Econometrics for Climate Change

Session 2 — Linear Regression & Applications to Climate Change

John Gómez-Mahecha

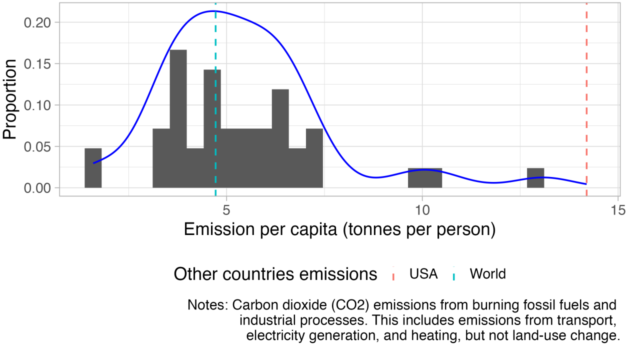
2026-03-03

An example of a mean test

Question 1 Are the mean emissions per capita across European countries less than the per capita emissions from the US? Is it different from the world average?

Data: Emission per capita for each country for 2024.

- Europe: Avr: 5.36, Var=4.56, n=42
- US Avr: 14.2
- World Avr: 4.72



Two or more variables

What we have done?

- Until now, we have only seen univariate probability functions - that is, one X at a time.
- We check what are possible forms of characterizing the information of those univariate probability functions: Mean, Variance, etc..
- But what we are really interested is on the relation between variables. We want to know if there are dependencies between variables
  - Does GDP depends on weather?
  - Does labor productivity depends on heat?

How do you state/describe the relation between two variables?

Relation between two variables - Linear regression

Law of Iterated Expectations

$$E[Y] = E[E[Y|X]]$$

- Y usually known as **dependendent variable**, **outcome variable** or **regressand**
- X usually known as **independent variable** or **regressor**
- **Independence**: $E[Y|X] = E[Y]$
- $E[Y|X]$ is the best predictor of Y
- This allow us to write Y as
$$Y = E[Y|X] + u$$
  with $E[u] = 0$
- u is known as the **error**.

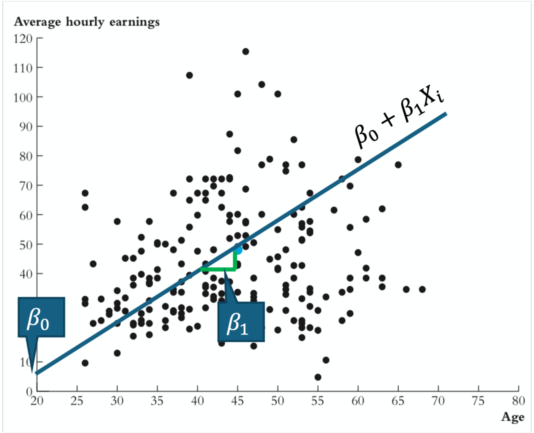
Conditional Expectations

The following table gives the joint probability distribution between employment status and college graduation among those either employed or looking for work (unemployed) in the working-age population of South Africa.

	Unemployed ($Y = 0$)	Employed ($Y = 1$)	Total
Non–college grads ($X = 0$)	0.078	0.673	0.751
College grads ($X = 1$)	0.042	0.207	0.249
Total	0.12	0.88	1.000

- What is the probability of being employed?
- What is the probability of being employed given that I am a college graduate?
- What is the $E[Y]$?
- What is the $E[Y|X = 1]$ and the $E[Y|X = 0]$?
- Is Y independent of X?

Linear Regression Model

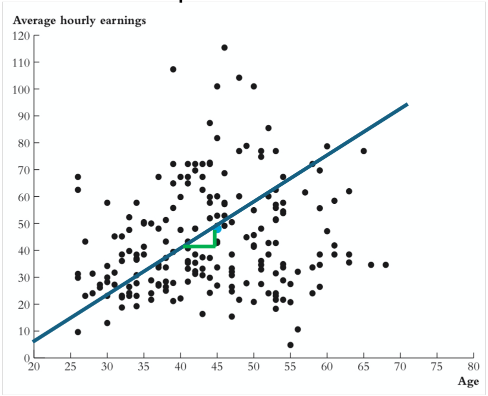


- Each dot correspond to a pair of (Y_i, X_i).
- Linear regression model takes advantage that exist a line $\beta_0 + \beta_1 \cdot X_i$ so that $E[u_i] = 0$
- Here it is easy to see the interpretation of the β_1 coefficient. Notice that:

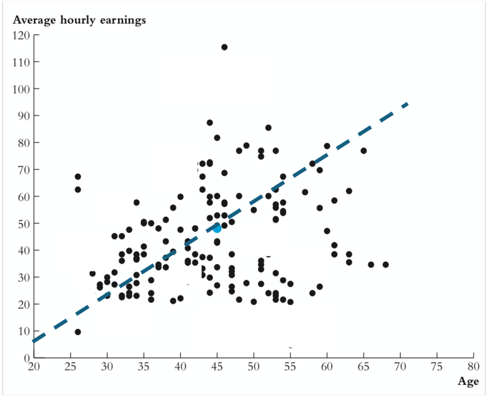
$$\beta_1 = \frac{d\ E[Y|X_1]}{d\ X_1}$$

From the Population to the Sample

Population



Sample (size n)



One possible estimator: Ordinary Least Square (OLS)

- The most known estimator, but not the only one.
- Our main interest is to get a line as close as possible to the linear regression line.
- The proposed line will have the form

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- $\hat{Y}_i$ is usually known as the **prediction** of Y_i
- The difference between the predicted value and the real prediction is usually known as the **residual** $(\hat{u}_i)$

$$\hat{u}_i = Y_i - \hat{Y}_i$$

OLS estimator

Concept: Ordinary Least Square Estimators for the Linear Regression

The estimators ($\hat{\beta}_0$,$\hat{\beta}_1$) are the OLS estimator if they are the values that solve for

$$Min_{\{\hat{\beta}_j\}} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

- A usual optimization problem
- Notice that we have not talked about their statistical features
- Can be extended to more independent variables?

OLS derivation

$$Min_{\{\hat{\beta}_i\}} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Is it a good estimation?

- How we check it is a good way to represent Y_i?

Is it a good estimation?

- How we check it is a good way to represent Y_i?
  - Sum Square of the Residuals (SSR)

$$SSR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Is it a good estimation?

- How we check it is a good way to represent Y_i?
  - Sum Square of the Residuals (SSR)

$$SSR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  - Root mean Square of the Residuals (RMSR)

$$RMSR = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

Is it a good estimation?

- How we check it is a good way to represent Y_i?
  - Sum Square of the Residuals (SSR)

$$SSR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  - Root mean Square of the Residuals (RMSR)

$$RMSR = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

  - Adjustment coefficient R^2

$$R^2 = 1 - \frac{MSR}{MST}$$

where $MST = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2$

Are the $\hat{\beta}$s a good estimator for the βs?

- Do you remember what it meant to be a good estimator?
- **Gauss Markov Theorem**

1. No multicollinearity
2. Homoskedasticity
3. iid sample
4. Linear model
5. Exogeneity (Independence of X from u)

Then the $\hat{\beta}$s, will be unbiased, efficient among linear estimators, consistent, and asymptotically distributed Normal (if n is big enough).

How to use this for <u>statistical inference?</u>

Can OLS be used for any type of variables?

- The question is not "can" but if "it make sense"
- X could be continuous or binary without problem
- If X is discrete, it requires further manipulation but it is still possible
  - For example if $X \in \{1, 2, 3\}$
  - Will create indicators variables: $X_i\{X == i\}$
  - Include all the indicators <u>except for one</u>
- If Y is continuous, you are OK
- If Y is discrete, but dense, you are OK
- If Y is discrete, but take only few possible numbers, maybe you will be better with other type of models.

Does it has to be always linear?

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 \tag{1}$$

Does it has to be always linear?

$$ln(Y_i) = \beta_0 + \beta_1 \cdot ln(X_i) \tag{1}$$

Does it has to be always linear?

$$ln(Y_i) = \beta_0 + \beta_1 \cdot ln(X_i) \tag{1}$$

Model	Dependent Variable	Independent Variable	Interpretation of β_1
level-level	y	x	$\Delta y = \beta_1 \Delta x$
level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

OLS and mean differences?

Assume that X is binary variables, taking only as values 1 or 0

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{2}$$

- It can be shown that $\beta_1 = E[Y_i|X_i = 1] - E[Y_i|X_i = 0]$
- Highly used in modern econometrics

Does it always have to be only one X?

We can think of models of the form:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + u_i \tag{3}$$

And estimate a prediction of the form

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + ... + \hat{\beta}_k X_{i,k} \tag{4}$$

- Not include variables that are linear combinations of the others.
- OLS estimators are not longer represented as sums, but in matrix form.
- Why more variables?

OLS under the violation of assumption

Concept: Gauss Markov Assumptions

1. No multicollinearity
2. Homoskedasticity
3. iid sample
4. Linear model
5. Exogeneity (Independence of X from u)

Omitted Variable bias

Lets suppose you estimate this:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1}$$

But the real population model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + u_i$$

It can be shown that

$$E[\beta_1] = \beta_1 + \beta_2 \cdot corr(Y, X_2)$$