0. Introductory econometrics

0.1 Structure of economic data

Cross-sectional data:

- Data which are collected from units of the underlying population at a given time period (which may vary occasionally) (the arrangement of the units in the dataset is irrelevant)
- Starting point is mostly the implicit assumption that the data have been collected by random sampling
- Examples: Individual or household data (e.g. income), firm data (e.g. sales), city or country data (e.g. unemployment)

Example:

Observation number	Country	Population density	GDP per capita	Labor force (rural)	GDP growth	Birth rate	Net migration
1	A	212.4	20116	9.8	53	8.4	-0.7
2	B	623.7	24966	3.4	73.1	6.1	3.4
3	C	93.1	19324	23.6	47.9	12.3	-1.9
:	:	:	:	:	:	:	:
10	J	287.4	23136	8.8	59.4	12.4	1.7
11	K	166.2	20707	14.1	74	13	3.6
12	L	388.1	23624	9.6	54.3	6.9	-0.4

Time series data:

- Data which are collected for one or more variables during several successive time periods
- Time is an important dimension (i.e. observations are often correlated over time) so that the arrangement of the observations in the data set contains potentially important information
- The frequency of data collection over time may vary strongly, e.g. daily, weekly, monthly, quarterly, and annual data with possible seasonal effects
- Examples: Macroeconomic data (e.g. income, consumption, investments, supply of money, price index), financial market data (e.g. stock prices)

Example:

Observation number	Year	Inflation US	Unemployment rate US
1	1948	8.1	3.8
2	1949	-1.2	5.9
3	1950	1.3	5.3
4	1951	7.9	3.3
:	:	:	:
54	2001	2.8	4.7
55	2002	1.6	5.8

Pooled cross-sectional data:

- Data which have cross-sectional as well as time series characteristics since several cross-sectional data sets are collected independently over different time periods and linked to increase the sample size
- Although the arrangement of the observations in the data set is not essential, the corresponding period is recognized as an important variable
- The data are mostly analyzed like conventional cross-sectional data
- Examples: Individual or household data (e.g. income, expenditures) during several years

---

Example:

Observation number	Year	House price	Wealth tax	House size
1	1993	85500	42	1600
2	1993	67300	36	1440
:	:	:	:	:
250	1993	243600	41	2600
251	1995	65000	16	1250
:	:	:	:	:
520	1995	57200	16	1100

Panel data:

- Data which have both a time series and a cross-sectional dimension where (in contrast to pooled cross-sectional data) the same units (e.g. individuals, firms, countries) are observed over several time periods
- The number of units is often much higher than the time dimension
- The data are often first sorted by units and then by periods
- The data provide the opportunity to control for unobserved characteristics of the units and to examine lagged variables
- Examples: Individual or household panel data (e.g. SOEP), firm panel data (e.g. MIP), country panel data

---

Example:

Observation number	Household	Year	Size	Net income	Smoking household
1	1	2000	5	3200	yes
2	1	2005	6	3500	yes
3	2	2000	2	2900	no
4	2	2005	2	3000	no
:	:	:	:	:	:
299	150	2000	3	1793	no
300	150	2005	4	2380	no

0.2 Linear regression models (for cross-sectional data)

Multiple linear regression model:

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

x_1, x_2, x_3,…, x_{k-1}, x_k: Explanatory variables
β_0: Intercept
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:

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

This assumption states that the error term ε is mean independent of the explanatory variables x_1, x_2,…, x_k.

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:

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

The inclusion of the observations i = 1,..., n leads to the following linear regression model:

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

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

Main task of regression analysis:

Estimation of the unknown regression parameters β_0, β_1, β_2,…

Optimization problem in the ordinary method of least squares (OLS method) for the multiple linear regression model:

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

The first order conditions for the k+1 estimated regression parameters are then:

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

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

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

$$\vdots$$

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

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:

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

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

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

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

In this case, the estimated parameter of the explanatory variable x_1 thus indicates the change of the estimated dependent variable if x_1 increases by one.

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

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

In this case, the estimated parameter of the explanatory variable x_k thus indicates the change of the estimated dependent variable if x_k increases by one.

Therefore, the estimated parameters can be interpreted as estimated partial effects, i.e. the estimated effect of an explanatory variable implies that it is controlled for all other explanatory variables. This partial effect interpretation is the strong advantage of regression analyses (and of econometric analyses in general), i.e. a ceteris paribus interpretation is generally possible without the necessity of conducting a corresponding controlled experiment.

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

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

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 equals the squared correlation coefficient between the actual 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)
- Therefore, R^2 is a poor measure to evaluate the quality of a linear regression model (also the adjusted coefficient of determination, which takes the number of explanatory variables into account, is not an appropriate measure for evaluating the quality of a linear regression model)

Example: Determinants of (the logarithm of) wages (I)

By using a linear regression model, the effect of the years of education (educ), the years of labor market experience (exper), and the years with the current employer (tenure) on the logarithm of hourly wage (logwage) is examined:

$$\text{logwage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure} + \varepsilon$$

The following OLS regression equation was estimated on the basis of n = 526 workers:

$$\widehat{\text{logwage}} = 0.284 + 0.092\text{educ} + 0.0041\text{exper} + 0.022\text{tenure}$$

Interpretation:

- Estimated positive effect of the years of education: If exper and tenure are held fixed, an increase of the years of education by one year leads to an estimated increase of the logarithm of wage by 0.092
- Exper and tenure (as expected) have also positive estimated effects, if the other explanatory variables are held fixed, respectively

---

Example: Determinants of (the logarithm of) wages (II)

```
reg logwage educ exper tenure
```

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   80.39
       Model |  46.8741806     3  15.6247269           Prob > F      =  0.0000
    Residual |  101.455582   522  .194359353           R-squared     =  0.3160
-------------+------------------------------           Adj R-squared =  0.3121
       Total |  148.329763   525  .282532881           Root MSE      =  .44086

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .092029   .0073299    12.56   0.000     .0776292    .1064288
       exper |   .0041211   .0017233     2.39   0.017     .0007357    .0075065
      tenure |   .0220672   .0030936     7.13   0.000     .0159897    .0281448
       _cons |   .2843595   .1041904     2.73   0.007     .0796755    .4890435
------------------------------------------------------------------------------
```

0.3 Expected value and variance of OLS estimators

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

- Assumption A1: Linear in parameters
  The relationship between the dependent variable y and the explanatory variables x_1, x_2,…, 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 with n observations 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 no exact linear relationship between the explanatory variables exists
- Assumption A4: Zero conditional mean
  $E(\varepsilon | x_1, x_2, \ldots, x_k) = 0$

Under these four assumptions all OLS estimators are unbiased:

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

Remark about assumption A4:

If this assumption holds, the explanatory variables are characterized as exogenous. In contrast, if it is violated, endogenous variables or endogeneity are present.

- A4 is e.g. violated in the case of measurement errors in the explanatory variables or if the functional relationship between the dependent and explanatory variables is misspecified
- One of the major violations of A4 is the omission of a relevant explanatory variable, which is correlated with other explanatory variables

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

The correct linear regression model is as follows (where assumptions A1 through A4 hold):

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

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

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

The OLS regression equations in the correct and misspecified linear regression models are as follows:

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

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

In this case, the following relationship exists (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, even when the direction of the bias is ambiguous. The estimator is only unbiased if β_k or $\tilde{\delta}_h$ is zero. If $\tilde{\delta}_h$ is zero, x_h and x_k are uncorrelated in the sample.

In contrast:

The inclusion of irrelevant explanatory variables (i.e. one or more explanatory variables that have no partial effect on the dependent variable) has no impact on the unbiasedness of the OLS estimators and thus does not lead to biases

→ However, the inclusion of irrelevant explanatory variables has an impact on the variance of the OLS estimators

Assumptions for the analysis of the variance of OLS estimators:

- Assumptions A1 through A4 from the analysis of the expected value of OLS estimators
- Assumption A5: Homoskedasticity
  The conditional variance of the error term ε is constant, i.e.
  $Var(\varepsilon|x_1, x_2,\ldots, x_k) = \sigma^2$. If the assumption is violated, i.e. if the variance depends on the explanatory variables, this leads to heteroskedasticity.

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

Under the assumptions A1 through 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 regressing x_h on all other explanatory variables (including a constant).

→ While the assumption of homoskedasticity is negligible for the unbiasedness of the estimated parameters, the above variance is only true with the homoskedasticity 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 the estimation of the variance of the OLS estimated regression parameters

Since $\sigma^2 = E(\varepsilon^2)$, the following estimator would be obvious:

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

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

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

Thus, the corresponding (consistent, but not unbiased) standard error of the regression (SER), which is an estimator of the standard deviation σ of the error term ε, 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:

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

Standard deviation of the OLS estimated slope parameters:

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

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 use of these estimates (also called standard errors of the estimated parameters) is particularly based on the homoskedasticity assumption A5. In contrast, the variance of the OLS estimated slope parameters is estimated with a bias in the case of heteroskedasticity (although heteroskedasticity has no influence on the unbiasedness of the estimated regression parameters).

Under the assumptions A1 through 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“ indicates that the estimator is not biased
- “Linear“ indicates that the estimator is a linear function of the data and the dependent variable
- “Best“ indicates that the estimator has the smallest 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 through A5 hold.

0.4 Testing of hypotheses about regression parameters

Additional assumption A6: Normality

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

→ The assumptions A1 through A6 are also called classical linear model assumptions. Therefore, the corresponding approach is also called classical linear regression model.

Under the assumptions A1 through A6 it follows for the dependent variable:

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

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 smallest variance not only in the class of all linear unbiased estimators, but also in the larger class of all unbiased estimators.

→ Indeed, if the error term ε is not normally distributed, the realization of statistical tests is no problem if the sample size n is large

If assumption A6 (normally distributed error term ε) holds, 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, each linear function of the OLS estimated regression parameters $\hat{\beta}_0, \hat{\beta}_1,\ldots, \hat{\beta}_k$ is normally distributed.

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

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

k+1 is the number of unknown regression parameters.

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 in this case is the following t statistic (t value), which includes the estimated standard deviation (standard error) of the estimated parameters:

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

The testing of H_0: $\beta_h = 0$ at a given significance level is based on the property that the t statistic is t distributed with n-k-1 degrees of freedom under the null hypothesis. In empirical analyses a two-tailed test is usually examined. The two-sided alternative hypothesis is:

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

The null hypothesis is thus rejected if:

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

More general null hypothesis:

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

The null hypothesis is rejected if $\hat{\beta}_h$ is strongly different 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 the 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: Effect of air pollution on housing prices (l)

By using a linear regression model, the effect of the logarithm of nitrogen oxides (in ppm) in the air (lognox), the logarithm of the weigthed distances (in miles) from five employment centers (logdist), the average number of rooms in houses (rooms), and the average ratio between teachers and pupils in schools (stratio) on the logarithm of the median housing prices (logprice) is examined with a sample of n = 506 communities. The following OLS regression equation was estimated ($R^2 = 0.584$):

$$\widehat{\log price} = 11.08 - 0.954\,\log nox - 0.134\,\log dist + 0.255\,rooms - 0.052\,stratio$$
$$(0.32)\quad (0.117)\qquad (0.043)\qquad (0.019)\qquad (0.006)$$

Due to the high common t values, all explanatory variables have an effect at common significance levels (e.g. 0.05, 0.01). Another interesting null hypothesis refers to the testing whether β_1 equals the value -1, i.e. H_0: $\beta_1 = -1$. In this case it follows t = (-0.954+1)/0.117 = 0.393. Therefore, the null hypothesis cannot be rejected at common significance levels (i.e. the estimated elasticity is not significantly different from the value -1).

Example: Effect of air pollution on housing prices (II)

```
reg logprice lognox logdist rooms stratio
```

```
      Source |       SS       df       MS              Number of obs =     506
-------------+------------------------------           F(  4,   501) =  175.86
       Model |  49.3987581     4  12.3496895           Prob > F      =  0.0000
    Residual |  35.1834907   501  .070226528           R-squared     =  0.5840
-------------+------------------------------           Adj R-squared =  0.5807
       Total |  84.5822488   505  .167489602           Root MSE      =    .265

------------------------------------------------------------------------------
    logprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lognox |  -.9535397   .1167418    -8.17   0.000    -1.182904   -.7241759
     logdist |    -.13434   .0431032    -3.12   0.002    -.2190254   -.0496547
       rooms |    .254527   .0185303    13.74   0.000     .2181203    .2909338
     stratio |  -.0524512   .0058971    -8.89   0.000    -.0640373   -.0408651
       _cons |   11.08386   .3181115    34.84   0.000     10.45887    11.70886
------------------------------------------------------------------------------
```

Also hypotheses about linear combinations of regression parameters can be tested. With arbitrary values $r_1, r_2,\ldots,r_k$ and c the corresponding 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 the 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:

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

An often considered null hypothesis refers to 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 is:

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

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

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

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

In order to test whether a group of q explanatory variables has no effect on the dependent variable, the null hypothesis can be stated 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 restricted regression model under H_0 is then:

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

For the corresponding F test the following F statistic (F value) is considered:

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

If $H_0: \beta_{k-q+1} = \beta_{k-q+2} = \ldots = \beta_k = 0$ is true, 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 the significance level α in favor of the alternative hypothesis if $F > F_{q;n-k-1;1-\alpha}$.

Alternative formulation with the coefficients of determination R^2_r and R^2_{ur} in 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 commonly considered 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:

$$y = \beta_0 + \varepsilon$$

The coefficient of determination R^2_r is zero in such restricted linear regression models so that the (R-squared form of the) F statistic in this case with q = k restrictions is (with R^2 as the ordinary coefficient of determination in a 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: Determinants of birth weigths (I)

By using a linear regression model, the effect 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 number of years of schooling for the mother (motheduc), and the number of years of schooling of the father (fatheduc) on the birth weight (in ounces = 28.3495 grams) of children (bwght) is examined:

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

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

- For n = 1191 births the unrestricted and restricted regression models are estimated by OLS. It follows $R^2_r = 0.0364$ and $R^2_{ur} = 0.0387$.
- Since n-k-1 = 1191 - 6 = 1185 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 1185 degrees of freedom is $F_{2;1185;0.95} = 3.00$. Therefore, the null hypothesis cannot be rejected at the 5% significance level.

Example: Determinants of birth weigths (II)

```
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: Determinants of birth weigths (III)

```
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 directly after the OLS estimation in the unrestricted regression model, differences are due to roundings):

```
test motheduc=fatheduc=0

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

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

0.5 Asymptotic properties

Definition of consistency:
Let W_n be the estimator of 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 for $n \rightarrow \infty$. In this case W_n converges stochastically to θ, i.e. $\text{plim}(W_n) = \theta$.

Consistency of OLS estimators:

- If the assumptions A1 through A4 hold, the OLS estimators $\hat{\beta}_h$ (h = 0,1,…, k) in linear regression models are consistent estimators of β_h, i.e. $\text{plim}(\hat{\beta}_h) = \beta_h$
- For the consistency of OLS estimators the same assumptions as for the unbiasedness are therefore required, i.e. e.g. the assumption A5 (homoskedasticity) can be violated. In fact, for the consistency of OLS estimators already a weakening of A4 is sufficient in addition to the assumptions A1 through A3, i.e. A4‘: $E(\varepsilon) = 0$ and $\text{Cov}(x_h, \varepsilon) = 0$ (h = 1, 2,…, k).

Inconsistency of OLS estimators:

- Remember: 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 not unbiased
- In the same manner, all OLS estimators are inconsistent if ε is correlated with any explanatory variable, i.e. if A4‘ is violated

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 t and F statistics) is based on assumption A6, i.e. $\varepsilon \sim N(0; \sigma^2)$. However, functions of the OLS estimators can also be asymptotically normally distributed if A6 is violated.

If the assumptions A1 through A5 hold, 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 according to which this function is exactly t distributed with n-k-1 degrees of freedom if the assumptions A1 through A6 hold since also the following notation is feasible (since the t distribution converges to the standard normal distribution if 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 for the case that the error term ε is not normally distributed, the previously considered t and F tests can be conducted and the confidence intervals can be constructed. The prerequisite for this is that the sample size n is sufficiently large. For small n (or a small number of degrees of freedom n-k-1), e.g. the approximation of the t statistic towards the t distribution is insufficient.

Asymptotic efficiency:

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

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

0.6 The structure of dependent and explanatory variables

Logarithmized und squared variables:

Linear regression models can comprise non-linear relationships by including (naturally) logarithmized und 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: Effect of air pollution on housing prices

By using a linear regression model, the effect of the logarithm of nitrogen oxides (in ppm) in the air (lognox) and the average number of rooms in houses (rooms) on the logarithm of the median housing prices (logprice) is examined with a sample of n = 506 communities. The OLS estimation with STATA leads to the following results ($R^2 = 0.514$):

```
reg logprice lognox rooms
```

logprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lognox	-.7176732	.0663397	-10.82	0.000	-.8480102	-.5873361
rooms	.3059183	.0190174	16.09	0.000	.268555	.3432816
_cons	9.233737	.1877406	49.18	0.000	8.864885	9.602589

It follows:

- An increase of nitrogen oxides by 1% (i.e. $\%\Delta nox = 1$) leads to an approximately estimated reduction of the median housing prices by 0.718% (if the variable rooms is held fixed)
- An increase of the average number of rooms by one (i.e. $\Delta rooms = 1$) leads to an approximately estimated increase of the median of the real estate prices by $0.306 \cdot 100 = 30.6\%$ (if nox is 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 (besides 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 does not indicate alone 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 constant, 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 also depends on $\hat{\beta}_2$ and the values of x_1.

Interaction terms:

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

Additional inclusion of an interaction term for x_1 and x_2 (besides the k-1 explanatory variables x_1, x_2,…, x_{k-1}):

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

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

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

If x_2,…, x_{k-1} are held constant, it follows:

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

Therefore, the estimated (partial) marginal effect of x_1 on y also depends on $\hat{\beta}_3$ and x_2. In this case, interesting values of x_2 are generally examined (e.g. the mean of x_2 in the sample). $\hat{\beta}_1$ alone only indicates the estimated (partial) effect of x_1 if x_2 is zero.

Qualitative explanatory variables:

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

Qualitative variables:

- Qualitative information on explanatory variables can be captured by corresponding binary (i.e. dummy) variables that have exactly two possible categories and thus take two values, namely one and 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 exclusive inclusion of quantitative explanatory variables

Single binary explanatory variables:

Inclusion of qualitative variables with two categories

On the basis of a multiple linear regression model with only quantitative explanatory variables, an additional binary explanatory variable x_0 is included (besides 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 of the constant for $x_0 = 1$ und $x_0 = 0$.

Note:

For one factor (e.g. gender) it is not possible to jointly include two dummy variables (e.g. one variable that takes the value one for women and one variable that takes the value one for men) in linear regression models since this would lead to perfect collinearity (simple version of „dummy variable trap“)

Example: Determinants of (the logarithm of) wages

By using a linear regression model, the effects of gender (female), the years of education (educ), the years of labor market experience (exper), the squared years of labor market experience (expersq), the years with the current employer (tenure) and the squared years with the current employer (tenuresq) on the logarithm of hourly wage (logwage) is examined. On the basis of n = 526 workers, the OLS estimation with STATA leads to the following results ($R^2 = 0.441$):

```
reg logwage female educ exper expersq tenure tenuresq
```

logwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.296511	.0358055	-8.28	0.000	-.3668524	-.2261696
educ	.0801967	.0067573	11.87	0.000	.0669217	.0934716
exper	.0294324	.0049752	5.92	0.000	.0196584	.0392063
expersq	-.0005827	.0001073	-5.43	0.000	-.0007935	-.0003719
tenure	.0317139	.0068452	4.63	0.000	.0182663	.0451616
tenuresq	-.0005852	.0002347	-2.49	0.013	-.0010463	-.0001241
_cons	.4166909	.0989279	4.21	0.000	.2223425	.6110394

The results imply that the estimated hourly wage for women is on average approximately $100 \cdot 0.297 = 29.7\%$ smaller (for equal education, labor market experience, and years with the current employer).

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) with q > 2 different categories is considered. In this case (maximal) q-1 dummy variables x_{01}, x_{02},…, $x_{0,q-1}$ can be included (besides now k-q+1 quantitative explanatory variables x_1, x_2,…, 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 considered as the base category. As a consequence, the estimated slope parameters $\hat{\beta}_1$, $\hat{\beta}_2$,…, $\hat{\beta}_{q-1}$ indicate for the corresponding group of the qualitative variable (i.e. for x_{01}, x_{02},…, $x_{0,q-1}$) the estimated average difference in the dependent variable y compared with the base category, i.e. compared with x_{0q}.

Note:

It is not possible to jointly include all q dummy variables x_{01}, x_{02},…, x_{0q} since this would lead to perfect collinearity (general version of „dummy variable trap“). Many econometric packages such as STATA automatically correct for this mistake.

Interaction terms with binary explanatory variables:

Interaction terms can also comprise dummy variables (and thus need not only refer to two quantitative explanatory variables)

Additional inclusion of an interaction term for two binary explanatory variables x_{01} und x_{02} (besides 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 these interaction terms (besides the separate inclusion of the corresponding dummy variables) is an alternative to the inclusion of three binary explanatory variables if 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 $x_{02} = 0$)
- 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 $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$

Inclusion of an interaction term for one binary explanatory variables x_0 and one quantitative explanatory variable x_1 (besides 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:

- These interaction terms allow the analysis of possible differences in the (partial) effect (or elasticity or semi elasticity) of the quantitative explanatory variable x_1 in linear regression models for the two categories of the binary explanatory variable x_0. If $\beta_3 = 0$, then there is no difference.
- If $x_0 = 0$, it follows for the OLS regression equation:

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

  The estimated constant in this case is $\hat{\beta}_0$ and the estimated (partial) effect of x_1 is $\hat{\beta}_2$.
- If $x_0 = 1$, it follows for the OLS regression equation:

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

  The estimated constant in this case is $\hat{\beta}_0 + \hat{\beta}_1$ and the estimated (partial) effect of x_1 is $\hat{\beta}_2 + \hat{\beta}_3$.

0.7 Heteroskedasticity

Previously, for the variance of the OLS estimators assumption A5 (homoskedasticity) 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 has an impact on the (estimated) variance of the OLS estimated slope parameters in linear regression models.
- As discussed above, it follows in the case of homoskedasticity, i.e. under the assumptions A1 through A5, for the variance of the OLS estimated slope parameters (with R_h^2 as the coefficient of determination of a 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, it follows the following estimated standard deviation with a consistent estimator of the standard deviation σ:

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

- Since this variance is only true in the case of homoskedasticity, but not under heteroskedasticity, this estimated standard deviation is a biased and inconsistent estimator of the standard deviation of the OLS estimators
- In the case of heteroskedasticity the estimated standard deviations are thus no longer valid for constructing confidence intervals as well as 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, the desirable BLUE property (or efficiency) of OLS estimators and the property of asymptotic efficiency are not valid in the case of heteroskedasticity. However, by knowing the form of heteroskedasticity, it is possible to construct more efficient estimators compared to the OLS estimators.

A standard approach to test homoskedasticity is the Breusch-Pagan test (an alternative is the White test). The null hypothesis is:

$$H_0: \mathrm{Var}(\varepsilon|x_1, x_2,\ldots,x_k) = \sigma^2 \quad \text{resp.} \quad H_0: E(\varepsilon^2|x_1, x_2,\ldots,x_k) = E(\varepsilon^2) = \sigma^2$$

If H_0 is not true, ε^2 is a function of one or more explanatory variables. If a linear function of all explanatory variables is considered, it follows in this case with an error term v with a (conditional) expected value of zero:

$$\varepsilon^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + v$$

The null hypothesis of homoskedasticity is then:

$$H_0: \delta_1 = \delta_2 = \cdots = \delta_k = 0$$

Since the ε_i are unknown, they are replaced by the corresponding estimators, i.e. the residuals $\hat{\varepsilon}_i$, so that the squared residuals are regressed on the explanatory variables:

$$\hat{\varepsilon}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + v$$

A high value of the coefficient of determination $R^2_{\hat{\varepsilon}^2}$ in this auxiliary regression suggests the validity of the alternative hypothesis, i.e. heteroskedasticity. One version of the Breusch Pagan test statistic is:

$$BP = nR^2_{\hat{\varepsilon}^2}$$

Under the null hypothesis (i.e. homoskedasticity) BP is asymptotically χ^2 distributed with k degrees of freedom, i.e.:

$$BP \overset{a}{\sim} \chi^2_k$$

Thus, the null hypothesis of homoskedasticity is (for a large sample size n) rejected in favor of the alternative hypothesis of heteroskedasticity at the significance level α if:

$$BP > \chi^2_{k;1-\alpha}$$

Example: Determinants of house prices (I)

By using a linear regression model, the effect of the lot size in square feet (lotsize), the living space size in square feet (sqrft), and the number of bedrooms (bdrms) on house prices in 1000 dollar (price) is examined. The following OLS regression equation was estimated:

$$\widehat{price} = -21.77 + 0.00207 lotsize + 0.123 sqrft + 13.85 bdrms$$
$$(29.48)\ (0.00064)\qquad (0.013)\qquad (9.01)$$

$$n = 88;\ R^2 = 0.672$$

On the basis of a Breusch Pagan test, the null hypothesis of homoskedasticity is tested at the 1% significance level:

- First, the residuals $\hat{\varepsilon}_i$ are calculated. The auxiliary regression of $\hat{\varepsilon}^2$ on lotsize, sqrft, and bdrms leads to a coefficient of determination in the amount of $R^2_{\hat{\varepsilon}^2} = 0.1601$.
- The corresponding Breusch Pagan test statistic amounts to the value of $BP = 88 \cdot 0.1601 = 14.09$
- With k = 3 the critical value is $\chi^2_{3;0.99} = 11.34$. Thus, the null hypothesis is rejected at the 1% significance level (the corresponding p-value is p = 0.0028)

Example: Determinants of house prices (II)

```
reg price lotsize sqrft bdrms
```

```
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =   57.46
       Model |  617130.702     3  205710.234           Prob > F      =  0.0000
    Residual |  300723.806    84  3580.04531           R-squared     =  0.6724
-------------+------------------------------           Adj R-squared =  0.6607
       Total |  917854.508    87  10550.0518           Root MSE      =  59.833

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |   .0020677   .0006421     3.22   0.002     .0007908    .0033446
       sqrft |   .1227782   .0132374     9.28   0.000     .0964541    .1491022
       bdrms |   13.85252   9.010145     1.54   0.128     -4.06514    31.77018
       _cons |  -21.77031   29.47504    -0.74   0.462    -80.38466    36.84404
------------------------------------------------------------------------------
```

Testing command and results in STATA (only possible directly after the OLS estimation):

```
estat hettest, rhs iid
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: lotsize sqrft bdrms

         chi2(3)      =    14.09
         Prob > chi2  =   0.0028
```

If the null hypothesis is rejected at a low significance level and thus heteroskedasticity is verified, this should be considered:

- One possibility is the use of alternative estimation methods instead of the OLS method such as the weighted least squares (WLS) method. However, for this estimation method, it is necessary to know the precise form of heteroskedasticity.
- In the case of heteroskedasticity the general question arises whether an alternative estimation method instead of the OLS estimation should really be applied: Since the OLS estimators are also unbiased and consistent in the case of heteroskedasticity (under the assumptions A1 through A4), the application of OLS can still be useful.
- However, for the construction of confidence intervals and the application of t or F tests in the case of heteroskedasticity, the estimated standard deviations of the OLS estimators should at least be corrected

The starting point of these corrections are the actual (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 stem from the original OLS estimation). In multiple linear regression models the estimated variance of the OLS estimated slope parameters generally is:

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

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

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

On this basis several further asymptotically equivalent estimators of standard deviations have been developed. By using these estimated standard deviations, heteroskedasticity robust confidence intervals and especially heteroskedasticity robust t statistics can be constructed.

Example: Determinants of (the logarithm of) wages (I)

By using a linear regression model, the effect of the years of education (educ), the years of labor market experience (exper), the squared years of labor market experience (expersq), the years with the current employer (tenure), the squared years with the current employer (tenuresq), as well as three combined variables for marital and gender status for married men (marrmale), married women (marrfem), and non-married women (singfem) on the logarithm of hourly wage (logwage) is examined. On the basis of n = 526 workers, the following OLS regression equation was estimated which also reports heteroskedasticity robust estimated standard deviations of the estimated parameters (in brackets) in addition to conventionally estimated standard deviations (with $R^2 = 0.461$):

$$\widehat{\log wage} = 0.321 + 0.213\ marrmale - 0.198 marrfem - 0.110 singfem + 0.0789 educ$$
$$(0.100)\ (0.055) \qquad (0.058) \qquad (0.056) \qquad (0.0067)$$
$$[0.109]\ [0.057] \qquad [0.059] \qquad [0.057] \qquad [0.0074]$$
$$+ 0.0268 exper - 0.00054 expersq + 0.0291 tenure - 0.00053 tenuresq$$
$$(0.0055) \qquad (0.00011) \qquad (0.0068) \qquad (0.00023)$$
$$[0.0051] \qquad [0.00011] \qquad [0.0069] \qquad [0.00024\ ]$$

Example: Determinants of (the logarithm of) wages (II)

```
reg logwage marrmale marrfem singfem educ exper expersq tenure tenuresq, robust

Linear regression                                      Number of obs =     526
                                                       F(  8,   517) =   51.70
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4609
                                                       Root MSE      =  .39329
------------------------------------------------------------------------------
             |               Robust
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    marrmale |   .2126756   .0571419     3.72   0.000     .1004167    .3249345
     marrfem |  -.1982677     .05877    -3.37   0.001    -.3137251   -.0828103
     singfem |  -.1103502   .0571163    -1.93   0.054    -.2225587    .0018583
        educ |   .0789103   .0074147    10.64   0.000     .0643437    .0934769
       exper |   .0268006   .0051391     5.22   0.000     .0167044    .0368967
     expersq |  -.0005352   .0001063    -5.03   0.000    -.0007442   -.0003263
      tenure |   .0290875   .0069409     4.19   0.000     .0154516    .0427234
    tenuresq |  -.0005331   .0002437    -2.19   0.029    -.0010119   -.0000544
       _cons |    .321378    .109469     2.94   0.003     .1063193    .5364368
------------------------------------------------------------------------------
```

0.8 Micro data and microeconometrics

→ The previous analysis focuses at least implicitly on cross-sectional or micro data, i.e. data from persons, households, firms, but also from regions, countries, or even supermarket-scanner data. An important feature of micro data is the independence between observations. Micro data are the basis of microeconometrics.

Microeconometrics:

This direction of empirical analyses uses econometric methods that have been developed to study microeconomic problems, i.e. they are motivated by an economic question and are often based on a microeconomic theory or model to select the dependent and explanatory variables

→ The previous analysis furthermore implicitly focuses on quantitative continuous dependent variables with an unrestricted range. In this case the application of linear regression models and the OLS estimation of the parameters to empirically examine the determinants of a variable is optimal.

However:

Microdata and thus microeconometrics are often not based on quantitative continuous dependent variables with an unrestricted range, but on other types of dependent variables

Qualitative (categorial) variables (which are always discrete):

- Binary variables: These variables have exactly two possible categories (e.g. employment of a person, household ownership of a certain insurance, profits of a firm are higher than a specific amount)
- Multinomial variables: These variables have more than two possible mutually exclusive categories which are not ordered (e.g. employment status of a person, individual choice among several means of transportation, portfolio structure of a household, innovation status of a firm)
- Ordinal (ordered) variables: These variables have more than two possible categories which are ordered (e.g. individual satisfaction with life, personal strength of agreement to a political program, credit rating of a firm)

Quantitative variables which are not continuous or with a restricted range:

- Count variables: These variables are discrete and restricted to non-negative integers (e.g. individual number of visits to a hospital, number of journeys of a household, number of patents of a firm)
- Continuous variables with a restricted range: Non-negative variables (e.g. duration of unemployment, wages), non-negative variables with many zeros (e.g. expenditures for a certain good), truncated variables where realizations below or above a threshold are excluded (e.g. incomes below a threshold), censored variables where values in a certain range are transformed to a single value (e.g. top-coding incomes such as social security earnings)

In the case of discrete dependent variables and continuous dependent variables with a restricted range, the use of linear regression models and the OLS method to estimate the corresponding parameters is not appropriate so that the microeconometric analysis should be adjusted:

- Continuous dependent variables with a restricted range (limited dependent variables):
  The corresponding restrictions should be taken into account since in the case of non-negative dependent variables the OLS fitted values in linear regression models can be outside the allowed range and in the case of the other limited dependent variables the exogeneity assumption A4 is violated so that the OLS estimators are generally biased
- Discrete dependent variables:
  The modeling should be completely shifted from the previous analysis of conditional expected values of dependent variables (or conditional expectation functions) with $E(y|x_1, x_2,\ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ towards conditional probability functions of the discrete dependent variables y
- Parameter estimation:
  In all these cases, the OLS method should be replaced by other estimation methods and particularly by the maximum likelihood method which is based on a parametric distribution of the dependent variable and which is the most important estimation approach in microeconometrics