GOV1368 Section 2

Mohit Karnani

Harvard

Fall 2024

Agenda

Introduction

Linear Regression

Application

Contents

Introduction

Linear Regression

Application

Recap

Last time we covered:

- The importance of statistics in social science
- Data and random variables
- Descriptive and inferential statistics
- The role of causal inference
- An application using real data

Recap

Last time we covered:

- The importance of statistics in social science
- Data and random variables
- Descriptive and inferential statistics
- The role of causal inference
- An application using real data

Today we will learn about a powerful tool related to all of the above: **Linear Regression**.

Recap

Last time we covered:

- The importance of statistics in social science
- Data and random variables
- Descriptive and inferential statistics
- The role of causal inference
- An application using real data

Today we will learn about a powerful tool related to all of the above: **Linear Regression**.

Have you heard of it before? What do you know about it?

Contents

Introduction

Linear Regression

Application

Modeling the Relationship between Variables

Suppose we have two random variables, X and Y.

For example, X_i could be the number of hours student i studied for an exam, and Y_i could be student i's exam score.

Call X the independent variable (or regressor) and Y the dependent variable (or outcome).

We want to *model* the relationship between them.

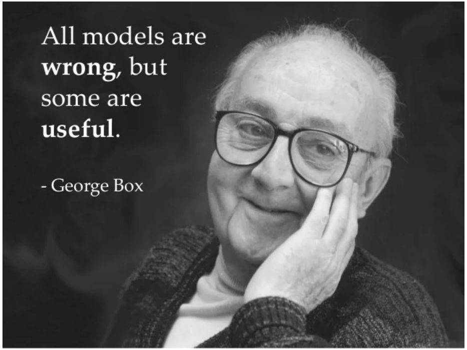
Modeling the Relationship between Variables

Suppose we have two random variables, X and Y.

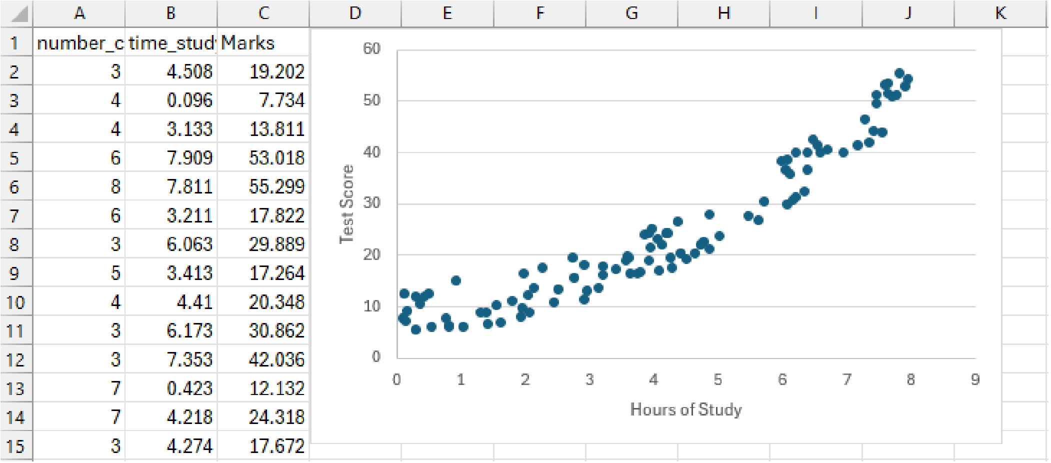
For example, X_i could be the number of hours student i studied for an exam, and Y_i could be student i's exam score.

Call X the independent variable (or regressor) and Y the dependent variable (or outcome).

We want to *model* the relationship between them.



Hours of Study and Test Scores



	A	B	C
1	number_c	time_stud	Marks
2	3	4.508	19.202
3	4	0.096	7.734
4	4	3.133	13.811
5	6	7.909	53.018
6	8	7.811	55.299
7	6	3.211	17.822
8	3	6.063	29.889
9	5	3.413	17.264
10	4	4.41	20.348
11	3	6.173	30.862
12	3	7.353	42.036
13	7	0.423	12.132
14	7	4.218	24.318
15	3	4.274	17.672

Simple Linear Regression Model

We can do this using a *linear regression* model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

We will assume only two things for now (you might find other assumptions elsewhere):

1. The relationship is linear in parameters. (for the method to work)
2. The errors are independent. (for the relationship to be causal)

These assumptions will be "hand-wavy" for now, but we will cover them in detail later. Some comments:

- Linear regression is an extremely powerful tool, but it is also very sensitive to violations of these assumptions. They seldom hold in non-experimental settings.
- It is by far the most used method in social science; people build careers on it.
- Practitioners often fail to understand the assumptions and limitations of the model.

Fitting the Model to the Data

The linear regression model is a mathematical abstraction, but we can pair it with data!

Intuitively, we want to find the "best" linear function $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ picking "optimal" values for $\hat{\beta}_0$ and $\hat{\beta}_1$ to fit a line representing our cloud of data points.

The terms "best" and "optimal" here refer to *minimizing the sum of squared errors* made by the linear regression model: $\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$ (no need to learn this)

The process of finding the best coefficients is called *estimation*: we are trying to estimate the "true" (population) parameters in the model by fitting a line using the data (sample).

But what do these estimated coefficients mean?

Interpreting the Coefficients

In general, we can interpret the coefficient paired to the regressor as an "association".

For example, if we estimate $\hat{\beta}_1 = 5$, this can be interpreted as "an hour of study is associated to an average increase of 5 points in the test score".

It's not really a correlation: $\hat{\beta}_1 \approx \dfrac{\text{cov}(X, Y)}{\sigma_X^2} = \text{corr}(X, Y) \cdot \dfrac{\sigma_Y}{\sigma_X}$.

The intercept coefficient is a "baseline" measure of the expected outcome when the regressor is 0. So if we estimate $\hat{\beta}_0 = 10$, that means an average student would score 10 points with zero hours of study.

Importantly, in general, these coefficients show **association, not causation**.

It is very hard for our second assumption to hold when using observational data, and without it we cannot claim any sort of causality. Next week we will learn more about this.

Multiple Linear Regression

A variation of the *simple* linear regression model is the *multiple* linear regression model.

In the multiple linear regression model we still have 1 dependent variable, but now we can have *multiple* independent variables. For example, we can augment our model of the relationship between hours of study and test scores to also incorporate the number of courses that each student is taking:

$$Score_i = \beta_0 + \beta_1 Hours_i + \beta_2 Courses_i + \varepsilon_i$$

Instead of fitting a *line*, now we are fitting a *plane*. Things can get hard to visualize...

Our estimates can be interpreted as $\hat{\beta}_1$ being the average change in test scores associated with studying one more hour, and $\hat{\beta}_2$ being the average change in test scores associated with taking one more course. Nothing causal!

Contents

Introduction

Linear Regression

Application

Application: NAEP trends and Scores vs Study Hours, Courses Taken

`Stata`

Empirical Methods in Education Reading Group (EMERG)

bit.ly/eme-rg