Introductory Econometrics

Lecture 4: Multiple Regression

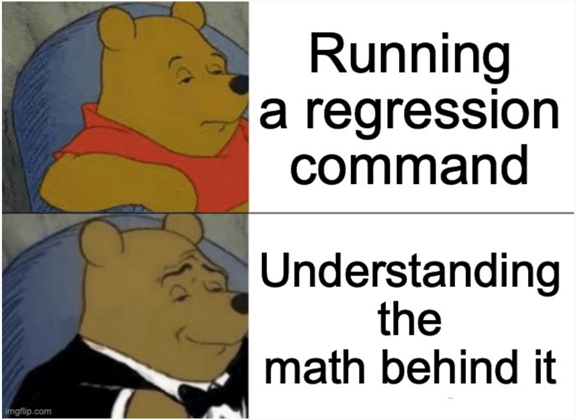
Alex Alekseev

November 7, 2022

University of Regensburg, Department of Economics

Previously on *Introductory Econometrics*...

- Algebraic properties of OLS
- Goodness-of-fit
- Statistical properties of OLS


Running
a regression
command
Understanding
the
math behind it
imgflip.com

Introduction to multiple regression

Motivation

- So far we have only considered a simple regression with just one predictor
- More commonly, we are working with more than one predictor
- There are several reasons for including multiple predictors

Motivation

- So far we have only considered a simple regression with just one predictor
- More commonly, we are working with more than one predictor
- There are several reasons for including multiple predictors
  - To improve a model's explanatory or predictive power

Motivation

- So far we have only considered a simple regression with just one predictor
- More commonly, we are working with more than one predictor
- There are several reasons for including multiple predictors
  - To improve a model's explanatory or predictive power

Explanation

More predictors typically means we can improve a model's fit, which might be what we want. We might not care, however, about theoretical considerations for why a given variable should be included in a model. This is a so-called **kitchen sink** approach.

Motivation

- So far we have only considered a simple regression with just one predictor
- More commonly, we are working with more than one predictor
- There are several reasons for including multiple predictors
  - To improve a model's explanatory or predictive power
  - Because our theoretical model says that there are more than one relevant factors affecting our outcome of interest

Motivation

- So far we have only considered a simple regression with just one predictor
- More commonly, we are working with more than one predictor
- There are several reasons for including multiple predictors
  - To improve a model's explanatory or predictive power
  - Because our theoretical model says that there are more than one relevant factors affecting our outcome of interest

Explanation

In this case we might not want to include **all** available predictors, but just a few that are motivated by the theory.

Motivation

- So far we have only considered a simple regression with just one predictor
- More commonly, we are working with more than one predictor
- There are several reasons for including multiple predictors
  - To improve a model's explanatory or predictive power
  - Because our theoretical model says that there are more than one relevant factors affecting our outcome of interest
  - To be able to make causal claims about the effect of a variable of interest on an outcome

Motivation

- So far we have only considered a simple regression with just one predictor
- More commonly, we are working with more than one predictor
- There are several reasons for including multiple predictors
  - To improve a model's explanatory or predictive power
  - Because our theoretical model says that there are more than one relevant factors affecting our outcome of interest
  - To be able to make causal claims about the effect of a variable of interest on an outcome

Explanation

Often the effect of a variable is **confounded** by other factors, which we need to control for. We would then typically call that variable of interest a **treatment** variable, and other predictors **controls**. This is the modern applied approach to regression.

Trade example

- Let's use the trade example to illustrate the second reason
- Typical trade models include not just the GDP but also the distance between countries as a relevant factor that affects trade
- According to the model, the trade should increase with the size of economies and decrease with distance
- These are the so-called **gravity models of trade** by an analogy with the gravity equation in physics
- Our new (population) trade model then would look like this

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + \beta_2 \ln(distance_i) + u_i$$

Regression output

Names of predictors

	simple	gravity
(Intercept)	−5.785	4.670
	(2.199)	(2.181)
log(gdp)	1.078	0.976
	(0.088)	(0.064)
log(distance)		−1.075
		(0.157)
Num.Obs.	48	48
RMSE	1.29	0.90
R2	0.767	0.886

Various model statistics

Different models

Coefficient estimates

Estimates of standard errors

Regression output

Names of predictors

	simple	gravity
(Intercept)	−5.785	4.670
	(2.199)	(2.181)
log(gdp)	1.078	0.976
	(0.088)	(0.064)
log(distance)		−1.075
		(0.157)
Num.Obs.	48	48
RMSE	1.29	0.90
R2	0.767	0.886

Various model statistics

Different models

Coefficient estimates

Estimates of standard errors

- The coefficient on GDP slightly decreased, although still positive
- The coefficient on distance is negative
- These signs are in line with our model predictions
- Including distance as a predictor improved the model's fit

Income and education

- Let's say we are interested in the effect of **education** on a person's **income**
- Our first naive attempt to answer this question could be to estimate a simple regression of income on education

$$income_i = \beta_0 + \beta_1 education_i + u_i$$

- However, we might worry that there are some factors that affect both income and education, and thus confound the effect of education
- One of such factors is **ability**
- Our second regression will include ability as an additional predictor

$$income_i = \beta_0 + \beta_1 education_i + \beta_2 ability_i + u_i$$

Regression output

	naive	**controls**
(Intercept)	54.947	49.976
	(0.321)	(0.320)
education	16.675	9.918
	(0.370)	(0.382)
ability		20.007
		(0.565)
Num.Obs.	5000	5000
RMSE	11.27	10.07
R2	0.289	0.431

- Including ability as a control variable significantly reduces the effect of education on income
- It also improves the model's fit

Population regression model and assumptions

Population regression model

- The population multiple regression model is a simple extension of the simple regression model for multiple predictors
- Recall our population regression model from before

$$Y = \mathbb{E}[Y \mid X] + U,$$

where we define $U \equiv Y - \mathbb{E}[Y \mid X]$ to be the error term

- Now instead of a single predictor X, let's consider the conditional expectation of Y given predictors $X_1, X_2, \ldots, X_k$
- The population model retains its structure

$$Y = \mathbb{E}[Y \mid X_1, X_2, \ldots, X_k] + U.$$

Assumptions

Assumption: Linear CEF

The CEF of Y given $X_1, X_2, \ldots, X_k$ is linear:

$$\mathbb{E}[Y \mid X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Assumptions

Assumption: Linear CEF

The CEF of Y given $X_1, X_2, \ldots, X_k$ is linear:

$$\mathbb{E}[Y \mid X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Explanation

This assumption allows us to write the population model as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U,$$

where U is mean-independent of $X_1, X_2, \ldots, X_k$:
$\mathbb{E}[U \mid X_1, X_2, \ldots, X_k] = 0$.

Assumptions

Assumption: Linear CEF

The CEF of Y given $X_1, X_2, \dots, X_k$ is linear:

$$\mathbb{E}[Y \mid X_1, X_2, \dots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Note

The Linear CEF assumption is equivalent to a set of two assumptions: that the population model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U,$$

and that $\mathbb{E}[U \mid X_1, X_2, \dots, X_k] = 0$.

Assumptions

Assumption: Linear CEF

The CEF of Y given $X_1, X_2, \ldots, X_k$ is linear:

$$\mathbb{E}[Y \mid X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Assumption: Random Sampling

Our sample $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)_{i=1}^n$ is a random sample from the population, i.e., the observations are pairwise independent and identically distributed

Assumptions

Assumption: Linear CEF

The CEF of Y given $X_1, X_2, \ldots, X_k$ is linear:

$$\mathbb{E}[Y \mid X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Assumption: Random Sampling

Our sample $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)_{i=1}^n$ is a random sample from the population, i.e., the observations are pairwise independent and identically distributed

Assumption: No Perfect Collinearity

None of the predictors is a linear combination of other predictors

Assumptions

Assumption: Linear CEF

The CEF of Y given $X_1, X_2, \ldots, X_k$ is linear:

$$\mathbb{E}[Y \mid X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Assumption: Random Sampling

Our sample $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)_{i=1}^n$ is a random sample from the population, i.e., the observations are pairwise independent and identically distributed

Assumption: No Perfect Collinearity

None of the predictors is a linear combination of other predictors

Explanation

The exact meaning of this assumption will be explained later, but the intuition is that it simply generalizes the previous assumption to a case of more than one predictor.

Interpretation of coefficients

- In a simple regression model, we had only one slope coefficient whose interpretation was that it is the change in the expected outcome following a unit change in the predictor
- What is the interpretation of each beta coefficient in the multiple regression?
- A given coefficient β_j is by how much the expected outcome increases following a unit change in X_j while keeping the **rest of the variables constant**
- Notice the last part: It is equivalent of the **ceteris paribus** condition that we often use in economic models

Formally

- Suppose we increase X_j by one unit

$$\mathbb{E}[Y \mid X_1, \ldots, X_j{+}1, \ldots, X_k] = \beta_0{+}\beta_1 X_1{+}\ldots{+}\beta_j(X_j{+}1){+}\ldots{+}\beta_k X_k$$

- Then

$$\beta_j = \mathbb{E}[Y \mid X_1, \ldots, X_j + 1, \ldots, X_k] - \mathbb{E}[Y \mid X_1, \ldots, X_k]$$

- Since our CEF is linear, each β_j is also a partial derivative of the CEF with respect to X_j

$$\beta_j = \frac{\partial \mathbb{E}[Y \mid X_1, X_2, \ldots, X_k]}{\partial X_j}$$

Alternative notation

Summation

- When we consider a multiple regression, writing each individual predictor out can get cumbersome
- One alternative could be to use the **summation** notation:

$$Y = \sum_{j=0}^{k} \beta_j X_j + U$$

Vectors and matrices

- Another option is to use the tools of **linear algebra**: vectors and matrices
- In fact, even our simple linear regression could benefit from it
- Before, we wrote the population model of a simple regression as

$$Y = \beta_0 + \beta_1 X + U$$

- Let's use a simple trick and rewrite it as

$$Y = \beta_0 X_0 + \beta_1 X_1 + U,$$

where $X_0 = 1$ is just a constant (in fact, the summation notation above already uses this trick)

Stacking

- Now instead of writing the sum on the right explicitly, we can **stack** the coefficients and predictors into vectors:

$$\mathbf{x} = \begin{pmatrix} X_0 \\ X_1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

- and use the rules of matrix multiplication to write

$$Y = \mathbf{x}'\beta + U.$$

Linear algebra: Multiplying two vectors

If you have two vectors

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix},$$

then

$$\mathbf{a}'\mathbf{b} = (a_1, \dots, a_n) \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = a_1 b_1 + \ldots + a_n b_n = \sum_{i=1}^{n} a_i b_i$$

Sample regression model

Sample regression model

- As before, our sample linear regression model takes the form

$$y_i = \hat{\beta}_0 x_{i0} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i, \quad i = 1, \ldots, n,$$

  where we use the convention that $x_{i0} = 1$ for all i

- The first part of the expression on the right is called the **fitted values**, $\hat{y}_i$

$$\hat{y}_i = \hat{\beta}_0 x_{i0} + \ldots + \hat{\beta}_k x_{ik}$$

- and the second part, $\hat{u}_i$ are the **residuals**:

$$\hat{u}_i = y_i - \hat{y}_i$$

OLS derivation

Definition

- We define the OLS estimator in exactly the same way as for the simple regression
- The estimator minimizes the sum of squared residuals

$$(\hat{\beta}_0, \ldots, \hat{\beta}_k) = \underset{\tilde{\beta}_0, \ldots, \tilde{\beta}_k}{\arg\min} \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - (\tilde{\beta}_0 x_{i0} + \ldots + \tilde{\beta}_k x_{ik}))^2.$$

- Denoting the objective function as $g(\tilde{\beta}_0, \ldots, \tilde{\beta}_k)$, we can write the FONC

$$\frac{\partial g}{\partial \tilde{\beta}_j} = 0 \text{ at } \tilde{\beta}_j = \hat{\beta}_j, \quad j = 0, \ldots, k$$

or

$$\sum_{i=1}^{n} x_{ij}(y_i - (\hat{\beta}_0 x_{i0} + \ldots + \hat{\beta}_k x_{ik})) = 0, \quad j = 0, \ldots, k.$$

- Unlike in the simple regression case, however, solving this system of k equations using similar methods becomes unwieldy.

Re-write in a matrix form

- This is where the tools of **linear algebra** can help us
- We can re-write the sample regression model by stacking all the observations using **matrices** and **vectors**

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{u}},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{10} \dots x_{1k} \\ \vdots \ddots \vdots \\ x_{n0} \dots x_{nk} \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}, \quad \hat{\mathbf{u}} = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n. \end{pmatrix}$$

Re-write in a matrix form

- This is where the tools of **linear algebra** can help us
- We can re-write the sample regression model by stacking all the observations using **matrices** and **vectors**

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{u}},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{10} \dots x_{1k} \\ \vdots \ddots \vdots \\ x_{n0} \dots x_{nk} \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}, \quad \hat{\mathbf{u}} = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n. \end{pmatrix}$$

Explanation

You can think of vector **y** as stacking (vertically) all of the observations of the outcome

Re-write in a matrix form

- This is where the tools of **linear algebra** can help us
- We can re-write the sample regression model by stacking all the observations using **matrices** and **vectors**

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{u}},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{10} \dots x_{1k} \\ \vdots \ddots \vdots \\ x_{n0} \dots x_{nk} \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}, \quad \hat{\mathbf{u}} = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n. \end{pmatrix}$$

Explanation

Matrix $\mathbf{X}$ is constructed by first stacking (vertically) all of the observations of predictor 0, then stacking (vertically) all of the observations of predictor 1, and so on, and then stacking all of those predictors horizontally

Re-write in a matrix form

- This is where the tools of **linear algebra** can help us
- We can re-write the sample regression model by stacking all the observations using **matrices** and **vectors**

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{u}},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{10} \dots x_{1k} \\ \vdots \ddots \vdots \\ x_{n0} \dots x_{nk} \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}, \quad \hat{\mathbf{u}} = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n. \end{pmatrix}$$

Explanation

Vector $\hat{\beta}$ stacks (vertically) all of the beta coefficients. And finally, vector $\hat{\mathbf{u}}$ stacks (vertically) all of the residuals.

Linear algebra: Matrix by vector multiplication

- Take the first observation on all of the predictors and multiple it by the vector of betas.

$$(x_{10} \ldots x_{1k}) \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \hat{\beta}_0 x_{10} + \ldots + \hat{\beta}_k x_{1k}$$

Linear algebra: Matrix by vector multiplication

- Take the first observation on all of the predictors and multiple it by the vector of betas.

$$(x_{10} \dots x_{1k}) \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \hat{\beta}_0 x_{10} + \dots + \hat{\beta}_k x_{1k}$$

- Adding a second observation leads to

$$\begin{pmatrix} x_{10} \dots x_{1k} \\ x_{20} \dots x_{2k} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 x_{10} + \dots + \hat{\beta}_k x_{1k} \\ \hat{\beta}_0 x_{20} + \dots + \hat{\beta}_k x_{2k} \end{pmatrix}$$

Linear algebra: Matrix by vector multiplication

- Take the first observation on all of the predictors and multiple it by the vector of betas.

$$(x_{10} \dots x_{1k}) \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \hat{\beta}_0 x_{10} + \dots + \hat{\beta}_k x_{1k}$$

- Adding a second observation leads to

$$\begin{pmatrix} x_{10} \dots x_{1k} \\ x_{20} \dots x_{2k} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 x_{10} + \dots + \hat{\beta}_k x_{1k} \\ \hat{\beta}_0 x_{20} + \dots + \hat{\beta}_k x_{2k} \end{pmatrix}$$

- And so on

$$\underbrace{\begin{pmatrix} x_{10} \dots x_{1k} \\ x_{20} \dots x_{2k} \\ \vdots \ddots \vdots \\ x_{n0} \dots x_{nk} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}}_{\hat{\beta}} = \underbrace{\begin{pmatrix} \hat{\beta}_0 x_{10} + \dots + \hat{\beta}_k x_{1k} \\ \hat{\beta}_0 x_{20} + \dots + \hat{\beta}_k x_{2k} \\ \vdots \\ \hat{\beta}_0 x_{n0} + \dots + \hat{\beta}_k x_{nk} \end{pmatrix}}_{\hat{\mathbf{y}}}$$

Linear algebra: The sum of squares

If you have a vector

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

then

$$\mathbf{a}'\mathbf{a} = \sum_{i=1}^{n} a_i^2$$

Objective function

- Using the matrix notation, our objective function becomes

$$
\begin{aligned}
g(\tilde{\beta}) &= (\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta}) \\
&= (\mathbf{y}' - \tilde{\beta}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\tilde{\beta}) \\
&= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\tilde{\beta} - \tilde{\beta}\mathbf{X}'\mathbf{y} + \tilde{\beta}'\mathbf{X}'\mathbf{X}\tilde{\beta}
\end{aligned}
$$

Differentiation

$$g(\tilde{\beta}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\tilde{\beta} - \tilde{\beta}\mathbf{X}'\mathbf{y} + \tilde{\beta}'\mathbf{X}'\mathbf{X}\tilde{\beta}$$

- How do we differentiate it with respect to $\tilde{\beta}$?
- First, let's define
$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}.$$
- Now we need to establish two properties

Linear algebra: Differentiating a linear form

If you have two vectors,

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

then

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}'\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$$

This property implies that

$$\frac{\partial \mathbf{y}'\mathbf{X}\tilde{\beta}}{\partial \tilde{\beta}} = \frac{\partial \tilde{\beta}'\mathbf{X}'\mathbf{y}}{\partial \tilde{\beta}} = \mathbf{X}'\mathbf{y}$$

Linear algebra: Differentiating a quadratic form

If you have a vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

and a square and symmetric matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} \dots a_{1n} \\ \vdots \ddots \vdots \\ a_{n1} \dots a_{nn} \end{pmatrix}$$

(symmetric means that $a_{ij} = a_{ji}$) then

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

This property implies that

$$\frac{\partial \tilde{\beta}'\mathbf{X}'\mathbf{X}\tilde{\beta}}{\partial \tilde{\beta}} = 2\mathbf{X}'\mathbf{X}\tilde{\beta}.$$

OLS

- Using these two properties, we can find that the derivative of the objective function is

$$\frac{\partial g}{\partial \tilde{\beta}} = \frac{\partial \mathbf{y}'\mathbf{y}}{\partial \tilde{\beta}} - \frac{\partial \mathbf{y}'\mathbf{X}\tilde{\beta}}{\partial \tilde{\beta}} - \frac{\partial \tilde{\beta}\mathbf{X}'\mathbf{y}}{\partial \tilde{\beta}} + \frac{\partial \tilde{\beta}'\mathbf{X}'\mathbf{X}\tilde{\beta}}{\partial \tilde{\beta}}$$

OLS

- Using these two properties, we can find that the derivative of the objective function is

$$\frac{\partial g}{\partial \tilde{\beta}} = 0 - \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\beta}$$

OLS

- Using these two properties, we can find that the derivative of the objective function is

$$\frac{\partial g}{\partial \tilde{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\beta}$$

OLS

- Using these two properties, we can find that the derivative of the objective function is

$$\frac{\partial g}{\partial \tilde{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\beta}$$

- All the partial derivates must be zero at the optimum, therefore

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}.$$

OLS

- Using these two properties, we can find that the derivative of the objective function is

$$\frac{\partial g}{\partial \tilde{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\beta}$$

- All the partial derivates must be zero at the optimum, therefore

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}.$$

- Multiplying by the inverse of the $\mathbf{X}'\mathbf{X}$ on the left, we get

OLS

- Using these two properties, we can find that the derivative of the objective function is

$$\frac{\partial g}{\partial \tilde{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\beta}$$

- All the partial derivates must be zero at the optimum, therefore

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}.$$

- Multiplying by the inverse of the $\mathbf{X}'\mathbf{X}$ on the left, we get

OLS estimator

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Inverse

Note

How do we now that the inverse of $\mathbf{X}'\mathbf{X}$ exists? This is guaranteed by our assumption of no perfect collinearity.

Next Time on *Introductory Econometrics*...

Properties of multiple regression