Lecture 1: A review of linear regression analysis

Rigissa Megalokonomou
University of Queensland

Course Information

- This is A course in Applied Micro-Econometrics. These methodologies are used most widely in fields like labour economics, development economics, health economics, economics of crime, economics of education among others.

- One Lecture (2hr) and one Practical (2hr) per week
- Lecture is online.
- Practical is online or face-to-face depending on delivery mode (flexible delivery or external delivery).

Course Information

- One Lecture (2hr) per week
  - Pre-recorded and uploaded on BB at least one day before the regular lecture time.
  - Link(s) will be available on BB.
- One Practical (2hr) per week
  - Online or face-to-face depending on delivery mode.
  - Online practical will be recorded and uploaded on BB.
- Students use STATA for practicals.
- Internal students will access STATA in the labs.
- External students will virtually access STATA on UQ's license (more info on BB).

Course's Consultation Times

Rigissa Megalokonomou

- Wednesdays 15:30-16:00 via zoom for questions related to this week's lecture.
- Fridays 16.00-18:00 via zoom
- Zoom id will be provided on BB.
- Email: R.Megalokonomou@uq.edu.au

Pablo and Patrick's consultation hours will be available on BB shortly along with their zoom details.

Course Information

- Assessment for ECON7360:

  1. Two Online Quizes (10% and 15%), due dates: 22nd of September (16:00) and 7th of November (16:00)
     - First online quiz: All material covered in lectures and tutorials up to and including lecture 7 are examinable.
     - Second online quiz:All material covered in lectures and tutorials in the whole course are examinable.

  Note:

  2. Two Problem Sets (20% each), due dates: 6th of September (16:00) and 18th of October (16:00)
     - First problem set consists of questions which require short answers and some calculations.
     - Second problem set is a STATA-based assessment.

  3. Article review (20%): A critical review of a journal article that uses methodologies covered in the course. Due date is 4th of October (16:00).

  4. Research Proposal (15%): It incorporates the submission of a complete project plan that includes a clearly defined research question, literature review, plan for data collection, and the methodology that addresses the question. Due date is 30th of October (16:00).

Course Information

- Required Resources:
  1) Wooldridge, Jeffrey M. (2013) **Introductory Econometrics: A Modern Approach** 5th (or 4th or 6th) edition.
  2) Angrist, Joshua and Jorn-Steffen Pischke. 2009. **Mostly Harmless Econometrics**. Princeton: Princeton University Press.
- Recommended Resources (the following are standard textbooks in micro-econometrics):
  - Cameron, Colin and Pravin Trivedi. 2009. **Microeconometrics Using Stata**. College Station, TX: Stata Press.
  - Wooldridge, Jeffrey M. 2010. **Econometric Analysis of Cross Section and Panel Data**. MIT Press.
  - Cameron, Colin and Pravin Trivedi. 2005. **Microeconometrics: Methods and Applications**. Cambridge University Press.

Course Outline

- This course covers the basic micro-econometric concepts and methods used in **modern** empirical economic research.
- The goal is to help you understand research papers (i.e. journal articles) in applied microeconomics and to give you those skills that are required for you to execute your own empirical project.
- One of the most important skills that you will be required to have in the workforce is the ability to convert a large and complex set of information into a nice neat package.
  - For this class, students will get some practice on this by reading academic articles and summarizing the content in a non-technical manner. Students will be asked to write a short (3 pages) paper that summarizes the key concepts of an article (article review) and their own research proposal.

Course Outline

- **Topics** include linear regression analysis, randomized controlled trial, instrumental variables estimation, linear panel data models, differences-in-differences method, differences-in-differences-in-differences method, simultaneous equations models, propensity score matching, regression discontinuity design, probit and logit models, quantile regressions etc.
- Look at various research papers that use those methods.
- Each lecture will include a theory part and then examples coming from academic papers (from lecture 2 onwards).
- I am also going to present a research paper in week 8.
- We will look at many examples (coming from experiments that are done or experiments that we would like to design) and do a fair amount of programming.

Course Outline

- We start with three workhorse methods that are in all applied econometrician's toolkit:
  - **Linear regression models** designed to control for variables that may mask the causal effects of interest.
  - **Instrumental variables** methods for the analysis of real and natural experiments.
  - **Differences-in-differences-type** of strategies that use repeated observations to control for unobserved omitted factors.
- The productive use of these techniques requires a solid conceptual foundation and a good understanding of machinery of statistical inference.

Reading for lecture 1:

A review of linear regression analysis in Introductory Econometrics (Courses like $ECON2300$, $ECON7310$)

- Please review chapters 1-4 and appendix D and E from the Introductory Econometrics: A Modern Approach textbook

- It is inevitable to use some **matrix algebra** to understand linear regression models in depth that are used in practice.

- Econometrics is the measurement of economic relations
- Need to know
  - **What is an economic relationship**?
  - **How do we measure such a relation**?
- Examples:
  - *production function*: relationship between output of firm and inputs of labor, capital, materials (output=f(inputs))
  - *earnings function*: relation between earnings and education, work experience, job tenure, worker's ability (earnings=f(education, experience, etc))
  - *education production function*: relation between student academic performance and inputs such as class-size, student, teacher, peer characteristics (score=f(class size, teacher charact., etc))
- All these relations can be expressed as mathematical functions:

$$y = f(x_1, x_2, ..., x_k)$$

can be approximated by linear regression model.

$$y = x_1\beta_1 + x_2\beta_2 + .. + x_k\beta_k + u$$
$$y = x\beta + u$$

Objective: Causal relation of economic variables

- Most empirical studies in economics are interested in **causal relations of economic variables**. The **goal** of most empirical studies in economics is to determine whether a change in one variable, say x_1, **causes** a change in another variable, say y, while we keep all other variables fixed **(ceteris paribus)**.
- Examples:
  - Does having another year of education cause an increase in monthly salary?
  - Does reducing class size cause an improvement in student performance?
  - Does attending a twenty-week job training program cause an increase in workers' productivity?

Simple association is not a proper measurement for a relation of economic variables

- Simply finding an association between two or more variables might be suggestive but it is rarely useful for policy analysis. In other words, **association does not imply causality**. See some examples:
  - If you know that there are two cities A and B and there are more police officers on the streets in city A would you expect the crime rate to be lower in city A compared to that of city B?
  - *No. City A might be dodgier! (The decision on the size of police officers might be correlated with other city-related factors that affect crime)*
  - Suppose that you want to examine the effect of hiring more teachers. Students in year t and grade g have a given performance. In year t+1 additional teachers are hired in grade g. If students' GPA go up next year, is it purely the effect of increasing the number of teachers?
  - *No. It could be that next cohort is smarter on average!*

How easy is it to think about the ceteris paribus assumption here?

Regression fundamentals We start with the linear regression framework because:

- Very robust technique that allows us to incorporate fairly general functional form relationships.
- Also provides a basis for more advanced empirical methods.
- A transparent and relatively easy to understand technique.
- Before we get into the important question of when a regression is likely to have a causal interpretation, let's review a number of regression facts and properties.
- The multiple linear regression model and its estimation using ordinary least squares (OLS) is the most widely used tool in econometrics.

$$y = \mathbf{x}\beta + u$$
$$\hat{\beta}_{ols} = (x'x)^{-1}x'y$$

Regression fundamentals

- Setting aside the relatively abstract causality problem for the moment, we start with the **mechanical** properties of the regression estimates.
- We start with the multiple linear regression model:

$$y = \mathbf{x}\beta + u$$

where $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)$ and $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)'$.

- y is called the dependent variable, outcome variable, response variable, explained variable, predicted variable and;
- X is called independent variable, explanatory variable, control variable, predictor variable, regressor, covariates;
- β_0 is the intercept parameter;
- β_j where j=1,2,..,k are slope parameters (our primary interest in most cases);
- u is called the error term or disturbance

OLS estimator

- For observation i = (1, 2, .., n),

$$y_i = x_i\beta + u_i \ (1)$$

$$min_\beta \Sigma_{i=1}^{n} u_i^2$$

OLS estimator for β chooses such β that minimizes the sum of error squares.

- In a matrix form, where $u = [u_1, u_2, ...u_n]$

$$min_\beta u'u(2)$$

From (1) you can substitute u with $u = (y - x\beta)$ in (2) and then set the derivative equal to 0. This gives us: $\hat{\beta} = (x'x)^{-1}x'y$

- What is the condition for $E(\hat{\beta}|x) = \beta$? (will prove that later)

The Multiple Linear Regression Model

- a structure in which one or more **explanatory variables** are considered to generate an outcome variable and so that is more amenable to **ceteris paribus** analysis as it allows us to explicitly control for many other factors that affect the outcome variable.
- Very robust technique that allows to incorporate fairly general functional form relationships.
- Also provides a **basis** for more advanced empirical methods.
- Transparent and relatively easy to understand technique.

Statistical Properties of the OLS estimators

$$y = \beta_0 + \mathbf{x}\beta + u$$

- $(y, \mathbf{x}, u)$ are random variables (a random variable is a variable taking on numerical values determined by the outcome of a random phenomenon).
- $(y, \mathbf{x})$ are observed variables (we can sample observations on them)
- u is unobservable (no statistical tests involving u)
- (β_0, β) are unobserved but can be estimated under certain conditions
- Model implies that u captures everything that determines y except for $\mathbf{x}$

β captures economic relationship between y and $\mathbf{x}$.

Statistical properties of the OLS estimators

In the case of a single covariate:

$$y = \beta_0 + \beta x + u$$

Estimators:

$$slope = \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; intercept = \hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}$$

where we have a sample of individuals $i = 1, 2, 3..., n$.

Population analogues:

$$slope = \frac{Cov(x, y)}{Var(x)}; intercept = E(y) - \hat{\beta}E(x)$$

Unbiasedness of OLS When is the OLS estimator $\hat{\beta}$ unbiased (i.e.,$E(\hat{\beta}) = \beta$)?

- A1. Model in the population is linear in parameters:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \beta + u$$

In matrix form: $y = \mathbf{x}\beta + u; \mathbf{x} = (1, x_1, x_2, ..., x_k)$

- A2. Have a random sample on $(y_i, \mathbf{x}_i)$. Draws are from iid. (hint: iid means independent and identically distributed random variables: if each random variable has the same probability distribution as the others and all are mutually independent.)

- A3. None of the independent variables is constant and there are no exact linear relationships among the independent variables.

- **A4**. Zero conditional mean of errors: The error has an expected value of zero given any values of the independent variables.

$$E(u|X) = 0 \Rightarrow cov(X, u) = 0$$

- **OLS is unbiased under A1-A4.**

Unbiasedness of OLS When is OLS unbiased (i.e.,$E(\hat{\beta}) = \beta$)?

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

where $\mathbf{X} \equiv (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n')'$ and $\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i$ and $y = X\beta + u$

$$\hat{\beta} = [(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'(X\beta + u)]$$

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

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

$$E(\hat{\beta}|\mathbf{x}) = \beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'E(u|\mathbf{x})) = \beta$$

(because of A4. $E(u|X) = 0$)

- Unbiasedness is a feature of sampling distribution of $\hat{\beta}_{OLS}$: central tendency to true parameter value

Omitted Variable Problem What if we exclude a relevant variable from the regression?

- This is an example of misspecification.
- For example *ability* is not observed and not included in the wage equation (missing $\beta_2 ability$): $wage = \beta_0 + \beta_1 educ + u$
- β_2 is probably positive

- The higher a student's *ability* is, the higher will be her wage.
- But also: the higher a student's *ability* is, the higher will be her education level so $cov(X_1, X_2) > 0, X_1 : educ, X_2 : ability$
- This means that β_1 has an **upward bias**

Omitted Variable Problem Suppose that the researcher mistakenly uses:

$$y = a^* + b_1^* X_1 + e$$

while X2 is mistakenly omitted from the model. So the model should have been:

$$y = a + b_1 X_1 + b_2 X_2 + u$$

- How does b_1 (the regression estimate from the correctly specified model) compare to b_1^* (the regression estimate from the mis-specified model)?

$$b_1^* = \frac{Cov(X_1, Y)}{Var(X_1)} = \frac{Cov(X_1, a + b_1 X_1 + b_2 X_2 + u)}{Var(X_1)} =$$

(hint: substitute the formula for Y from the correctly specified model)

$$\frac{Cov(X_1, a) + b_1 Cov(X_1, X_1) + b_2 Cov(X_1, X_2) + Cov(X_1, u)}{Var(X_1)} =$$

(hint: Cov(a+b,c+d) = Cov(a,c) + Cov(a,d) + Cov(b,c) + Cov(b,d))

$$\frac{0 + b_1 Var(X_1) + b_2 Cov(X_1, X_2) + 0}{Var(X_1)}$$

(hint: Recall that Cov(variable, constant) = 0. Also, Xs are uncorrelated with the residuals. (A4.)

$$b_1^* = b_1 + b_2 \frac{Cov(X_1, X_2)}{Var(X_1)}$$

Sampling variance of OLS slope estimator

- A5. (Homoskedasticity) The error u has the same variance given any values of explanatory variables

$$Var(u|\mathbf{x}) = \sigma^2$$

  - Variance in the error term u, conditional on the explanatory variables, is the same for all combinations of the explanatory variables

- Under A1–A5, conditional on the sample values of the independent variables,

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

for $j = 1, 2, .., k$ where $SST_j = \Sigma_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j, and R_j^2 is the R-squared from regression x_j to all other independent variables including an intercept.

- The size of $Var(\hat{\beta}_j)$ is important: a larger variance means a less precise estimator.

Aside: The components of the OLS variance

- Variance of $\hat{\beta}_j$ depends on three factors: σ^2, SST_j, R_j^2
  - (1) The error variance, σ^2: A larger σ^2 means larger variance $->$ more noise in the equation $->$ makes it hard to precisely estimate β. For a given dependent variable y, there is only one way to reduce the error variance, and that is to add more explanatory variables to the equation. It is not always possible to find additional legitimate factors that affect y.
  - (2) The total sample variation in x_j, SST_j: The larger the variation in x_j (SST_j) $->$the smaller the variance. There is a way to increase the sample variation in each of the independent variables: increase the sample size.
  - (3) The linear relationships among the independent variables, R_j^2: high multicollinearity between x_j and other independent variables leads to imprecise estimate of x_j (e.g perfect multicollinearity means $R_j^2 = 1$ and the variance is infinite) Note that $R_j^2 = 1$ is ruled out by assumption 3.

2. Gauss-Markov Theorem: OLS is BLUE. Under A1-A5, the OLS estimator $\hat{\beta}$ is the best linear unbiased estimator (BLUE) of true parameter β.
Best means the most efficient (i.e. smallest variance) estimator.

Assumptions again...

- A1: Model is linear
- A2: Have random sample of (yi,xi)
- A3: None of the Xs is constant and there is no perfect multicollinearity
- A4: Zero conditional mean of errors (exogeneity)
- A5: Error has the same variance for all Xs (homoskedasticity)
- A6: ...

3. Inference with OLS estimator

- Under Gauss-Markov (GM) assumptions, the distribution of $\hat{\beta}_j$ can have virtually any shape.

- To make the sampling distribution tractable, we add an assumption on the distribution of the errors:
  - **A6. Normality**. The population error u is normally distributed with zero mean and constant variance: $u \sim N(0, \sigma^2)$

- The assumption of normality, as we have stated it, subsumes both the assumption of the error process being independent of the explanatory variables (A4), and that of homoskedasticity (A5). For cross-sectional regression analysis, these assumptions define the classical linear model (CLM).

What does A6 add?

- The CLM assumptions A1-A6 allow us to obtain the exact sampling distributions of t-statistics and F-statistics, so that we can carry out exact **hypotheses tests**.
- The rationale for A6: we can appeal to the **Central Limit Theorem** to suggest that the sum of a large number of random factors will be approximately normally distributed.
- The assumption of normally distributed error is probably not a bad assumption.

Testing hypotheses about single parameter: the t test Under the CLM assumptions, a test statistic formed from the OLS estimates may be expressed as:

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

This test statistic allows us to test the null hypothesis:

$$H_0 : \beta_j = 0$$

- We have $(n-k-1)$ degrees of freedom. Where n is not that large relative to k, the resulting t distribution will have considerably fatter tails than the standard normal.

- Where $(n-k-1)$ is a large number - greater than 100, for instance the t distribution will essentially be the standard normal. $->$ That is why big n helps.

Summary

- A1-A4: OLS is unbiased
- A1-A5: We derive $var(\hat{\beta}_j)$
- A1-A5: Gauss-Markov holds and the OLS estimator is BLUE
- A1-A6: We obtain the exact distributions of t-statistics and F-statistics

The idea is:

- Broadly speaking, empirical micro-econometric methodologies can be viewed as tools to solve this problem (i.e. $cov(x_1^*, u^*) \neq 0$), which is also called **endogeneity problem** or violation of exogeneity assumption. Typically we use two approaches:
  - (i) Try to include as much information as possible in $\mathbf{x}_2$ so that u is as small as possible. The examples of approaches include DID method, fixed effects estimator in panel data model, among many others.
  - (ii) Try to design the setting and the dataset such that $cov(x_1)$ is random or isolate the part of variation in $cov(x_1)$ that is uncorrelated to unobserved factors u. The examples of this type of approaches include random experiment design, instrumental variables approach among others.
  - Of course, we can combine these approaches.

Summary of Linear Regression Model

- Advantages:
  - Very robust technique that allows to incorporate fairly general functional form relationships
  - Also provides a **basis** for more advanced empirical methods
  - Transparent and relatively easy to understand technique
- Economists use econometric methods to effectively hold **other factors** fixed.
- Causality and **ceteris paribus**: In economic theory, economic relations hold ceteris paribus (i.e. holding all other relevant variables fixed); but since the econometrician does not observe **all of the factors** that might be important, we **cannot** always make sensible inferences about potentially causal factors.
- Our best hope is that we might control for many of the factors, and be able to use our empirical findings to examine whether systematic/important factors have been omitted.

Supplementary slides

Main issue here is: when a regression is likely to have a causal interpretation

- Consider the following model and suppose that we are interested in estimating β_1:

$$y = f(x_1, x_2, \ldots, x_k)$$
$$y = \beta_1 x_1 + \mathbf{x}_2 \beta_2 + u$$
$$E(y|\mathbf{x}) = \beta_1 x_1 + \mathbf{x}_2 \beta_2$$
$$E(u|\mathbf{x}) = 0$$

- Suppose we estimate $y = \beta_1 x_1 + u_1$ instead and we implement OLS to get $\hat{\beta}_1$.
- Note that $\hat{\beta}_1 = \frac{cov(x_1, y)}{cov(x_1, x_1)}$.

$$\hat{\beta}_1 = \frac{cov(x_1, \beta_1 x_1)}{cov(x_1, x_1)} + \frac{cov(x_1, \mathbf{x}_2 \beta_2)}{cov(x_1, x_1)} + \frac{cov(x_1, u)}{cov(x_1, x_1)}$$
$$\hat{\beta}_1 = \beta_1 + \frac{cov(x_1, \mathbf{x}_2)}{cov(x_1, x_1)} \beta_2 + \frac{cov(x_1, u)}{cov(x_1, x_1)}$$

- What is the condition for $E(\hat{\beta}_1|\mathbf{x}) = \beta_1$?

Main issue here is:

- Even if we assume that $E(\frac{cov(x_1,u)}{cov(x_1,x_1)}|\mathbf{x}) = 0$, we still have $E(\frac{cov(x_1,\mathbf{x}_2)}{cov(x_1,x_1)}|\mathbf{x}) \neq 0$ where $\mathbf{x} = (x_1, \mathbf{x}_2)$.

$$E(\hat{\beta}_1|\mathbf{x}) = \beta_1 + E(\frac{cov(x_1,\mathbf{x}_2)}{cov(x_1,x_1)}|\mathbf{x})\beta_2$$

- In the previous examples, $cov(x_1, \mathbf{x}_2) \neq 0$ so $E(\hat{\beta}_1|\mathbf{x}) \neq \beta_1$.

$$wage = \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 tenure + u$$

- $wage = \beta_1 educ + u_1$

Supplementary slides: What is the importance of assuming normality for the error process? Under the assumptions of the classical linear model, normally distributed errors give rise to normally distributed OLS estimators:

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$$

where $Var(\hat{\beta}_j)$ is provided in p.19 and which will then imply that:

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

where $sd(\hat{\beta}_j) = \sigma_{\hat{\beta}_j}$

- This follows since each of the $\hat{\beta}_j$ can be written as a linear combination of the errors in the sample.
- Since we assume that the errors are independent, identically distributed normal random variates, any linear combination of those errors is also normally distributed.
- Any linear combination of the $\hat{\beta}_j$ is also normally distributed, and a subset of these estimators has a joint normal distribution.

Supplementary slides: Heteroskedasticity robust variance: Under A1-A4 assumptions, heteroskedasticity robust variance for $\hat{\beta}_j$ is provided as

$$\widehat{Avar}(\hat{\beta}_j|\mathbf{x}) = \frac{\Sigma_{i=1}^n \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2} \tag{1}$$

where $\hat{r}_{ij}$ denotes the ith residual from regression x_j on all other independent variables, and SSR_j is the sum of squared residuals from this regression.

In matrix form:

$$\widehat{Avar}(\hat{\beta}|\mathbf{x}) = (\Sigma_{i=1}^n \hat{x}_i' \hat{x}_i)^{-1} \cdot (\Sigma_{i=1}^N \hat{x}_i' \hat{x}_i \hat{u}_i^2) \cdot (\Sigma_{i=1}^N \hat{x}_i' \hat{x}_i)^{-1}$$

$$Avar(\hat{\beta}|\mathbf{x}) = E((\mathbf{x}'\mathbf{x})^{-1}(\mathbf{x}' uu' \mathbf{x})(\mathbf{x}'\mathbf{x})^{-1}|\mathbf{x})$$

where the square roots of the diagonal elements of this matrix are the heteroskedasticity robust standard errors as the square roots of variance. Under homoskedasticity, $\sigma^2 \cdot (\mathbf{x}'\mathbf{x})^{-1}$.

- Most statistical packages now support the calculation of these robust standard errors when a regression is estimated.
- The heteroskedasticity robust standard errors may be used to compute the heteroskedasticity-robust t-statistic and, likewise, F-statistics.

OLS estimator

- For observation i = (1, 2, .., n),

$$y_i = x_i\beta + u_i$$
$$min_\beta \Sigma_{i=1}^n u_i^2$$

OLS estimator for β chooses such β that minimizes the sum of error squares.

- In a matrix form, where u = $[u_1, u_2, ...u_n]$

$$min_\beta u'u$$

$$(y - x\beta)'(y - x\beta) = 0$$

$$y'y - \hat{\beta}'x'y - y'x\hat{\beta} + \hat{\beta}'x'x\hat{\beta} = 0$$

$$y'y - 2\hat{\beta}'x'y + \hat{\beta}'x'x\hat{\beta} = 0$$

because the transpose of a scalar is the scalar i.e $y'x\hat{\beta} = (\hat{\beta}'x'y)' = \hat{\beta}'x'y$
So you need to take the derivative w.r.t. $\hat{\beta}$ and so:
$-2x'y + 2x'x\hat{\beta} = 0 => 2x'x\hat{\beta} = 2x'y => \hat{\beta} = (x'x)^{-1}x'y$

- If x=x_1 is scalar, $\hat{\beta}_1 = \frac{cov(x_1,y)}{cov(x_1,x_1)}$.