Lecture 4: Multiple Regression Model: Inference

Xi Qu

Fall 2019

LEARNING OBJECTIVE

- Sampling distribution of the OLS estimators
- Perform a hypothesis testing
  - Single parameter
  - A linear combination of parameters
  - Multiple linear restrictions

Multiple Regression Analysis: Inference

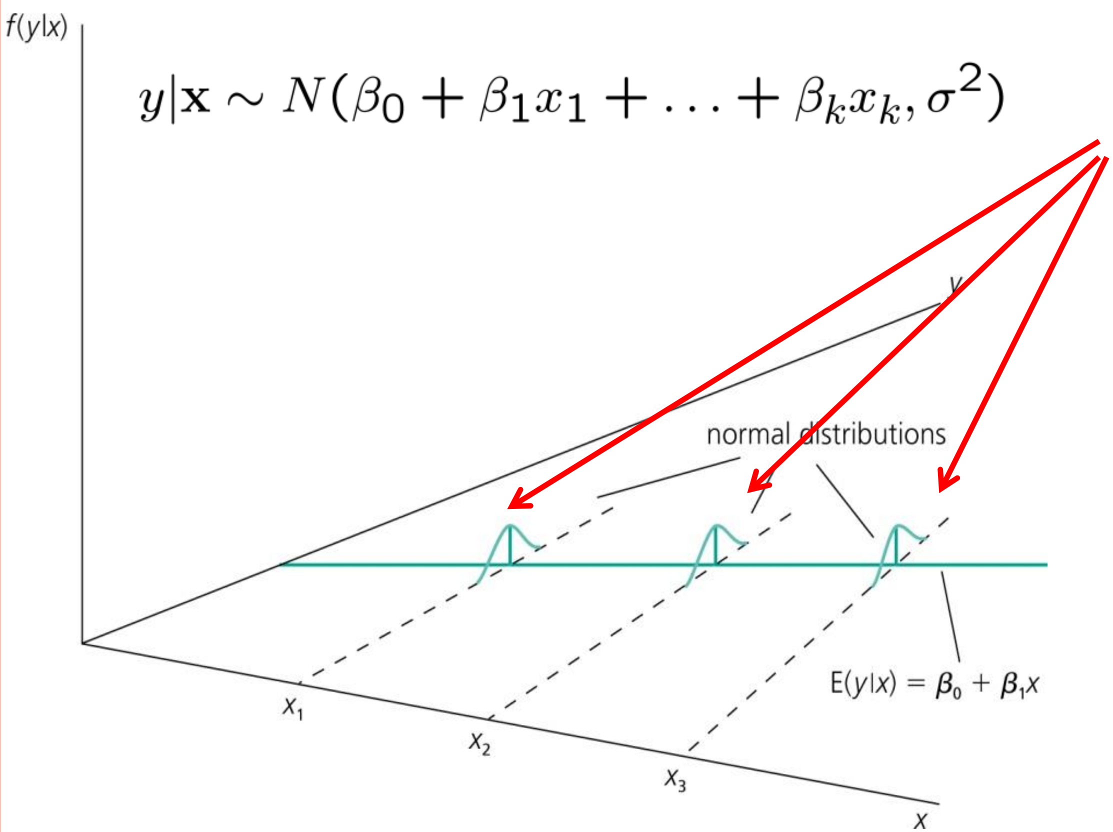
- **Statistical inference in the regression model**
  - Hypothesis tests about population parameters
  - Construction of confidence intervals
- **Sampling distributions of the OLS estimators**
  - The OLS estimators are random variables
  - We already know their expected values and their variances
 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

Assumption

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

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

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



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.

Discussion

Discussion of the normality assumption

- The error term is the collection of all unobserved factors
- Problems:
  - How many different factors?
  - Heterogeneous distributions of individual factors
  - Independence?
- 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

Discussion (Con‘t)

- 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 <u>transformations</u> of the dependent variable (e.g. use log(wage) instead of wage)

- Under normality, OLS is the **best unbiased estimator** (BLUE?)

- For the purposes of statistical inference, the assumption of normality can be replaced by a large sample size.

QUESTION

- 1. The normality assumption implies that:

a. the population error u is independent on the explanatory variables and is normally distributed with mean equal to zero and variance σ^2.

b. the population error u is independent of the explained variables and is normally distributed with mean equal to one and variance σ^2.

c. the population error u is dependent of each other and is normally distributed with mean zero and variance σ^2.

d. the residual $\hat{u}$ is independent of the explanatory variables and is normally distributed with mean zero and variance σ^2.

- 2. Which of the following statements is true?

a. Taking a log of a nonnormal distribution might yield a distribution that might be closer to normal.

b. The mean of a nonnormal distribution is 0 and the variance is σ^2.

c. The CLT assumes that the dependent variable is unaffected by unobserved factors.

d. OLS estimators have the highest variance among unbiased estimators.

Assumptions

- **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:

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

The estimators are normally distributed around the true parameters with its variance

The standardized estimators follow a standard normal distribution

- **<u>Testing hypotheses about a single population parameter</u>**
- **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 <u>estimated</u> standard deviation (= standard error), the normal distribution is replaced by a t distribution (σ^2 known or unknown)

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

- **Null hypothesis**

$$H_0: \ \beta_j = 0$$

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

t statistic

- **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.

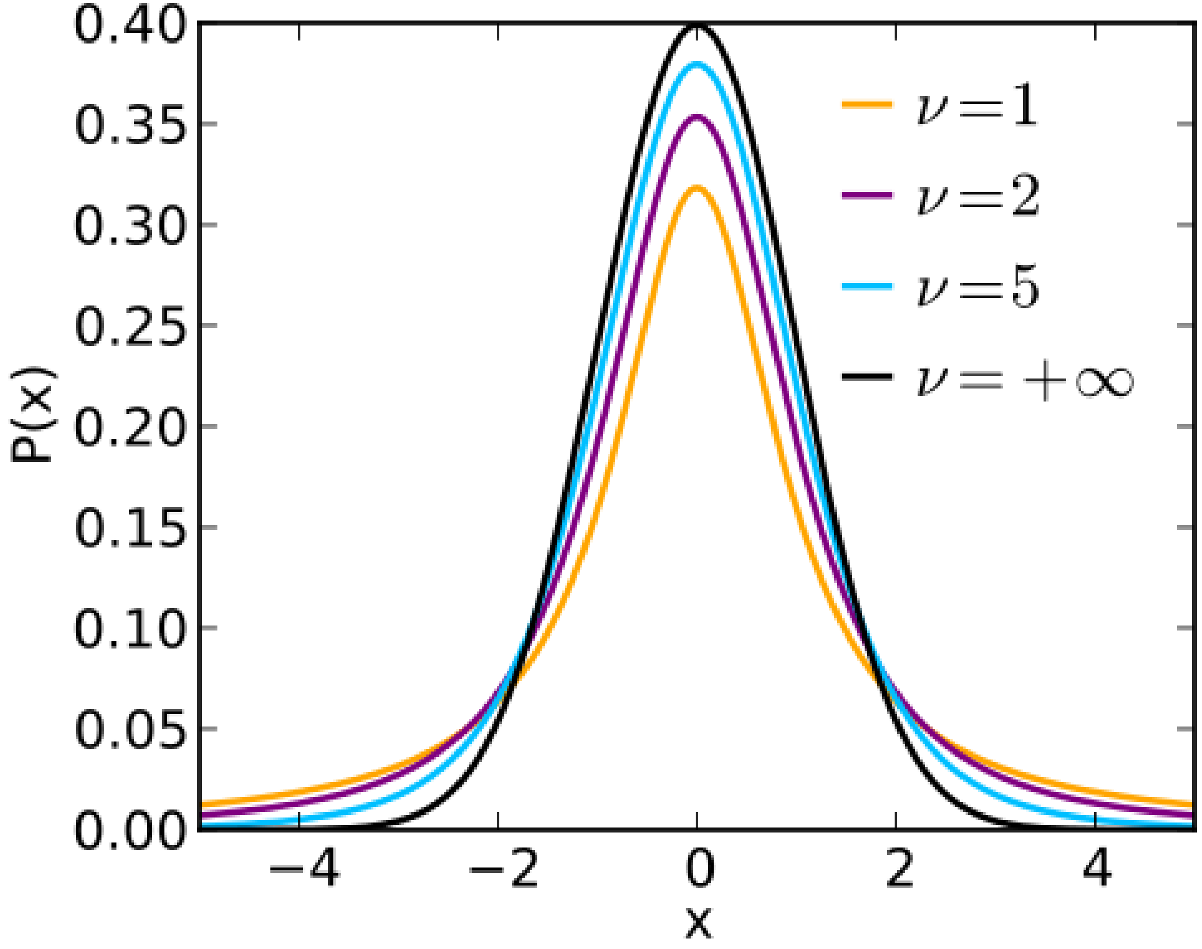
This depends on the variability of the estimated coefficient, i.e. its standard deviation. <u>The t-statistic measures how many estimated standard deviations the estimated coefficient is away from zero.</u>

- **Distribution of the t-statistic <u>if the null hypothesis is true</u>**

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

- **<u>Goal</u>: Define a rejection rule so that, if it is true, H_0 is rejected with a small probability (= significance level, e.g. 5% or 1%)**

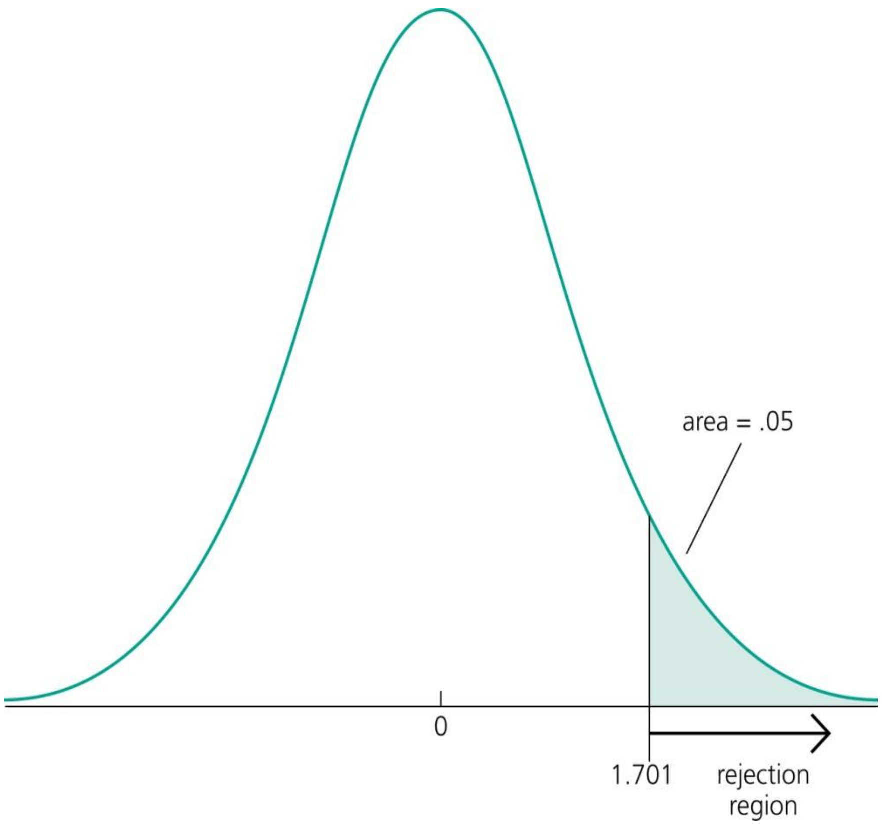
Student's t distribution



Rejection Rule (1)

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

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



Reject H_0 in favour of H_1 if the estimated coefficient is <u>"too large"</u> (i.e. larger than a critical value).

Construct the critical value so that, if H_0 is true, it is rejected in, for example, 5% of the cases.

In the example of n-k-1=28, this is the point of the t-distribution with 28 degrees of freedom that is exceeded in 5% of the cases.

<u>→ Reject if t-statistic greater than 1.701</u>

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

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

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

Example (con't)

- **The test statistic and rejection rule**

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

t-statistic

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

Degrees of freedom; 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.

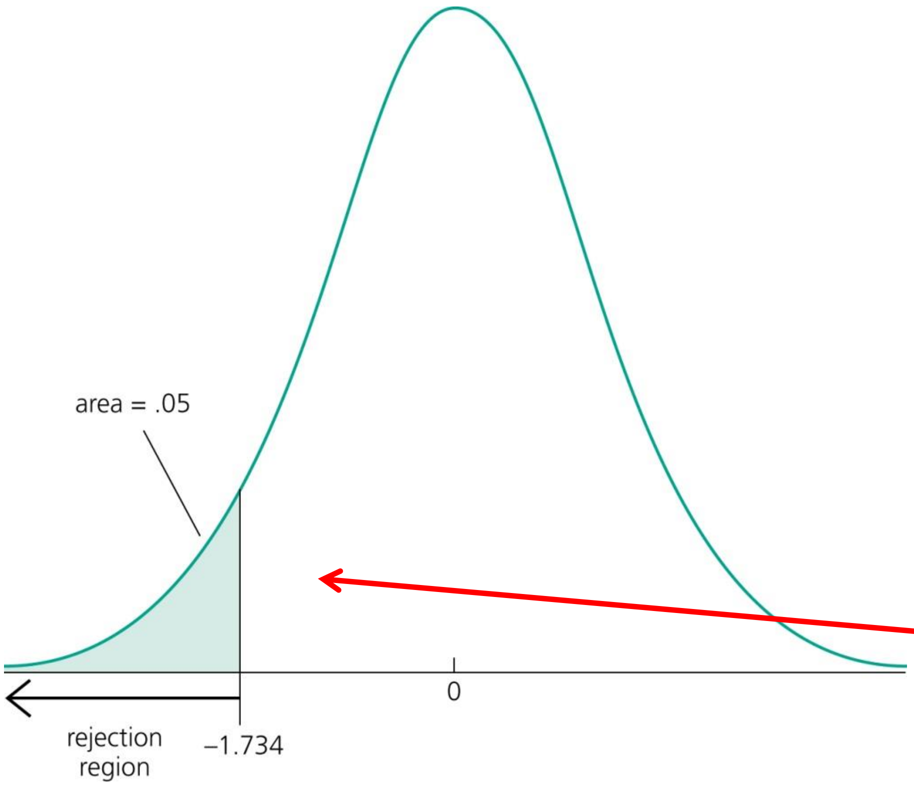
The null hypothesis is rejected because the t-statistic exceeds the critical value.

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

Rejection Rule (2)

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

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



Reject H_0 in favour of H_1 if the estimated coefficient is "too small" (i.e. smaller than a critical value).

Construct the critical value so that, if H_0 is true, it is rejected in, for example, 5% of the cases.

In the example of n-k-1=18, 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

Example: Student behavior

- Test whether smaller school size leads to better student performance

Percentage of students passing maths test

Staff per one thousand students

$$\widehat{math}10 = + \underset{(6.113)}{2.274} + \underset{(.00010)}{.00046}\ totcomp + \underset{(.040)}{.048}\ staff - \underset{(.00022)}{.00020}\ enroll$$

Average annual teacher compensation

School enrollment(= school size)

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

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

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

Example (con't)

- **t statistic**

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

t-statistic

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

Degrees of freedom; the standard normal approximation applies

$$c_{0.05} = -1.65$$

$$c_{0.15} = -1.04$$

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

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

<u>Conclusion:</u>
<u>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>

Example (con't)

- 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 $H_1 : \beta_{\log(enroll)} < 0$

t statistic

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

t-statistic

$$c_{0.05} = -1.65$$

Critical value for the 5% significance level
$\rightarrow$ reject null hypothesis

Conclusion:
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 $\rightarrow$ -0.129 percentage points students pass test

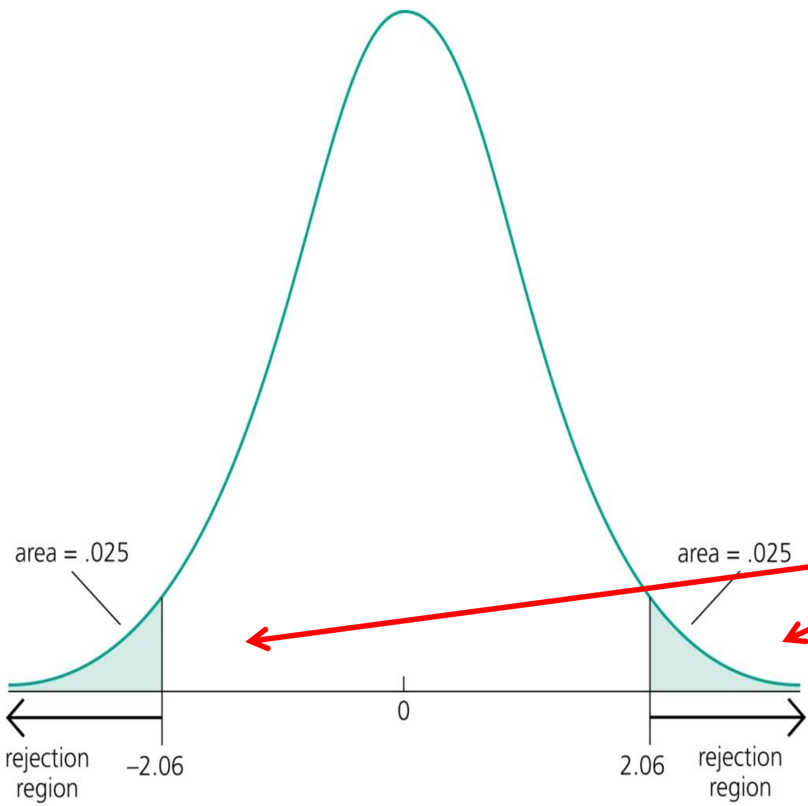
$$-1.29 = \frac{\partial math10}{\partial \log(enroll)} = \frac{math10}{\frac{\partial enroll}{enroll}} = \frac{\frac{-1.29}{100}}{\frac{1}{100}} = \frac{-0.0129}{+1\%}$$

(small effect)

Rejection Rule (3)

- **Testing against two-sided alternatives**

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



Reject H_0 in favour of H_1 if the absolute value of the estimated coefficient is too large.

Construct the critical value so that, if H_0 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.

$\rightarrow$ Reject if absolute value of t-statistic is less than -2.06 or greater than 2.06

Example: College GPA

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.

Statistical Significance

- 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:

$|t - ratio| > 1.645$ $\rightarrow$ statistically significant at 10 % level

$|t - ratio| > 1.96$ $\rightarrow$ statistically significant at 5 % level

$|t - ratio| > 2.576$ $\rightarrow$ statistically significant at 1 % level

QUESTION

- 3. A normal variable is standardized by:

a. subtracting off its mean from it and multiplying by its standard deviation.

b. adding its mean to it and multiplying by its standard deviation.

c. subtracting off its mean from it and dividing by its standard deviation.

d. adding its mean to it and dividing by its standard deviation.

- 4. Consider the equation, $Y = \beta_1 + \beta_2 X_2 + u$. A null hypothesis, H_0: $\beta_2 = 0$ states that:

a. X_2 has no effect on the expected value of β_2.

b. X_2 has no effect on the expected value of Y.

c. β_2 has no effect on the expected value of Y.

d. Y has no effect on the expected value of X_2.

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
- <u>The fact that a coefficient is statistically significant does not necessarily mean it is economically or practically significant!</u>
  - 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

Testing more general hypotheses

- **Null hypothesis**

$$H_0:\ \beta_j = a_j$$

Hypothesized value of the coefficient

- **t-statistic**

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

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

Example

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

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

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

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

$$\text{Test } H_0 : \beta_{\log(enroll)} = 1 \text{ against } H_1 : \beta_{\log(enroll)} \neq 1$$

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

The hypothesis is rejected at the 5% level

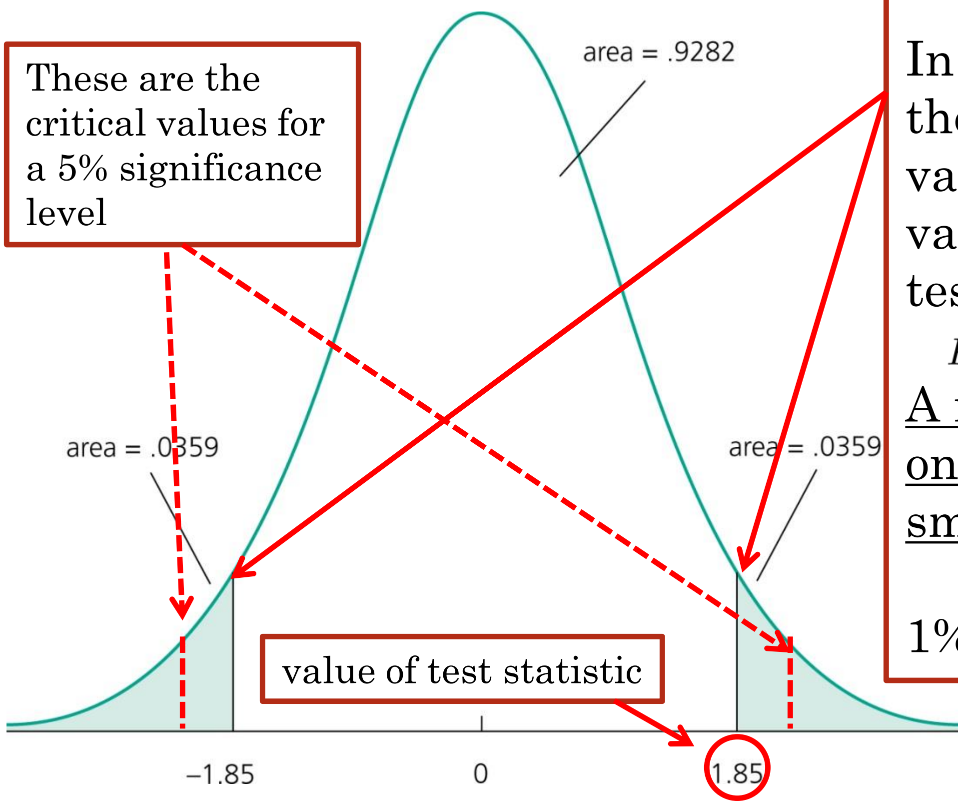
Alternative rejection rule

- **Computing p-values for t-tests**
  - If the significance level is made smaller and smaller (10%-5%-1%), there will be a cutoff point (a probability) 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

Compute the p-value

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

These are the critical values for a 5% significance level



value of test statistic

The p-value is the significance level at which one is indifferent between rejecting and not rejecting H_0

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

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

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

1%? 5%? 10%?

Confidence Intervals

- **Theorem 4.2 implies that**

Critical value of two-sided test

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

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 (better than saying "the population parameter falls in the interval in 95% of the cases")

- **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 in large samples

$$c_{0.01} = 2.576, c_{0.05} = 1.96, c_{0.10} = 1.645$$

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

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

Example

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

Spending on R&D

Annual sales

Profits as percentage of sales

$$\widehat{\log(rd)} = \underset{(.47)}{-4.38} + \underset{(.060)}{1.084} \log(sales) + \underset{(.0218)}{.0217}\, profmarg$$

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

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

$$.0217 \pm 2.045(.0218) = \boxed{(-0.023, 0.067)}$$

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.

Question

- 5. The significance level of a test is:

a. the probability of rejecting the null hypothesis when it is false.

b. one minus the probability of rejecting the null hypothesis when it is false.

c. the probability of rejecting the null hypothesis when it is true.

d. one minus the probability of rejecting the null hypothesis when it is true.

- 6. Which of the following statements is true?

a. When the standard error of an estimate increases, the confidence interval for the estimate narrows down.

b. Standard error of an estimate does not affect the confidence interval for the estimate.

c. The lower bound of the confidence interval for a regression coefficient, say β_j, is given by $\hat{\beta}_j$ -[standard error × ($\hat{\beta}_j$)]

d. The upper bound of the confidence interval for a regression coefficient, say β_j, is given by $\hat{\beta}_j$ + [Critical value × standard error ($\hat{\beta}_j$)]

More on hypotheses test

- **<u>Testing hypotheses about a linear combination of parameters</u>**
- **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

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

$$\text{Test } H_0 : \beta_1 - \beta_2 = 0 \text{ against } H_1 : \beta_1 - \beta_2 < 0$$

A possible test statistic would be:

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

The difference between the estimates is normalized by the estimated standard deviation of the difference. H_0 is rejected if the statistic is "too negative" to believe that the true difference between the parameters is equal to zero.

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

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

- **Alternative method**

Usually not available in regression output

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)

Estimation results

Total years of college

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

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

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

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

- To implement the test of linear combinations using software, either
- Run the original regression, and the next step is to do the hypothesis test; or
- Run the transformed model and directly get the t-statistic and p-value.

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

Salary of major league base ball 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

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

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

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 is statistically significant when tested individually

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

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

- **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?

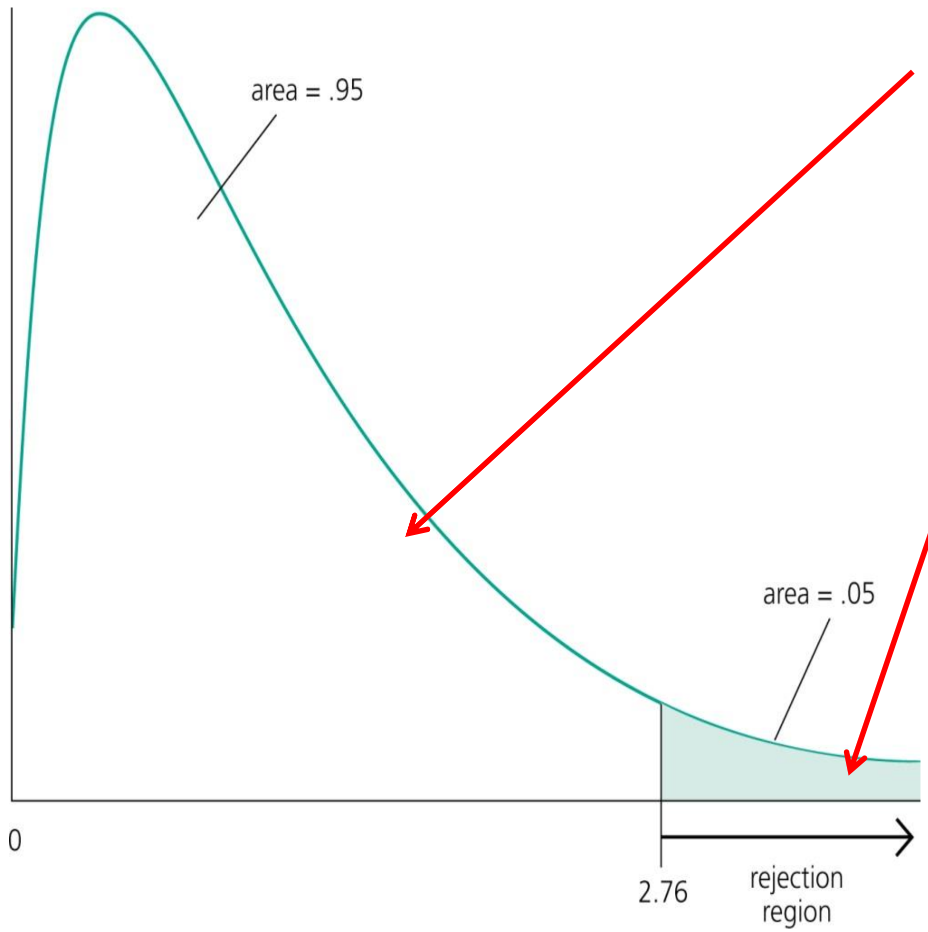
- **Test statistic**

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)

- **Rejection rule (Figure 4.7)**



An 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.

- **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

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

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

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

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

$$y = \beta_0 + u \quad \leftarrow$$

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**

- **<u>Testing general linear restrictions with the F-test</u>**
- **Example: Test whether house price assessments are rational**

Actual house price

The assessed housing value (before the house was sold)

Size of lot (in feet)

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize)$$
$$+ \beta_3 \log(sqrft) + \beta_4 bdrms + u$$

Square footage

Number of bedrooms

$$\boxed{H_0 : \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0}$$

In addition, other known factors should not influence the price once the assessed value has been controlled for.

If house price assessments are rational, a 1% change in the assessment should be associated with a 1% change in price.

- **Unrestricted regression**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_3 x_3 + \beta_4 x_4 + u$$

- **Restricted regression**

$$y = \beta_0 + x_1 + u \quad \Rightarrow \quad [y - x_1] = \beta_0 + u$$

The restricted model is actually a regression of $[y\text{-}x_1]$ on a constant

- **Test statistic**

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(1.880 - 1.822)/4}{1.822/(88-4-1)} \approx .661$$

$$F \sim F_{4,83} \quad \Rightarrow \quad c_{0.05} = 2.50 \quad \Rightarrow \quad H_0 \quad \text{cannot be rejected}$$

- **Regression output for the unrestricted regression**

$$\widehat{\log}(price) = \underset{(.570)}{.264} + \underset{(.151)}{1.043} \log(assess) + \underset{(.0386)}{.0074} \log(lotsize)$$

$$- \underset{(.1032)}{.1384} \log(sqrft) + \underset{(.0221)}{.0338}\, bdrms$$

$$n = 88, SSR = 1.822, R^2 = .773$$

When tested individually, there is also no evidence against the rationality of house price assessments

- **The F-test works for general multiple linear hypotheses**
- **<u>For all tests and confidence intervals, validity of assumptions MLR.1 – MLR.6 has been assumed.</u>**
- **<u>Tests may be invalid otherwise.</u>**