# Multiple Regression Analyses: ***<u>Statistical Inference</u>***

October 18, 2020

Dali Laxton

# Today's Lecture

- We are going to discuss how hypotheses about coefficients can be tested in regression models
- We will explain what significance of coefficients mean
- We will learn how to read regression output

  - Wooldridge Chapter 4;
  - Studenmund Chapter 5.1-5.4

# Multiple Regression Analyses: ***Inference***

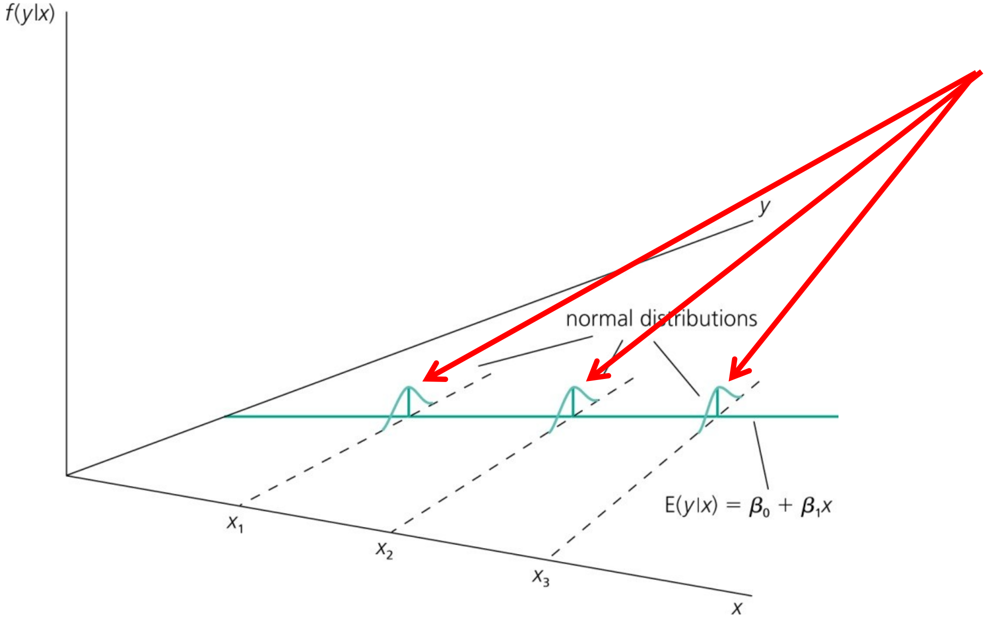
- **Statistical inference in the regression model**
  - Hypothesis tests about population parameters
  - Construction of confidence intervals

- **Sampling distributions of the OLS estimators**
  - The OLS estimators are random variables
  - We already know their expected values and their variances
  - For hypothesis testing we need to know their distribution

# *Inference*: Sampling distributions of the OLS Estimators

- **Assumption 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.

# Show normality of the error terms in GRETL

- Open GRETL load sample data “Engel”
- Run regression Ols foodexp income
- Generate residuals:

genr resid =foodexp-( $coeff(const) + $coeff(income)*income)

- Display distribution of residuals:

freq resid --plot=display

# *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? Observations 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)
    - Unemployment (indicator variable, takes on only 1 or 0)
  - In some cases, normality can be achieved through transformations of the dependent variable
  - 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 (CLT)

# Multiple Regression Analyses: ***Hypothesis Testing***

- We cannot prove that a given hypothesis is "correct" using hypothesis testing
- All we can do is to state that a particular sample conforms to a particular hypothesis
- We can often reject a given hypothesis with a certain degree of confidence
- In such a case, we conclude that it is very unlikely the sample result would have been observed if the hypothesized theory were correct

# Multiple Regression Analyses: ***Hypothesis Testing***

- Step 1: state explicitly the hypothesis to be tested
- ***Null hypothesis***: statement of the range of values of the regression coefficient that would be expected to occur if the researcher's ***theory were not correct***
- ***Alternative hypothesis****:* specification of the range of values of the coefficient that would be expected to occur if the researcher's ***theory were correct***
- In other words, we define the null hypothesis as the result we do not expect

# Type I and Type II Errors

- It would be unrealistic to think that conclusions drawn from regression analysis will always be right
- There are two types of errors we can make:
  - Type I: we reject a true null hypothesis
  - Type II: We fail to reject a false null hypothesis

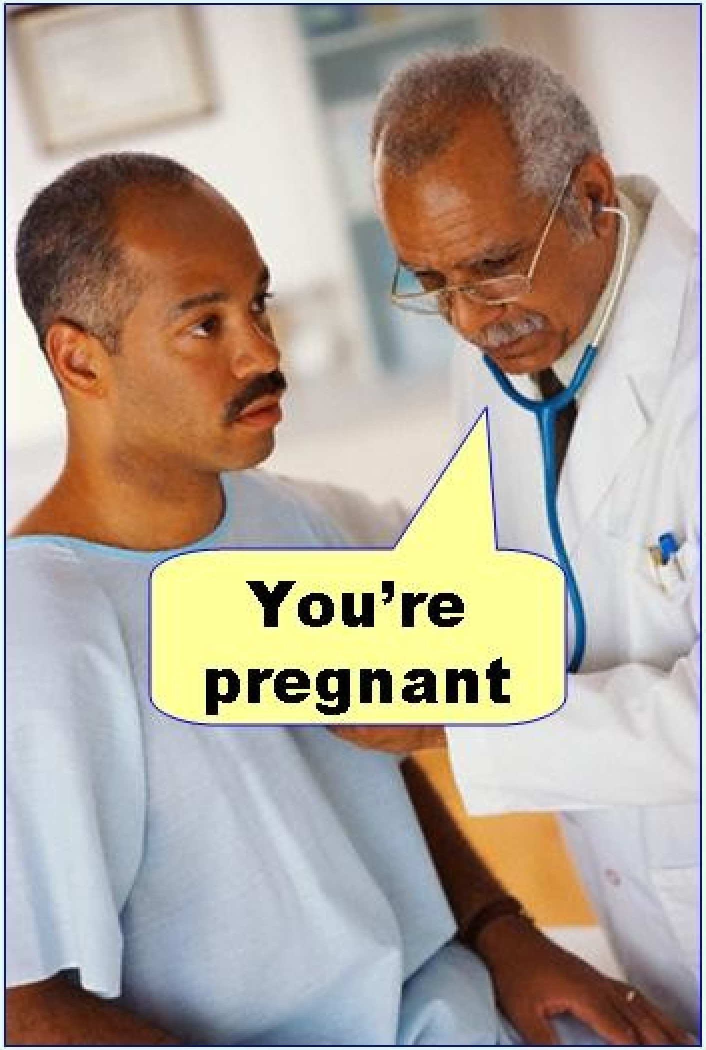
# Type I and Type II Errors

Example:

- $H_0$: The defendant is innocent
- $H_A$: The defendant is guilty
  - Type I error: sending an innocent person to jail
  - Type II error: freeing a guilty person
- Lowering the probability of Type I error means increasing the probability of Type II error;
- In hypothesis testing, we focus on Type I error and we ensure that its probability is not unreasonably large

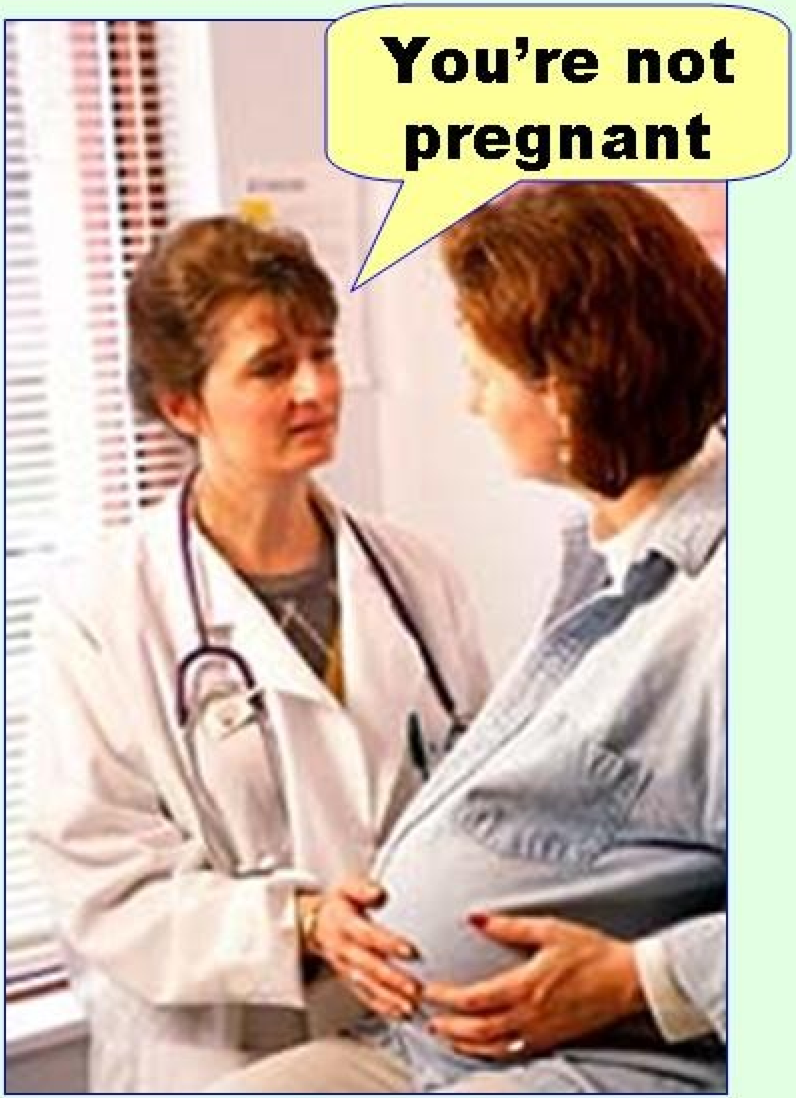
**Type I error**

(false positive)




**Type II error**

(false negative)



# *Inference*: The *t* Test

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

Under assumptions 1 – 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.

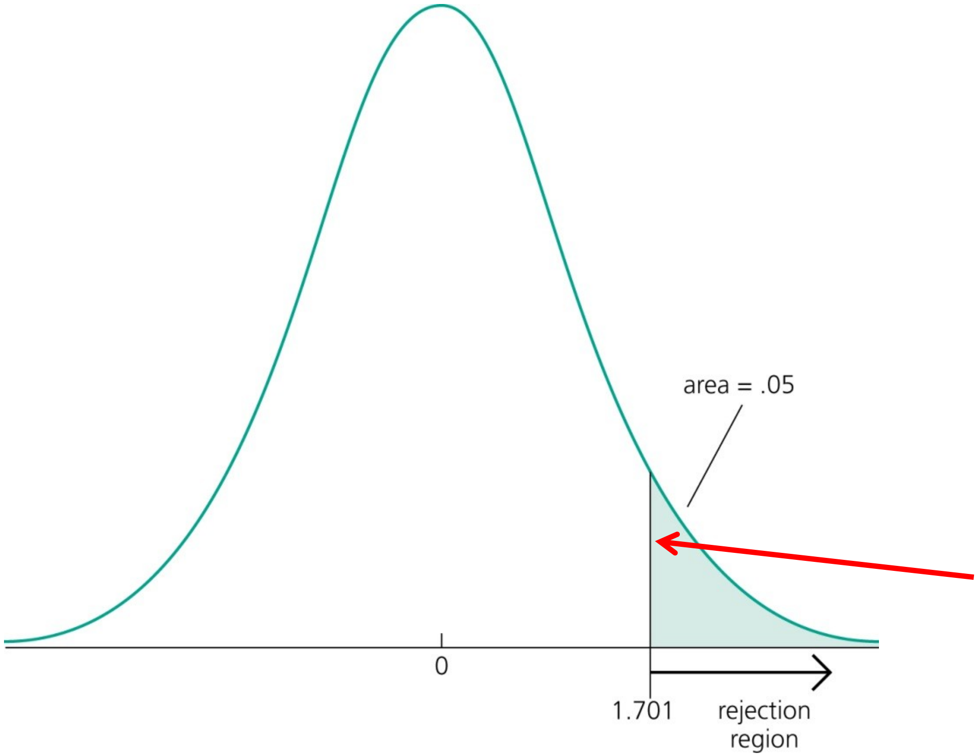
- **Distribution of the t-statistic if the null hypothesis is true**

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

- **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.)**

t-statistic

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

Degrees of freedom; here the standard normal approximation applies

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

$$c_{0.05} = 1.645$$

$$c_{0.01} = 2.326$$

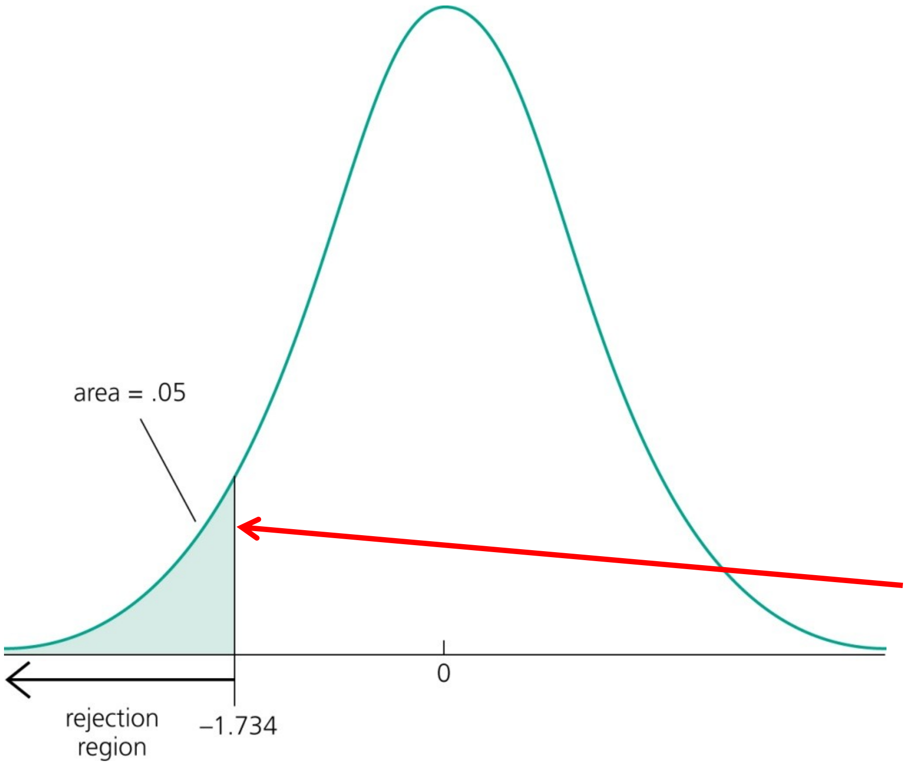
Critical values for the 5% and the 1% significance level (these are conventional significance levels).

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 coef-ficient is „too small" (i.e. smaller than a criti-cal 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)

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

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

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.)**

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

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.

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

# *Inference*: The *t* Test

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

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

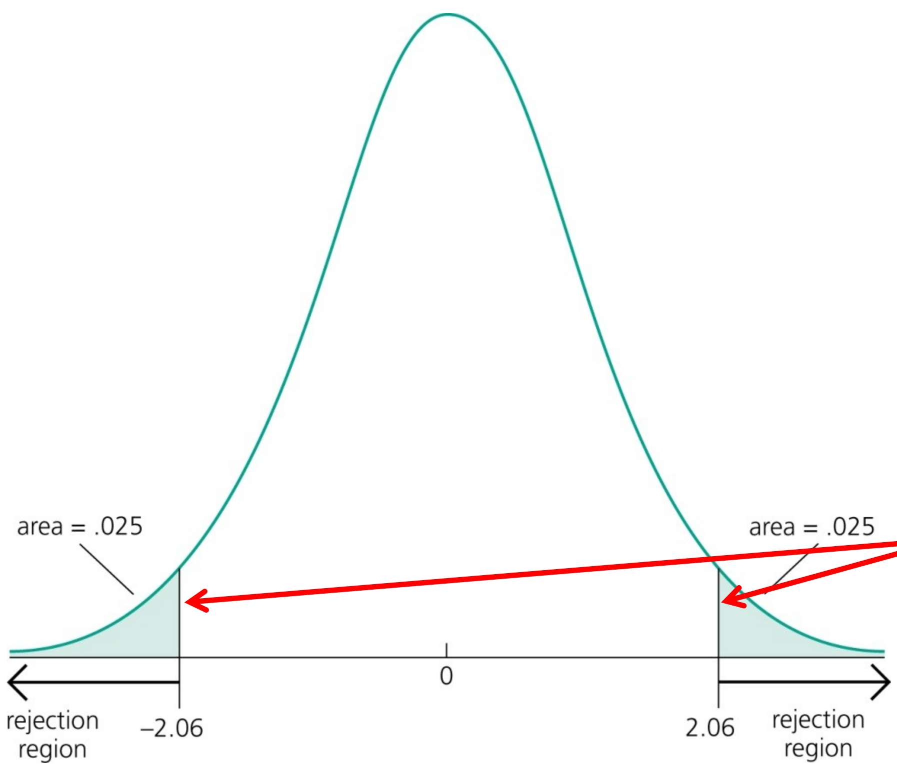
$$+ \underset{(4.19)}{3.98}\log(staff) - \underset{(0.69)}{1.29}\log(enroll)$$

$$n = 408, \ R^2 = .0654$$ ← R-squared slightly higher

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

# *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 (→ skipped)

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

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

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**

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

- **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?

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

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

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

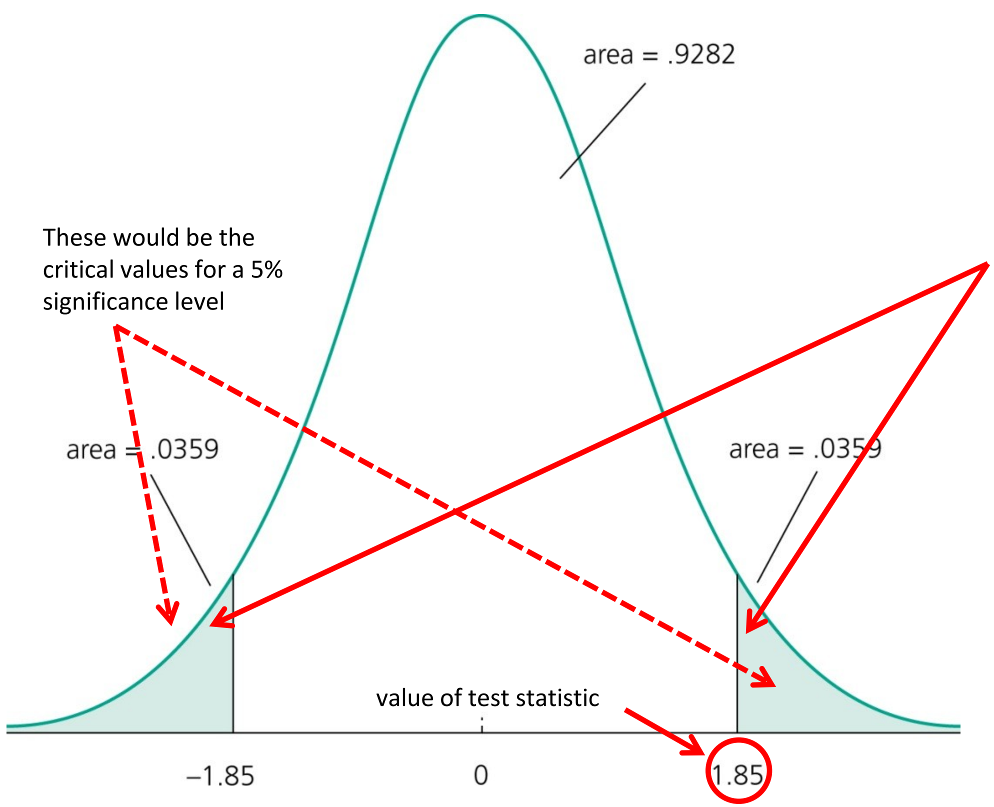
The hypothesis is rejected at the 5% level

# *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 <u>a null hypothesis is rejected if and only if the corresponding p-value is smaller than the significance level.</u>

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**

Critical value of two-sided test

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

Lower bound of the Confidence interval

Upper bound of the Confidence interval

Confidence level

- **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**

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

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

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

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**

$$a_j \notin interval \;\Rightarrow\; \text{reject } H_0 : \beta_j = a_j \text{ in favor of } H_1 : \beta_j \neq 0$$

# *Inference*: Confidence Intervals

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

Spending on R&D — Annual sales — Profits as percentage of sales

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

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

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

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

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 → $jc$

Years of education at 4 year colleges → $univ$

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

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**

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

Usually not available in regression output

- **Alternative method**

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

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

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

Insert into original regression

a new regressor (= total years of college)

# *Inference*: Testing hypotheses about a linear combination of parameters

- **Estimation results**

Total years of college

$$\widehat{\log}(wage) = \underset{(.021)}{1.472} - \underset{(.0069)}{.0102}\, jc + \underset{(.0023)}{.0769}\, totcoll + \underset{(.0002)}{.0049}\, exper$$

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

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

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

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

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

- **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

$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ against $H_1 : H_0$ is not true

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

# *Inference*: The *F* Test

- **Estimation of the unrestricted model**

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

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

None of these variabels are statistically significant when tested individually

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

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

# *Inference*: The *F* Test

- **Estimation of the restricted model**

$$\widehat{\log}(salary) = \underset{(0.11)}{11.22} + \underset{(.0125)}{.0713}\ years + \underset{(.0013)}{.0202}\ gamesyr$$

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

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

- **Test statistic**

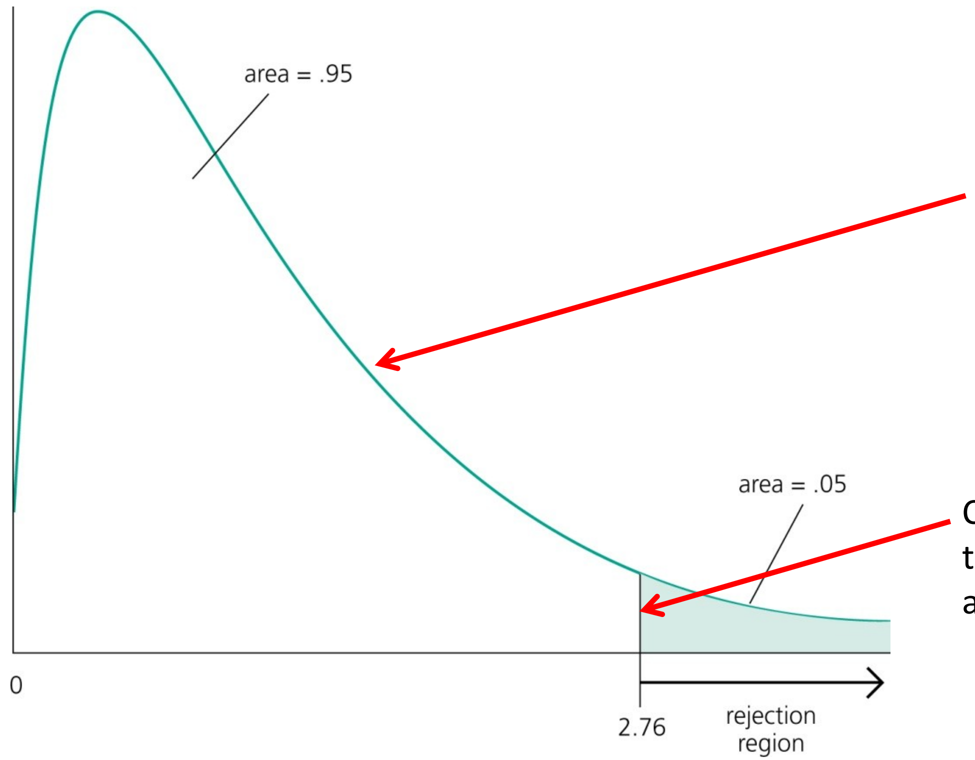
Number of restrictions

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

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**




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

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

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**

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

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

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**