Econ 103: Introduction to Econometrics
Week 1

Francisco Díaz-Valdés*

Fall 2025

1 R installation tutorial

1.1 R and Rstudio

Here is a step-by-step guide for installing R and Rstudio. It is important that you first install R and then R Studio.

1. Click here to get R, the actual programming language. Choose your operating system, and download the program.

2. Click here to access the Rstudio website. Rstudio is an excellent IDE (Integrated Development Environment). To put it simply, Rstudio is an interface used to interact with R.

Now, open R Studio and, in the console, type the following command:

```
print("Hello world!")
```

If you see the message "Hello world!" in the console, you have successfully installed R and R Studio.

1.2 Basic Calculations

To familiarize ourselves with R, let's do some simple calculations.

1.2.1 Addition, Subtraction, Multiplication, and Division

Math	R code	Result
$3 + 2$	`3 + 2`	5
$3 - 2$	`3 - 2`	1
$3 \cdot 2$	`3 * 2`	6
$3/2$	`3/2`	1.5

[*] Many thanks to all previous TAs for providing the notes. All mistakes are my own. Please get in touch with me at fdiazvaldes@g.ucla.edu if you spot any typos or mistakes. Some parts of this note rely on https://scpoecon.github.io/ScPoEconometrics.

1.2.2 Exponents and logarithms

Math	R code	Result
3^2	3 ∧ 2	9
$2^{(-3)}$	2 ∧ (−3)	0.125
$100^{1/2}$	100 ∧ (1/2)	10
$\sqrt{100}$	sqrt(100)	10

1.2.3 Mathematical Constants

Math	R code	Result
π	pi	3.1415927
e	exp(1)	2.7182818

1.3 Getting Help

To get documentation about a function in R, put a question mark in front of the function name or call the function `help(function)`. Some examples:

```
?exp
help(exp) # help() is equivalent
help(ggplot,package="ggplot2") # show help from a certain package
```

1.4 Installing packages

One of the most important packages is the PoEdata, which contains the data we will use in this course. To install it, type the following command in the console:

```
# Install devtools to get access to the POE package
install.packages("devtools")

# Install PoEdata
devtools::install_git("https://github.com/ccolonescu/PoE5Rdata",force=TRUE)
```

Check if PoE5Rdata is installed by typing the following command:

```
# Load PoEdata
library(POE5Rdata)

# Load food dataset
data("food")

# Check first rows of the dataset
head(food)
```

If you see the first rows of the dataset, you have successfully installed the PoEdata package.

Now, let's install the other packages that we might use in this course:

```
# Install other packages
install.packages(c("tidyverse","bookdown", "knitr", "xtable", "printr","stargazer","
    rmarkdown"))
```

1.5 R Markdown

R Markdown is a file format for creating documents containing R code and text. You can use R Markdown to create reports, presentations, and websites. In this course, the two projects must be typed in R Markdown format and submitted in HTML or PDF format.

1.5.1 Creating an R Markdown file

- To create an R Markdown file, open R Studio and click on File − > New File − > *R* Markdown.
- Choose the output format (HTML or PDF) and give a title to your document.
- A new file with some example text and code will be created. Modify it to fit your needs, and click ”Knit” to create the document.

1.5.2 R Markdown syntax

R markdown uses a specific syntax to include R code in the markdown document. Markdown is a markup language for creating formatted text using a plain-text editor. The following is an example of an R Markdown file:

```
# Title
This is a paragraph.
‘‘‘{r }
# This is an R code chunk
print("Hello world!")
‘‘‘
```

Note that the R code is enclosed in triple backticks and the R code chunk starts with {r}. We can choose to display the code or not by using the option {r, echo=TRUE} or {r, echo=FALSE}.

2 Resources for learning R

Great Website for beginners Bookdown R guide is a great website for beginners. Lots of examples.

Another Website to get started with R ScPoEconomics. Many econometrics applications

R for data science book This is a free book that you can access here.

Ryan Longmuir's webpage Ryan is another TA for the course and he put together some materials about R. You can check it here.

General resources Whenever you have a question, you can google it and check the Stack Overflow website. It is a great resource for R and other programming languages. Another resource is to ask ChatGPT or other Language Models, they are really good at answering simple questions. There are also excellent youtube tutorials for R.

3 Theory

3.1 Parameters, estimators and estimate

Parameter Numerical characteristic of the distribution. Example: expected value μ, variance σ^2, lower quartile p_{25}, etc. Example in joint distribution: $\text{Corr}(X,Y)$

Estimator Random variable that is a function of the sample $(X_1,\dots,X_n)$ and estimates an unknown parameter. It is not function of the data. Example: $\bar{X}_n$ for the expected value μ. We generally want estimators that are consistent, unbiased and have low variance.

Unbiased Estimator An estimator $\hat{\theta}$ is unbiased for the parameter θ if $\mathbb{E}[\hat{\theta}]=\theta$. Example: $\mathbb{E}[\bar{X}_n]=\mu$

Consistent Estimator An estimator $\hat{\theta}$ is consistent for the parameter θ if the distribution of $\hat{\theta}$ converges to θ when the sample size goes to infinity. Example: Since $\mathbb{E}[\bar{X}_n]=\mu$ and $\text{Var}(\bar{X}_n)=\frac{\sigma^2}{n}\xrightarrow{n\to\infty}0$, then distribution of $\bar{X}_n$ converges to μ

Estimate Realization of an estimator given a sample of data. It is function of the data. Example: $\bar{x}_n$.

3.1.1 Common estimators

Estimator for the mean:

$$\bar{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$$

(Most common) Estimator for the variance [1]:

$$S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)^2$$

Estimator for the covariance:

$$\widehat{\text{Cov}}(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)\left(Y_i-\bar{Y}_n\right)$$

Estimator for the correlation:

$$\widehat{\text{Corr}}(X,Y)=\frac{\widehat{\text{Cov}}(X,Y)}{S_X S_Y}$$

3.1.2 Confidence intervals

Objective We want to find an interval $\left[\underline{L},\bar{L}\right]$ defined by the random variables $\underline{L}$ and $\bar{L}$ such that $\Pr\left(\theta\in\left[\underline{L},\bar{L}\right]\right)=1-\alpha$, where θ is the parameter of interest, $1-\alpha$ is the level of confidence we want and $\underline{L},\bar{L}$ are random variables for which their distributions are known (or at least their approximations).

Estimator for the mean

$$\left[\bar{X}_n-t_{\frac{\alpha}{2}}^{(n-1)}\frac{S}{\sqrt{n}},\bar{X}_n+t_{\frac{\alpha}{2}}^{(n-1)}\frac{S}{\sqrt{n}}\right]$$

Estimate for the mean

$$\left[\bar{x}_n-t_{\frac{\alpha}{2}}^{(n-1)}\frac{s}{\sqrt{n}},\bar{x}_n+t_{\frac{\alpha}{2}}^{(n-1)}\frac{s}{\sqrt{n}}\right]$$

[1] We divide by $(n-1)$ instead of n to make the estimator unbiased

Interpretation We have $1-\alpha$ confidence that the parameter μ belongs to the interval above. Note that, it is not true that $\Pr\left(\theta \in \left[\underline{l}, \bar{l}\right]\right) = 1-\alpha$. Since $\underline{l}$ and $\bar{l}$ are numbers, $\Pr\left(\theta \in \left[\underline{l}, \bar{l}\right]\right)$ is zero or one. This is true only when we have random variables $\underline{L}$ and $\bar{L}$.
It is important to realize that, if we want to have more confidence, with a fixed sample size, the length of the interval is going to increase, so we are less sure about the value of the parameter. There is a trade off between how sure we are about the level of the parameter, the length of the intervals, and how sure we are that the theoretical interval $\underline{L}$ and $\bar{L}$ contains the parameter.

3.2 Introduction to Linear Regression

As economists, we are frequently interested in the relationship between two economic variables X and Y, say $X =$ years of schooling and $Y =$ wages. First, we should investigate the summary statistics we reviewed in the first week, for example, mean, variance, and correlation. However, focusing on the summary statistics won't get us far in questions regarding prediction and causality. That's when we need an economic model. As a starting point, we will focus on a **linear model**; we model the relationship between X and Y as a line. Suppose we are given an sample $\{X_i, Y_i\}$ for $i = 1, \cdots, n$ of n observations, and we have the model

$$\underbrace{Y_i}_{\text{dependent variable}} = \underbrace{\beta_1}_{\text{intercept}} + \underbrace{\beta_2}_{\text{slope}} \times \underbrace{X_i}_{\text{independent variable}} + \underbrace{e_i}_{\text{error/residual term}}$$

where the e_i are the error terms. Here, we assume that this model is the **true** relationship between X and Y, and we are interested in the **parameters** β_1 and β_2. These parameters are **unknown**.

- Why do we need this model?
  - Econometric methodology examines and analyzes a **sample of data** from the **population**. After analyzing the data, we make statistical inferences. These are conclusions or judgments about a population based on the data analysis. Great care must be taken when drawing inferences. The inferences are conclusions about the particular population from which the data were collected.
  - **Prediction:** Predicting the value of one variable given the value of another, or others, is one of the primary uses of regression analysis.
  - **Causality:** A second primary use of regression analysis is to attribute, or relate, changes in one variable to changes in another variable.

3.2.1 Assumptions of the Simple Linear Regression Model

- **SR1 Econometric Model:** All data pairs (y_i, x_i) collected from a population satisfy the relationship

$$Y_i = \beta_1 + \beta_2 X_i + e_i,\ i = 1, ..., N$$

- **SR2 Zero mean error :** $\mathbb{E}[e_i] = 0$ for any i. This assumption is just a normalization.
- **SR3 Constant X :** $P(X_i = x_i) = 1$ for any i. It helps when one is learning regression for the first time, but it can be substituted by the exogeneity assumption $\mathbb{E}[e_i \mid X_i] = 0$ for any i.
- **SR4 Homoscedasticity:** $Var(e_i) = \sigma^2$ for any i. If X is not constant, this assumption becomes $Var(e_i \mid X = x) = \sigma^2$ for any x. Only important for the BLUE result and to simplify exposition of the material. In practice, no researcher assumes that this assumption holds
- **SR5 Uncorrelated errors:** $Cov(e_i, e_j) = 0$ for any $i \neq j$. Unobserved variables are uncorrelated for different entities i
- **SR6 Normal errors:** $e_i \sim \mathcal{N}(0, \sigma^2)$ for any i. Only to simplify exposition of the material. Versions of the CLT will guarantee that this is true assymptotically.

3.2.2 Interpretation of Parameters in Linear Models

How do we interpret β_2 in the model $Y_i = \beta_1 + \beta_2 X_i + e_i$? Note that under assumptions

$$E[Y|X = x] = \beta_1 + \beta_2 x$$

which implies

$$\beta_2 = \frac{\partial E[Y|X = x]}{\partial x}$$

That is, roughly, on average, one unit increase in X is associated with β_2 units change in Y.

3.3 Estimating the Model with OLS

One common method for estimating β_1 and β_2 is called Ordinary Least Squares, or **OLS**. This idea behind this method is to **minimize the sum of squared residuals**.
Let b_1 and b_2 be some arbitrary constants used as estimates for β_1 and β_2, respectively. We can rearrange the model given above to calculate the residual (or error term) for each observation, (Y_i, X_i).[2]

$$e_i = Y_i - (b_1 + b_2 X_i)$$

Summing up the squared error for each observation gives us the **sum of squared residuals/errors (SSE)**, conditional on our arbitrary choice of b_1 and b_2:

$$SSE(b_1, b_2) = \sum_{i=1}^{n} (Y_i - (b_1 + b_2 X_i))^2$$

Our job as econometricians is to find the best estimators of β_1 and β_2. Under the OLS methodology, this means we need b_1 and b_2 such that the SSE is **minimized**. As a mathematical convention, we denote the estimator with "hat" symbol.

$$(\hat{b}_1, \hat{b}_2) = \arg\min_{b_1, b_2} \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i)^2$$

This line should be read as "$\hat{b}_1$ and $\hat{b}_2$ are the estimators that minimizes the SSE among all possible values b_1 and b_2". ($\arg\min$ denotes the values/arguments that minimize the object.) So we know that

$$SSE(\hat{b}_1, \hat{b}_2) \leq SSE(b_1, b_2) \quad \text{for all } b_1, b_2$$

3.3.1 Explicit Formula of the Estimators

One can prove that

$$\begin{aligned} \hat{b}_1 &= \overline{Y} - \hat{b}_2 \overline{X} \\ \hat{b}_2 &= \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\text{Var}}(X)} \end{aligned}$$

where $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.
Try to prove this on your own (**hint:** start with the equation for $SSE(b_1, b_2)$ and take partial derivatives with respect to b_1 and b_2).

[2] Note that there will always be some error, unless the data points form an exact straight line. If that is the case, we can solve for β_1 and β_2 using any two points, and have no need for more complicated econometrics.

3.3.2 Some Terminology

- **Estimators** are formulas/functions of random variables, so they themselves are random. In the section above, we solved for $\hat{b}_1$ and $\hat{b}_2$, which are estimators for β_1 and β_2.
- **Estimates** are **realizations** of the estimator that depend on the data a researcher uses. For example, once you obtain some data $\{X_i = x_i, Y_i = y_i\}_{i=1}^n$, you can plug them into the formulas for $\hat{b}_1$ and $\hat{b}_2$ to get estimates - these values, unlike the estimator, are not random.
- The **fitted line** is given by $y = \hat{b}_1 + \hat{b}_2 x$.
- The **fitted value**, $\hat{Y}_i$ is given by $\hat{Y}_i = \hat{b}_1 + \hat{b}_2 X_i$.
- The difference between the true value, Y_i, and the fitted value, $\hat{Y}_i$, is the **fitted residual**

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - \underbrace{(\hat{b}_1 + \hat{b}_2 X_i)}_{\hat{Y}_i}$$

3.3.3 Estimators are BLUE

Gauss-Markov Theorem If the assumptions SR1 to SR5 are true, then estimators $\hat{b}_1$ and $\hat{b}_2$ are the best linear unbiased estimators. Best means that the variance of these estimators is smaller than any other linear and unbiased estimators. Linear comes from the fact that one can rewrite $\hat{b}_1$ and $\hat{b}_2$ as a linear function of $(y_1, \ldots, y_n)$

4 Extra Practice Problems

Problem 1

[3] An ice cream vendor at UCLA football games wants to know how many ice creams to stock before each game so that she doesn't run out, and isn't left over with too many unsold ice creams. She determines that weather is likely a large factor in the variation of ice creams sold, and estimates the following relationship between ice cream sales and the temperature based on 32 home games

$$\hat{y} = -240 + 8x,$$

where $\hat{y}$ is the predicted number of ice creams sold, and x is the temperature in degrees Fahrenheit.

(a) Interpret the estimated slope and intercept. Do the estimates make sense? Why, or why not?

Solution: -240 is the intercept; it is the estimate of ice creams sold when the temperature is zero degrees Fahrenheit. This doesn't make a ton of sense since we probably won't be operating the ice cream stand when the temperature is zero degrees!

8 is the slope term. In this context, on average, an increase of 1 degree (change in X) is associated with an increase of 8 in sales (change in Y).

(b) When the temperature on game day is predicted to be 80 degrees Fahrenheit, how many ice creams is the vendor expected to sell?

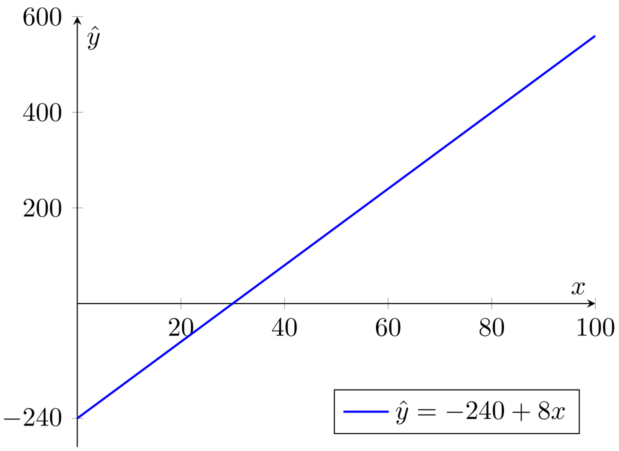
Solution: Plug in $x = 80$ to find that the vendor expects to sell $-240 + 8 \cdot 80 = 400$ ice creams.

(c) Is there a temperature at which she is not expected to sell any ice creams? If so, what is it?

Solution: Here, we want to solve for the temperature x such that we expect to have zero sales y. So $0 = -240 + 8x$ yields $x = 30$. So if it is thirty degrees outside, the vendor expects to make no sales.

(d) Sketch a graph of the estimated regression line.

Solution:



4.1 Practice Midterm 1

[4] Consider the following linear model based on the Cubic function:

$$wage = \beta_1 + \beta_2 \cdot educ3 + e,$$

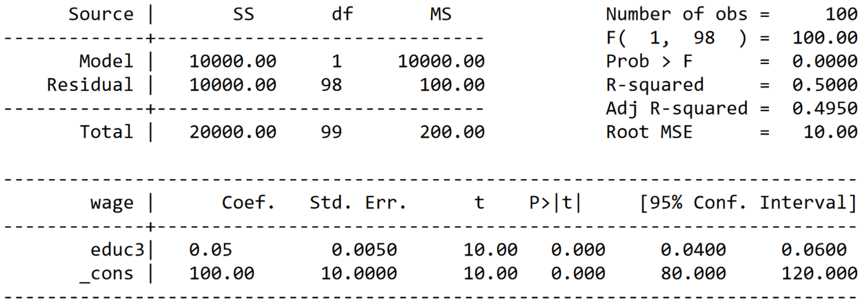
where

[3] This problem is taken, or modified, from the fourth edition of "Principles of Econometrics", by Hill, Griffiths, and Lim.

[4] These questions were taken from the practice midterm provided by Professor Pinto during Spring 2024

1. *wage* means daily wage measured in dollars.
2. *educ* means years of education, measured in years of schooling.
3. *educ*3 means the cubic of education, that is $(educ)^3$.

The STATA Output for this linear regression is given below:



```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,   98  ) =  100.00
       Model |   10000.00     1     10000.00           Prob > F      =  0.0000
    Residual |   10000.00    98       100.00           R-squared     =  0.5000
-------------+------------------------------           Adj R-squared =  0.4950
       Total |   20000.00    99       200.00           Root MSE      =   10.00

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ3|   0.05         0.0050      10.00   0.000     0.0400     0.0600
       _cons |   100.00      10.0000      10.00   0.000     80.000     120.000
------------------------------------------------------------------------------
```

Answer the following questions based on the regression output:
Question 1. What is the expected wage for a person with 10 years of education?

(a) 15

(b) 50

(c) 100

(d) 150

(e) 500

Solution: The estimated coefficients from the STATA output are $\hat{\beta}_1 = 100$ and $\hat{\beta}_2 = 0.05$. So, the expected wage for a person with ten years of education is given by

$$\begin{aligned}\mathbb{E}[Y|x_0 = 10] =& \hat{b}_1 + \hat{b}_2 \cdot x_0^3 \\ =& \hat{b}_1 + \hat{b}_2 \cdot 10^3 \\ =& 100 + 0.05 \cdot 10^3 \\ =& 150\end{aligned}$$

Question 2. What is the *marginal effect* of education for a person with ten years of education?

(a) 1.5

(b) 3

(c) 5

(d) 10

(e) 15

Solution: Here, we are asked about the *marginal effect* we need to take a derivative with respect to years of education:

$$\begin{aligned}\frac{\Delta \mathbb{E}[Y]}{\Delta x} =& 3 \cdot \hat{b}_2 \cdot x_0^2 \\ =& 3 \cdot 0.05 \cdot 10^2 \\ =& 15\end{aligned}$$

4.2 Question 15

. Regarding the Simple Regression Model $y = \beta_1 + \beta_2 \cdot x + e$, which of the following is FALSE?

(a) $\overline{y} - \hat{b}_1 - \hat{b}_2 \overline{x} = 0$, where $\overline{x}, \overline{y}$ denote sample means.

(b) The LS estimates for the quadratic regression $Y = \beta_1 + \beta_2 \cdot x^2 + e$ is not BLUE because the relation between Y and x is not linear, so the linearity assumption is violated.

(c) In the Simple Regression Model, $y = \beta_1 + \beta_2 \cdot x + e, \hat{b}_2 = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$, where $\text{Cov}(x, y)$ is the sample covariance and $\text{Var}(x)$ is the sample variance of x.

(d) Let $\hat{b}_1^*, \hat{b}_2^*$ be estimates for β_1, β_2 other than the least squares estimates $\hat{b}_1, \hat{b}_2$, then it must be that:

$$\sum_{i=1}^{N} \left(\hat{b}_1^* + \hat{b}_2^* x_i - y_i\right)^2 \geq \sum_{i=1}^{N} \left(\hat{b}_1 + \hat{b}_2 x_i - y_i\right)^2$$

Solution: The correct answer is (b). Options (a) and (c) are the standard equations that are used to estimate parameters in the simple regression model. Option (d) is correct because least squares estimates $\hat{b}_1, \hat{b}_2$ minimize the sum of squared errors, so the sum of squared errors must be larger when using different estimates $\hat{b}_1^*, \hat{b}_2^*$ than when using least squares estimates $\hat{b}_1, \hat{b}_2$. Option (b) is false because the equation is quadratic in x but is linear in x^2, so $Y = \beta_1 + \beta_2 \cdot x^2 + e$ is BLUE.