Econometrics
The Simple Regression Model

João Valle e Azevedo

Faculdade de Economia
Universidade Nova de Lisboa

Spring Semester
Objectives

Given the model

\[ y = \beta_0 + \beta_1 x + u \]

- Where \( y \) is earnings and \( x \) is years of education
- Or \( y \) is sales and \( x \) is spending in advertising
- Or \( y \) is the market value of an apartment and \( x \) is its area

Our **objectives** for the moment will be:

- Estimate the unknown parameters \( \beta_0 \) and \( \beta_1 \)
- Assess how much \( x \) explains \( y \)
The Gauss-Markov Assumptions

- **Assumption SLR.1 (Linearity in Parameters)**

\[ y = \beta_0 + \beta_1 x + u \]

- \( y \) is the dependent variable (or explained variable, or response variable, or the regressand)
- \( x \) is the independent variable (or explanatory variable, or control variable, or the regressor)
- \( u \) is the error term or disturbance (\( y \) is allowed to vary for the same level of \( x \))

This is a **population** equation. It is **linear in the (unknown) parameters**.
The Gauss-Markov Assumptions

• Assumption **SLR.2 (Random Sampling)** Random sample of size $n$, \( \{(x_i, y_i): i=1,2,...,n\} \) such that:

\[
y_i = \beta_0 + \beta_1 x_i + u_i, \ i = 1, 2, ..., n
\]

• Assumption **SLR.3 (Sample Variation in the Explanatory Variable)**
  The sample outcomes on $x$, namely \( \{x_i: i=1,2,...,n\} \), are not all the same value
The Gauss-Markov Assumptions

- Assumption **SLR.4 (Zero Conditional Mean)** The error $u$ has an expected value of zero given any value of the explanatory variable

\[ E(u|x) = 0 \]

**Crucial Assumption!**
Remember the previous examples:

\[ Earnings = \beta_0 + \beta_1 Education + u \]

\[ CrimeRate = \beta_0 + \beta_1 Policeman + u \]

In these cases **SLR.4** most likely fail
The Gauss-Markov Assumptions

- Assumption **SLR.4’ (Zero Mean)** The error $u$ has an expected value of zero given any value of the explanatory variable

$$E(u) = 0$$

Not really necessary given that **SLR.4 implies SLR.4’**

- This is not a restrictive assumption since we can always use $\beta_0$ so that **SLR.4’** holds
Zero Conditional Mean

\[ E(u|x) = 0 \Rightarrow E(y|x) = \beta_0 + \beta_1 x \]

**Figure:** For any \( x \), the distribution of \( y \) is centered about \( E(y|x) \)
Graph of $y_i = \beta_0 + \beta_1 x_i + u_i$

**Figure:** Population regression line, data points and the associated error terms
**OLS Estimators**

- Want to estimate the population parameters from a sample
- Let \( \{(x_i, y_i): i=1,2,...,n\} \) denote a random sample of size \( n \) from the population
- For each observation in this sample:
  \[
y_i = \beta_0 + \beta_1 x_i + u_i, \ i = 1, 2, ..., n
\]
- To derive the OLS estimates we need to realize that our main assumption of \( E(u|x) = E(u) = 0 \) also implies:
  \[
  Cov(x, u) = E(xu) = 0
  \]
- Remember from basic probability that \( Cov(x,y) = E(xy) - E(x)E(y) \)
OLS Estimators

- **Two Restrictions:**
  \[ E(u|x) = E(u) = 0 \]
  \[ \text{Cov}(x,u) = E(xu) = 0 \]

- **Since** \( u = y - \beta_0 - \beta_1 x \):
  - **Two Moment Conditions:**
    \[ E(y - \beta_0 - \beta_1 x) = 0 \]
    \[ E[x(y - \beta_0 - \beta_1 x)] = 0 \]
  - **Sample versions:**
    \[ n^{-1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \]
    \[ n^{-1} \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \]
OLS Estimators

We have two equations and two unknowns $\hat{\beta}_0$, $\hat{\beta}_1$

- Pick $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the sample moments match the population moments
- From the **first sample moment condition** and given the definition of a sample mean and properties of summation:

$$n^{-1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\iff \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\iff \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
OLS Estimators

From the second sample moment condition:

\[ n^{-1} \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \]

Plug-in \( \hat{\beta}_0 \) yields:

\[ \sum_{i=1}^{n} x_i (y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) = 0 \]

This simplifies to:

\[ \sum_{i=1}^{n} x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) \]

Rearranging gives:

\[ \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

Thus, \( \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \)
OLS Estimators

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]

- Need \( x \) to vary in our sample so that the slope parameter is well-defined!

\[ \sum_{i=1}^{n} (x_i - \bar{x})^2 > 0 \]

- Sample covariance between \( x \) and \( y \) divided by the sample variance of \( x \)

- If \( x \) and \( y \) are positively correlated, the slope will be positive

- If \( x \) and \( y \) are negatively correlated, the slope will be negative
  - It is an estimator if we treat \((x_i, y_i)\) as random variables
  - It is an estimate once we substitute \((x_i, y_i)\) by the actual observations
Example

\[ \hat{\text{wage}}_i = 245.073 + 52.513\text{educ}_i \]

\[ n = 11064, \text{ Data from the Employment Survey (INE), 2003} \]

- \textit{wage} is net monthly wage and \textit{educ} measures the number of years of schooling completed

- So, we estimate a positive relation between these two variables:
  - An additional year of schooling leads to an estimated expected increase of 52.513 euros in the wage on average, ceteris paribus
  - Caution about the interpretation of these estimates!
Some Definitions

Intuitively, OLS is fitting a line through the sample points such that the SSR is as small as possible, hence the term least squares.

The residual, \( \hat{u} \), is an estimate of the error term, \( u \), and is the difference between the fitted line (sample regression function) and the sample point:

- **Fitted Value**
  \[
  \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i
  \]

- **Residual**
  \[
  \hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i
  \]

- **Sum of Squared Residuals (SSR)**
  \[
  \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
  \]
Graph of $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

**Figure:** Sample regression line, sample data points and the associated estimated error terms - the residuals
Alternative approach to derive OLS

- We want to choose our parameters such that we **minimize the SSR**:

\[
\text{SSR} = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
\]

\[
\frac{\delta \text{RSS}}{\delta \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0
\]

\[
\frac{\delta \text{RSS}}{\delta \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0
\]

- These equations are equivalent to the same equations we had before: thus, same solution!
More Terminology

Think of each observation as being made up of an explained part and an unexplained part:

\[ y_i = \hat{y}_i + \hat{u}_i \]

- **Total Sum of Squares (SST)**
  \[ \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

- **Explained Sum of Squares (SSE)**
  \[ \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \]

- **Residual Sum of Squares (SSR)**
  \[ \sum_{i=1}^{n} \hat{u}_i^2 \]
$$SST = SSE + SSR$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

$$= \sum_{i=1}^{n} [\hat{u}_i + (\hat{y}_i - \bar{y})]^2$$

$$= \sum_{i=1}^{n} \hat{u}_i^2 + 2 \sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$= SSR + SSE$$
Goodness of Fit

- How well does our sample regression line fit our sample data?
  - If SSR is very small relative to SST (or SSE very large relative to SST), then a lot of the variation in the dependent variable y is explained by the model.

- Can compute the fraction of the total sum of squares (SST) that is explained by the model: the \textbf{R-squared} of the regression

\[
R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}
\]

\[
0 \leq R^2 \leq 1
\]

- The lower the SSR, the higher will be the \( R^2 \)
- OLS maximises the \( R^2 \) (minimises SSR)
Goodness of Fit - Example (Cont.)

\[ \hat{\text{wage}}_i = 245.073 + 52.513 \text{educ}_i \]

\[ n = 11064, \text{ Data from the Employment Survey (INE), 2003} \]

\[ R^2 = 0.262 \]

- A lot of variation in the dependent variable (wage) is left unexplained by the model
- This is not related to the fact that educ is most probably correlated with \( u \)
- Can have low \( R^2 \) even if all the assumptions hold true
Properties of OLS Estimators

- Thought experiment:
  - Get a sample of size \( n \) from the population many times (infinite times)
  - Compute the OLS estimates
  - Is the average of these estimates equal to the true parameter values?
  - Under SLR.1 to SLR.4 it turns out that the answer is positive

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]

- These are estimators if we treat \((x_i, y_i)\) as random variables

Theorem

*Under these assumptions:*

\[ E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1 \]
Proof of Unbiasedness of OLS

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{SST_x}
\]

- where:

\[
SST = \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

- Note that:

\[
\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \quad \text{and} \quad \sum_{i=1}^{n} (x_i - \bar{x})x_i = \sum_{i=1}^{n} (x_i - \bar{x})^2
\]
Proof of Unbiasedness of OLS (Cont.)

- Focusing on the numerator:
  \[ \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})y_i \]
  \[ = \sum_{i=1}^{n} (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i) \]
  \[ = \beta_1 SST_x + \sum_{i=1}^{n} (x_i - \bar{x})u_i \]

- Putting together:
  \[ \hat{\beta}_1 = \frac{\beta_1 SST_x + \sum_{i=1}^{n} (x_i - \bar{x})u_i}{SST_x} \]
  \[ = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})u_i}{SST_x} \]
Proof of Unbiasedness of OLS (Cont.)

\[ \hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})u_i}{SST_x} \]

- Now take expectations conditional on the values of the \( x \)'s, that is, treat the \( x \)'s as constants. To avoid heavy notation that isn't done explicitly

\[ E(\hat{\beta}_1) = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})E(u_i)}{SST_x} \]

\[ = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})0}{SST_x} \]

\[ = \beta_1 \]

- And if the conditional expectation is a constant, the unconditional expectation is also that constant...
Proof of Unbiasedness of OLS (Cont.)

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]

- Now, it’s really easy to prove unbiasedness of \( \hat{\beta}_0 \). Try it yourself!

Unbiasedness Summary

- The OLS estimators of \( \beta_0 \) and \( \beta_1 \) are unbiased
- However, in a given sample we may be close or far from the true parameter, unbiasedness is a minimal requirement
- Unbiasedness depends on the 4 assumptions: if any assumption fails, then OLS is not necessarily unbiased
Variance of the OLS Estimators

- We know already that the sampling distribution of our estimators is centered around the true parameter.
- But, is the distribution **concentrated** around the true parameter?
- To answer this question, we add an additional assumption:

  - **Assumption SLR.5 (Homoskedasticity)** The error $u$ has the same variance given any value of the explanatory variable.

  $$\text{Var}(u|x) = \sigma^2$$
Variance of the OLS Estimators

- Remember that:

\[ \sigma^2 = Var(u|x) = E(u^2|x) - [E(u|x)]^2 \]

\[ = E(u^2|x), \text{given that } E(u|x)=0 \]

\[ = E(u^2) = \text{Var}(u) \]

- So, \( \sigma^2 \) is also the unconditional variance, called the variance of the error.
- \( \sigma \), the square root of the error variance, is called the standard deviation of the error.

- Can write:

\[ E(y|x) = \beta_0 + \beta_1x \text{ and } \text{Var}(y|x) = \sigma^2 \]
Homoskedastic Case

\[ E(y|x) = \beta_0 + \beta_1 x \]

**Figure**: How spread out is the distribution of the estimator
Heteroskedastic Case

Figure: How spread out is the distribution of the estimator
Variance of the OLS Estimators (Cont.)

\[ \hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}) u_i}{SST_x}, \text{where } SST_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

\[ \text{Var}(\hat{\beta}_1) = \text{Var} \left( \beta_1 + \frac{1}{SST_x} \sum_{i=1}^{n} (x_i - \bar{x}) u_i \right) \]

\[ = \left( \frac{1}{SST_x} \right)^2 \text{Var} \left( \sum_{i=1}^{n} (x_i - \bar{x}) u_i \right) \]

\[ = \left( \frac{1}{SST_x} \right)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{Var}(u_i) \]

\[ = \left( \frac{1}{SST_x} \right)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma^2 = \sigma^2 \left( \frac{1}{SST_x} \right)^2 SST_x \]
Variance of the OLS Estimators (Cont.)

\[ Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}, \text{where } SST_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

- The larger the error variance, \( \sigma^2 \), the larger the variance of the slope estimator
- The larger the variability in the \( x_i \), the smaller the variance of the slope estimate
- \( SST_x \) tends to increase with the sample size \( n \), so variance tends to decrease with the sample size

Can also show:

\[ Var(\hat{\beta}_0) = \frac{n^{-1} \sigma^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]
Variance of the OLS

Theorem

Under Assumptions **SLR.1 through SLR.5**

\[
\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

\[
\text{Var}(\hat{\beta}_0) = \frac{n^{-1} \sigma^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

conditional on the sample values \( \{x_1, x_2, \ldots \} \)
Estimating the Error Variance

- In practice, the error variance $\sigma^2$ is unknown since we don’t observe the errors, we must estimate it...

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2} = \frac{SSR}{n-2}$$

- Can be show that this is an unbiased estimator of the error variance.

- The so-called standard error of the regression (SER) is given by:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

- Recall that:

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}}, \text{and similarly for } \hat{\beta}_0$$
Example (Cont.)

\[ \hat{\text{wage}}_i = 245.073 + 52.513\text{educ}_i \]

\( n = 11064 \), Data from the Employment Survey (INE), 2003

\[ R^2 = 0.262 \]

\[ \hat{\sigma}^2 = 155366 \text{ and } \hat{\sigma} = 394.164 \]

\[ \text{se}(\hat{\beta}_0) = 7.575 \text{ and } \text{se}(\hat{\beta}_1) = 0.837 \]