Wirtschafts- und Sozialstatistik
Universität Trier

Elements of Statistics

Chapter 10:
Regression analysis

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier
Faculty IV
Economic and Social Statistics Department

Winter term 2021/22

Aim of regression analysis

Let (x_i, y_i) be pairs of observations from a survey. At first, the attributes can be analysed individually. Furthermore the reciprocal relations of the variables can be examined.

Symmetric relation: Is there a statistical correlation between the variables?

- Contingency coefficient
- Rank correlation coefficient
- Bravais-Pearson correlation coefficient

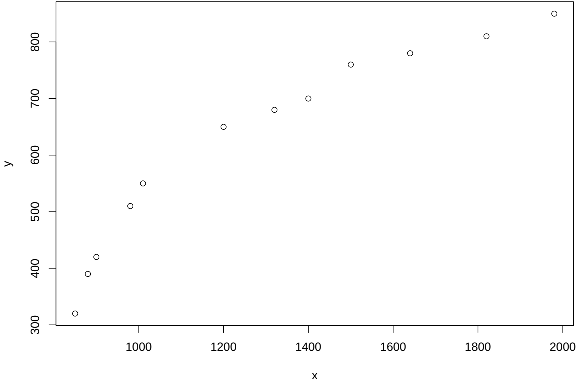
Asymmetric realtion: Does one variable influence the other, e.g. is there a functional correlation?

y is called the response variable (regressand / explained variable).
x is called the explanatory variable (regressor).

Example 10.1: Income and consumption (1)

In a survey for habitudes of income and consumption, $n = 12$ households were sampled (see HABE in Switzerland). x is the net income of the household and y are the expenditures for nutrition.

x_i	y_i
1010	550
1200	650
1400	700
980	510
850	320
1640	780
1320	680
880	390
1500	760
900	420
1980	850
1820	810



The income level seems to have an influence on consumption, but there seems to be a saturation as well.

Example 10.1: Income and consumption (2)

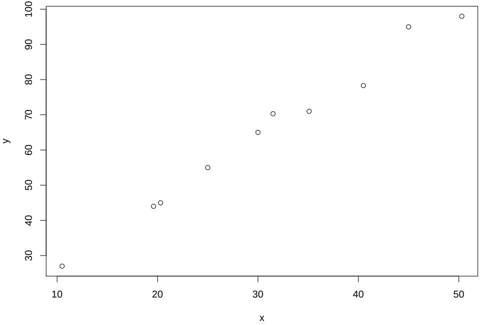
Data input and scatter plot in R:

```
x10_1 <- c(1010, 1200, 1400, 980, 850, 1640,
           1320, 880, 1500, 900, 1980, 1820)

y10_1 <- c(550, 650, 700, 510, 320, 780,
           680, 390, 760, 420, 850, 810)

plot(x = x10_1, y = y10_1)
```

Example 10.2: School readiness test (1)

In order to judge the school readiness of kindergartners, two different tests were used. A SRS of size $n = 10$ was taken. The values for the points are given through (x_i, y_i).

x_i	y_i
20.3	45.0
31.5	70.3
40.5	78.3
19.6	44.0
25.0	55.0
10.5	27.0
30.0	65.0
35.1	71.0
50.3	98.0
45.0	95.0



We have a strong linear correlation. A functional relation in the sense of independent and dependent variable is however not necessarily plausible.

Example 10.2: School readiness test (2)

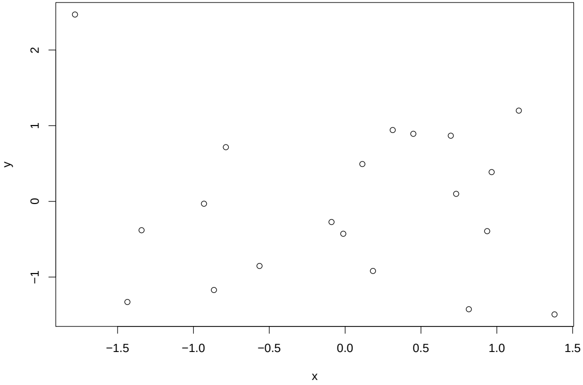
Data input and scatter plot in R:

```
x10_2 <- c(20.3, 31.5, 40.5, 19.6, 25.0,
           10.5, 30.0, 35.1, 50.3, 45.0)

y10_2 <- c(45.0, 70.3, 78.3, 44.0, 55.0,
           27.0, 65.0, 71.0, 98.0, 95.0)

plot(x = x10_2, y = y10_2)
```

Example 10.3: Two random variables (1)

Let the random variables X and Y be standard normally distributed and stochastically independent. A two-dimensional random sample of size $n = 20$ resulted in:

x_i	y_i
0.815	-1.425
-1.342	-0.381
1.380	-1.493
-1.435	-1.329
0.936	-0.393
0.450	0.893
-0.786	0.716
-0.090	-0,273
0.314	0.943
-0.930	-0,031
-0.012	-0.428
1.145	1.199
0.184	-0.919
0.732	0.100
-1.781	2.469
0.114	0.493
0.966	0.387
-0.865	-1.171
-0.564	-0.853
0.696	0.868



There is no correlation and thus no relation between the two variables.

Example 10.3: Two random variables (2)

Data input and scatter plot in R:

```
x10_3 <- c(0.815, -1.342, 1.38, -1.435, 0.936,
           0.45, -0.786, -0.09, 0.314, -0.93,
          -0.012, 1.145, 0.184, 0.732, -1.781,
           0.114, 0.966, -0.865, -0.564, 0.696)

y10_3 <- c(-1.425, -0.381, -1.493, -1.329, -0.393,
            0.893, 0.716, -0.273, 0.943, -0.031,
           -0.428, 1.199, -0.919, 0.1, 2.469,
            0.493, 0.387, -1.171, -0.853, 0.868)

plot(x = x10_3, y = y10_3)
```

Problems for regression ideas

- Quality and quantity of the involved variables
  MZ: salary $\sim$ age, gender, education, experience
  Rent index: rent $\sim$ living area, neighbourhood, ...
- Type of correlation
  - Linear
  - Polynomial
  - Exponential
- Sample size (and sampling design)
- Solution in the sense of inferential statistics
  - **Estimation** of the parameters of interest
  - Checking of parameters
    (Example: growth or stagnation)

Simple linear regression model

We assume the following linear model:

$$Y = \alpha + \beta \cdot X + \varepsilon$$

with error term ε. Hence, there might be more than one value of y corresponding to any given value of x (random error).

Using the data at hand, we would like to determine two parameters: the intercept a and the slope b of

$$\widehat{y}_i = a + b \cdot x_i \quad ,$$

where $\widehat{y}_i$ is the vertical projection of observation y_i onto the regression line. Let $e_i = y_i - \widehat{y}_i$ be the residual corresponding to observation x_i.

Using the method of ordinary least squares (OLS), we can reach estimates for the parameters a and b:

$$Z(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - (a + b \cdot x_i)\right)^2 \rightarrow \min \quad .$$

Solution to the minimisation problem

Using the two first order conditions, we get

$$n \cdot a + b \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{I}$$

$$a \cdot \sum_{i=1}^{n} x_i + b \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \,. \tag{II}$$

Solving for a in (I) and then substituting into (II), we get

$$b = \frac{\sum\limits_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum\limits_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}^*}{s_x^{*\,2}}$$

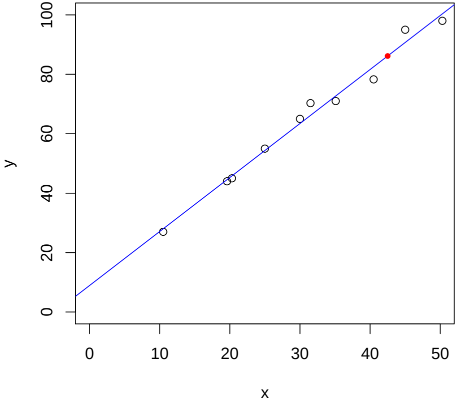
and $a = \overline{y} - b \cdot \overline{x}$. Finally, for the sample regression line we have:

$$\widehat{y} = \overline{y} + \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} \cdot (x - \overline{x}) \quad .$$

Example 10.4: see Ex. 10.2 (1)

We want to obtain estimates for the parameters α and β using the OLS method. Furthermore, we want to find a suitable estimation for the second school readiness criterion, if $x_0 = 42{,}5$ was observed for the first school readiness criterion.

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
20.3	45.0	412.09	2025.00	913.50
31.5	70.3	992.25	4942.09	2214.45
40.5	78.3	1640.25	6130.89	3171.15
19.6	44.0	384.16	1936.00	862.40
25.0	55.0	625.00	3025.00	1375.00
10.5	27.0	110.25	729.00	283.50
30.0	65.0	900.00	4225.00	1950.00
35.1	71.0	1232.01	5041.00	2492.10
50.3	98.0	2530.09	9604.00	4929.40
45.0	95.0	2025.00	9025.00	4275.00
307.8	648.6	10851.10	46682.98	22466.50



Example 10.4: see Ex. 10.2 (2)

From the table above we get:

$$\overline{x} = \frac{1}{10} \cdot 307.8 = 30{,}78 \qquad \text{sowie} \qquad \overline{y} = 64.86 \quad .$$

$\overline{x}$ and $\overline{y}$ in R:

```
SpMean_x <- mean(x10_2); SpMean_y <- mean(y10_2)
```

```
SpMean_x
```

```
[1] 30.78
```

```
SpMean_y
```

```
[1] 64.86
```

With this, we get:

$$b = \frac{22466.50 - 10 \cdot 30.78 \cdot 64.86}{10851.10 - 10 \cdot 30.78^2} = 1.817$$

$$a = 64.86 - 1{,}817 \cdot 30.78 = 8.933 \quad .$$

Example 10.4: see Ex. 10.2 (3)

Regression analysis in R:

```
reg_analysis <- lm(formula = y10_2 ~ x10_2)
summary(reg_analysis)
```

```
Call:
lm(formula = y10_2 ~ x10_2)
Residuals:
    Min      1Q  Median      3Q     Max
-4.2252 -1.5342 -0.6775  1.3293  4.2965

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.92036    2.56067   3.484  0.00828 **
x10_2         1.81740    0.07774  23.379 1.19e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.885 on 8 degrees of freedom
Multiple R-squared:  0.9856,Adjusted R-squared:  0.9838
F-statistic: 546.6 on 1 and 8 DF,  p-value: 1.191e-08
```

Example 10.4: see Ex. 10.2 (4)

Obtaining b and a in R:

```
b <- summary(reg_analysis)$coeff[2,1]
a <- summary(reg_analysis)$coeff[1,1]
```

```
b
```

```
[1] 1.817402
```

```
a
```

```
[1] 8.920358
```

Furthermore we have

$$r = \frac{22466.50 - 10 \cdot 30.78 \cdot 64.86}{\sqrt{\left(10851.10 - 10 \cdot 30.78^2\right) \cdot \left(46682.98 - 10 \cdot 64.86^2\right)}} = 0.9927 \quad .$$

Determination of r in R:

```
r <- sqrt(summary(reg_analysis)$r.squared)

r
```

```
[1] 0.9927614
```

Example 10.4: see Ex. 10.2 (5)

As estimation for the second school readiness criterion we get:

$$\widehat{y}_0 = 8.933 + 1.817 \cdot 42.5 = 86.159 \quad .$$

Construction of the estimation value $\widehat{y}_0$ in R:

```
x_0 <- 42.5
y_hat_0 <- predict(object = reg_analysis,
                   newdata = data.frame(x10_2 = x_0))

y_hat_0
```

```
       1
86.15995
```

Graphic in R:

```
plot(x = x10_2, y = y10_2)
abline(reg = reg_analysis, col = "blue") # Point not shown
```

Further issues

- From (I) follows:

$$\sum_{i=1}^{n}(y_i - a - b \cdot x_i) = \sum_{i=1}^{n}(y_i - \widehat{y}_i) = \sum_{i=1}^{n} e_i = 0$$

  and thus $\overline{y} = \overline{\widehat{y}}$.

- We have:

$$s_y^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2 \quad \text{Variation of } y$$

$$s_{\widehat{y}}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \quad \text{Variation of } \hat{y}$$

$$s_e'^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} e_i^2 \quad \text{Variation of residuals}$$

  It is: $s_y^2 = s_{\widehat{y}}^2 + s_e'^2$ and $1 = \dfrac{s_{\widehat{y}}^2}{s_y^2} + \dfrac{s_e'^2}{s_y^2}$

Coefficient of determination

The ratio $r_{xy}^2 = \frac{s_{\widehat{y}}^2}{s_y^2} = 1 - \frac{s_e'^2}{s_y^2}$ is called coefficient of determination. It is equal to the squared Bravais-Pearson correlation coefficient. Of special interest are the exceptional cases $r_{xy}^2 = 0$ and $r_{xy}^2 = 1$. It is

$$\frac{s_{\widehat{y}}^2}{s_y^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{s_y^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\big(a + bx_i - (a + b\overline{x})\big)^2}{s_y^2}$$

$$= b^2 \cdot \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}{s_y^2} = \left(\frac{s_{xy}}{s_x^2}\right)^2 \cdot \frac{s_x^2}{s_y^2} = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = r_{xy}^2 \qquad \square$$

Example 10.4: see Ex. 10.2 (6)

As a value for the coefficient of determination, we get $r_{xy}^2 = 0{,}9927^2 = 0.9855$, e.g. 98.55% of the variance of the target variable are explained through the variance of the exogeneous variable.

The coefficient of determination r_{xy}^2 in R:

```
r_q <- summary(reg_analysis)$r.squared

r_q
```

```
[1] 0.9855751
```

Inferential statistical properties of OLS

Regression line of the universe: $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$
Regression line of the sample: $y_i = a + b \cdot x_i + e_i$

Since we are drawing a random sample, A and B are random variables as estimators for α and β.

System of assumptions:

1. The error terms have the expected value 0:
$$\mathrm{E}\left(\varepsilon_i\right) = 0\,.$$
2. The error terms have a constant variance (homoskedasticity)
$$\mathrm{Var}\left(\varepsilon_i\right) = \sigma_\varepsilon^2\,.$$
3. The error terms are uncorrelated
$$\mathrm{Cov}\left(\varepsilon_i, \varepsilon_j\right) = 0 \qquad \text{for all } i \neq j\,.$$
4. Normal distribution assumption
$$\varepsilon_i \sim N(0; \sigma_\varepsilon^2)$$

Statements to the OLS regression line

Following from the assumptions 1. – 3. we have:

- A and B are best linear unbiased estimators for α and β.
- The estimator $S_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} E_i^2$ is unbiased for σ_ε^2.
- $\widehat{Y} = A + B \cdot x_0$ is the best linear unbiased estimator for the value of the target variable corresponding to x_0.

If additionally assumption 4 also holds, we have:

- A and B are ML-estimators for α and β.
- The estimator $S_e^{*2} = \frac{1}{n} \sum_{i=1}^{n} E_i^2$ is an ML-estimator for σ_ε^2.
- $\widehat{Y} = A + B \cdot x_0$ is an ML-estimator for the value of the target variable corresponding to x_0.

Point and interval estimation

Let $\operatorname{E}(A) = \alpha$ and $\operatorname{E}(B) = \beta$ as well as $\operatorname{E}(S_e^2) = \sigma_\varepsilon^2$. Furthermore, we have:

$$A \sim N\left(\alpha; \sigma_\varepsilon^2 \cdot \frac{\sum\limits_{i=1}^{n} x_i^2}{n \sum\limits_{i=1}^{n} (x_i - \overline{x})^2}\right)$$

$$B \sim N\left(\beta; \frac{\sigma_\varepsilon^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}\right)$$

$$\frac{(n-2) \cdot S_e^2}{\sigma_\varepsilon^2} = \frac{\sum\limits_{i=1}^{n} E_i^2}{\sigma_\varepsilon^2} \sim \chi_{n-2}^2$$

A and $\sum\limits_{i=1}^{n} E_i^2 / \sigma_\varepsilon^2$ resp. B and $\sum\limits_{i=1}^{n} E_i^2 / \sigma_\varepsilon^2$ are stochastically independent.

Hence, we obtain the following confidence intervals:

Confidence intervals

CI for α $$\left[a \pm t(1-\tfrac{\alpha}{2}; n-2) \cdot \sqrt{s_e^2 \cdot \frac{\sum_i x_i^2}{n\sum_i (x_i - \overline{x})^2}}\,\right]$$

CI for β $$\left[b \pm t(1-\tfrac{\alpha}{2}; n-2) \cdot \sqrt{\frac{s_e^2}{\sum_i (x_i - \overline{x})^2}}\,\right]$$

CI for σ_ε^2 $$\left[\frac{(n-2)\cdot s_e^2}{\chi^2_{n-2}(1-\frac{\alpha}{2})}; \frac{(n-2)\cdot s_e^2}{\chi^2_{n-2}(\frac{\alpha}{2})}\right]$$

CI for the mean of an observation

$$\left[\widehat{y}_0 \pm t(1-\frac{\alpha}{2}; n-2) \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}\,\right]$$

CI for the single value of an observation

$$\left[\widehat{y}_0 \pm t(1-\frac{\alpha}{2}; n-2) \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}}\,\right]$$

Example 10.5: see Ex. 10.4 (1)

We want to determine the confidence intervals for α, β and σ_ε^2 as well as for $x_0' = 35$ and $x_0'' = 50$.
At first, we have:

$$r^2 = 1 - \frac{s_e'^2}{s_y^2} \quad \Leftrightarrow \quad s_e'^2 = s_y^2 \cdot (1 - r^2) \quad \Rightarrow \quad s_e^2 = \frac{n-1}{n-2} \cdot s_y^2 \cdot (1 - r^2) \quad .$$

Using the current results, we get: $s_e^2 = 8.3210$.

Standard deviations in R:

```
Std_a <- summary(reg_analysis)$coeff[1,2]
Std_b <- summary(reg_analysis)$coeff[2,2]
Std_e <- summary(reg_analysis)$sigma
```

```
round(Std_a,4)
```

```
[1] 2.5607
```

```
round(Std_b,4)
```

```
[1] 0.0777
```

```
round(Std_e,4)
```

```
[1] 2.8846
```

Example 10.5: see Ex. 10.4 (2)

With the needed quantiles $t(0.975; 8) = 2{,}306$, $\chi^2(0.975; 8) = 17.535$ and $\chi^2(0.025; 8) = 2.180$ we get:

$$\text{CI}_\alpha : [3{,}0154; 14{,}8253] \ ; \ \text{CI}_\beta : [1{,}6381; 1{,}9967] \quad .$$

Determination of CI_α und CI_β in R:

```
alpha <- 0.05

CI_a_and_b <- confint(object = reg_analysis,
                      level = 1 - alpha
                      )
CI_a_and_b
```

```
               2.5 %   97.5 %
(Intercept) 3.015438 14.82528
x10_2       1.638145 1.99666
```

Example 10.5: see Ex. 10.4 (3)

Additionally, we get:

$$\mathrm{CI}_{\sigma_\varepsilon^2} : [3.7964, 30.5394] \quad .$$

Determination of $\mathrm{CI}_{\sigma_\varepsilon^2}$ in R:

```
n <- length(x10_2)
df <- summary(reg_analysis)$df[2]

CI_sigma_epsilon <- vector()
CI_sigma_epsilon[1] <- ((n - 2) * Std_e^2) /
                       qchisq(p = 1-(alpha/2), df = df)
CI_sigma_epsilon[2] <- ((n - 2) * Std_e^2) /
                       qchisq(p = (alpha/2), df = df)
CI_sigma_epsilon
```

```
[1]  3.79637   30.53938
```

Example 10.5: see Ex. 10.4 (4)

Finally, we get:

CI	Mean	Single value
35	[70.2940,74.7648]	[65.5120,79.5469]
50	[95.7538,103.827]	[92.0095,107.571]

Determination of the confidence intervals above in R:

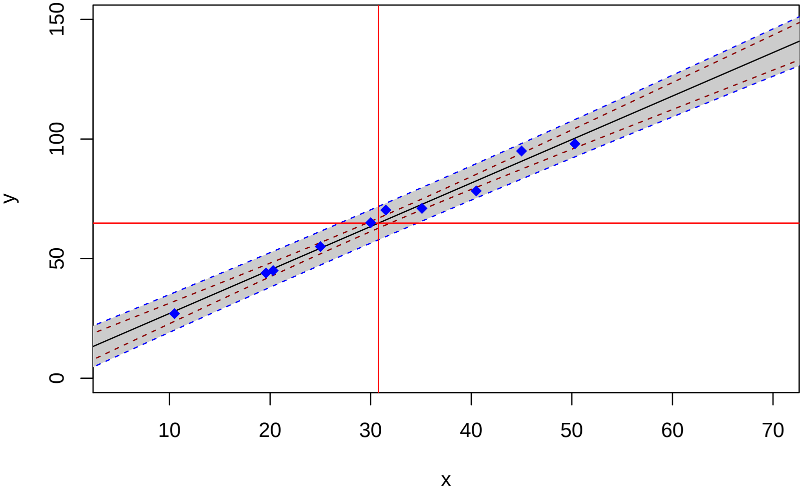
```
x_0 <- c(35, 50)

CI_Mean_Obs <- predict(reg_analysis,
                        newdata = data.frame(x10_2 = x_0),
                        interval = "confidence")

CI_Obs <- predict(reg_analysis,
                   newdata = data.frame(x10_2 = x_0),
                   interval = "prediction")
```

```
CI_Mean_Obs
       fit      lwr       upr
1 72.52944 70.29403  74.76484
2 99.79047 95.75376 103.82719
```

```
CI_Obs
       fit      lwr       upr
1 72.52944 65.51196  79.54692
2 99.79047 92.00953 107.57141
```

Example 10.5: see Ex. 10.4 (5)

Confidence bands for means (red) and single values (blue)



Hypothesis testing

- $H_0 : \alpha = \alpha_0$ versus $H_1 : \alpha \neq \alpha_0$

  Test statistic and test distribution:

$$\frac{A - \alpha_0}{S_e} \cdot \sqrt{\frac{n \sum_i (x_i - \overline{x})^2}{\sum_i x_i^2}} \sim t(n-2)$$

- $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$

  Test statistic and test distribution:

$$\frac{B - \beta_0}{S_e} \cdot \sqrt{\sum_i (x_i - \overline{x})^2} \sim t(n-2)$$

Example 10.6: see Ex. 10.4 (1)

Repetition of the regression analysis in R:

```
summary(reg_analysis)
```

```
Call:
lm(formula = y10_2 ~ x10_2)
Residuals:
Min       1Q  Median      3Q     Max
-4.2252 -1.5342 -0.6775  1.3293  4.2965

Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept)  8.92036     2.56067   3.484  0.00828 **
x10_2        1.81740     0.07774  23.379 1.19e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.885 on 8 degrees of freedom
Multiple R-squared:  0.9856,Adjusted R-squared:  0.9838
F-statistic: 546.6 on 1 and 8 DF,  p-value: 1.191e-08
```

Example 10.6: see Ex. 10.4 (2)

For the *p* value of the estimations of the coefficients, there are symbols of significance given for different values in the R-Output. We have:

$$\texttt{t value} = \texttt{Estimate}/\texttt{Std. Error} \quad .$$

This is equal to the values of the test statistics for $\alpha_0 = 0$ resp. $\beta_0 = 0$. We are especially interested in the corresponding null hypotheses, thus

- $H_0 : \alpha = \alpha_0 = 0$ and above all
- $H_0 : \beta = \beta_0 = 0$.

We test for *significance* of the single parameters. At a significance level of 5%, we reject both null hypotheses, e.g. that the parameters α resp. β are zero (*p* values are substantially smaller than 0.05). This means that the intercept and the exogenous variable X have a significant influence in the model.

From the output, we can obtain crucial parts of the beforehand calculated confidence intervals.

Example 10.6: see Ex. 10.4 (3)

Hypothesis test regarding $H_0 : \alpha = \alpha_0 = 0$ in R:

```
alpha <- 0.05

Teststat_a <- summary(reg_analysis)$coeff[1,3]
p_value_a <- summary(reg_analysis)$coeff[1,4]

c_stat_a <- vector()
c_stat_a[1] <- qt(p = alpha/2, df = df)
c_stat_a[2] <- qt(p = 1-alpha/2, df = df)
```

```
Teststat_a
```

```
[1] 3.483601
```

```
c_stat_a[1]
```

```
[1] -2.306004
```

```
c_stat_a[2]
```

```
[1] 2.306004
```

```
Teststat_a < c_stat_a[1] | Teststat_a > c_stat_a[2]
```

```
[1] TRUE
```

Alternative test decision in R:

```
p_value_a < alpha
```

```
[1] TRUE
```

```
0 < CI_a_and_b[1,1] | 0 > CI_a_and_b[1,2]
```

```
[1] TRUE
```

Example 10.6: see Ex. 10.4 (4)

Hypothesis test regarding $H_0 : \beta = \beta_0 = 0$ in R:

```
Teststat_b <- summary(reg_analysis)$coeff[2,3]
p_value_b <- summary(reg_analysis)$coeff[2,4]

c_stat_b <- vector()
c_stat_b[1] <- qt(p = alpha/2, df = df)
c_stat_b[2] <- qt(p = 1-alpha/2, df = df)
```

```
Teststat_b
```

```
[1] 23.37944
```

```
c_stat_b[1]
```

```
[1] -2.306004
```

```
c_stat_b[2]
```

```
[1] 2.306004
```

```
Teststat_b < c_stat_b[1] | Teststat_b > c_stat_b[2]
```

```
[1] TRUE
```

Alternative test decision in R:

```
p_value_b < alpha
```

```
[1] TRUE
```

```
0 < CI_a_and_b[2,1] | 0 > CI_a_and_b[2,2]
```

```
[1] TRUE
```

Example 10.6: see Ex. 10.4 (5)

Now, we're not interested in the *significance* of β, but whether the value of the parameter is at least 1. By negating the working hypothesis, we obtain the null hypothesis $H_0 : \beta \leq \beta_0 = 1$ and the corresponding alternative hypothesis.
We obtain the test statistic with the R outputs via

$$\frac{B-\beta_0}{S_e} \cdot \sqrt{\sum_i (x_i - \overline{x})^2} = \frac{B-\beta_0}{\underbrace{S_e \Big/ \sqrt{\sum_i (x_i - \overline{x})^2}}_{=0.07774}} = \frac{1.81740 - 1}{0.07774} = 10.5 \quad .$$

At a significance level of $\alpha = 0.05$, we derive from $t(0.95|8) = 1.860$ the rejection of the null hypothesis.

Pracal example: The consumption rate is at most 0.8.

Example 10.6: see Ex. 10.4 (6)

Test decision regarding $H_0 : \beta \leq \beta_0 = 1$ in R:

```
b0 <- 1

Teststat_b0 <- (b - b0) / Std_e *
               sqrt(sum((x10_2 - SpMean_x)^2))

c_stat_b0 <- qt(p = 1-alpha, df = df)
```

```
Teststat_b0              c_stat_b0
[1] 10.51523             [1] 1.859548
```

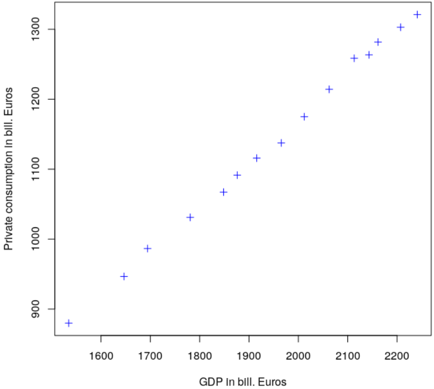
```
Teststat_b0 > c_stat_b0
```

```
[1] TRUE
```

Example 10.7: GDP and cons. expenditure

The following table contains the gross domestic product (bill. Euro) and the private consumption expenditure (bill. Euro) (**dependent variable**) for the years 1991 until 2005.

Year	GPD_i	C_i
1991	1534.60	879.86
1992	1646.62	946.60
1993	1694.37	986.54
1994	1780.78	1031.10
1995	1848.45	1067.19
1996	1876.18	1091.50
1997	1915.58	1115.78
1998	1965.38	1137.51
1999	2012.00	1175.01
2000	2062.50	1214.16
2001	2113.16	1258.57
2002	2143.18	1263.46
2003	2161.50	1281.76
2004	2207.20	1302.94
2005	2241.00	1321.06



You can find the data in the file `Example10-7.RData`.

Estimate α and β using OLS and calculate r_{xy}^2.

Change of measuring units

- Multiplying the dependent variable y with a constant c also multiplies a and b with this constant.
- Multiplying the independent variable x with a constant c changes nothing for a. b on the other hand is divided by this constant.
- The value of the coefficient of determination is independent of changing the measuring units and stays the same in both cases.

Why multiple regressors?

In reality, a variable y is depends seldom on only one regressor x.

For example, income doesn't only depend on the level of education, but also on other determinates like gender, age and job tenure.

The *multiple regression* extends the model of simple linear regression by letting more than one independent variable determine the dependent variable.

The assumptions of the simple linear regression must be kept also here. Additionally, there are some further assumptions to be made.

The multiple regression model (1)

We have the following regression model:
$y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \beta_{p-1} \cdot X_{p-1} + \varepsilon$

- y is metric in the linear regression model.
- The independent variables however don't need to be. For example, income (metric) depends on age (metric) as well as on job tenure (metric) and gender (categorial).
- In order to include gender for example in the multiple regression model, a *dummy variable D_i* is created.
- It will take the values 0 for the *reference category*and 1 for the category of interest.

The multiple regression model (2)

Let now the reference category for the variable gender be *male* ($D_i = 0$). Then, we set $D_i = 1$ for *female* observations.

From this, we implicitly obtain **two** regression equations:

Model for men (reference group):

$$y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \varepsilon$$

Model for women (group of interest):

$$y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \delta + \varepsilon$$

If a categorial variable has more than two domains, we have to generate multiple dummy variables.

In this case, we take one domain as reference category. For m domains, we get $m - 1$ dummy variables $D_1, \ldots, D_{(m-1)}$.

If the i-th observation is part of the reference category, we have $D_{ij} = 0, \forall j = 1, \ldots, m-1$ and the following model:
$y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \delta_1 \cdot D_1 + \ldots + \delta_{(m-1)} \cdot D_{(m-1)} + \varepsilon$

Example 10.8: see Ex. 10.6

Multiple regression model in R:

```
g10_8 <- factor(x = c("w", "w", "m", "w", "m", "w", "w",
                      "m", "m", "w"), levels = c("m", "w"))
reg_analysis <- lm(formula = y10_2 ~ x10_2 + g10_8)
summary(reg_analysis)
```

```
Call:
lm(formula = y10_2 ~ x10_2 + g10_8)
Residuals:
Min      1Q  Median      3Q     Max
-2.5504 -1.4267 -0.5005  1.1902  3.6159
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.85787    2.82465   1.366    0.214
x10_2        1.90105    0.06877  27.643 2.08e-08 ***
g10_8w       4.14633    1.64727   2.517    0.040 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.234 on 7 degrees of freedom
Multiple R-squared:  0.9924,Adjusted R-squared:  0.9903
F-statistic: 458.7 on 2 and 7 DF,  p-value: 3.777e-08
```

F-test

Testing the whole model requires the simultaneous test pf all $p-1$ parameters, excluding the intercept.

$H_0 : \beta_1 = \ldots = \beta_{p-1} = 0$ versus
H_1 : there is at least one $j \in \{1, \ldots, p-1\}$ with $\beta_j \neq 0$.

For linear regression, we use the F-test:

$$F = \frac{\frac{1}{p-1}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y}_i)^2}{\frac{1}{n-p}\sum_{i=1}^{n} e_i^2} = \frac{\frac{1}{p-1} r_{xy}^2}{\frac{1}{n-p}(1 - r_{xy}^2)}$$

The test statistic is F-distributed with $(p-1, n-p)$ degrees of freedom.

Common violations of assumptions

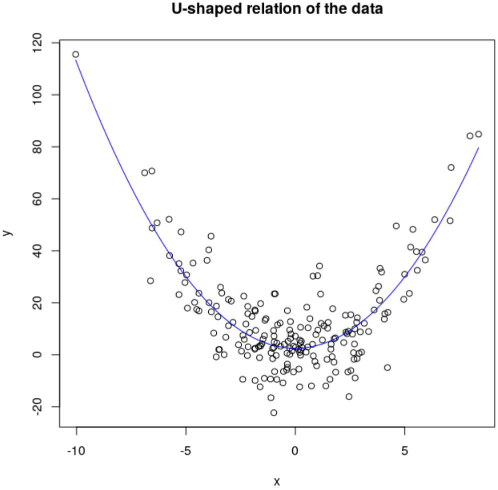
Linear regression models are based on certain assumptions (see above). If those assumptions are not met, the quality of the estimates decreases. In order to *cure* these violations, a transformation of the model can be useful.

Some empirical problems are:

- Autocorrelation of independent variables
- Non-linearity of independent variables
- Heteroskedasticity of error terms
- Non-normality of the dependent variable
- Multicollinearity

Quadratic correlation

In order to model a u-shaped correlation, one variable can be squared
Model: $y = \beta_0 + \beta_1 x_1^2 + \varepsilon$



Transformations

Depending on the relation of dependent and independent variable, there exist a number of different transformations.

Transformation	Formula	Linearisation
Linear	$Y = \alpha + \beta \cdot x$	
Logarithmic	$Y = \alpha + \beta \cdot ln(x)$	
Inverse	$Y = \alpha + \beta / x$	$Y = \alpha + \beta \cdot 1/x$
Quadratic	$Y = \alpha + \beta_1 \cdot x + \beta_2 \cdot x^2$	
Cubic	$Y = \alpha + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3$	
Power	$Y = \alpha \cdot x^{\beta}$	$ln(Y) = ln(\alpha) + \beta \cdot ln(x)$
Composite	$Y = \alpha \cdot \beta^{x}$	$ln(Y) = ln(\alpha) + ln(\beta) \cdot x$
S-curve	$Y = e^{\alpha + \beta / x}$	$ln(Y) = \alpha + \beta \cdot 1/x$
Logistic	$Y = \frac{1}{1/M + \alpha \cdot \beta^{x}}$	$ln(\frac{1}{Y} - \frac{1}{M}) = ln(\alpha) + ln(\beta) \cdot x$
Buildup	$Y = e^{\alpha + \beta \cdot x}$	$ln(Y) = \alpha + \beta \cdot x$
Exponential	$Y = \alpha \cdot e^{\beta \cdot x}$	$ln(Y) = ln(\alpha) + \beta \cdot x$

Overview from Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R. (2011): Multivariate Analysemethoden, Springer-Verlag Berlin Heidelberg, Aufl. ?, p. 141.

Linear regression with heteroskedasticity

If we have heteroskedasticity, OLS is not efficient. Furthermore, the standard errors of the coefficients are biased and thus inconsistent. Particularly, the parameters of the model can't be tested like before.

Two transformations to be applied in this context are

- to logarithmise the dependent variable and
- to square the dependent variable.

With this, we want to reach homoskedastic error terms. Then, we can apply our familiar tests. We have to proof, if the transformation was successful in the particular case.

Alternative approaches are using *robust* standard errors and a weighted estimation.

Linear regression to calculate (constant) elasticities

An elasticity gives the relative change of a dependent variable due to a change in the independent variable.

Elasticities are of interest for example in economics (price elasticity) and are defined by:

$$E = \frac{\Delta y}{\Delta x} \cdot \frac{x}{y}$$

As

$$\beta = \frac{\Delta ln(y)}{\Delta ln(x)} \approx \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}} = E$$

the β of a *log-log-model* can be directly interpreted as (constant) elasticity.

Example 10.9: Elasticity estimation

In order to estimate the elasticities, a log-log-model is stated as follows:

$$ln(y) = \alpha + \beta \cdot ln(x)$$

We obtain an estimation of $\widehat{\alpha} = 1$ and $\widehat{\beta} = 3$.

x	$\Delta x/x$	$\%\Delta x$	$y = e^{1+3\cdot ln(x)}$	$\Delta y/y$	$\%\Delta y$
2			21.74625		
2.02	$\frac{2.02-2}{2} = 0.01$	1%	22.40519	0.030301	$\approx 3\%$

This means that a relative change of x by 1% results approximately, according to the model, in a relative change of y by 3%.