2. Linear regression models based on cross-sectional data

2.1. Basic structure

Multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \varepsilon$$

y: Dependent variable

x_1, x_2, x_3,…, x_{k-1}, x_k: Explanatory variables

β_0: Intercept (constant)

β_1: This parameter measures the effect of an increase of x_1 on y, holding all other observed and unobserved factors fixed

β_2: This parameter measures the effect of an increase of x_2 on y, holding all other observed and unobserved factors fixed

⋮

β_k: This parameter measures the effect of an increase of x_k on y, holding all other observed and unobserved factors fixed

ε: Error term

Key assumption for the error term ε:

$$E(\varepsilon | x_1, x_2, \ldots, x_k) = 0$$

It states that the error term ε is mean independent of the explanatory variables.

Example: Determinants of wages

We consider a multiple linear regression model that explains wage by education time (educ), work experience (exper), and other unobserved factors (terminology: "wage is regressed on educ and exper"):

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \varepsilon$$

The following assumption must be met:

$$E(\varepsilon|\text{educ}, \text{exper}) = 0$$

We also consider a multiple linear regression model that explains wage by education time (educ), work experience (exper), on-the-job training (training), education of the mother (motheduc), and other unobserved factors (terminology: "wage is regressed on educ, exper, training, and motheduc"):

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{training} + \beta_4 \text{motheduc} + \varepsilon$$

Compared to the previous linear regression model with two explanatory variables, the factors on-the-job training and education of the mother are removed from the error term and explicitly included in the regression model.

The following assumption must now be met:

$$E(\varepsilon|\text{educ}, \text{exper}, \text{training}, \text{motheduc}) = 0$$

On the linearity of regression models:

The linearity of the corresponding regression models refers to the unknown regression parameters β_0, β_1, β_2,..., but not to the structure of the explanatory and dependent variables. Thus, nonlinearities can also be included in linear regression models.

Popular approaches of incorporating nonlinear relationships into linear regression models include the use of logarithmized and/or squared variables such as:

$$\log y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \varepsilon$$

$$y = \beta_0 + \beta_1 \log x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \cdots + \beta_{k-1} x_{k-2} + \beta_k x_{k-1} + \varepsilon$$

$$\log y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 \log x_2 + \beta_4 x_3 + \cdots + \beta_{k-1} x_{k-2} + \beta_k x_{k-1} + \varepsilon$$

The last example can be transformed into the original structure of the multiple linear regression model with $y = \log y$, $x_2 = x_1^2$, $x_3 = \log x_2$, $x_4 = x_3$,...., $x_k = x_{k-1}$.

Example: Determinants of the logarithm of wages

Now, we consider a multiple linear regression model that explains the logarithm of wage (logwage) by education time (educ), squared education time (sqeduc), work experience (exper), on-the-job training (training), education of the mother (motheduc), and other unobserved factors:

$$\text{logwage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{sqeduc} + \beta_3 \text{exper} + \beta_4 \text{training} + \beta_5 \text{motheduc} + \varepsilon$$

Another example is a linear regression model, where the effect of age and the logarithm of the years with the current employer (logtenure) on wage is examined:

$$\text{wage} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{logtenure} + \varepsilon$$

The handling of such specified linear regression models will be discussed below. Nonlinear regression models (with respect to the regression parameters) will not be examined in this course.

2.2 Estimation of the unknown parameters with the OLS method

For the further analysis of linear regression models a sample of size n from the population is required.

Multiple linear regression model with k explanatory variables:

$\{(x_{i1}, x_{i2},\ldots, x_{ik}, y_i), i = 1,\ldots, n\}$

Including the observations i = 1,..., n leads to the following linear regression models:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

For example, x_{ik} is the value of explanatory variable number k for observation i.

Essential task of regression analyses:

Estimation of unknown regression parameters $\beta_0, \beta_1, \beta_2,\ldots$

Optimization approach in the ordinary least squares (OLS) method for multiple linear regression models:

$$\min_{b_0, b_1, b_2,..., b_k} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik})^2$$

It follows for the first order conditions for the k+1 estimated regression parameters:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

OLS fitted values (predicted values) are the estimated values of the dependent variable:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} \quad \text{for } i = 1, \ldots, n$$

OLS regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

Interpretation of the OLS estimated parameters in multiple linear regression models (under the aforementioned key assumption for ε):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \cdots + \hat{\beta}_k \Delta x_k$$

If the explanatory variables x_2, x_3, x_4,…, x_k are held fixed, it follows:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1$$

The estimated parameter of the explanatory variable x_1 thus indicates in this case the change of the fitted value if x_1 increases by one unit.

If x_1, x_2, x_3,…, x_{k-1} are held fixed, it follows:

$$\Delta\hat{y} = \hat{\beta}_k \Delta x_k$$

The estimated parameter for the explanatory variable x_k thus indicates in this case the change of the fitted value if x_k increases by one unit.

The estimated parameters can thus be interpreted as estimated partial effects, i.e. when estimating the effect of one variable, the other explanatory variables are controlled for. This is the great advantage of regression analyses (or general econometric analyses), i.e. a *ceteris paribus* consideration is possible without having to conduct a randomized controlled experiment. However, this interpretation is only possible for $E(\varepsilon|x_1, x_2,\ldots, x_k) = 0$ (see below).

Example 1: Determinants of net income

By using a linear regression model, the effects of age in years (age), the years with the current employer (tenure), and working hours per week (workinghours) on the monthly net income in Euro (nincomemonth) are examined:

$$\text{nincomemonth} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{tenure} + \beta_3 \text{workinghours} + \varepsilon$$

The following OLS regression equation was estimated on the basis of n = 1,456 employees subject to social security contributions (Earnings Structure Survey 2018):

$$\hat{\text{nincomemonth}} = -682.021 + 7.705\,\text{age} + 21.746\,\text{tenure} + 56.951\,\text{workinghours}$$

Interpretation:

- Estimated positive effect of age: If tenure and workinghours are held fixed, an increase of age by one year leads to an estimated increase of the monthly net income by 7.71 Euro
- As expected, tenure and workinghours also have estimated positive effects if the other explanatory variables are held fixed, respectively

---

Example 1: Determinants of net income (Stata output)

The OLS estimation with Stata leads to the following results:

```
reg nincomemonth age tenure workinghours
```

Source	SS	df	MS	Number of obs	=	1,456
Model	754937870	3	251645957	F(3, 1452)	=	431.82
Residual	846161629	1,452	582755.943	Prob > F	=	0.0000
Total	1.6011e+09	1,455	1100412.03	R-squared	=	0.4715
				Adj R-squared	=	0.4704
				Root MSE	=	763.38

nincomemonth	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	7.70454	1.956777	3.94	0.000	3.866128	11.54295
tenure	21.74587	2.194552	9.91	0.000	17.44104	26.0507
workinghours	56.95107	1.860032	30.62	0.000	53.30244	60.59971
_cons	-682.0207	109.3244	-6.24	0.000	-896.4714	-467.5701

Example 2: Determinants of (gross) income

By using a linear regression model, the effects of education in years (educ), age in years (age), and household size (hhsize) on the monthly (gross) income in Euro (income) are examined:

$$\text{income} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{age} + \beta_3 \text{hhsize} + \varepsilon$$

The following OLS regression equation was estimated on the basis of n = 1,652 full- and parttime workers (SOEP 2019):

$$\hat{\text{income}} = -1{,}616.504 + 270.974\,\text{educ} + 26.473\,\text{age} + 64.370\,\text{hhsize}$$

Interpretation:

- Estimated positive effect of educ: If age and hhsize are held fixed, an increase of educ by one year leads to an estimated increase of the monthly income by 270.97 Euro
- Age and household size also have estimated positive effects if the other explanatory variables are held fixed, respectively

---

Example 2: Determinants of (gross) income (Stata output)

The OLS estimation with Stata leads to the following results:

```
reg income educ age hhsize
```

```
      Source |       SS           df       MS      Number of obs   =     1,652
-------------+----------------------------------   F(3, 1648)      =    126.17
       Model |  1.1615e+09         3   387170345   Prob > F        =    0.0000
    Residual |  5.0571e+09     1,648  3068619.67   R-squared       =    0.1868
-------------+----------------------------------   Adj R-squared   =    0.1853
       Total |  6.2186e+09     1,651  3766563.44   Root MSE        =    1751.7

------------------------------------------------------------------------------
      income | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   270.9736    15.2336    17.79   0.000     241.0943    300.8528
         age |   26.47332   4.011118     6.60   0.000     18.60589    34.34074
      hhsize |   64.36995   30.95926     2.08   0.038      3.64631    125.0936
       _cons |  -1616.504   289.1115    -5.59   0.000    -2183.569   -1049.439
------------------------------------------------------------------------------
```

Residuals (estimated error terms): Difference between the actually observed values of the dependent variable and the OLS fitted values

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik} \quad \text{for } i = 1, \ldots, n$$

If the residual is positive:

Actual observed dependent variable is larger than the corresponding fitted value and is thus underestimated

If the residual is negative:

Actual observed dependent variable is smaller than the corresponding fitted value and is thus overestimated

Therefore, the OLS method can be alternatively characterized as follows: The OLS estimators are obtained by minimizing the sum of squared residuals

$$\min_{\hat{\beta}_0, \hat{\beta}_1, \ldots} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min_{\hat{\beta}_0, \hat{\beta}_1, \ldots} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik})^2$$

Example: Determinants of executive board salaries

By using a linear regression model, the effect of corporate return on equity in % (roe) on the salary of executive board members of these companies in 1,000 dollars (salary) is examined:

$$salary = \beta_0 + \beta_1 roe + u$$

Using a sample of n = 209 executive board members, the following OLS regression equation was estimated:

$$\hat{salary} = 963.191 + 18.501 roe$$

The OLS estimation with Stata leads to the following results:

```
reg salary roe
      Source |       SS       df       MS              Number of obs =     209
-------------+------------------------------           F(  1,   207) =    2.77
       Model |  5166419.33     1  5166419.33           Prob > F      =  0.0978
    Residual |   386566563   207  1867471.32           R-squared     =  0.0132
-------------+------------------------------           Adj R-squared =  0.0084
       Total |   391732982   208  1883331.64           Root MSE      =  1366.6

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         roe |   18.50119   11.12325     1.66   0.098    -3.428195    40.43057
       _cons |   963.1913   213.2403     4.52   0.000     542.7902    1383.592
------------------------------------------------------------------------------
```

Example: Determinants of executive board salaries (continued)

It follows for the first five board members:

Observation number	roe	salary	$\widehat{salary}$	$\hat{\varepsilon}$
1	14.1	1,095	1,224.058	-129.058
2	10.9	1,001	1,164.854	-163.854
3	23.5	1,122	1,397.969	-275.969
4	5.9	578	1,072.348	-494.348
5	13.8	1,368	1,218.508	149.492

For the first board member, for example, the following applies to the residual:

1,095 - (963.191 + 18.501·14.1) = 1,095 - 1,224.058 = -129.058

The first four board members thus receive a lower salary than implied by the corresponding fitted values so that the salary is overestimated. In contrast, the salary of the fifth board member is underestimated.

Alternative formulation of linear regression models:

$$y_i = \hat{y}_i + \hat{\varepsilon}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} + \hat{\varepsilon}_i \quad \text{for } i = 1, \ldots, n$$

Total sum of squares:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Explained sum of squares:

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Residual sum of squares (or sum of squared residuals):

$$SSR = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$$

General rule:

$$SST = SSE + SSR$$

$$\frac{SSR}{SST} + \frac{SSE}{SST} = 1$$

Coefficient of determination: Ratio between the explained variation and the total variation (of the dependent variable y_i)

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

The coefficient of determination also corresponds to the squared correlation coefficient between the dependent variables and the OLS fitted values:

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)} = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right)}$$

Properties of the coefficient of determination:

- $0 \leq R^2 \leq 1$
- R^2 never decreases if an additional (and possibly irrelevant) explanatory variable is included (since SSR never rises in this case)
- For this reason, R^2 is a poor measure for assessing the quality or goodness of a linear regression model (also the adjusted coefficient of determination, which takes the number of explanatory variables into account, is an inappropriate measure for evaluating the quality of a linear regression model)

Example: Determinants of net income

As before, we consider the OLS regression equation examined to analyze monthly net income:

$$\widehat{nincome}month = -682.021 + 7.705age + 21.746tenure + 56.951workinghours$$

The R^2 is 0.4715, i.e. the variation in age, tenure, and working hours explains 47.15% of the variation in monthly net income.

```
reg nincomemonth age tenure workinghours
      Source |       SS           df       MS      Number of obs   =     1,456
-------------+----------------------------------   F(3, 1452)      =    431.82
       Model |   754937870         3   251645957   Prob > F        =    0.0000
    Residual |   846161629     1,452  582755.943   R-squared       =    0.4715
-------------+----------------------------------   Adj R-squared   =    0.4704
       Total |  1.6011e+09     1,455  1100412.03   Root MSE        =    763.38
------------------------------------------------------------------------------
nincomemonth | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |    7.70454   1.956777     3.94   0.000     3.866128    11.54295
      tenure |   21.74587   2.194552     9.91   0.000     17.44104     26.0507
workinghours |   56.95107   1.860032    30.62   0.000     53.30244    60.59971
       _cons |  -682.0207   109.3244    -6.24   0.000    -896.4714   -467.5701
------------------------------------------------------------------------------
```

2.3 Expected value and variance of OLS estimators

Assumptions for the analysis of the expected value of OLS estimators:

- Assumption A1: Linear in the parameters
  The relationship between the dependent variable y and the explanatory variables $x_1, x_2,\ldots, x_k$ is linear in the parameters with
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$
- Assumption A2: Random sampling
  The OLS estimation is based on a random sample of size n from the population with $\{(x_{i1}, x_{i2},\ldots, x_{ik}, y_i), i = 1,\ldots, n\}$ so that it follows for a particular observation i: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$
- Assumption A3: No perfect collinearity
  In the sample (and therefore also in the population) none of the explanatory variables is constant, and there is no exact linear relationship between the explanatory variables
- Assumption A4: Conditional expected value of ε is zero
  $E(\varepsilon|x_1, x_2,\ldots, x_k) = 0$

Under these four assumptions, all OLS estimated parameters are unbiased:

$$E(\hat{\beta}_h) = \beta_h \quad \text{for } h = 0, 1,\ldots, k$$

About assumption A4:

If this assumption is satisfied, the explanatory variables are characterized as exogenous. If A4 is violated, endogenous explanatory variables or endogeneity are present.

- In this case, the interpretation of the parameters as causal effects is no longer valid
- Although there are still true theoretical parameter values in the correct structural causal model, OLS does not estimate these values correctly, i.e. the OLS estimated parameters are biased
- In the presence of endogeneity, the causal effect is therefore not identifiable and cannot be correctly estimated with OLS
- To obtain more reliable estimates of causal effects, alternative estimation methods such as instrumental variable (IV) estimation or panel data approaches that address different types of endogeneity should be applied
- A violation of A4 can occur e.g. in the presence of measurement errors in the explanatory variables or if the functional relationship between the dependent and explanatory variables is misspecified
- One important violation of A4 arises when a relevant explanatory variable that is correlated with other explanatory variables is omitted
- Another major violation of A4 occurs in the case of reverse causality or simultaneity

Possible bias due to reverse causality and simultaneity

The effect of the explanatory variable x_1 on the dependent variable y is considered and estimated. However, suppose that y (also) has an effect on x_1. The corresponding linear regression models are as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \varepsilon$$

$$x_1 = \gamma_0 + \gamma_1 y + \gamma_2 x_2 + \cdots + \gamma_{k-1} x_{k-1} + \gamma_k x_k + u$$

Since x_1 now depends on y and y depends on ε, x_1 is correlated with ε. Therefore, x_1 is endogenous, violating A4. As a result, the estimator $\hat{\beta}_1$, which is intended to capture the estimated causal effect of x_1 on y, is biased.

In the presence of (true) simultaneity, y and x_1 are jointly and simultaneously determined within the same system (e.g. a market equilibrium). Endogeneity arises since both variables are determined at the same point of time together with the same structural error term.

Examples:

- Price and quantity (at the macro level): Price and quantity of goods are determined simultaneously by supply and demand in the market equilibrium
- Wages and working hours in negotiated contracts: The employer sets the wage depending on the hours the employee is willing to work, while the employee decides how many hours to work depending on the wage offered
- Time spent with child and working hours each day

In contrast, reverse causality refers to situations in which causality only runs in the opposite direction or in both directions, but not necessarily simultaneously (bidirectional causality). The two causal mechanisms are conceptually or behaviorally distinct. Although there may be feedback between y and x_1, the variables are not jointly determined in a simultaneous system. Instead, there is a temporal or structural sequence between the variables, leading to endogeneity even in the absence of true simultaneity.

Examples:

- Income and health: While a better health can enhance productivity and thus increase income, a higher income enables access to improved healthcare, which can lead to better health
- Education and cognitive ability: While greater cognitive ability can facilitate educational attainment, which leads to more education, education exposes individuals to cognitively demanding tasks, which can strengthen skills and thus increase cognitive ability
- Wages and training: While training programs can enhance productivity, leading to higher wages, more productive or better-paid workers are also more likely to be offered training opportunities
- Happiness and income: While higher income can increase happiness by improving living standards, happier individuals can be more motivated, which can lead to higher productivity, better employment prospects, and ultimately a higher income

Interpretation of OLS estimated parameters in the presence of reverse causality (but in the absence of omitted variable bias, see below):

- Although the estimated parameters are biased with respect to the causal effects of the explanatory variables, they can still be interpreted as estimated statistical correlations, specifically as estimated partial (linear) correlations between the relevant variables (e.g. happiness and income)
- If the same control variables are included in both reverse linear regression models (e.g. when happiness is regressed on income and income is regressed on happiness), the estimated parameters of the relevant variables have the same sign (and the corresponding t-values are identical, see below)
- However, the magnitudes of the estimated parameters differ due to different scales and units of the variables
- In contrast, the underlying estimated partial correlations are identical in both regression models since they represent non-scaled linear associations between the variables, controlling for the same variables
- However, if different control variables are included in the two regression models, the estimated partial correlations (and the t-values) differ, which can even lead to a change in the sign of the estimated parameters

Example: Reverse causality between happiness and income

By using a linear regression model, the effects of the monthly (gross) income in 1,000 Euro (income1000), education in years (educ), and age in years (age) on happiness (happiness), measured on an eleven-stage scale, are examined:

$$\text{happiness} = \beta_0 + \beta_1 \text{income1000} + \beta_2 \text{educ} + \beta_3 \text{age} + \varepsilon$$

However, as discussed above, it is possible that happiness can affect income. The corresponding linear regression model with the same control variables is:

$$\text{income1000} = \gamma_0 + \gamma_1 \text{happiness} + \gamma_2 \text{educ} + \gamma_3 \text{age} + u$$

The following OLS regression equations were estimated:

$$\widehat{\text{happiness}} = 7.388 + 0.086\,\text{income1000} + 0.021\,\text{educ} - 0.008\,\text{age}$$

$$\widehat{\text{income1000}} = -2.254 + 0.122\,\text{happiness} + 0.266\,\text{educ} + 0.026\,\text{age}$$

Interpretation:

- Estimated positive partial correlation between happiness and income1000 when controlling for educ and age (i.e. holding educ and age fixed)
- An increase of income1000 by one unit (i.e. an increase of income by 1,000 Euro) is associated with an estimated increase of happiness by 0.086 points
- An increase of happiness by one point is associated with an estimated increase of income1000 by 0.122 units (i.e. 122 Euro)

Example: Reverse causality between happiness and income (Stata output)

The OLS estimations with Stata lead to the following results:

```
reg happiness income1000 educ age
```

Source	SS	df	MS	Number of obs	=	1,652
				F(3, 1648)	=	10.59
Model	68.4645471	3	22.8215157	Prob > F	=	0.0000
Residual	3550.20858	1,648	2.15425278	R-squared	=	0.0189
				Adj R-squared	=	0.0171
Total	3618.67312	1,651	2.19180686	Root MSE	=	1.4677

happiness	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
income1000	.0861289	.0206124	4.18	0.000	.0456996	.1265583
educ	.0214697	.0139334	1.54	0.124	-.0058592	.0487987
age	-.0076401	.0033641	-2.27	0.023	-.0142385	-.0010418
_cons	7.387825	.2226399	33.18	0.000	6.951138	7.824512

```
reg income1000 happiness educ age
```

Source	SS	df	MS	Number of obs	=	1,652
				F(3, 1648)	=	131.54
Model	1201.40028	3	400.466762	Prob > F	=	0.0000
Residual	5017.19592	1,648	3.044415	R-squared	=	0.1932
				Adj R-squared	=	0.1917
Total	6218.5962	1,651	3.76656342	Root MSE	=	1.7448

income1000	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
happiness	.1217184	.0291297	4.18	0.000	.0645832	.1788536
educ	.2656497	.0152293	17.44	0.000	.2357788	.2955206
age	.0259021	.0039543	6.55	0.000	.0181461	.0336581
_cons	-2.254379	.3372995	-6.68	0.000	-2.91596	-1.592799

Remarks:

- If the effects are positive in both directions (e.g. income has a positive effect on happiness and happiness has a positive effect on income), the true positive (causal) effects are overestimated. Similarly, if the effects are negative in both directions, the true negative effects are overestimated, i.e. the absolute values of the estimated negative parameters are too large.
- If one effect is positive and the other is negative, the true positive or negative effects are underestimated, i.e. the absolute values of the estimated parameters are too small
- However, rather than the direction of the bias, it is usually the magnitude that is of greater concern. The exact size of the bias cannot typically be determined or estimated without additional information or assumptions.

Possible bias due to the omission of relevant explanatory variables („omitted variable bias“)

The correct linear regression model is specified as follows (where the assumptions A1 to A4 above are satisfied):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \varepsilon$$

However, the following misspecified linear regression model, which omits the explanatory variable x_k (e.g. due to ignorance or lack of data), is estimated:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon$$

This leads to the following OLS regression equations in the correctly specified and misspecified linear regression models:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_{k-1} x_{k-1} + \hat{\beta}_k x_k$$

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \cdots + \tilde{\beta}_{k-1} x_{k-1}$$

The following relationship applies (h = 1,…, k-1):

$$\tilde{\beta}_h = \hat{\beta}_h + \hat{\beta}_k \tilde{\delta}_h$$

$\tilde{\delta}_h$ (h = 1,…, k-1) is the OLS estimated slope parameter for x_h from a regression of x_k on all other explanatory variables (including a constant). It follows:

$$E(\tilde{\beta}_h) = \beta_h + \beta_k \tilde{\delta}_h$$

The OLS estimator of the slope parameter is thus usually biased unless either β_k or $\tilde{\delta}_h$ is zero. If $\tilde{\delta}_h$ is zero, x_h and x_k are uncorrelated in the sample. If β_k is zero, the "misspecified" linear regression model is correctly specified.

As a rule, however, x_h and x_k are correlated, leading to omitted variable bias. The direction of the bias can be summarized as follows:

- Positive bias: $\beta_k > 0$ and corr $(x_h, x_k) > 0$ or $\beta_k < 0$ and corr $(x_h, x_k) < 0$
- Negative bias: $\beta_k < 0$ and corr $(x_h, x_k) > 0$ or $\beta_k > 0$ and corr $(x_h, x_k) < 0$

As discussed above, it is rather the magnitude of the bias that is of greater concern. The size of the bias cannot generally be calculated or estimated.

Example 1: Determinants of net income

As before, we consider the OLS regression equation examined to analyze monthly net income:

$$\widehat{nincomemonth} = -682.021 + 7.705age + 21.746tenure + 56.951workinghours$$

In the next step, we exclude tenure from the linear regression model. The following OLS regression equation was estimated:

$$\widehat{nincomemonth} = -1048.895 + 18.261age + 61.278workinghours$$

Interpretation:

- The effect of age appears to be strongly overestimated in the second regression model
- One plausible explanation for this overestimation is the strong positive correlation between age and tenure
- These interpretations are based on the assumption that the first regression model is correctly specified, i.e. that no endogeneity problems are present, e.g. due to further omitted variables (the possibility of reverse causality between age and nincomemonth can be practically excluded)

Example 1: Determinants of net income (Stata output)

The OLS estimations with Stata lead to the following results:

```
reg nincomemonth age tenure workinghours
      Source |       SS           df       MS      Number of obs   =     1,456
-------------+----------------------------------   F(3, 1452)      =    431.82
       Model |   754937870         3   251645957   Prob > F        =    0.0000
    Residual |   846161629     1,452  582755.943   R-squared       =    0.4715
-------------+----------------------------------   Adj R-squared   =    0.4704
       Total |  1.6011e+09     1,455  1100412.03   Root MSE        =    763.38
------------------------------------------------------------------------------
nincomemonth | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |    7.70454   1.956777     3.94   0.000     3.866128    11.54295
      tenure |   21.74587   2.194552     9.91   0.000     17.44104     26.0507
workinghours |   56.95107   1.860032    30.62   0.000     53.30244    60.59971
       _cons |  -682.0207   109.3244    -6.24   0.000    -896.4714   -467.5701
------------------------------------------------------------------------------

reg nincomemonth age workinghours
      Source |       SS           df       MS      Number of obs   =     1,456
-------------+----------------------------------   F(2, 1453)      =    561.10
       Model |   697717779         2   348858890   Prob > F        =    0.0000
    Residual |   903381720     1,453  621735.526   R-squared       =    0.4358
-------------+----------------------------------   Adj R-squared   =    0.4350
       Total |  1.6011e+09     1,455  1100412.03   Root MSE        =     788.5
------------------------------------------------------------------------------
nincomemonth | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |    18.2608   1.695369    10.77   0.000     14.93516    21.58643
workinghours |   61.27775   1.867546    32.81   0.000     57.61438    64.94112
       _cons |  -1048.895   106.2486    -9.87   0.000    -1257.312   -840.4778
------------------------------------------------------------------------------
```

Example 2: Determinants of total rents

By using a linear regression model, the effects of the total area of the rental apartment in square meters (space), household net income in Euro (nethhincome), and household size (hhsize) on the total rent in Euro (totalrent) is examined with a sample of 4,573 rental households (main income earners):

$$\text{totalrent} = \beta_0 + \beta_1 \text{space} + \beta_2 \text{nethhincome} + \beta_3 \text{hhsize} + \varepsilon$$

The following OLS regression equation was estimated:

$$\widehat{\text{totalrent}} = 80.795 + 4.910\text{space} + 0.041\text{nethhincome} + 4.912\text{hhsize}$$

In the next step, we first exclude nethhincome and then space from the linear regression model. The following OLS regression equations were estimated:

$$\widehat{\text{totalrent}} = 76.111 + 5.836\text{space} + 15.316\text{hhsize}$$

$$\widehat{\text{totalrent}} = 276.639 + 0.074\text{nethhincome} + 49.873\text{hhsize}$$

Interpretation:

- The effect of household size appears to be strongly overestimated in the latter regression models
- One plausible explanation for this overestimation is the strong positive correlation between hhsize and nethhincome, and especially space

Example 2: Determinants of total rents (Stata output)

The OLS estimation of the first regression model with Stata leads to the following results:

```
reg totalrent nethhincome space hhsize
```

Source	SS	df	MS	Number of obs	=	4,573
Model	127229802	3	42409934	F(3, 4569)	=	1754.18
Residual	110462356	4,569	24176.4842	Prob > F	=	0.0000
Total	237692158	4,572	51988.661	R-squared	=	0.5353
				Adj R-squared	=	0.5350
				Root MSE	=	155.49

totalrent	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
nethhincome	.0411824	.0019216	21.43	0.000	.0374151	.0449496
space	4.909524	.1112115	44.15	0.000	4.691496	5.127553
hhsize	4.912236	2.673225	1.84	0.066	-.3285778	10.15305
_cons	80.79485	6.563701	12.31	0.000	67.92682	93.66287

Example 2: Determinants of total rents (Stata output continued)

The OLS estimations of the latter two regression models with Stata lead to the following results:

```
reg totalrent space hhsize
      Source |       SS           df       MS      Number of obs   =     4,573
-------------+----------------------------------   F(2, 4570)      =   2182.73
       Model |   116125511         2  58062755.6   Prob > F        =    0.0000
    Residual |   121566647     4,570  26601.0169   R-squared       =    0.4886
-------------+----------------------------------   Adj R-squared   =    0.4883
       Total |   237692158     4,572   51988.661   Root MSE        =     163.1

------------------------------------------------------------------------------
   totalrent | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       space |   5.836396   .1074724    54.31   0.000     5.625698    6.047094
      hhsize |   15.31582   2.757444     5.55   0.000     9.909898    20.72175
       _cons |   76.11138   6.881142    11.06   0.000     62.62102    89.60175
------------------------------------------------------------------------------

reg totalrent nethhincome hhsize
      Source |       SS           df       MS      Number of obs   =     4,573
-------------+----------------------------------   F(2, 4570)      =   1161.70
       Model |  80113319.8         2  40056659.9   Prob > F        =    0.0000
    Residual |   157578839     4,570  34481.1463   R-squared       =    0.3370
-------------+----------------------------------   Adj R-squared   =    0.3368
       Total |   237692158     4,572   51988.661   Root MSE        =    185.69

------------------------------------------------------------------------------
   totalrent | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
 nethhincome |   .0741717   .0021142    35.08   0.000     .0700268    .0783166
      hhsize |   49.87345   2.951713    16.90   0.000     44.08667    55.66024
       _cons |   276.6393   5.777168    47.88   0.000     265.3132    287.9653
------------------------------------------------------------------------------
```

Remarks:

- In the case of omitting relevant explanatory variables, the estimated parameters are not only biased with respect to the causal effects of the explanatory variables, but also to partial correlations between the relevant variables (e.g. between age and nincomemonth in the previous example)
- Specifically, the estimated parameters reflect biased effects or biased partial correlations that incorporate the effect of the omitted variable(s)
- In contrast, in the case of reverse causality without omitted variable bias, the partial correlations between the variables are statistically correctly estimated, although the estimated parameters are biased as causal effects
- Omitted variable bias and reverse causality can naturally also occur simultaneously (e.g. reverse causality between happiness and income, with health omitted from the explanation of happiness)

Inclusion of irrelevant explanatory variables

In contrast to the biased parameter estimators that result from the omission of relevant explanatory variables, the inclusion of irrelevant explanatory variables that have no (partial) effect on the dependent variable, does not affect the unbiasedness of the OLS estimators and thus does not lead to bias

→ However, the inclusion of irrelevant explanatory variables does affect the variance of the OLS estimators, typically increasing their variance

Assumptions for considering the variance of OLS estimators:

- Assumptions A1 to A4 when considering the expected value of OLS estimators
- Assumption A5: Homoskedasticity
  The conditional variance of the error term ε is constant, i.e. $Var(\varepsilon|x_1, x_2,..., x_k) = \sigma^2$. If this condition does not hold, i.e. if the variance depends on explanatory variables, the error terms are heteroskedastic.

→ The assumptions A1 to A5 are also known as the Gauss-Markov assumptions (in the case of regression analyses with cross-sectional data)

Under the assumptions A1 to A5, the variance of the OLS estimated slope parameters in linear regression models is:

$$Var(\hat{\beta}_h) = \frac{\sigma^2}{(1-R_h^2)\sum_{i=1}^{n}(x_{ih} - \bar{x}_h)^2} = \frac{\sigma^2}{(1-R_h^2)SST_h} \quad \text{for } h = 1,\ldots, k$$

R_h^2 is the coefficient of determination from the auxiliary regression of x_h on all other explanatory variables (including a constant).

→ While the assumption of homoskedasticity is not required for the unbiasedness of the estimated parameters, the above variance is valid only under this assumption, but not in the case of heteroskedasticity

Estimation of the variance σ^2 of the error term ε:

The estimation of σ^2 is the basis for estimating the variance of the OLS estimated regression parameters

Since $\sigma^2 = E(\varepsilon^2)$, the following estimator for σ^2 could be used:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2} = \frac{SSR}{n}$$

However, this estimator is biased. An unbiased estimator is given by the ratio between SSR and the difference between the sample size n and the number k+1 of regression parameters:

$$\hat{\sigma}^2 = \frac{1}{n\text{-}k\text{-}1}\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2} = \frac{SSR}{n\text{-}k\text{-}1}$$

The corresponding (consistent, but biased) estimator of the standard deviation σ of the error term ε, also known as standard error of the regression (SER), is:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n\text{-}k\text{-}1}\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}}$$

An unbiased estimator of the variance of the OLS estimated slope parameters in linear regression models is therefore:

$$\hat{\mathrm{Var}}(\hat{\beta}_h) = \frac{\hat{\sigma}^2}{(1-R_h^2)SST_h} \quad \text{for } h = 1,\ldots, k$$

The corresponding standard deviation of the OLS estimated slope parameters is:

$$\sqrt{\mathrm{Var}(\hat{\beta}_h)} = \frac{\sigma}{\sqrt{(1-R_h^2)SST_h}} \quad \text{for } h = 1,\ldots, k$$

This standard deviation can then be estimated as follows:

$$\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_h)} = \frac{\hat{\sigma}}{\sqrt{(1-R_h^2)SST_h}} \quad \text{for } h = 1,\ldots, k$$

The correct use of these estimates (also called standard errors of the estimated parameters) is particularly based on the homoskedasticity assumption A5. In the presence of heteroskedasticity (see below), the OLS estimated slope parameters remain unbiased, but their estimated variances and standard deviations are biased.

Under the assumptions A1 to A5, it follows:

The OLS estimators are the best linear unbiased estimators (BLUE) of the regression parameters in linear regression models

Components of BLUE:

- “Unbiased“ means that the estimator is not biased
- “Linear“ means that the estimator is a linear function of the data and the dependent variable
- “Best“ means that the estimator has the lowest variance

In accordance with the Gauss-Markov theorem, OLS estimators have the smallest variance in the class of all linear and unbiased estimators. However, the prerequisite for this property is that the assumptions A1 to A5 are satisfied.

2.4 Testing of hypotheses about regression parameters

Additional assumption A6: Normality

The error term ε is independent of the explanatory variables x_1, x_2, …, x_k and is normally distributed with an expected value of zero and a variance of σ^2:
$\varepsilon \sim N(0; \sigma^2)$

→ The assumptions A1 to A6 are commonly referred to as the classical linear model assumptions. The corresponding regression model is therefore also called the classical linear regression model.

Under the assumptions A1 to A6, it follows for the dependent variable:

$$y|x_1, x_2,\ldots, x_k \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k; \sigma^2)$$

In particular, it follows:

The OLS estimators are the best unbiased estimators (BUE) of the regression parameters in linear regression models. The OLS estimators thus have the lowest variance not only in the class of all linear unbiased estimators, but also among the broader class of all unbiased estimators.

→ While we assume in the following that the error term ε is normally distributed, statistical tests can still be conducted when this assumption A6 does not hold, provided that the sample size n is sufficiently large (see below)

However, if the assumption A6 (normally distributed error term ε) is satisfied, the OLS estimated slope parameters in linear regression models are also normally distributed, i.e. it follows (h = 1,…, k):

$$\hat{\beta}_h \sim N[\beta_h;\ Var(\hat{\beta}_h)] \quad \text{resp.} \quad \hat{\beta}_h \sim N\left[\beta_h;\ \frac{\sigma^2}{(1-R_h^2)\sum_{i=1}^{n}(x_{ih} - \bar{x}_h)^2}\right]$$

It follows (h = 1,…, k):

$$\frac{\hat{\beta}_h - \beta_h}{\sqrt{Var(\hat{\beta}_h)}} \sim N(0;\ 1) \quad \text{resp.} \quad \frac{\hat{\beta}_h - \beta_h}{\frac{\sigma}{\sqrt{(1-R_h^2)\sum_{i=1}^{n}(x_{ih} - \bar{x}_h)^2}}} \sim N(0;\ 1)$$

In addition, any linear function of the OLS estimated regression parameters $\hat{\beta}_0$, $\hat{\beta}_1$,…, $\hat{\beta}_k$ is also normally distributed.

However, the variances and standard deviations of the OLS estimated slope parameters in linear regression models are usually unknown and must therefore be estimated. Under the assumptions A1 to A6, it follows:

$$\frac{\hat{\beta}_h - \beta_h}{\sqrt{\hat{Var}(\hat{\beta}_h)}} \sim t_{n-k-1} \quad \text{resp.} \quad \frac{\hat{\beta}_h - \beta_h}{\frac{\hat{\sigma}}{\sqrt{(1-R_h^2)\sum_{i=1}^{n}(x_{ih} - \bar{x}_h)^2}}} \sim t_{n-k-1}$$

k+1 is the number of unknown regression parameters (including the constant).

The main null hypothesis that is tested in empirical applications is:

$$H_0: \beta_h = 0 \quad \text{for } h = 1,\ldots, k$$

The null hypothesis about the slope parameter β_h implies that the explanatory variable x_h has no partial effect on the dependent variable y. The test statistic of the corresponding t-test is the following t-statistic (t-value), which incorporates the estimated standard deviation (standard error) of the estimated parameter:

$$t = t_{\hat{\beta}_h} = t_h = \frac{\hat{\beta}_h}{\sqrt{\hat{Var}(\hat{\beta}_h)}}$$

Testing H_0: $\beta_h = 0$ at a given significance level α is based on the fact that the t-statistic is t-distributed with n-k-1 degrees of freedom under the null hypothesis. In empirical analyses, two-tailed t-tests are usually examined. This leads to the following alternative hypothesis:

$$H_1: \beta_h \neq 0 \quad \text{for } h = 1,\ldots, k$$

The null hypothesis is thus rejected if:

$$|t| > t_{n-k-1;1-\alpha/2}$$

More general null hypothesis:

$$H_0: \beta_h = a_h \quad \text{for } h = 1,\ldots, k$$

This null hypothesis is rejected if $\hat{\beta}_h$ differs substantially from a_h. The appropriate test statistic is the following more general t-statistic:

$$t = \frac{\hat{\beta}_h - a_h}{\sqrt{\hat{Var}(\hat{\beta}_h)}}$$

If H_0: $\beta_h = a_h$ is true, this t-statistic is again t-distributed with n-k-1 degrees of freedom. The null hypothesis is rejected at a significance level α in favor of the alternative hypothesis H_1: $\beta_h \neq a_h$ if $|t| > t_{n-k-1;1-\alpha/2}$.

Example 1: Determinants of electricity consumption

By using a linear regression model, the effects of age in years (Age), household size (HHSize), and the logarithm of household living space in square meters (LogApartmentSize) on the logarithm of household electricity consumption (LogElecUse) are examined with a sample of n = 3,705 individuals. The following OLS regression equation was estimated, including the estimated standard deviation of the estimated parameters:

$$\widehat{\text{LogElecUse}} = -5.387 + 0.006\text{Age} + 0.174\text{HHSize} + 0.395\text{LogApartSize}$$
$$(0.072) \quad (0.0005) \quad (0.008) \quad (0.018)$$

Interpretation:

- Due to the high conventional t-values, all explanatory variables have a significant effect at common significance levels (e.g. 0.05, 0.01)
- Another interesting null hypothesis tests whether β_3 equals one, i.e. H_0: $\beta_3 = 1$. In this case, it follows t = (0.395-1)/0.018 = -32.718. Therefore, the null hypothesis is rejected at all common significance levels (indicating that the estimated elasticity is strongly significantly different from the value of one).
- These (and the following) interpretations are based on the assumptions A1 to A6

Example 1: Determinants of electricity consumption (Stata output)

The OLS estimation with Stata leads to the following results:

```
reg LogElecUse Age HHSize LogApartSize

      Source |       SS           df       MS      Number of obs   =     3,705
-------------+----------------------------------   F(3, 3701)      =    741.40
       Model |  418.672194         3  139.557398   Prob > F        =    0.0000
    Residual |   696.65628     3,701  .188234607   R-squared       =    0.3754
-------------+----------------------------------   Adj R-squared   =    0.3749
       Total |  1115.32847     3,704  .301114599   Root MSE        =    .43386

------------------------------------------------------------------------------
  LogElecUse | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         Age |   .0056057   .0005076    11.04   0.000     .0046105    .0066009
      HHSize |   .1742857   .0079202    22.01   0.000     .1587574    .1898141
LogApartSize |    .395196   .0184851    21.38   0.000      .358954     .431438
       _cons |   5.387003   .0716573    75.18   0.000     5.246511    5.527494
------------------------------------------------------------------------------
```

Remarks about “significance”:

- Statistical significance should be assessed first and is a necessary condition, but the true relevance of an explanatory variable depends on its economic significance, i.e. the magnitude of the estimated slope parameter
- If a null hypothesis is not rejected at a given significance level, this does not imply that it is true. Therefore, the correct phrasing is that the "null hypothesis cannot be rejected at that significance level".

Example 2: Determinants of happiness

As before, we consider the OLS estimation results with Stata to analyze happiness:

```
reg happiness income1000 educ age
```

Source	SS	df	MS	Number of obs	=	1,652
Model	68.4645471	3	22.8215157	F(3, 1648)	=	10.59
Residual	3550.20858	1,648	2.15425278	Prob > F	=	0.0000
Total	3618.67312	1,651	2.19180686	R-squared	=	0.0189
				Adj R-squared	=	0.0171
				Root MSE	=	1.4677

happiness	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
income1000	.0861289	.0206124	4.18	0.000	.0456996	.1265583
educ	.0214697	.0139334	1.54	0.124	-.0058592	.0487987
age	-.0076401	.0033641	-2.27	0.023	-.0142385	-.0010418
_cons	7.387825	.2226399	33.18	0.000	6.951138	7.824512

Interpretation:

- The null hypothesis that the parameter of educ is zero cannot be rejected at common significance levels. Therefore, educ has no significant effect on happiness.
- The null hypothesis that the parameter of age is zero can be rejected at the 10% and 5% significance levels, but not at the 1% significance level

In addition to individual parameter tests, hypotheses about linear combinations of slope parameters can also be tested. For any values $r_1, r_2,\ldots,r_k$ and c, the null hypothesis can be specified as follows:

$$H_0: r_1\beta_1 + r_2\beta_2 + \cdots + r_k\beta_k = c \quad \text{resp.} \quad H_0: r_1\beta_1 + r_2\beta_2 + \cdots + r_k\beta_k - c = 0$$

The inclusion of the estimated variance of this linear combination of the slope parameters leads to the following t-statistic, which is t-distributed with n-k-1 degrees of freedom under the null hypothesis:

$$t = \frac{r_1\hat{\beta}_1 + \cdots + r_k\hat{\beta}_k - c}{\sqrt{\hat{Var}(r_1\hat{\beta}_1 + \cdots + r_k\hat{\beta}_k)}}$$

A common example is testing the equality of two parameters, e.g.:

$$H_0: \beta_1 = \beta_2 \quad \text{resp.} \quad H_0: \beta_1 - \beta_2 = 0$$

The corresponding t-statistic becomes:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\hat{Var}(\hat{\beta}_1 - \hat{\beta}_2)}}$$

The null hypothesis H_0 is thus rejected at a significance level α (in the case of a two-tailed test) if $|t| > t_{n-k-1;1-\alpha/2}$.

Example: Determinants of total rents

As before, we consider the OLS estimation results with Stata to analyze total rent and additionally test the null hypothesis that the parameters of nethhincome and hhsize are equal:

```
reg totalrent nethhincome space hhsize
```

Source	SS	df	MS	Number of obs	=	4,573
Model	127229802	3	42409934	F(3, 4569)	=	1754.18
Residual	110462356	4,569	24176.4842	Prob > F	=	0.0000
Total	237692158	4,572	51988.661	R-squared	=	0.5353
				Adj R-squared	=	0.5350
				Root MSE	=	155.49

totalrent	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
nethhincome	.0411824	.0019216	21.43	0.000	.0374151	.0449496
space	4.909524	.1112115	44.15	0.000	4.691496	5.127553
hhsize	4.912236	2.673225	1.84	0.066	-.3285778	10.15305
_cons	80.79485	6.563701	12.31	0.000	67.92682	93.66287

```
test nethhincome = hhsize
 ( 1)  nethhincome - hhsize = 0
       F(  1,  4569) =    3.32
            Prob > F =    0.0685
```

Finally, multiple linear exclusion restrictions can also be tested. The starting point is the following (unrestricted) multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

To test whether a group of q explanatory variables jointly has no effect on the dependent variable, the null hypothesis can be formulated as follows:

$$H_0: \beta_{k-q+1} = 0, \beta_{k-q+2} = 0, \ldots, \beta_k = 0 \quad \text{resp.} \quad H_0: \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0$$

The corresponding restricted regression model under H_0 is then:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-q} x_{k-q} + \varepsilon$$

The following F-statistic (F-value) is considered for the corresponding F-test:

$$F = \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{n-k-1}} = \frac{SSR_r - SSR_{ur}}{SSR_{ur}} \frac{n-k-1}{q}$$

Under H_0: $\beta_{k-q+1} = \beta_{k-q+2} = \ldots = \beta_k = 0$, this test statistic is F-distributed with q (i.e. the number of exclusion restrictions) and n-k-1 degrees of freedom, i.e.:

$$F \sim F_{q;n-k-1}$$

The null hypothesis is rejected at a significance level α in favor of the alternative hypothesis if $F > F_{q;n-k-1;1-\alpha}$.

Alternative formulation of the F-statistic with the coefficients of determination R^2_r and R^2_{ur} from the restricted and unrestricted linear regression models:

$$F = \frac{\frac{R^2_{ur} - R^2_r}{q}}{\frac{1-R^2_{ur}}{n-k-1}} = \frac{R^2_{ur} - R^2_r}{1-R^2_{ur}} \frac{n-k-1}{q}$$

The most common application of the F-test in empirical analyses refers to the following null hypothesis:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

This leads to the following restricted linear regression model:

$$y = \beta_0 + \varepsilon$$

In this restricted model, the coefficient of determination R^2_r is zero so that the F-statistic in its R-squared form for this special case with q = k restrictions becomes (where R^2 is the ordinary coefficient of determination from the unrestricted linear regression model with k explanatory variables):

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} = \frac{R^2}{1-R^2} \frac{n-k-1}{k}$$

Example 1: Determinants of birth weigths

By using a linear regression model, the effects of the average number of cigarettes the mother smoked per day during pregnancy (cigs), the birth order of the child (parity), the annual family income (faminc), the years of schooling of the mother (motheduc), and the years of schooling of the father (fatheduc) on the birth weight (in ounces, one ounce = 28.3495 grams) of children (bwght) are examined:

$$\text{bwght} = \beta_0 + \beta_1 \text{cigs} + \beta_2 \text{parity} + \beta_3 \text{faminc} + \beta_4 \text{motheduc} + \beta_5 \text{fatheduc} + \varepsilon$$

Based on the significance level of 0.05, the null hypothesis that the years of schooling of the parents has no effect on the birth weight is considered, i.e. H_0: $\beta_4 = \beta_5 = 0$:

- With a sample of n = 1,191 births, the unrestricted and restricted regression models are estimated by OLS, yielding $R^2_r = 0.0364$ and $R^2_{ur} = 0.0387$
- Since n-k-1 = 1,191 - 6 = 1,185 and q = 2, it follows for the F-statistic: F = [(0.0387-0.0364)/(1-0.0387)](1185/2) = 1.42
- The critical value from the F-distribution with 2 and 1,185 degrees of freedom is $F_{2;1185;0.95} = 3.00$. Since 1.42 < 3.00, the null hypothesis cannot be rejected at the 5% significance level. Therefore, there is no evidence that either motheduc or fatheduc have an effect on the birth weight.

Example 1: Determinants of birth weigths (continued)

In the unrestricted linear regression model, the OLS estimation with Stata leads to the following results:

```
reg bwght cigs parity faminc motheduc fatheduc
```

```
      Source |       SS       df       MS              Number of obs =    1191
-------------+------------------------------           F(  5,  1185) =    9.55
       Model |  18705.5567     5  3741.11135           Prob > F      =  0.0000
    Residual |  464041.135  1185  391.595895           R-squared     =  0.0387
-------------+------------------------------           Adj R-squared =  0.0347
       Total |  482746.692  1190  405.669489           Root MSE      =  19.789

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5959362   .1103479    -5.40   0.000    -.8124352   -.3794373
      parity |   1.787603   .6594055     2.71   0.007     .4938709    3.081336
      faminc |   .0560414   .0365616     1.53   0.126    -.0156913    .1277742
    motheduc |  -.3704503   .3198551    -1.16   0.247    -.9979957    .2570951
    fatheduc |   .4723944   .2826433     1.67   0.095    -.0821426    1.026931
       _cons |   114.5243   3.728453    30.72   0.000     107.2092    121.8394
------------------------------------------------------------------------------
```

Example 1: Determinants of birth weigths (continued)

OLS estimation results with Stata in the restricted linear regression model:

```
reg bwght cigs parity faminc
```

```
      Source |       SS       df       MS              Number of obs =    1191
-------------+------------------------------           F(  3,  1187) =   14.95
       Model |  17579.8997     3  5859.96658           Prob > F      =  0.0000
    Residual |  465166.792  1187  391.884408           R-squared     =  0.0364
-------------+------------------------------           Adj R-squared =  0.0340
       Total |  482746.692  1190  405.669489           Root MSE      =  19.796

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5978519   .1087701    -5.50   0.000    -.8112549   -.3844489
      parity |   1.832274   .6575402     2.79   0.005     .5422035    3.122345
      faminc |   .0670618   .0323938     2.07   0.039     .0035063    .1306173
       _cons |   115.4699   1.655898    69.73   0.000     112.2211    118.7187
------------------------------------------------------------------------------
```

Testing command and results in Stata (only possible immediately after the OLS estimation of the unrestricted regression model, where minor differences are due to rounding):

```
test motheduc=fatheduc=0
```

```
( 1)  motheduc - fatheduc = 0
( 2)  motheduc = 0

      F(  2,  1185) =    1.44
           Prob > F =    0.2380
```

Example 2: Determinants of happiness

As before, we consider the OLS estimation results with Stata to analyze happiness:

```
reg happiness income1000 educ age
```

```
      Source |       SS           df       MS      Number of obs   =     1,652
-------------+----------------------------------   F(3, 1648)      =     10.59
       Model |  68.4645471         3  22.8215157   Prob > F        =    0.0000
    Residual |  3550.20858     1,648  2.15425278   R-squared       =    0.0189
-------------+----------------------------------   Adj R-squared   =    0.0171
       Total |  3618.67312     1,651  2.19180686   Root MSE        =    1.4677

------------------------------------------------------------------------------
   happiness | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  income1000 |   .0861289   .0206124     4.18   0.000     .0456996    .1265583
        educ |   .0214697   .0139334     1.54   0.124    -.0058592    .0487987
         age |  -.0076401   .0033641    -2.27   0.023    -.0142385   -.0010418
       _cons |   7.387825   .2226399    33.18   0.000     6.951138    7.824512
------------------------------------------------------------------------------
```

The following test results in Stata refer to the null hypothesis that the parameters of educ and age are jointly zero:

```
test educ age
```

```
 ( 1)  educ = 0
 ( 2)  age = 0

       F(  2,  1648) =    3.77
            Prob > F =    0.0233
```

2.5 Asymptotic properties

Definition of consistency:
Let W_n be an estimator of the parameter θ based on a sample $y_1, y_2,\ldots,y_n$. Then W_n is a consistent estimator of θ if $P(|W_n - \theta| > \xi)$ converges (for $\xi > 0$) to zero as $n \rightarrow \infty$. In this case. W_n converges stochastically to θ, i.e. $\text{plim}(W_n) = \theta$.

Consistency of OLS estimators:

- Under the assumptions A1 to A4, the OLS estimators $\hat{\beta}_h$ ($h = 0,1,\ldots, k$) in linear regression models are consistent estimators of the true parameter β_h, i.e. $\text{plim}(\hat{\beta}_h) = \beta_h$
- Therefore, the same assumptions that ensure unbiasedness of OLS estimators also guarantee their consistency, i.e. e.g. assumption A5 (homoskedasticity) is not required for consistency. In fact, consistency of OLS estimators requires only a weaker version of A4 in addition to the assumptions A1 to A3, i.e. A4‘: $E(\varepsilon) = 0$ and $\text{Cov}(x_h, \varepsilon) = 0$ ($h = 1, 2,\ldots, k$).

Inconsistency of OLS estimators:

- Recall that if $E(\varepsilon|x_1, x_2,\ldots, x_k) \neq 0$, i.e. if A4 is violated, the OLS estimators in linear regression models are biased
- Similarly, all OLS estimators are inconsistent if the error term ε is correlated with any explanatory variable, i.e. if A4‘ is violated (e.g. in the case of reverse causality or omitted variable bias)

Asymptotic distributions of OLS estimators:

The exact normal distribution of the OLS estimators in linear regression models (and therefore the exact t- and F-distributions of the corresponding t- and F-statistics) is based on assumption A6, i.e. $\varepsilon \sim N(0; \sigma^2)$. However, even if A6 is violated, certain functions of the OLS estimators are asymptotically normally distributed.

Under the assumptions A1 to A5, it follows for the OLS estimated slope parameters in linear regression models (even without assumption A6):

$$\frac{\hat{\beta}_h - \beta_h}{\sqrt{\hat{Var}(\hat{\beta}_h)}} \overset{a}{\sim} N(0; 1)$$

This property does not contradict the previous property that this function is exactly t-distributed with n-k-1 degrees of freedom under the assumptions A1 to A6 since also the following notation is feasible (as the t-distribution converges to the standard normal distribution as the number of degrees of freedom increases):

$$\frac{\hat{\beta}_h - \beta_h}{\sqrt{\hat{Var}(\hat{\beta}_h)}} \overset{a}{\sim} t_{n-k-1}$$

It follows:

Even if the error term ε is not normally distributed, the previously introduced t- and F-tests can still be applied, provided that the sample size n is sufficiently large. However, for a small sample size n (or a small number of degrees of freedom n-k-1), the approximation of the t-statistic towards the t-distribution is insufficient.

Asymptotic efficiency:

Under the Gauss-Markov assumptions (and thus under the assumptions A1 to A5), the OLS estimators $\hat{\beta}_h$ (h = 0, 1,…, k) are asymptotically efficient within the class of consistent estimators $\tilde{\beta}_h$ of the regression parameters in linear regression models, i.e. it follows for the asymptotic variance Avar:

$$\text{Avar}[\sqrt{n}(\hat{\beta}_h - \beta_h)] \leq \text{Avar}[\sqrt{n}(\tilde{\beta}_h - \beta_h)]$$

2.6 The structure of dependent and explanatory variables

Logarithmized und squared variables:

Linear regression models can capture non-linear relationships by including (naturally) logarithmized and/or squared variables

Overview of the inclusion of logarithmized variables:

Linear regression model	Dependent variable	Explanatory variable	Interpretation of the estimated slope parameter
Level-level	y	x_h	$\Delta\hat{y} = \hat{\beta}_h \Delta x_h$
Level-log	y	$\log x_h$	$\Delta\hat{y} \approx (\hat{\beta}_h/100)\%\Delta x_h$
Log-level	log y	x_h	$\%\Delta\hat{y} \approx (100\hat{\beta}_h)\Delta x_h$
Log-log	log y	$\log x_h$	$\%\Delta\hat{y} \approx \hat{\beta}_h \%\Delta x_h$

Example 1: Determinants of electricity consumption

As before, we consider the OLS estimation results with Stata to analyze electricity consumption:

```
reg LogElecUse Age HHSize LogApartSize
```

LogElecUse	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Age	.0056057	.0005076	11.04	0.000	.0046105	.0066009
HHSize	.1742857	.0079202	22.01	0.000	.1587574	.1898141
LogApartSize	.395196	.0184851	21.38	0.000	.358954	.431438
_cons	5.387003	.0716573	75.18	0.000	5.246511	5.527494

It follows:

- An increase of household living space (ApartSize) by 1% (i.e. $\%\Delta$ApartSize = 1) leads to an approximately estimated increase of household electricity consumption (ElecUse) by 0.395% (if the variables Age and HHSize are held fixed)
- An increase of household size by one person (i.e. ΔHHSize = 1) leads to an approximately estimated increase of household electricity consumption (ElecUse) by $0.174 \cdot 100 = 17.4\%$ (if Age and LogApartSize are held fixed)

Squared explanatory variables:

These variables allow for increasing or decreasing (partial) marginal effects in linear regression models

Additional inclusion of a squared explanatory variable x_1^2 (in addition to the k-1 explanatory variables $x_1, x_2, \ldots, x_{k-1}$):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \cdots + \beta_{k-1} x_{k-2} + \beta_k x_{k-1} + \varepsilon$$

In this case, β_1 alone does not represent the change of y with respect to x_1. The OLS regression equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \hat{\beta}_3 x_2 + \cdots + \hat{\beta}_{k-1} x_{k-2} + \hat{\beta}_k x_{k-1}$$

If $x_2, \ldots, x_{k-1}$ are held fixed, it follows the approximation:

$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x_1)\Delta x_1 \quad \text{resp.} \quad \frac{\Delta\hat{y}}{\Delta x_1} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x_1$$

Therefore, the estimated (partial) marginal effect of x_1 on y depends on both $\hat{\beta}_2$ and the value of x_1.

Example: Determinants of net income

As before, we consider a linear regression model to analyze monthly net income. However, in addition to age, tenure, and workinghours, we include the squared value of tenure (tenuresquared) to explain nincomemonth. The OLS estimation with Stata leads to the following results:

```
reg nincomemonth tenure tenuresquared age workinghours
```

```
      Source |       SS           df       MS      Number of obs   =     1,456
-------------+----------------------------------   F(4, 1451)      =    329.10
       Model |   761614778         4   190403694   Prob > F        =    0.0000
    Residual |   839484721     1,451  578555.976   R-squared       =    0.4757
-------------+----------------------------------   Adj R-squared   =    0.4742
       Total |  1.6011e+09     1,455  1100412.03   Root MSE        =    760.63

------------------------------------------------------------------------------
 nincomemonth | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
--------------+---------------------------------------------------------------
       tenure |   41.35841   6.173452     6.70   0.000     29.24856    53.46825
tenuresquared |  -.5451525   .1604734    -3.40   0.001    -.8599371   -.2303679
          age |   6.819813    1.96703     3.47   0.001     2.961288    10.67834
 workinghours |   56.29534   1.863342    30.21   0.000     52.64021    59.95048
        _cons |  -710.6947   109.2562    -6.50   0.000    -925.0117   -496.3776
------------------------------------------------------------------------------
```

Example: Determinants of net income (continued)

Interpretation:

- The first additional year of tenure (from tenure = 0) leads to an approximately estimated increase of nincomemonth by 41.358 Euro
- An increase of tenure from one to two years leads to the following approximately estimated increase of nincomemonth:
  41.358 - 2·0.545 = 40.268 Euro
- An increase of tenure from ten to eleven years leads to an even lower approximately estimated increase of nincomemonth:
  41.358 - 2·0.545·10 = 30.455 Euro
- Estimated inverse U-shaped relationship between tenure and nincomeonth: Beyond the turning point of $|\hat{\beta}_{tenure}/(2\cdot\hat{\beta}_{tenuresquared})|$ = 41.358/ (2·0.545) = 37.933 years, tenure has an estimated negative effect on nincomemonth
- If $\hat{\beta}_1$ (i.e. the estimated parameter of the unsquared variable) were negative and $\hat{\beta}_2$ (i.e. the estimated parameter of the squared variable) were positive, an estimated U-shaped relationship would occur. Thus, the variable would have an estimated negative marginal effect (decreasing with the value of the variable) before the turning point and an estimated positive marginal effect (increasing with the value of the variable) after the turning point.

Interaction terms:

These terms allow the (partial) effect (or elasticity or semi-elasticity) of an explanatory variable in linear regression models to depend on different values of another explanatory variable

Additional inclusion of an interaction term between x_1 and x_2 (in addition to the k-1 explanatory variables x_1, x_2,…, x_{k-1}):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_3 + \cdots + \beta_{k-1} x_{k-2} + \beta_k x_{k-1} + \varepsilon$$

In this case, β_1 again does not represent the change of y with respect to x_1 alone. The OLS regression equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_3 + \cdots + \hat{\beta}_{k-1} x_{k-2} + \hat{\beta}_k x_{k-1}$$

If x_2,…, x_{k-1} are held fixed, it follows the approximation:

$$\Delta\hat{y} \approx (\hat{\beta}_1 + \hat{\beta}_3 x_2)\Delta x_1 \quad \text{resp.} \quad \frac{\Delta\hat{y}}{\Delta x_1} \approx \hat{\beta}_1 + \hat{\beta}_3 x_2$$

Therefore, the estimated (partial) marginal effect of x_1 on y also depends on both $\hat{\beta}_3$ and the value of x_2. In practice, the estimated effect is often examined at interesting values of x_2 (e.g. the mean of x_2 in the sample). $\hat{\beta}_1$ alone indicates the estimated (partial) effect of x_1 only when x_2 is zero.

Example: Determinants of (gross) income

As before, we consider a linear regression model to analyze monthly gross income. However, in addition to educ, age, and hhsize, we include the interaction of educ and age (educage) to explain income. The OLS estimation with Stata leads to the following results:

```
reg income hhsize educ age educage
```

Source	SS	df	MS	Number of obs	=	1,652
Model	1.2161e+09	4	304023960	F(4, 1647)	=	100.10
Residual	5.0025e+09	1,647	3037340.86	Prob > F	=	0.0000
Total	6.2186e+09	1,651	3766563.44	R-squared	=	0.1956
				Adj R-squared	=	0.1936
				Root MSE	=	1742.8

income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hhsize	52.923	30.91921	1.71	0.087	-7.722102	113.5681
educ	-5.671606	66.99484	-0.08	0.933	-137.0756	125.7324
age	-51.49744	18.82052	-2.74	0.006	-88.41211	-14.58277
educage	6.023479	1.420883	4.24	0.000	3.236552	8.810406
_cons	1986.282	897.2191	2.21	0.027	226.472	3746.093

Example: Determinants of (gross) income (continued)

Interpretation:

- Considering the estimated parameter of educ alone would suggest an estimated negative effect. However, the estimated effect corresponds to age = 0, which is practically irrelevant.
- A more interesting value for considering the effect of educ is the sample mean of age, which is 45.725 years (other values such as the minimum or the maximum would also be interesting). At this age, the estimated marginal effect of educ is: -5.672 + 6.023·45.725 = 269.749 Euro. This implies that for an age of 45.725 years, an increase of educ by one year leads to an approximately estimated increase of income by 269.749 Euro.
- Whether this estimated effect of educ at the average age is significant can be assessed by estimating a reparameterized linear regression model, replacing age with age-45.725 and educage with (age-45.725)*educ. The corresponding t-value is 17.80, implying a highly significantly positive effect.
- Overall, the estimated parameter of the interaction term implies that the estimated marginal effect of educ increases with increasing age
- The same approach can be used to estimate the effect of age on income for different values of educ

Example: Determinants of (gross) income (continued)

The marginal effects of educ at different values of age can be estimated with Stata as follows:

```
reg income hhsize c.age##c.educ
```

Source	SS	df	MS	Number of obs	=	1,652
				F(4, 1647)	=	100.10
Model	1.2161e+09	4	304023960	Prob > F	=	0.0000
Residual	5.0025e+09	1,647	3037340.86	R-squared	=	0.1956
				Adj R-squared	=	0.1936
Total	6.2186e+09	1,651	3766563.44	Root MSE	=	1742.8

income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hhsize	52.923	30.91921	1.71	0.087	-7.722102	113.5681
age	-51.49744	18.82052	-2.74	0.006	-88.41211	-14.58277
educ	-5.671606	66.99484	-0.08	0.933	-137.0756	125.7324
c.age#c.educ	6.023479	1.420883	4.24	0.000	3.236552	8.810406
_cons	1986.282	897.2191	2.21	0.027	226.472	3746.093

```
margins, dydx(educ) at(age=(20 40 60))
```

	Delta-method					
	dy/dx	std. err.	t	P>|t|	[95% conf.	interval]
educ						
_at						
1	114.798	39.83605	2.88	0.004	36.66333	192.9326
2	235.2676	17.33896	13.57	0.000	201.2588	269.2763
3	355.7371	25.08973	14.18	0.000	306.526	404.9483

Qualitative explanatory variables:

So far, the focus has been implicitly on quantitative continuous dependent and explanatory variables in linear regression models (with an unrestricted range) such as income, age, or yearsof education. However, in empirical analyses, qualitative explanatory variables often play an important role such as gender, race, sectoral effects, or regional effects.

Qualitative variables:

- Qualitative information in explanatory variables can be captured by corresponding binary (i.e. dummy) variables, which have exactly two possible categories and thus take the values one or zero
- The OLS estimation and the testing of hypotheses about regression parameters in linear regression models with qualitative explanatory variables is fully equivalent to the case with the exclusive inclusion of quantitative explanatory variables

Single binary explanatory variables:

Inclusion of qualitative variables with two categories

Based on a multiple linear regression model with only quantitative explanatory variables, an additional binary explanatory variable x_0 is included (in addition to now k-1 quantitative explanatory variables x_1, x_2,…, x_{k-1}):

$$y = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_2 + \cdots + \beta_k x_{k-1} + \varepsilon$$

With $E(\varepsilon|x_0,x_1,x_2,\ldots, x_{k-1}) = 0$, it follows:

$$E(y|x_0, x_1, x_2,\ldots, x_{k-1}) = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_2 + \cdots + \beta_k x_{k-1}$$

It follows:

$$\beta_1 = E(y|x_0 = 1, x_1, x_2,\ldots, x_{k-1}) - E(y|x_0 = 0, x_1, x_2,\ldots, x_{k-1})$$

β_1 thus is the difference in the expected value of y between $x_0 = 1$ und $x_0 = 0$, given the same values of $x_1, x_2,\ldots, x_{k-1}$ and ε.

β_0 is therefore the constant for $x_0 = 0$. For $x_0 = 1$, the constant is $\beta_0 + \beta_1$ so that β_1 is the difference in the constant for $x_0 = 1$ und $x_0 = 0$.

Remark:

For a single binary factor (e.g. gender) it is not possible to include two dummy variables simultaneously (e.g. one variable that takes the value one for females and another variable that takes the value one for non-females) in linear regression models since this would lead to perfect collinearity (simple version of the „dummy variable trap“)

Binary explanatory variables for multiple categories:

Inclusion of qualitative variables with more than two categories

The basis is again a multiple linear regression model with only quantitative variables. Now, an additional qualitative (nominal or ordinal) explanatory variable (e.g. sector, region, education, gender with more than two categories) with $q > 2$ different categories is considered. In this case, (up to) q-1 dummy variables $x_{01}, x_{02},\ldots, x_{0,q-1}$ can be included (in addition to now k-q+1 quantitative explanatory variables $x_1, x_2,\ldots, x_{k-q+1}$):

$$y = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_{q-1} x_{0,q-1} + \beta_q x_1 + \beta_{q+1} x_2 + \cdots + \beta_k x_{k-q+1} + \varepsilon$$

Category q of the qualitative variable (i.e. the dummy variable x_{0q}) is treated as the base category. As a consequence, the estimated slope parameters $\hat{\beta}_1, \hat{\beta}_2,\ldots, \hat{\beta}_{q-1}$ indicate the estimated average difference in the dependent variable y for each corresponding group of the qualitative variable (i.e. for $x_{01}, x_{02},\ldots, x_{0,q-1}$) compared with the base category, i.e. compared with x_{0q}.

Remark:

It is not possible to jointly include all q dummy variables $x_{01}, x_{02},\ldots, x_{0q}$ since this would lead to perfect collinearity (general version of the „dummy variable trap“). Most econometric packages including Stata automatically correct for this issue by omitting one category, i.e. one of the q dummy variables.

Example: Determinants of (gross) income

As before, we consider linear regression models to analyze monthly gross income. However, in addition to educ, age, and hhsize, the dummy variable fulltime that takes the value one if a person works full-time, and the dummy variable female that takes the value one if the worker is female, are included to explain income. Furthermore, we consider the ordinal variable health, which takes the values from one to five, with five indicating the highest level for subjective health. On this basis, five dummy variables for the five categories of health were generated (health1, health2, health3, health4, health5).The OLS estimations of two regression models with Stata lead to the following results:

```
reg income educ age hhsize fulltime female health5
```

Source	SS	df	MS	Number of obs	=	1,652
Model	2.3084e+09	6	384734766	F(6, 1645)	=	161.86
Residual	3.9102e+09	1,645	2377013.77	Prob > F	=	0.0000
Total	6.2186e+09	1,651	3766563.44	R-squared	=	0.3712
				Adj R-squared	=	0.3689
				Root MSE	=	1541.8

income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	296.8215	13.57365	21.87	0.000	270.1981	323.445
age	28.92672	3.589096	8.06	0.000	21.88704	35.9664
hhsize	99.65515	27.91729	3.57	0.000	44.89798	154.4123
fulltime	1318.201	98.0169	13.45	0.000	1125.95	1510.452
female	-751.2202	87.00182	-8.63	0.000	-921.8662	-580.5742
health5	379.5957	124.0253	3.06	0.002	136.3316	622.8598
_cons	-2797.88	285.7242	-9.79	0.000	-3358.302	-2237.459

Example: Determinants of (gross) income (continued)

```
reg income educ age hhsize fulltime female health2 health3 health4 health5
```

Source	SS	df	MS	Number of obs	=	1,652
Model	2.3268e+09	9	258534940	F(9, 1642)	=	109.08
Residual	3.8918e+09	1,642	2370147.25	Prob > F	=	0.0000
Total	6.2186e+09	1,651	3766563.44	R-squared	=	0.3742
				Adj R-squared	=	0.3707
				Root MSE	=	1539.5

income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	293.3754	13.63442	21.52	0.000	266.6327	320.1181
age	29.87959	3.615005	8.27	0.000	22.78908	36.9701
hhsize	98.05561	27.90375	3.51	0.000	43.32493	152.7863
fulltime	1318.278	97.89543	13.47	0.000	1126.265	1510.291
female	-747.0686	86.89066	-8.60	0.000	-917.4968	-576.6404
health2	631.7705	356.5834	1.77	0.077	-67.6357	1331.177
health3	549.7177	342.9262	1.60	0.109	-122.9011	1222.336
health4	728.0416	340.857	2.14	0.033	59.4813	1396.602
health5	1033.307	356.5223	2.90	0.004	334.0204	1732.593
_cons	-3440.015	439.2034	-7.83	0.000	-4301.473	-2578.557

Example: Determinants of (gross) income (continued)

Interpretation:

- The estimated monthly gross income is on average about 1,318 Euro higher for full-time workers than for part-time workers
- The estimated monthly gross income is on average 751.220 or 747.069 Euro lower for female workers than for non-female workers
- The estimated monthly gross income is on average 379.596 Euro higher for workers with the highest health level than for workers with the four lower health levels
- All these estimated effects are highly significant
- The estimated monthly gross income is on average 1,033.307 Euro higher for workers with the highest health level and 631.771 Euro higher for workers with the second lowest health level compared with workers with the lowest health level (= base category), respectively
- The latter estimated effect is only weakly significant

Interaction terms with binary explanatory variables:

Interaction terms can also involve dummy variables (and thus do not necessarily refer only to two quantitative explanatory variables)

Additional inclusion of an interaction term between two binary explanatory variables x_{01} and x_{02} (in addition to now k-3 quantitative explanatory variables x_1, x_2,…, x_{k-3}):

$$y = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_3 x_{01} x_{02} + \beta_4 x_1 + \beta_5 x_2 + \cdots + \beta_k x_{k-3} + \varepsilon$$

Interpretation:

- The inclusion of such interaction terms (in addition to the separate inclusion of the corresponding dummy variables) is an alternative to the inclusion of three binary explanatory variables when four categories are examined
- $\hat{\beta}_1$ (resp. $\hat{\beta}_2$) indicates for $x_{02} = 0$ (resp. $x_{01} = 0$) the estimated average difference in the dependent variable y between $x_{01} = 1$ and $x_{01} = 0$ (resp. between $x_{02} = 1$ and $x_{02} = 0$)
- $\hat{\beta}_3$ indicates the additional estimated effect for $x_{01} = 1$ and $x_{02} = 1$, i.e. the estimated average difference in the effect of x_{01} on y between $x_{02} = 1$ and $x_{02} = 0$ (resp. in the effect of x_{02} on y between $x_{01} = 1$ and $x_{01} = 0$).
- For $x_{01} = 1$ and $x_{02} = 0$ (resp. for $x_{01} = 0$ and $x_{02} = 1$), the estimated constant is $\hat{\beta}_0 + \hat{\beta}_1$ (resp. $\hat{\beta}_0 + \hat{\beta}_2$)
- For $x_{01} = 1$ and $x_{02} = 1$, the estimated constant is $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$

Example: Determinants of net income

As before, we consider a linear regression model to analyze monthly net income. However, in addition to age, tenure, and workinghours, we include the dummy variables female and east that take the value one if the worker is female or lives in Eastern Germany, respectively, as well as the interaction of female and east (femaleeast) to explain nincomemonth. The OLS estimation with Stata leads to the following results:

```
reg nincomemonth age tenure workinghours female east femaleeast
```

Source	SS	df	MS	Number of obs	=	1,456
Model	803415873	6	133902645	F(6, 1449)	=	243.24
Residual	797683626	1,449	550506.298	Prob > F	=	0.0000
Total	1.6011e+09	1,455	1100412.03	R-squared	=	0.5018
				Adj R-squared	=	0.4997
				Root MSE	=	741.96

nincomemonth	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	8.627047	1.90543	4.53	0.000	4.88935	12.36474
tenure	21.52807	2.133708	10.09	0.000	17.34258	25.71356
workinghours	54.61518	1.890608	28.89	0.000	50.90656	58.3238
female	-277.8267	45.60745	-6.09	0.000	-367.2904	-188.363
east	-456.888	66.62907	-6.86	0.000	-587.5877	-326.1882
femaleeast	241.7843	95.9221	2.52	0.012	53.62323	429.9453
_cons	-439.8791	112.7948	-3.90	0.000	-661.1377	-218.6205

Example: Determinants of net income (continued)

Interpretation:

- For workers in Western Germany (east = 0), the estimated average monthly net income of females is 277.827 Euro lower than that of non-females
- For non-female workers (female = 0), the estimated average monthly net income in Eastern Germany is 456.888 Euro lower than in Western Germany
- The estimated average difference in the gender effect between Eastern and Western Germany is 241.784 Euro, i.e. the estimated (absolute) average difference in monthly net income between females and non-females is 241.784 Euro smaller in Eastern Germany than in Western Germany
- Similarly, the estimated average difference in monthly net income between Eastern and Western Germany is about 241.784 Euro smaller among females than among non-females
- The t-values of -6.09 (female), -6.86 (east), and 2.52 (femaleeast) indicate that all three estimated effects are different from zero at the 1% or 5% significance levels, with the first two showing significantly negative effects and the latter a significantly positive interaction effect

Additional inclusion of an interaction term between a binary explanatory variables x_0 and a quantitative explanatory variable x_1 (in addition to now k-2 quantitative explanatory variables x_1, x_2,…, x_{k-2}):

$$y = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_0 x_1 + \beta_4 x_2 + \cdots + \beta_k x_{k-2} + \varepsilon$$

Interpretation:

- This interaction term allows the (partial) effect (or elasticity or semi-elasticity) of the quantitative explanatory variable x_1 on the dependent variable y to differ across the two categories of the binary explanatory variable x_0 in linear regression models
- If $\beta_3 = 0$, there is no difference in the effect of x_1 between the two groups defined by x_0
- $\hat{\beta}_1$ indicates for $x_1 = 0$ the estimated average difference in the dependent variable y between $x_0 = 1$ and $x_0 = 0$
- $\hat{\beta}_2$ indicates the estimated (partial) effect of the quantitative variable x_1 when $x_0 = 0$
- $\hat{\beta}_3$ indicates the additional estimated (partial) effect of the quantitative variable x_1 when $x_0 = 1$, i.e. the estimated average difference in the effect of x_1 on y between $x_0 = 1$ and $x_0 = 0$

Example: Determinants of (gross) income

As before, we consider a linear regression model to analyze monthly gross income. However, in addition to educ, age, hhsize, fulltime, female, and health5, the interaction of educ and female (educfemale) is included to explain income. The OLS estimation of the regression model with Stata leads to the following results:

```
reg income age hhsize fulltime health5 educ female educfemale
```

Source	SS	df	MS	Number of obs	=	1,652
Model	2.3719e+09	7	338840280	F(7, 1644)	=	144.81
Residual	3.8467e+09	1,644	2339850.54	Prob > F	=	0.0000
Total	6.2186e+09	1,651	3766563.44	R-squared	=	0.3814
				Adj R-squared	=	0.3788
				Root MSE	=	1529.7

income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	27.74351	3.568168	7.78	0.000	20.74488	34.74215
hhsize	97.7236	27.70068	3.53	0.000	43.39127	152.0559
fulltime	1328.887	97.2693	13.66	0.000	1138.102	1519.671
health5	348.2585	123.1989	2.83	0.005	106.6151	589.9019
educ	363.7433	18.61337	19.54	0.000	327.2349	400.2517
female	1078.504	361.754	2.98	0.003	368.9568	1788.051
educfemale	-139.6717	26.81678	-5.21	0.000	-192.2703	-87.07302
_cons	-3597.947	322.426	-11.16	0.000	-4230.356	-2965.539

Example: Determinants of (gross) income (continued)

Interpretation:

- For female workers with zero years of education (educ = 0), the estimated average monthly gross income is 1,078.504 Euro higher than that of non-female workers with zero years of education
- For non-female workers (female = 0), an increase of educ by one year leads to an estimated increase of monthly gross income by 363.743 Euro
- The estimated interaction parameter implies that the estimated increase of monthly gross income if educ increases by one year is 139.672 Euro smaller for female workers than for non-female workers
- Therefore, an increase of educ by one year leads to an estimated increase of monthly gross income by 363.743 - 139.672 = 224.072 Euro for female workers (female = 1)
- The t-values of 2.98 (female), 19.54 (educ), and -5.21 (educfemale) indicate that all three estimated effects are different from zero at all common significance levels, with the first two showing significantly positive effects and the latter a significantly negative interaction effect

---

2.7 Heteroskedasticity

Previously, assumption A5 (homoskedasticity), which refers to the variance of the OLS estimators, was discussed in detail:

- If $Var(\varepsilon|x_1, x_2,\ldots, x_k) \neq \sigma^2$, this leads to heteroskedasticity
- Unlike e.g. omitting relevant explanatory variables, heteroskedasticity has no impact on the unbiasedness or consistency of OLS estimators. However, heteroskedasticity affects the (estimated) variance of the OLS estimated slope parameters in linear regression models.
- As discussed above, in the case of homoskedasticity, i.e. under the assumptions A1 to A5, it follows for the variance of the OLS estimated slope parameters (with R_h^2 denoting the coefficient of determination from the auxiliary regression of x_h on all other explanatory variables):

$$Var(\hat{\beta}_h) = \frac{\sigma^2}{(1-R_h^2)\sum_{i=1}^{n}(x_{ih} - \bar{x}_h)^2} = \frac{\sigma^2}{(1-R_h^2)SST_h} \quad \text{for } h = 1,\ldots, k$$

- Thus, in the case of homoskedasticity, the estimated standard deviation with a consistent estimator of the standard deviation σ is:

$$\sqrt{\hat{Var}(\hat{\beta}_h)} = \frac{\hat{\sigma}}{\sqrt{(1-R_h^2)SST_h}} \quad \text{for } h = 1,\ldots, k$$

- Since this variance is valid only in the case of homoskedasticity, the corresponding estimated standard deviation becomes a biased and inconsistent estimator of the true standard deviation of the OLS estimators in the presence of heteroskedasticity
- Under heteroskedasticity, the estimated standard deviations are thus no longer valid for constructing confidence intervals or for calculating t- and F-statistics. The corresponding t-statistics are therefore no longer t-distributed (even for a large sample size n), and the corresponding F-statistics are no longer F-distributed in the case of heteroskedasticity.
- Finally, under heteroskedasticity, the desirable BLUE (or efficiency) property of OLS estimators, as well as their property of asymptotic efficiency, no longer hold. However, if the form of heteroskedasticity is known, it is possible to construct more efficient estimators than the usual OLS estimators.
- Under heteroskedasticity, the general question arises whether an alternative estimation method should really be used instead of the OLS estimation: Since the OLS estimators remain unbiased and consistent in the presence of heteroskedasticity (under the assumptions A1 to A4), the application of OLS can still be useful.
- However, for the construction of confidence intervals and the application of t- or F-tests under heteroskedasticity, the estimated standard deviations of the OLS estimators should at least be corrected

The starting point for these corrections are the actual (but unknown) variances of the OLS estimators. The unknown variances σ_i^2 of the error term ε_i are replaced by the corresponding squared residuals $\hat{\varepsilon}_i^2$ (which are obtained from the original OLS estimation). In multiple linear regression models, the estimated variance of the OLS estimated slope parameters generally is:

$$\widehat{\mathrm{Var}}(\hat{\beta}_h) = \frac{\sum_{i=1}^{n} \hat{r}_{ih}^2 \hat{\varepsilon}_i^2}{\mathrm{SSR}_h^2}$$

$\hat{r}_{ih}$ denotes the residual of observation i from the auxiliary regression of x_h on all other explanatory variables, and SSR_h denotes the corresponding sum of squared residuals from this regression. The estimated standard deviation of the OLS estimated slope parameters according to White (1980) is then:

$$\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_h)} = \frac{\sqrt{\sum_{i=1}^{n} \hat{r}_{ih}^2 \hat{\varepsilon}_i^2}}{\mathrm{SSR}_h}$$

On this basis, several further asymptotically equivalent estimators of standard deviations have been developed, especially by including the small-sample correction factor n/(n-k) in the estimated variance in Stata. By using these corrected estimated standard deviations, heteroskedasticity-robust confidence intervals and especially t-statistics can be constructed.

Example: Determinants of (gross) income

As before, we consider a linear regression model to analyze monthly gross income, including educ, age, hhsize, and educage as explanatory variables. The following OLS regression equation was estimated, with heteroskedasticity-robustly estimated standard deviations of the estimated parameters (in brackets) now also reported in addition to the conventionally estimated standard deviations (in parentheses):

$$\widehat{income} = 1986.282 + 52.923\,hhsize - 5.672\,educ - 51.497\,age + 6.023\,educage$$

(897.219) (30.919) (66.995) (18.821) (1.421)

[763.059] [36.117] [62.234] [16.774] [1.391]

Example: Determinants of (gross) income (Stata output)

The OLS estimation of the regression model with heteroskedasticity-robustly estimated standard deviations of the estimated parameters with Stata leads to the following results:

```
reg income hhsize educ age educage, robust

Linear regression                               Number of obs     =      1,652
                                                F(4, 1647)        =      79.89
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1956
                                                Root MSE          =     1742.8

------------------------------------------------------------------------------
             |               Robust
      income | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      hhsize |     52.923   36.11676     1.47   0.143    -17.91662    123.7626
        educ |  -5.671606   62.23355    -0.09   0.927    -127.7368    116.3936
         age |  -51.49744   16.77361    -3.07   0.002    -84.39728   -18.59759
     educage |   6.023479   1.390927     4.33   0.000     3.295308     8.75165
       _cons |   1986.282    763.059     2.60   0.009     489.6144     3482.95
------------------------------------------------------------------------------
```