Chapter 4: Multiple Regression Analysis: Inference

Introductory Econometrics: A Modern Approach

Objectives

In this chapter, we are testing hypothesis about parameters in the population regression model:

- find distribution of the OLS estimators
- cover hypothesis testing about individual parameters
- construction of confidence intervals
- how to test a single hypothesis involving more than one parameter
- determine whether a group of independent variables can be omitted from a model

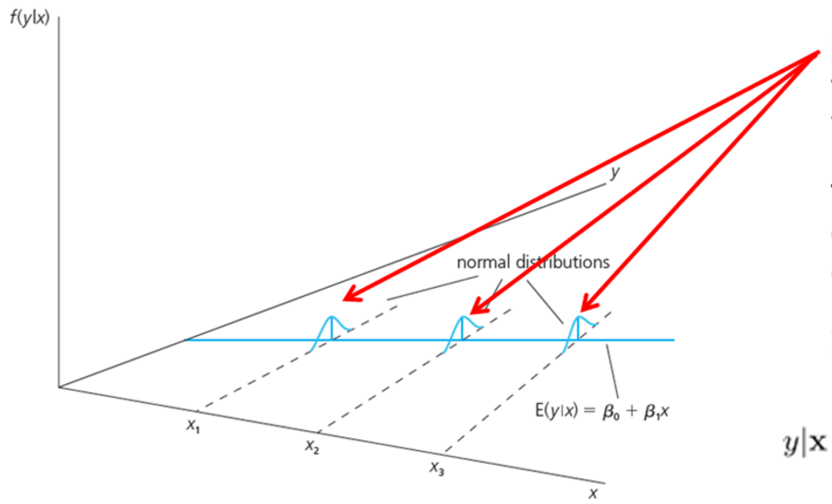
4.1 Sampling Distributions of the OLS Estimators

- 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

MLR.6 Normality of error terms

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

Normality of error terms



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:

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

Discussion of the normality assumption

The error term is the sum of “many” different unobserved factors. Sums of independent factors are normally distributed.

Problems:

- How many different factors? Number large enough?
- Possibly very heterogeneous 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.

Discussion of the normality assumption

Examples where normality cannot hold:

- Wages (non-negative; 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.

Law of large numbers states that as the sample size gets larger, the sample mean approaches the true population mean.

Law of large numbers--Simulations

Terminology

MLR.1- MLR.5 "Gauss- Markov Assumption"

MLR.1- MLR.6 "Classical Lineal Model (CLM) assumptions"

Theorem 4.1 (Normal sampling distributions), under MLR.1- MLR.6 assumptions

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

$$\widehat{\beta}_j \sim \mathrm{Normal}\left(\beta_j, \mathrm{Var}\left(\hat{\beta}_j\right)\right)$$

The standardized estimators follow a standard normal distribution

$$\frac{\widehat{\beta}_j - \beta_j}{sd\left(\widehat{\beta}_j\right)} \sim \mathrm{Normal}(0, 1)$$

4.2 Testing hypothesis about a single population parameter

Theorem 4.2 `t distribution` for the standardized estimators

If the standardization is done using the estimated standard deviation (called standard error), the normal distribution is replace by a `t-distribution`

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

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

Null hypothesis

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

$$H_0 : \beta_j = 0$$

t-statistic (or t-ratio)

$$t_{\hat{\beta}_j} \equiv \frac{\hat{\beta}_j}{\operatorname{se}\left(\widehat{\beta}_j\right)}$$

- The t-statistic will be used to test the above null hypothesis.
- The farther the t-statistic is away from zero, the less likely it is that the null hypothesis holds true.

- But what does "far" away from zero mean?

    - How many standard deviations?

- **The t-statistic measures how many estimated standard deviations the estimated coefficient is away from zero.**

Distribution of the t-statistic

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

$$t_{\widehat{\beta}_j} \equiv \widehat{\beta}_j / \operatorname{se}\left(\widehat{\beta}_j\right) \equiv \left(\widehat{\beta}_j - \beta_j\right) / \operatorname{se}\left(\widehat{\beta}_j\right) \sim t_{n-k-1}$$

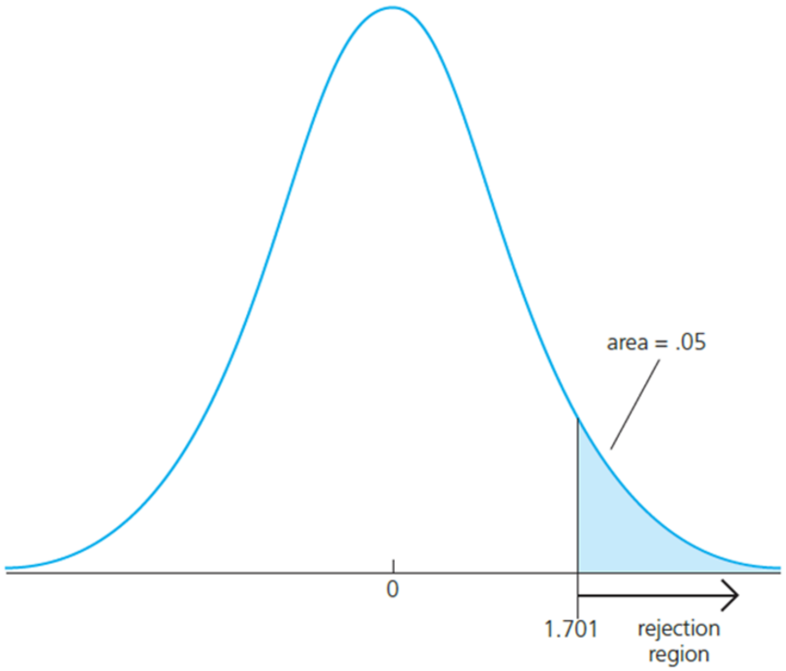
How should we choose a rejection rule?

- Decide on a **significance level**: the probability of rejecting H_0 when it is in fact true.
- Suppose we decide on 5% significance level.

```
- That means that we are mistakenly reject `\(H_0\)` when it is true, 5% of the time.
```

a) Testing against One- Sided Alternatives

- In the given example, the *critical value* of the t-distribution with 28 degrees of freedom at 5% level (one tail) is 1.701.



- Reject the H_0 in favor of the H_1, if the `t-statistic` is in the rejection region.
- **Rejection rule of H_0:** `t-statistic > critical value`

Example: wage equation

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

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

Standard errors

- Test whether, after controlling for education and tenure, higher work experience leads to higher hourly wages.
- One would either expect a positive effect of experience on hourly wage or no effect at all.

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

- Degrees of freedom= n-k-1=526-3-1
- t distribution approximates the standard normal distribution, so we use `Z-statistic` instead of `t-statistic`

Example: wage equation

Step 1. Find the critical value in the table

- Critical values (these are conventional significance levels):

10% significance level $c_{0.10} = 1.282$

5% significance level $c_{0.05} = 1.645$

1% significance level $c_{0.01} = 2.326$

Step 2. Calculate the `t-statistic` (or `z-statistic` for a large sample size)

- Use the coefficient and standard error from the regression

z-statistic: $t_{exper} = \frac{.0041}{.0017} \approx 2.41$

Example: wage equation

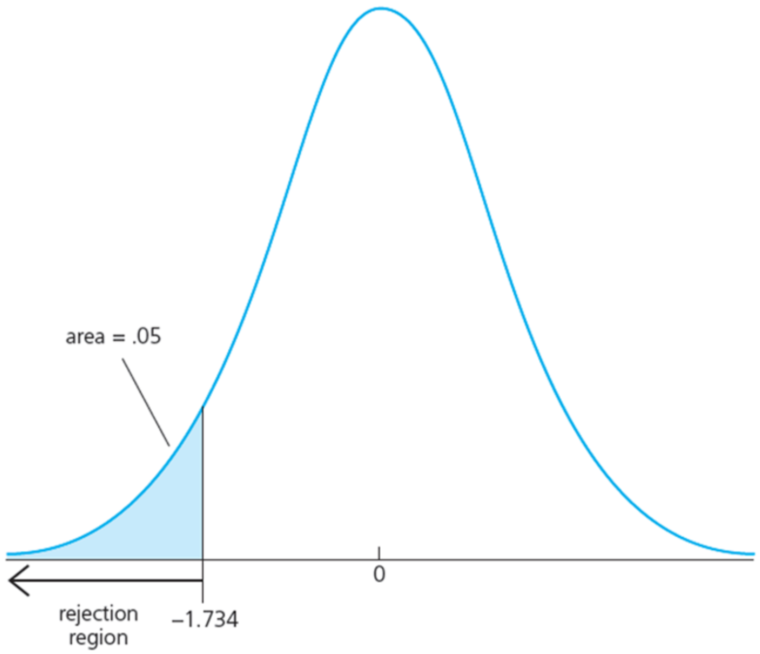
Step 3. Compare the `t-statistic` with the critical value, and decide if you reject or not the H_0

In this case: t-statistic (2.41) > critical value, so the H_0 is rejected

The effect of experience on hourly wage is statistically greater than zero at the 10% (and even at the, 5% and 1%) significance level.

Testing against one-sided alternatives (less than zero)

- In the given example, the critical value of the t-distribution with 18 degrees of freedom at 5% level is -1.734.



- Reject the H_0 in favor of the H_1, if the `t- statistic` is in the rejection region.
- The critical values in the table are reported as positive numbers. We compare the `t- statistic` with the NEGATIVE critical value.
- **Rejection rule of H_0:** `t- statistic < - critical value`

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

Student enrollment (= school size)

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

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

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

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

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

Step 1: Critical values for the 5% significance level : $-c_{0.05} = -1.65$

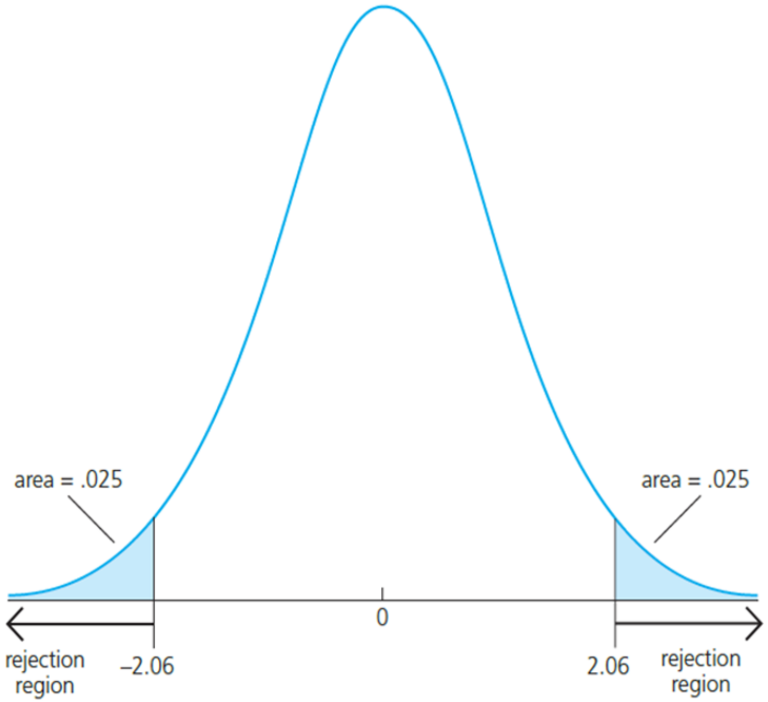
Step 2: z-statistic: $t_{enroll} = \frac{-.00020}{.00022} \approx -.91$

Step 3: -1.65> -.91

We fail to reject the null hypothesis because the value of the t-statistic -.91 is larger than the value of the critical value -1.65. This t-statistics lies outside of the rejection region.

b) Testing against Two-Sided Alternatives

- In the given example, the critical value of the t-distribution with 25 degrees of freedom is 2.06 .



- Reject the H_0 in favor of the H_1, if the `t statistic` is in the rejection regions.

Rejection rule of H_0: `|t-statistic| > critical value`

Example: Determinant of colllege GPA

Lectures missed per week

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

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

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

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

The effects of `hsGPA` and `skipped` are significantly different from zero at the 1% significance level.

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

The effect of `ACT` is not significantly different from zero at the 10% significance level.

“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:

- $|t \text{ ratio } | > 1.645 \longrightarrow$ "statistically significant at 10% level"
- $|t \text{ ratio } | > 1.96 \longrightarrow$ "statistically significant at 5% level"
- $|t \text{ ratio } | > 2.576 \longrightarrow$ "statistically significant at 1% level"

c) Testing other hypothesis about β_j

Null hypothesis

$H_0 : \beta_j = a_j \leftarrow$ Hypothesized value of the coefficient

t-statistic

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

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

Example: Campus crime and enrollment

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

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

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

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

$$t = \frac{1.27 - 1}{.11} \approx 2.45 > 1.985 = c_{0.05}$$

We reject the H_0 in favor of the H_1 at 5% significance level.

d) Computing P-values for t-tests

Rather than testing at different significance levels is more informative to answer the question:

- Given the value of the t- statistic, **what is the smallest significance level** at which the null hypothesis would be rejected? This is known as **P-value**
- or given that the null hypothesis is true, how unlikely is that we get a sample like the one we got?

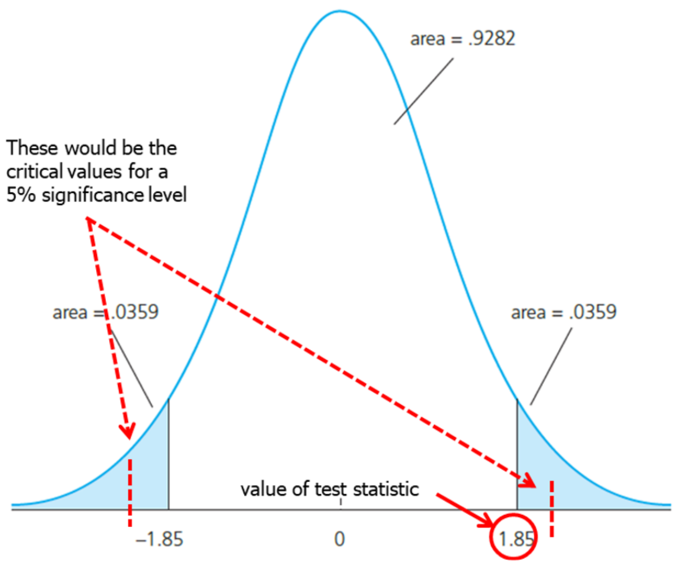
The P-value is the probability of observing a t-statistic as extreme as we did if the null hypothesis is true

"when your P-value is low, it says that: it's very unlikely that we would have got these results by chance, so they must be reflecting a real finding"

d) 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.

Computing P-values for t-tests



- 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:

```
P(|t|>1.85) = 2*(0.0359) = .0718
```

- **Reject the H_0 if P-value < the significance level.**
  - For example, the P value (P(|t|>1.85) = 2*(0.0359) = .0718) is larger than the significance level of 5% (0.05). Thus we fail to reject the null hypothesis.

e) 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 the sample size is small, effects might be imprecisely estimated so that the case for dropping insignificant variables is less strong.

4.3 Confidence Intervals

Under CLM assumption we can construct confidence intervals for the population parameter β_j

A manipulation of the Theorem 4.2

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

implies that

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

Interpretation of the confidence interval:

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

Confidence intervals for typical confidence levels

$$P\left(\widehat{\beta}_j - c_{0.01} \cdot \mathrm{se}\left(\widehat{\beta}_j\right) \leq \beta_j \leq \widehat{\beta}_j + c_{0.01} \cdot \mathrm{se}\left(\widehat{\beta}_j\right)\right) = 0.99$$
$$P\left(\widehat{\beta}_j - c_{0.05}\right) \cdot \mathrm{se}\left(\widehat{\beta}_j\right) \leq \beta_j \leq \widehat{\beta}_j + c_{0.05} \cdot \mathrm{se}\left(\widehat{\beta}_j\right)\Big) = 0.95$$
$$P\left(\widehat{\beta}_j - c_{0.10} \cdot \mathrm{se}\left(\widehat{\beta}_j\right) \leq \beta_j \leq \widehat{\beta}_j + c_{0.10} \cdot \mathrm{se}\left(\widehat{\beta}_j\right)\right) = 0.90$$

- 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 $\Rightarrow$ reject $H_0 : \beta_j = a_j$ in favor of $H_1 : \beta_j \neq a_j$

- Example: suppose $a_j = 0$, then if 0 is NOT in the confidence interval, we reject the null hypothesis H_0

Example: Model of firms' R&D expenditures

$$\underset{\text{Spending on R\&D}}{\widehat{\log}(rd)} = -\underset{(.47)}{4.38} + \underset{(.060)}{1.084}\ \underset{\text{Annual sales}}{\log(sales)} + \underset{(.0128)}{.0217}\ \underset{\text{Profits as percentage of sales}}{profmarg}$$

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

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

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.

$$.0217 \pm 2.045(.0128)$$
$$= (-.0045, .0479)$$

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

4.4 Testing hypotheses about a linear combination of the parameters

Example: Return to education at two-year vs. at four-year colleges

Years of education at two year colleges

Years of education at four year colleges

Months in the workforce

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

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

A possible test statistic would be:

$$t = \frac{\left(\hat{\beta}_1 - \hat{\beta}_2\right)}{\operatorname{se}\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)}$$

The difference between the estimates is normalized by the estimated standard deviation of the difference.

The standard error of the difference in parameters is impossible to calculate with standard regression output

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

$\operatorname{Cov}(\widehat{\beta_1}, \widehat{\beta_2})$ is usually not available in regression output

An alternative method is to make a substitution in variables.

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

We get $\beta_1 = \theta_1 + \beta_2$

$$\begin{aligned} \log(\text{ wage }) &= \beta_0 + (\theta_1 + \beta_2)\, jc + \beta_2 univ + \beta_3 \text{ exper } + u \\ &= \beta_0 + \theta_1 jc + \beta_2 (jc + univ) + \beta_3 \text{ exper } + u \end{aligned}$$

Estimation results

```
data(twoyear, package='wooldridge')

res<-feols(lwage~ jc+univ+exper, data = twoyear)

modelsummary(res,output = "markdown")
```

	Model 1
(Intercept)	1.472
	(0.021)
jc	0.067
	(0.007)
univ	0.077
	(0.002)
exper	0.005
	(0.000)
Num.Obs.	6763

Estimation results

```r
data(twoyear, package='wooldridge')

res<-feols(lwage~ jc+totcoll+exper, data = twoyear)

modelsummary(res,output = "markdown")
```

Total years of college

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

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

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

$$p-\text{value} = P(t-\text{ratio} < -1.48) = .070$$

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

Note: This method works always for single linear hypotheses

4.5 Testing multiple linear restrictions: The F-test

Multiple hypotheses: whether a group of variables has no effect on the dependent variables

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

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

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

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

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

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

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

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

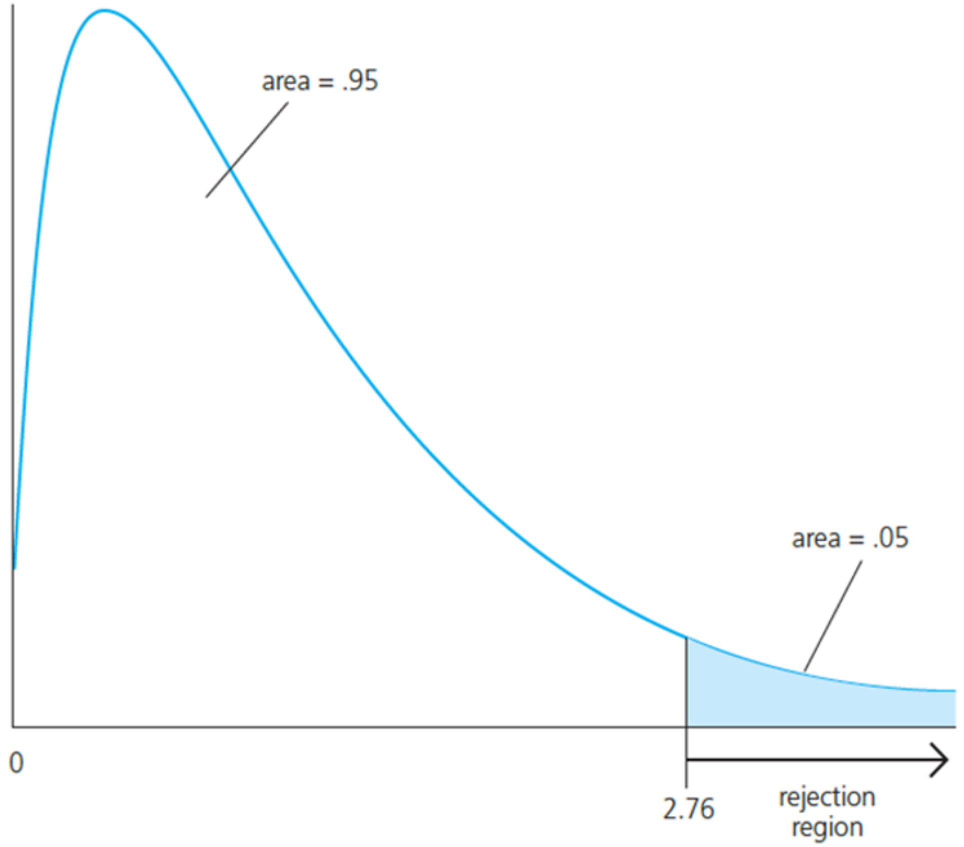
The sum of squared residuals necessarily increases, but is the increase statistically significant?

Test statistic:

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

q- number of restrictions F statistic measures the relative increase in SSR when moving from UNRESTICTED to RESTRICTED model.

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_0 to H_1.
- Choose the critical value so that we incorrectly reject the null hypothesis in, for example, only 5% of the cases.

Test statistic

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

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

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

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

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

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

$$P(F-\text{ statistic} > 9.55) = 0.000$$

Discussion:

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

```r
data(mlb1, package='wooldridge')

res.ur<- lm(log(salary)~ years+gamesyr+bavg+hrunsyr+rbisyr, data = mlb1) # Unrestricted OLS regress
res.r<- lm(log(salary)~ years+gamesyr, data = mlb1) # Restricted OLS regression

(r2.ur<-summary(res.ur)$r.squared) # R^2
```

```
## [1] 0.6278028
```

```r
r2.ur
```

```
## [1] 0.6278028
```

```r
(r2.r<-summary(res.r)$r.squared)
```

```
## [1] 0.5970716
```

```r
r2.r
```

```
## [1] 0.5970716
```

```r
(F<- (r2.ur-r2.r)/(1-r2.ur)*347/3) # F statistic
```

```
## [1] 9.550254
```

Test of overall significance of a regression

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

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \longleftarrow$ The null hypothesis states that the explanatory variables are not useful at all in explaining the dependent variable

$y = \beta_0 + u \leftarrow$ Restricted model (regression on constant)

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

Testing general linear restrictions with the F-test

Example: Test whether house price assessments are rational

Actual house price | The assessed housing value (before the house was sold) | Size of lot (in square 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

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

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

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

Testing general linear restrictions with the F-test

Unrestricted regression

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

Restricted regression

$$\log(\text{ price }) = \beta_0 + \log(\text{ assess }) + u$$

The restricted model is actually a regression of $log(price) - log(assess)$ on a constant

$$\log(\text{ price }) - \log(\text{ assess }) = \beta_0 + u$$

Test Statistic

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

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