3.3 Confidence intervals for regression coefficients

Confidence intervals (interval estimation):
Specifying the region within which the "true" regression coefficient will lie with a large probability

Construction of a confidence interval for the unknown regression coefficients β_j

To this end, we start with the t-value of the regression coefficient $\hat{\beta}_j$ (standardized regression coefficient):

$$\text{(3.15)} \qquad t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{x^{jj}}}$$

which is t-distributed with n-k degrees of freedom. The probability that the standardized regression coefficient lies within a symmetric interval around its expected value 0, is given by

$$P\left(-t_{n-k;1-\alpha/2} \le t \le t_{n-k;1-\alpha/2}\right) = 1-\alpha$$

$t_{nk;\,1-\alpha/2}$: (1-α/ 2)-quantile of the t-distribution with n-k degrees of freedom

With (3.15), we obtain

$$P\left(-t_{n-k;1-\alpha/2} \le \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{x^{jj}}} \le t_{n-k;1-\alpha/2}\right) = 1-\alpha$$

and, after a further transformation

$$P\left(\hat{\beta}_j - t_{n-k;1-\alpha/2} \cdot \hat{\sigma}\sqrt{x^{jj}} \le \beta_j \le \hat{\beta}_j + t_{n-k;1-\alpha/2} \cdot \hat{\sigma}\sqrt{x^{jj}}\right) = 1-\alpha \tag{3.16}$$

Equation (3.16) indicates the confidence interval for β_j at a confidence level (= confidence probability) of 1-α. In the case of a normally distributed error term u, the unknown regression coefficient β_j lies with a probability of 1-α within the interval

$$\hat{\beta}_j \pm t_{n-k;1-\alpha/2} \cdot \hat{\sigma}\sqrt{x^{jj}}$$

Confidence intervals for β_1 and β_2 in the simple regression:

$$\hat{\beta}_1 \pm t_{n-2;1-\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{\frac{\sum x_t^2}{n\sum x_t^2 - (\sum x_t)^2}}$$

and

$$\hat{\beta}_2 \pm t_{n-2;1-\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{\frac{n}{n\sum x_t^2 - (\sum x_t)^2}}$$

Example:

As a point estimate for the marginal propensity to consume c_1 of the macro-economic consumption function the value 0,9335 has been determined. For economic policy measures it is however also of interest to know, at a high level of confidence, the range within which the marginal propensity to consume will likely lie. For this reason the point estimate for the marginal propensity to consume is to be supplemented with a **confidence interval**.

A **95% confidence interval** for the marginal propensity is given by

$$\hat{c}_1 \pm t_{n-2;0,975} \cdot \hat{\sigma} \cdot \sqrt{\frac{n}{n\sum x_t^2 - \left(\sum x_t\right)^2}}$$

The numerical values that are required in order to calculate the standard deviation of c_1 are already known:

$$\hat{\sigma} = 8{,}1103, \quad n = 19, \quad \sum x_t^2 = 37472832, \quad \sum x_t = 26529{,}9$$

Using the 0.975-quantile $t_{17;\,0.975} = 2.110$ of the t-distribution with 17 degrees of freedom one obtains the **confidence interval**

$$\left[0{,}9335 - 2{,}110 \cdot 8{,}11 \cdot \sqrt{\frac{19}{19 \cdot 37472832 - 26529{,}9^2}};\quad 0{,}9335 + 2{,}110 \cdot 8{,}11 \cdot \sqrt{\frac{19}{19 \cdot 37472832 - 26529{,}9^2}}\right]$$

$$\left[0{,}9335 - 2{,}110 \cdot 0{,}0124;\quad 0{,}9335 + 2{,}110 \cdot 0{,}0124\right] = \left[0{,}9074\,;\,0{,}9596\right]$$

for the marginal propensity to consume at the **95% confidence level**. □

Example:

In monetary policy ranges of the magnitude of monetary variables are being used for different purposes. Here we will undertake an interval estimation of the elasticities of income and the interest rate of the demand for money. The **95% confidence intervals** for the income and interest rate elasticities are:

$$\hat{\beta}_2 \pm t_{0,975;n-3} \cdot \sqrt{\widehat{\text{VAR}}\left(\hat{\beta}_2\right)}$$

and

$$\hat{\beta}_3 \pm t_{0,975;n-3} \cdot \sqrt{\widehat{\text{VAR}}\left(\hat{\beta}_3\right)}$$

The 0.975-quantile of the t-distribution with 16 degrees of freedom, resulting from the sample size of n = 19, is 2.120. With the estimates

$$\hat{\beta}_2 = 2{,}81 \qquad \text{und} \qquad \hat{\beta}_3 = -0{,}189$$

and the standard deviations

$$\sqrt{\widehat{\text{Var}}\left(\hat{\beta}_2\right)} = 0{,}159 \quad \text{und} \qquad \sqrt{\widehat{\text{Var}}\left(\hat{\beta}_3\right)} = 0{,}0297$$

one obtains for the income elasticity the 95% confidence interval

$$\left[2,81-2,120\cdot 0,159\,;\quad 2,81+2,120\cdot 0,159\right]=\left[2,4747;\quad 3,1478\right]$$

and for the interest rate elasticity the 95% confidence interval

$$\left[-0,189-2,120\cdot 0,0297\,;\quad -0,189+2,120\cdot 0,0297\right]=\left[-0,2521\,;\quad -0,1261\right] \qquad \square$$

Choice and interpretation of the confidence level 1-α:

Commonly used values for 1-α: 0.95 and 0.99 (as may be necessary 0.90)

When selecting a confidence level of 0.95, one has a 95% chance that the confidence interval includes the unknown regression coefficient β_j.

A trade off exists between accuracy (= length of the confidence interval) and confidence level (= confidence probability):
The higher the confidence level the more imprecise will be the interval estimation.

→ For a given sample, an improvement in the accuracy of the interval estimation can only be achieved by a reduction in the confidence level (for example from 99% to 95%).

In contrast, the risk that the confidence interval does not include the unknown regression coefficient β_j can only be reduced by an increase in the confidence level (for example from 95% to 99%).

If a confidence interval does not include 0, the implication is at the same time, that the estimated regression coefficient is significantly different from 0.

3.4 Analysis of variance and the significance of the overall model

The coefficient of determination is a measure of the goodness of fit that mainly is to be interpreted in a descriptive mode. However, the herein entered squared sums are random variables that allow for a variance analytical test. In an analysis of variance (= ANOVA) of an econometric model, the sum of the squared deviations of the values of the endogenous variable y from its mean $\bar{y}$ is partitioned into the sum of squares „explained" by the regression and the sum of deviation squares of the residuals. An ANOVA table, which forms the basis for a significance test of the overall model, displays also in addition the degrees of freedom and the mean squared deviations.

Overview: ANOVA table for a regression analysis

Source of dispersion	Sum of squared deviations	Degrees of freedom	Mean deviation squares
Regression (explained)	$\sum(\hat{y}_t - \bar{y})^2 = \hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{y}^2 = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$	k-1	$\left(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2\right) / (k-1)$
Residuals (unexplained)	$\sum \hat{u}_t^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}$	nk	$\hat{\mathbf{u}}'\hat{\mathbf{u}} \;/\; (n-k)$
Total	$\sum(y_t - \bar{y})^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2$	n-1	

Null hypothesis for a test of the overall model:

(3.17) $$H_0: \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

Test statistic of the test:

(3.18a) $$F = \frac{\left(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2\right) / (k-1)}{\hat{\mathbf{u}}'\hat{\mathbf{u}} / (n-k)}$$ or (3.18b) $$F = \frac{R^2/(k-1)}{\left(1-R^2\right)/(n-k)}$$

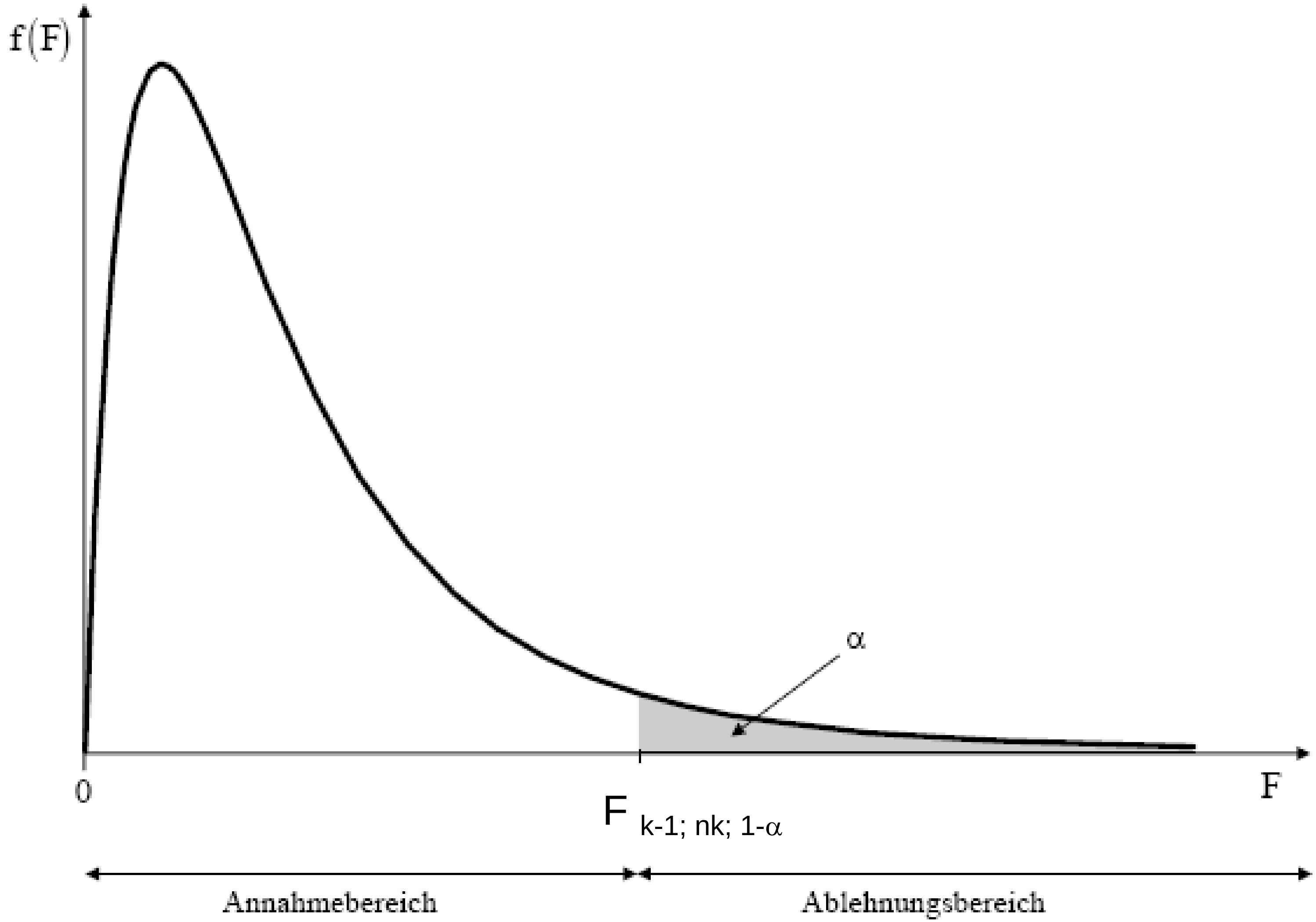
The test statistic is the ratio of the explained sum of squares to the residual sum of squares whereby the measures of dispersion are constructed in the form of mean deviation square, taking into account the different degrees of freedom. Under the null hypothesis the test statistics (3.18a) and (3.18b) of the significance test is F-distributed with k-1 and n-k degrees of freedom.

Test decision rule:

$$F > F_{1-\alpha;k-1;n-k} \rightarrow H_0 \text{ reject}$$

In the case of a rejection of the null hypothesis and with an error probabilty of α, all regressors can no longer be considered as irrelevant. The overall regression model viewed as a whole can then be said to provide a significant contribution to the explanation of the dependent variable y.

Figure: Acceptance and rejection regions of the F-test



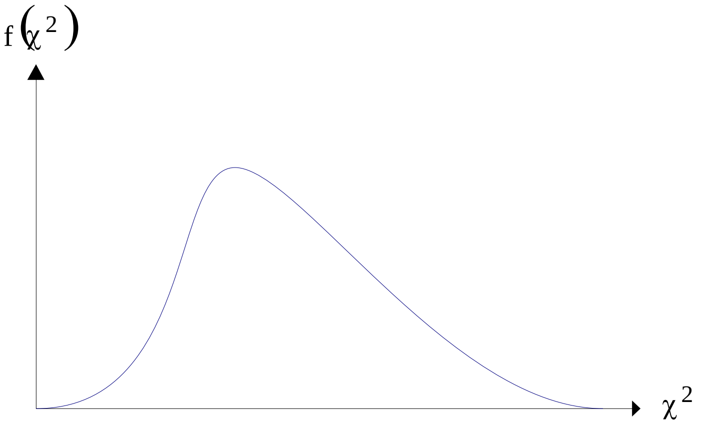
Excursion: Chi-square and F-distribution

- **Chi-Square distribution**

Let $Z_1, Z_2, \ldots, Z_n$,

$$Z_i = \frac{X_i - \mu}{\sigma}\;2',$$

be n independently, normally distribute random variables. Then the square sum



$$(*)\ \chi^2 = Z_1^2 + Z_2^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$$

follows a χ^2-distribution with n degrees of freedom. Thereby the square sum can be written in the form

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2}\ .$$

If μ is estimated by the sample mean, the sum of deviation squares

$$\chi^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sigma^2}$$

is χ^2-distributed with n-1 degrees of freedom. In general, a square sum of the form (*) is χ^2-distributed with n-k degrees of freedom if k parameters are estimated.

- **F-Verteilung**

Let χ^2_{v1} and χ^2_{v2} be two independently χ^2 distributed random variables with v_1 and v_2 degrees of freedom. Then the proportion

$$F = \frac{\chi_1^2 / v_1}{\chi_2^2 / v_2}$$

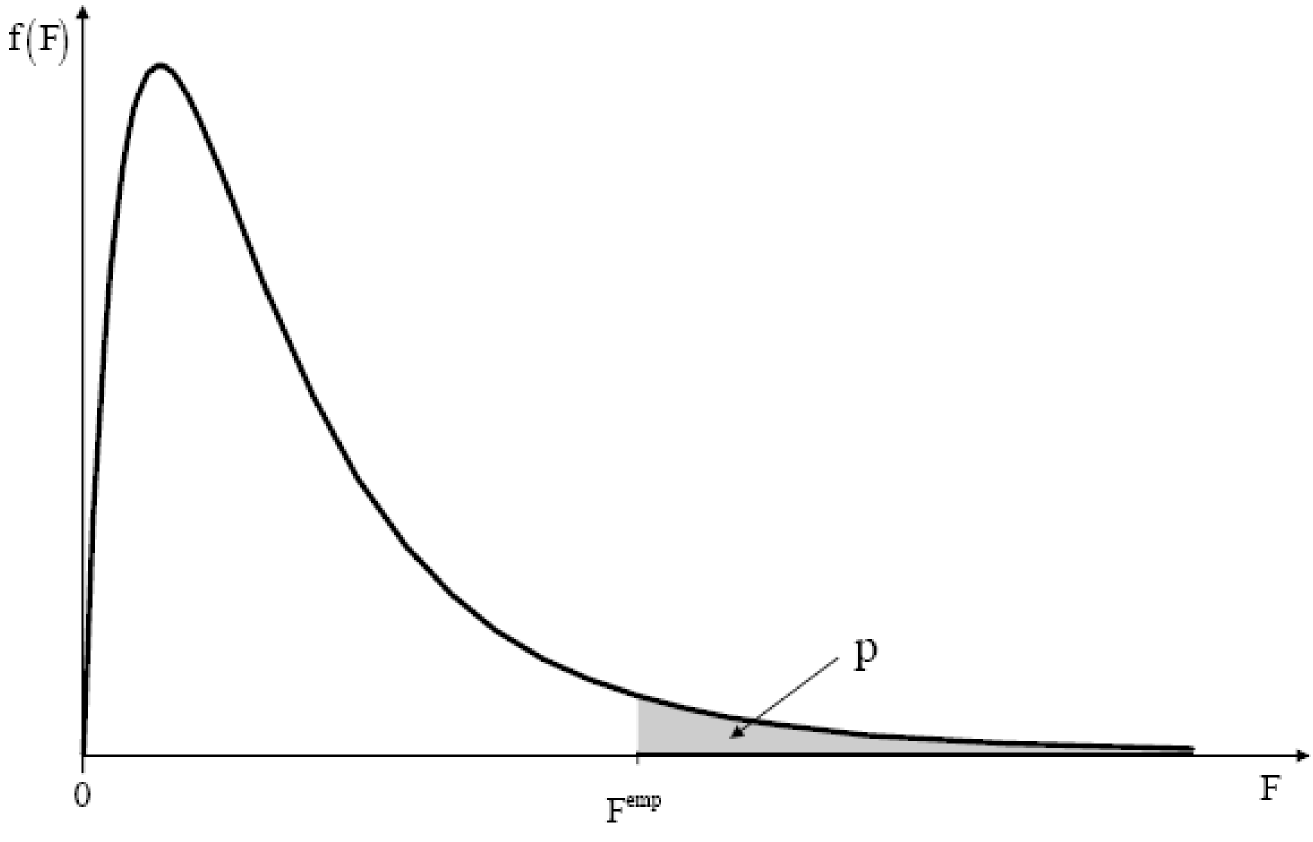
is F distributed with v_1 and v_2 degrees of freedom.

F-test and p-value

Significant deviations from the null hypothesis always show up in the F-test as a large positive empirical F-value F_{emp} (= value of the test variable F). Qua being a quotient of two independent χ^2-distributed random variables, the test value always takes on a positive sign. For this reason, the F-test is always a one-sided test. The **p-value** then corresponds to the **one-sided exceedance probability** ,

$$p = P(F > F_{emp}) = 1 - P(F \le F_{emp}).$$

Fig .: p-value at F test



Example:

The econometric analysis of the aggregate consumption function of the Federal Republic of Germany in the support period yields the following ANOVA table:

Source of dispersion	Sum of squared deviations	Degrees of Freedom	Mean deviation squares
Regression (explained)	373741,1	1	373741,1
Residuals (unexplained)	1,118,2	17	65,8
Total	395684,8	18	20825,5

From this, the test statistic (3.11a) and (3.11b) can be calculated:

$$F = \frac{373741{,}1}{65{,}8} = 5679{,}96$$ and $$F = \frac{0{,}997/1}{(1-0{,}997)/17} = 5649{,}67$$ [differences due to rounding]

Test decision (α = 0.01):

$F = 5679.96 > F_{1;\,17;\,0.99} = 8.40 \rightarrow H_0$ reject

The overall model is therefore statistically significant.

Example:

In a capital market model with three exogenous variables (incl. the dummy variable) the coefficient of determination is 0.303. The regression model has been estimated from 16 related observations. Now the question arises whether the exogenous variables do have an overall impact on the endogenous variable or whether they are substantially meaningless.

Clarification on this point at issue can be provided by an F-test for the significance of the determination coefficient. At a significance level of 5%, the test statistic is

$$F = \frac{R^2 / (k-1)}{(1-R^2)/(n-k)} = \frac{0{,}303/(3-1)}{(1-0{,}303)/(16-3)} = 2{,}826$$

which is to be compared with the 0,95 quantile $F_{0,95;\,2;\,15}$ of the F-distribution with 2 and 13 degrees of freedom. Because of

$$F = 2{,}826 < F_{0,95;2;13} = 3{,}82$$

the null hypothesis (3.10) can not be rejected, i.e. the influence of the exogenous variables on the endogenous variable of the capital market model are statistically not secured. □