Chapter 7: Linear regression

Contents

1 Simple linear regression

1.1 Fitting a line

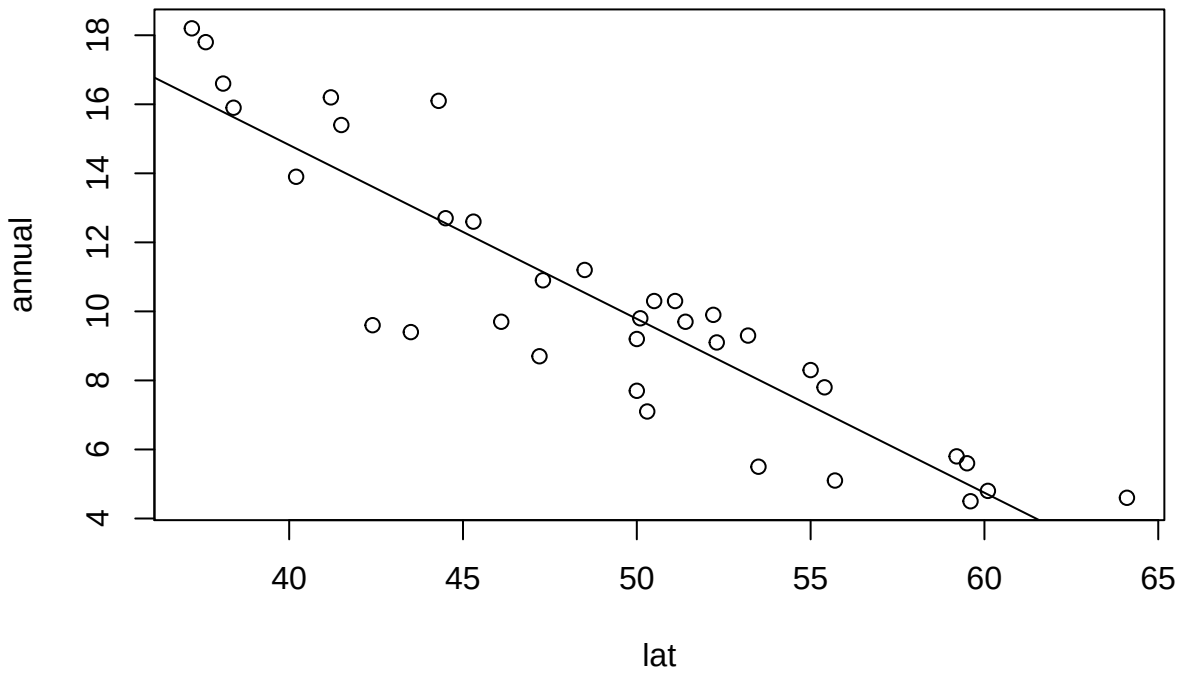
Line fitting is the process of constructing a straight line that has the best fit to a series of data points.

1.1.1 Illustrative example

Example *Temperature*
Import the data set *temp_warm.txt* as `temperature`.

```
annual <- temperature$annual
lat <- temperature$Latitude
long <- temperature$Longitude
```

From the previous chapter, we know that there is a linear relationship between *annual* and *Latitude*. We now will try to model this linear relationship.



We now try to **fit a linear model** for the variables *annual* and *Latitude*. This means, we have to find estimates for two parameters α and β such that the observed data points (x_i, y_i) are as close as possible to the fitted line of
$annual = \alpha + \beta \cdot Latitude$.

1.1.2 In general

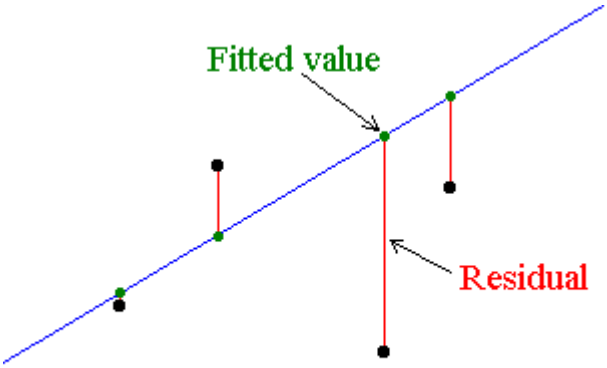
In general, a population model is described by

$$Y = \alpha + \beta x$$

With:

- Y: Dependent variable
- α: Intercept
- β: Slope parameter
- x: Independent variable

An appropriate measure of the quality of the fit of our model is to look at the **residuals**, or the differences between the observed values for the response and the predicted values. The straight line represents the predicted values. The vertical distance from the straight line to the observed data value is the residual. Hence, if the actual response value is less than the value predicted by the straight line, the residual is negative. Values of the residuals near zero indicate a good fit.



In spirit of the sample variance, an appropriate measure of the quality of the fit is the **sum of squares for the residuals**, defined by

$$\begin{aligned} Q &= \sum_i (y_i - \hat{y}_i)^2 \\ &= \sum_i (y_i - (a + b \cdot x_i))^2 \end{aligned} \tag{1}$$

with $\hat{y}_i$ the fitted value of y_i, and a and b estimates for respectively the parameters α and β.

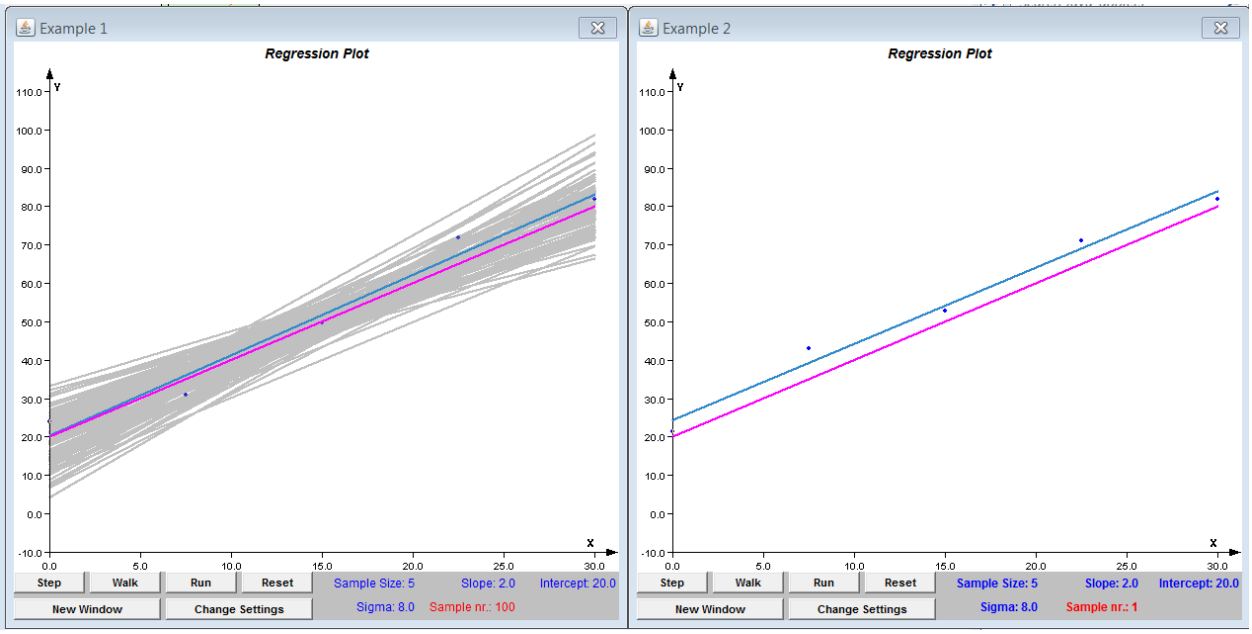
We estimate the parameters α and β by the method of least squares, which minimizes the sum of squares for the residuals.

The solution of this least squares problem provides the **least squares estimators** a and b:
$a = \overline{y} - b \cdot \overline{x}$
$b = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$

The estimates a and b are the least squares estimators for α and β.

The line $Y = \alpha + \beta x$ is the **population regression line**.
The line $\hat{y} = a + bx$ is the **estimated regression line** where a and b are estimates for α and β.

Take a look at http://lstat.kuleuven.be/java/index.htm. Select *REGRESSION* $\rightarrow$ *Simple Linear Regression* and run the applet to get following figures:



1.2 In R

Example *Temperature* in R

Using the least squares estimators, we can predict the annual temperature (variable *annual*) for a city with a given value of *Latitude*:
$\hat{annual} = a + b \cdot Latitude$

```
res.lm1 <- lm(annual ~ lat)
summary(res.lm1)
```

```
##
## Call:
## lm(formula = annual ~ lat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0116 -0.7017  0.4748  1.0526  3.4454
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.96769    2.06934   16.90  < 2e-16 ***
## lat         -0.50368    0.04177  -12.06 1.23e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.73 on 33 degrees of freedom
## Multiple R-squared:  0.815,  Adjusted R-squared:  0.8094
## F-statistic: 145.4 on 1 and 33 DF,  p-value: 1.226e-13
```

The estimated regression model for annual temperature can be written as
$\hat{annual} = 35 - 0.5 \cdot Latitude$

- For a city with $Latitude = 0$ (on the equator), this model predicts an annual temperature of 35°.
- Since slope < 0, there is a negative linear relationship between *Latitude* and *annual*. The larger the value for *Latitude*, the lower the annual temperature.

- For an increase of 1° in *Latitude*, the regression model predicts an average decrease in annual temperature of 0.5°.

Remark:
So far, this calculation was done without any statistical assumptions.

Assumptions in regression analysis are formulated on the residuals. For any given data set, **residuals** can be computed as:
$r_i = y_i - \hat{y}_i$

1.3 The statistical model

In real cases, the data (x_i, y_i) (with $i = 1, 2, ..., n$) will never be on one straight line.

This means that we will never have the situation
$\forall i : y_i = \alpha + \beta \cdot x_i$

Our observations will satisfy:

$\forall i : y_i = \alpha + \beta \cdot x_i + \varepsilon_i$ **where ε_i is the random error.**

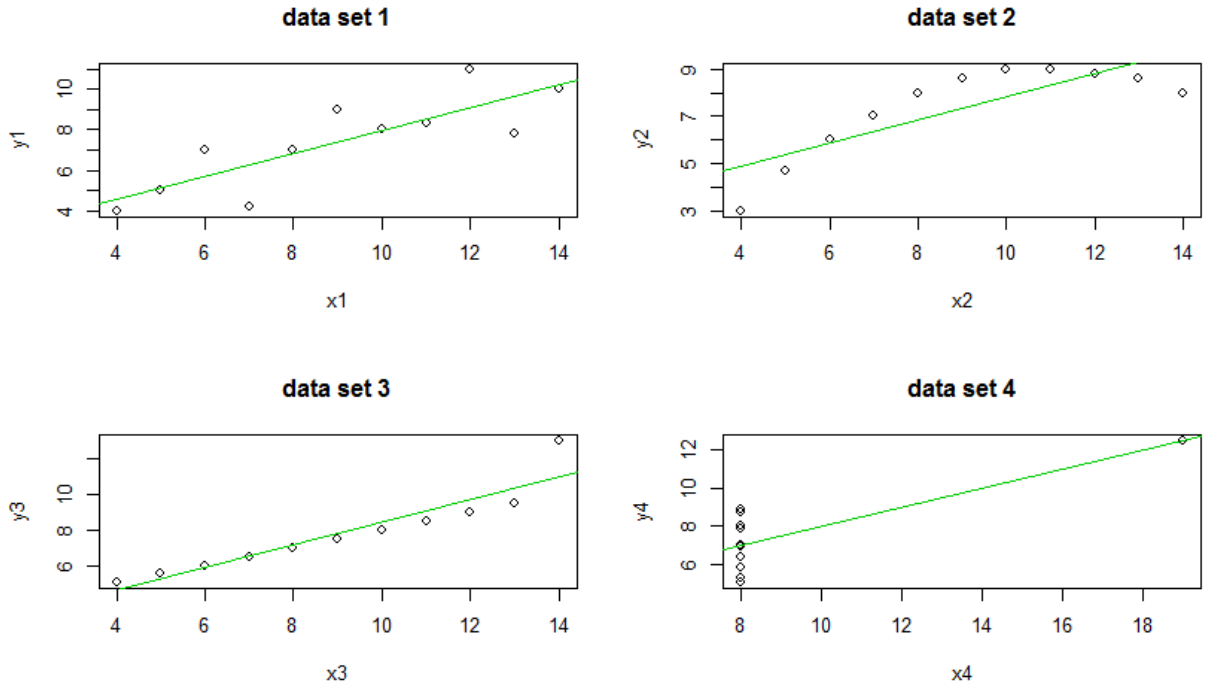
We assume that the random errors are independent normally distributed with mean 0 and variance σ^2, i.e., $\varepsilon_i \sim N(0, \sigma^2)$

Remark:

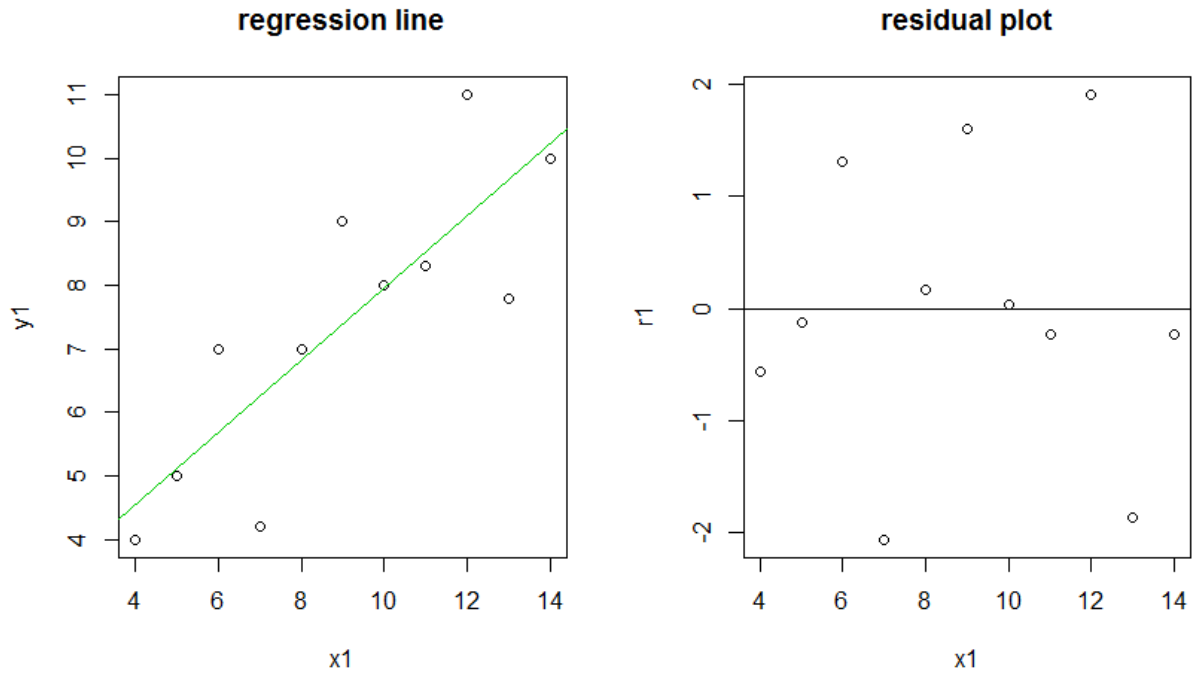
1. These ε_i are the population values of the residuals r_i.
2. The **regression model assumes**:
   a) Linearity: there is a linear relationship between x and Y
   b) Normality of the residuals
   c) Constant variance

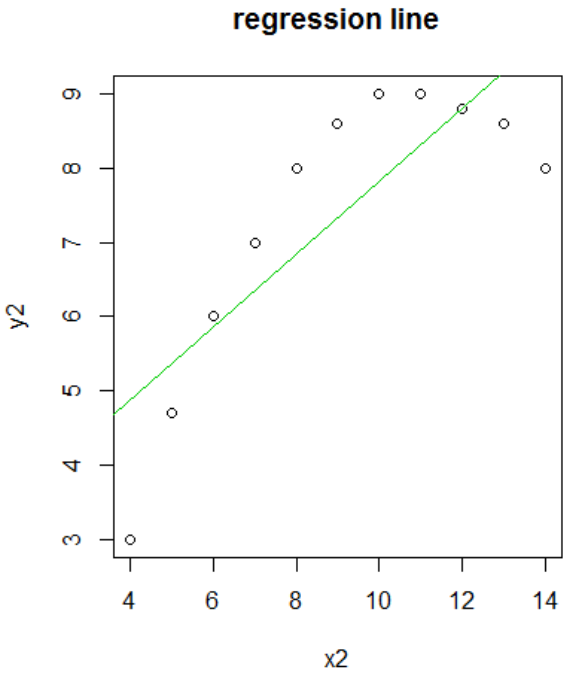
Example

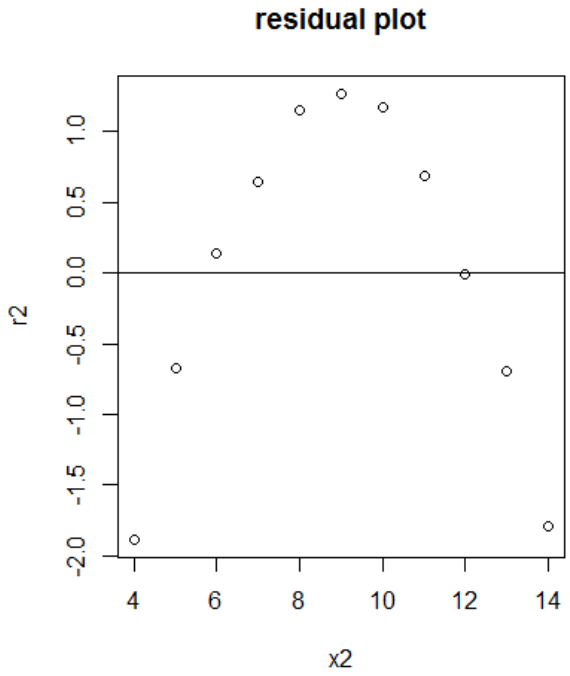
Consider the next 4 data sets with their regression line. Whether there exists a useful fit will depend on the behavior of the residuals $r_i = y_i - \hat{y}_i$

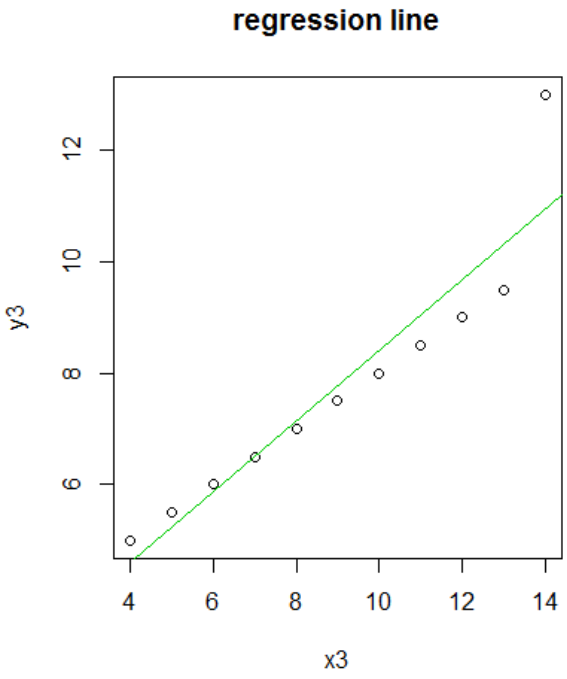


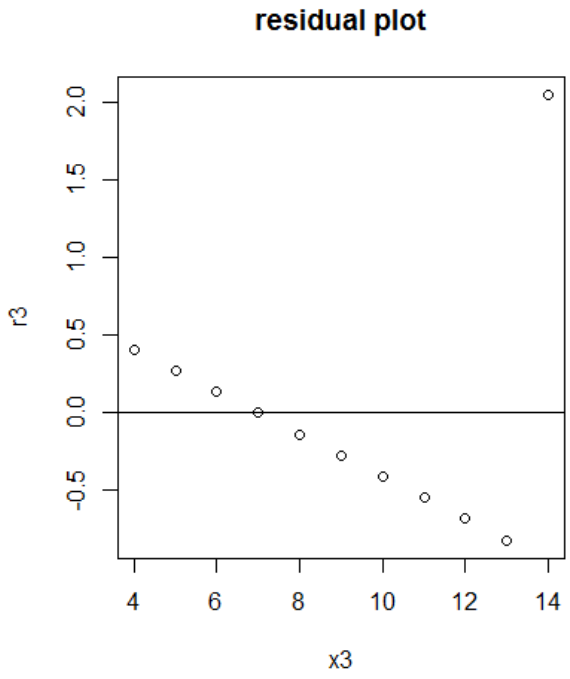
Task:
Sketch the corresponding residual plot for every data set.

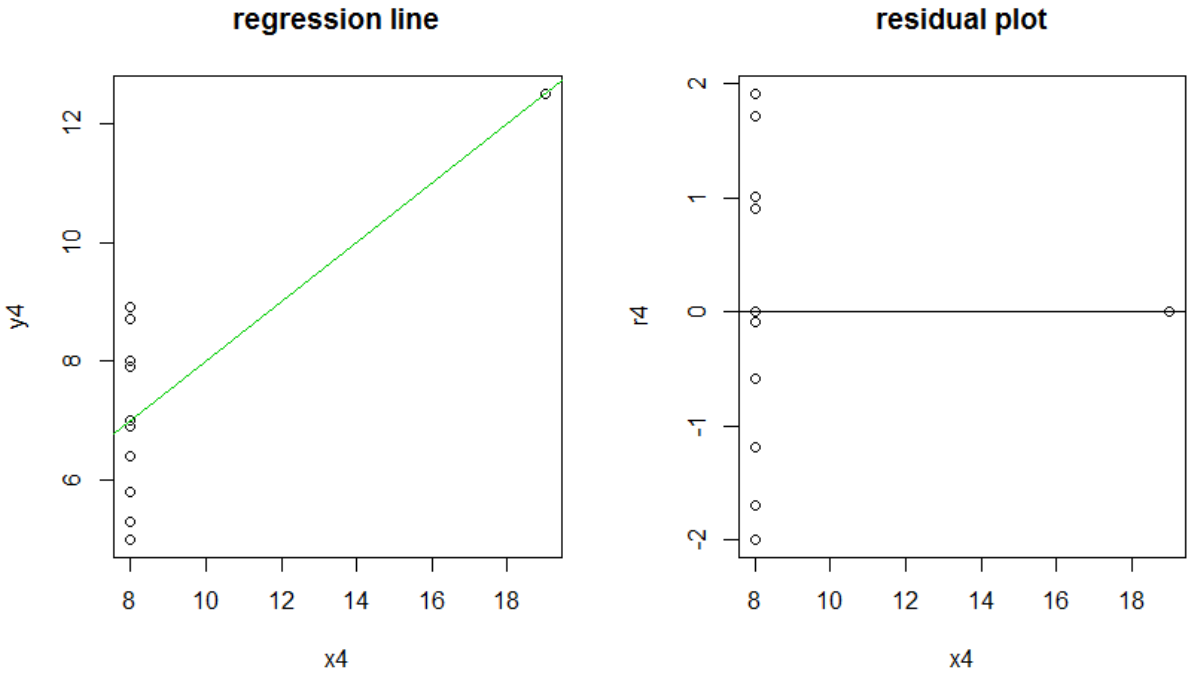



regression line
y2
x2
3
4
5
6
7
8
9
4
6
8
10
12
14


residual plot
r2
x2
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
4
6
8
10
12
14


regression line
y3
x3
6
8
10
12
4
6
8
10
12
14


residual plot
r3
x3
-0.5
0.0
0.5
1.0
1.5
2.0
4
6
8
10
12
14



Checking the residuals will be an important part of a regression analysis. The residuals measure how well the model predicts the data. In a sense, the residuals represent the failure of the model to predict the given data. As a result, the residuals provide information on the quality of the analysis. Large residuals indicate either an outlier or a poor model.

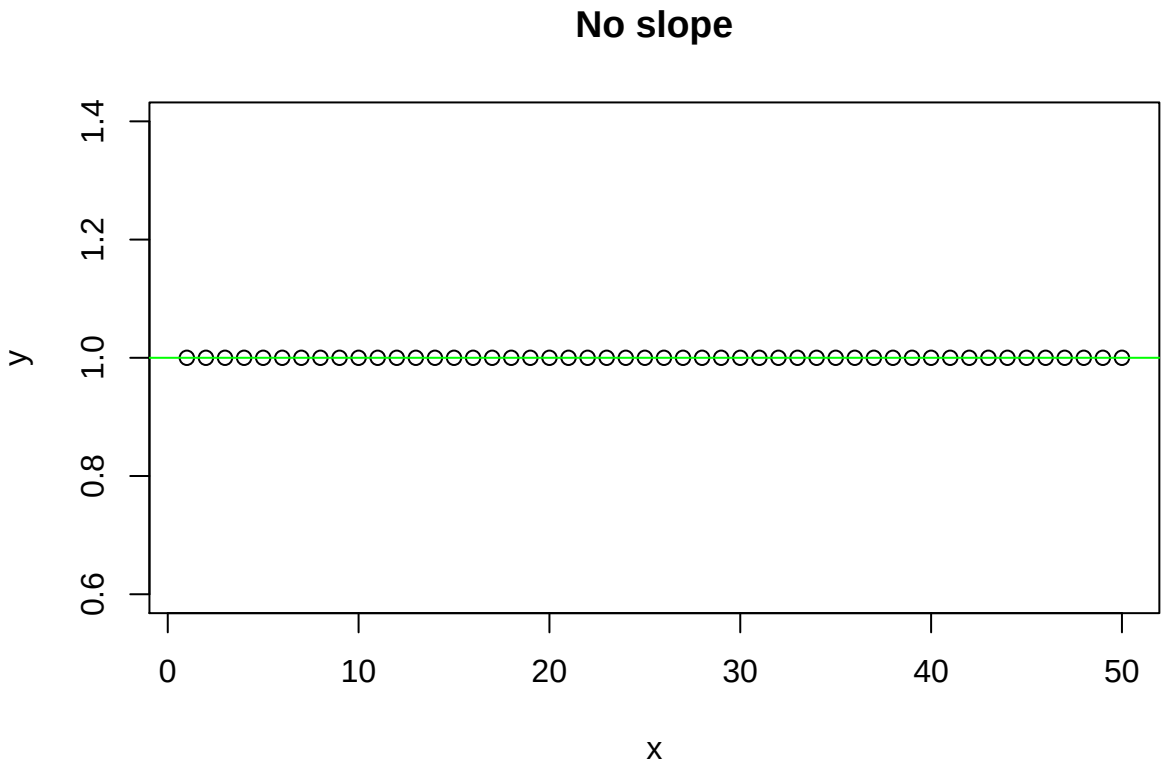
2 Hypothesis test for α and β

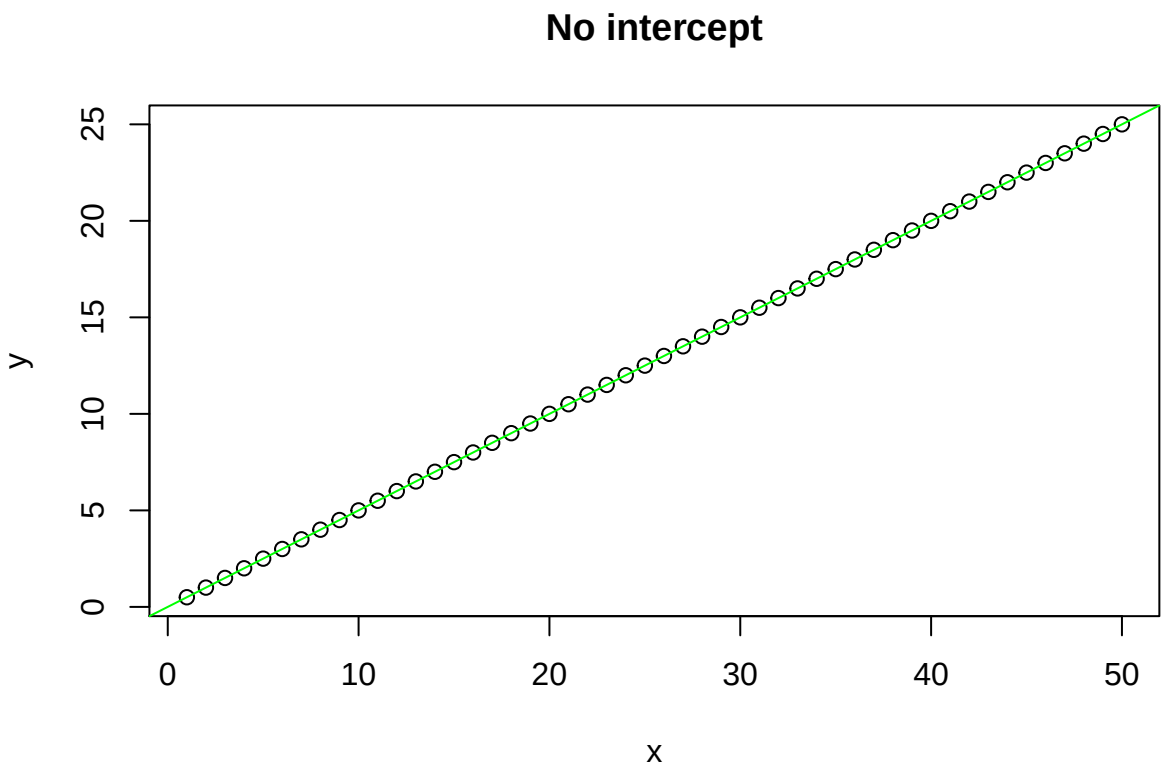
2.1 General

Our statistical regression model can be written as
$Y_i = \alpha + \beta \cdot x_i + \varepsilon_i$ with $i = 1, 2, ..., n$.

Two extreme cases can be considered:

1. No slope or $\beta = 0$
   Then $Y_i = \alpha + \varepsilon_i$ (for $i = 1, ..., n$) and there is no linear relation between x and Y.
2. No intercept or $\alpha = 0$
   Then $Y_i = \beta \cdot x_i + \varepsilon_i$ (for $i = 1, ..., n$). This is visualized as a straight line through the origin.


No slope
y
1.4
1.2
1.0
0.8
0.6
0
10
20
30
40
50
x



1. Is there a linear relationship between x and Y?
The slope parameter β represents the expected change in the response Y given a one-unit change in the regressor x and determines the nature of the relationship between the response and the regressor.

If $\beta = 0$ then the response really does not depend on the regressor at all. In that case, the expected value of the response does not change as we change the values of the regressor.

To determine if there is a linear relationship between x and Y, we test the hypothesis
$H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.
We can use the theoretical result that under H_0 (thus if $\beta = 0$)
$\frac{b}{se(b)} \sim t_{n-2}$
with $se(b)$ the standard error of b.

2. Is the intercept significant different from 0?
Usually we are not interested in this hypothesis, but if we would like to test it, we test this with the hypothesis
$H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$.
We can use the theoretical result that under H_0 (thus if $\alpha = 0$)
$\frac{a}{se(a)} \sim t_{n-2}$.

2.2 In R

Example *Temperature* in R

Our population regression model is $annual = \alpha + \beta \cdot Latitude$.

```
res.lm1 <- lm(annual ~ lat)
summary(res.lm1)
```

```
##
```

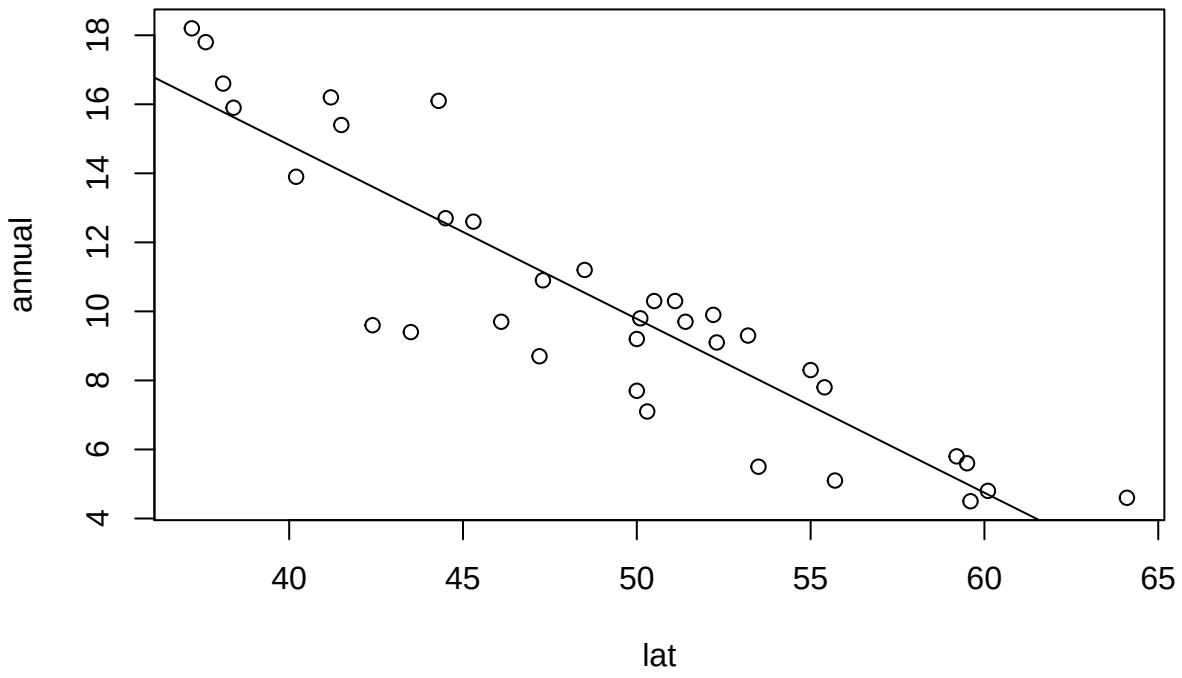
```
## Call:
## lm(formula = annual ~ lat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0116 -0.7017  0.4748  1.0526  3.4454
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.96769    2.06934   16.90  < 2e-16 ***
## lat         -0.50368    0.04177  -12.06 1.23e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.73 on 33 degrees of freedom
## Multiple R-squared:  0.815,  Adjusted R-squared:  0.8094
## F-statistic: 145.4 on 1 and 33 DF,  p-value: 1.226e-13
```

Parameter estimates for α and β:
Estimate for intercept $= \alpha = 35$
Estimate for coefficient for $Latitude = \beta = -0.5$
The estimated model is $\hat{annual} = 35 - 0.5 \cdot Latitude$

1. The $H_0 : \beta = 0$ is rejected by the t-test. The $p - value < 0.05$ and hence, we can conclude that the slope is significant different from 0.
2. The $H_0 : \alpha = 0$ is rejected since the $p - value < 0.05$. We see that the intercept is significant different from 0.

The estimated regression model is $annual = 35\breve{}0.5 \cdot Latitude$ and can be visualized by

```
plot(annual ~ lat)
abline(res.lm1)
```



2.3 Variability components

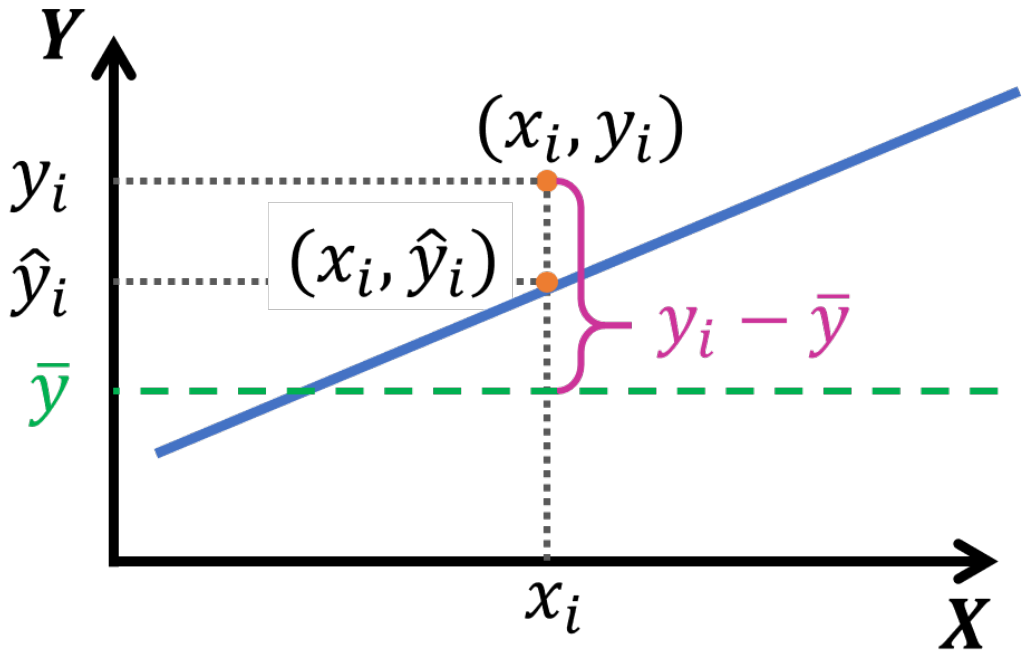
ANOVA (*Analysis of variance*) is a testing procedure to analyze the differences among group means. The term *analysis of variance* refers to analysis of the variance explained by the model.

The **overall variability** in the data is $\sum_{i=1}^{n}(y_i - \overline{y})^2$.
We can partition the overall variability in the data into two components: A part which is explained by the regression model and a part which is unexplained.

$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$
Thus, *Total SS = Error SS + Model SS*
where:

- Total SS = $\sum_{i=1}^{n}(y_i - \overline{y})^2$
  (SS = Sum of Squares)
- Error SS (= SSE) or residual SS = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  This is the **unexplained variability**.

- Model SS = $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$
  This is the **explained variability**.



2.4 How good is the linear fit?

The **coefficient of determination,** R^2 uses the relative sizes of the variability explained by the regression model and the total variability to measure the overall adequacy of the model. It is defined by

$R^2 = \frac{Model\ SS}{Total\ SS} = \frac{Explained\ variation}{Total\ variation}$

which guarantees that $0 \leq R^2 \leq 1$. We may interpret R^2 as the proportion of the total variability explained by the regression model.

For models that fit the data well, R^2 is near 1. Models that poorly fit the data have R^2 near 0.

Remark:

- $R^2 = 0 \Rightarrow$ Model SS $= 0$. The regression line is a horizontal line (or $\beta = 0$).
- $R^2 = 1 \Rightarrow$ Model SS $=$ Total SS $\Rightarrow$ Error SS $= 0$. All pairs (x_i, y_i) are on one straight line.
- $R^2 = r^2$ in case of simple linear regression.

2.5 In R

Example *Temperature* in R
We want to fit a linear relation between *annual* and *Latitude*: $annual = \alpha + \beta \cdot Latitude$.
Estimates for α and β are obtained using the *Least Squares Method*.

```
res.lm1 <- lm(annual ~ lat)
summary(res.lm1)
```

```
##
## Call:
## lm(formula = annual ~ lat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0116 -0.7017  0.4748  1.0526  3.4454
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.96769    2.06934   16.90  < 2e-16 ***
## lat         -0.50368    0.04177  -12.06 1.23e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.73 on 33 degrees of freedom
## Multiple R-squared:  0.815,  Adjusted R-squared:  0.8094
## F-statistic: 145.4 on 1 and 33 DF,  p-value: 1.226e-13
```

The **Anova table** can be obtained by

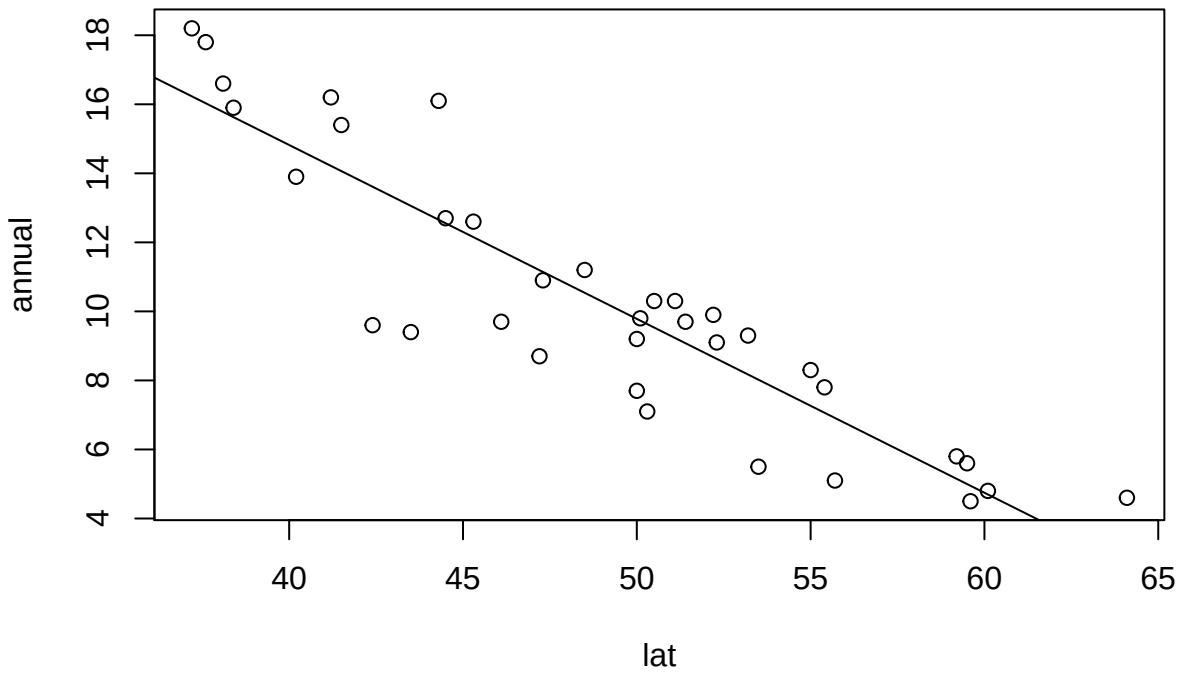
```
anova(res.lm1)
```

```
## Analysis of Variance Table
##
## Response: annual
##           Df Sum Sq Mean Sq F value    Pr(>F)
## lat        1 435.22  435.22   145.4 1.226e-13 ***
## Residuals 33  98.78    2.99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

1. Parameter estimates for α and β:
   Estimate for intercept $= \alpha = 35$
   Estimate for coefficient for $Latitude = \beta = -0.5$
2. Is the regression good?
   We look at R^2. *Latitude* can explain 81.5% of the total variability in *annual*. This is very good.
   $R^2 = r^2$ with r the Pearson correlation coefficient between *annual* and *Latitude*. From previous chapter, we observed a $r = -0.9$.
3. The estimated regression model is $\hat{annual} = 35 - 0.5 \cdot Latitude$.
4. Visualization of the points and the fitted regression line:

```
plot(annual ~ lat)
abline(res.lm1)
```



3 Model diagnostics

3.1 In general

The obtained $p - values$ for the parameter estimates are correct if the underlying assumptions are satisfied.

The behavior of the observations is correctly described by $\forall i = 1, ..., n : Y_i = \alpha + \beta \cdot x_i + \varepsilon_i$ (with ε_i the random error).
We assume that the random errors are independent normally distributed with mean 0 and variance σ^2, i.e., $\varepsilon_i \sim N(0, \sigma^2)$

Thus, the **assumptions** are:

a) Linearity: there is a linear relationship between x and Y
b) Normality of the residuals
c) Constant variance

Therefore, the **assumptions need to be checked** are:

a. **We have to check linearity:**
Make a plot of (standardized) residuals versus fitted response.
Make a plot of (standardized) residuals versus each regressor.

b. **Check normality of the (standardized) residuals:**
Perform a Shapiro Wilk test (and make a histogram)
c. **Check for influential points:**
Make a plot of the Cooks distance

3.2 Checking for linearity (check if whether the model is correct specified).

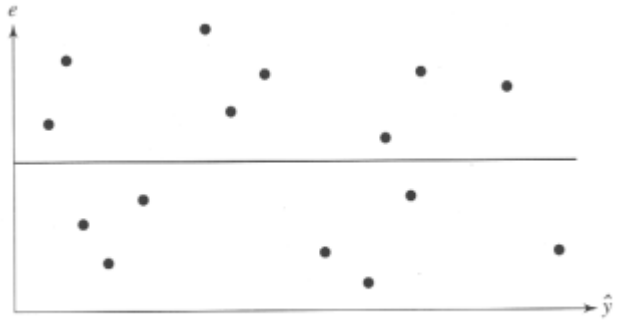
3.2.1 Introduction

Non-linearity (or model misspecification) often reveals itself by a systematic pattern in the residuals. Two plots often prove useful for identifying this problem:

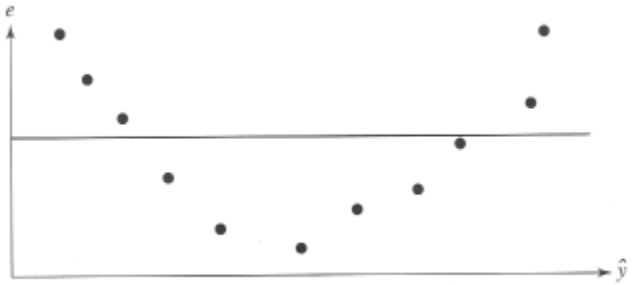
- The residuals against the fitted values
- The residuals against each regressor (independent variable)

In case of simple linear regression, the plots are essentially the same.

If our model is a **linear model**, we expect the **residual plots to look like a random pattern** as is the next graph.



If the residual plots is like below, then this is an indication of a non-linear model.



We can try to correct this problem by adding appropriate term (e.g., an x^2 term)

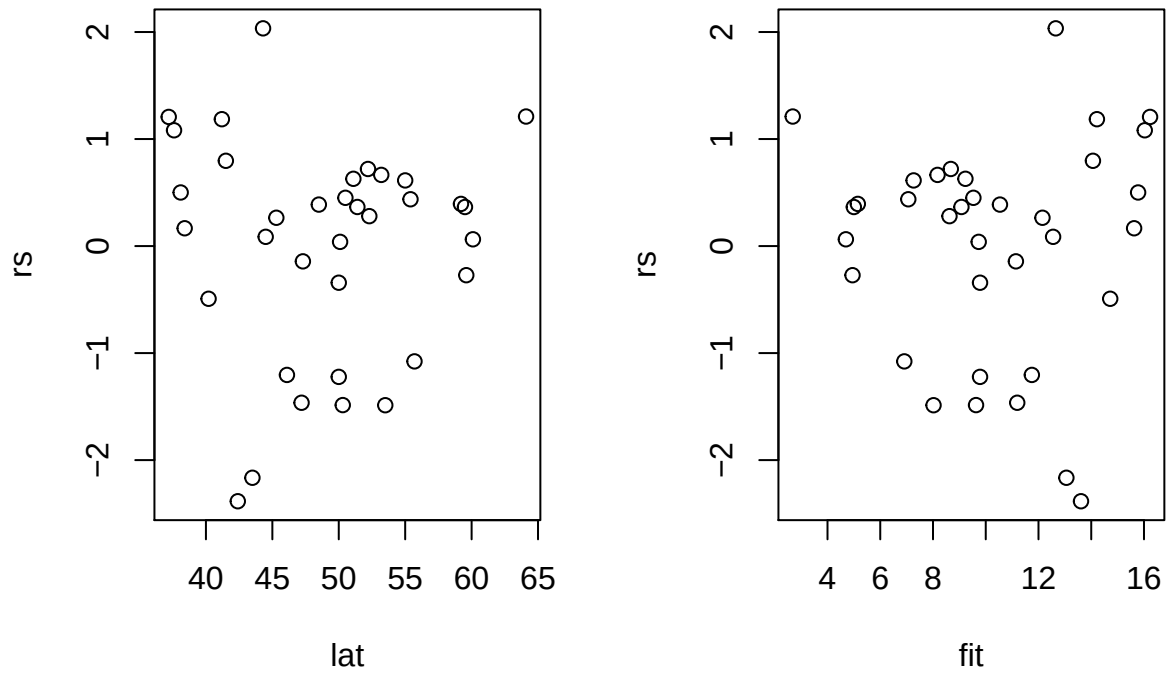
3.2.2 Examples

Example ***Temperature***

```
res.lm1 <- lm(annual ~ lat)
fit <- fitted(res.lm1)
rs <- rstandard(res.lm1)
```

We first apply the function `lm` to fit a linear model. The fitted values are extracted with the function `fitted`. The function `rstandard` returns the standard residuals which are the residuals divided by their standard deviation. Let's now plot these standard residuals:

```
par(mfrow = c(1,2))
plot(rs ~ lat)
plot(rs ~ fit)
```



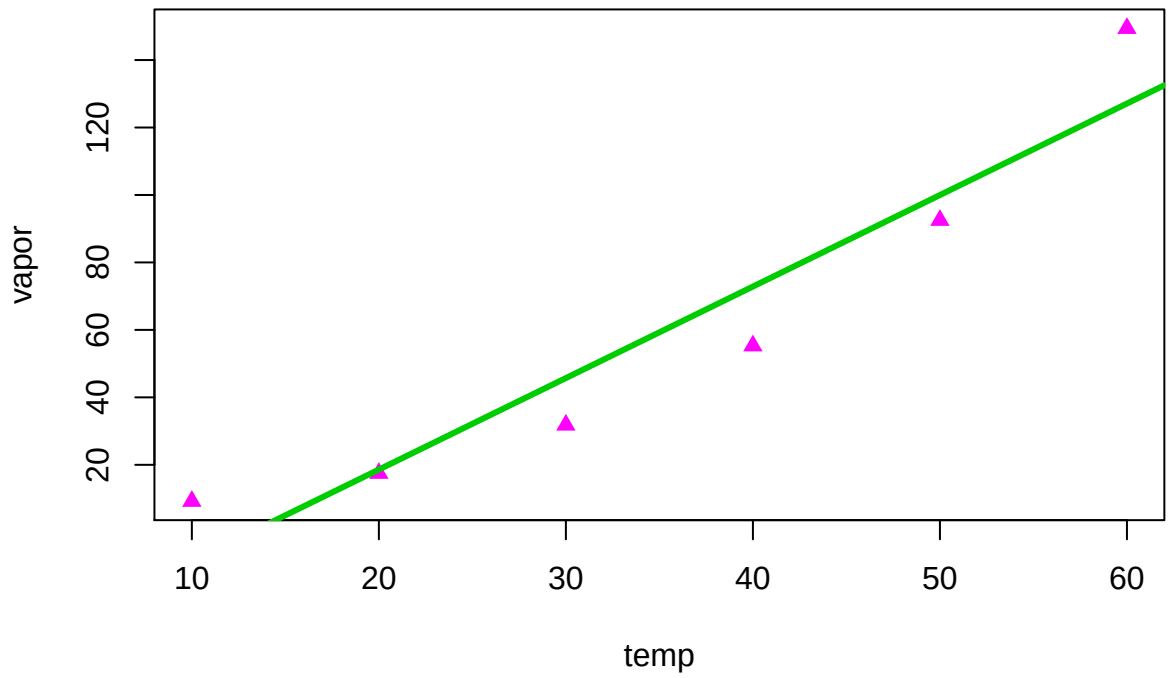
There is an ad random pattern. The assumption of linearity is satisfied.

Example ***Vapor pressure of water***

Chemical and mechanical engineers often need to know the vapor pressure of water for specific temperatures. One approach requires the engineer to always refer to the steam tables. Another approach seeks to use a simple model to predict the vapor pressure given the temperature. Thus temperature is the regressor, and vapor pressure is the response. The next table lists the vapor pressures of water for various temperatures from $10°C$ to $60°C$.

Temperature	Vapor pressure
10	9.2
20	17.5
30	31.5
40	55.3
50	92.5
60	149.4

```
temp <- c(10, 20, 30, 40, 50, 60)
vapor <- c(9.2, 17.5, 31.8, 55.3, 92.5, 149.4)
plot(vapor ~ temp, col = 6, pch = 17)
abline(lm(vapor ~ temp), col=3, lwd=3)
```



This plot indicates that we may be able to model the vapor pressures using a straight line in the temperature. Although the data display some curvature, a straight line, may serve as a useful first approximation.

Results of a regression fit:

```
res.lm <- lm(vapor ~ temp)
summary(res.lm)
```

```
##
## Call:
## lm(formula = vapor ~ temp)
##
## Residuals:
##       1       2       3       4       5       6
##  17.738  -1.090 -13.919 -17.548  -7.476  22.295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.6667    17.2317  -2.070  0.10725
## temp          2.7129     0.4425   6.131  0.00359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.51 on 4 degrees of freedom
## Multiple R-squared:  0.9038, Adjusted R-squared:  0.8798
## F-statistic: 37.59 on 1 and 4 DF,  p-value: 0.003586
```

Estimated regression line:

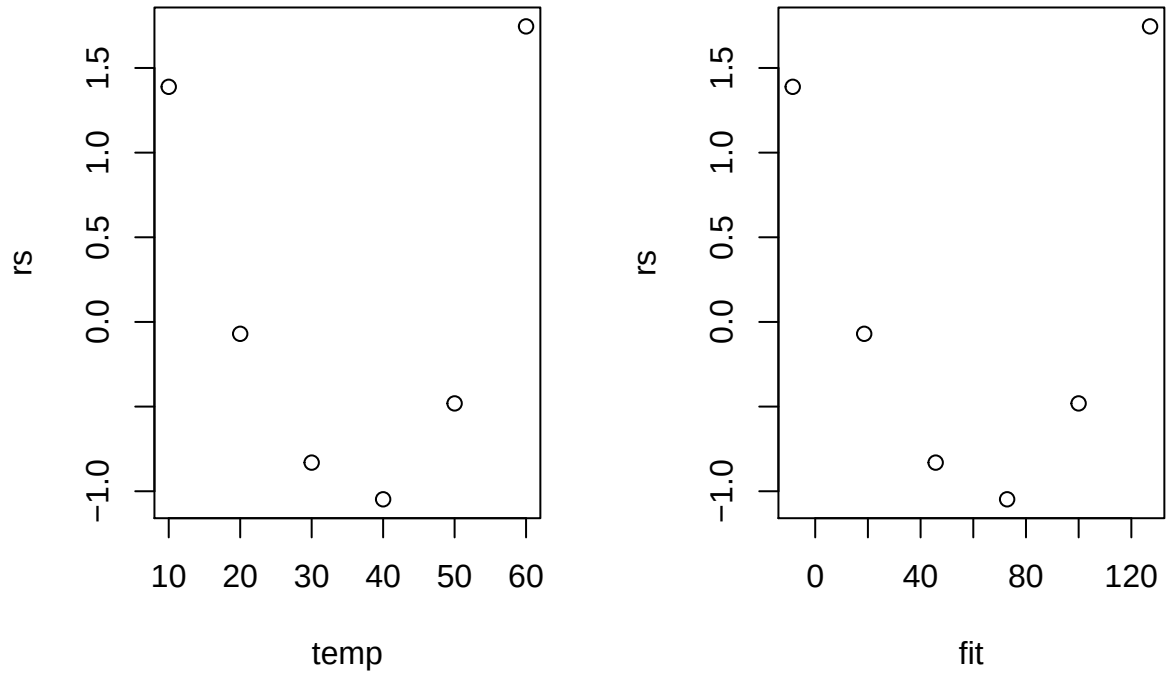
$\hat{vapor} = -36 + 2.7 \cdot temperature$

The slope is significant different from 0 (since $p - value = 0.0036$)
The coefficient of determination $R^2 = 0.90$ which indicates that a large proportion of the variability is explained by the regression model.

Residual plots

```
fit <- fitted(res.lm)
rs <- rstandard(res.lm)

par(mfrow = c(1,2))
plot(rs ~ temp)
plot(rs ~ fit)
```



There is a quadratic pattern in the residual plots. This suggest that the relation between temperature and vapor a quadratic function is (rather than a linear model).

What to do when linearity assumption does not hold?

Transforming x or adding a quadratic term of x (here temperature) to the model (polynomial model) might help.

Here, we will add a quadratic term of temperature to the regression model. Then population regression model is now: $vapor = \alpha + \beta_1 \cdot temperature + \beta_2 \cdot temperature^2$

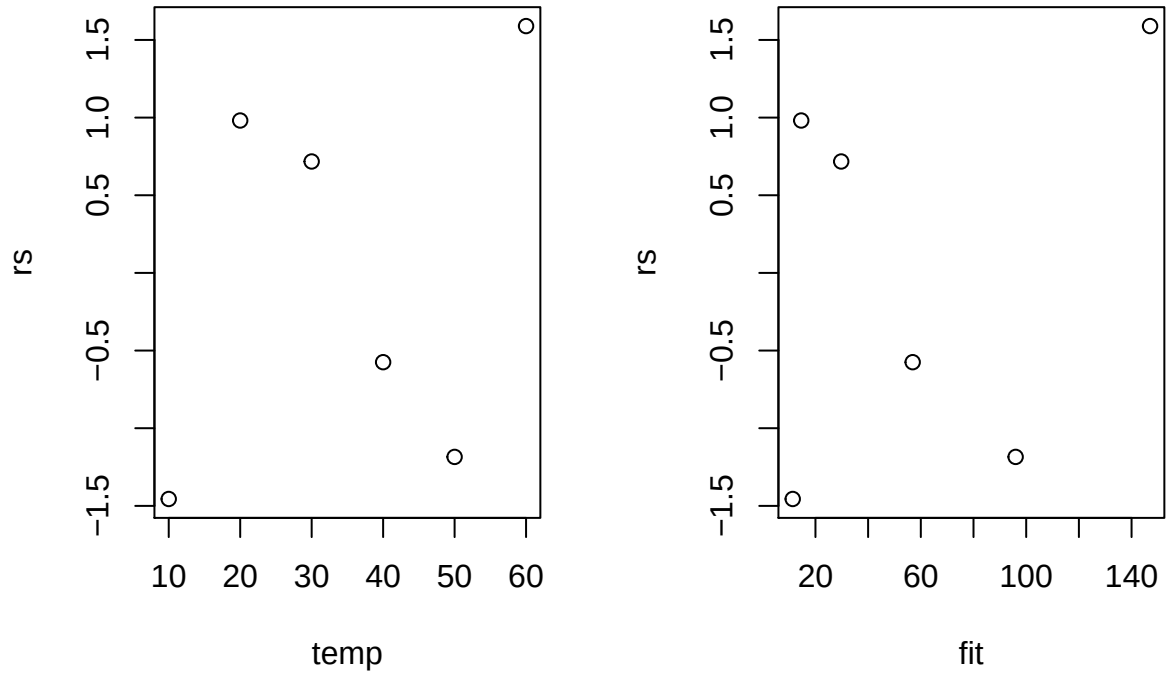
```
res.lm.q <- lm(vapor ~ temp + I(temp^2))
summary(res.lm.q)
```

```
##
## Call:
```

```
## lm(formula = vapor ~ temp + I(temp^2))
##
## Residuals:
##      1      2      3      4      5      6
## -2.179  2.893  2.014 -1.614 -3.493  2.379
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.100000   6.335989   3.172  0.05039 .
## temp        -1.469643   0.414518  -3.545  0.03822 *
## I(temp^2)    0.059750   0.005797  10.307  0.00195 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.542 on 3 degrees of freedom
## Multiple R-squared:  0.9974, Adjusted R-squared:  0.9956
## F-statistic: 566.4 on 2 and 3 DF,  p-value: 0.0001357
```

```
fit <- fitted(res.lm.q)
rs <- rstandard(res.lm.q)

par(mfrow = c(1,2))
plot(rs ~ temp)
plot(rs ~ fit)
```



Remark:
Instead of using a quadratic term, transformations of the regressor may also be used : `ln(x)`, `ln(x+1)`,

`sqrt(x)`, `1/x`, `exp(x)`, ...

3.3 Checking normality of the residuals

3.3.1 Introduction

- We can look at a histogram of the (standardized) residuals for a graphical check.
- We can use the Shapiro Wilk test on the (standardized) residuals as formal test for normality.

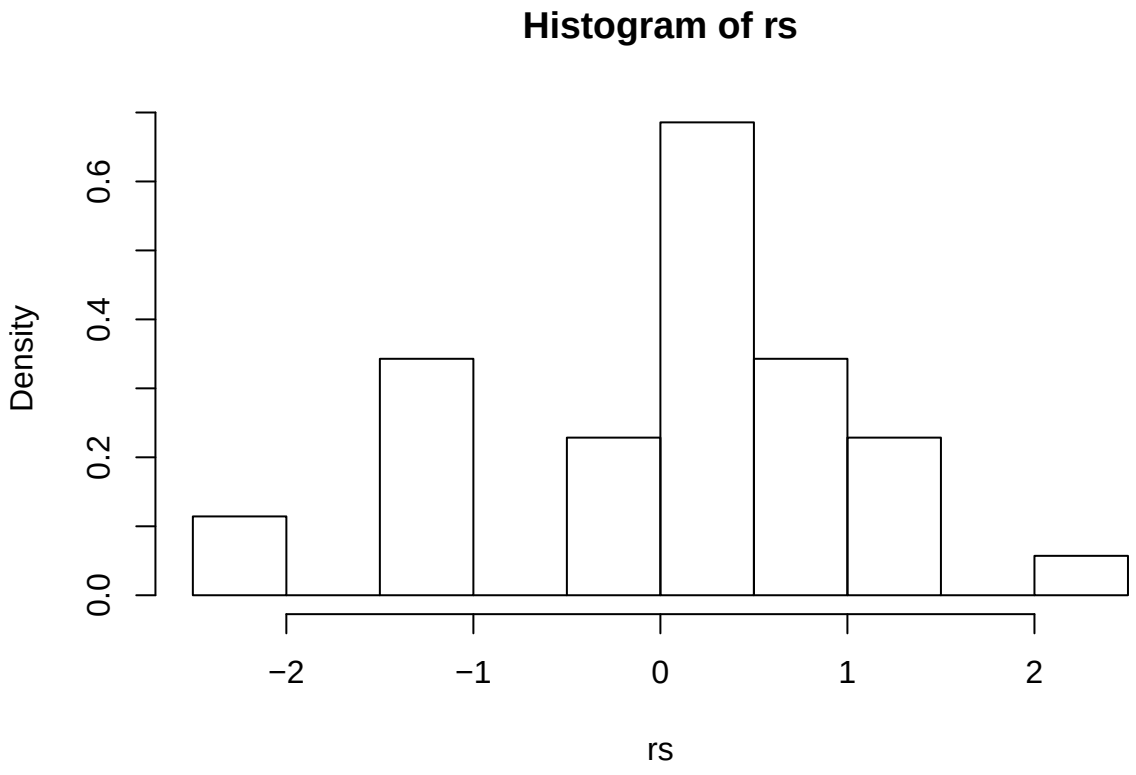
3.3.2 Examples

Example *Temperature*

```
res.lm1 <- lm(annual ~ lat)
rs <- rstandard(res.lm1)
shapiro.test(rs)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rs
## W = 0.93756, p-value = 0.04717
```

```
hist(rs, prob = TRUE)
```



Example *bacteria deaths*

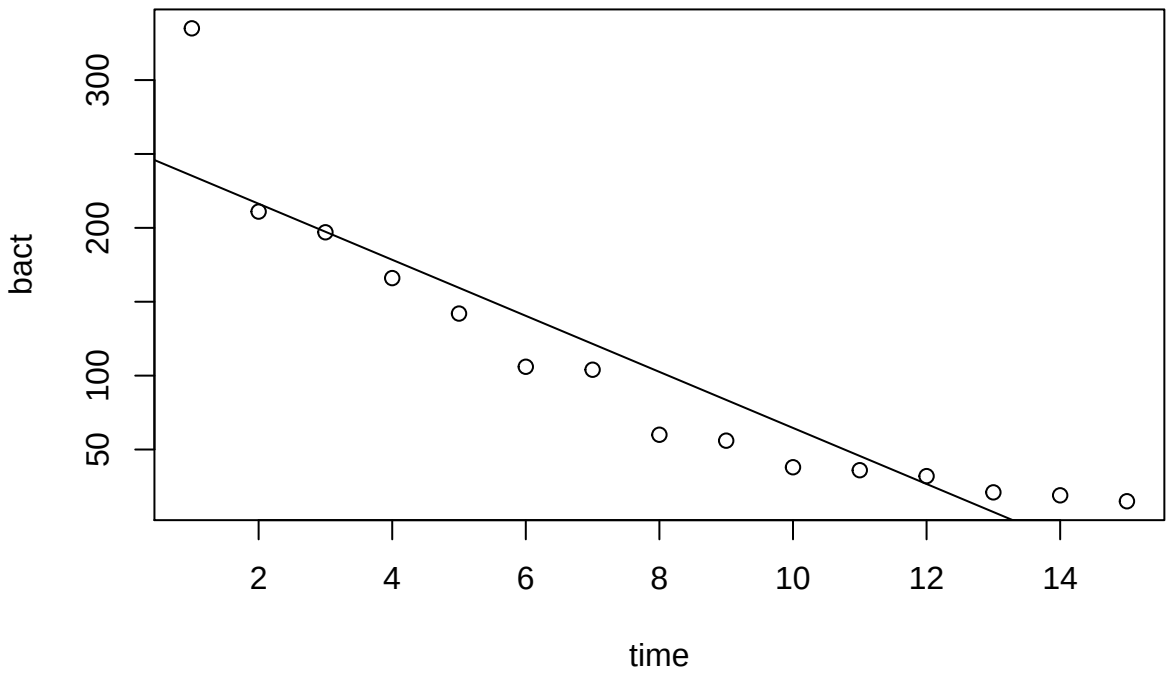
The following data represent the number of surviving bacteria (in hundreds) after exposure time t. Bacteria are exposed to 200-kilovolt X-rays for periods ranging from $t = 1$ to 15 intervals of 6 minutes.

```r
time <- 1:15
bact <- c(335, 211, 197, 166, 142, 106, 104, 60, 56, 38, 36, 32, 21, 19, 15)


res.lm <- lm(bact ~ time)
summary(res.lm)
```

```
##
## Call:
## lm(formula = bact ~ time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.533 -22.051  -9.640   9.306  99.717
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  254.248     20.474  12.418 1.38e-08 ***
## time         -18.964      2.252  -8.422 1.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.68 on 13 degrees of freedom
## Multiple R-squared:  0.8451, Adjusted R-squared:  0.8332
## F-statistic: 70.93 on 1 and 13 DF,  p-value: 1.269e-06
```

```r
plot(bact ~ time)
abline(res.lm)
```



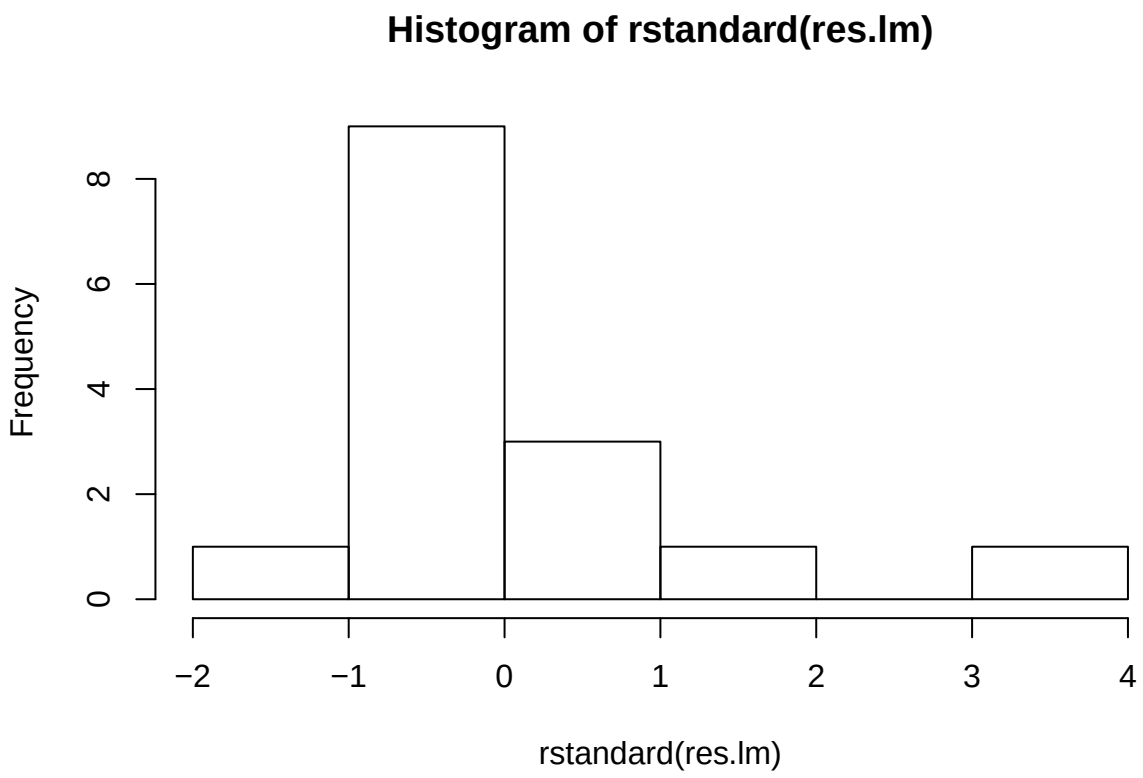

```
rstandard(res.lm)
```

```
##           1           2           3           4           5           6
##  3.03895573 -0.15735729 -0.01024803 -0.35129780 -0.48717036 -0.95401887
##           7           8           9          10          11          12
## -0.48159257 -1.16841662 -0.75879172 -0.73650731 -0.26951124  0.15094194
##          13          14          15
##  0.38385389  0.89497839  1.37801887
```

```
shapiro.test(rstandard(res.lm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(res.lm)
## W = 0.84505, p-value = 0.01479
```

```
hist(rstandard(res.lm))
```



Standard residuals are not normally distributed.

What to do when normality of residuals does not hold?
We transform the response: `ln(Y)`, `sqrt(Y)`, `1/Y`, `exp(Y)`, `ln(Y+1)`,...

Example ***bacteria deaths*** $ln(bacteria)$.
The population regression model is know:
$\ln(bacteria) = \alpha + \beta \cdot time$

```
ln_bact <- log(bact)
reslog.lm <- lm(ln_bact ~ time)
summary(reslog.lm)
```

```
##
## Call:
## lm(formula = ln_bact ~ time)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17189 -0.05803  0.01255  0.04890  0.20552
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.957697   0.057638  103.36  < 2e-16 ***
## time        -0.216976   0.006339  -34.23 3.99e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1061 on 13 degrees of freedom
## Multiple R-squared:  0.989,  Adjusted R-squared:  0.9882
## F-statistic:  1171 on 1 and 13 DF,  p-value: 3.994e-14
```

```
shapiro.test(rstandard(reslog.lm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(reslog.lm)
## W = 0.95694, p-value = 0.6394
```

Residuals are now normally distributed.
The estimated regression model is now: $ln(\hat{bacteria}) = 6 - 0.22 \cdot time$

- R^2 has increased from 0.82 to 0.98
- The effect of time stays significant
- The obtained model is a nonlinear model in *bacteria* but a linear model in $ln(bacteria)$.

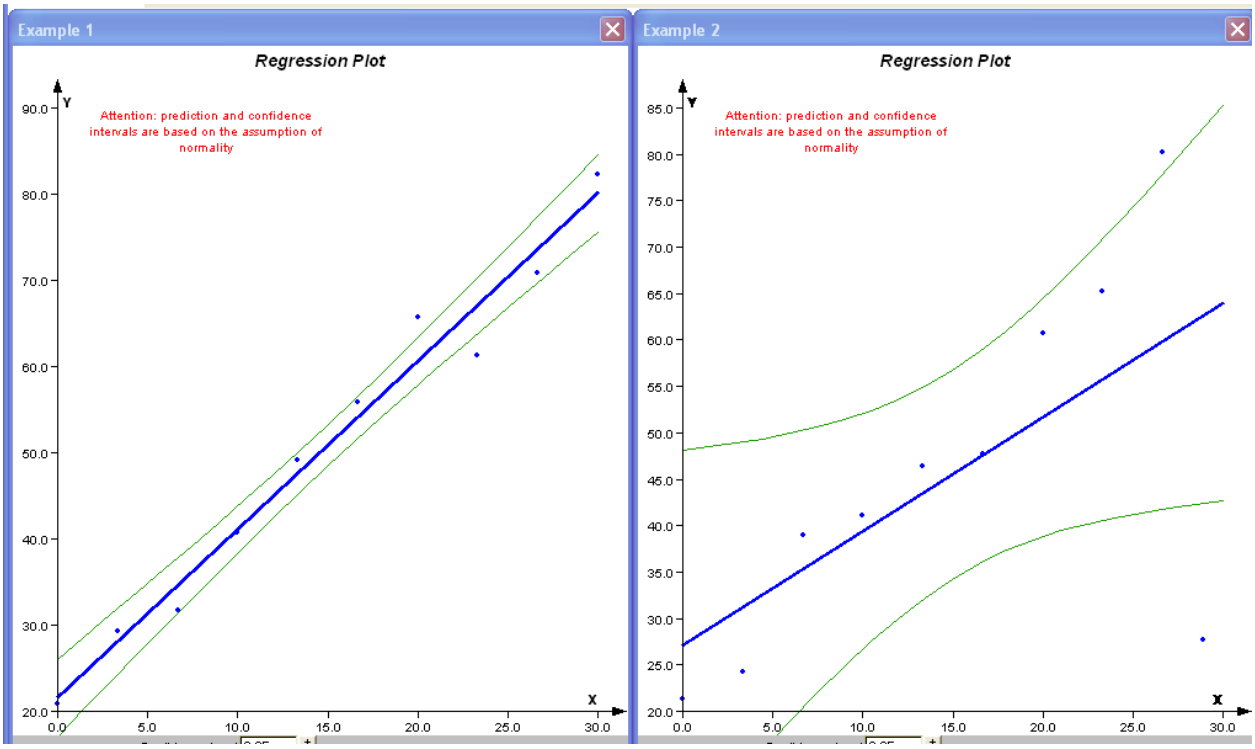
3.4 Influential points

3.4.1 Introduction

We have already observed that outliers are data values that appear to be distinctly different from the rest of the data. In the regression setting, that may appear to be distinctly different in terms of the response or in terms of the regressors. Thus there are three possibilities for the different points:

1. **Outliers**, which are data points where the observed response does not appear to follow the pattern established by the rest of the data.
2. **Leverage points**, which are data points that are distant from the other data points in terms of the regressors.
3. **Influential points**, which try to combine the concepts of both leverage points and outliers.

Outliers are extreme data values in terms of the y-direction. Leverage points are extreme data values in terms of the x's. Influential points are extreme in a combined sense. (see http://lstat.kuleuven.be/java/index.htm)



3.4.2 How to detect: Cook's distance

Cook's distance is a way to detect points that negatively affect the regression model.[1] The calculation of Cook's distance C_i for the i^{th} observation requires the creation of two different regression models:

1. Regression with all observations $\Rightarrow$ vector of regression estimates B
2. Regression with all observations excepts the i^{th} observation $\Rightarrow$ vector of regression estimates $B_{(i)}$

Cook's distance C_i is the distance between the vector of regression coefficients B and the vector of regression coefficients $B_{(i)}$

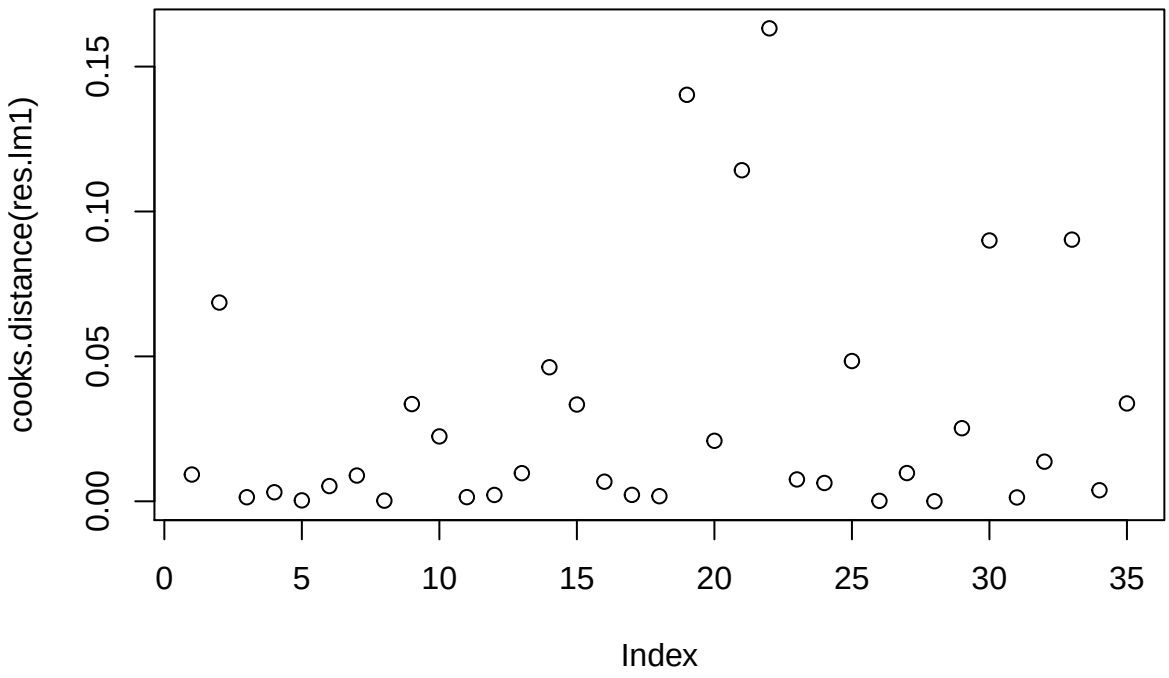
- When the C_i values are all about the same, no action need to be taken!
- If there are data points with C_i values that stand out from the rest, these points should be flagged and examined! The model may then be refitted without that point to see the effect of that point.

3.4.3 Examples

Example *Temperature*

```
res.lm1 <- lm(annual ~ lat)
plot(cooks.distance(res.lm1))
```

[1] From https://www.statisticshowto.com/cooks-distance/



There are no influential points.

Example ***Surface tension of water-based coatings***

The laboratory manager of a pant manufacturer asked a summer intern to make a series of water-based coatings by changing only the amount of the surfactant. The manager wanted the intern to see exactly what the surfactant does to the surface tension of the coating.

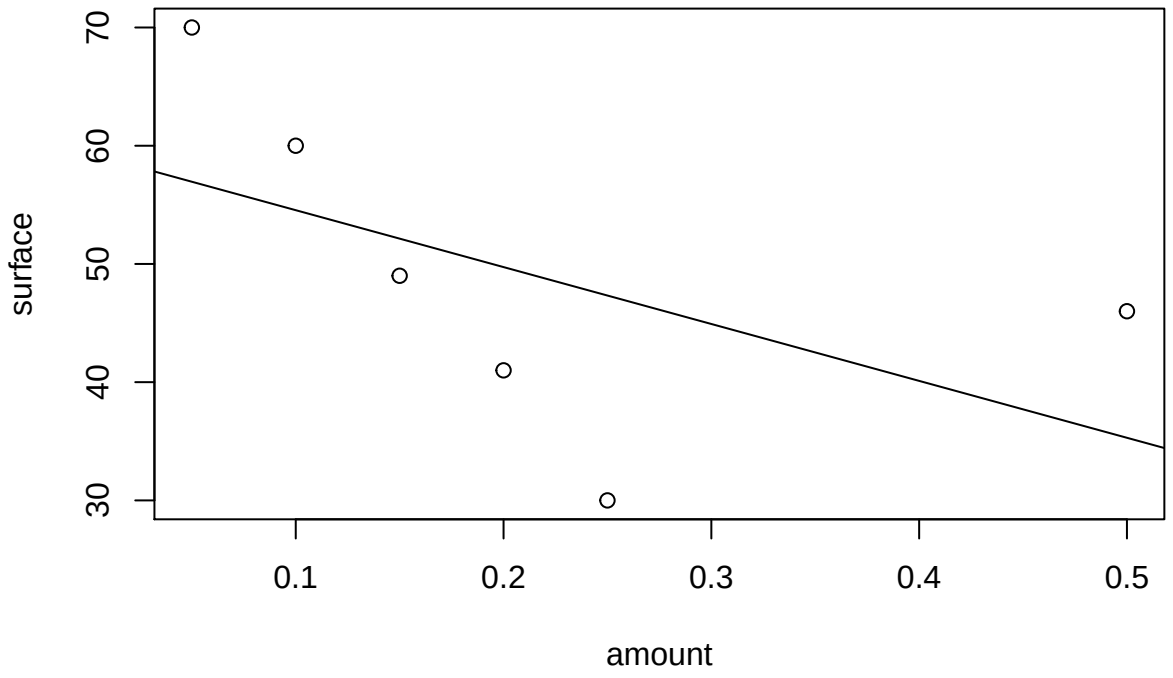
Amount surfactant	Surface tension
0.05	70
0.10	60
0.15	49
0.20	41
0.25	30
0.50	46

```
amount <- c(0.05,0.1,0.15,0.2,0.25,0.5)
surface <-c(70,60,49,41,30,46)
ex2 <- data.frame(amount, surface)
res.lm3 <- lm(surface ~ amount)
summary(res.lm3)
```

```
##
## Call:
## lm(formula = surface ~ amount)
##
## Residuals:
```

```
##       1       2       3       4       5       6
##  13.046   5.452  -3.141  -8.734 -17.328  10.705
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.361      9.437   6.290  0.00326 **
## amount       -48.131     37.133  -1.296  0.26464
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.24 on 4 degrees of freedom
## Multiple R-squared:  0.2958, Adjusted R-squared:  0.1197
## F-statistic:  1.68 on 1 and 4 DF,  p-value: 0.2646
```

```
plot(surface ~ amount)
abline(res.lm3)
```



In this case, the data point with an amount of 0.50 appears to be highly influential because its response does not seem to follow the same trend as the rest of the data and it is extreme relative to the other amounts. *When we use least squares to estimate our model, influential points tend to draw the line to themselves* as can be seen in the graph.

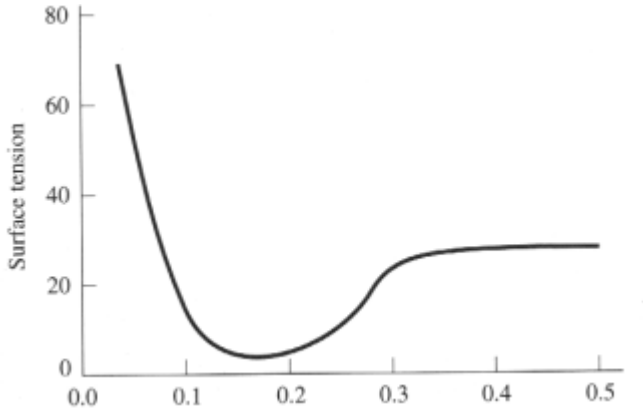
In general there are three reasons why an influential point can be present:

1. Someone made a recording error
2. Someone made a fundamental error collecting the observation
3. The data point is perfectly valid, in which case the model cannot account for the behavior.

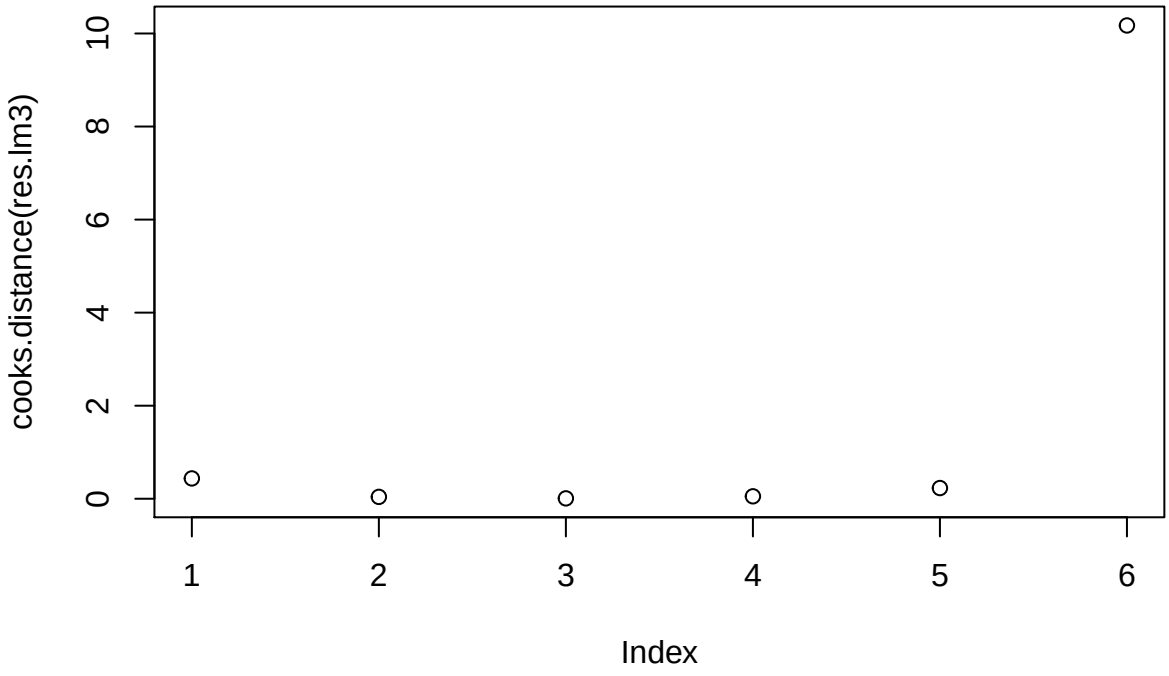
Remark:
Be careful by throwing out data: **"Sometimes we throw out perfectly good data when we should be throwing out questionable models".**

Continuation of example *Surface tension of water based coatings*

The next figure is a typical plot of the surface tension of water-based coating as we add surfactant. Although a straight-line model works well for part of the curve, it cannot work well over the entire range of interest.




```
plot(cooks.distance(res.lm3))
```




```
ex2[cooks.distance(res.lm3)>8,]
```

```
##   amount surface
## 6    0.5      46
```

The last data point is influential. We can show this by deleting that point and showing that the obtained regression parameters are quite different.

Delete last point and redo the analysis:

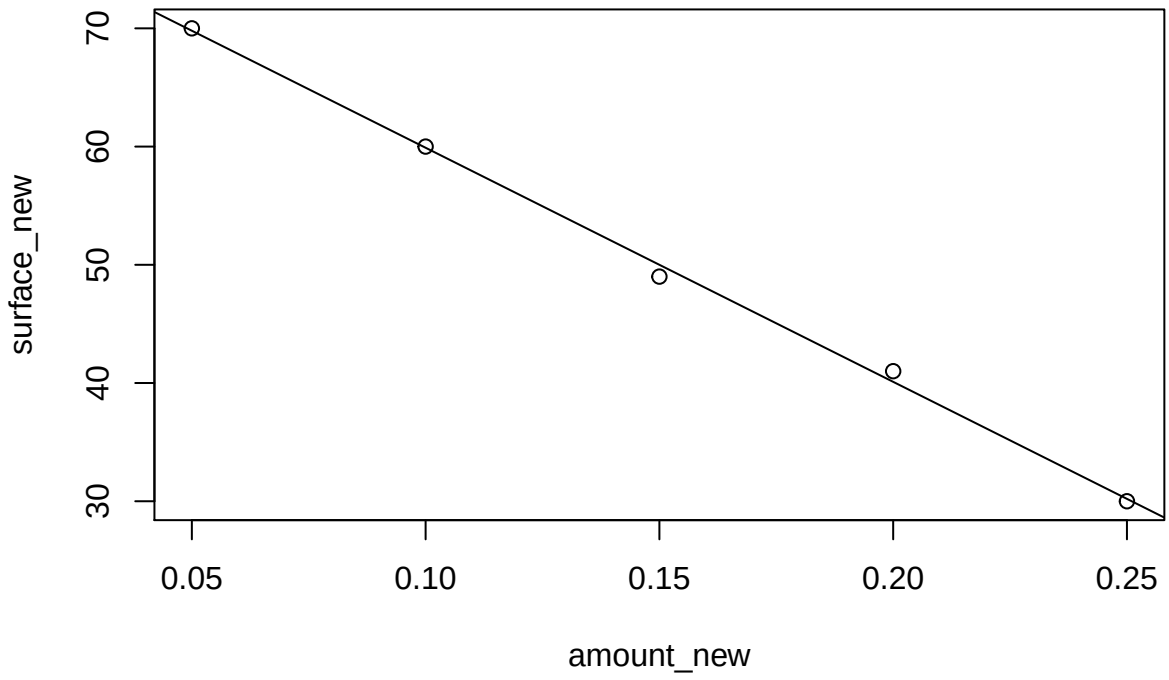
```
amount_new <- amount[-6]
surface_new <- surface[-6]
res.lm3_new <- lm(surface_new ~ amount_new)
summary(res.lm3_new)
```

```
##
## Call:
## lm(formula = surface_new ~ amount_new)
##
## Residuals:
##    1    2    3    4    5
##  0.2  0.1 -1.0  0.9 -0.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   79.7000     0.8347   95.49 2.53e-06 ***
## amount_new  -198.0000     5.0332  -39.34 3.61e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7958 on 3 degrees of freedom
## Multiple R-squared:  0.9981, Adjusted R-squared:  0.9974
## F-statistic:  1548 on 1 and 3 DF,  p-value: 3.614e-05
```

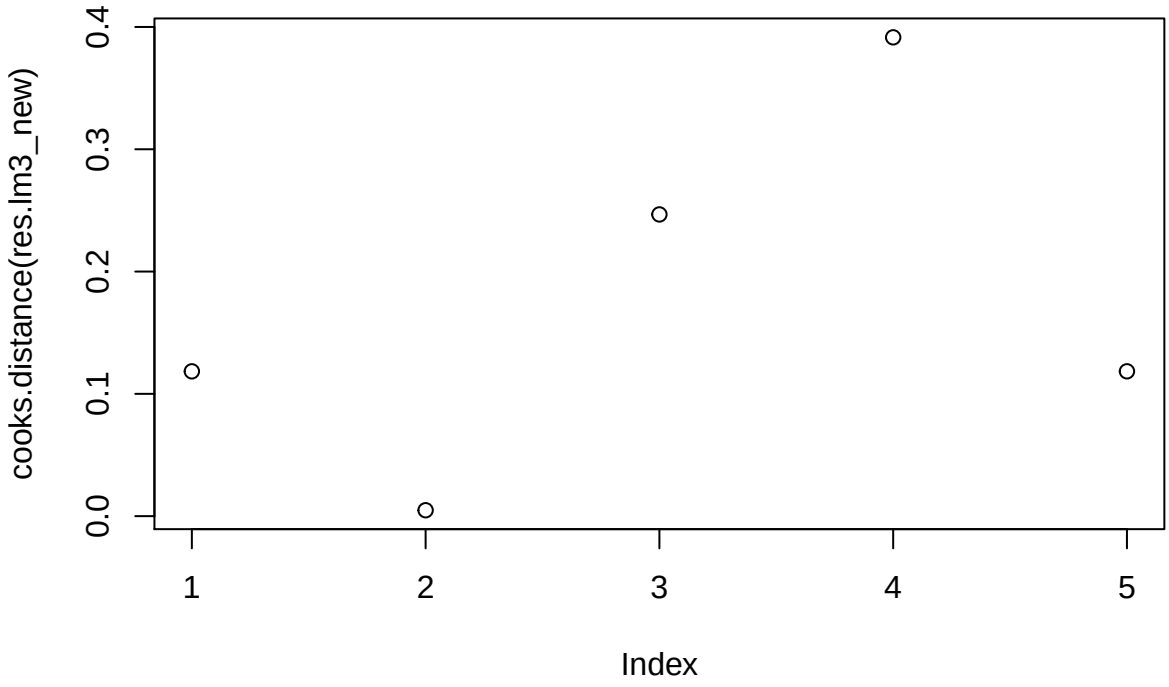
Comparison of intercept and slope before and after removing the influential point:

Number of observations	Intercept	Slope
N = 6	59.4	-48
N = 5	79.7	-198

```
plot(surface_new ~ amount_new)
abline(res.lm3_new)
```




```
plot(cooks.distance(res.lm3_new))
```



We observe that the regression estimates are totally different! Data point 6 is an influential point.

4 Prediction

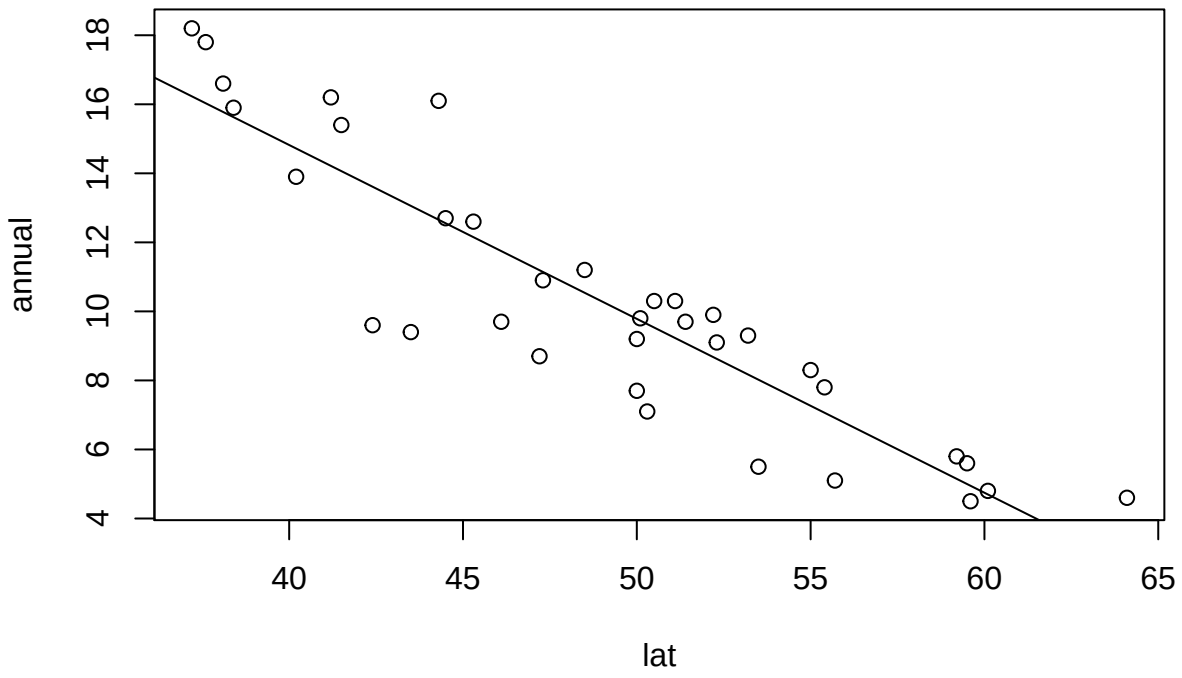
Once we are satisfied with our model, we often use it for prediction.

Example *Temperature*
The estimated regression model is
$\hat{annual} = 35 - 0.5 \cdot Latitude$
This model can be improved (see later), but we want to use this to illustrate the concept of prediction.

Assume we want to predict the annual temperature for Belgium (`lat=50`), Athens (`lat=37`), Ankara (`lat=39`) and Bangkok (`lat=13`). (visit the following website to have latitude values of many world cities http://www.infoplease.com/ipa/A0001769.html)

We might be interested in

1. Predict the average annual temperature for a city with given value for the variable *Latitude*.
2. Predict the annual temperature for a city with given value for the variable *Latitude*.



4.1 Prediction of the expected response for some specific value of the regressor.

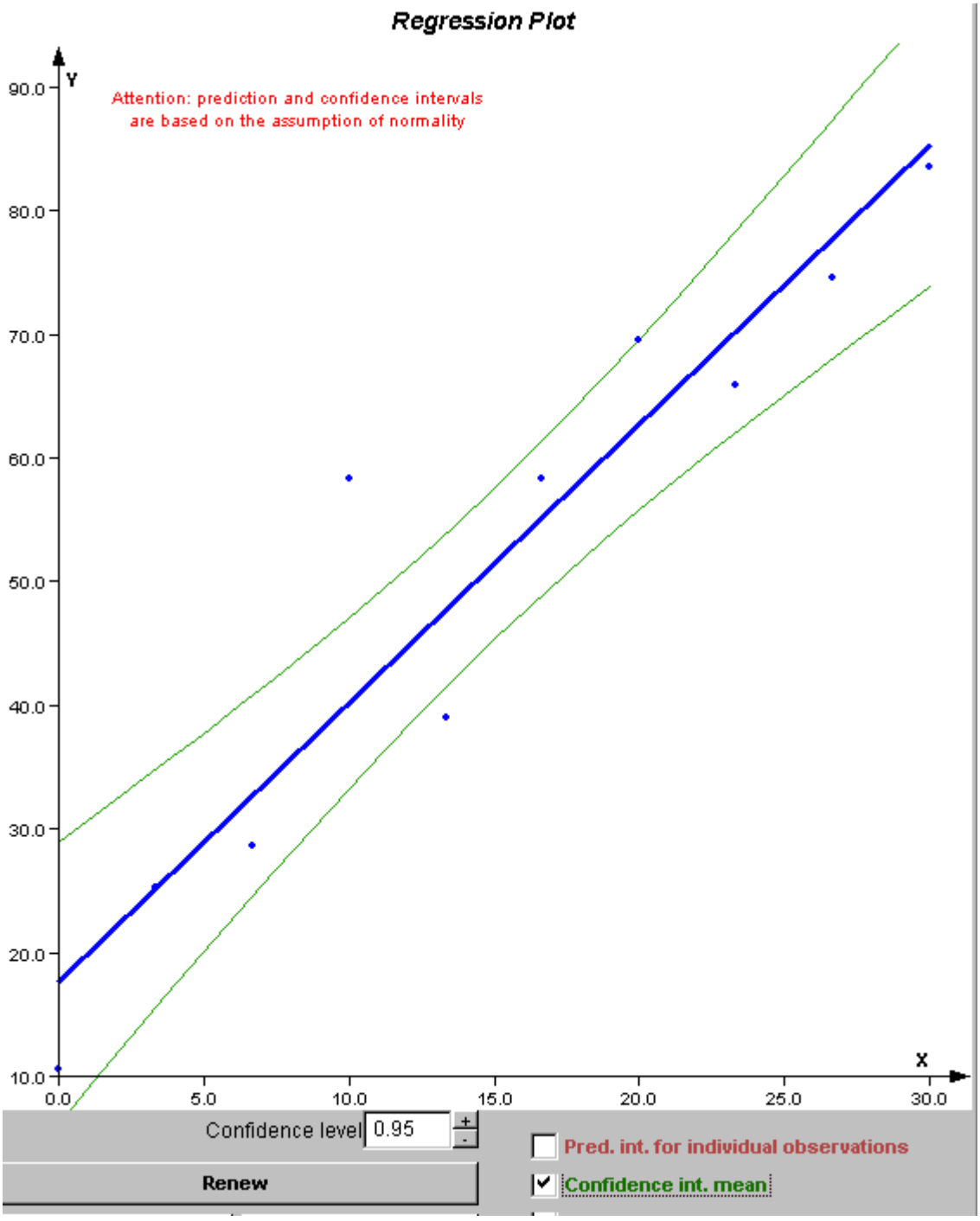
Often we wish to **estimate the expected response** for some specific value of the regressor. Let x_0 be such a specific value, not necessarily one used to estimate the model, and let $\hat{y}(x_0) = a + bx_0$ be the resulting predicted value for the response from the estimated model.

A $(1-\alpha)100\%$ **confidence interval for the expected response** when the regressor is set to x_0 is

$$\left[a + bx_0 - t_{n-2,\frac{\alpha}{2}} \cdot S\sqrt{\tfrac{1}{n} + \tfrac{(x_0-\overline{x})^2}{\sum(x_0-\overline{x})^2}}, a + bx_0 + t_{n-2,\frac{\alpha}{2}} \cdot S\sqrt{\tfrac{1}{n} + \tfrac{(x_0-\overline{x})^2}{\sum(x_0-\overline{x})^2}}\right]$$

The width of this confidence interval depends on the distance of x_0 from $\overline{x}$. The width is minimum when $x_0 = \overline{x}$. As x_0 gets farther from the center of the x's used to estimate the model, the confidence interval becomes wider. In some cases, we may wish to use our model for extrapolation, where we predict the response for a x_0 outside the range of x's used to estimate the model. In such a case, the interval can become so wide that almost any value is plausible for the prediction. *As a result, we generally do not recommend using the estimated model for extrapolation.*

Usually we are interested in the confidence interval for several values of the regressor. A $(1-\alpha)100\%$ confidence band for the expected values of the response is the plot of the $(1-\alpha)100\%$ confidence intervals for the values of the regressor over the region of interest.



(see http://lstat.kuleuven.be/java/index.htm)

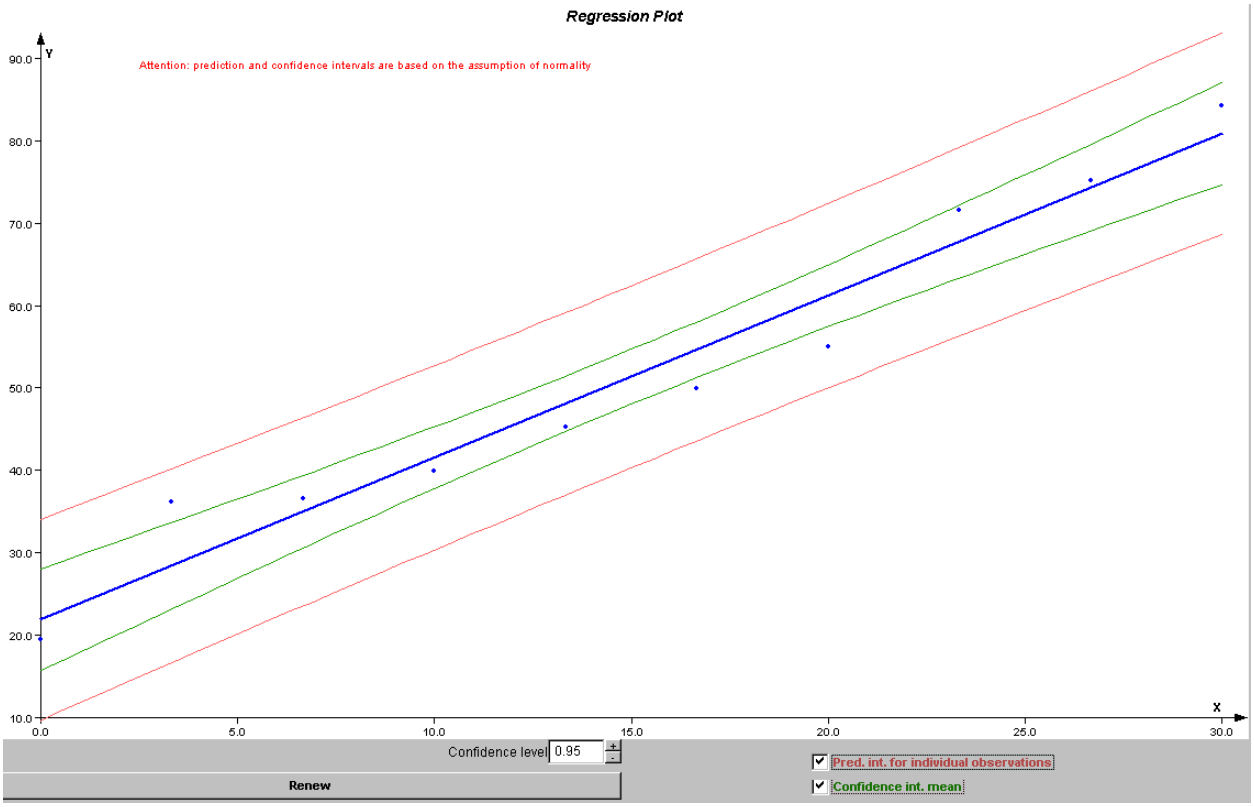
4.2 Prediction of the response for some specific value of the regressor: Prediction interval.

Whereas the confidence interval focuses on the estimation of the expected value or the mean of the response, the **prediction interval** considers the estimation of the individual responses. Such an interval must consider both the variability in the prediction of the expected value of the response and the variability of the individual responses around the expected value.

A $(1-\alpha)100\%$ **prediction interval** is

$$\left[a+bx_0-t_{n-2,\frac{\alpha}{2}}\cdot S\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum(x_0-\overline{x})^2}}, a+bx_0+t_{n-2,\frac{\alpha}{2}}\cdot S\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum(x_0-\overline{x})^2}}\right]$$

The constant 1 in this expression takes care of the variability associated with the individual responses. The $(1-\alpha)100\%$ prediction band is the plot of the $(1-\alpha)100\%$ prediction intervals for the values of the regressor over the region of interest.



(see http://lstat.kuleuven.be/java/index.htm)

4.3 In R

Example *Temperature* in R

Assume we want to predict the annual temperature for Belgium (`lat=50`), Athens (`lat=37`), Ankara (`lat=39`) and Bangkok (`lat=13`).

```
res.lm1 <- lm(annual ~ lat)
city <- c("Belgium", "Athens", "Ankara", "Bangkok")

res.pred_CI <- predict(res.lm1, list(lat= c(50, 37, 39, 13)), interval = "confidence")
pred_CI <- data.frame(city, res.pred_CI)
pred_CI
```

```
##      city       fit       lwr      upr
## 1 Belgium  9.783619  9.183106 10.38413
## 2  Athens 16.331479 15.147652 17.51531
## 3  Ankara 15.324116 14.283717 16.36451
## 4 Bangkok 28.419835 25.299523 31.54015
```

```
res.pred_PI <- predict(res.lm1, list(lat= c(50, 37, 39, 13)), interval = "prediction")
pred_PI <- data.frame(city, res.pred_PI)
pred_PI
```

```
##      city       fit       lwr      upr
## 1 Belgium  9.783619  6.212823 13.35442
## 2  Athens 16.331479 12.617799 20.04516
## 3  Ankara 15.324116 11.653639 18.99459
## 4 Bangkok 28.419835 23.715973 33.12370
```

Remark:

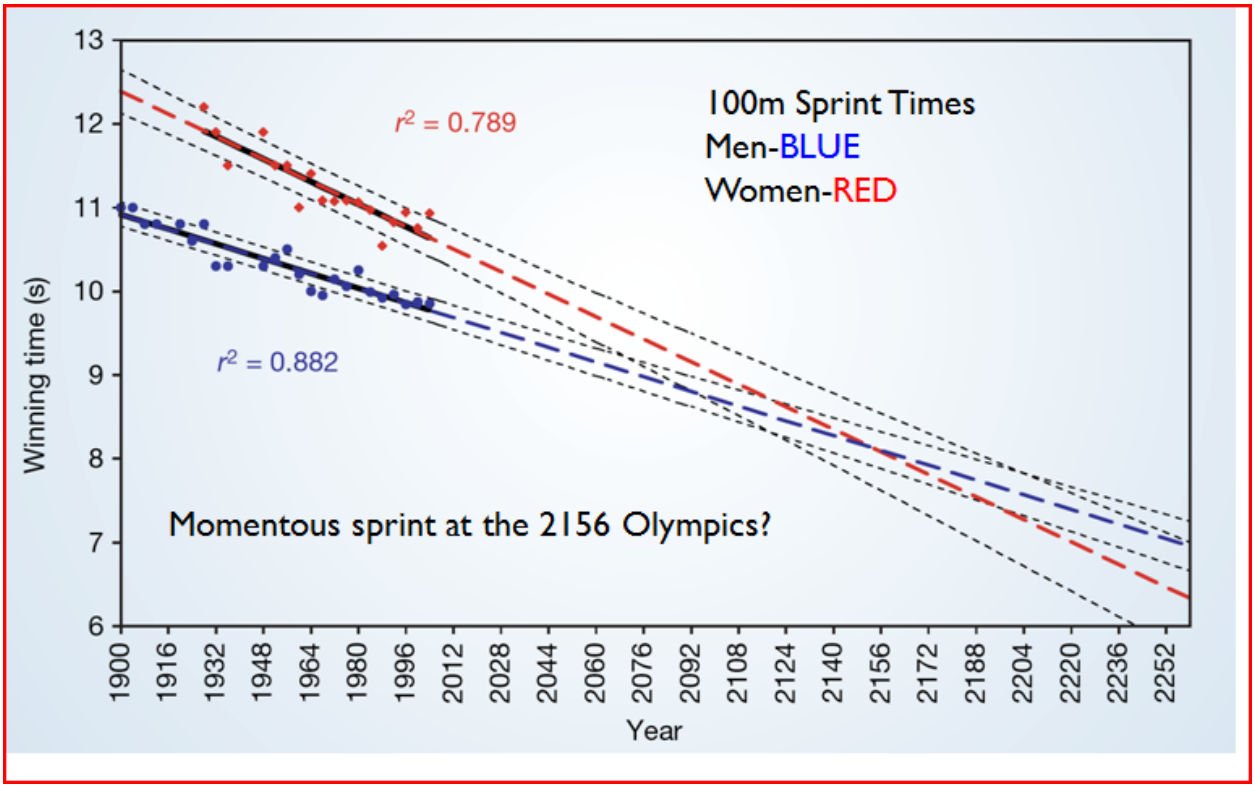


Figure 1: Jonathan Gelfond (September 2013, Flames)

5 Overview: Global structure for regression analysis

1. Investigate scatter plots between Y (response) and the x's (explanatory variables) and perform a correlation analysis.
2. Propose a linear regression between Y and x's
   $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$
3. Check if all the x's are significant.
4. If yes, go to step 6
   If no, go to step 5
5. Remove that variable with the highest p-value. Rerun the regression and go to step 3.
6. Check the underlying assumptions of regression analysis:
   a) Normality of the residuals (if not, use a transformation and go back to step 2)
   b) Check for non-linearity (if you observed a quadratic pattern, use e.g. a quadratic term and go to step 2)

c) Check for influential observations (remove outlier if this is allowed and go to step 2)

7. Based on the value of R^2, you can use this model for prediction or not.

6 Multiple linear regression

In most engineering problems, the response depends on several regressors. Multiple regression is a straightforward extension of the simple linear regression to more than one regressor.

6.1 Illustrative example

Example *Temperature*

So far, we only considered the variable *Latitude* to explain the annual temperature. We do know that there is also the variable *Longitude*.

```
##    annual Amplitude Latitude Longitude
## 1     9.9      14.6     52.2       4.5
## 2    17.8      18.3     37.6      23.5
## 3     9.1      18.5     52.3      13.2
## 4    10.3      14.4     50.5       4.2
## 5    10.9      23.1     47.3      19.0
## 6     7.8      17.5     55.4      12.3
## 7     9.3      10.2     53.2       6.1
## 8     4.8      23.4     60.1      25.0
## 9     7.1      25.3     50.3      30.3
## 10    7.7      22.1     50.0      19.6
```

We now try to fit a plane
$annual = \alpha + \beta_1 \cdot Latitude + \beta_2 \cdot Longitude$

A regression model is then
$annual_i = \alpha + \beta_1 \cdot Latitude_i + \beta_2 \cdot Longitude_i + \varepsilon_i$
for $i = 1, 2, ..., n$ and with ε_i the random error. **We assume the random errors to be independent normally distributed with mean** 0 **and variance** σ^2**.**

6.2 Multiple regression model in general

Multiple linear regression means that the model is linear in terms of the coefficients.

In general we consider a model with p independent variables: $x_1, x_2, ..., x_n$.

The response (dependent) variable Y can be modeled as:

$$Y_i = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + ... + \beta_1 p \cdot x_{pi} + \varepsilon_i$$

for $i = 1, 2, ..., n$

and where ε_i are independent normally distributed with mean 0 and constant variance σ^2.

In a multiple regression setting, we must be very careful with our interpretation of the coefficients.
The coefficient β_2 represents the expected change in the response for a one-unit change in x_2 given that all the other regressors are held constant.

Analog as in simple linear regression we have in multiple regression,
Total SS = Error SS + Model SS

In case the dataset has the same number of observations (n), then:

- Total SS (simple regression) = Total SS (multiple regression)
- Model SS (simple regression) < Model SS (multiple regression)
- Error SS (simple regression) > Error SS (multiple regression)

6.3 Evaluate multiple regression model

General population regression model with p explanatory variables:
$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$.

6.3.1 Is there regression?

H_0 : **There is no regression**
versus
H_1 : **There is regression**
This is similar to

$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$
versus
H_1 : **Not all** $\beta_j = 0$ **with** $j = 1, 2, ..., p$

The **test statistic**
$F = \frac{\text{Model SS}/p}{\text{Error SS}/(n-p-1)}$
has an F distribution with p and $n - p - 1$ degrees of freedom under the null hypothesis H_0.

6.3.2 How good is the regression?

The coefficient of determination, R^2 is used.
$R^2 = \frac{\text{Model SS}}{\text{Total SS}} = \frac{\text{Explained variation}}{\text{Total variation}}$

Next to the R^2 *value*, the *ADJ* R^2 *value* is important to evaluate the overall adequacy of the model.
$R^2_{adj} = 1 - \frac{n-1}{n-p} \cdot (1 - R^2)$

One problem with R^2 is that it cannot decrease as we add more independent variables to our model. The *ADJ* R^2 adjusts R^2 for the degrees of freedom we use for the model. As p gets larger, $n - p$ gets smaller. As we add terms to our model, *ADJ* R^2 will increase only if the new term reduces *Error SS* enough to compensate for the decrease in $n - p$.

6.3.3 Individual parameter estimates

Once we have determined that the response depends on at least one of the regressors, we need to determine which specific ones. Therefore t-tests are used.
$H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$

Under the null hypothesis,
$t_j = \frac{b_j}{se(b_j)}$
has a t-distribution with $n - p - 1$ degrees of freedom.

It is important to note that these hypotheses test the relationship given that the other regressors are held constant. We thus must use some care in interpreting the results of these tests!

6.4 Multiple regression model in R

Example *Temperature* in R

Population regression model: $annual = \alpha + \beta_1 \cdot Latitude + \beta_2 \cdot Longitude$

```
res.lm5 <- lm(annual ~ lat + long)
summary(res.lm5)
```

```
##
## Call:
## lm(formula = annual ~ lat + long)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9880 -0.6970  0.2245  0.6792  3.3085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.26034    1.86857   18.34  < 2e-16 ***
## lat         -0.46612    0.03943  -11.82 3.26e-13 ***
## long        -0.08723    0.02889   -3.02  0.00494 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.55 on 32 degrees of freedom
## Multiple R-squared:  0.856,  Adjusted R-squared:  0.847
## F-statistic: 95.14 on 2 and 32 DF,  p-value: 3.402e-14
```

1. Is there regression?
The F-test tests the hypothesis
$H_0 : \beta_1 = \beta_2 = 0$ versus $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$ (or both different from 0).

Since the $p - value < 0.0001$ we can reject the H_0.

2. How good is the regression?
R^2 has a value of 86%. This means, 86% of the total variability in *annual* can be explained by using the variables *Latitude* and *Longitude*. In the simple regression model (with only *Latitude*) we had a R^2 value of 0.815.

3. The obtained parameter estimates are:

- Estimate for the intercept $\alpha = 34$
- Estimate of β_1: (coeff. for *Latitude*) $= -0.47$
- Estimate of β_2: (coeff. for *Longitude*) $= -0.087$

These parameter estimates are all significant different from 0. We conclude this from the results of the t-test which all have $p - values < 0.05$.
The estimated regression model is:
$annual = 34 - 0.47 \cdot Latitude - 0.087 \cdot Longitude$

6.5 Interpretation of the coefficients

Be careful in the interpretation of the coefficients! Compare the estimates from the multiple regression model with those from the two simple regression models. This means that the interpretation of the coefficients is different.

Continuation example *Temperature* in R

	Simple model (with *Latitude*)	Simple model (with *Longitude*)	Multiple model (with *Latitude* and *Longitude*)
a	35	13	34
b_1 (*Latitude*)	-0.50		-0.47
b_2 (*Longitude*)		-0.19	-0.087

Interpretation of the value of the coefficients of the last column:

1. How to interpret $b_1 = -0.47$ for *Latitude*?
When comparing cities with the same *Longitude*, the expected annual temperature decreases with 0.47° for

each increase of 1° in *Latitude*.

Example

City	Latitude	Longitude	Expected annual temperature
Aberdeen (Scotland)	57	2	7.5
Barcelona (Spain)	41	2	14.9

Those cities have the same value for *Longitude* and a different value for *Latitude*. Their expected annual temperature is given by

$\hat{annual}(Aberdeen) = 34 - 0.47 \cdot 57 - 0.087 \cdot 2 = 7.5$ $\hat{annual}(Barcelona) = 34 - 0.47 \cdot 41 - 0.087 \cdot 2 = 14.9$

The difference is
$\hat{annual}(Aberdeen) - \hat{annual}(Barcelona) = -0.47(57 - 41) = -0.47 \cdot 16$

Hence, if we only compare cities with the same *Longitude*, then the expected annual temperature decreases with 0.47° for every increase of 1° in *Latitude*.

2. How to interpret $b_2 = -0.087$ for *Longitude*?
When comparing cities with the same *Latitude*, the expected annual temperature decreases with 0.087° for each increase of 1° in *Longitude*.

Example

City	Latitude	Longitude	Expected annual temperature
Berlin (Germany)	52	13	8.88
Birmingham (Great Britain)	52	1	9.94

Those cities have the same value for *Latitude* and a different value for *Longitude*. Their expected annual temperature is given by

$\hat{annual}(Berlin) = 34 - 0.47 \cdot 52 - 0.087 \cdot 13 = 8.88$ $\hat{annual}(Birmingham) = 34 - 0.47 \cdot 52 - 0.087 \cdot 1 = 9.94$

The difference is
$\hat{annual}(Berlin) - \hat{annual}(Birmingham) = -0.087(13 - 1) = -0.087 \cdot 12$

Hence, if we only compare cities with the same *Latitude*, then the expected annual temperature decreases with 0.087° for every increase of 1° in *Longitude*.

6.6 Checking assumptions in R

The multiple regression model makes assumptions. These assumptions need to be checked:

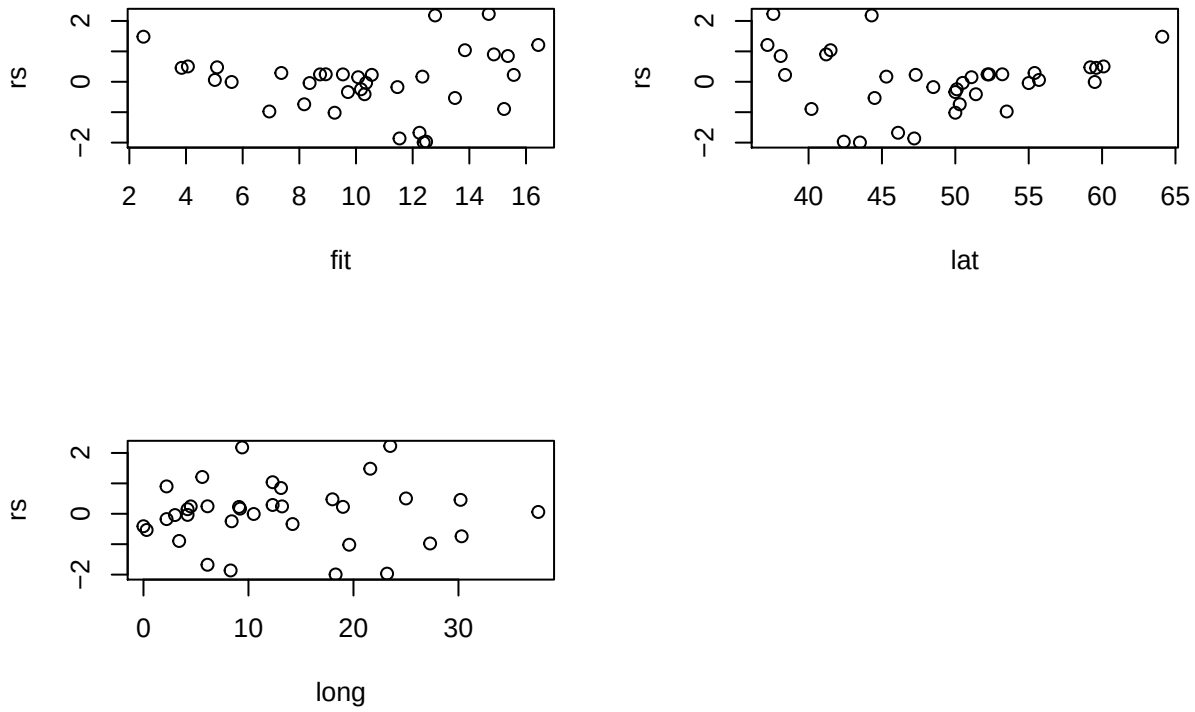
a. We have to check linearity.
   Make a plot of (standardized) residuals versus fitted response Make a plot of (standardized) residuals versus each regressor
b. Check normality of the (standardized) residuals Shapiro Wilk test (+ histogram)
c. Check for influential points Make a plot of the Cooks distance

6.6.1 Checking for linearity

Checking for linearity (checking whether the model is correct specified)

```
res.lm5 <- lm(annual ~ lat + long)
fit <- fitted(res.lm5)
rs <- rstandard(res.lm5)
```

```
par(mfrow=c(2,2))
plot(rs ~ fit)
plot(rs ~ lat)
plot(rs ~ long)
```



There is a random pattern, hence the linearity is satisfied.

6.6.2 Check normality of (standardized) residuals

H_0 : **Standard residuals are normal distributed**
versus
H_1 : **Standard residuals are not normal distributed**

```
shapiro.test(rs)
```

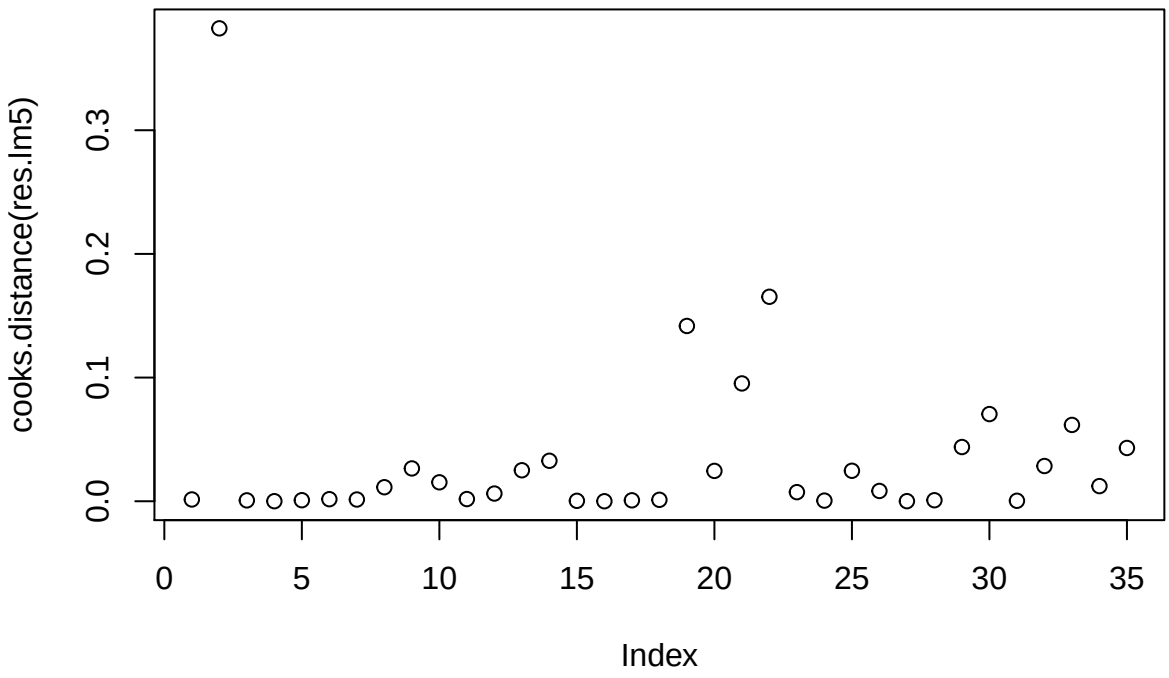
```
##
##  Shapiro-Wilk normality test
##
## data:  rs
## W = 0.96088, p-value = 0.2429
```

Normality can be assumed.

6.6.3 Check for influential points

Cook's distance

```
plot(cooks.distance(res.lm5))
```




```
temperature[which(cooks.distance(res.lm5) > 0.3),]
```

```
##     City January February March April  May June July August September October
## 2 Athens     9.1      9.7  11.7  15.4 20.1 24.5 27.4   27.2      23.8    19.2
##   November December annual1 Amplitude Latitude Longitude  Area annual warm
## 2     14.6       11    17.8      18.3     37.6      23.5 South   17.8    1
```

We can always check the influence of that observation on the regression model by performing the regression analysis without that point.

Delete point number 2 and redo the analysis:

```
annual_new <- annual[-2]
lat_new <- lat[-2]
long_new <- long[-2]
res.lm6 <- lm(annual_new ~ lat_new + long_new)
summary(res.lm6)
```

```
##
## Call:
## lm(formula = annual_new ~ lat_new + long_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7594 -0.7529  0.1360  0.6910  3.5072
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 32.73002    1.85869  17.609  < 2e-16 ***
## lat_new     -0.43170    0.03954 -10.918 3.78e-12 ***
## long_new    -0.10776    0.02831  -3.807 0.000624 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.447 on 31 degrees of freedom
## Multiple R-squared:  0.8635, Adjusted R-squared:  0.8547
## F-statistic: 98.03 on 2 and 31 DF,  p-value: 3.944e-14
```

$\hat{annual} = 32.7 - 0.43 \cdot Latitude - 0.10 \cdot Longitude$

	Complete data set	Reduced data set
n	35	34
a	34	32.7
b_1 ($Latitude$)	-0.47	-0.43
b_2 ($Longitude$)	-0.087	-0.107

That observation is not influential.

6.7 Predictions

Compute the expected annual temperature in following cities:

City	Latitude	Longitude
Aberdeen (Scotland)	57	2
Barcelona (Spain)	41	2
Berlin (Germany)	52	13
Birmingham (Great Britain)	52	1

```
new <- data.frame(list(lat = c(52, 52 ,57, 41), long = c(13, 1, 2, 2)))
res.pred5 <- predict(res.lm5, new, interval = "confidence")
city1 <- c("Berlin","Birmingham", "Aberdeen", "Barcelona")
pred5 <- data.frame(city1, res.pred5)
pred5
```

```
##        city1       fit       lwr       upr
## 1     Berlin  8.888340  8.304128  9.472552
## 2 Birmingham  9.935085  8.962138 10.908031
## 3   Aberdeen  7.517277  6.344936  8.689618
## 4  Barcelona 14.975130 14.048927 15.901334
```

7 Polynomial model

A **polynomial regression model** is a model where the response variable Y can be expressed as a linear function of powers of one or more independent variables.

7.1 A polynomial model with one regressor.

A polynomial model with one regressor can be expressed as

$$Y_i = \alpha + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + ... + \beta_p \cdot x_i^p + \varepsilon_i$$

The highest exponent of x is the order of the polynomial model. Usually one also keeps the terms of lower order into the model. Because it is linear in the coefficients, it can be considered as a multiple regression model. Sometimes, the interpretation is more difficult.

In the search for the optimal model one often starts with the most simply model and try to add successively a next power of x. It is important to consider the increase in R^2 and the decrease in the SS *error*.

7.2 A polynomial model in R

Example *Temperature* in R

```
res.lm7 <- lm(annual ~ lat + I(lat^2))
summary(res.lm7)
```

```
##
## Call:
## lm(formula = annual ~ lat + I(lat^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0061 -1.0920  0.2194  1.1137  3.6824
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.139111  12.474601   4.741 4.22e-05 ***
## lat         -1.503022   0.510750  -2.943  0.00601 **
## I(lat^2)     0.010121   0.005157   1.963  0.05843 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.66 on 32 degrees of freedom
## Multiple R-squared:  0.8349, Adjusted R-squared:  0.8246
## F-statistic: 80.91 on 2 and 32 DF,  p-value: 3.049e-13
```

The quadratic term seems not to be significant, hence we drop it from the model.

8 Interaction models

8.1 Illustrative example

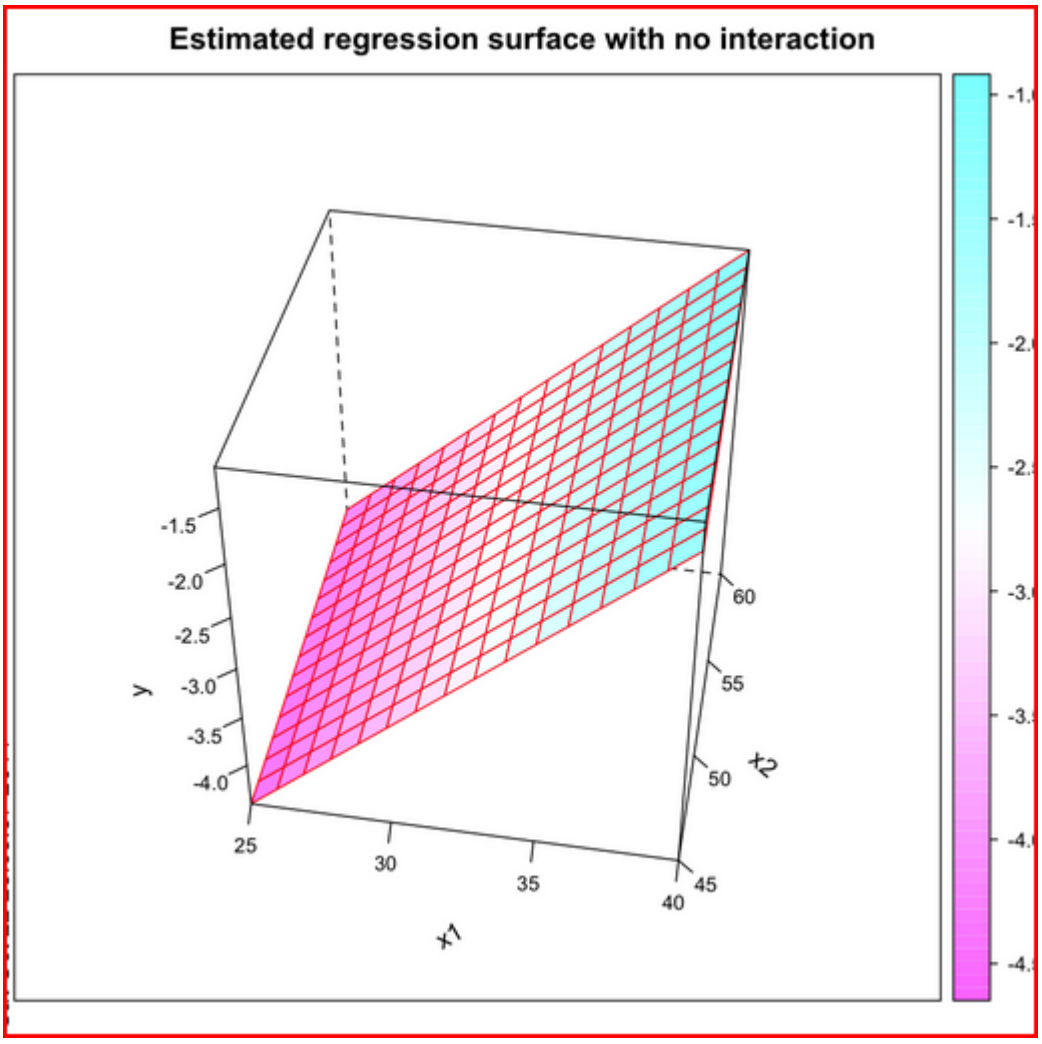
Example *Temperature*

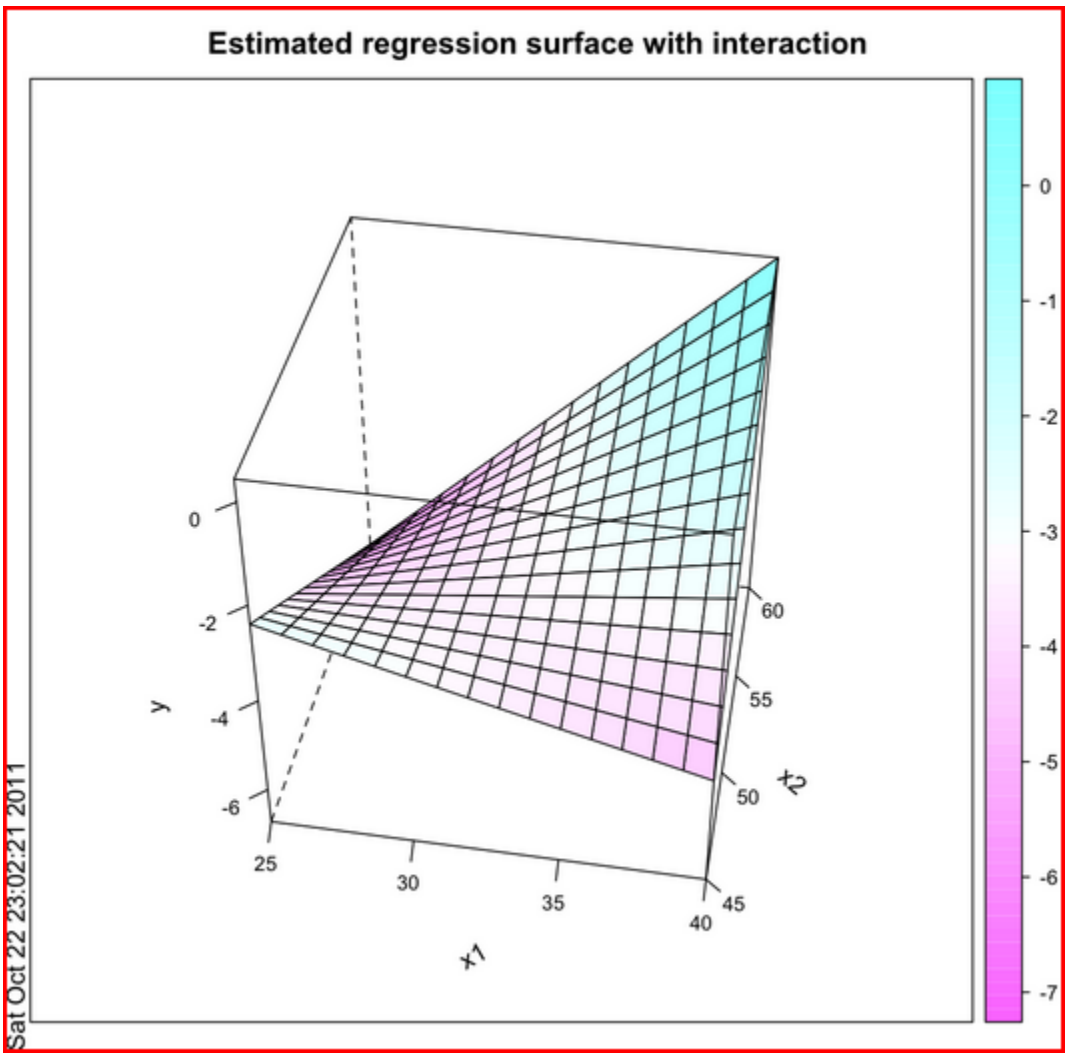
We retake the multiple regression model for annual temperature:
$\widehat{annual} = 34 - 0.47 \cdot Latitude - 0.087 \cdot Longitude$

1. This model assumes that the effect of *Latitude* on the annual temperature is independent from the value of *Longitude*. For each fixed value of *Longitude*, we know that an increase of 1° in *Latitude* will have a decrease in the average annual temperature of 0.47°.
2. This model assumes that the effect of *Longitude* on the annual temperature is independent form the value of *Latitude*. For each fixed value of *Latitude*, we know that an increase of 1° in *Longitude* will have a decrease in the average annual temperature of 0.087°.
3. A regression model without these assumptions is a model with the interaction term between *Latitude* and *Longitude*:
   $annual_i = \alpha + \beta_1 \cdot Latitude_i + \beta_2 \cdot Longitude_i + \beta_3 \cdot (Latitude_i \cdot Longitude_i) + \varepsilon_i$
4. We can e.g. rewrite this equation as a linear regression in *Longitude*. The intercept and the slope depends on *Latitude*.

$annual_i = [\alpha + \beta_1 \cdot Latitude_i] + [\beta_2 + \beta_3 \cdot Latitude_i] \cdot Longitude_i + \varepsilon_i$
with intercept: $\alpha + \beta_1 \cdot Latitude_i$
with slope: $\beta_2 + \beta_3 \cdot Latitude_i$

Notice that, although this model is a linear regression model, the shape of the surface is not linear!





8.2 In R

Example *Temperature* in R

with interaction term:

```
res.lm8 <- lm(annual ~ lat + long + I(lat*long))
summary(res.lm8)
```

```
##
## Call:
## lm(formula = annual ~ lat + long + I(lat * long))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8357 -0.7158  0.2229  0.5532  3.6156
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.961464   3.857578   9.582 8.81e-11 ***
## lat           -0.520978   0.079087  -6.587 2.33e-07 ***
```

```
## long          -0.272761   0.233220  -1.170    0.251
## I(lat * long)  0.003646   0.004547   0.802    0.429
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.559 on 31 degrees of freedom
## Multiple R-squared:  0.859,  Adjusted R-squared:  0.8453
## F-statistic: 62.94 on 3 and 31 DF,  p-value: 2.764e-13
```

The interaction term seems not to be significant here, hence we drop this term from the model. We then obtain the multiple regression model from previous pages.

$\hat{annual} = 34 - 0.47 \cdot Latitude - 0.087 \cdot Longitude$

9 Qualitative independent variable

9.1 Introduction

Example *Wine*

Import the data set *wine.txt* as `wine`.

```
head(wine)
```

```
##   flavor quality region indicator
## 1    3.9    13.6      3         1
## 2    4.5    14.4      3         1
## 3    4.3    12.3      2         0
## 4    7.0    16.1      3         1
## 5    6.7    16.1      3         1
## 6    5.8    15.5      3         1
```

This data set presents data on the quality of Pinot Noir wine. Is there a relationship between quality of the wine and the flavor of the wine? Wines from two different regions (region 2 (`indicator=0`) and region 3 (`indicator=1`)) are selected.

The population regression model is:
$quality = \alpha + \beta \cdot flavor$

```
quality <- wine$quality
flavor <- wine$flavor
res.lm <- lm(quality ~ flavor)
summary(res.lm)
```

```
##
## Call:
## lm(formula = quality ~ flavor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9494 -1.2568 -0.2273  0.9580  3.0050
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.3109     1.6714   1.981   0.0623 .
## flavor        1.8677     0.3222   5.797 1.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.511 on 19 degrees of freedom
## Multiple R-squared:  0.6388, Adjusted R-squared:  0.6198
## F-statistic:  33.6 on 1 and 19 DF,  p-value: 1.386e-05
```

$\widehat{quality} = 3.3 + 1.87 \cdot flavor$
$R^2 = 64\%$

Does the region (given by the indicator variable) has an impact on wine quality, when we have information about the flavor of the wine?

Remark:
Indicator variables (dummy's) can take the value 0 and 1. It takes the value 1 if the observation belongs to a certain category and it takes the value 0 else.

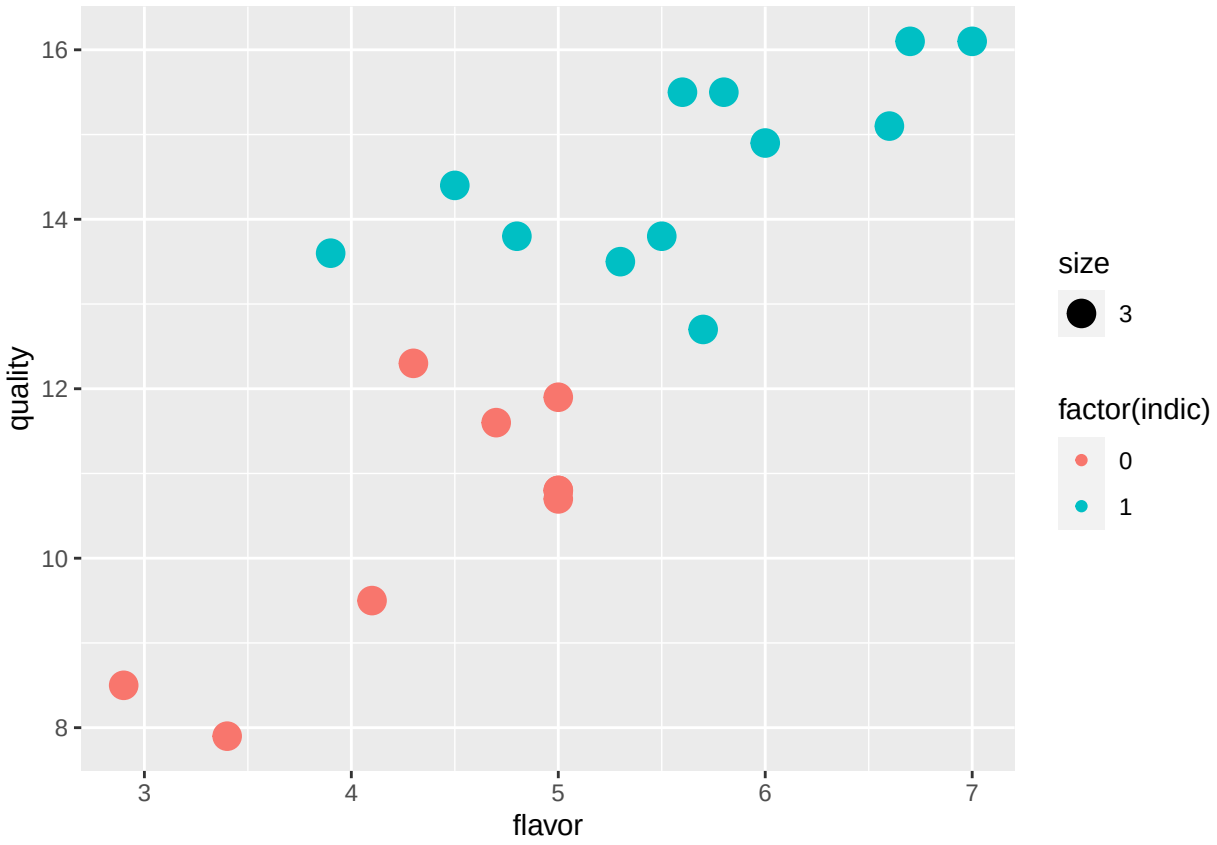
Examples:

- gender (1: woman, 0: man)
- higher education (1: yes, 0: no)
- drive license (1: yes, 0: no)

We visualize the relation between quality of the wine and flavor, taking into account the region.

```
flavor <- wine$flavor
quality <- wine$quality
indic <- wine$indicator

# visualise the data
library(ggplot2)
pl1 <- qplot(flavor, quality, colour=factor(indic), data= wine, size=3)
pl1
```



It seems that the quality for wines from region 1 is higher than for wines from region 0.

9.2 Regression model with indicator variable (but without interaction term)

$\widehat{quality} = \alpha + \beta_1 \cdot flavor + \beta_2 \cdot indicator$

```
res.lm1 <- lm(quality ~ flavor + indic)
summary(res.lm1)
```

```
##
## Call:
## lm(formula = quality ~ flavor + indic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97021 -0.50853 -0.08299  0.80642  1.93664
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.8803     1.1662   5.042 8.47e-05 ***
## flavor        1.0426     0.2561   4.070 0.000718 ***
## indic         2.8473     0.5296   5.376 4.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9615 on 18 degrees of freedom
## Multiple R-squared:  0.8614, Adjusted R-squared:  0.846
```

```
## F-statistic: 55.93 on 2 and 18 DF,  p-value: 1.889e-08
```

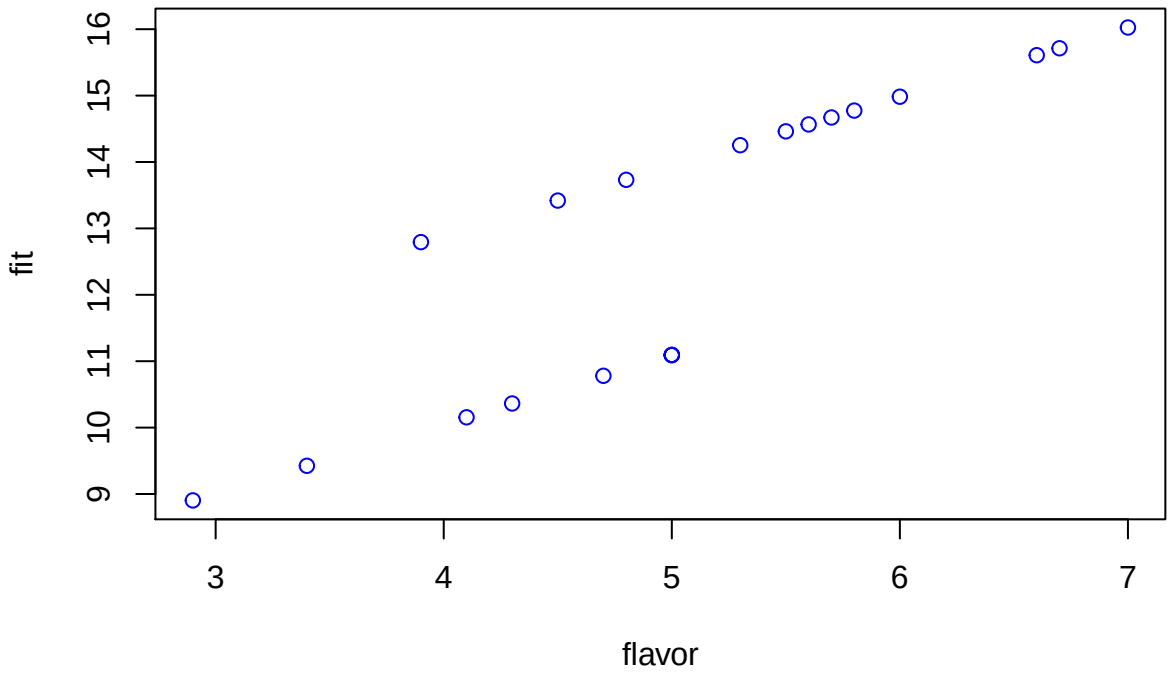
Interpretation:

1. *Is there regression?*
   F statistic is significant. H_0 is rejected.
2. *How good is the regression?*
   R^2 is 86%. R^2 has improved a lot compared with the simple linear regression model.

3. *Individual parameter estimates*
   $flavor$ and $indicator$ are significant different from 0.
   The estimated regression model is $\hat{quality} = 6 + 1.04 \cdot flavor + 2.8 \cdot indicator$
   - Regression line for wine from region 2 (`indicator = 0`)
     $\hat{quality} = 6 + 1.04 \cdot flavor$
   - Regression line for wine from region 3 (`indicator = 1`) $\hat{quality} = 6 + 1.04 \cdot flavor + 2.8 = 8.8 + 1.04 \cdot flavor$

We can visualize this as follows

```
fit <- fitted(res.lm1)
plot(flavor, fit, lty = 2, col = 4)
```



The curve (lowest) is the predicted regression line for region 2 ($indicator = 0$). The other curve is the predicted regression line for region 3 ($indicator = 1$). **There is only a difference in intercept.** The influence of flavor on quality is the same for both regions.

9.3 Regression model with indicator variable and interaction term

Now, assume we want to see whether the two regression lines possibly have also a different slope.
$quality = \alpha + \beta_1 \cdot flavor + \beta_2 \cdot indicator + \beta_3 \cdot (flavor \cdot indicator)$

- For wine from region 2 ($indicator = 0$):
  $quality = \alpha + \beta_1 \cdot flavor$
- For wine from region 3 ($indicator = 1$):
  $quality = (\alpha + \beta_2) + (\beta_1 + \beta_3) \cdot flavor$
  This is a regression model with intercept $(\alpha + \beta_2)$ and with slope $(\beta_1 + \beta_3)$.

Special cases:

- $\beta_2 = 0$
  Then it means that both regression lines have same intercept. This can be tested by
  $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$
- $\beta_3 = 0$
  Then it means that both regression lines have same intercept. This can be tested by
  $H_0 : \beta_3 = 0$ versus $H_1 : \beta_3 \neq 0$

Example *Temperature* in R
Model with interaction

```
res.lm2 <- lm(quality ~ flavor + indic + I(flavor * indic))
summary(res.lm2)
```

```
##
## Call:
## lm(formula = quality ~ flavor + indic + I(flavor * indic))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94964 -0.58359  0.01166  0.65470  1.97295
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.8369     1.8896   2.031  0.05824 .
## flavor              1.5093     0.4257   3.546  0.00248 **
## indic               6.2775     2.5822   2.431  0.02641 *
## I(flavor * indic)  -0.7137     0.5263  -1.356  0.19285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9399 on 17 degrees of freedom
## Multiple R-squared:  0.8749, Adjusted R-squared:  0.8528
## F-statistic: 39.64 on 3 and 17 DF,  p-value: 6.854e-08
```

The interaction term is not significant. When we drop the interaction term, we obtain the previous model again.

10 Exercises

10.1 Run test

We have some data available from a run test. There were 31 participants. The variables are:

`age`: age of the person

`weight`: weight of the person (in kg)
`runtime`: time necessary to run 1.5 miles
`rstpulse`: pulse rate in rest
`runpulse`: pulse rate at the end of the 1.5 miles
`maxpulse`: maximum pulse rate while running
`oxygen`: oxygen consumption, expressed in mL/kg/min

We are interested to model the oxygen consumption by using the other variables.

The data set `runtest.txt` is available on Toledo.

```
##    age weight runtime rstpulse maxpulse oxygen
## 1   57  73.37   12.63       58      176 39.407
## 2   54  79.38   11.17       62      165 46.080
## 3   52  76.32    9.63       48      166 45.441
## 4   50  70.87    8.92       48      155 54.625
## 5   51  67.25   11.08       48      172 45.118
## 6   54  91.63   12.88       44      172 39.203
## 7   51  73.71   10.47       59      188 45.790
## 8   57  59.08    9.93       49      155 50.545
## 9   49  76.32    9.40       56      188 48.673
## 10  48  61.24   11.50       52      176 47.920
## 11  52  82.78   10.50       53      172 47.467
## 12  44  73.03   10.13       45      168 50.541
## 13  45  87.66   14.03       56      192 37.388
## 14  45  66.45   11.12       51      176 44.754
## 15  47  79.15   10.60       47      164 47.273
## 16  54  83.12   10.33       50      170 51.855
## 17  49  81.42    8.95       44      185 49.156
## 18  51  69.63   10.95       57      172 40.836
## 19  51  77.91   10.00       48      168 46.672
## 20  48  91.63   10.25       48      164 46.774
## 21  49  73.37   10.08       76      168 50.388
## 22  44  89.47   11.37       62      182 44.609
## 23  40  75.07   10.07       62      185 45.313
## 24  44  85.84    8.65       45      168 54.297
## 25  42  68.15    8.17       40      172 59.571
## 26  38  89.02    9.22       55      180 49.874
## 27  47  77.45   11.63       58      176 44.811
## 28  40  75.98   11.95       70      180 45.681
## 29  43  81.19   10.85       64      170 49.091
## 30  44  81.42   13.08       63      176 39.442
## 31  38  81.87    8.63       48      186 60.055
```

10.2 Cholesterol data

We are interested to know how age, height and weight influence the cholesterol value of a man. 200 men are selected and the age, height, weight and cholesterol is registered. Analyze these data. Predict the cholesterol value for a person with an age of 40, a height of 170 and a weight of 65.

The data file`Chol.txt` is available on Toledo.