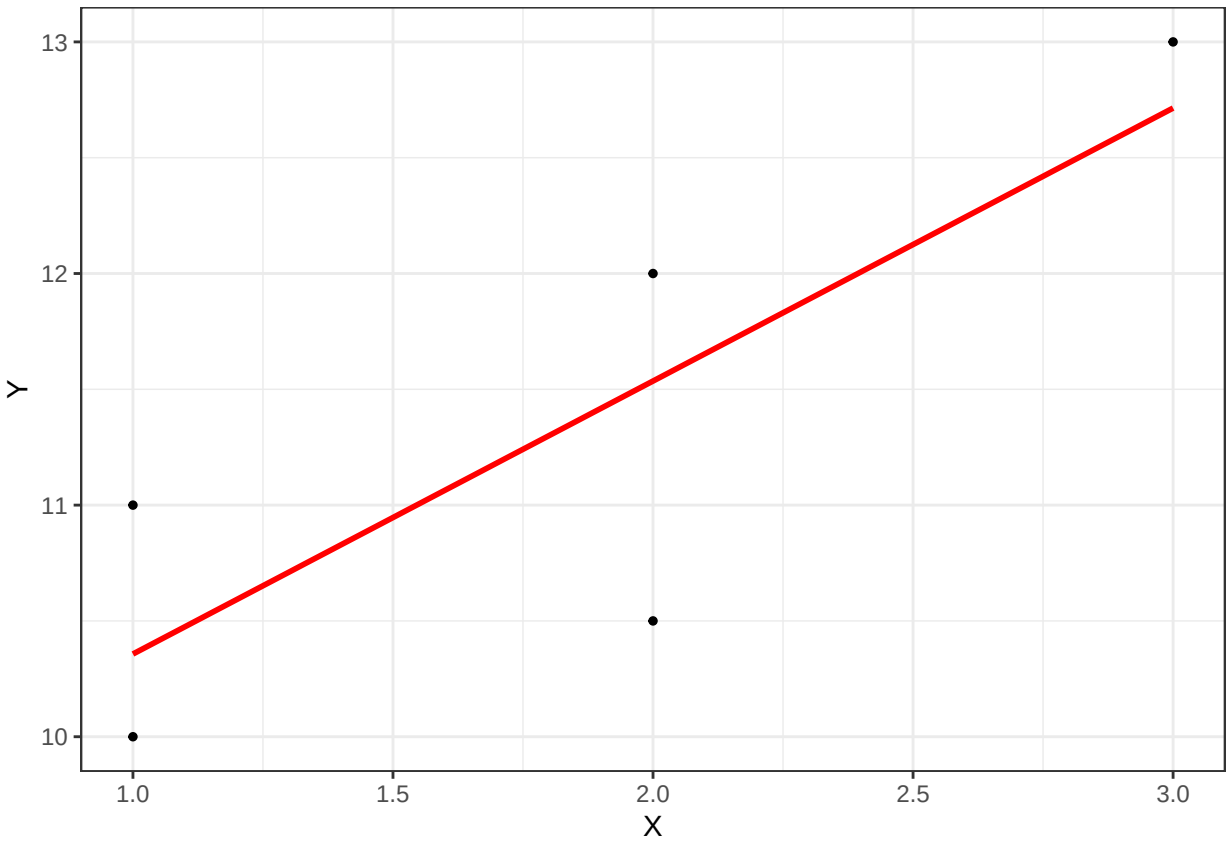
```
x=c(1,1,2,2,3)
y=c(10,11,10.5,12,13)

data = data.frame(x,y)

mod = lm(y~x, data=data)
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##       1       2       3       4       5
## -0.3571  0.6429 -1.0357  0.4643  0.2857
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.1786     0.9297   9.872  0.00221 **
## x             1.1786     0.4769   2.471  0.08997 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7981 on 3 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.5608
## F-statistic: 6.107 on 1 and 3 DF,  p-value: 0.08997
```

```
ggplot2::ggplot(data,ggplot2::aes(x=x,y=y)) +
  ggplot2::geom_point(size=1) +
```

```
  ggplot2::xlab("X") +
  ggplot2::ylab("Y") +
  ggplot2::geom_smooth(color="red", method="lm", se=FALSE) +
  ggplot2::theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
n=5
p=2
c=3
n1 = 2
n2 = 2
n3 = 1
SST = (n-1)*var(y)
SSR = as.double(mod[[1]][2]^2*(n-1)*var(x))
SSE = SST - SSR
SSPE = (n1-1)*var(data[data$x == 1,2]) + (n2-1)*var(data[data$x == 2,2]) # + (n3-1)*var(data[data$x ==
SSLF = SSE - SSPE
Fstar = (SSLF/(c-p))/(SSPE/(n-c))
qf(0.95,c-p,n-c)
```

```
## [1] 18.51282
```

```
Fstar < qf(0.95,c-p,n-c)
```

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

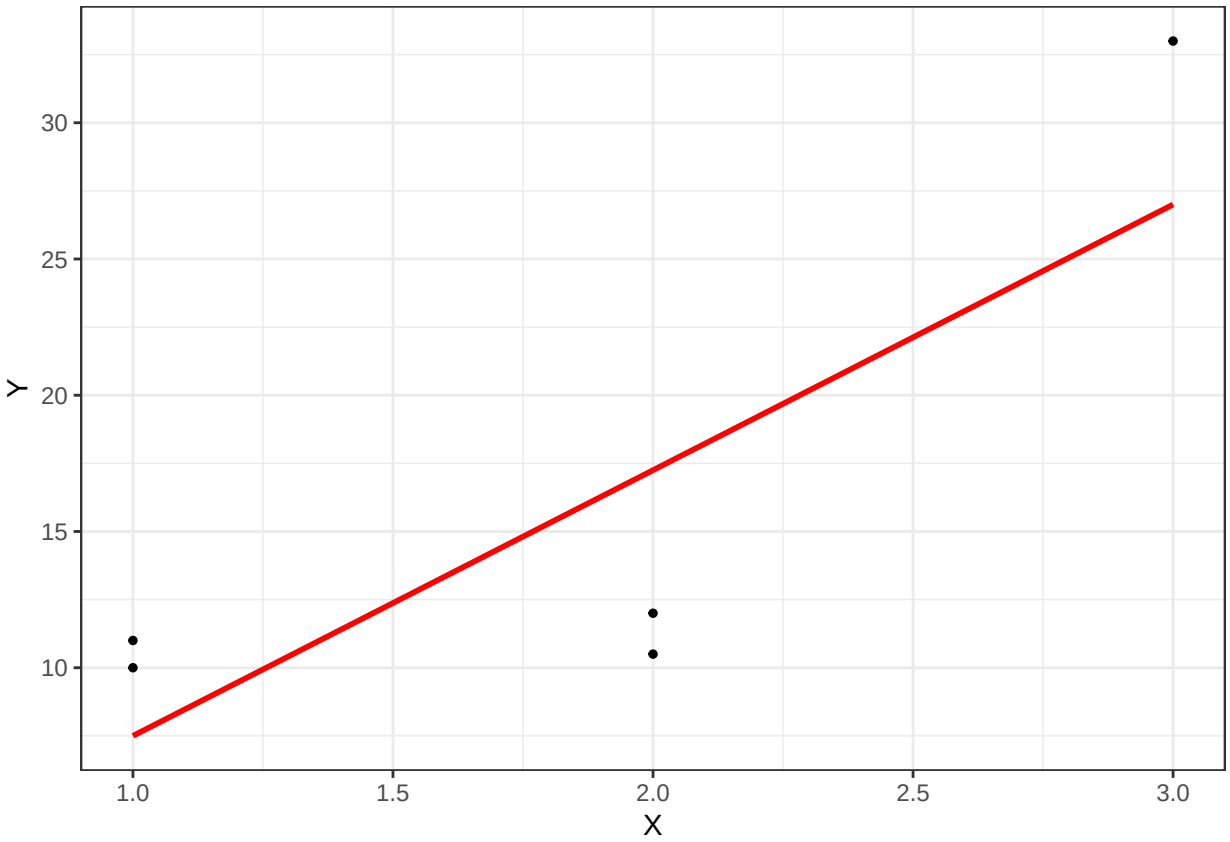
```
x=c(1,1,2,2,3)
y=c(10,11,10.5,12,33)
```

```r
data = data.frame(x,y)

mod = lm(y~x, data=data)
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     1     2     3     4     5
##  2.50  3.50 -6.75 -5.25  6.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.250      7.598  -0.296   0.7865
## x              9.750      3.898   2.501   0.0876 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.522 on 3 degrees of freedom
## Multiple R-squared:  0.6759, Adjusted R-squared:  0.5679
## F-statistic: 6.257 on 1 and 3 DF,  p-value: 0.0876
```

```r
ggplot2::ggplot(data,ggplot2::aes(x=x,y=y)) +
  ggplot2::geom_point(size=1) +
  ggplot2::xlab("X") +
  ggplot2::ylab("Y") +
  ggplot2::geom_smooth(color="red", method="lm", se=FALSE) +
  ggplot2::theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
n=5
p=2
c=3
n1 = 2
n2 = 2
n3 = 1
SST = (n-1)*var(y)
SSR = as.double(mod[[1]][2]^2*(n-1)*var(x))
SSE = SST - SSR
SSPE = (n1-1)*var(data[data$x == 1,2]) + (n2-1)*var(data[data$x == 2,2]) # + (n3-1)*var(data[data$x == 
SSLF = SSE - SSPE
Fstar = (SSLF/(c-p))/(SSPE/(n-c))
qf(0.95,c-p,n-c)
```

```
## [1] 18.51282
```

```
Fstar < qf(0.95,c-p,n-c)
```

```
## [1] FALSE
```

12.2 Polynomial regression

```
X=c(1,1,2,4,3,6)
X2 = X^2
Xm = X-mean(X)
X2m = (X-mean(X))^2
Y=c(0.8,1.3,4.1,15.3,8.8,36)
data=data.frame(X,X2,Xm,X2m,Y)
```

```r
par(mfrow = c(1,2))

summary(lm(Y~X, data=data))
```

```
##
## Call:
## lm(formula = Y ~ X, data = data)
##
## Residuals:
##      1      2      3      4      5      6
##  2.020  2.520 -1.373 -3.558 -3.365  3.756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.913      2.734  -2.895  0.04435 *
## X              6.693      0.818   8.182  0.00122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.55 on 4 degrees of freedom
## Multiple R-squared:  0.9436, Adjusted R-squared:  0.9295
## F-statistic: 66.94 on 1 and 4 DF,  p-value: 0.001215
```

```r
plot(data$X,data$Y)
abline(lm(Y~X, data=data), col='red')

summary((mod1<-lm(Y~X+X2, data=data)))
```

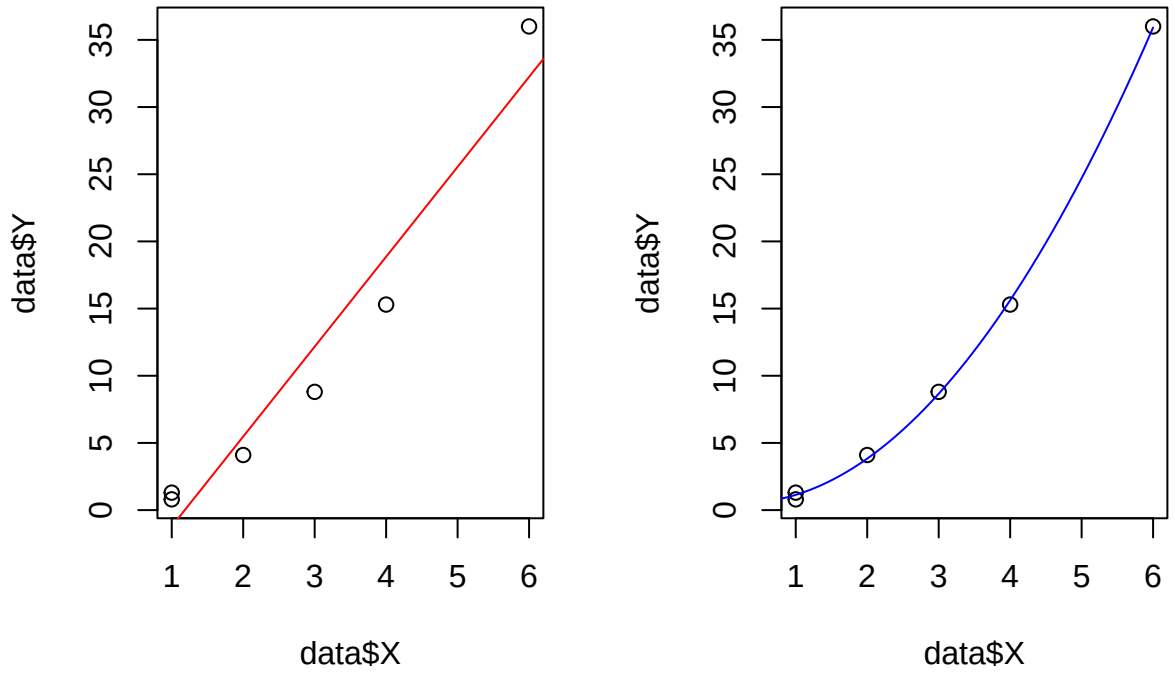
```
##
## Call:
## lm(formula = Y ~ X + X2, data = data)
##
## Residuals:
##        1        2        3        4        5        6
## -0.33510  0.16490  0.26683 -0.31731  0.13942  0.08125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.56635    0.47768   1.186 0.321128
## X           -0.49591    0.34935  -1.420 0.250809
## X2           1.06466    0.05046  21.101 0.000233 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3354 on 3 degrees of freedom
## Multiple R-squared:  0.9996, Adjusted R-squared:  0.9994
## F-statistic:  3973 on 2 and 3 DF,  p-value: 7.331e-06
```

```r
modX = lm(X~X2, data=data)
VIF1 = 1/(1-summary(modX)$r.squared)


summary((mod2<-lm(Y~Xm+X2m, data=data)))
```

```
##
```

```
## Call:
## lm(formula = Y ~ Xm + X2m, data = data)
##
## Residuals:
##        1        2        3        4        5        6
## -0.33510  0.16490  0.26683 -0.31731  0.13942  0.08125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.70814    0.20935   36.82 4.41e-05 ***
## Xm           5.53718    0.09472   58.46 1.10e-05 ***
## X2m          1.06466    0.05046   21.10 0.000233 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3354 on 3 degrees of freedom
## Multiple R-squared:  0.9996, Adjusted R-squared:  0.9994
## F-statistic:  3973 on 2 and 3 DF,  p-value: 7.331e-06
```

```
t <- seq(0, 6, 0.1)
y <- predict(mod1,list(X=t, X2=t^2))
plot(data$X, data$Y)

#add predicted lines based on quadratic regression model
lines(t, y, col='blue')
```

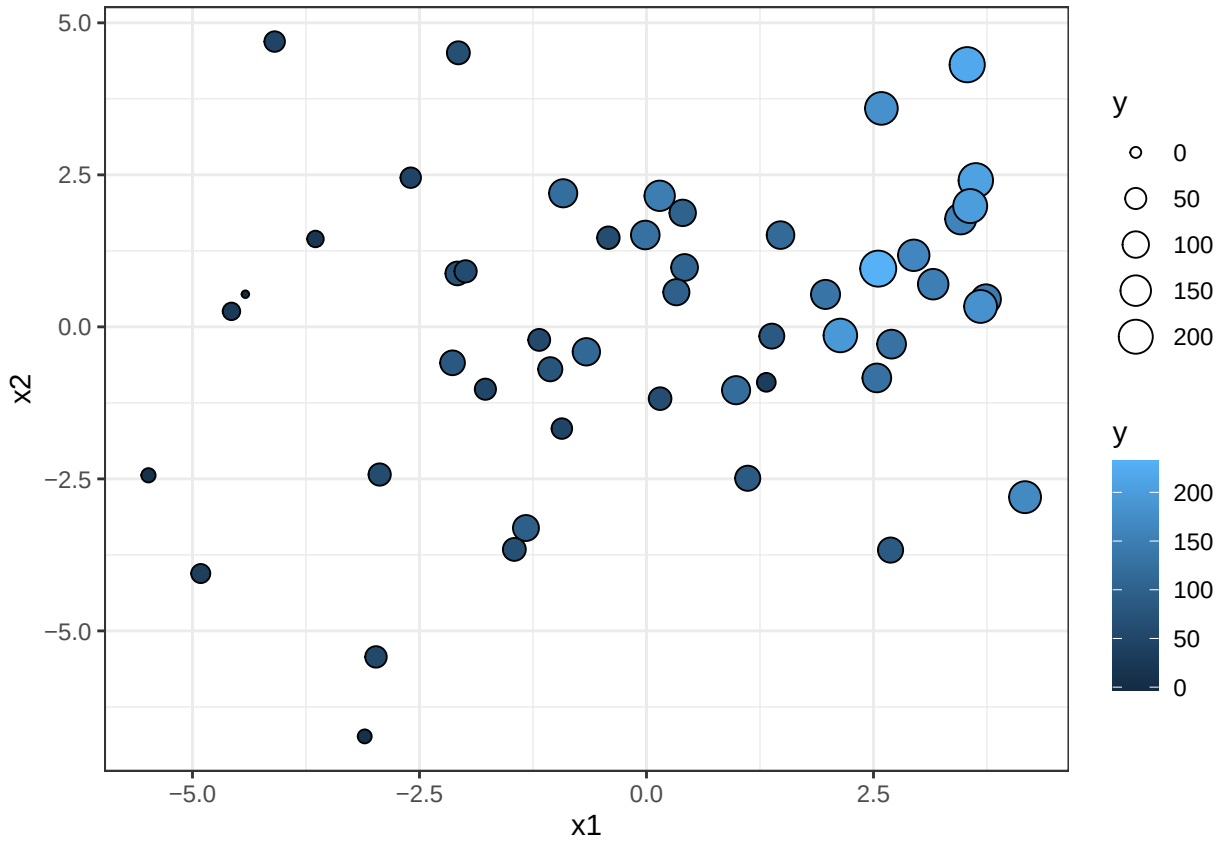



```
modXm = lm(Xm~X2m, data=data)
VIF1m = 1/(1-summary(modXm)$r.squared)
```

12.3 Interaction terms

```
x1 <- runif(50, 0, 10)
x2 <- rnorm(50, 10, 3)
```

```
modmat <- model.matrix(~x1 * x2, data.frame(x1 = x1, x2 = x2))
coeff <- c(1, 2, -1, 1.5)
y <- rnorm(50, mean = modmat %*% coeff, sd = 25)
dat <- data.frame(y = y, x1 = x1, x2 = x2)
dat2 = dat
dat2[,c(2:3)] <- scale(dat[,c(2:3)], scale=FALSE)

library(ggplot2)
ggplot(dat2,aes(x=x1,y=x2,fill=y,size=y)) + geom_point(pch=21) + theme_bw()
```

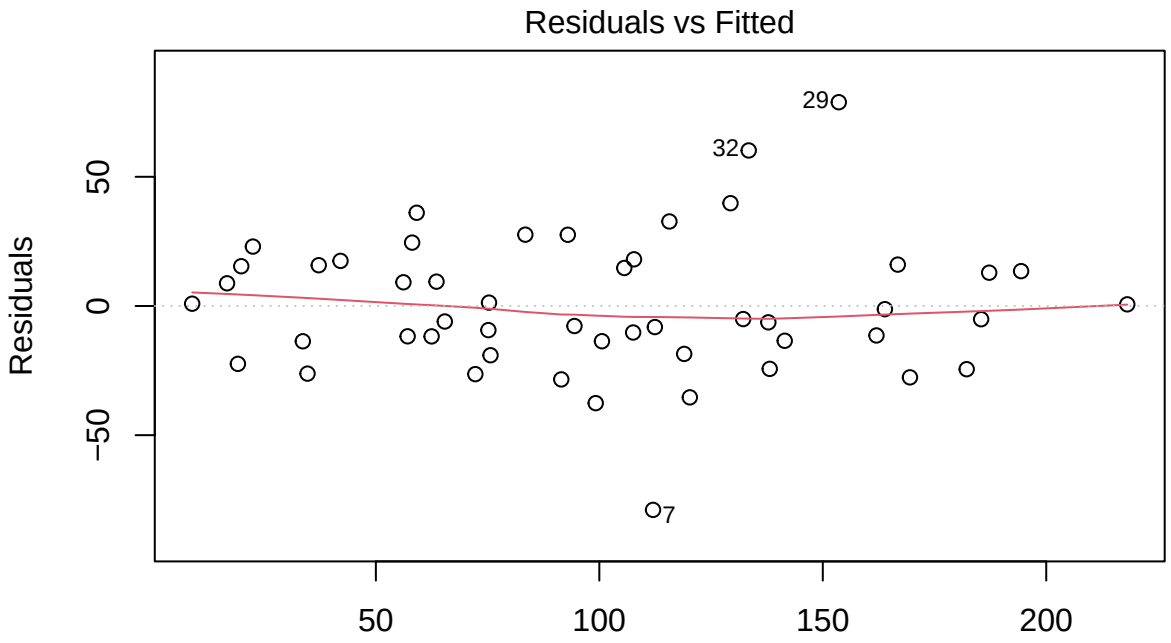


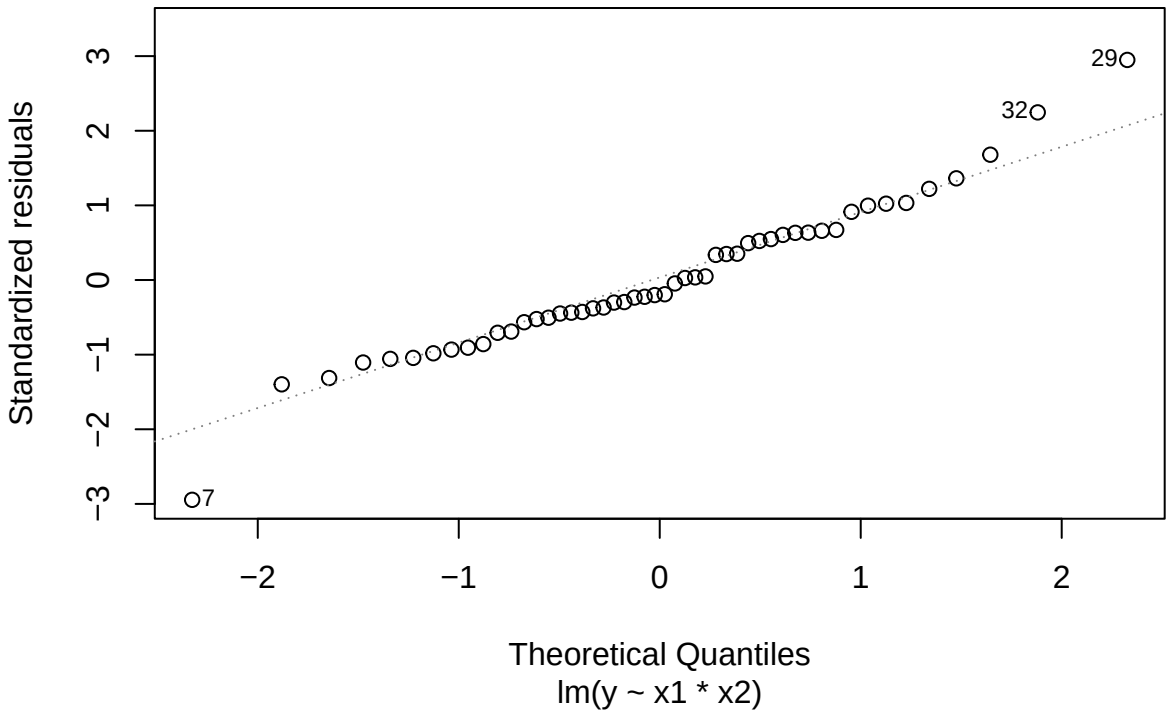

```
summary(lm(y ~ x1 * x2, data=dat2))
```

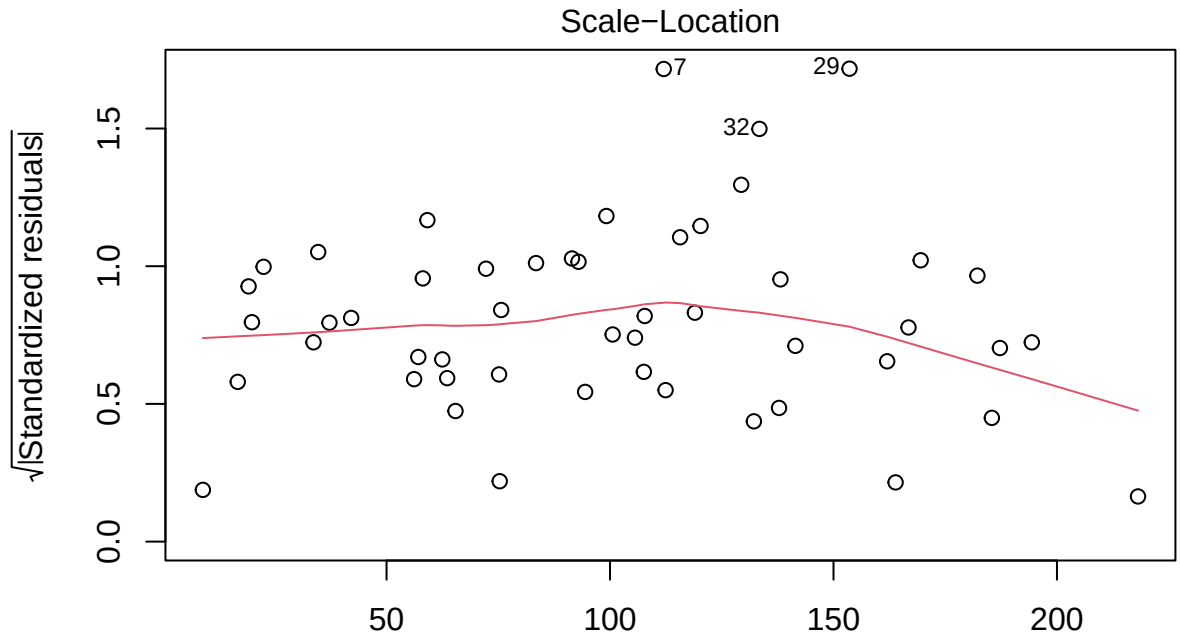
```
##
## Call:
## lm(formula = y ~ x1 * x2, data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.921 -13.657  -5.093  15.678  78.872
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  97.4472     3.9601  24.608  < 2e-16 ***
## x1           17.5792     1.4979  11.736 1.97e-15 ***
## x2            7.0033     1.7950   3.902  0.00031 ***
## x1:x2         1.8649     0.5814   3.208  0.00244 **
## ---
```

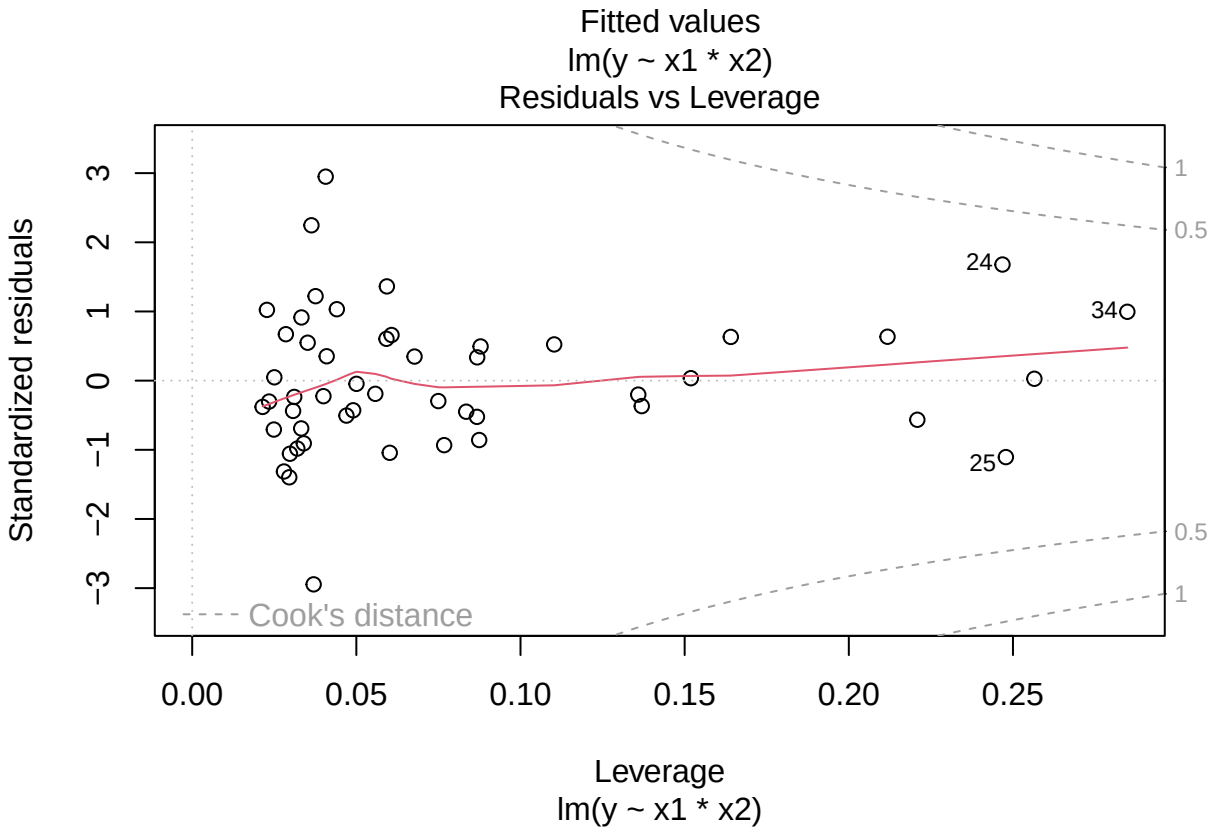
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.31 on 46 degrees of freedom
## Multiple R-squared:  0.8009, Adjusted R-squared:  0.7879
## F-statistic: 61.69 on 3 and 46 DF,  p-value: 3.729e-16
```

```
plot(lm(y ~ x1 * x2, data=dat2))
```







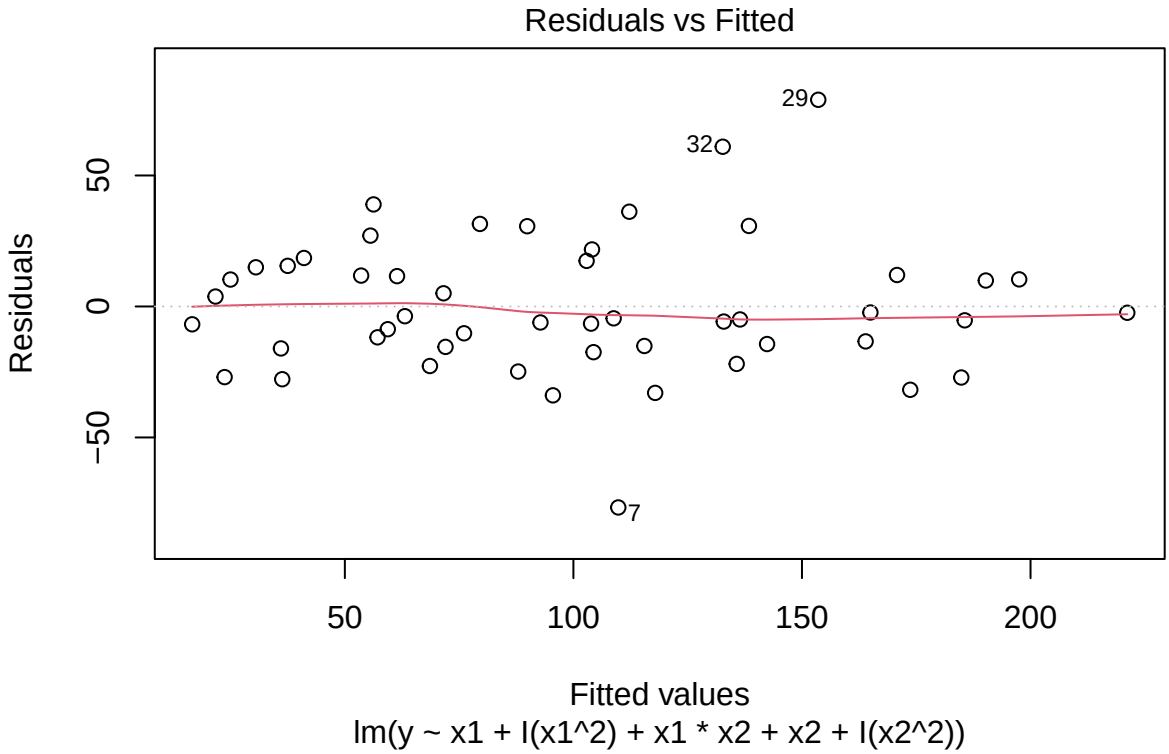


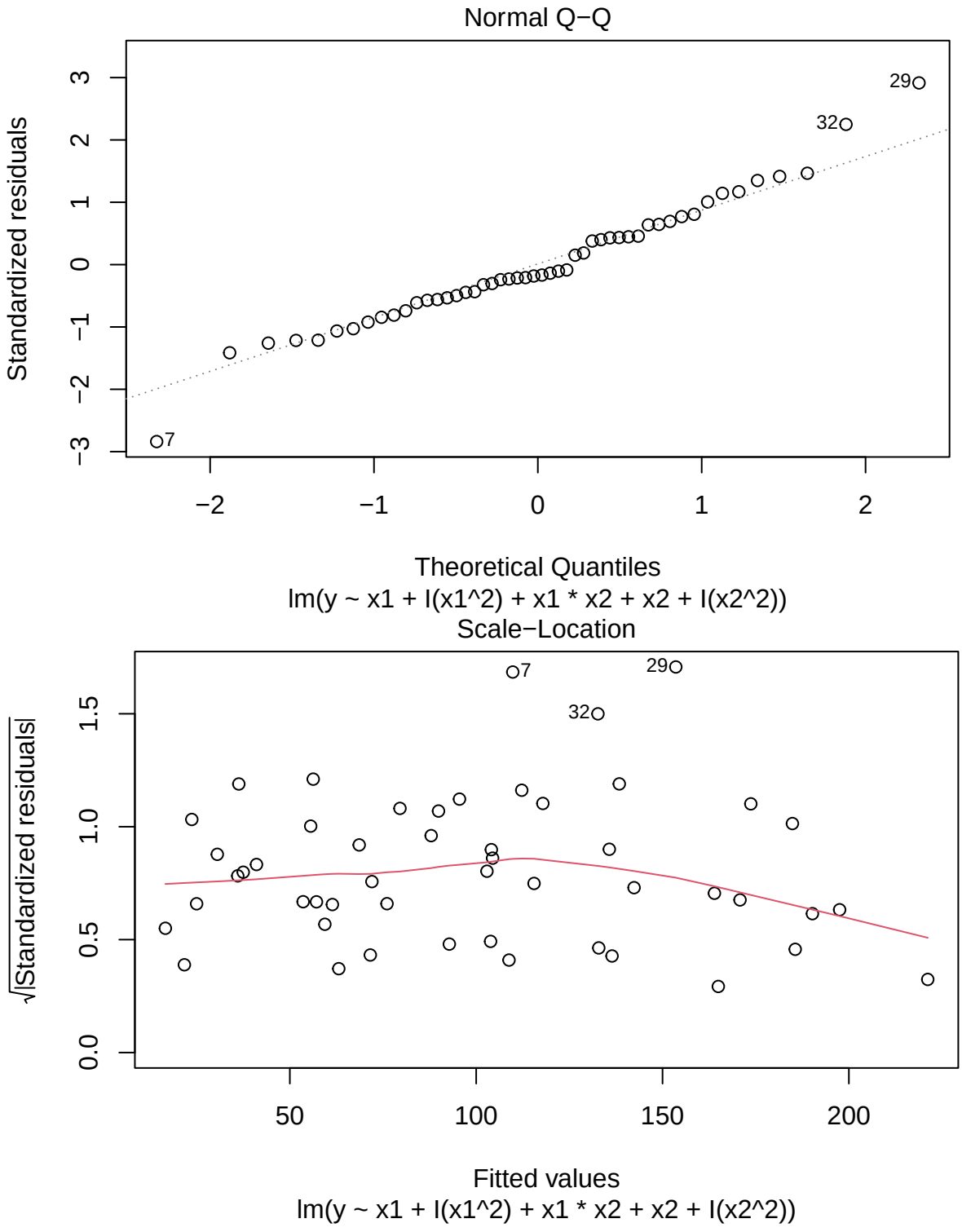

```
summary(lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2), data=dat2))
```

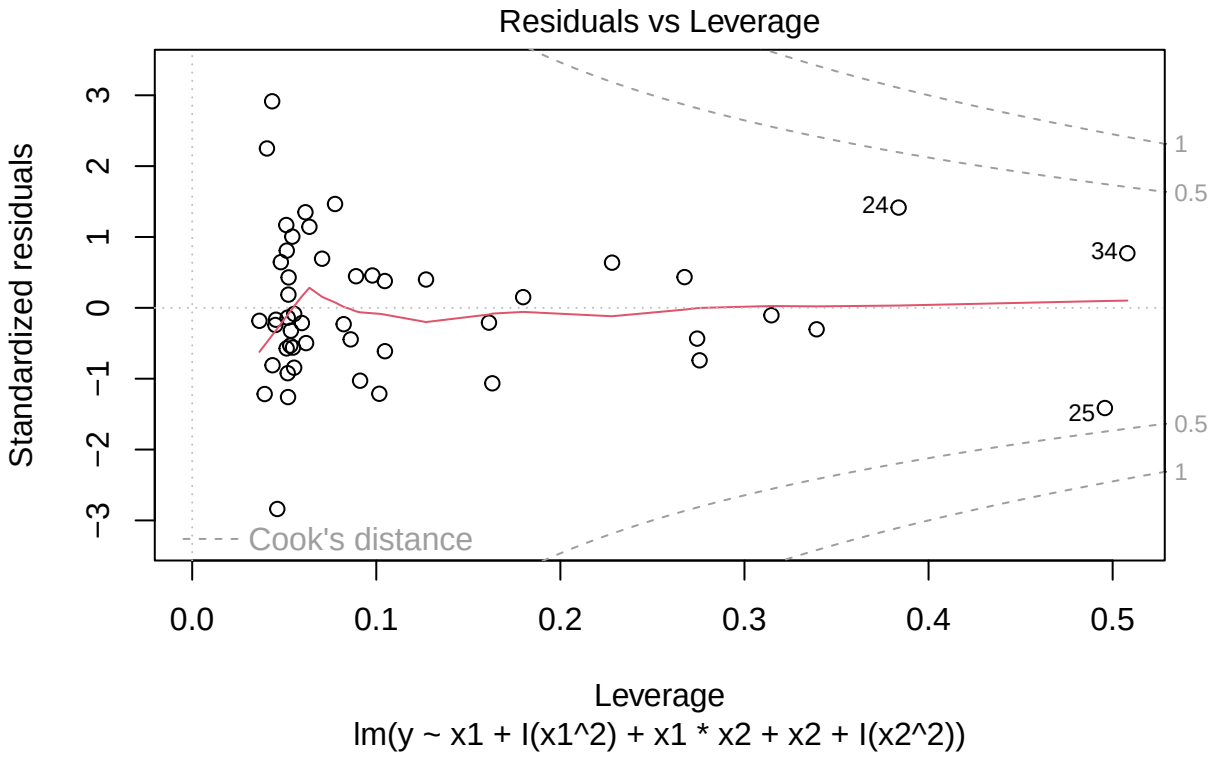
```
##
## Call:
## lm(formula = y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2), data = dat2)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -76.719 -15.351  -4.749  14.215  78.890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  93.5831     6.0467  15.477  < 2e-16 ***
## x1           17.9469     1.5930  11.266 1.49e-14 ***
## I(x1^2)       0.4959     0.5946   0.834 0.408720
## x2            6.9600     1.8479   3.766 0.000488 ***
## I(x2^2)       0.1054     0.5110   0.206 0.837598
## x1:x2         1.7100     0.6290   2.719 0.009347 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.67 on 44 degrees of freedom
## Multiple R-squared:  0.8044, Adjusted R-squared:  0.7822
## F-statistic: 36.19 on 5 and 44 DF,  p-value: 1.585e-14
```

```
plot(lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2), data=dat2))
```

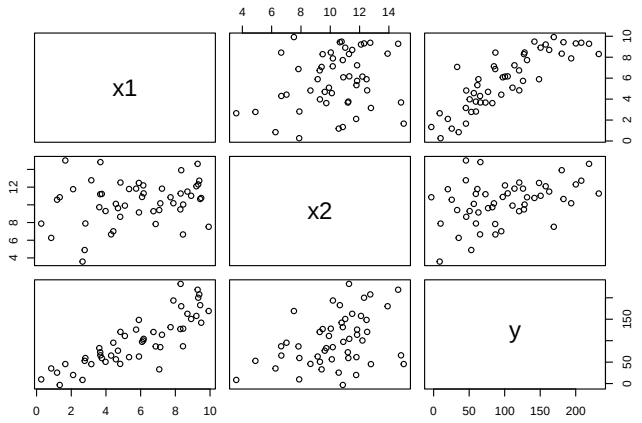



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2))
Scale−Location
√|Standardized residuals|
Fitted values
lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2))



12.4 Weighted least squares

```
n = nrow(x1)
p = 2
plot(data.frame(x1,x2,y))
```

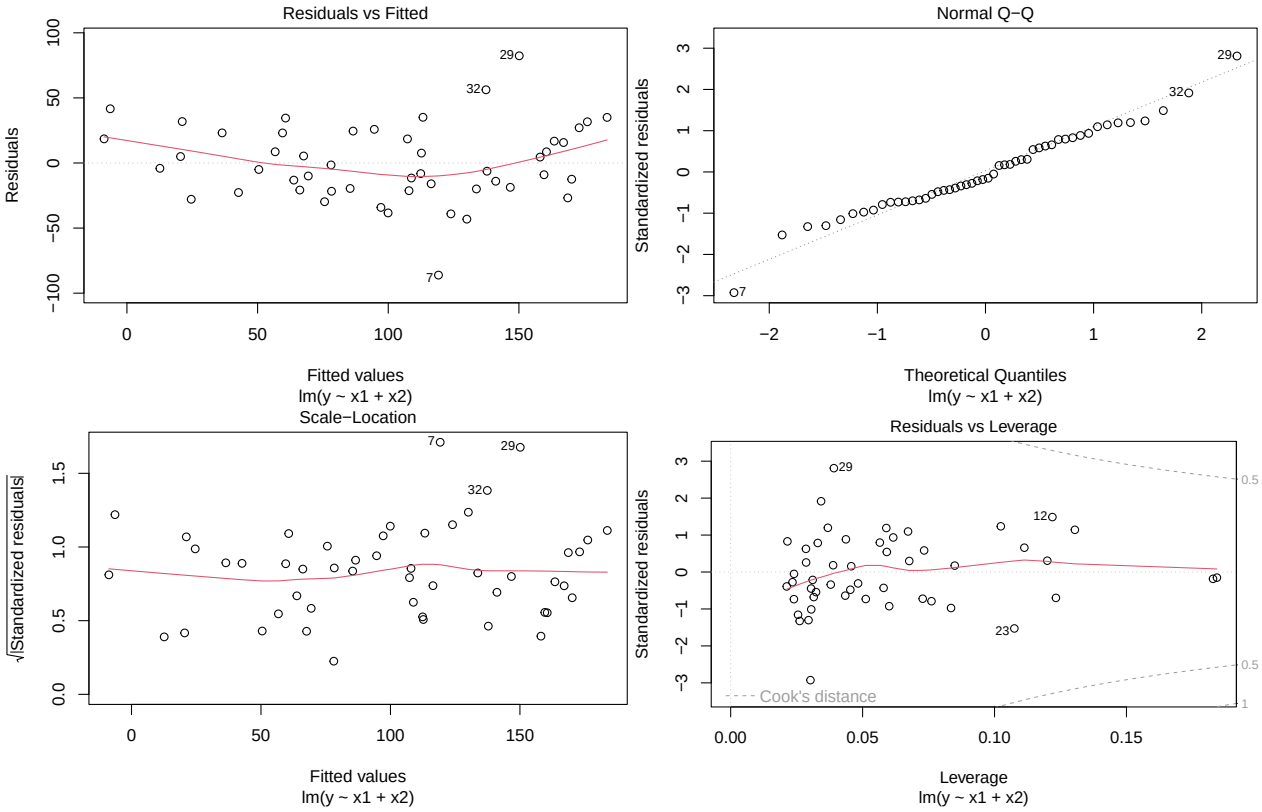



```
mod = lm(y ~ x1 + x2)
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -86.122 -19.858  -4.545  21.948  82.331
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -51.706     19.388  -2.667   0.0105 *
## x1            17.716      1.639  10.812 2.43e-14 ***
## x2             4.850      1.822   2.662   0.0106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.88 on 47 degrees of freedom
## Multiple R-squared:  0.7564, Adjusted R-squared:  0.746
## F-statistic: 72.97 on 2 and 47 DF,  p-value: 3.864e-15
```

```
plot(mod)
```




```
(MSE = sum(mod$residuals^2)/(n-p))
```

```
## numeric(0)
```

```
poids <- 1 / lm(abs(mod$residuals) ~ x1 + x2)$fitted.values^2
mod.wls <- lm(y ~ x1 + x2, weights=poids)

(MSE.w = sum(mod.wls$residuals^2)/(n-p))
```

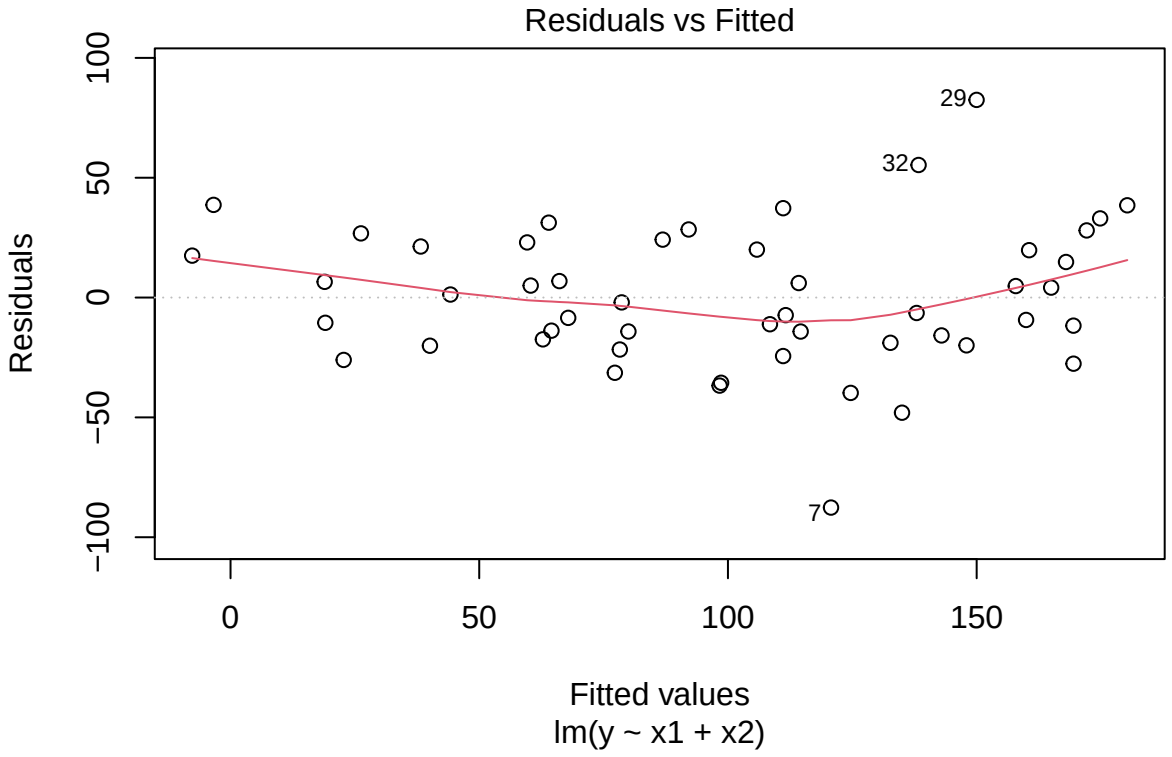
```
## numeric(0)
```

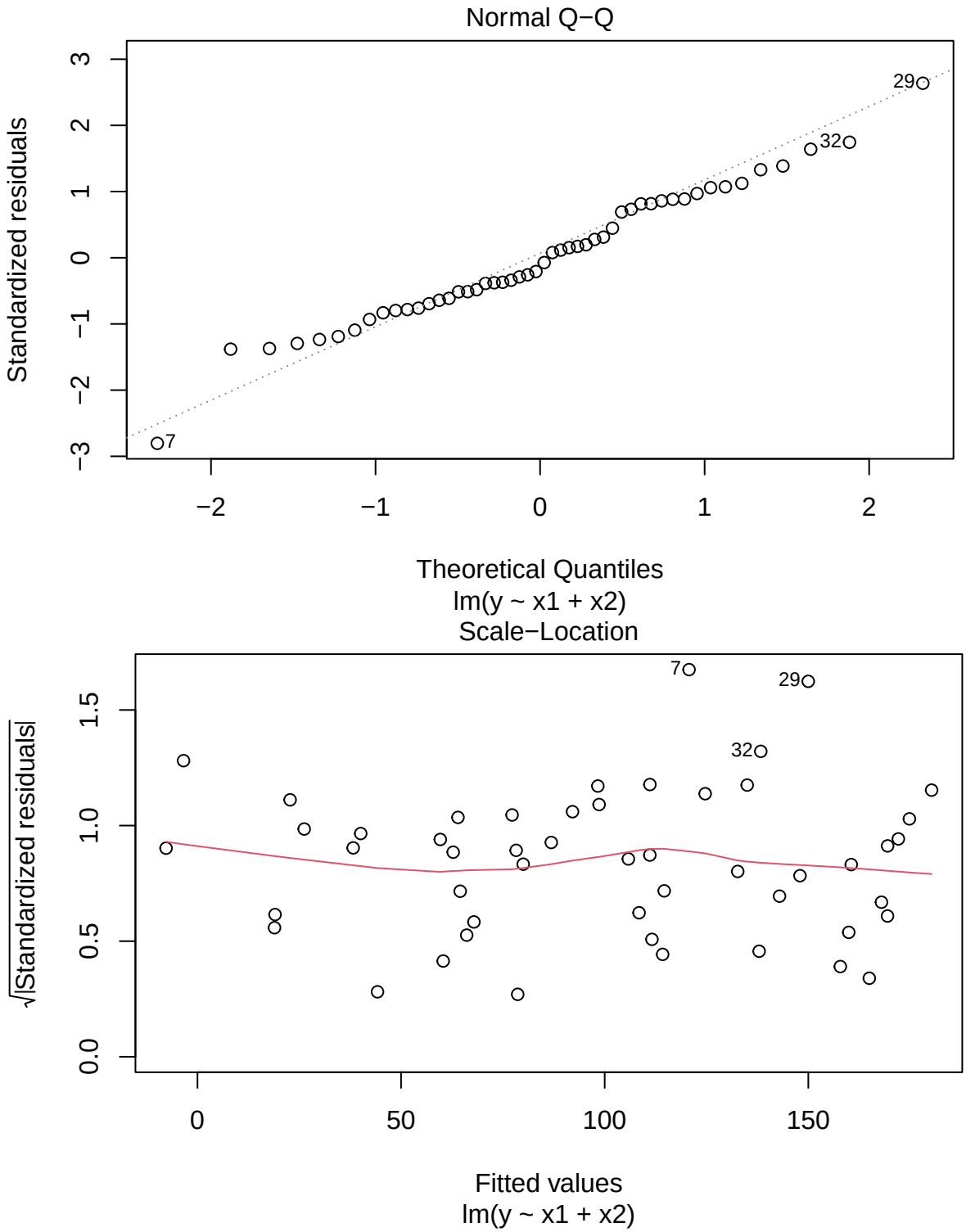
```
summary(mod.wls)
```

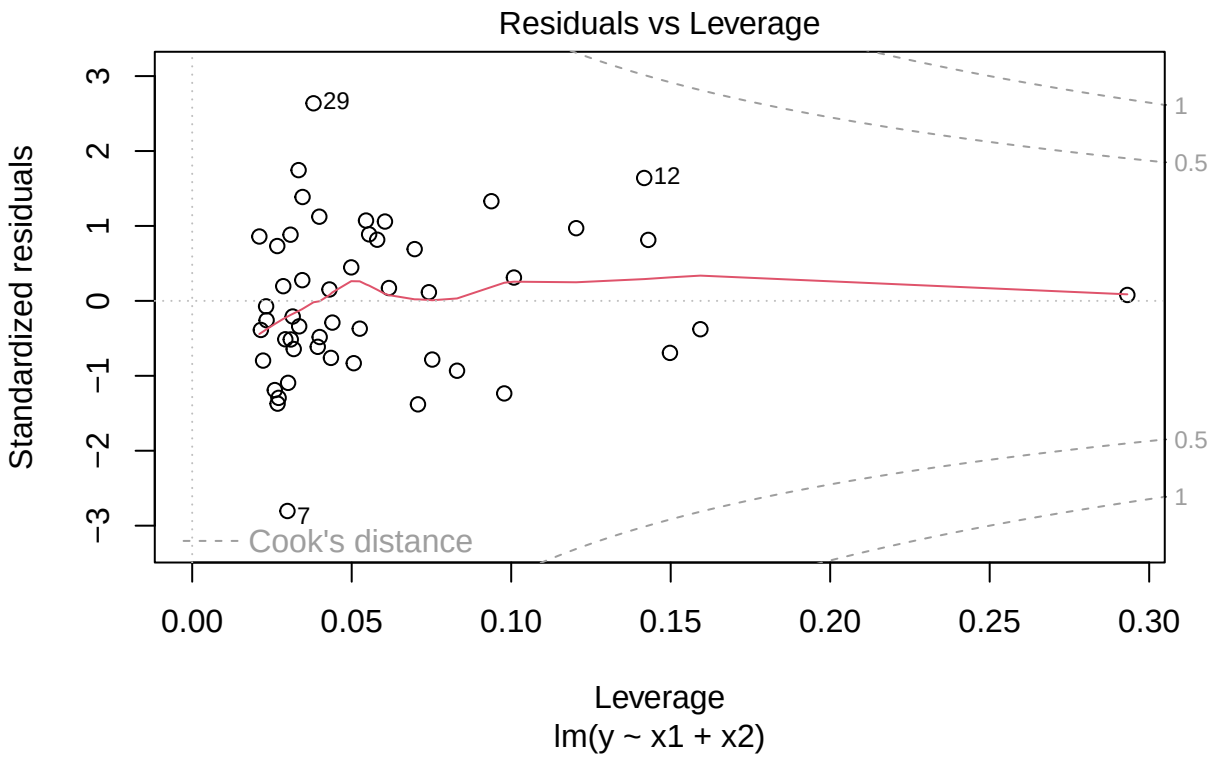
```
##
## Call:
## lm(formula = y ~ x1 + x2, weights = poids)
```

```
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4491 -0.7966 -0.1730  0.9773  3.2297
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -42.216     18.235  -2.315    0.025 *
## x1            18.027      1.493  12.073 5.22e-16 ***
## x2             3.767      1.638   2.299    0.026 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.249 on 47 degrees of freedom
## Multiple R-squared:  0.7767, Adjusted R-squared:  0.7672
## F-statistic: 81.76 on 2 and 47 DF,  p-value: 4.978e-16
```

```
plot(mod.wls)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x1 + x2)
Scale−Location
√|Standardized residuals|
Fitted values
lm(y ~ x1 + x2)
7
29
32


Residuals vs Leverage
Standardized residuals
Leverage
lm(y ~ x1 + x2)
Cook's distance
29
12
7
1
0.5
0.5
1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−3
−2
−1
0
1
2
3