Lecture 2

What is a Regression?

Ananya Iyengar

January 2026

St. Stephen's College

This corresponds to Chapter 1 of Studenmund (2017).

Fundamentally, we want to characterise the relationship between some variable of interest and other variables that are related to it. Some examples: schooling and wages, cash transfers and asset ownership, income and spending, GDP and inequality. (*Can you think of some?*) Colloquially, we call the variable of interest the *dependent variable* and the related variables the *independent* variable. Don't interpret this notation as causal, it is **correlational**! Think of defining $y = f(x)$, which can be equivalently defined as some $x = g(y)$. The choice of *what* is independent and what is dependent come from economic intuition, not from the math!

1 The Regression

1. One main variable of focus (the "dependent variable" Y) and one related variable (the "independent" variable X).
2. The simple linear regression expresses Y as an affine function of X.

$$Y = \beta_0 + \beta_1 X \tag{1}$$

   Here, β_0 is the intercept and β_1 is the slope.

3. When we say *linear*, we mean *linear in parameters* β. Then, the slope coefficient gives the relationship between X and Y.

4. A regression is essentially calculating a mean – finding some average relationship. This deterministic relationship is called the **Conditional Expectation Function** $E[Y/X]$. For example, wages by gender.

5. Remember, reality does not mimic models! $\implies$ ϵ, the **stochastic error term**. Stochastic = random – it is here that thinking about the model as an experiment and observed data as realisations of a random experiment comes in!

6. The regression equation is given by:

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

   Here, ϵ_i is the stochastic component and $\beta_0 + \beta_1 X_i$ is the deterministic component.

7. We need ϵ because (1) omitted variables; (2) measurement error; (3) different functional form of true relationship; (4) chance.

8. Above, i is used to denote each realisation in the sample i.e. $i \in \{1, ..., n\}$. Here, n is called the sample size.

9. *Aside:* The regression is essentially an orthogonal projection of Y on the space defined by the independent variables!

10. In the real world, we do not know β. We only observe imperfect measures of X and Y. Therefore, there is some true relationship in the population which we do not know but assume some functional form of. We aim to ***estimate*** the parameters of (2) (i.e. the regression coefficients) to quantify the relationship between X and Y.

11. We will look at estimation in the next chapter, but introduce notation now. The estimated parameter is denoted by $\hat{\beta}$. This is also called a *fitted value*. Then $\hat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i$ is our *estimate of* $E[Y/X]$!

12. The error term $\epsilon = Y - E[Y/X]$ in the population and $e_i = Y_i - \hat{Y_i}$ in the sample.

13. Interpreting the coefficient β_1: A one unit change in X is *associated* with, on an *average*, a β_1 unit change in Y.

14. All of the above can be extended to more than one X variable. Such a model is called the

multiple linear regression. The idea of *controls* is important here!

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

Recall out discussion of replacing experimental controls with statistical controls – it is as if we are keeping other things constant and isolating the relationship of each X_j with Y. This is called the *ceteris paribus* assumption in economic modelling.

The question that remains is: how do we estimate this regression equation? That is the topic of our next discussion.

For the tutorial: Attempt from the back of the chapter: Q1, Q3, Q6, Q7