Multiple Regression Analyses: ***Estimation***, ***Inference***

3,4 Chapter

Ketevani Kapanadze

Brno, 2020

From previous class

- **Properties of OLS on any sample of data**
- **Fitted values and residuals**

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad \widehat{u}_i = y_i - \widehat{y}_i$$

Fitted or predicted values

Deviations from regression line (= residuals)

- **Algebraic properties of OLS regression**

$$\sum_{i=1}^{n} \widehat{u}_i = 0 \qquad \sum_{i=1}^{n} x_i \widehat{u}_i = 0 \qquad \bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}$$

Deviations from regression line sum up to zero

Correlation between deviations and regressors is zero

Sample averages of y and x lie on regression line

Properties of OLS

- **Goodness-of-Fit**

„How well does the explanatory variable explain the dependent variable?"

- **Measures of Variation**

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

Total sum of squares, represents total variation in dependent variable

Explained sum of squares, represents variation explained by regression

Residual sum of squares, represents variation not explained by regression

Properties of OLS

- **Decomposition of total variation**

$$SST = SSE + SSR$$

Total variation Explained part Unexplained part

- **Goodness-of-fit measure (R-squared)**

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

R-squared measures the fraction of the total variation that is explained by the regression

Properties of OLS: Examples

- **CEO Salary and return on equity**

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

$$n = 209, \quad R^2 = 0.0132$$

The regression explains only 1.3 % of the total variation in salaries

- **Voting outcomes and campaign expenditures**

$$\widehat{voteA} = 26.81 + 0.464\ shareA$$

$$n = 173, \quad R^2 = 0.856$$

The regression explains 85.6 % of the total variation in election outcomes

- **Caution: A high R-squared does not necessarily mean that the regression has a causal interpretation!**

Expected Values and Variance of the OLS

- **The estimated regression coefficients are random variables because they are calculated from a random sample**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Data is random and depends on particular sample that has been drawn

- **The question is what the estimators will estimate on average and how large their variability in repeated samples is**

$$E(\hat{\beta}_0) = ?,\ E(\hat{\beta}_1) = ? \qquad Var(\hat{\beta}_0) = ?,\ Var(\hat{\beta}_1) = ?$$

Expected Values and Variance of the OLS

- **Standard assumptions for the linear regression model**
- **Assumption SLR.1 (Linear in parameters)**

$$y = \beta_0 + \beta_1 x + u$$

In the population, the relationship between y and x is linear

- **Assumption SLR.2 (Random sampling)**

$$\{(x_i, y_i) : \ i = 1, \ldots n\}$$

The data is a random sample drawn from the population

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Each data point therefore follows the population equation

Expected Values and Variance of the OLS

- **<u>Assumptions for the linear regression model (cont.)</u>**
- **Assumption SLR.3 (Sample variation in explanatory variable)**

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$$

The values of the explanatory variables are not all the same (otherwise it would be impossible to study how different values of the explanatory variable lead to different values of the dependent variable)

- **Assumption SLR.4 (Zero conditional mean)**

$$E(u_i|x_i) = 0$$

The value of the explanatory variable must contain no information about the mean of the unobserved factors

Expected Values and Variance of the OLS

- <u>**Theorem 2.1 (Unbiasedness of OLS)**</u>

$$SLR.1 - SLR.4 \quad \Rightarrow \quad E(\widehat{\beta}_0) = \beta_0,\ E(\widehat{\beta}_1) = \beta_1$$

- **Interpretation of unbiasedness**
  - The estimated coefficients may be smaller or larger, depending on the sample that is the result of a random draw
  - However, on average, they will be equal to the values that characterize the true relationship between y and x in the population
  - „On average“ means if sampling was repeated, i.e. if drawing the random sample and doing the estimation was repeated many times
  - In a given sample, estimates may differ considerably from true values

Expected Values and Variance of the OLS

- **<u>Variances of the OLS estimators</u>**
  - Depending on the sample, the estimates will be nearer or farther away from the true population values
  - How far can we expect our estimates to be away from the true population values on average (= sampling variability)?
  - Sampling variability is measured by the estimator‘s variances

$$Var(\widehat{\beta}_0),\ Var(\widehat{\beta}_1)$$

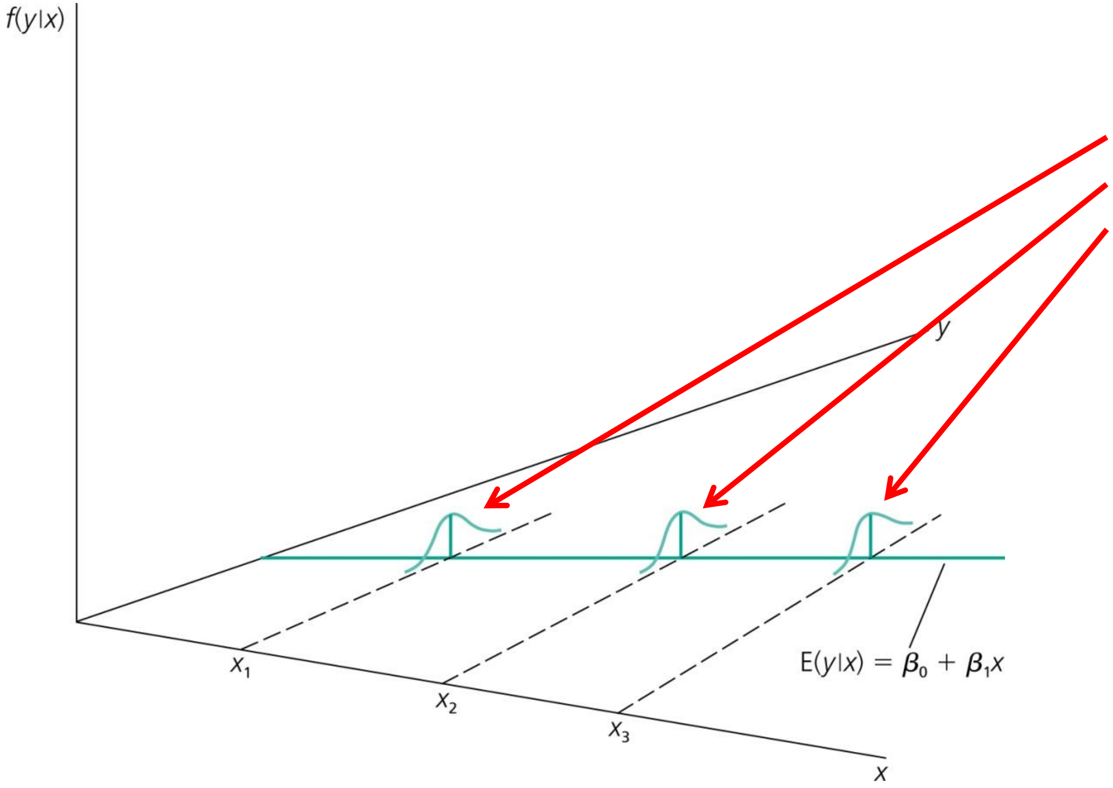
- **Assumption SLR.5 (Homoskedasticity)**

$$Var(u_i|x_i) = \sigma^2$$

The value of the explanatory variable must contain no information about the <u>variability</u> of the unobserved factors

Expected Values and Variance of the OLS

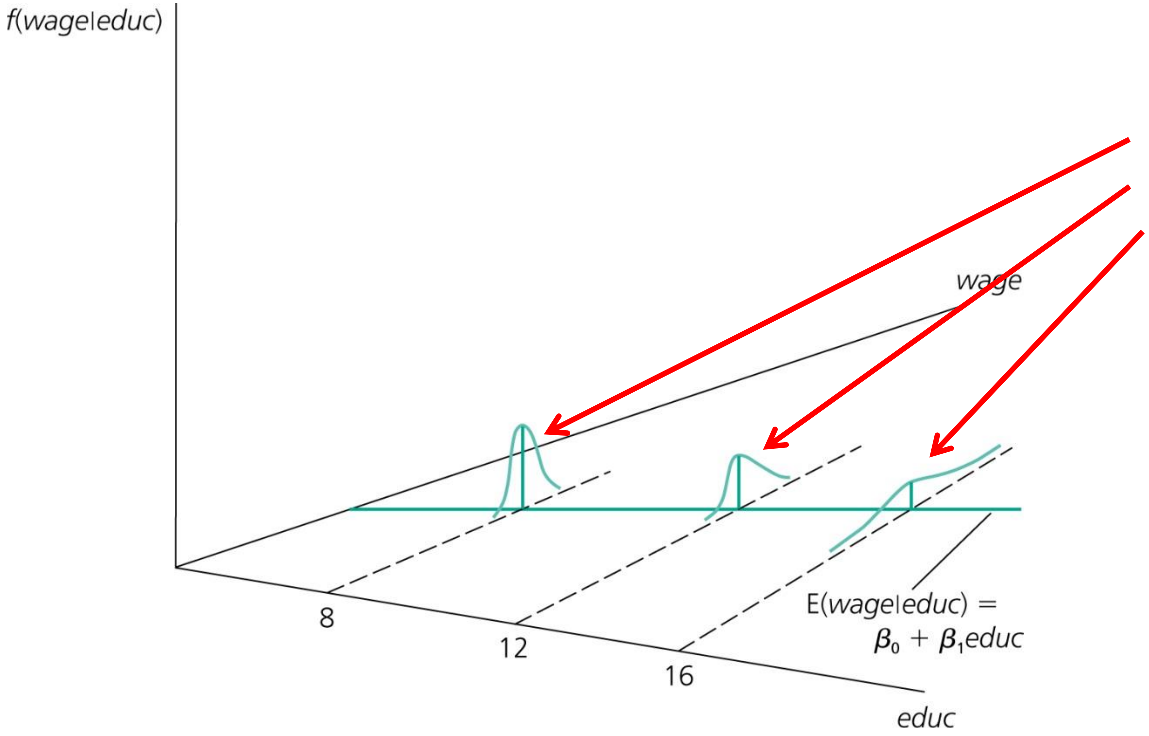
- **Graphical illustration of homoskedasticity**



The variability of the unobserved influences does not dependent on the value of the explanatory variable

Expected Values and Variance of the OLS

- **An example for heteroskedasticity: Wage and education**



The variance of the unobserved determinants of wages increases with the level of education

Expected Values and Variance of the OLS

- **<u>Theorem 2.2 (Variances of OLS estimators)</u>**

Under assumptions SLR.1 – SLR.5:

$$Var(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

$$Var(\widehat{\beta}_0) = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{SST_x}$$

- **Conclusion:**
  - The sampling variability of the estimated regression coefficients will be the higher the larger the variability of the unobserved factors, and the lower, the higher the variation in the explanatory variable

Expected Values and Variance of the OLS

- **Estimating the error variance**

$$Var(u_i|x_i) = \sigma^2 = Var(u_i)$$

The variance of u does not depend on x, i.e. is equal to the unconditional variance

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$$

One could estimate the variance of the errors by calculating the variance of the residuals in the sample; unfortunately this estimate would be biased

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

An unbiased estimate of the error variance can be obtained by substracting the number of estimated regression coefficients from the number of observations

Expected Values and Variance of the OLS

- <u>**Theorem 2.3 (Unbiasedness of the error variance)**</u>

$$SLR.1 - SLR.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2$$

- **Calculation of standard errors for regression coefficients**

$$se(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)} = \sqrt{\hat{\sigma}^2/SST_x}$$

$$se(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} x_i^2/SST_x}$$

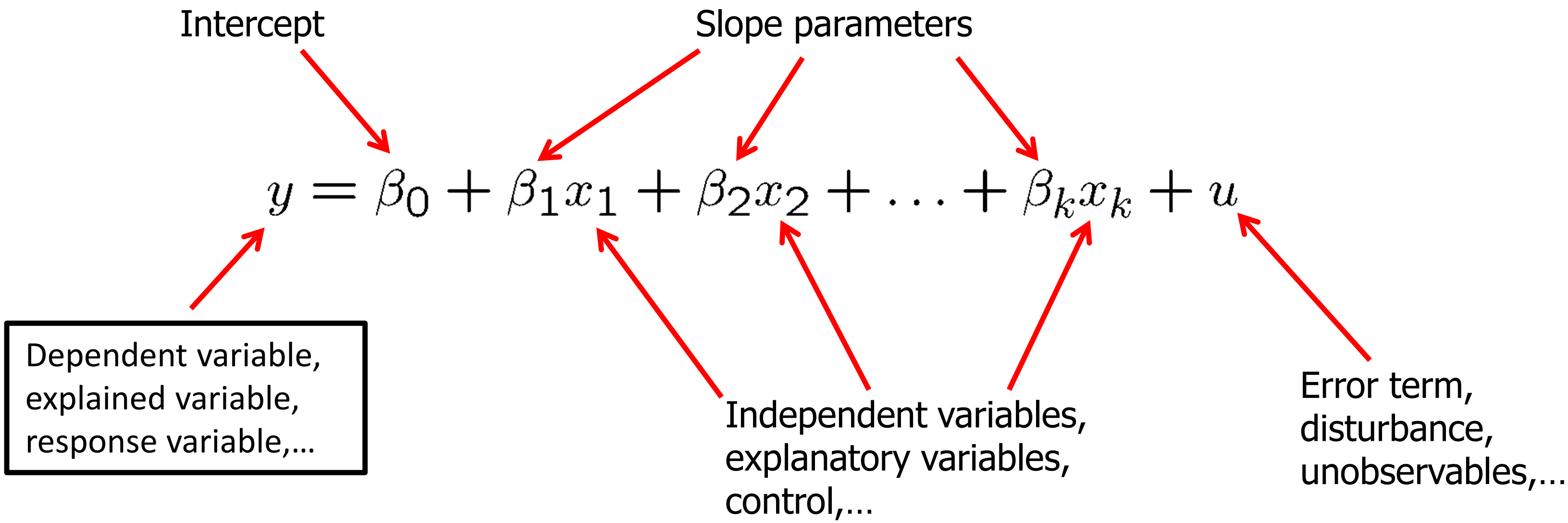
Plug in $\hat{\sigma}^2$ for the unknown σ^2

The estimated standard deviations of the regression coefficients are called „standard errors“.
<u>They measure how precisely the regression coefficients are estimated.</u>

- MULTIPLE REGRESSION MODELS

Multiple Regression Analyses: ***Estimation***

- **Definition of the multiple linear regression model**



Intercept

Slope parameters

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Dependent variable, explained variable, response variable,…

Independent variables, explanatory variables, control,…

Error term, disturbance, unobservables,…

Estimation: Motivation

- **Motivation for multiple regression**
  - Incorporate more explanatory factors into the model
  - Explicitly hold fixed other factors that otherwise would be in u
  - Allow for more flexible functional forms
- **Example: Wage equation**

Now measures effect of education explicitly holding experience fixed

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

All other factors…

Hourly wage

Years of education

Labor market experience

Estimation: Motivation

- **Example: Average test scores and per student spending**

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u$$

Other factors

Average standardized test score of school

Per student spending at this school

Average family income of students at this school

  - Per student spending is likely to be correlated with average family income at a given high school because of school financing
  - Omitting average family income in regression would lead to biased estimate of the effect of spending on average test scores
  - In a simple regression model, effect of per student spending would partly include the effect of family income on test scores

Estimation: Motivation

- **Example: Family income and family consumption**

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$$

Other factors

Family consumption

Family income

Family income squared

- Model has two explanatory variables: inome and income squared
- Consumption is explained as a quadratic function of income
- One has to be very careful when interpreting the coefficients:

By how much does consumption increase if income is increased by one unit?

$$\frac{\partial cons}{\partial inc} = \beta_1 + 2\beta_2 inc$$

Depends on how much income is already there

Estimation: Motivation

- **Example: CEO salary, sales and CEO tenure**

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u$$

Log of CEO salary

Log sales

Quadratic function of CEO tenure with firm

  - Model assumes a constant elasticity relationship between CEO salary and the sales of his or her firm
  - Model assumes a quadratic relationship between CEO salary and his or her tenure with the firm

- **Meaning of „linear" regression**
  - ***The model has to be linear in the parameters (not in the variables)***

Estimation: OLS estimation of the MLR

- **Random sample**

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

- **Regression residuals**

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}$$

- **Minimize sum of squared residuals**

$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$$

Minimization will be carried out by computer

Estimation: OLS estimation of the MLR

- **Interpretation of the multiple regression model**

$$\beta_j = \frac{\partial y}{\partial x_j}$$

By how much does the dependent variable change if the j-th independent variable is increased by one unit, **holding all other independent variables and the error term constant**

- The multiple linear regression model manages to hold the values of other explanatory variables fixed even if, in reality, they are correlated with the explanatory variable under consideration
- „Ceteris paribus“-interpretation
- It has still to be assumed that unobserved factors do not change if the explanatory variables are changed

Estimation: OLS estimation of the MLR

- **Example: Determinants of college GPA**

$$\widehat{colGPA} = 1.29 + .453 hsGPA + .0094 ACT$$

Grade point average at college

High school grade point average

Achievement test score

- **Interpretation**
  - Holding ACT fixed, another point on high school grade point average is associated with another .453 points college grade point average
  - Or: If we compare two students with the same ACT, but the hsGPA of student A is one point higher, we predict student A to have a colGPA that is .453 higher than that of student B
  - Holding high school grade point average fixed, another 10 points on ACT are associated with less than one point on college GPA

Estimation: OLS estimation of the MLR

- **Properties of OLS on any sample of data**
- **Fitted values and residuals**

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_k x_{ik} \qquad \widehat{u}_i = y_i - \widehat{y}_i$$

Fitted or predicted values

Residuals

- **Algebraic properties of OLS regression**

$$\sum_{i=1}^{n} \widehat{u}_i = 0 \qquad \sum_{i=1}^{n} x_{ij}\widehat{u}_i = 0 \qquad \bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}_1 + \ldots + \widehat{\beta}_k \bar{x}_k$$

Deviations from regression line sum up to zero

Correlations between deviations and regressors are zero

Sample averages of y and of the regressors lie on regression line

Estimation: OLS estimation of the MLR

- **Goodness-of-Fit**
- **Decomposition of total variation**

$$STT = SSE + SSR$$

- **R-squared**

$$R^2 = SSE/SST = 1 - SSR/SST$$

Notice that R-squared can only increase if another explanatory variable is added to the regression

- **Alternative expression for R-squared**

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

R-squared is equal to the squared correlation coefficient between the actual and the predicted value of the dependent variable

Estimation: The Expected Value of the OLS Estimators

- **Standard assumptions for the multiple regression model**
- **Assumption MLR.1 (Linear in parameters)**
- **Assumption MLR.2 (Random sampling)**
- **Assumption MLR.3 (No perfect collinearity)**

  „In the sample (and therefore in the population), none of the independent variables are constant and there are no exact linear relationships among the independent variables“

- **Assumption MLR.4 (Zero conditional mean)**

Estimation: The Expected Value of the OLS Estimators

- **Example for perfect collinearity: small sample**

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u$$

In a small sample, avginc may accidentally be an exact multiple of expend; it will not be possible to disentangle their separate effects because there is exact covariation

- **Example for perfect collinearity: relationships between regressors**

$$voteA = \beta_0 + \beta_1 shareA + \beta_2 shareB + u$$

Either shareA or shareB will have to be dropped from the regression because there is an exact linear relationship between them: shareA + shareB = 1

Estimation: The Expected Value of the OLS Estimators

- **Assumption MLR.4 (Zero conditional mean)**

$$E(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$$

The value of the explanatory variables must contain no information about the mean of the unobserved factors

  - In a multiple regression model, the zero conditional mean assumption is much more likely to hold because fewer things end up in the error

- **Example: Average test scores**

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u$$

If avginc was not included in the regression, it would end up in the error term; it would then be hard to defend that expend is uncorrelated with the error

Estimation: The Expected Value of the OLS Estimators

- **Discussion of the zero mean conditional assumption**
  - Explanatory variables that are correlated with the error term are called endogenous; endogeneity is a violation of assumption MLR.4
  - Explanatory variables that are uncorrelated with the error term are called exogenous; MLR.4 holds if all explanat. var. are exogenous
  - Exogeneity is the key assumption for a causal interpretation of the regression, and for unbiasedness of the OLS estimators
- **Theorem 3.1 (Unbiasedness of OLS)**

$$MLR.1 - MLR.4 \quad \Rightarrow \quad E(\widehat{\beta}_j) = \beta_j, \quad j = 0, 1, \ldots, k$$

➢ Unbiasedness is an average property in repeated samples; in a given sample, the estimates may still be far away from the true values

Estimation: The Expected Value of the OLS Estimators

- **Including irrelevant variables in a regression model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3} x_3 + u$$

= 0 in the population

No problem because $E(\widehat{\beta}_3) = \beta_3 = 0.$

However, including irrevelant variables may increase sampling variance.

- **Omitting relevant variables: the simple case**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

True model (contains x_1 and x_2)

$$y = \alpha_0 + \alpha_1 x_1 + w$$

Estimated model (x_2 is omitted)

Estimation: The Expected Value of the OLS Estimators

- **Omitted variable bias**

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

If x_1 and x_2 are correlated, assume a linear regression relationship between them

$$\Rightarrow \quad y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u$$

$$= \underbrace{\boxed{(\beta_0 + \beta_2\delta_0)}}_{\text{If y is only regressed on } x_1 \text{ this will be the estimated intercept}} + \underbrace{\boxed{(\beta_1 + \beta_2\delta_1)}}_{\text{If y is only regressed on } x_1\text{, this will be the estimated slope on } x_1} \enclose{circle}{x_1} + \underbrace{\boxed{(\beta_2 v + u)}}_{\text{error term}}$$

- **<u>Conclusion:</u> All estimated coefficients will be biased**

Estimation: The Expected Value of the OLS Estimators

- **Example: Omitting ability in a wage equation**

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u$$

Both will be positive

$$abil = \delta_0 + \delta_1 educ + v$$

$$wage = (\beta_0 + \beta_2\delta_0) + \boxed{(\beta_1 + \beta_2\delta_1)} educ + (\beta_2 v + u)$$

The return to education β_1 will be overestimated because $\beta_2\delta_1 > 0$. It will look as if people with many years of education earn very high wages, but this is partly due to the fact that people with more education are also more able on average.

- **When is there no omitted variable bias?**
  - If the omitted variable is irrelevant or uncorrelated

Estimation: The Expected Value of the OLS Estimators

- **Omitted variable bias: more general cases**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

← True model (contains x_1, x_2 and x_3)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + w$$

← Estimated model (x_3 is omitted)

  - No general statements possible about direction of bias
  - Analysis as in simple case if one regressor uncorrelated with others

- **Example: Omitting ability in a wage equation**

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

If exper is approximately uncorrelated with educ and abil, then the direction of the omitted variable bias can be as analyzed in the simple two variable case.

Estimation: The Variance of the OLS Estimators

- **Standard assumptions for the multiple regression model (cont.)**
- **Assumption MLR.5 (Homoscedasticity)**
- **Short hand notation**

All explanatory variables are collected in a random vector

$$Var(u_i|\mathbf{x}_i) = \sigma^2 \quad \text{with} \quad \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$$

Under assumptions MLR.1 – MLR.5:

$$Var(\widehat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \qquad j = 1, \ldots, k$$

Estimation: The Variance of the OLS Estimators

- **Components of OLS Variances:**
- **1) The error variance**
  - A high error variance increases the sampling variance because there is more „noise“ in the equation
  - A large error variance necessarily makes estimates imprecise
  - The error variance does not decrease with sample size
- **2) The total sample variation in the explanatory variable**
  - More sample variation leads to more precise estimates
  - Total sample variation automatically increases with the sample size
  - Increasing the sample size is thus a way to get more precise estimates

Estimation: The Variance of the OLS Estimators

- **3) Linear relationships among the independent variables**

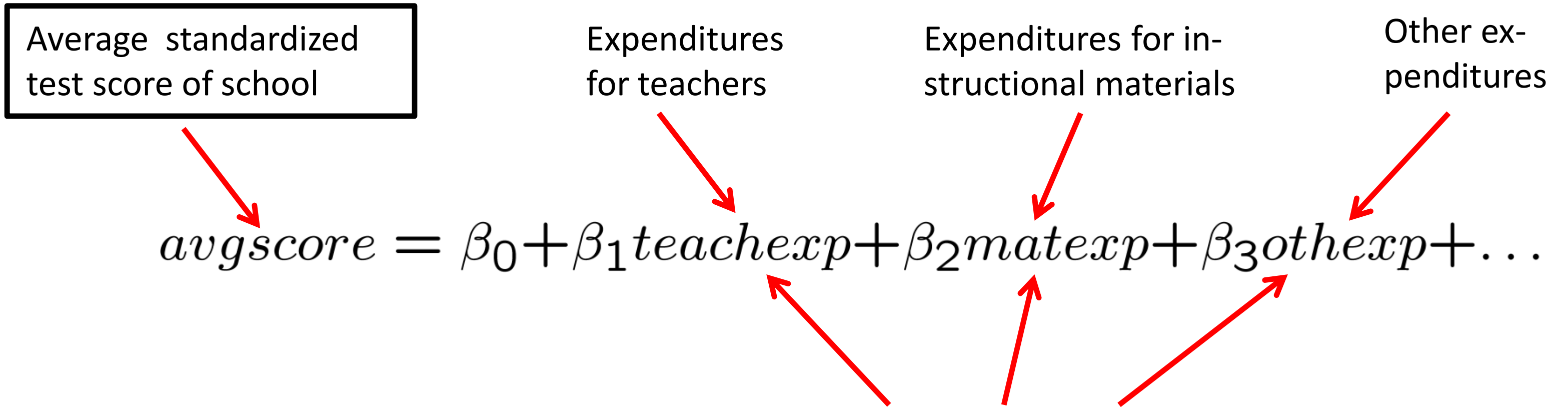
Regress x_j on all other independent variables (including a constant)

The R-squared of this regression will be higher the better x_j can be linearly explained by the other independent variables

- Sampling variance of $\widehat{\beta}_j$ will be the higher the better explanatory variable x_j can be linearly explained by other independent variables
- The problem of almost linearly dependent explanatory variables is called multicollinearity (i.e. $R_j \rightarrow 1$ for some j)

Estimation: The Variance of the OLS Estimators

- **An example for multicollinearity**



Average standardized test score of school

Expendítures for teachers

Expenditures for instructional materials

Other expenditures

$$avgscore = \beta_0 + \beta_1 teachexp + \beta_2 matexp + \beta_3 othexp + \ldots$$

The different expenditure categories will be strongly correlated because if a school has a lot of resources it will spend a lot on everything.

It will be hard to estimate the differential effects of different expenditure categories because all expenditures are either high or low. For precise estimates of the differential effects, one would need information about situations where expenditure categories change differentially.

As a consequence, sampling variance of the estimated effects will be large.

Estimation: The Variance of the OLS Estimators

- **Discussion of the multicollinearity problem**
  - In the above example, it would probably be better to lump all expenditure categories together because effects cannot be disentangled
  - In other cases, dropping some independent variables may reduce multicollinearity (but this may lead to omitted variable bias)
  - Only the sampling variance of the variables involved in multicollinearity will be inflated; the estimates of other effects may be very precise
  - Note that multicollinearity is not a violation of MLR.3 in the strict sense
  - Multicollinearity may be detected through „variance inflation factors“

$$VIF_j = 1/(1 - R_j^2)$$

As an (arbitrary) rule of thumb, the variance inflation factor should not be larger than 10

Estimation: The Variance of the OLS Estimators

- **Estimating the error variance**

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} \hat{u}_i^2 \right) / \, [n - k - 1]$$

An unbiased estimate of the error variance can be obtained by substracting the number of estimated regression coefficients from the number of observations. The number of observations minus the number of estimated parameters is also called the degrees of freedom. The n estimated squared residuals in the sum are not completely independent but related through the k+1 equations that define the first order conditions of the minimization problem.

- **Theorem 3.3 (Unbiased estimator of the error variance)**

$$MLR.1 - MLR.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2$$

Estimation: The Variance of the OLS Estimators

- **Estimation of the sampling variances of the OLS estimators**

The true sampling variation of the estimated β_j

$$sd(\widehat{\beta}_j) = \sqrt{Var(\widehat{\beta}_j)} = \sqrt{\sigma^2 / \left[SST_j(1 - R_j^2)\right]}$$

Plug in $\widehat{\sigma}^2$ for the unknown σ^2

The estimated sampling variation of the estimated β_j

$$se(\widehat{\beta}_j) = \sqrt{\widehat{Var}(\widehat{\beta}_j)} = \sqrt{\widehat{\sigma}^2 / \left[SST_j(1 - R_j^2)\right]}$$

- **Note that these formulas are only valid under assumptions MLR.1-MLR.5 (in particular, there has to be homoscedasticity)**

Estimation: Efficiency of OLS

- **Efficiency of OLS: The Gauss-Markov Theorem**
  - Under assumptions MLR.1 - MLR.5, OLS is unbiased
  - However, under these assumptions there may be many other estimators that are unbiased
  - Which one is the unbiased estimator with the smallest variance?
  - In order to answer this question one usually limits oneself to linear estimators, i.e. estimators linear in the dependent variable

$$\tilde{\beta}_j = \sum_{i=1}^{n} \boxed{w_{ij}} y_i$$

Maybe an arbitrary function of the sample values of all the explanatory variables; the OLS estimator can be shown to be of this form

Estimation: Efficiency of OLS

- **Theorem 3.4 (Gauss-Markov Theorem)**
  - Under assumptions MLR.1 - MLR.5, the OLS estimators are the best linear unbiased estimators (BLUEs) of the regression coefficients, i.e.

$$Var(\widehat{\beta}_j) \leq Var(\tilde{\beta}_j) \quad j = 0, 1, \ldots, k$$

for all $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$ for which $E(\tilde{\beta}_j) = \beta_j, j = 0, \ldots, k$.

- **OLS is only the best estimator if MLR.1 – MLR.5 hold; if there is heteroscedasticity for example, there are better estimators.**

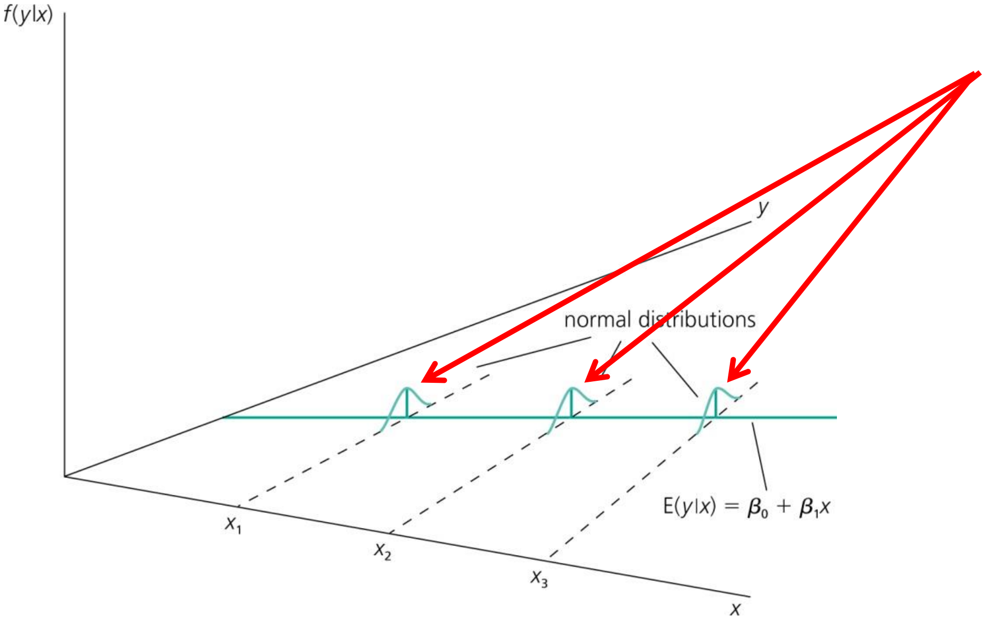
Multiple Regression Analyses: ***Inference***

- **Statistical inference in the regression model**
  - Hypothesis tests about population parameters
  - Construction of confidence intervals
- **Sampling distributions of the OLS estimators**
  - The OLS estimators are random variables
  - We already know their expected values and their variances
  - However, for hypothesis tests we need to know their distribution
  - In order to derive their distribution we need additional assumptions
  - Assumption about distribution of errors: normal distribution

Inference: Sampling distributions of the OLS Estimators

- **Assumption MLR.6 (Normality of error terms)**

$$u_i \sim N(0, \sigma^2) \quad \text{independently of} \quad x_{i1}, x_{i2}, \ldots, x_{ik}$$



It is assumed that the unobserved factors are normally distributed around the population regression function.

The form and the variance of the distribution does not depend on any of the explanatory variables.

It follows that:

$$y|\mathbf{x} \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$$

Inference: Sampling distributions of the OLS Estimators

- **Discussion of the normality assumption**
  - The error term is the sum of „many“ different unobserved factors
  - Sums of independent factors are normally distributed (CLT)
  - Problems:
    - How many different factors? Number large enough?
    - Possibly very heterogenuous distributions of individual factors
    - How independent are the different factors?
  - The normality of the error term is an empirical question
  - At least the error distribution should be „close“ to normal
  - In many cases, normality is questionable or impossible by definition

Inference: Sampling distributions of the OLS Estimators

- **Discussion of the normality assumption (cont.)**
  - Examples where normality cannot hold:
    - Wages (nonnegative; also: minimum wage)
    - Number of arrests (takes on a small number of integer values)
    - Unemployment (indicator variable, takes on only 1 or 0)
  - In some cases, normality can be achieved through transformations of the dependent variable (e.g. use log(wage) instead of wage)
  - Under normality, OLS is the best (even nonlinear) unbiased estimator
  - Important: For the purposes of statistical inference, the assumption of normality can be replaced by a large sample size

Inference: Sampling distributions of the OLS Estimators

- **Terminology**

$$\underbrace{MLR.1 - MLR.5}_{\text{„Gauss-Markov assumptions“}} \qquad \underbrace{MLR.1 - MLR.6}_{\text{„Classical linear model (CLM) assumptions“}}$$

- **Theorem 4.1 (Normal sampling distributions)**

Under assumptions MLR.1 – MLR.6:

$$\widehat{\beta}_j \sim N(\beta_j, Var(\widehat{\beta}_j)) \qquad \frac{\widehat{\beta}_j - \beta_j}{sd(\widehat{\beta}_j)} \sim N(0,1)$$

The estimators are normally distributed around the true parameters with the variance that was derived earlier

The standardized estimators follow a standard normal distribution

Inference: The *t* Test

- **Testing hypotheses about a single population parameter**
- **Theorem 4.2 (t-distribution for standardized estimators)**

Under assumptions MLR.1 – MLR.6:

$$\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \sim t_{n-k-1}$$

If the standardization is done using the estimated standard deviation (= standard error), the normal distribution is replaced by a t-distribution

Note: The t-distribution is close to the standard normal distribution if n-k-1 is large.

- **Null hypothesis (for more general hypotheses, see below)**

$$H_0 : \ \beta_j = 0$$

The population parameter is equal to zero, i.e. after controlling for the other independent variables, there is no effect of x_j on y

Inference: The *t* Test

- **t-statistic (or t-ratio)**

$$t_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}$$

The t-statistic will be used to test the above null hypothesis. The farther the estimated coefficient is away from zero, the less likely it is that the null hypothesis holds true. But what does „far“ away from zero mean?

This depends on the variability of the estimated coefficient, i.e. its standard deviation. The t-statistic measures how many estimated standard deviations the estimated coefficient is away from zero.

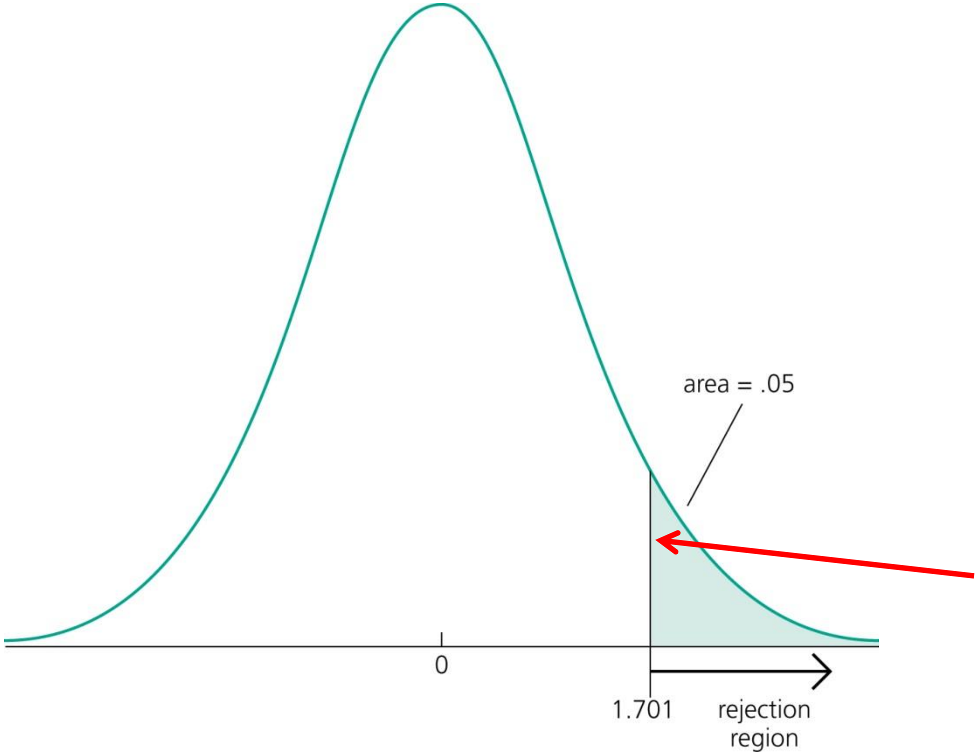
- **Distribution of the t-statistic if the null hypothesis is true**

$$t_{\widehat{\beta}_j} = \widehat{\beta}_j / se(\widehat{\beta}_j) = (\widehat{\beta}_j - \beta_j)/se(\widehat{\beta}_j) \sim t_{n-k-1}$$

- **Goal: Define a rejection rule so that, if it is true, H_0 is rejected only with a small probability (= significance level, e.g. 5%)**

Inference: The *t* Test

- **Testing against one-sided alternatives (greater than zero)**



Test $H_0 : \beta_j = 0$ against $H_1 : \beta_j > 0$

Reject the null hypothesis in favour of the alternative hypothesis if the estimated coefficient is „too large“ (i.e. larger than a critical value).

Construct the critical value so that, if the null hypothesis is true, it is rejected in, for example, 5% of the cases.

In the given example, this is the point of the t-distribution with 28 degrees of freedom that is exceeded in 5% of the cases.

! Reject if t-statistic greater than 1.701

Inference: The *t* Test

- **Example: Wage equation**
  - Test whether, after controlling for education and tenure, higher work experience leads to higher hourly wages

$$\widehat{\log}(wage) = \underset{(.104)}{.284} + \underset{(.007)}{.092}\ educ + \underset{(.0017)}{.0041}\ exper + \underset{(.003)}{.022}\ tenure$$

$$n = 526,\ R^2 = .316$$

Standard errors

Test $H_0 : \beta_{exper} = 0$ against $H_1 : \beta_{exper} > 0$.

One would either expect a positive effect of experience on hourly wage or no effect at all.

Inference: The *t* Test

- **Example: Wage equation (cont.)**

$$\boxed{t_{exper} = .0041/.0017 \approx 2.41}$$

t-statistic

$$df = n - k - 1 = 526 - 3 - 1 = 522$$

Degrees of freedom;
here the standard normal
approximation applies

$$c_{0.05} = 1.645$$

$$c_{0.01} = 2.326$$

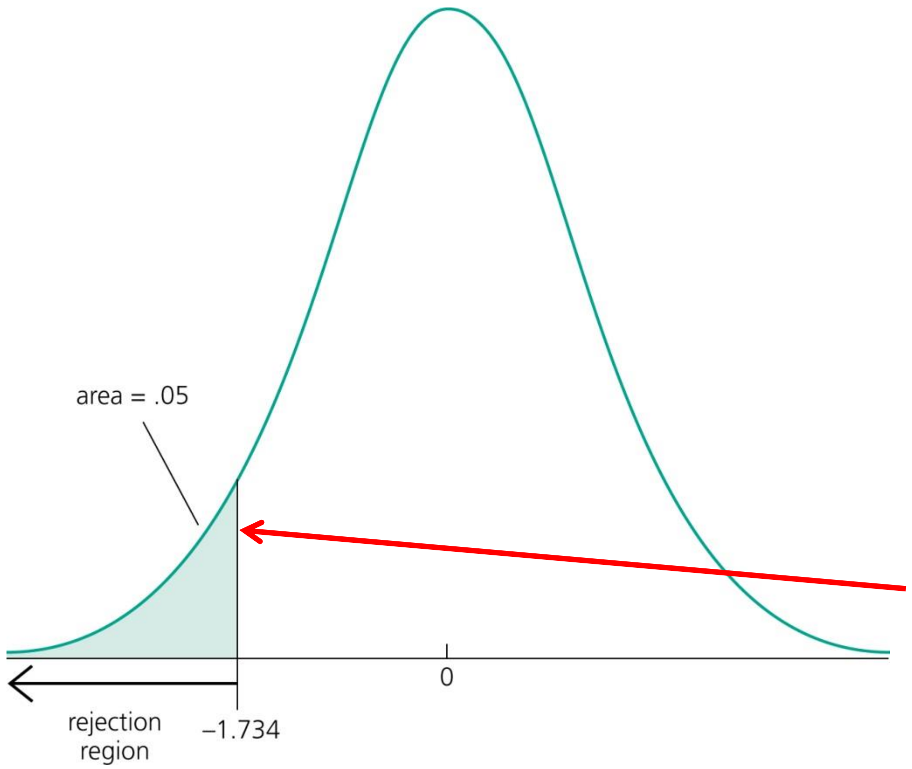
Critical values for the 5% and the 1% significance level (these are conventional significance levels).

The null hypothesis is rejected because the t-statistic exceeds the critical value.

„The effect of experience on hourly wage is statistically greater than zero at the 5% (and even at the 1%) significance level.“

Inference: The *t* Test

- **Testing against one-sided alternatives (less than zero)**



Test $H_0 : \beta_j = 0$ against $H_1 : \beta_j < 0$

Reject the null hypothesis in favour of the alternative hypothesis if the estimated coefficient is „too small“ (i.e. smaller than a critical value).

Construct the critical value so that, if the null hypothesis is true, it is rejected in, for example, 5% of the cases.

In the given example, this is the point of the t-distribution with 18 degrees of freedom so that 5% of the cases are below the point.

! Reject if t-statistic less than -1.734

Inference: The *t* Test

- **Example: Student performance and school size**
    - Test whether smaller school size leads to better student performance

Percentage of students passing maths test

Average annual teacher compensation

Staff per one thousand students

School enrollment (= school size)

$$\widehat{math10} = + \underset{(6.113)}{2.274} + \underset{(.00010)}{.00046}\ totcomp + \underset{(.040)}{.048}\ staff - \underset{(.00022)}{\boxed{.00020}}\ enroll$$

$$n = 408, \ R^2 = .0541$$

Test $H_0 : \beta_{enroll} = 0$ against $H_1 : \beta_{enroll} < 0$.

Do larger schools hamper student performance or is there no such effect?

Inference: The *t* Test

- **Example: Student performance and school size (cont.)**

$$\boxed{t_{enroll} = -.00020/.00022 \approx -.91}$$

t-statistic

$$df = n - k - 1 = 408 - 3 - 1 = 404$$

Degrees of freedom;
here the standard normal
approximation applies

$$c_{0.05} = -1.65$$

$$c_{0.15} = -1.04$$

Critical values for the 5% and the 15% significance level.

The null hypothesis is not rejected because the t-statistic is not smaller than the critical value.

One cannot reject the hypothesis that there is no effect of school size on student performance (not even for a lax significance level of 15%).

<u>Inference</u>: The *t* Test

- **Example: Student performance and school size (cont.)**
  - Alternative specification of functional form:

$$\widehat{math10} = -\underset{(48.70)}{207.66} + \underset{(4.06)}{21.16}\,\log(totcomp)$$

$$+ \underset{(4.19)}{3.98}\,\log(staff) - \underset{(0.69)}{1.29}\,\log(enroll)$$

$$n = 408,\ \ R^2 = .0654 \longleftarrow$$

R-squared slightly higher

Test $H_0 : \beta_{\log(enroll)} = 0$ against $H_1 : \beta_{\log(enroll)} < 0$.

Inference: The *t* Test

- **Example: Student performance and school size (cont.)**

$$\boxed{t_{\log(enroll)} = -1.29/.69 \approx -1.87}$$

t-statistic

$$c_{0.05} = -1.65$$

Critical value for the 5% significance level ! reject null hypothesis

The hypothesis that there is no effect of school size on student performance can be rejected in favor of the hypothesis that the effect is negative.

How large is the effect?

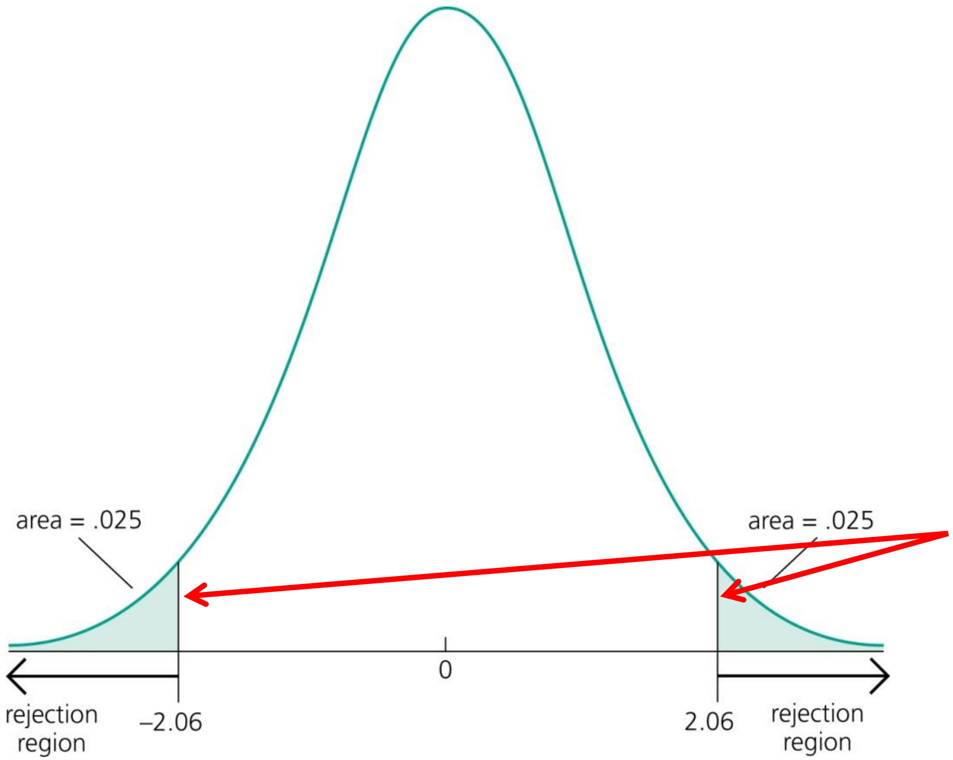
+ 10% enrollment ! -0.129 percentage points students pass test

$$-1.29 = \frac{\partial math10}{\partial \log(enroll)} = \frac{math10}{\frac{\partial enroll}{enroll}} = \frac{\frac{-1.29}{100}}{\frac{1}{100}} = \frac{-0.0129}{+1\%}$$

(small effect)

Inference: The *t* Test

- **Testing against two-sided alternatives**



Test $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$

Reject the null hypothesis in favour of the alternative hypothesis if the absolute value of the estimated coefficient is too large.

Construct the critical value so that, if the null hypothesis is true, it is rejected in, for example, 5% of the cases.

In the given example, these are the points of the t-distribution so that 5% of the cases lie in the two tails.

! Reject if absolute value of t-statistic is less than -2.06 or greater than 2.06

Inference: The *t* Test

- **Example: Determinants of college GPA**

Lectures missed per week

$$\widehat{collGPA} = \underset{(.33)}{1.39} + \underset{(.094)}{.412}\ hsGPA + \underset{(.011)}{.015}\ ACT - \underset{(.026)}{.083}\ skipped$$

$$n = 141,\ R^2 = .234$$

For critical values, use standard normal distribution

$$t_{hsGPA} = 4.38 > c_{0.01} = 2.58$$

$$t_{ACT} = 1.36 < c_{0.10} = 1.645$$

$$|t_{skipped}| = |-3.19| > c_{0.01} = 2.58$$

The effects of hsGPA and skipped are significantly different from zero at the 1% significance level. The effect of ACT is not significantly different from zero, not even at the 10% significance level.

Inference: The *t* Test

- **„Statistically significant“ variables in a regression**
  - If a regression coefficient is different from zero in a two-sided test, the corresponding variable is said to be „statistically significant“
  - If the number of degrees of freedom is large enough so that the normal approximation applies, the following rules of thumb apply:

$|t - ratio| > 1.645$ → „statistically significant at 10 % level“

$|t - ratio| > 1.96$ → „statistically significant at 5 % level“

$|t - ratio| > 2.576$ → „statistically significant at 1 % level“

Inference: The *t* Test

- **Guidelines for discussing economic and statistical significance**
  - If a variable is statistically significant, discuss the magnitude of the coefficient to get an idea of its economic or practical importance
  - The fact that a coefficient is statistically significant does not necessarily mean it is economically or practically significant!
  - If a variable is statistically and economically important but has the „wrong“ sign, the regression model might be misspecified
  - If a variable is statistically insignificant at the usual levels (10%, 5%, 1%), one may think of dropping it from the regression
  - If the sample size is small, effects might be imprecisely estimated so that the case for dropping insignificant variables is less strong

Inference: The *t* Test

- **Testing more general hypotheses about a regression coefficient**
- **Null hypothesis**

$$H_0 : \ \beta_j = a_j$$

Hypothesized value of the coefficient

- **t-statistic**

$$t = \frac{(estimate - hypothesized\ value)}{standard\ error} = \frac{(\hat{\beta}_j - a_j)}{se(\hat{\beta}_j)}$$

- **The test works exactly as before, except that the hypothesized value is substracted from the estimate when forming the statistic**

Inference: The *t* Test

- **Example: Campus crime and enrollment**
  - An interesting hypothesis is whether crime increases by one percent if enrollment is increased by one percent

$$\widehat{\log}(crime) = -\underset{(1.03)}{6.63} + \underset{(0.11)}{1.27}\log(enroll)$$

Estimate is different from one but is this difference statistically significant?

$$n = 97,\ R^2 = .585$$

$$H_0 : \beta_{\log(enroll)} = 1,\ H_1 : \beta_{\log(enroll)} \neq 1$$

$$t = (1.27 - 1)/.11 \approx 2.45 > 1.96 = c_{0.05}$$

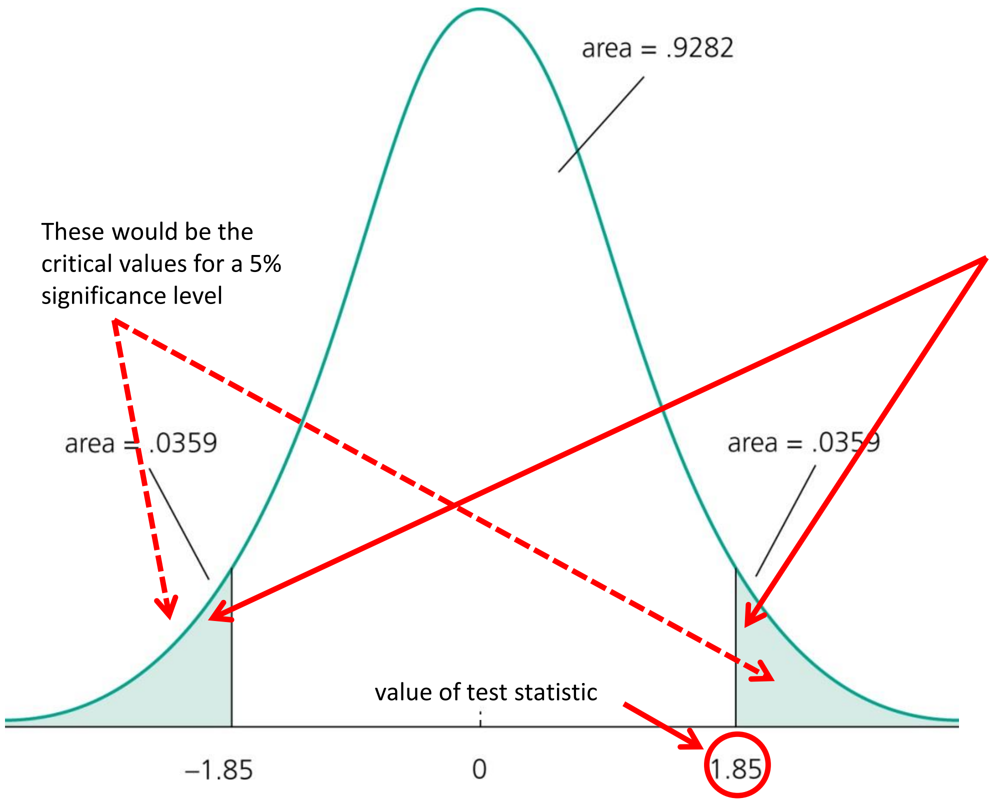
The hypothesis is rejected at the 5% level

Inference: The *t* Test

- **Computing p-values for t-tests**
  - If the significance level is made smaller and smaller, there will be a point where the null hypothesis cannot be rejected anymore
  - The reason is that, by lowering the significance level, one wants to avoid more and more to make the error of rejecting a correct H_0
  - The smallest significance level at which the null hypothesis is still rejected, is called the p-value of the hypothesis test
  - A small p-value is evidence against the null hypothesis because one would reject the null hypothesis even at small significance levels
  - A large p-value is evidence in favor of the null hypothesis
  - P-values are more informative than tests at fixed significance levels

Inference: The *t* Test

- **How the p-value is computed (here: two-sided test)**



The p-value is the significance level at which one is indifferent between rejecting and not rejecting the null hypothesis.

In the two-sided case, the p-value is thus the probability that the t-distributed variable takes on a larger absolute value than the realized value of the test statistic, e.g.:

$$P(|t - ratio| > 1.85) = 2(.0.359) = .0718$$

From this, it is clear that a null hypothesis is rejected if and only if the corresponding p-value is smaller than the significance level.

For example, for a significance level of 5% the t-statistic would not lie in the rejection region.

Inference: Confidence Intervals

- **Confidence intervals**
- **Simple manipulation of the result in Theorem 4.2 implies that**

$$P\left(\underbrace{\hat{\beta}_j - c_{0.05} \cdot se(\hat{\beta}_j)}_{\text{Lower bound of the Confidence interval}} \le \beta_j \le \underbrace{\hat{\beta}_j + \overset{\text{Critical value of two-sided test}}{c_{0.05}} \cdot se(\hat{\beta}_j)}_{\text{Upper bound of the Confidence interval}}\right) = \underset{\text{Confidence level}}{0.95}$$

- **Interpretation of the confidence interval**
  - The bounds of the interval are random
  - In repeated samples, the interval that is constructed in the above way will cover the population regression coefficient in 95% of the cases

Inference: Confidence Intervals

- **Confidence intervals for typical confidence levels**

$$P\left(\widehat{\beta}_j - \boxed{c_{0.01}} \cdot se(\widehat{\beta}_j) \leq \beta_j \leq \widehat{\beta}_j + c_{0.01} \cdot se(\widehat{\beta}_j)\right) = 0.99$$

$$P\left(\widehat{\beta}_j - \boxed{c_{0.05}} \cdot se(\widehat{\beta}_j) \leq \beta_j \leq \widehat{\beta}_j + c_{0.05} \cdot se(\widehat{\beta}_j)\right) = 0.95$$

$$P\left(\widehat{\beta}_j - \boxed{c_{0.10}} \cdot se(\widehat{\beta}_j) \leq \beta_j \leq \widehat{\beta}_j + c_{0.10} \cdot se(\widehat{\beta}_j)\right) = 0.90$$

Use rules of thumb $c_{0.01} = 2.576, c_{0.05} = 1.96, c_{0.10} = 1.645$

- **Relationship between confidence intervals and hypotheses tests**

$a_j \notin interval \quad \Rightarrow \quad$ reject $H_0 : \beta_j = a_j$ in favor of $H_1 : \beta_j \neq 0$

Inference: Confidence Intervals

- **Example: Model of firms‘ R&D expenditures**

Spending on R&D

Annual sales

Profits as percentage of sales

$$\widehat{\log(rd)} = \underset{(.47)}{-\ 4.38} + \underset{(.060)}{1.084} \log(sales) + \underset{(0.0128\ )}{0.0217}\ profmarg$$

$$n = 32,\ R^2 = .918,\ df = 32-2-1 = 29 \Rightarrow c_{0.05} = 2.045$$

$$1.084 \pm 2.045(.060) = \boxed{(.961, 1.21)}$$

$$.0217 \pm 2.045\ (0.0128\ ) = \boxed{(-.0045, .0479)}$$

The effect of sales on R&D is relatively precisely estimated as the interval is narrow. Moreover, the effect is significantly different from zero because zero is outside the interval.

This effect is imprecisely estimated as the interval is very wide. It is not even statistically significant because zero lies in the interval.

Inference: Testing hypotheses about a linear combination of parameters

- **Example: Return to education at 2 year vs. at 4 year colleges**

Years of education at 2 year colleges

Years of education at 4 year colleges

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

Test $H_0 : \beta_1 - \beta_2 = 0$ against $H_1 : \beta_1 - \beta_2 < 0$.

A possible test statistic would be:

$$t = \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{se(\widehat{\beta}_1 - \widehat{\beta}_2)}$$

The difference between the estimates is normalized by the estimated standard deviation of the difference. The null hypothesis would have to be rejected if the statistic is „too negative“ to believe that the true difference between the parameters is equal to zero.

Inference: Testing hypotheses about a linear combination of parameters

- **Impossible to compute with standard regression output because**

$$se(\hat{\beta}_1-\hat{\beta}_2)=\sqrt{\widehat{Var}(\hat{\beta}_1-\hat{\beta}_2)}=\sqrt{\widehat{Var}(\hat{\beta}_1)+\widehat{Var}(\hat{\beta}_2)-2\widehat{Cov}(\hat{\beta}_1,\hat{\beta}_2)}$$

Usually not available in regression output

- **Alternative method**

Define $\theta_1=\beta_1-\beta_2$ and test $H_0:\theta_1=0$ against $H_1:\theta_1<0$.

$$\log(wage)=\beta_0+(\theta_1+\beta_2)jc+\beta_2 univ+\beta_3 exper+u$$

$$=\beta_0+\theta_1 jc+\beta_2\boxed{(jc+univ)}+\beta_3 exper+u$$

Insert into original regression

a new regressor (= total years of college)

Inference: Testing hypotheses about a linear combination of parameters

Total years of college

- **Estimation results**

$$\begin{array}{llllll} \widehat{\log}(wage) = & 1.472 & -\ .0102\ jc & +\ .0769\ totcoll & +\ .0049\ exper \\ & (.021) & (.0069) & (.0023) & (.0002) \end{array}$$

$$n = 6,763, \;\; R^2 = .222$$

$$t = -.0102/.0069 = -1.48$$

Hypothesis is rejected at 10% level but not at 5% level

$$p - value = P(t - ratio < -1.48) = .070$$

$$-.0102 \pm 1.96(.0069) = (-.0237, .0003)$$

- **This method works <u>always</u> for single linear hypotheses**

Inference: The *F* Test

- **Testing multiple linear restrictions: The F-test**
- **Testing exclusion restrictions**

Salary of major league baseball player

Years in the league

Average number of games per year

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr$$

$$+\beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

Batting average

Home runs per year

Runs batted in per year

$$\boxed{H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0} \text{ against } H_1 : H_0 \text{ is not true}$$

Test whether performance measures have no effect/can be exluded from regression.

Inference: The *F* Test

- **Estimation of the unrestricted model**

$$\widehat{\log}(salary) = \underset{(0.29)}{11.19} + \underset{(.0121)}{.0689}\ years + \underset{(.0026)}{.0126}\ gamesyr$$

$$+ \underset{(.00110)}{.00098}\ \boxed{bavg} + \underset{(.0161)}{.0144}\ \boxed{hrunsyr} + \underset{(.0072)}{.0108}\ \boxed{rbisyr}$$

None of these variabels are statistically significant when tested individually

$$n = 353,\ SSR = \boxed{183.186},\ R^2 = .6278$$

Idea: How would the model fit be if these variables were dropped from the regression?

Inference: The *F* Test

- **Estimation of the restricted model**

$$\begin{array}{rcccc}\widehat{\log}(salary) = & 11.22 & + .0713\ years & + & .0202\ gamesyr \\ & (0.11) & (.0125) & & (.0013)\end{array}$$

$$n = 353,\ SSR = 198.311,\ R^2 = .5971$$

The sum of squared residuals necessarily increases, but is the increase statistically significant?

- **Test statistic**

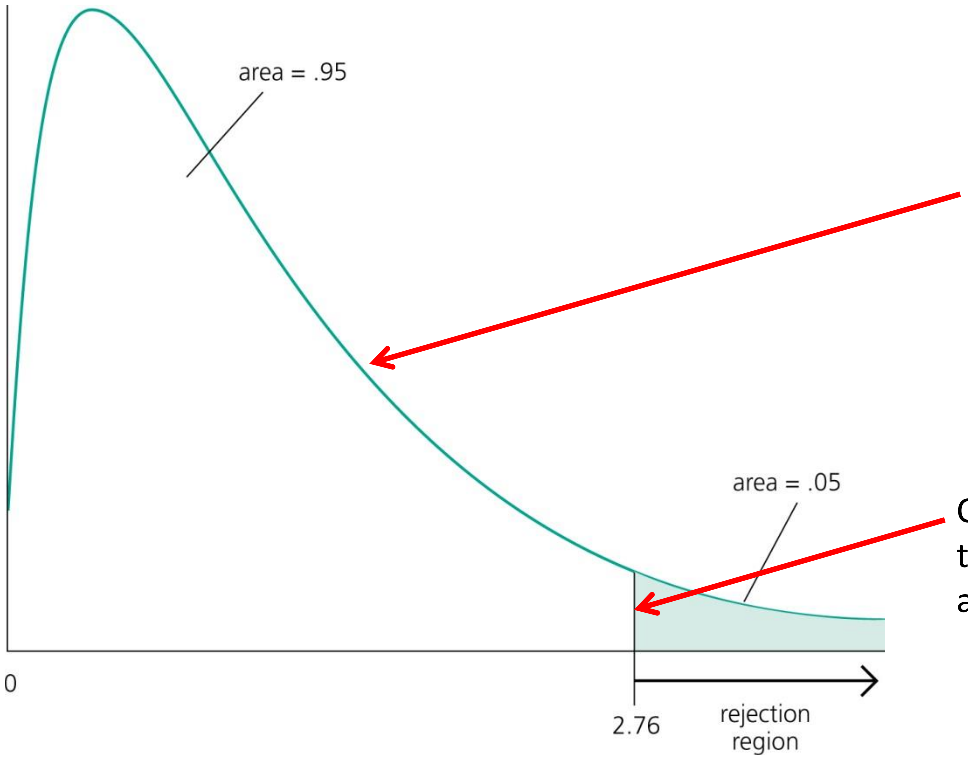
Number of restrictions

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$$

The relative increase of the sum of squared residuals when going from H_1 to H_0 follows a F-distribution (if the null hypothesis H_0 is correct)

Inference: The *F* Test

- **Rejection rule (Figure 4.7)**



A F-distributed variable only takes on positive values. This corresponds to the fact that the sum of squared residuals can only increase if one moves from H_1 to H_0.

Choose the critical value so that the null hypothesis is rejected in, for example, 5% of the cases, although it is true.

Inference: The *F* Test

- **Test decision in example**

Number of restrictions to be tested

$$F = \frac{(198.311 - 183.186)/3}{183.186/\boxed{(353-5-1)}} \approx 9.55$$

Degrees of freedom in
the unrestricted model

$$F \sim F_{3,347} \;\Rightarrow\; c_{0.01} = 3.78$$

$$P(F - statistic > 9.55) = 0.000$$

The null hypothesis is overwhelmingly rejected (even at very small significance levels).

- **Discussion**
  - The three variables are „jointly significant“
  - They were not significant when tested individually
  - The likely reason is multicollinearity between them

Inference: The *F* Test

- **Test of overall significance of a regression**

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u$$

$$\boxed{H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0}$$

The null hypothesis states that the explanatory variables are not useful at all in explaining the dependent variable

$$y = \beta_0 + u$$

Restricted model
(regression on constant)

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1}$$

- **The test of overall significance is reported in most regression packages; the null hypothesis is usually overwhelmingly rejected**

Next Class

- **Freeing up the classical assumptions (heteroskedasticity)**