# Linear Regression with R/Stata

## Day 1: Recap and Basics

Constantin Manuel Bosancianu

February 25, 2019

Wissenschaftszentrum Berlin
*Institutions and Political Inequality*
bosancianu@icloud.com

# Preamble

# Basic setup for course

Lecture (slide-based) + lab (code-based) + consultations.

Pace: fairly slow, but we can customize to your needs.

Difficulty: light, but we can adapt as we go here as well.

Credit system: 2 = attendance and readings; 2+1 = attendance, readings, and short class assignment; 2+2 = all of the above, plus final take-home assignment.

# Outline

- ✓ Basic notation in statistics;
- ✓ Basic concepts: mean, variance, covariance, correlation, *t*-test;
- ✓ (it might extend slightly into the laboratory session): Introduction to simple linear regression: fitting a line to data.

The first two I'll keep short, but if you feel we went too fast through them, let's talk about this more during the consultation sessions.

# Notation and basic concepts

# Why notation

You will encounter it in a lot of quantitative literature, so it's good to get familiar with it early.

Some statistical topics you will have to learn on your own, so it's best if you get used with the symbols which many books use.

It slows you down a bit in the short term, but makes things faster in the long term.

# Building blocks I

Capital letters refer to variables ($X$, $Y$). Small letters refer to specific observations ($x_i$ is the $i$th observation on variable $X$).

Sample size: $n$.

Sum of elements: $\sum_{i=1}^{n}$ (read: “sum of all elements from 1 to n”).

What would $\frac{\sum_{i=1}^{n} x_i}{n}$ do, then?

I denote the mean by $\bar{x}$.

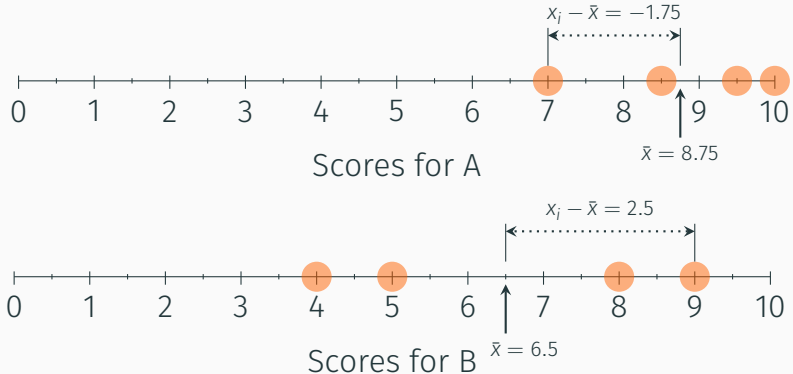
# Building blocks II

Variance: a measure of how spread out observations on a variable are, from the mean of the variable. Denoted by $\sigma^2$.

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{1}$$

- ✓ Why are the elements in the numerator squared before adding them up?[1]
- ✓ Why is the denominator $n-1$ instead of $n$, as for the mean?

# Example




Scores for 2 countries, A and B, on 4 dimensions of democratization.

# Example

$$\sigma_A^2 = \frac{(-1.75)^2 + (-0.25)^2 + (0.75)^2 + (1.25)^2}{4-1} = \frac{5.25}{3} = 1.75 \tag{2}$$

$$\sigma_B^2 = \frac{(-2.5)^2 + (-1.5)^2 + (1.5)^2 + (2.5)^2}{4-1} = \frac{17}{3} = 5.667 \tag{3}$$

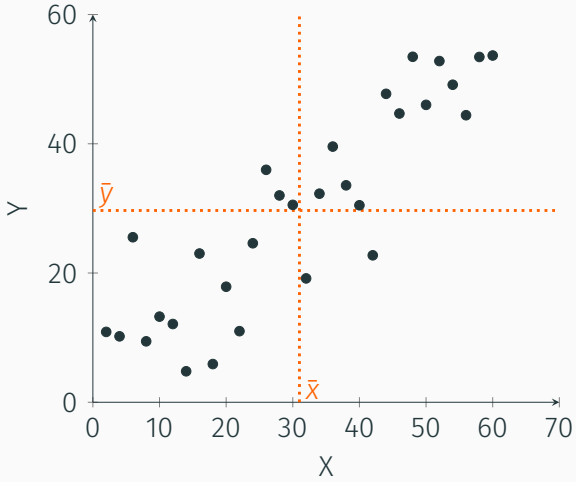
From variance, a quick derivate measure is the standard deviation: $\sigma_X = \sqrt{\sigma_X^2}$.

# Building blocks III

Covariance: measure of association between two variables. It describes how they vary together—when observation $j$ has a high value on $X$, how is its value on $Y$?

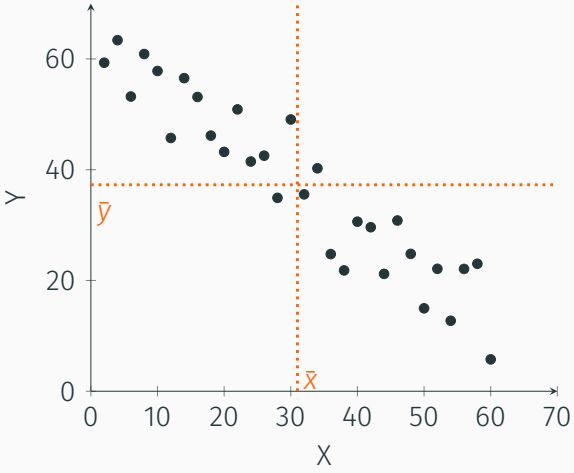
$$cov(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{4}$$

# Example



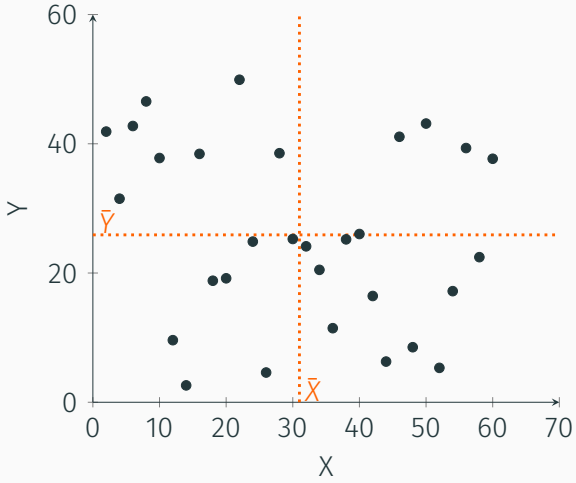

Relationship between two variables (I). Covariance is 250.445.

# Example




Relationship between two variables (II). Covariance is -262.678.

# Example




Relationship between two variables (III). Covariance is -50.273.

# Building blocks IV

Covariance is an imperfect measure, though, because it's sensitive to the scale of measurement.

IQ and income: from 0 to 60,000 EUR, or from 0 to 60 (in 1,000s of EUR).

Correlation tackles this problem by standardizing covariance:

$$r_{XY} = \frac{cov(X, Y)}{\sigma_x \sigma_y} = \frac{\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sigma_x \sigma_y} \tag{5}$$

# Building blocks V

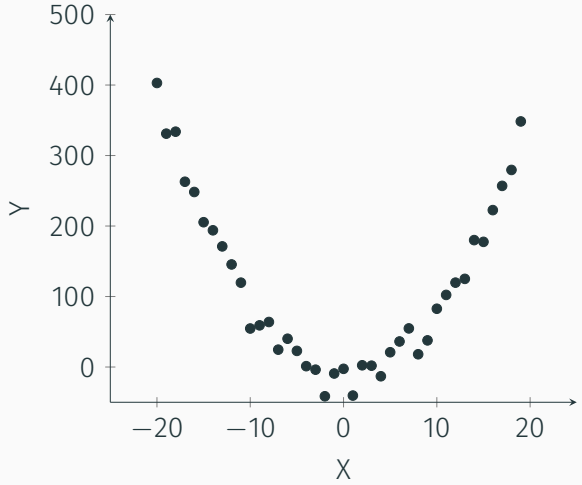
Characteristics of correlation:

- ✓ always ranges between -1 and 1;
- ✓ 0 indicates lack of any association between *X* and *Y*;
- ✓ indicates strength of relationship, as well as direction (negative vs. positive);

Limitations:

- ✓ requires continuous variables (Pearson's *r*);
- ✓ can only capture linear relationships.

# Example




Curvilinear relationship. Pearson's *r* is -0.158.

# Building blocks VI

Just as a variable $X$ can have a standard deviation, so can the mean of $X$.

Such a quantity would tell us how spread out $\bar{x}$ would be, if we took repeated samples of size $n$ from the population, and computed $\bar{x}$ again and again.

This quantity is called a standard error (SE).

$$SE_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \tag{6}$$

# Building blocks VII

A *t*-test is a way of comparing two means, to check whether they are statistically significantly different from each other.

- ✓ one-sample *t*-test: a mean of a sample is compared with the population mean;
- ✓ two-sample *t*-test: means of two samples are compared to each other;

Take the first type (the sample is $x$ and the population mean is $\mu$):

$$t = \frac{\bar{x} - \mu}{SE_{\bar{x}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma_x^2}{n}}} \tag{7}$$

# Building blocks VIII

This $t$ value has a well-behaved $t$ distribution (*https://rpsychologist.com/d3/tdist/*), with $n - 1$ degrees of freedom ($df$).

If the $t$ value surpasses the critical value for these degrees of freedom, at a pre-specified level of confidence, then the test is statistically significant.

The mean in the sample is statistically significantly different than the population mean.

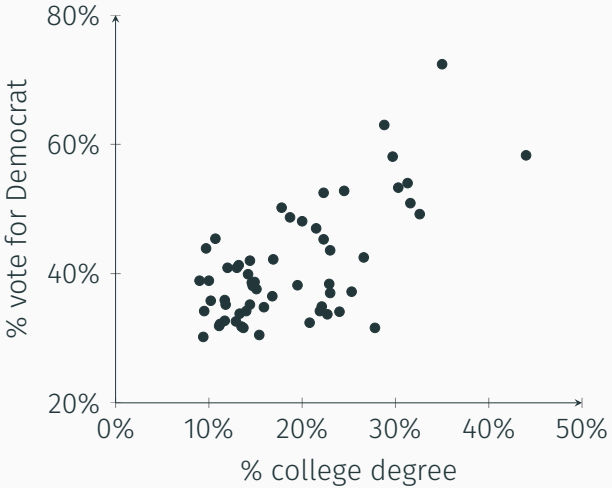
# Simple linear regression: intro

# Simple linear regression: benefits

Like correlation or covariance, it describes the relationship between two variables. It goes far beyond this, though:

- ✓ gives us very precise details regarding how the two variables are associated;
- ✓ allows us to quantify uncertainty about this association as well;
- ✓ allows us to make predictions about $Y$ for any level of $X$ we might want;
- ✓ produces a measure of how well we're describing $Y$ with $X$.

# Example: California counties in 1992




Relationship between education and vote choice.

# Example: California counties in 1992

A line would appear to fit the relationship between the two variables.

We need two pieces of information to uniquely identify the line:

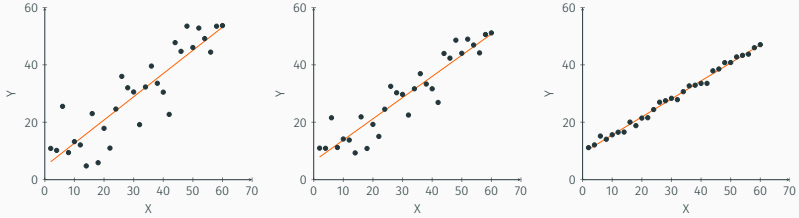
- ✓ the point at which it intersects $Y$;
- ✓ the slope of the line.

$$Y = a + bX \tag{8}$$

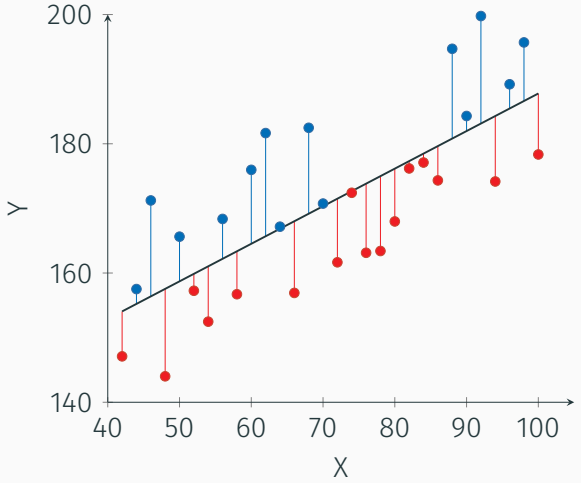
# The role of residuals

The relationship isn't perfect, though, so we need one more element: the error term.

$$Y = a + bX + e \tag{9}$$



# Residuals




Residuals (blue is positive, red is negative)

# How to construct the line?

There are numerous ways of thinking about this.

One approach is to choose the line that minimizes the total size of the (absolute values of) residuals.

$$\sum_{i=1}^{n} |e_i| = |e_1| + |e_2| + \cdots + |e_n| \tag{10}$$

Why not just add up the residuals? (think back to the way variance was constructed out of deviances).

# How to construct the line?

In fact, we will be minimizing the total size of the *squared* residuals (this is also called the *sum of squared errors* (SSE) in some texts).

$$\sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \cdots + e_n^2 \tag{11}$$

This is what gives OLS its name: *ordinary least squares*.

# Terminology

In covariance or correlation, we didn't make a big distinction between $X$ and $Y$.

In regression, we do:

- ✓ $Y$ is called: outcome, response variable, or dependent variable;
- ✓ $X$ is called: predictor, or independent variable.

In this understanding, there is a causal relationship between $X$ and $Y$, although regression can only offer some clues about the precise direction of causality.

# Coefficients

$$Y = a + bX + e \tag{12}$$

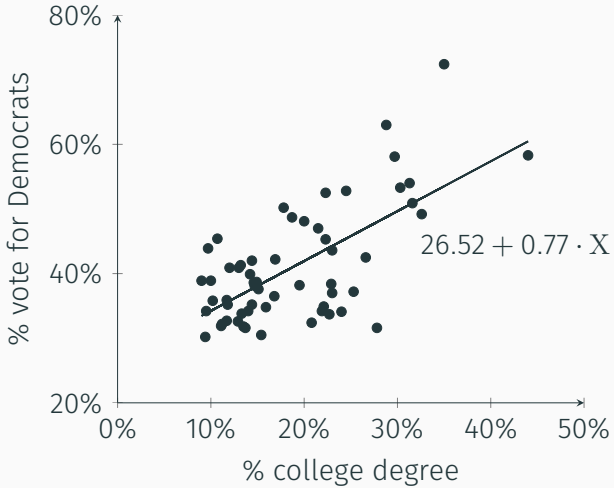
$a$ = intercept. The value of $Y$ when $X$ is 0.

$b$ = slope. The *change* in $Y$ when $X$ increases by 1-unit.

For the simple regression case:

$$\begin{aligned} b &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ a &= \bar{y} - b\bar{x} \end{aligned} \tag{13}$$

# California counties in 1992




OLS estimates: education and vote choice (California 1992)

# Interpretation for California 1992

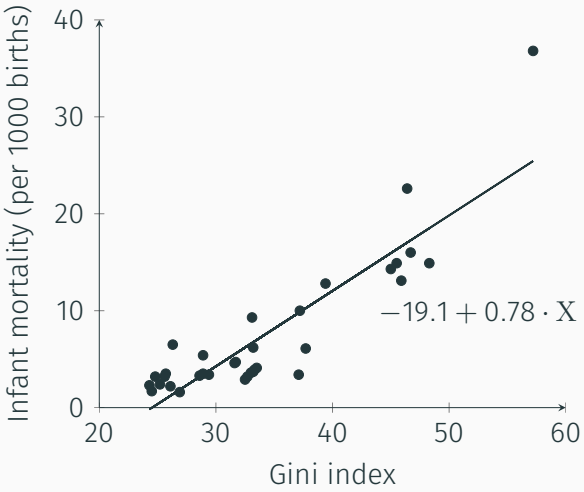
*a*—"baseline" value of % Democratic vote for 0% college educated in the county.

26.5% for Democrats for 0% college educated.

*b*—increase in % vote for Democratic candidate when % college educated increases by 1.

0.77 percentage points increase in % Democratic vote for 1 point increase in % college educated.

# Interpretation II




OLS estimates: income inequality and infant mortality

# Interpretation II

-19.1—infant mortality level for a 0 value for Gini (perfect equality). Because the intercept can sometimes have these absurd interpretations, it is occasionally ignored.

0.78—increase in number of infants who die associated with a 1-point increase in Gini.

If it feels odd to interpret a negative *a* in this case, you can easily correct this: subtract, say, 30 points from each observation's value on Gini.

It's called rescaling; it won't impact the value for *b*.

Thank you for the kind attention!