The Regression Model

Recall from the first lecture that **econometrics** is the unification of economic theory and statistical methodology.

What we try and do is quantify a model that describes the underlying structure of the behaviour of economic variables. We do this using **regression analysis**.

Regression analysis is a statistical technique used to explain variation in one variable, called the **dependent/endogenous** variable (the $\boldsymbol{Y}$) as a function of the variation in a set of other variables, called the **independent/explanatory/exogenous** variables (the $\boldsymbol{X}$s).

Let's consider a **linear model** with only one explanatory variable, X:

$$Y = \beta_0 + \beta_1 X.$$

This is an equation of a straight line.

We can graph this equation very easily:

When we use the word **linear** we mean linear in the coefficients, not X. Thus we will consider models like $Y = \beta_0 + \beta_1 X^2$ but not like $Y = \beta_0 + X^{\beta_1}$.

The variation in Y will not be perfectly explained by the variation in X. There are probably other factors that affect the variation in Y.

Like what?

- Omitted variables
- Nonlinearities
- Measurement errors
- Unpredictable effects

These factors are captured by a **stochastic error term** (or **disturbance term**) denoted by ε:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

This model is referred to as a **simple linear regression model**.

The term $\beta_0 + \beta_1 X$ is the **deterministic part** of the model and it is the expected value of Y given X, or the **conditional mean of *Y* given *X***, that is

$$E(Y|X) = \beta_0 + \beta_1 X,$$

when $E(\varepsilon|X) = 0$. This is also called the **population regression function**.

The error term ε is the **stochastic** or **random part** of the model and is the difference between Y and the deterministic part $\beta_0 + \beta_1 X$. Think of it as a combination of the four factors discussed above.

Example

Let's think about final grades (out of 100) for BUEC 333 students (Y). Some of the variation is predictable and some is not.

We can extend the model to allow for more X variables to affect Y

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon.$$

This model is referred to as a **multiple linear regression model**.

The coefficients associated with the X variables can be interpreted as **partial derivatives**:

We can also think of the regression model in terms of the individual observations:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \qquad i = 1,2,\dots,n$$

Now, as we know, it is unrealistic to conduct a complete census of the population in order to derive the **population regression function**. So, we do not know: $\beta_0, \beta_1, \dots, \beta_k, \varepsilon_i$.

In practice, we use a sample of data. This gives us: $Y_i, X_{1i}, X_{2i}, \dots, X_{ki}$

Given this sample, we derive an **estimated regression equation**. Our goal is to estimate the unknown coefficients and errors.

For the linear regression model with k explanatory variables we have

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i,$$

we cannot observe β_0 and $\beta_1, \dots, \beta_k$, we must obtain estimates of them, say $\hat{\beta}_0$ and $\hat{\beta}_1, \dots, \hat{\beta}_k$ using our sample.

The sample regression equation then becomes

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \ + \hat{\beta}_k X_{ki}$$

$\hat{Y}_i$ is the **estimated** (or **fitted** or **predicted) value** of Y_i, it is our prediction of $E(Y_i|X_i)$. The $\hat{\beta}_0$ and $\hat{\beta}_1, \dots, \hat{\beta}_k$ coefficients are the **estimated regression coefficients**.

The **residual** is defined as the observed value of Y minus the predicted value of Y:

$$e_i = Y_i - \hat{Y}_i.$$

In contrast, the error is defined

$$\varepsilon_i = Y_i - E(Y_i|X_i).$$

Notice that the model error ε_i can never be observed.