Econometrics I
Introduction

Austin M. Mitchell

Hiroshima University

April 8, 2025

HIROSHIMA
UNIVERSITY

Introductions

Instructor:
Austin MITCHELL
Assistant Professor, IEDP and SmaSo

Teaching Assistant (TA):
Chengyi JIANG
PhD Student, IEDP

Today's course materials: https://www.austin-mitchell.com
Go to: Teaching-Econometrics

Overview

Course information

- Format of classes
- Topics covered in this course
- Homework and exams
- Grading
- Labs

Basics of econometrics

Course information

Format of classes

Classes involve 3 components:

- Lecture on the methodology for the day.
- Discussion of applied research papers.
- Lab to learn how to implement the methodology for the day.

The first half of class is devoted to lecture and discussion.
The second half of class is for the lab.

Topics covered in this course

Econometrics I covers methods for experiment and quasi-experimental data.

- Class 1: Basics of econometrics
- Class 2: Data and measurement
- Class 3: Research designs
- Class 4: Random control trials (RCT)
- Class 5: Conditionality
- Class 6: Intent to treat (ITT)
- Class 7: Random encouragement designs (RED)
- Class 8: Exam

Econometrics II (next term) covers methods for observational data.

Homework and exams

There are weekly homework assignments for this class.

The homework assignments involve applying the methodology we learn in the lectures and labs.

Homework assignments will be either individual or group-based. The Instructor will provide details for each homework assignment during the lab sessions.

The final exam will have an in-class exam component.

Grading

The distribution of grading:

Attendance	10%
Homeworks	40%
Exam	50%

Attendance will be collected twice daily. Once for lecture and once for lab.

If you must be absent for a class, you must inform the Instructor.

Labs

The labs will be conducted using Stata and R.

The Stata do files and R scripts for the labs will be provided before each class.

Students must bring their computers to class to participate in labs.

For homework assignments, students may choose whether to use Stata or R.

Stata is not provided for this course. Students without a Stata license can use R.

Basics of econometrics

Today's class

- Ordinal least squares (OLS)
- Coefficients
- Standard errors
- Hypothesis tests
- Test (t) statistics
- t tests
- Statistical significance
- Confidence intervals
- p-values

Ordinal least squares (OLS)

OLS fits a straight line through through the data. OLS minimizes the distance (error) between between each data point and the line.

A bivariate regression (with just one covariate) is simply:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

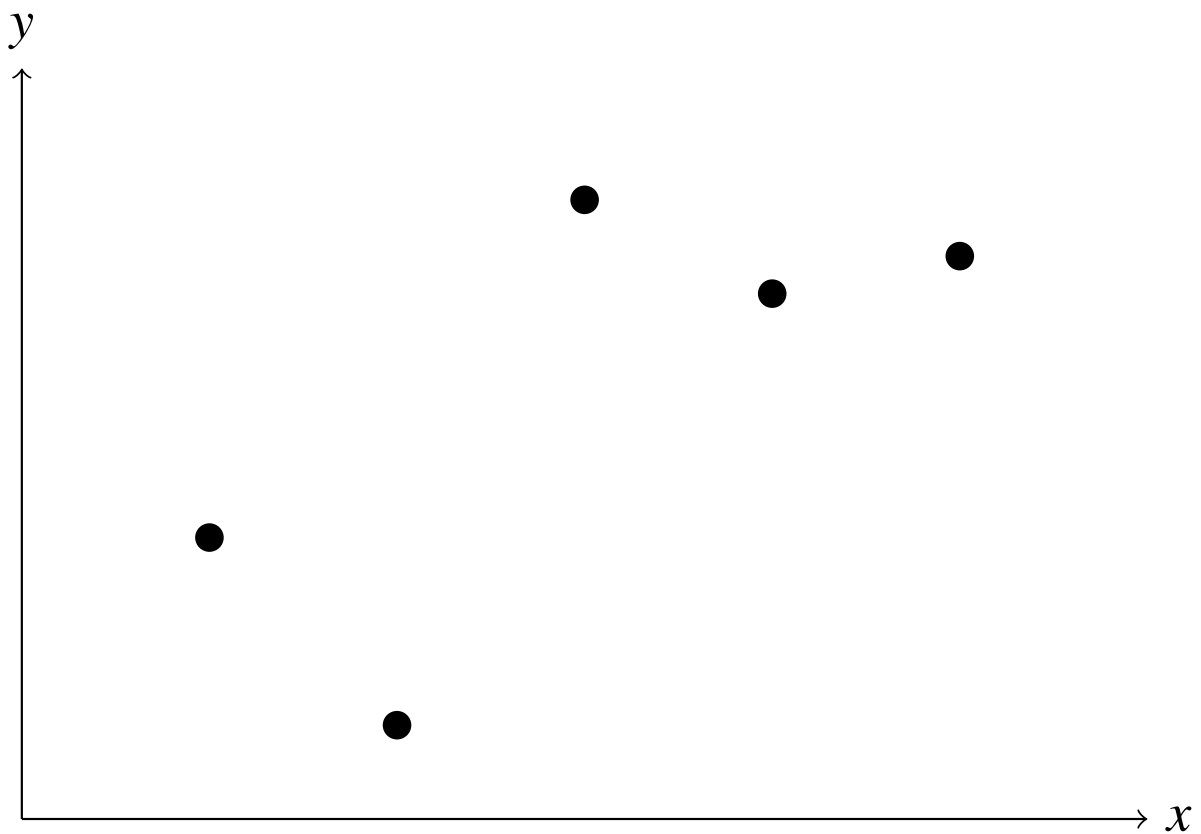
- β_0 is the constant, which is where the regression line crosses 0 on the y axis.
- β_1 is the coefficient on x. A one unit change in x leads to a β_1 unit change in y.
- ε is the error term. It represents the variation in the data that is not explained by the model. We assume ε is normally distributed and centered at 0.

You may have learned this in basic math courses as:

$$y = mx + b \tag{2}$$

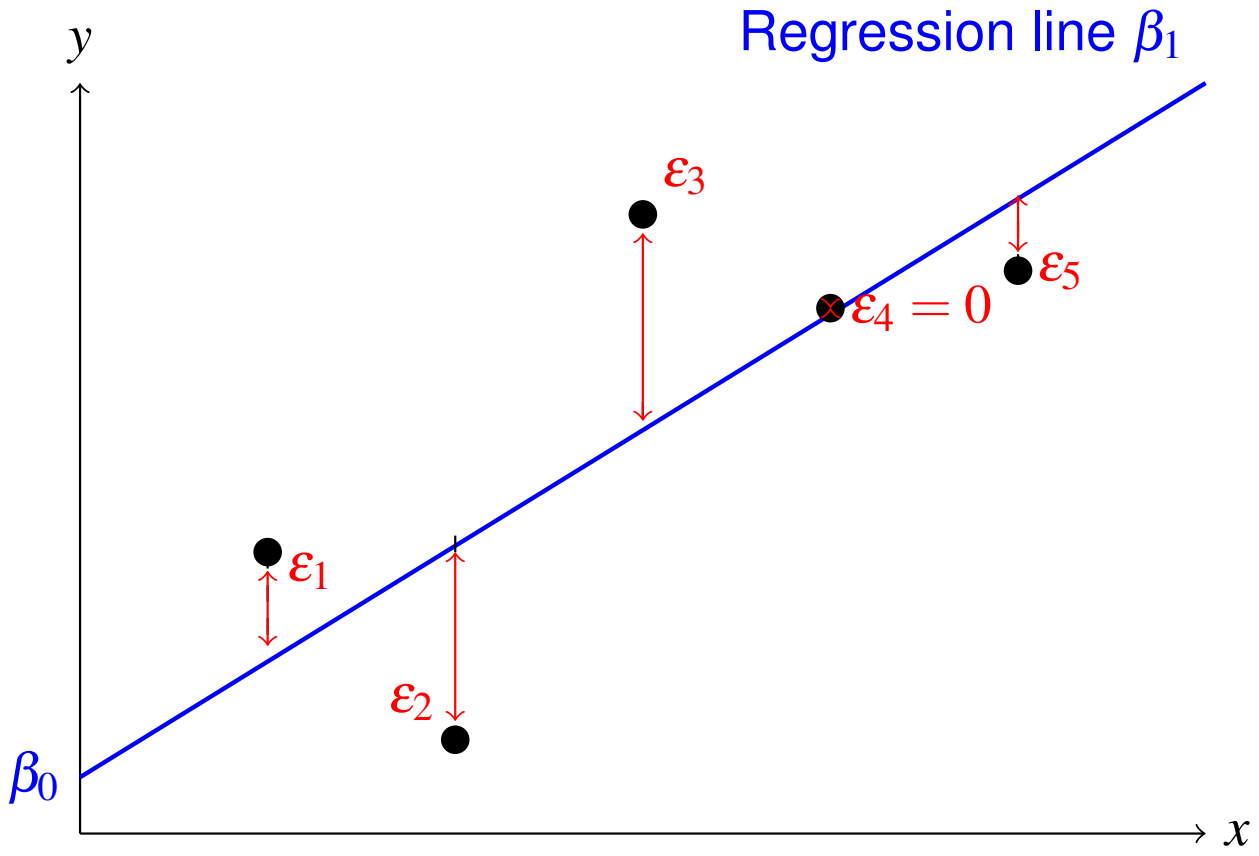
Example data for y and x

x is the treatment and y is the outcome.



Example data with a regression line

Regression model: $y = \beta_0 + \beta_1 x + \varepsilon$



OLS minimizes the errors between each data point and the line.
The sum of the errors for OLS is always 0.

Multivariate OLS (multiple regression)

A regression with just two covariates:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{3}$$

When there are two or more covariates, the interpretation of the β_1 and β_2 coefficients changes.

The β_1 is the partial effect of x_1 on y, *holding* x_2 *constant*.

The interpretation of β_2 is similar.

The *holding* x_2 *constant* part means that the regression accounts for the covariation between x_1 and x_2 when calculating β_1 and β_2.

This is important when we consider confounding.

Coefficients

Coefficients in OLS are the βs.

In the regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{4}$$

β_0, β_1, and β_2 are each coefficients.

Example regression output: coefficients

Outcome: Probability of coup attempt	
Decentralization	-0.153
	(0.059)
GDP per capita	-0.014
	(0.009)
Constant	0.164
	(0.055)
Observations	7,174
Country	164

There are two covariates in the table.
Decentralization and GDP per capita.

There are three coefficients in the table.
$\beta_1 =$ -0.153
$\beta_2 =$ -0.014
$\beta_0 =$ 0.164

A one unit increase in decentralization is associated with a -15 percentage point decrease in the probability of a coup attempt.

Standard errors (SE)

Standard errors are a measure of uncertainty about the coefficients.

More accurate regressions have smaller standard errors. When the errors between the data and the regression line decrease, the standard error decreases.

Larger samples also decrease standard errors by reducing the uncertainty of the coefficients.

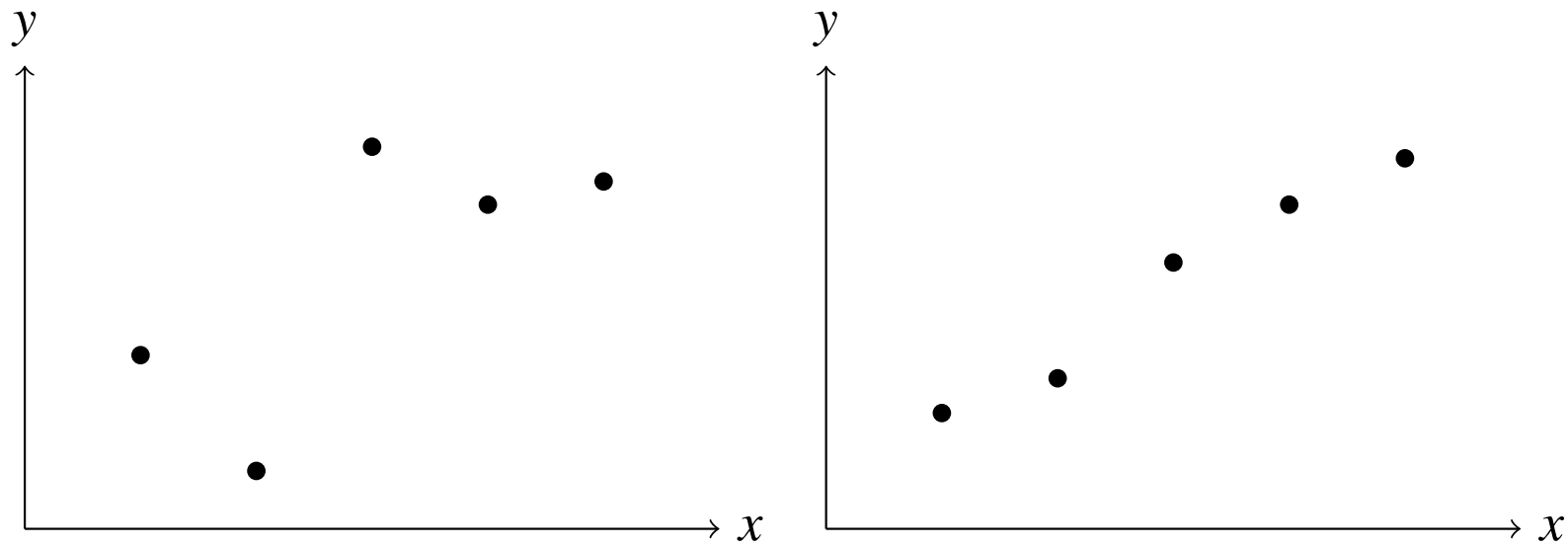
SE equation for a coefficient

$$\text{SE}(\beta_1) = \frac{\sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- i is an observation identifier
- $\hat{y}$ is predicted y values
- $\bar{x}$ is the mean of x
- n is the number of observations
- k is the number of parameters

Note that this is the bivariate case. Multivariate SEs are more complicated.

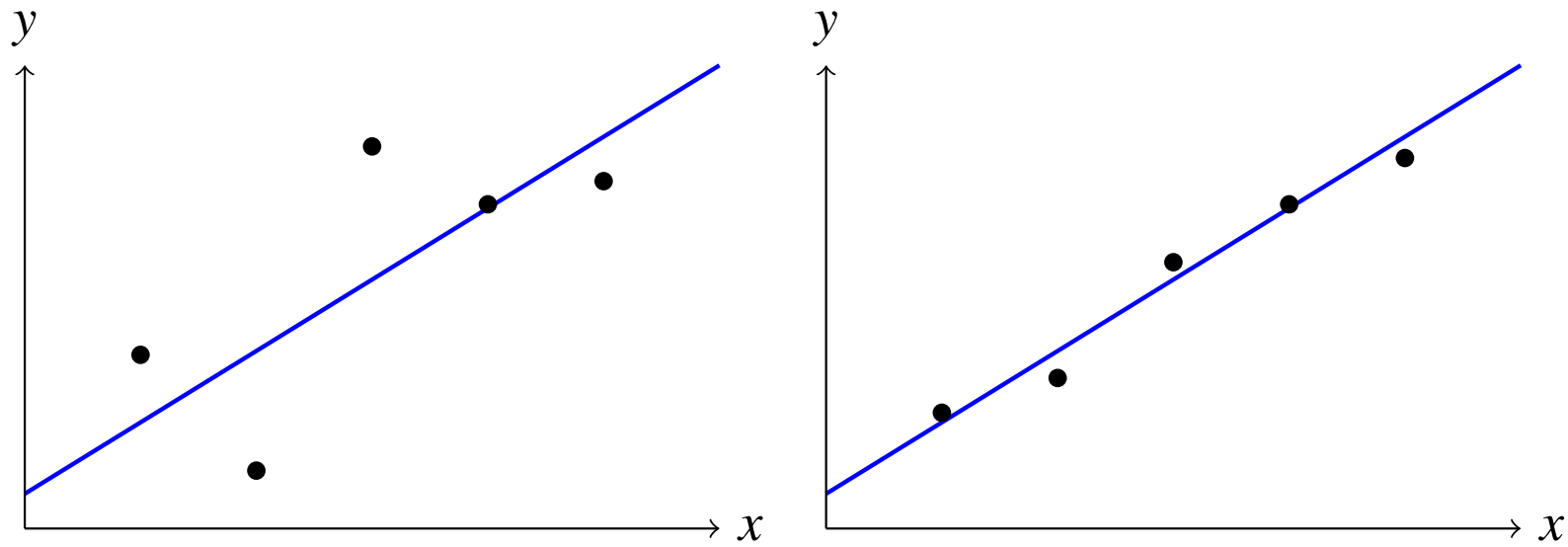
Data variance



Which of these data plots will have lower error variance if we fit a regression line?

Data variance with regression lines

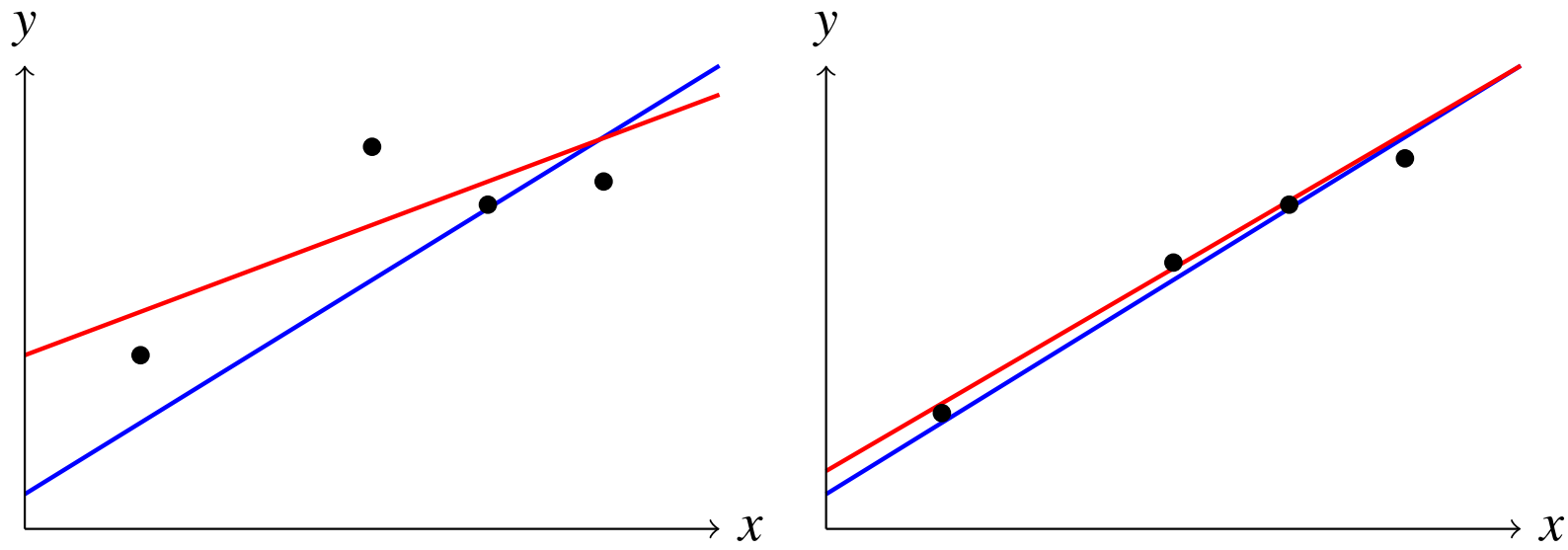
The regression lines are the same but the data on the right have lower error variance.



The standard errors for the regression on the right will be smaller (higher statistical significance).

Errors and regression line uncertainty

Imagine we remove one data point. How does that change the regression line?



When we remove one data point, the regression line changes more dramatically when the data have high variance.

This is why larger errors and smaller datasets reduce statistical significance.

Example regression output: standard errors

Outcome: Probability of coup attempt	
Decentralization	-0.153
	(0.059)
GDP per capita	-0.014
	(0.009)
Constant	0.164
	(0.055)
Observations	7,174
Country	164

Standard errors (SE) are typically found in parentheses.

There are three standard errors in the table.
β_1 $SE = 0.059$
β_2 $SE = 0.009$
β_0 $SE = 0.055$

Hypothesis testing

In econometrics, our goal is to test hypotheses.

We construct 1) a null hypothesis and 2) an alternative hypothesis.

The null hypothesis always represents no relationship between the treatment and outcome.

$$\beta = 0$$

The alternative hypothesis is set by the researcher. Typically we consider either a positive or negative hypothesis:

$$\beta > 0$$

or

$$\beta < 0$$

Note about hypothesis testing

A full understanding of hypothesis testing requires knowledge of sampling distributions, degrees of freedom, asymptotics, and the Central Limit Theorem. We omit the underlying statistical theory and begin with *t* statistics.

Additionally, in this class we will not focus on hypothesis testing by *t* statistics and *t* tests. Instead we utilize *p*-values, which are derived from *t* tests.

For equations and more technical details about hypothesis testing, see Wooldridge. *Introductory Econometrics: A Modern Approach* (Any edition).

Test statistics (*t*-statistics)

We conduct hypothesis tests of our coefficients according to their t statistics.

The test statistic for β is simply:

$$t = \frac{\beta}{SE_\beta} \tag{5}$$

SE is the standard error.

t tests

We test whether or not to reject the null hypothesis ($\beta = 0$) according to whether the t statistic is less than or greater than some critical value.

The value of the critical value is determined by multiple factors

1. a **significance level** that we define (i.e 95%)
2. the value of β
3. the standard error SE_β
4. the number of observations
5. the number of parameters in the model

The result of the *t* test determines the result of the hypothesis test. We reject the null hypothesis if the coefficient is **statistically significant** (t>critical value).

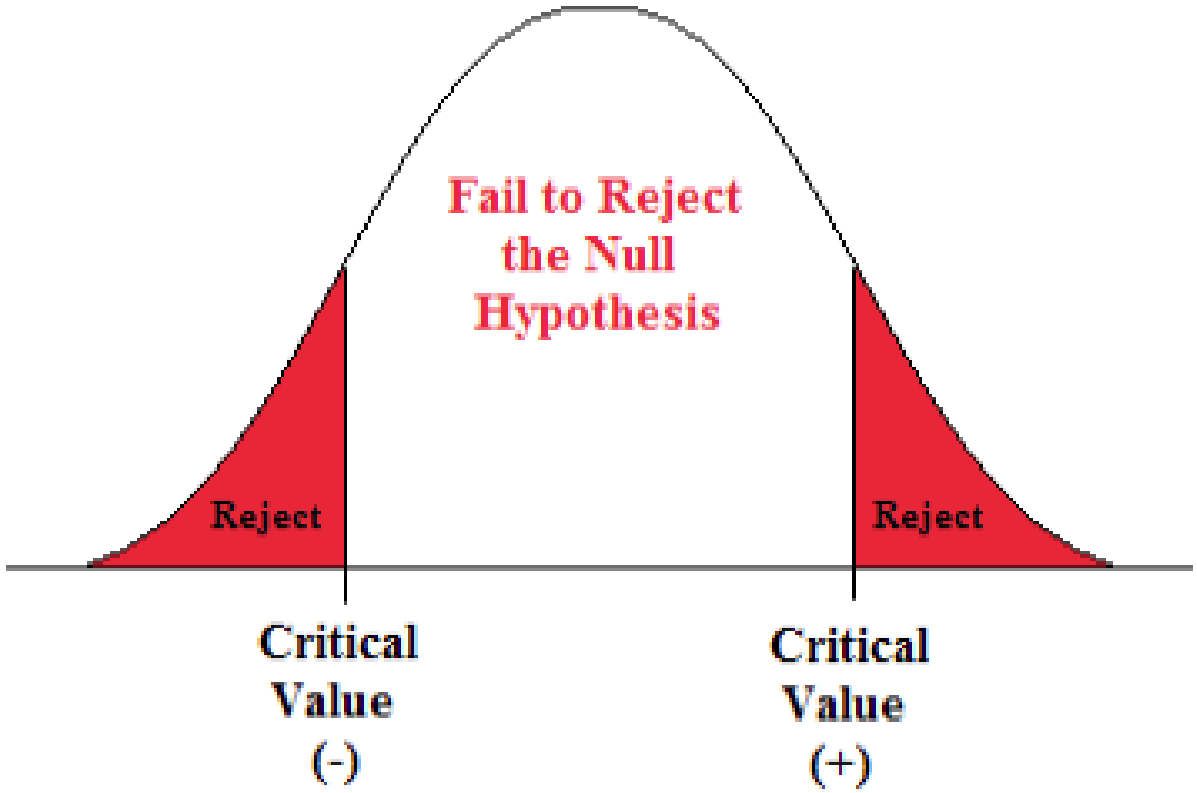
Statistical significance

When we calculate our coefficients, we test the null versus alternative hypotheses according to the statistical significance of the coefficient.

Statistically significant means (informally) that the coefficient is statistically different from zero.

A more technical definition is: it would be improbable to obtain the estimated β coefficient if the null hypotheses were actually true ($\beta = 0$).

Intervals for null hypothesis tests (two-tailed)



The figure represents the sampling distribution for β_1. The distribution centers at 0.

If β_1 is large (or small) enough, then we reject the null hypothesis. In that case we say β_1 is statistically significantly different from 0 (null).

Statistical significance and confidence levels

Statistical significance is judged according to confidence levels. A confidence level is a standard for what is, and is not, considered statistically significant.

Social sciences tend to use 90%, 95%, and 99% confidence levels. In IEDP we use 95% as a standard.

Confidence levels are are expressed as critical values, which are just 1 - confidence level. For instance, the 90%, 95%, and 99% confidence levels have critical values 0.1, 0.05, and 0.01, respectively.

Natural sciences tend to use 99%, 99.9%, and 99.99% because they have larger datasets and study more easily predictable subjects.

Example regression output: standard errors with statistical significance stars

Stars indicate statistical significance *at specific confidence levels.*

Outcome: Probability of coup attempt	
Decentralization	-0.153**
	(0.059)
GDP per capita	-0.014
	(0.009)
Constant	0.164***
	(0.055)
Observations	7,174
Country	164

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The coefficient on decentralization is statistically significant at the 95% level. This means it is also statistically significant at the 90% level but not the 99% level.

Example regression output: standard errors with statistical significance stars

At IEDP and SmaSo, we only use 95% confidence levels.

Outcome: Probability of coup attempt	
Decentralization	-0.153*
	(0.059)
GDP per capita	-0.014
	(0.009)
Constant	0.164*
	(0.055)
Observations	7,174
Country	164

Standard errors are in parentheses. * $p < 0.05$

Confidence intervals (CI)

We can judge statistical significance according to the size of the coefficient compared to the standard error.

At the 95% confidence level, the coefficient must be 1.96 times larger than its standard error (according to statistical theory). However, this is a bit cumbersome to calculate.

We can also use confidence intervals to assess statistical significance.

A confidence interval is the upper and lower estimates of a coefficient according to a specific confidence level.

Standard errors determine confidence intervals. As the standard error increases, the confidence interval increases.

Refer to Wooldridge or other econometrics texts for the equation to calculate CIs from SEs.

Confidence intervals (CI) and statistical significance

Confidence intervals can be used to determine statistical significance.

If a confidence interval overlaps 0, then the coefficient is not statistically significant.

If a confidence interval does not overlap 0, the coefficient is statistically significant.

Confidence intervals are set according to the confidence level (such as 95%).

Example regression output: 95% confidence intervals

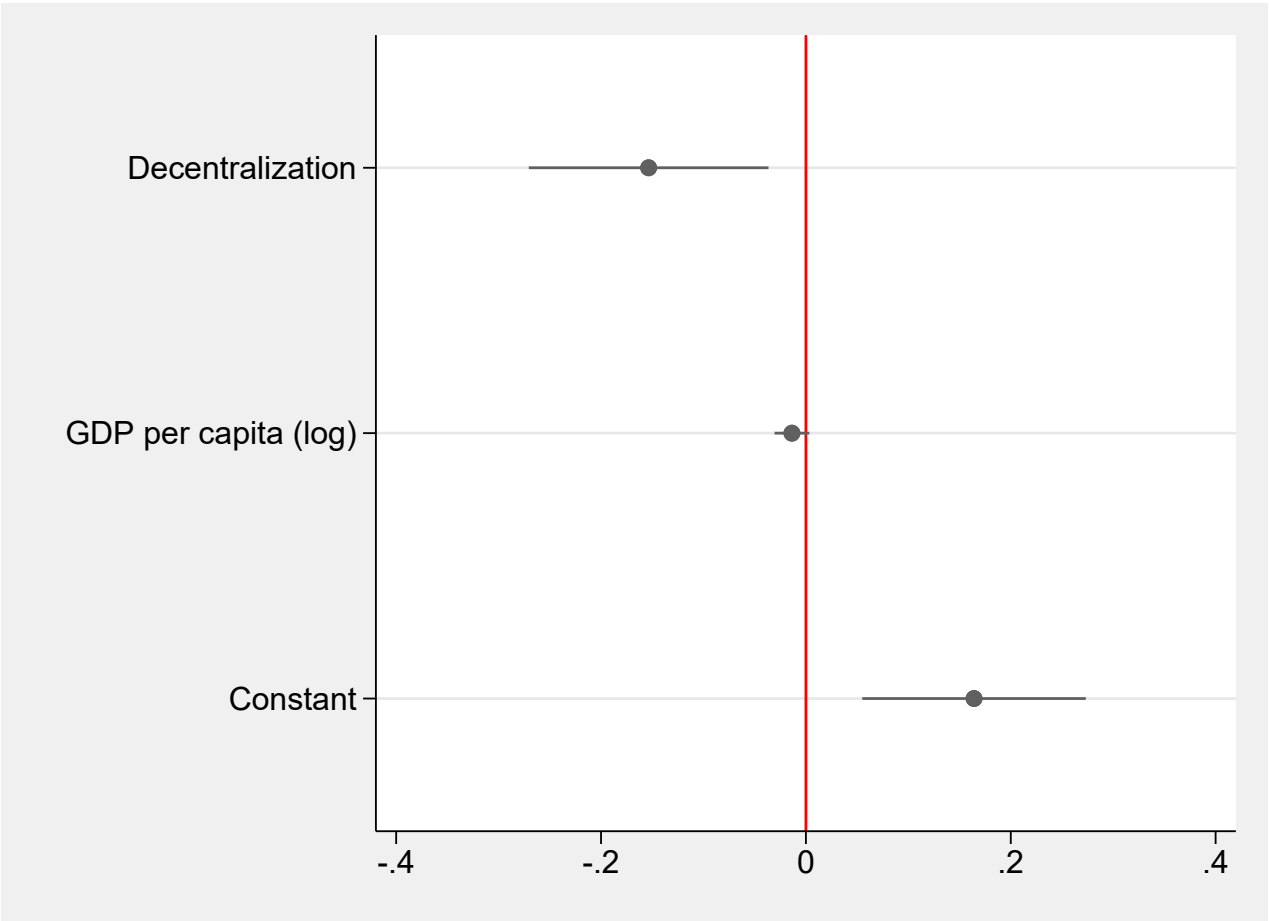
Outcome: Probability of coup attempt	
Decentralization	-0.153*
	[-0.271,-0.036]
GDP per capital (log)	-0.014
	[-0.031,0.003]
Constant	0.164*
	[0.055,0.273]
Observations	7,174
Country	164

Confidence intervals are in brackets. * $p < 0.05$

The 95% CIs for Decentralization do not overlap 0, which means the coefficient *is statistically significant at the 95% level*.

The CIs for GDP per capita do overlap 0, which means the coefficient is not statistically significant.

Example regression output: plots of coefficients and 95% CIs



Example regression output: 99% confidence intervals

If we change the confidence level, the confidence intervals change. Note that we have included stars for both 95% and 99% statistical significance.

Outcome: Probability of coup attempt	
Decentralization	-0.153**
	[-0.308,0.001]
GDP per capital (log)	-0.014
	[-0.036,0.009]
Constant	0.164***
	[0.020,0.308]
Observations	7,174
Country	164

Confidence intervals are in brackets. ** $p < 0.05$, *** $p < 0.01$

The 99% CIs for Decentralization do overlap 0, which means the coefficient is *not* statistically significant at the 99% level.

p values

p values are the easiest way to determine whether a statistic (coefficient) is statistically significant.

The *p* value is a probability, but its exact meaning is difficult to understand. It is the probability that a statistic is at least as extreme as the observed value *if the null hypothesis were true*.

In practical applications, we will simply use it to determine statistical significance at the 95% level.

p values and statistical significance

p values range from 0-1 since it they are probabilities.

p values relate to confidence levels. A 95% confidence level has a critical level for the *p* value of 0.05, which is $1 - 0.95 = 0.05$.

If a *p* value is *smaller* than the critical level of 0.05, then the statistic is statistically significant.

If *p* value is *greater* than 0.05, then we do not reject the null hypothesis.

Example regression output: *p* values with statistical significance stars

The table below includes 90%, 95%, and 99% thresholds.

At IEDP and SmaSo, we only use 95% confidence levels.

Outcome: Probability of coup attempt	
Decentralization	-0.153**
	(0.010)
GDP per capital (log)	-0.014
	(0.114)
Constant	0.164***
	(0.003)
Observations	7,174
Country	164

p values are in parentheses. ** $p < 0.05$, *** $p < 0.01$

Terms

- Ordinary least squares
- Regression
- Coefficient
- Constant
- Error
- Standard error
- *t* statistic
- *t* test
- Confidence level
- Critical value
- Statistical significance
- Confidence interval
- *p* value