Chapter 5

Simple Linear regression

Contents

5.1 Linear regression and least squares

In engineering applications we are often faced with the problem of determining the best model of some outcome given a known input

$$y = f(x), \tag{5-1}$$

hence x is the input and the function f is the model. The task is now to find the best model given the input variables (x) and the outcome (y). The simplest model, besides just a mean value (covered in Chapters 3 and 4), would be a model where f is a linear function of x

$$y = \beta_0 + \beta_1 x. \tag{5-2}$$

When the outcome y is the result of some experiment, the model will not be perfect, and we need to add an error term

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = \{1, \ldots, n\}, \tag{5-3}$$

where ε_i is called the *error* and is a (independent) random variable with expectation equal zero (i.e. the mean $\mathrm{E}(\varepsilon_i) = 0$ and some variance ($\mathrm{V}(\varepsilon_i) = \sigma^2$). The statistical interpretation of (5-2) is therefore that it expresses the expected value of the outcome

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i, \tag{5-4}$$

also called the *model prediction*.

It is of course a very unusual situation that we actually know the values of β_0 and β_1 and we will have to rely on estimates based on some observations ($y_1, \ldots, y_n$). As usual we express this by putting a "hat" on the parameters

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{5-5}$$

meaning that we expect or predict $\hat{y}_i$ (in mean or average) under the conditions given by x_i.

|||| Example 5.1

A car manufacturer wants to find the relation between speed and fuel consumption, to do so she sets up the following model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{5-6}$$

here $\mathrm{E}(Y_i)$ is the expected fuel consumption at the speed x_i. Further, there will be uncontrollable variations, e.g. due to differences in weather condition, but also non-linear effects not included in the model might be present. These variations are captured by the ε_i's. We see that speed is something we control here, and we then observe the outcome (here fuel consumption), at different experimental conditions (speeds).

In this chapter we will deal with estimation and inference of β_0, β_1, and prediction of Y_i given x_i. At some point we will have realizations (or observations) of the outcome, in this case we write

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = \{1, \ldots, n\}. \tag{5-7}$$

Now y_i is a realization and e_i is the deviation between the model prediction and the actual observation: a realization of the error ε_i, it is called a *residual*. Clearly, we want the residuals to be small in some sense, the usual choice (and the one treated in this chapter) is in the Residual Sum of Squares (RSS) sense, i.e. we want to minimize the residual sum of squares

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2, \tag{5-8}$$

where we have emphasized that the residual sum of squares is a function of the parameters (β_0, β_1). The parameter estimates $(\hat{\beta}_0, \hat{\beta}_1)$ are the values of β_0 and β_1 which minimize RSS. Note, that we use Y_i and ε_i rather than the observed values (y_i and e_i), this is to emphasize that the estimators are random variables, in actual calculations after the experiments are carried out we will just replace Y_i with y_i and ε_i with e_i. Figure 5.1 sketches the linear regression problem.

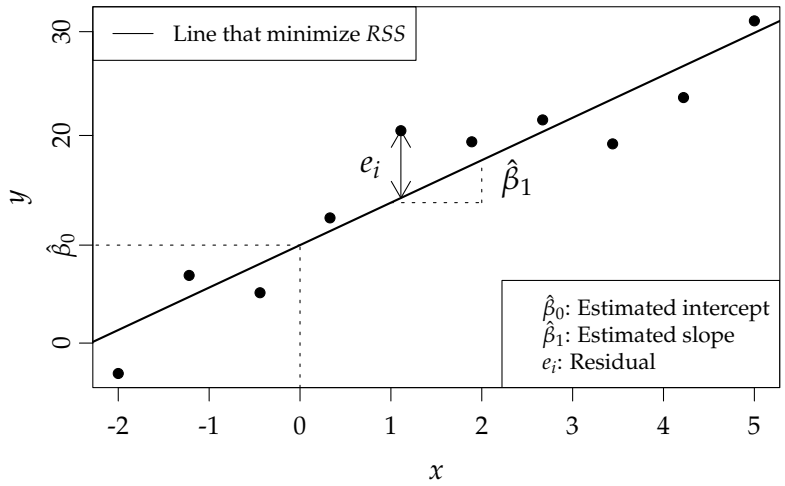


Figure 5.1: Conceptual diagram for the simple linear regression problem.

|||| Remark 5.2 Estimates and estimators

In (5-8) the *RSS* is a function of the random variables (Y_i), thus making *RSS* a random variable. If we replace Y_i with the realizations y_i then *RSS* is also a realization.
In this chapter the result of optimizing *RSS* with respect to β_0 and β_1 will be denoted $\hat{\beta}_0$ and $\hat{\beta}_1$. Sometimes $\hat{\beta}_0$ and $\hat{\beta}_1$ will be functions of Y_i and sometimes they will be functions of the realizations y_i, they are referred to as:

1. **Estimators**: before the experiment has been carried out, then $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of Y_i and they are also random variables, and we call them *estimators*.

2. **Estimates**: after the experiment had been carried out, then $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of y_i and they are also realizations of random variables, and we call them *estimates*.

Remark 5.3 Two types of examples

In this chapter we will use two types of examples, one is labelled "Simulation", which are simulation studies intended to illustrate the consequences of theorems and general results. While the other type of examples (not labelled "Simulation"), are intended to illustrate the use of the theorems on pratical examples.

5.2 Parameter estimates and estimators

When $\hat{\beta}_0$ and $\hat{\beta}_1$ is a result of minimizing the function in Equation (5-8), we refer to the estimators as *least squares estimators*. The least squares estimators are given in the following theorem:

Theorem 5.4 Least squares estimators

The least squares estimators of β_0 and β_1 are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}}, \tag{5-9}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \tag{5-10}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

As we can see above the estimators ($\hat{\beta}_1$ and $\hat{\beta}_2$) are functions of random variables (Y_i and $\bar{Y}$), and thus the estimators are themselves random variables. We can therefore talk about the expectation, variance and distribution of the estimators. In analyses with data we will of course only see realizations of Y_i and we just replace Y_i and $\bar{Y}$ with their realizations y_i and $\bar{y}$. In this case we speak about *estimates* of β_0 and β_1.

Before we go on with the proof of Theorem 5.4, the application of the theorem is illustrated in the following example:

Example 5.5 Student height and weight

Consider the student height and weight data presented in Chapter 1,

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

We want to find the best least squares regression line for these points, this is equivalent to calculating the least squares estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$.

We start by finding the two sample means

$$\bar{x} = \frac{1}{10}(168 + 161 + \ldots + 179) = 178,$$
$$\bar{y} = \frac{1}{10}(65.5 + 58.3 + \ldots + 78.9) = 78.11.$$

The value of S_{xx} is calculated by

$$S_{xx} = (168 - 178)^2 + \ldots + (179 - 178)^2 = 1342.$$

We can now calculate $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \frac{1}{1342}\big((65.5 - 78.11)(168 - 179) + \ldots + (79.9 - 78.11)(179 - 178)\big) = 1.11,$$

and finally, we can calculate $\hat{\beta}_0$ as

$$\hat{\beta}_0 = 78.11 - 1.11 \cdot 178 = -120.$$

In R the calculation above can be done by:

```
# Read data
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)

# Calculate averages
xbar <- mean(x)
ybar <- mean(y)

# Parameters estimates
Sxx <- sum((x - xbar)^2)
beta1hat <- sum((x - xbar)*(y - ybar)) / Sxx
beta0hat <- ybar - beta1hat * xbar
```

Rather than using "manual" calculations in R, we can use the build in R-function `lm`

```
D <- data.frame(x=x, y=y)
fitStudents <- lm(y ~ x, data=D)
summary(fitStudents)


Call:
lm(formula = y ~ x, data = D)

Residuals:
   Min     1Q Median     3Q    Max
-5.876 -1.451 -0.608  2.234  6.477

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -119.958     18.897   -6.35   0.00022 ***
x              1.113      0.106   10.50 0.0000059 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.88 on 8 degrees of freedom
Multiple R-squared:  0.932,Adjusted R-squared:  0.924
F-statistic:  110 on 1 and 8 DF,  p-value: 0.00000587
```

As we can see the two calculations give the same results regarding the parameter estimates. We can also see that the direct calculation in `R` (`lm`) gives some more information. How to interpret and calculate these numbers will be treated in the following pages.

Before we go on with the analysis of the result from `lm`, the proof of Theorem 5.4 is presented:

Proof

Of Theorem 5.4: In order to find the minimum of the function RSS we differentiate the residual sum of squares with respect to the parameters

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)), \tag{5-11}$$

now equating with zero we get

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) \\ &= -2n\bar{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\bar{x}, \end{aligned} \tag{5-12}$$

solving for $\hat{\beta}_0$ gives

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{5-13}$$

and by similar calculations we get

$$\begin{aligned} \frac{\partial RSS}{\partial \hat{\beta}_1} &= \frac{\partial}{\partial \hat{\beta}_1} \left(\sum_{i=1}^{n} (y_i - (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i))^2 \right) \\ &= \frac{\partial}{\partial \hat{\beta}_1} \left(\sum_{i=1}^{n} ((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}))^2 \right) \\ &= -2 \sum_{i=1}^{n} ((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}))(x_i - \bar{x}) \\ &= -2 \left[\sum_{i=1}^{x} (y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}_1 \sum_{i=1}^{x} (x_i - \bar{x})^2 \right], \end{aligned} \tag{5-14}$$

equating with zero and solving for $\hat{\beta}_1$ gives

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{S_{xx}}. \end{aligned} \tag{5-15}$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are called **least squares estimates**, because they minimize the sum of squared residuals (i.e. RSS). Replacing y_i with Y_i give the estimators in the theorem.

■

When we have obtained parameter estimates in the linear regression model above, we would like to make quantitative statements about the uncertainty of the parameters, and in order to design tests we will also need the probability distribution of the parameter estimators. The usual assumption is that the errors are normal random variables

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{where } \varepsilon_i \sim N(0, \sigma^2), \tag{5-16}$$

or in other words the errors are independent identically distributed (i.i.d.) normal random variables with zero mean and variance σ^2. When random variables are involved we know that repeating the experiment will result in different values of the response (Y_i), and therefore in different values of the parameter estimates. To illustrate this we can make simulation experiments to analyse the behaviour of the parameter estimates. Recall that the role of simulation examples are to illustrate probabilistic behaviour of e.g. estimators, not how actual data is analysed.

Remark 5.6 How to write a statistical model

In Remark 3.2 it was explained how to write the model behind the t-tests, i.e.

$$X_i \sim N(\mu, \sigma^2) \text{ and i.i.d.} \tag{5-17}$$

Remember, that i.i.d. is short for independently and identically distributed, which essentially means that the observations are selected randomly from population, see the text after Example 1.2.

Using this notation the linear regression model could be written

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \text{ and i.i.d.,} \tag{5-18}$$

however we will write models as above in Equation 5-16.
Note, if $\beta_1 = 0$ the model is

$$Y_i = \beta_0 + \varepsilon_i, \quad \text{where } \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.,} \tag{5-19}$$

which is exactly the model above in Equation 5-17, and the estimate of the mean of the population, from which the sample (i.e. $(y_1, \ldots, y_n)$) was taken, is then

$$\hat{\mu} = \hat{\beta}_0. \tag{5-20}$$

Example 5.7 Simulation of parameter estimation

Consider the linear model

$$Y_i = 10 + 3x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, 5^2) \tag{5-21}$$

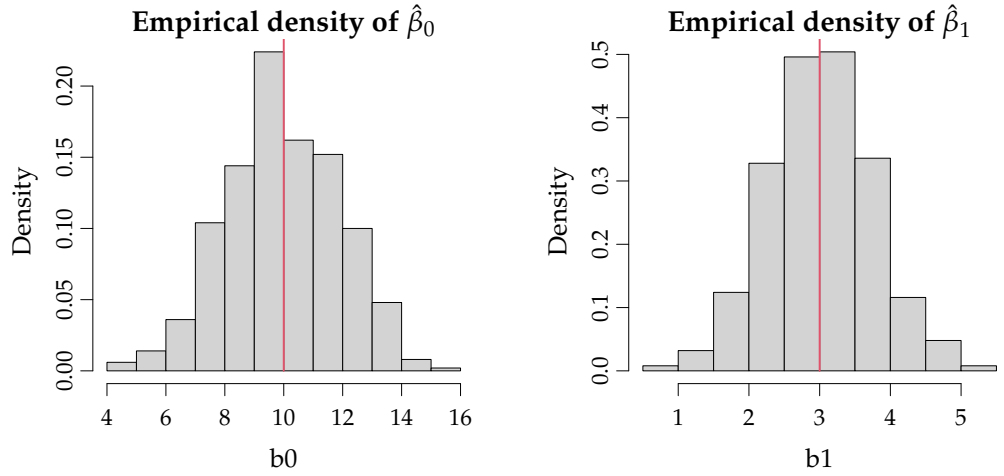
We can make repetitions of this experiment in R

```
n <- 10; k <- 500
beta0 <- 10;  beta1 <- 3;  sigma <- 5
x <- seq(-2, 5, length=n)
y <- matrix(0, ncol=k, nrow=n)
y <- y + beta0 + beta1*x + rnorm(n*k, sd=sigma)
```

The variable y now contains n rows and k columns, representing k experiments, for each of the experiment we can calculate the parameter estimates:

```
b0 <- numeric(k);  b1 <- numeric(k)
for(i in 1:k){
  b <- coef(lm(y[ ,i] ~ x))
  b0[i] <- b[1]
  b1[i] <- b[2]
}
c(mean(b0),  mean(b1))

[1] 9.955 3.008
```

As we can see the average of the parameter estimates (`mean(b0)` and `mean(b1)`) are very close to the true parameter values ($\beta_0 = 10$ and $\beta_1 = 3$). We can of course also look at the empirical density (the normalized histogram, see Section 1.6.1) of the parameter estimates:



The estimates seem to be rather symmetrically distributed around the true parameter values. It is also clear that there is some variation in the estimates: the estimates of β_0 range from about 4 to about 16 and the estimates of β_1 range from about 1 to 5.

Try changing the R code (see the accompanying chapter script):

What happens to the mean value of the estimates if you change the number of data points (n)?

What happens to the empirical density and the scatter plot of the parameter estimates if you change:

- The number of data points (n)?
- The range of x-values?
- The residual variance (σ^2)?
- The values of β_0 and β_1?

In the example above we saw that the average of the parameter estimates were very close to the true values, this is of course a nice property of an estimator. When this is the case in general, i.e. when $\mathrm{E}[\hat{\beta}_i] = \beta_i$ we say that the estimator is central. The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are in fact central, and we show this in Section 5.2.1 below.

In order to test hypothesis about β_0 and β_1 we will also need to give exact statements about the distribution of the parameters. We saw in Example 5.7 above that the distributions seem to be symmetric around the true values, but we will need more precise statements about the distributions and their variances. This important part will be dealt with in the Sections 5.3 and 5.4.

5.2.1 Estimators are central

In the linear regression model we assume that the observed values of Y_i can be split into two parts: the prediction (the part explained by the regression line $(\beta_0 + \beta_1 x_i)$) and the error (a random part (ε_i)). As usual we view our estimators as functions of random variables (the ε_i's), so it makes sense to calculate the expectation of the estimators. The assumption $\mathrm{E}(\varepsilon_i) = 0$ is central for the presented arguments, and will be used repeatedly.

In order to find the expectation of the parameter estimators we rewrite our estimators as functions of the true parameters (β_0 and β_1)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}}, \tag{5-22}$$

inserting $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$ gives

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \left[(\beta_0 + \beta_1 x_i + \varepsilon_i - (\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon})\right](x_i - \bar{x})}{S_{xx}}, \tag{5-23}$$

now the sum is divided into a part which depends on ε_i (the random part) and a part which is independent of ε_i

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n \beta_1 (x_i - \bar{x})^2}{S_{xx}} + \frac{\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x})}{S_{xx}} \\ &= \beta_1 + \frac{\sum_{i=1}^n \varepsilon_i (x_i - \bar{x})}{S_{xx}} - \frac{\bar{\varepsilon}\sum_{i=1}^n (x_i - \bar{x})}{S_{xx}}, \end{aligned} \tag{5-24}$$

now observe that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ to get

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} \varepsilon_i (x_i - \bar{x})}{S_{xx}}, \tag{5-25}$$

for $\hat{\beta}_0$ we get

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + \varepsilon_i) - \left(\beta_1 + \frac{\sum_{i=1}^{n} \varepsilon_i (x_i - \bar{x})}{S_{xx}}\right)\bar{x} \\
&= \beta_0 + \beta_1 \bar{x}_i + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i - \left(\beta_1 + \frac{\sum_{i=1}^{n} \varepsilon_i (x_i - \bar{x})}{S_{xx}}\right)\bar{x} \\
&= \beta_0 + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i - \left(\frac{\sum_{i=1}^{n} \varepsilon_i (x_i - \bar{x})}{S_{xx}}\right)\bar{x}.
\end{aligned} \tag{5-26}$$

Since expectation is a linear operation (see Chapter 2) and the expectation of ε_i is zero we find that $\mathrm{E}[\hat{\beta}_0] = \beta_0$ and $\mathrm{E}[\hat{\beta}_1] = \beta_1$, and we say that $\hat{\beta}_0$, $\hat{\beta}_1$ are central estimators.

5.3 Variance of estimators

In order for us to be able to construct confidence intervals for parameter estimates, talk about uncertainty of predictions and test hypothesis, then we will need the variance of the parameter estimates as well as an estimator of the error variance (σ^2).

Parameter variance and covariance of estimators are given in the following theorem:

|||| Theorem 5.8 Variance of estimators

The variance and covariance of the estimators in Theorem 5.4 are given by

$$\mathrm{V}[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}}, \tag{5-27}$$

$$\mathrm{V}[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}, \tag{5-28}$$

$$\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}\sigma^2}{S_{xx}}, \tag{5-29}$$

where σ^2 is usually replaced by its estimate ($\hat{\sigma}^2$). The central estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2}. \tag{5-30}$$

When the estimate of σ^2 is used the variances also become estimates and we'll refer to them as $\hat{\sigma}^2_{\hat{\beta}_0}$ and $\hat{\sigma}^2_{\hat{\beta}_1}$.

The variance of $\hat{\beta}_1$ is a function of the true error variance (σ^2) and S_{xx}. For most (all reasonable) choices of the regressors (x), S_{xx} will be an increasing function of n, and the variance of $\hat{\beta}_1$ will therefore decrease as n increases. This expresses that we will be more certain about the estimates as we increase the number of points in our sample. The same is true for the variance of $\hat{\beta}_0$, and the covariance between $\hat{\beta}_1$ and $\hat{\beta}_0$. The error variance estimate ($\hat{\sigma}$) is the residual sum of squares divided by $n-2$, the intuitive explanation for the $n-2$ (rather than n or $n-1$) is that if we only have two pairs (x_i, y_i, i.e. $n = 2$), it will not be possible to say anything about the variation (the residuals will be zero). Or another phrasing is that; we have used 2 degrees of freedom to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$.

Before we turn to the proof of Theorem 5.8, we will take a look at a couple of examples.

|||| Example 5.9 (Example 5.5 cont.)

In Example 5.5 we found the parameter estimates

$$\hat{\beta}_0 = -119.96, \quad \hat{\beta}_1 = 1.113,$$

we can now find predicted values of the dependent variable by

$$\hat{y}_i = -119.96 + 1.114 \cdot x_i,$$

and the values of the residuals

$$e_i = y_i - \hat{y}_i,$$

and finally the error variance estimate is

$$\hat{\sigma}^2 = \frac{1}{10-2}\sum_{i=1}^{10} e_i^2.$$

In R we can find the results by:

```
beta0 <- coef(fitStudents)[1]
beta1 <- coef(fitStudents)[2]
e <- y - (beta0 + beta1 * x)
n <- length(e)
sigma <- sqrt(sum(e^2) / (n - 2))
sigma.beta0 <- sqrt(sigma^2 * (1 / n + xbar^2 / Sxx))
sigma.beta1 <- sqrt(sigma^2 / Sxx)
c(sigma, sigma.beta0, sigma.beta1)

[1] 132.946 645.983   3.629
```

As usual we use standard deviations rather than variances, this also means that we can compare with the results from `lm` (see Example 5.5). Again we can find our estimates in the R-output, the parameter standard deviations are given in the second column of the coefficient matrix and the estimated standard deviation of the error is called `residual standard error`.

The simulation example (Example 5.7) can also be extended to check the equations of Theorem 5.8:

|||| Example 5.10 Simulation

In Example 5.7 we looked at simulation from the model

$$Y_i = 10 + 3x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, 5^2)$$

In order to calculate the variance estimates we need to calculate $\bar{x}$ and S_{xx}:

```
Sxx <- (n-1)*var(x)
c(mean(x), Sxx)

[1]  1.50 49.91


y <- matrix(0, ncol=k, nrow=n)
```

and we would expect to obtain the variance estimates close to

$$\mathrm{V}[\hat{\beta}_0] = 5^2 \left(\frac{1}{10} + \frac{1.50^2}{49.91} \right) = 3.63$$
$$\mathrm{V}[\hat{\beta}_1] = \frac{5^2}{49.91} = 0.501$$

With simulations we find:

```
b0 <- numeric(k);  b1 <- numeric(k)
sigma <- numeric(k)
for(i in 1:k){
  fit <- lm(y[ ,i] ~ x)
  b <- coef(fit)
  b0[i] <- b[1]
  b1[i] <- b[2]
  sigma[i] <- summary(fit)$sigma
 }
 c(var(b0), var(b1), mean(sigma))

[1] 3.7755 0.5427 4.8580
```

We can see that the simulated values are close to the theoretical values. You are invited to play around with different settings for the simulation, in particular increasing k will increase the accuracy of the estimates of the variances and covariance.

The example above shows how the Theorem 5.8 can be illustrated by simulation, a formal proof is given by:

Proof

Of Theorem 5.8. Using (5-26) we can write the variance of $\hat{\beta}_0$ as

$$\mathrm{V}(\hat{\beta}_0) = \mathrm{V}\left[\beta_0 + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i - \left(\frac{\sum_{i=1}^{n} \varepsilon_i (x_i - \bar{x})}{S_{xx}} \right) \bar{x} \right], \tag{5-31}$$

using the definition of the variance ($\mathrm{V}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$) and $\mathrm{E}(\varepsilon) = 0$ we get

$$\begin{aligned} \mathrm{V}(\hat{\beta}_0) = \mathrm{V}\left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \right] + \mathrm{V}\left[\left(\frac{\sum_{i=1}^{n} \varepsilon_i (x_i - \bar{x})}{S_{xx}} \right) \bar{x} \right] - \\ 2\,\mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(\frac{\sum_{i=1}^{n} \varepsilon_i (x_i - \bar{x})}{S_{xx}} \right) \bar{x} \right], \end{aligned} \tag{5-32}$$

now use independence between ε_i and ε_j ($i \neq j$) to get

$$\begin{aligned} \mathrm{V}(\hat{\beta}_0) &= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2 \sum_{i=1}^n (x_i - \bar{x})^2}{(S_{xx})^2} + \frac{\bar{x}\sigma^2}{nS_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \\ &= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}}. \end{aligned} \tag{5-33}$$

Finally, the variance of $\hat{\beta}_1$ is (again using the definition of variance and independence of the ε's)

$$\begin{aligned} \mathrm{V}(\hat{\beta}_1) &= \mathrm{V}\left[\beta_1 + \frac{\sum_{i=1}^n \varepsilon_i (x_i - \bar{x})}{S_{xx}}\right] \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \, \mathrm{V}(\varepsilon_i)}{(S_{xx})^2} \\ &= \frac{\sigma^2}{S_{xx}}, \end{aligned} \tag{5-34}$$

and the covariance between the parameters estimates becomes

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= \mathrm{E}\left[(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)\right] \\ &= \mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i - \frac{\sum_{i=1}^n \varepsilon_i (x_i - \bar{x})}{S_{xx}}\right) \bar{x} \frac{\sum_{i=1}^n \varepsilon_i (x_i - \bar{x})}{S_{xx}}\right] \\ &= \frac{\bar{x}}{nS_{xx}} \mathrm{E}\left[\sum_{i=1}^n \varepsilon_i \sum_{i=1}^n \varepsilon_i (x_i - \bar{x})\right] - \frac{\bar{x}}{(S_{xx})^2} \mathrm{E}\left[\sum_{i=1}^n \varepsilon_i^2 (x_i - \bar{x})^2\right] \\ &= \frac{\bar{x}\sigma^2(n\bar{x} - n\bar{x})}{nS_{xx}} - \frac{\bar{x}}{(S_{xx})^2}\sigma^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= -\frac{\bar{x}\sigma^2}{S_{xx}}. \end{aligned} \tag{5-35}$$

To get an estimate of the residual variance we calculate the expected value of the residual sum of squares

$$\mathrm{E}(RSS) = \mathrm{E}\left[\sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2\right], \tag{5-36}$$

inserting $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and rearranging gives

$$\begin{aligned} \mathrm{E}(RSS) &= \sum_{i=1}^n \mathrm{E}\left[(-(\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)x_i + \varepsilon_i)^2\right] \\ &= \sum_{i=1}^n \left\{\mathrm{E}\left[(\hat{\beta}_0 - \beta_0)^2\right] + \mathrm{E}\left[(\hat{\beta}_1 - \beta_1)^2\right] x_i^2 + \mathrm{E}[\varepsilon_i^2] + \right. \\ &\quad \left. 2\,\mathrm{E}\left[(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)\right] x_i - 2\,\mathrm{E}\left[(\hat{\beta}_0 - \beta_0)\varepsilon_i\right] - 2\,\mathrm{E}\left[(\hat{\beta}_1 - \beta_1)\varepsilon_i\right] x_i \right\}, \end{aligned} \tag{5-37}$$

now observe that $\mathrm{E}\left[(\hat{\beta}_0-\beta_0)^2\right] = \mathrm{V}[\hat{\beta}_0]$, $\mathrm{E}\left[(\hat{\beta}_1-\beta_1)^2\right] = \mathrm{V}[\hat{\beta}_1]$, $\mathrm{E}(\varepsilon_i^2) = \sigma^2$, and $\mathrm{E}\left[(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)\right] = \mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1)$, and insert $\hat{\beta}_0$ and $\hat{\beta}_1$ in the last two terms

$$\begin{aligned}
\mathrm{E}(RSS) &= n\,\mathrm{V}(\hat{\beta}_0) + \mathrm{V}(\hat{\beta}_1)\sum_{i=1}^{n} x_i^2 + n\sigma^2 + 2\sum_{i=1}^{n} Cov\left(\hat{\beta}_0,\hat{\beta}_1\right) x_i - \\
&\quad 2\sum_{i=1}^{n}\left\{\mathrm{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j - \frac{\sum_{j=1}^{2}\varepsilon_j(x_j-\bar{x})}{S_{xx}}\right)\varepsilon_i\right] - \mathrm{E}\left[\frac{\sum_{j=1}^{n}\varepsilon_j(x_j-\bar{x})}{S_{xx}}\varepsilon_i\right] x_i\right\} \\
&= \sigma^2 + \frac{n\bar{x}^2\sigma^2}{S_{xx}} + \frac{\sigma^2\sum_{i=1}^{n} x_i^2}{S_{xx}} + n\sigma^2 - 2\sum_{i=1}^{n}\frac{\bar{x}\sigma^2}{S_{xx}}x_i - \\
&\quad 2\sum_{i=1}^{n}\left(\frac{\sigma^2}{n} - \frac{\sigma^2(x_i-\bar{x})}{S_{xx}}\right) - 2\sum_{i=1}^{n}\frac{\sigma^2(x_i-\bar{x})x_i}{S_{xx}},
\end{aligned} \tag{5-38}$$

now collect terms and observe that $\sum x_i = n\bar{x}$

$$\begin{aligned}
\mathrm{E}(RSS) &= \sigma^2(n+1) + \frac{\sigma^2}{S_{xx}}\sum_{i=1}^{n}(x_i^2+\bar{x}^2) - 2\frac{n\bar{x}^2\sigma^2}{S_{xx}} - 2\sigma^2 - 2\frac{\sigma^2\sum_{i=1}^{n}(x_i^2 - x_i\bar{x})}{S_{xx}} \\
&= \sigma^2(n-1) + \frac{\sigma^2}{S_{xx}}\sum_{i=1}^{n}(-x_i^2 - \bar{x}^2 + 2x_i\bar{x}) \\
&= \sigma^2(n-1) - \frac{\sigma^2}{S_{xx}}S_{xx} \\
&= \sigma^2(n-2),
\end{aligned} \tag{5-39}$$

and thus a central estimator for σ^2 is $\hat{\sigma}^2 = \frac{RSS}{n-2}$.

■

Before we continue with parameter distributions and hypothesis testing, the next example illustrates the behaviour of the parameter variance estimates:

|||| Example 5.11 Simulation

Consider the following model

$$Y_i = 1 + x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,1), \tag{5-40}$$

also assume that $x_i = \frac{i-1}{n-1}$, $i = 1,\ldots,n$ where n is the number of pairs (x_i, y_i). We want to make a simulation experiment for increasing number of pairs, and extract the parameter variance, parameter covariance and residual variance estimates. In order to do so we need to extract these numbers from a linear model i R. This can be done by:

```r
x <- seq(0, 1, length=10)
y <- 1 + x + rnorm(10)
# Fit the model (estimate parameter)
fit <- lm(y ~ x)
# Print summary of model fit
summary(fit)
```

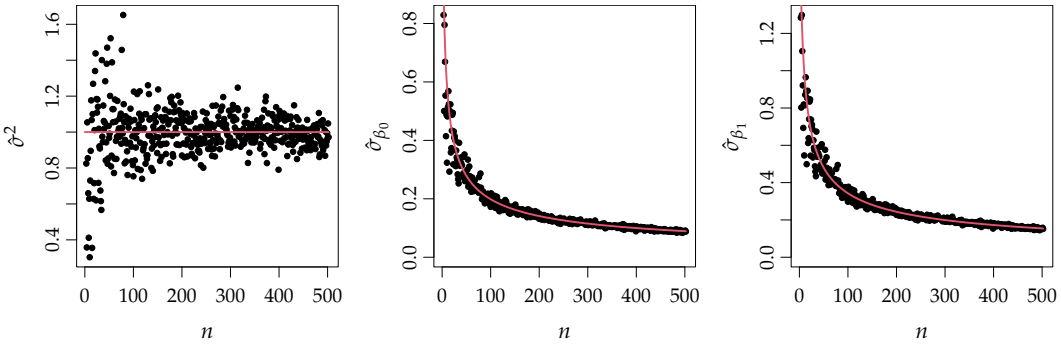
```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-0.867 -0.596  0.232  0.374  1.295

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.454      0.434    1.05   0.3256
x              2.521      0.731    3.45   0.0087 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.738 on 8 degrees of freedom
Multiple R-squared:  0.598,Adjusted R-squared:  0.548
F-statistic: 11.9 on 1 and 8 DF,  p-value: 0.0087
```

```r
# Residual standard deviation
sigma <- summary(fit)$sigma
# Estimated standard deviation of parameters
summary(fit)$coefficients[ ,2]
```

```
(Intercept)           x
     0.4336      0.7310
```

Now let's return to the simulation example, the number of independent variables (x) is increased and we draw the residual from the standard normal distribution, in this particular case we can find S_{xx} as a function of n, and compare the expected values (fix $\sigma^2 = 1$) with the simulation results

```
sigma.beta <- matrix(nrow=k,ncol=2)
sigma <- numeric(k);
n <- seq(3, k+2)
for(i in 1:k){
  x <- seq(0,1,length=n[i])
  y <- 1+x+rnorm(n[i])
  fit <- lm(y ~ x)
  sigma[i] <- summary(fit)$sigma
  sigma.beta[i, ] <- summary(fit)$coefficients[ ,2]
}
```



We see that the residual variance converge to the true value with smaller and smaller variation, while the parameter variances converge to zero. In a plot like this we can therefore see the gain from obtaining more observations of the model.

Again you are encouraged to change some of the specifications of the simulation set up and see what happens.

5.4 Distribution and testing of parameters

The regression model is given by

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{5-41}$$

where the estimators of the parameters and their variances are given by Theorems 5.4 and 5.8. Since the estimators are linear functions of normal random variables (ε_i) they will also be normal random variables. To give the full stochastic model we need to use the estimate of the residual variance, and take the uncertainty of this estimator into account when constructing tests.

As we already saw in Example 5.7 we cannot expect to get the true value of the parameter, but there will be some deviations from the true value due to the stochastic nature of the model/real world application. The purpose of this

section is to give the precise description of the parameter distributions. We aim at testing hypothesis of the type

$$H_{0,i}: \quad \beta_i = \beta_{0,i}, \tag{5-42}$$

against some alternatives. The general remarks on hypothesis testing from Chapter 3 still apply, but we will go through the specific construction for linear regression here.

The central estimator of σ^2 (Equation (5-30)) is χ^2-distributed with $n-2$ degrees of freedom. In order to test the hypothesis in Equation (5-42) we need the normalized distance to a null hypothesis (i.e the distance from the observed estimate $\hat{\beta}_{0,i}$ to the value under the null hypothesis $\beta_{0,i}$). From Theorem 5.8 the standard deviations of the parameter estimates are found to

$$\hat{\sigma}_{\beta_0} = \sqrt{\frac{\hat{\sigma}^2}{n} + \frac{\bar{x}^2\hat{\sigma}^2}{S_{xx}}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \tag{5-43}$$

$$\hat{\sigma}_{\beta_1} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \tag{5-44}$$

under the null hypothesis the normalized (with standard deviations) distance between the estimators and the true values are both t-distributed with $n-2$ degrees of freedom, and hypothesis testing and confidence intervals are based on this t-distribution:

|||| Theorem 5.12 Test statistics

Under the null hypothesis ($\beta_0 = \beta_{0,0}$ and $\beta_1 = \beta_{0,1}$) the statistics

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}, \tag{5-45}$$

$$T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}}, \tag{5-46}$$

are t-distributed with $n-2$ degrees of freedom, and inference should be based on this distribution.

|||| Proof

The proof is omitted, but rely on the fact that $\hat{\beta}_j$ is normally distributed, $\hat{\sigma}^2_{\hat{\beta}_j}$ is χ^2 distributed, and a normal random variable divided by the square root of a χ^2 distributed random variable is t-distributed.

■

In this material we only test two-sided hypothesis. The hypothesis can be concluded using p-values or critical values, in the same way as we saw for hypothesis regarding mean values in Chapter 3 Section 3.1.7.

|||| Example 5.13 Example 5.9 cont.

We continue with the data from Examples 5.5 and 5.9, where we found the parameter estimates and the variance estimates. We want to test the hypotheses

$$H_{00} : \beta_0 = 0 \quad \text{vs.} \quad H_{10} : \beta_0 \neq 0, \tag{5-47}$$

$$H_{01} : \beta_1 = 1 \quad \text{vs.} \quad H_{11} : \beta_1 \neq 1, \tag{5-48}$$

on confidence level $\alpha = 0.05$. With reference to Examples 5.5 and 5.9, and Theorem 5.12, we can calculate the t-statistics as

$$t_{\text{obs},\beta_0} = \frac{-119.96}{18.897} = -6.35, \tag{5-49}$$

$$t_{\text{obs},\beta_1} = \frac{1.113 - 1}{0.1059} = 1.07. \tag{5-50}$$

H_{00} is rejected if $|t_{\text{obs},\beta_0}| > t_{1-\alpha/2}$, and H_{01} is rejected if $|t_{\text{obs},\beta_1}| > t_{1-\alpha/2}$, as usual we can find the critical values in R by:

```
qt(0.975,df=10-2)

[1] 2.306
```

and we see that with significance level $\alpha = 0.05$, then H_{00} is rejected and H_{01} isn't. If we prefer p-values rather than critical values, these can be calculated by:

```
p.v0 <- 2 * (1 - pt(abs(-6.35), df=10-2))
p.v1 <- 2 * (1 - pt(abs(1.07), df=10-2))
c(p.v0,p.v1)

[1] 0.0002206 0.3158371
```

The p-value for the intercept (β_0) is less than 0.05, while the p-value for β_1 is greater than 0.05, hence we conclude that $\beta_0 \neq 0$, but we cannot reject that $\beta_1 = 1$. The summary of linear model in R, also give t-statistics and p-values (see Example 5.5). The test statistic and the p-value for H_{01} is different from the one we obtained above. The reason for this is that `summary()` tests the default hypothesis $H_{0i} : \beta_i = 0$ against the alternative $H_{1i} : \beta_i \neq 0$. Even though this choice is reasonable in many situations it does not cover all situations, and Play MoviesPlay Movies we need to calculate p-values from the summary statistics ourselves if the hypotheses are different from the default ones.

Method 5.14 Level α t-tests for parameter

1. Formulate the *null hypothesis*: $H_{0,i} : \beta_i = \beta_{0,i}$, and the alternative hypothesis $H_{1,i} : \beta_i \neq \beta_{0,i}$
2. Compute the test statistic $t_{\text{obs},\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}}$
3. Compute the evidence against the *null hypothesis*
$$p\text{-value}_i = 2 \cdot P(T > |t_{\text{obs},\beta_i}|) \tag{5-51}$$
4. If p-value$_i < \alpha$ reject $H_{0,i}$, otherwise accept $H_{0,i}$

In many situations we will be more interested in quantifying the uncertainty of the parameter estimates rather than testing a specific hypothesis. This is usually given in the form of confidence intervals for the parameters:

Method 5.15 Parameter confidence intervals

$(1 - \alpha)$ confidence intervals for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_0}, \tag{5-52}$$
$$\hat{\beta}_1 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_1}, \tag{5-53}$$

where $t_{1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of a t-distribution with $n - 2$ degrees of freedom. Where $\hat{\sigma}_{\beta_0}$ and $\hat{\sigma}_{\beta_1}$ are calculated from the results in Theorem 5.8, and Equations (5-43) and (5-44).

Remark 5.16

We will not show (prove) the results in Method 5.15, but see Remark 3.34.

Example 5.17 Example 5.13 cont.

Based on Method 5.15 we immediately find the 95% confidence intervals for the parameters

$$I_{\beta_0} = -119.96 \pm t_{0.975} \cdot 18.897 = [-163.54, -76.38],$$
$$I_{\beta_1} = 1.113 \pm t_{0.975} \cdot 0.1059 = [0.869, 1.357],$$

with the degrees of freedom for the t-distribution equal 8, and we say with high confidence that the intervals contain the true parameter values. Of course R can find these directly from the result returned by `lm()`:

```
confint(fitStudents, level=0.95)

                2.5 %  97.5 %
(Intercept) -163.5348 -76.381
x              0.8684   1.357
```

5.4.1 Confidence and prediction intervals for the line

It is clearly of interest to predict outcomes of future experiments. Here we need to distinguish between prediction intervals, where we predict the outcome of one single experiment, and confidence intervals, where we predict the mean value of future outcomes. In the latter case we only need to account for the uncertainty in the parameter estimates while in the first case we will also need to account for the uncertainty of the error (the random part ε_i).

If we conduct a new experiment with $x_i = x_{\text{new}}$ the *expected outcome* is

$$\hat{y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \tag{5-54}$$

where the only source of variation comes from the variance of the parameter estimates, and we can calculate the variance of $\hat{Y}_{\text{new}}$

$$\begin{aligned} \mathrm{V}(\hat{Y}_{\text{new}}) &= \mathrm{V}(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}) \\ &= \mathrm{V}(\hat{\beta}_0) + \mathrm{V}(\hat{\beta}_1 x_{\text{new}}) + 2\,\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 x_{\text{new}}), \end{aligned} \tag{5-55}$$

now use the calculation rules for variances and covariances (Section 2.7), and insert the variances and the covariance from Theorem 5.8

$$\begin{aligned} \mathrm{V}(\hat{Y}_{\text{new}}) &= \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{S_{xx}} + \frac{\sigma^2 x_{\text{new}}^2}{S_{xx}} - 2\frac{\sigma^2 \bar{x} x_{\text{new}}}{S_{xx}} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}} \right), \end{aligned} \tag{5-56}$$

to find the variance of a single new point, we are using

$$Y_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} + \varepsilon_{\text{new}}, \tag{5-57}$$

and therefore need to add the variance of the residuals (ε_{new} is independent from $\hat{\beta}_0$ and $\hat{\beta}_1$)

$$\mathrm{V}(Y_{\text{new}}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}\right). \tag{5-58}$$

When we construct confidence and prediction intervals we need to account for the fact that σ^2 is estimated from data and thus use the t-distribution:

Method 5.18 Intervals for the line

The (1-α) **confidence interval** for the line $\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}, \tag{5-59}$$

and the (1-α) **prediction interval** is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}, \tag{5-60}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the t-distribution with $n-2$ degrees of freedom.

Remark 5.19

We will not show the results in Method 5.18, but use Equations (5-54)-to-(5-58) and Remark 3.34.

As illustrated in Figure 5.2 the confidence interval width will approach zero for an increasing number of data points (n) increase or as S_{xx} increase (actually, in most situations S_{xx} will also increase as n increase). Note also, that the confidence and prediction interval widths are smallest when $x_{\text{new}} = \bar{x}$. The prediction interval width will approach $2z_{1-\alpha/2} \cdot \sigma$ as $n \to \infty$. The difference between the intervals are that the prediction interval covers a new observation in $(1-\alpha) \cdot 100\%$ of the times, while the confidence interval is expected to cover the true regression line $(1-\alpha) \cdot 100\%$ of the times. One important point

book-IntroStatistics

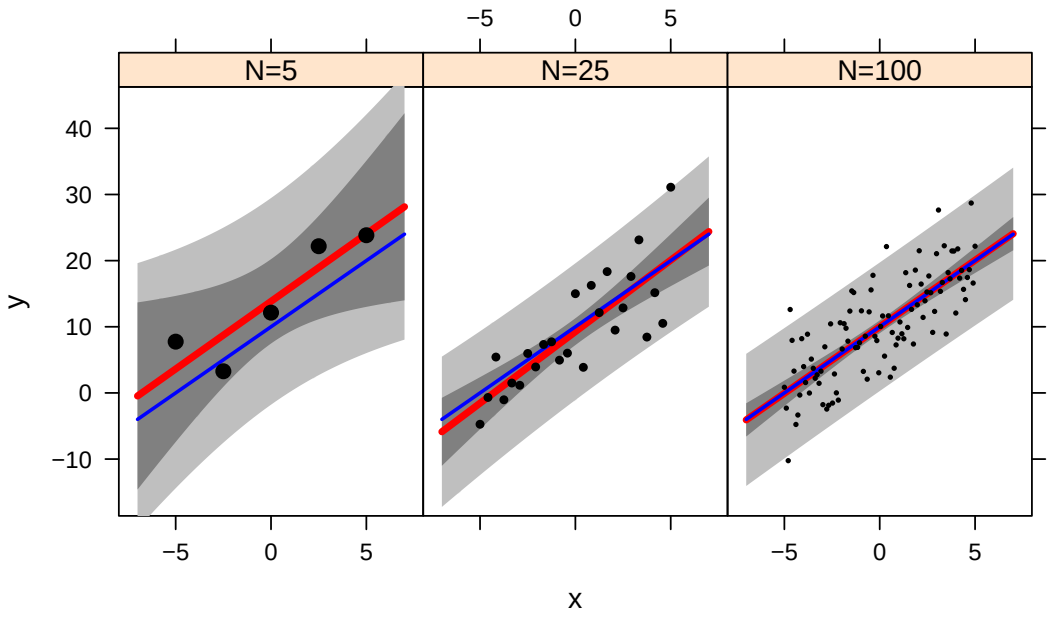


Figure 5.2: Best linear fit (red line), truth (blue line), 95% prediction interval for the points (light grey area), 95 % CI for the line (dark grey area), and observed values (black dots), for simulated data (see Example 5.21).

is: "when we have calculated the prediction interval based on some particular sample, then we actually don't know the probability of this interval covering new observations". What we know is: if we repeat the experiment, then in $(1-\alpha)\cdot 100\%$ of the times a new observation will be covered (we make a new observation each time). Same goes for the confidence interval: we don't know if the true regression line is covered by a particular interval, we only know that if we repeat the experiment, then in $(1-\alpha)\cdot 100\%$ of the times the true regression line will be covered.

In the following: first an example on calculating confidence and prediction intervals, second an example on the width of the intervals, and finally Example 5.22 on the prediction interval coverage, are given.

Example 5.20 Student height and weight Example (5.17 cont.)

With reference to Example 5.17 suppose we want to calculate prediction and confidence intervals for the line for a new student with $x_{\text{new}} = 200$ cm, the prediction is $\hat{y}_{\text{new}} = 102.6$ kg and the 95% confidence and prediction intervals become

$$I_{\text{pred}} = -120 + 1.113 \cdot 200 \pm t_{0.975}(8) \cdot 3.88\sqrt{1 + \frac{1}{10} + \frac{(178-200)^2}{1342}} = [91.8, 113], \tag{5-61}$$

$$I_{\text{conf}} = -120 + 1.113 \cdot 200 \pm t_{0.975}(8) \cdot 3.88\sqrt{\frac{1}{10} + \frac{(178-200)^2}{1342}} = [96.5, 109], \tag{5-62}$$

where $t_{0.975}$ is the 0.975-quantile of a t-distribution with $n-2$ degrees of freedom.

In R the intervals can be calculated by:

```
predict(fitStudents, newdata=data.frame(x=200), interval="confidence",
        level=0.95)

    fit   lwr   upr
1 102.6 96.52 108.7


predict(fitStudents, newdata=data.frame(x=200), interval="prediction",
        level=0.95)

    fit   lwr   upr
1 102.6 91.77 113.4
```

Example 5.21 Simulation

Figure 5.2 illustrates the difference between the confidence and prediction intervals for simulated data, with different numbers of observations: The model simulated is

$$y_i = 10 + 2x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, 5^2) \tag{5-63}$$

When n increases the width of the confidence interval for the line narrows and approaches 0, while the prediction interval width does not approach 0, but rather $2z_{1-\alpha/2}\sigma$. Further, the width of the prediction interval will always be larger than the width of the confidence interval.

Example 5.22 Prediction interval coverage

In this example it is illustrated that we actually don't know the probability that a prediction interval covers new observations, when the it is calculated using a sample (i.e. we have a realization of the prediction interval). First, a prediction interval is calculated using a single sample and it is investigated how many of k new observations falls inside it:

```
# The number of observations and the parameters
n <- 30
beta0 <- 10;  beta1 <- 3; sigma <- 0.5
# Generate some input values
x <- runif(n, -10, 10)
# Simulate output values
y <- beta0 + beta1*x + rnorm(n, sd=sigma)
# Fit a simple linear regression model to the sample
fit <- lm(y ~ x)

# The number of new observations
k <- 10000
# Generate k new input values
xnew <- runif(k, -10, 10)
# Calculate the prediction intervals for the new input values
PI <- predict(fit, newdata=data.frame(x=xnew), interval="pred")
# Simulate new output observations
ynew <- beta0 + beta1*xnew + rnorm(k, sd=sigma)
# Calculate the fraction of times the prediction interval covered the
# new observation
sum(ynew > PI[ ,"lwr"] & ynew < PI[ ,"upr"]) / k

[1] 0.8784
```

We see that the interval covered only 87.8% of the new observations, which quite less than 95% (per default predict use $\alpha = 5\%$).

Now, lets repeat the sampling, so we make a new sample k times and each time calculate a new fit and prediction interval, and each time check if a new observation falls inside it:

```r
# The number of simulated samples
k <- 10000
# Repeat the sampling k times
covered <- replicate(k, {
  # The number of observations and the parameters
  n <- 30
  beta0 <- 10;  beta1 <- 3; sigma <- 0.5
  # Generate some input values
  x <- runif(n, -10, 10)
  # Simulate output values
  y <- beta0 + beta1*x + rnorm(n, sd=sigma)
  # Fit a simple linear regression model to the sample
  fit <- lm(y ~ x)

  # Generate a new input value
  xnew <- runif(1, -10, 10)
  # The prediction interval for the new value
  PI <- predict(fit, newdata=data.frame(x=xnew), interval = "pred")
  # Simulate a single new observation
  ynew <- beta0 + beta1*xnew + rnorm(1, sd=sigma)
  # Check if the new observation was inside the interval
  ynew > PI[1,"lwr"] & ynew < PI[1,"upr"]
})
# The fraction of covered new observations
sum(covered)/k
```

```
[1] 0.9524
```

It is found that coverage is now very close to the expected 95% and this is indeed the way the coverage probability should be interpreted: with repeated sampling the probability is $1-\alpha$ that a prediction interval will cover a randomly chosen new observation. Same goes for confidence intervals (of any kind): with repeated sampling the probability is $1-\alpha$ that a confidence interval will cover the true value.

5.5 Matrix formulation of simple linear regression

The simple linear regression problem can be formulated in vector-matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{5-64}$$

or

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \varepsilon_i \sim N(0, \sigma^2) \tag{5-65}$$

One of the advantages of the matrix formulation is that the analysis generalize to higher dimensions in a straight forward way (i.e. more xs and parameters as in the following chapter). The residual sum of squares is given by

$$RSS = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{5-66}$$

and the parameter estimators are given by:

Theorem 5.23

The estimators of the parameters in the simple linear regression model are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}, \tag{5-67}$$

and the covariance matrix of the estimates is

$$\mathrm{V}[\hat{\boldsymbol{\beta}}] = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}, \tag{5-68}$$

and central estimate for the error variance is

$$\hat{\sigma}^2 = \frac{RSS}{n-2}. \tag{5-69}$$

Here $\mathrm{V}[\hat{\boldsymbol{\beta}}]$ is a matrix with elements $(\mathrm{V}[\hat{\boldsymbol{\beta}}])_{11} = \mathrm{V}[\hat{\beta}_0]$, $(\mathrm{V}[\hat{\boldsymbol{\beta}}])_{22} = \mathrm{V}[\hat{\beta}_1]$, and $(\mathrm{V}[\hat{\boldsymbol{\beta}}])_{12} = (\mathrm{V}[\hat{\boldsymbol{\beta}}])_{21} = \mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1]$.

When we want to find the minimum of RSS, we again need to differentiate RSS with respect to the parameters

$$\begin{aligned} \frac{\partial RSS}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{X}^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= -2(\boldsymbol{X}^T \boldsymbol{Y} - \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{5-70}$$

Solving for $\boldsymbol{\beta}$ gives

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}, \tag{5-71}$$

taking the expectation of $\hat{\boldsymbol{\beta}}$ we get

$$\begin{aligned} \mathrm{E}[\hat{\boldsymbol{\beta}}] &= \mathrm{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathrm{E}[\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{5-72}$$

The variance of the parameters estimates are given by

$$\begin{aligned} \mathrm{V}[\hat{\boldsymbol{\beta}}] &= \mathrm{V}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathrm{V}[\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-T} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\mathrm{V}[\boldsymbol{X}\boldsymbol{\beta}]+\mathrm{V}[\boldsymbol{\varepsilon}])\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-T} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-T} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \end{aligned} \tag{5-73}$$

Again a central estimate for σ^2 is

$$\hat{\sigma}^2 = \frac{RSS(\hat{\boldsymbol{\beta}})}{n-2}, \tag{5-74}$$

and the estimate of the parameter covariance matrix is

$$\hat{\boldsymbol{\Sigma}}_\beta = \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{5-75}$$

Marginal tests ($H_0 : \beta_i = \beta_{i,0}$) are constructed by observing that

$$\frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{(\hat{\boldsymbol{\Sigma}}_\beta)_{ii}}} \sim t(n-2). \tag{5-76}$$

The matrix calculations in R are illustrated in the next example.

Example 5.24 Student height and weight

To illustrate how the matrix formulation works in R, the student height and weight data is worked through below:

```
# Data
X <- cbind(1, x)
n <- length(y)
# Parameter estimates and variance
beta     <- solve(t(X) %*% X) %*% t(X) %*% y
e        <- y - X %*% beta
s        <- sqrt(sum(e^2) / (n - 2))
Vbeta    <- s^2 * solve(t(X) %*% X)
sbeta    <- sqrt(diag(Vbeta))
T.stat   <- beta / sbeta
p.value  <- 2 * (1 - pt(abs(T.stat), df = n-2))
# Print the results
coef.mat <- cbind(beta, sbeta, T.stat, p.value);
colnames(coef.mat) <- c("Estimates","Std.Error","t.value","p.value")
rownames(coef.mat) <- c("beta0", "beta1")
coef.mat;  s
```

```
      Estimates Std.Error t.value p.value
beta0     9.815   0.07773   126.3       0
beta1     3.039   0.01363   222.9       0
[1] 0.42
```

```
# Prediction and confidence interval
xnew  <- matrix(c(1, 200), ncol=2)
ynew  <- xnew %*% beta
Vconf <- xnew %*% Vbeta %*% t(xnew)
Vpred <- Vconf + s^2
sqrt(c(Vconf, Vpred))
```

```
[1] 2.740 2.772
```

5.6 Correlation

In the analysis above we focus on situations where we are interested in one variable (y) as a function of another variable (x). In other situations we might be more interested in how x and y vary together. Examples could be ecosystems, where the number of predators is a function of the number of preys, but

the reverse relation is also true, further both of these numbers are affected by random variations and knowledge of one only gives partial knowledge of the other. Another example is individual student grade in 2 different courses, before any grade has been given we will expect that a high grade in one course will imply a high grade in the other course, but none of them is controlled or known in advance.

In the cases above we talk about correlation analysis and to this end we will need the sample correlation coefficient, as defined in Section 1.4.3

$$\hat{\rho} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right). \tag{5-77}$$

In Section 1.4.3 we notated sample correlation with r, but here we use $\hat{\rho}$, since it is an estimate for the correlation ρ (see Section 2.8), and imply that there is a meaningful interpretation of the ρ.

5.6.1 Inference on the sample correlation coefficient

In order to answer the question: are X and Y correlated? We will be interested in constructing a test of the type

$$H_0 : \rho = 0, \quad H_1 : \rho \neq 0. \tag{5-78}$$

Consider the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{5-79}$$

in this case we can rewrite the sample correlation as

$$\begin{aligned}
\hat{\rho} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \\
&= \frac{S_{xx}}{n-1}\frac{1}{S_{xx}}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \\
&= \frac{S_{xx}}{n-1}\frac{1}{s_x s_y}\hat{\beta}_1 \\
&= \frac{s_x}{s_y}\hat{\beta}_1,
\end{aligned} \tag{5-80}$$

implying that the hypothesis (5-78) can be tested by testing the hypothesis

$$H_0 : \beta_1 = 0; \quad H_1 : \beta_1 \neq 0. \tag{5-81}$$

since clearly the relationship in Equation (5-79) can be reversed. It should be noted that we cannot use the test to construct a confidence interval for ρ.

It should be stressed that correlation does not imply causality, it just implies that the variables x and y vary together. As an example consider the number of beers sold at the university bar and the number of students attending the introductory course in statistics. Let's say that both numbers have increased and therefore have a high correlation coefficient, but it does not seem reasonable to conclude that students are more interested in statistics when drinking beers. A closer look might reveal that the number of enrolled students have actually increased and this can indeed explain the increase in both numbers.

5.6.2 Correlation and regression

In the linear regression models we would like to measure how much of the variation in the outcome (Y) is explained by the input (x). A commonly used measure for this is the coefficient of determination (explanation) or R^2-value (see also the R summary in Example 5.5).

|||| Definition 5.25 Coefficient of determination R^2

The coefficient of determination expresses the proportion of variation in the outcome (Y) explained by the regression line

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{5-82}$$

In order to find this we will split the variance of y into a component due to the regression line and a component due to the residual variation

$$\begin{aligned} s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i + e_i - \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i + e_i))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\beta}_1(x_i - \bar{x}) + e_i)^2 \\ &= \hat{\beta}_1^2 s_x^2 + \frac{n-2}{n-1}\hat{\sigma}^2, \end{aligned} \tag{5-83}$$

where the first term on the right hand side is the variability explained by the regression line and the second term is the residual variation. Dividing with the variance of Y gives a splitting in the relative variation from each of the terms. If we write out the variation explained by the regression line we get

$$\begin{aligned}
\frac{\hat{\beta}_1^2 s_x^2}{s_y^2} &= \left(\frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \frac{n-1}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \left(\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \right)^2 \frac{n-1}{\sum_{i=1}^n (x_i - \bar{x})^2} \frac{n-1}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \left(\frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right)^2 \\
&= \hat{\rho}^2.
\end{aligned} \tag{5-84}$$

We can therefore conclude that the proportion of variability (R^2) in Y explained by the regression line is equal to the squared sample correlation coefficient ($\hat{\rho}^2$).

|||| Example 5.26 Student weight and height (Example 5.20 cont.)

With reference to Example 5.20 above we can calculate the correlation coefficient in R:

```
cor(x, y)^2

[1] 0.9994
```

or we can base our calculations on the estimated slope:

```
#  fit <- lm(y ~ x)
coef(fitStudents)[2]^2 * var(x) / var(y)

    x
0.134
```

or we can find it directly in the summary of the regression model (see Example 5.5): where the number is called `Multiple R-squared`.

5.7 Model validation

So far we have discussed how to estimate parameters, predict future values, make inference etc. in the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \tag{5-85}$$

In all we have done so far the basic assumption is that the residuals are normally distributed with zero mean and constant variance, and further the residuals are mutually independent. These are assumptions which should be checked and if the assumptions are not fulfilled some actions should be taken in order to fix this. This is called *model validation* or *residual analysis* and is exactly the same idea behind the validation needed for the mean model used for t-tests in Section 3.1.8, though here including a few more steps.

The normality assumption can be checked by a normal q-q plot, and the constant variance assumption may be checked by plotting the residuals as a function of the fitted values. The normal q-q plot have been treated in Section 3.1.8 and should be applied equivalently. Plotting the residuals as a function of the fitted values should not show a systematic behaviour, this means that the range should be constant and the mean value should be constant, as illustrated in the following example:

|||| Example 5.27 Simulation

We consider data generated from the following three models

$$Y_{1,i} = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0,1), \tag{5-86}$$
$$Y_{2,i} = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0,1), \tag{5-87}$$
$$Y_{3,i} = e^{\beta_0 + \beta_1 x_{1,i} + \varepsilon_i}, \quad \varepsilon_i \sim N(0,1) \tag{5-88}$$

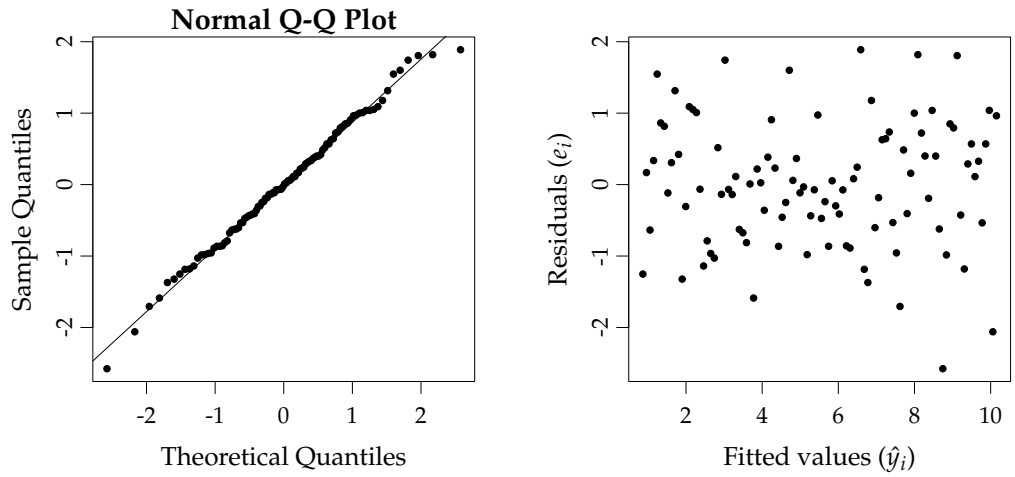
In all cases we fit the model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{5-89}$$

to the data: from the first model we would expect that the residual analysis do not show any problems, for the second model we have a linear dependence which is not included in the model and we should see this in the residual analysis, and the third is a non-linear function of the residuals as well as the regressors and one way to handle this will be discussed.

The first model is simulated, estimated and analysed by ($\beta_0 = 0$, $\beta_1 = 1$, and $\sigma^2 = 1$):

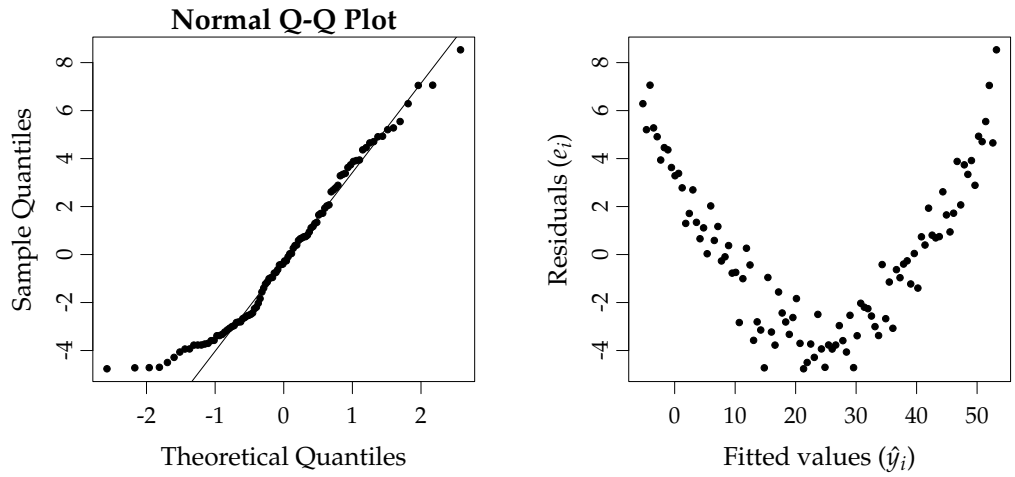
```
n <-100
x1 <- seq(1, 10, length=n)
y <- x1 + rnorm(n)
fit <- lm(y ~ x1)
qqnorm(fit$residuals, pch=19, cex=0.5)
qqline(fit$residuals)
plot(fit$fitted.values, fit$residuals, pch=19, cex=0.5,
     xlab="Fitted values ($\\hat{y}_i$)", ylab="Residuals ($e_i$)")
```



As we can see there is no serious departure from normality and there are no patterns in the residuals as a function of the fitted values.

The second model (with $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 0.5$ and $\sigma^2 = 1$) is simulated, estimated and analysed by (plot functions omitted):

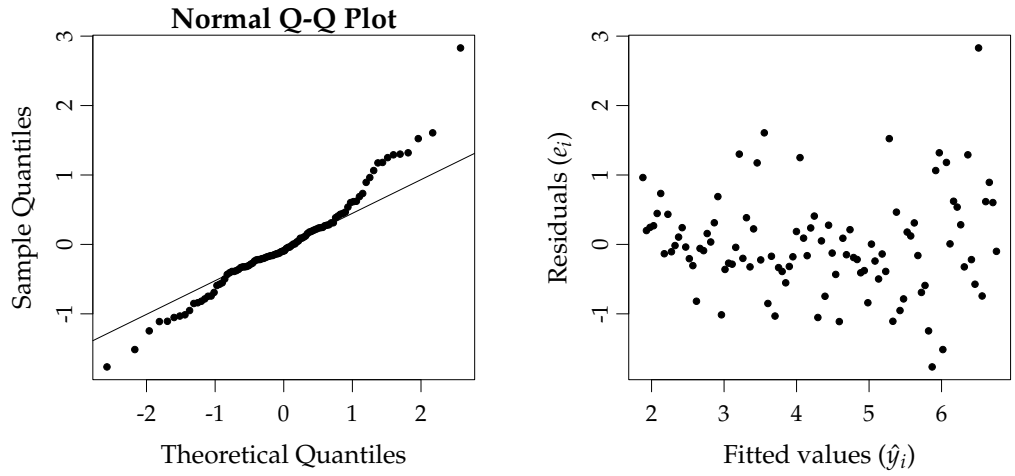
```
x1 <- seq(1, 10, length=n)
x2 <- seq(1, 10, length=n)^2
y <- x1 + 0.5 * x2 + rnorm(n)
fit <- lm(y ~ x1)
```



We see some departure from normality, but also that the residuals are related to the fitted values with a clear pattern. In the next chapter we will learn that we should find the hidden dependence (x_2) and include it in the model.

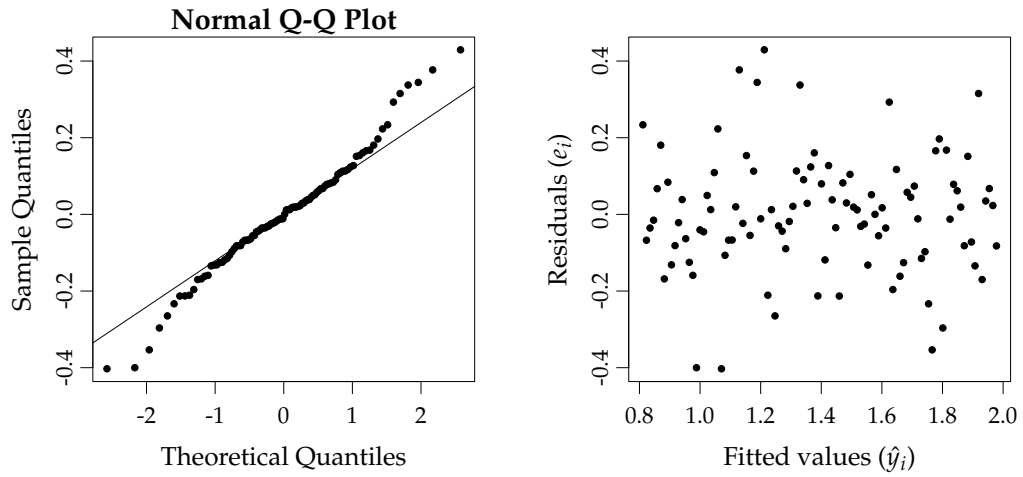
The third model (with $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 0.5$ and $\sigma^2 = 1$) is simulated, estimated and analysed by (plot function omitted):

```
x1 <- seq(4, 10, length=100)
y <- exp( 0.2 * x1 + rnorm(length(x1), sd=0.15))
fit <- lm(y ~ x1)
```



We see quite some departure from normality, and also that the variance increases as a function of the fitted values. When the variance is clearly related with the fitted values one should try to transform the dependent variable. The following R do the analysis based in log-transformed data:

```
y <- log(y)
fit <- lm(y ~ x1)
```



From the q-q plot it is found that the distribution is now quite symmetric, however still with slightly heavy tales, hence less departure from normality, compared to previous q-q plot. And, as we can see the residuals are no longer related clearly to the fitted values.

Method 5.28 Model validation (or residual analysis)

1. Check the normality assumption with a q-q plot of the residuals
2. Check the systematic behaviour by plotting the residuals e_i as a function of fitted values $\hat{y}_i$

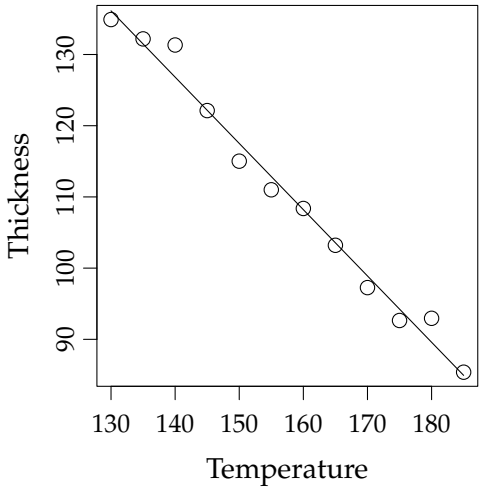
Remark 5.29 Independence

In general independence should also be checked, while there are ways to do this we will not discuss them here.

5.8 Exercises

|||| Exercise 5.1 Plastic film folding machine

On a machine that folds plastic film the temperature may be varied in the range of 130-185 °C. For obtaining, if possible, a model for the influence of temperature on the folding thickness, $n = 12$ related set of values of temperature and the fold thickness were measured that is illustrated in the following figure:



a) Determine by looking at the figure, which of the following sets of estimates for the parameters in the usual regression model is correct:

1) $\hat{\beta}_0 = 0, \hat{\beta}_1 = -0.9, \hat{\sigma} = 36$
2) $\hat{\beta}_0 = 0, \hat{\beta}_1 = 0.9, \hat{\sigma} = 3.6$
3) $\hat{\beta}_0 = 252, \hat{\beta}_1 = -0.9, \hat{\sigma} = 3.6$
4) $\hat{\beta}_0 = -252, \hat{\beta}_1 = -0.9, \hat{\sigma} = 36$
5) $\hat{\beta}_0 = 252, \hat{\beta}_1 = -0.9, \hat{\sigma} = 36$

b) What is the only possible correct answer:

1) The proportion of explained variation is 50% and the correlation is 0.98
2) The proportion of explained variation is 0% and the correlation is -0.98

3) The proportion of explained variation is 96% and the correlation is -1

4) The proportion of explained variation is 96% and the correlation is 0.98

5) The proportion of explained variation is 96% and the correlation is -0.98

|||| Exercise 5.2 Linear regression life time model

A company manufactures an electronic device to be used in a very wide temperature range. The company knows that increased temperature shortens the life time of the device, and a study is therefore performed in which the life time is determined as a function of temperature. The following data is found:

Temperature in Celcius (t)	10	20	30	40	50	60	70	80	90
Life time in hours (y)	420	365	285	220	176	117	69	34	5

a) Calculate the 95% confidence interval for the slope in the usual linear regression model, which expresses the life time as a linear function of the temperature.

b) Can a relation between temperature and life time be documented on level 5%?

|||| Exercise 5.3 Yield of chemical process

The yield y of a chemical process is a random variable whose value is considered to be a linear function of the temperature x. The following data of corresponding values of x and y is found:

Temperature in °C (x)	0	25	50	75	100
Yield in grams (y)	14	38	54	76	95

The average and standard deviation of temperature and yield are

$$\bar{x} = 50,\ s_x = 39.52847,\ \bar{y} = 55.4,\ s_y = 31.66702,$$

In the exercise the usual linear regression model is used

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, 5$$

a) Can a significant relationship between yield and temperature be documented on the usual significance level $\alpha = 0.05$?

b) Give the 95% confidence interval of the expected yield at a temperature of $x_{\text{new}} = 80$ °C.

c) What is the upper quartile of the residuals?

|||| Exercise 5.4 Plastic material

In the manufacturing of a plastic material, it is believed that the cooling time has an influence on the impact strength. Therefore a study is carried out in which plastic material impact strength is determined for 4 different cooling times. The results of this experiment are shown in the following table:

Cooling times in seconds (x)	15	25	35	40
Impact strength in kJ/m^2 (y)	42.1	36.0	31.8	28.7

The following statistics may be used:

$$\bar{x} = 28.75,\ \bar{y} = 34.65,\ S_{xx} = 368.75.$$

a) What is the 95% confidence interval for the slope of the regression model, expressing the impact strength as a linear function of the cooling time?

b) Can you conclude that there is a relation between the impact strength and the cooling time at significance level $\alpha = 5\%$?

c) For a similar plastic material the tabulated value for the linear relation between temperature and impact strength (i.e the slope) is -0.30. If the following hypothesis is tested (at level $\alpha = 0.05$)

$$H_0 : \beta_1 = -0.30$$
$$H_1 : \beta_1 \neq -0.30$$

with the usual t-test statistic for such a test, what is the range (for t) within which the hypothesis is accepted?

|||| Exercise 5.5 Water polution

In a study of pollution in a water stream, the concentration of pollution is measured at 5 different locations. The locations are at different distances to the pollution source. In the table below, these distances and the average pollution are given:

Distance to the pollution source (in km)	2	4	6	8	10
Average concentration	11.5	10.2	10.3	9.68	9.32

a) What are the parameter estimates for the three unknown parameters in the usual linear regression model: 1) The intercept (β_0), 2) the slope (β_1) and 3) error standard deviation (σ)?

b) How large a part of the variation in concentration can be explained by the distance?

c) What is a 95%-confidence interval for the expected pollution concentration 7 km from the pollution source?

|||| Exercise 5.6 Membrane pressure drop

When purifying drinking water you can use a so-called membrane filtration. In an experiment one wishes to examine the relationship between the pressure drop across a membrane and the flux (flow per area) through the membrane. We observe the following 10 related values of pressure (x) and flux (y):

	1	2	3	4	5	6	7	8	9	10
Pressure (x)	1.02	2.08	2.89	4.01	5.32	5.83	7.26	7.96	9.11	9.99
Flux (y)	1.15	0.85	1.56	1.72	4.32	5.07	5.00	5.31	6.17	7.04

Copy this into R to avoid typing in the data:

```
D <- data.frame(
    pressure=c(1.02,2.08,2.89,4.01,5.32,5.83,7.26,7.96,9.11,9.99),
    flux=c(1.15,0.85,1.56,1.72,4.32,5.07,5.00,5.31,6.17,7.04)
)
```

a) What is the empirical correlation between pressure and flux estimated to? Give also an interpretation of the correlation.

b) What is a 90% confidence interval for the slope β_1 in the usual regression model?

c) How large a part of the flux-variation ($\sum_{i=1}^{10}(y_i - \bar{y})^2$) is not explained by pressure differences?

d) Can you at significance level $\alpha = 0.05$ reject the hypothesis that the line passes through $(0, 0)$?

e) A confidence interval for the line at three different pressure levels: $x^{\mathrm{A}}_{\mathrm{new}} = 3.5$, $x^{\mathrm{B}}_{\mathrm{new}} = 5.0$ and $x^{\mathrm{C}}_{\mathrm{new}} = 9.5$ will look as follows:

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot x^{U}_{\mathrm{new}} \pm C_U$$

where U then is either A, B or C. Write the constants C_U in increasing order.

|||| Exercise 5.7 Membrane pressure drop (matrix form)

This exercise uses the data presented in Exercise 6 above.

a) Find parameters values, standard errors, t-test statistics, and p-values for the standard hypotheses tests.

Copy this into R to avoid typing in the data:

```
D <- data.frame(
  pressure=c(1.02,2.08,2.89,4.01,5.32,5.83,7.26,7.96,9.11,9.99),
  flux=c(1.15,0.85,1.56,1.72,4.32,5.07,5.00,5.31,6.17,7.04)
)
```

b) Reproduce the above numbers by matrix vector calculations. You will need some matrix notation in R:

- Matrix multiplication (XY): `X%*%Y`
- Matrix transpose (X^T): `t(X)`
- Matrix inverse (X^{-1}): `solve(X)`
- Make a matrix from vectors ($X = [x_1^T; x_2^T]$): `cbind(x1,x2)`

See also Example 5.24.

|||| Exercise 5.8 Independence and correlation

Consider the layout of independent variable in Example 5.11,

a) Show that $S_{xx} = \frac{n\cdot(n+1)}{12\cdot(n-1)}$.

Hint: you can use the following relations

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2},$$
$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

b) Show that the asymptotic correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$ is

$$\lim_{n\to\infty} \rho_n(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sqrt{3}}{2}.$$

Consider a layout of the independent variable where $n = 2k$ and $x_i = 0$ for $i \leq k$ and $x_i = 1$ for $k < i \leq n$.

c) Find S_{xx} for the new layout of x.

d) Compare S_{xx} for the two layouts of x.

e) What is the consequence for the parameter variance in the two layouts?

f) Discuss pro's and cons for the two layouts.

Glossaries

Correlation [Korrelation] The sample correlation coefficient are a summary statistic that can be calculated for two (related) sets of observations. It quantifies the (linear) strength of the relation between the two. See also: Covariance 31–33

Covariance [Kovarians] The sample covariance coefficient are a summary statistic that can be calculated for two (related) sets of observations. It quantifies the (linear) strength of the relation between the two. See also: Correlation 11, 12, 14–16, 22, 28, 29

Critical value *Kritisk værdi* As an alternative to the p-value one can use the so-called critical values, that is the values of the test-statistic which matches exactly the significance level 20

Degrees of freedom [Frihedsgrader] The number of "observations" in the data that are free to vary when estimating statistical parameters often defined as $n - 1$ 12, 19, 21, 23, 25

Expectation [Forventningsværdi] A function for calculating the mean. The value we expect for a random variable (or function of random variables), hence of the population 1, 4, 10, 11, 29

Histogram [Histogram] The default histogram uses the same width for all classes and depicts the raw frequencies/counts in each class. By dividing the raw counts by n times the class width the density histogram is found where the area of all bars sum to 1 9

Independence [Uafhængighed] 15, 37

(Statistical) Inference [Statistisk inferens (følgeslutninger baseret på data)] 2, 19, 34

Least squares [Mindste kvadraters (metode)] 4, 5, 7

Linear regression [Lineær regression (-sanalyse)] 2, 7, 10, 19, 27, 28, 32

Normal distribution [Normal fordeling] 17

Null hypothesis [Nulhypotese (H_0)] 19, 21

***P*-value** [*p*-værdi (for faktisk udfald af en teststørrelse)] 20, 21

Two-sided (test) [Tosidet test (test med tosidet alternativ)] Is also called non-directional (test) 20

Acronyms

ANOVA Analysis of Variance *Glossary:* Analysis of Variance

cdf cumulated distribution function *Glossary:* cumulated distribution function

CI confidence interval 11, 19, 21–23, 25, 32, *Glossary:* confidence interval

CLT Central Limit Theorem *Glossary:* Central Limit Theorem

IQR Inter Quartile Range *Glossary:* Inter Quartile Range

LSD Least Significant Difference *Glossary:* Least Significant Difference

pdf probability density function *Glossary:* probability density function