8 Regression

8.1 The simple linear regression model

Often we want to investigate the relationship between two variables beyond just finding the correlation coefficient between them. We may want to ask 'how does Y vary with changes in X?' We can answer this question using *simple linear regressions*.

Let us think of the two variables as **wages** (Y) and **education** (X). We want to ask how wages change with the change in the years of education a person has undergone. Before we start to analyse the data, we write down our assumption of what the nature of this relationship can be. For now, we will work with the simplest possibility - a linear relationship. We write this relationship down as a *regression model*.

$$Y = \alpha + \beta X$$

Here the coefficient β is the slope of the line, and the coefficient α is the y-intercept. [Think of the familiar equation of a line $y = mx + c$. In this case β represents m and α represents c.] This can be interpreted as saying that an increase in education by one year leads to an increase in wage by β units.

This is called a deterministic relationship. For a given value of education X, there can be only one passible value of wages Y. But in reality, wages are affected by many factors other than education. Hence, two people with the same level of education can have different wages. To reflect this, we add an *error term* u to the regression model.

$$Y = \alpha + \beta X + u$$

Observation	Wage	Years of education
1	y_1	x_1
2	y_2	x_2
.	.	.
.	.	.
.	.	.
n	y_n	x_n

Table 1: Representation of a dataset of two variables

Now we look at the data. Let us think of a survey of n individuals recording their wages and years of education. Hence the data may be represented as Table 1.

We can visualise this data as a scatter plot. Education, which is the *explanatory variable* or the *independent variable*, will be on the horizontal axis, and wage, which is the *dependent variable*, will be on the vertical axis.

Since, the relationship described by our model is linear, a best-fit line drawn using the points on the scatter plot should be a good estimate of what the actual relationship is. The best fit line can also be described using a slope coefficient and a y-intercept. Let us call these $\hat{\beta}$ and $\hat{\alpha}$ respectively.

8.2 OLS estimation

There are many ways in which we can draw a best-fit line. One way is to take a candidate line and measure the *vertical* distance of each point from that line. This distance is called the residual $\hat{u}_i$. Hence we can represent the data as

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{u}_i$$

Here, $\hat{\alpha} + \hat{\beta} x_i$ represents the *predicted values* of y that fall on the best-fit line.

Let us represent these values as $\hat{y}_i$. Therefore, the residuals are the difference between the actual value y_i and the corresponding predicted value on the line $\hat{y}_i$.

$$\hat{u}_i = y_i - \hat{y}_i$$

It is clear that for a given dataset, i.e. given values of $x_1, x_2, .., x_n$ and $y_1, y_2, ..., y_n$, different candidate lines represented by different values of $\hat{\beta}$ and $\hat{\alpha}$ will give different values of residuals $\hat{u}_1, \hat{u}_2, ..., \hat{u}_n$. The best-fit line would be one that, in some sense, *minimises* these residuals. We cannot just add up the residuals as some of them will be negative and some positive. Hence, we square and add them up. We can now define the best-fit line as one that minimises the sum of squared residuals (SSR). Hence, the problem is to choose $\hat{\beta}$ and $\hat{\alpha}$ such that the SSR $\sum_{i=1}^{n} \hat{u}_i^2$ is minimised.

$$\min_{\hat{\beta},\hat{\alpha}} \sum_{i=1}^{n} \hat{u}_i^2$$

Or,

$$\min_{\hat{\beta},\hat{\alpha}} \sum_{i=1}^{n} (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2$$

We can solve this and find the values of $\hat{\beta}$ and $\hat{\alpha}$ that minimise this expression. The results we get are -

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

This method of finding the best-fit line is called Ordinary Least Squares (OLS).

Since the survey is just a sample, these are sample statistics. If we did another survey, i.e. took another sample, the values of $\hat{\beta}$ and $\hat{\alpha}$ may turn out to be different. Thus, $\hat{\beta}$ and $\hat{\alpha}$, are the estimators for the actual β and α respectively. These estimators are called OLS estimators.

Now we need to see under what conditions the OLS estimator will have the desirable properties of estimators that we had listed earlier - unbiasedness, efficiency and consistency. We will only deal with unbiasedness now.

8.3 Unbiasedness of OLS estimator

Before thinking about unbiasedness, let us refine our regression model. As we mentioned earlier, the error term u represents all the other factors that affect the dependent variable Y. By this definition, u can be positive or negative. Returning to the wage-education example, we can think of u as representing innate ability. For the same level of education, wage is higher for people with higher ability.

Now, ability can be thought of as a variable with positive values. But in the regression model, we will represent u as the *deviation from average ability*. The effect of average ability will be incorporated in α. Hence, we think of every individual's ability as the average ability plus or minus a deviation from the average. This deviation is incorporated in the error term. This allows us to assume that the mean of the error term will always be zero.

$$E[u] = 0$$

This is consistent with the way we find the best fit line - the deviations from the best fit lines - the residuals - are positive and negative and their mean is zero.

Now, to think of unbiasedness. Imagine that in reality there is no actual effect of education on wages, i.e. $\beta = 0$. But wages do depend on ability. Also, it may be the case that people's ability determines their level of education - people with low ability drop out after school while people with high ability go on to do post-graduate studies. Since, ability affects wages, people with high ability will earn more than people with low ability. But if we conduct a regression of wage on education, we will find that people with high education get higher wages than people with low education i.e. $\hat{\beta} > 0$. In this case $\hat{\beta}$ is not an unbiased estimator

of β.

$$E[\hat{\beta}] \neq \beta$$

The condition for unbiasedness is that the unobserved factors that we have left out, that form the error term u, should be independent of the explanatory variable X.

$$Cov(u, X) = 0 \Rightarrow E[\hat{\beta}] = \beta$$

In other words, for a given value of X, the expected value of the error term u, should be the same as everywhere else.

$$E[u|X] = E[u] = 0$$

In our example, this means that for any given level of education, one should be as likely to find a person of above average ability as a person of below average ability.

This understanding of unbiasedness has important implications for how we interpret the regression results. If the independence condition is not true, we cannot say that education *causes* an increase in wages. We can still state the fact that the regression shows that people with higher education get higher wages. Hence, we can state the result in terms of *correlation* but not in terms of *causation*, unless we are sure that the independence condition is true. This will be discussed in more detail in a separate subsection.

8.4 Hypothesis testing

If the estimator $\hat{\beta}$ is an unbiased estimator of the population parameter β, then we can conduct a hypothesis test in much the same way we conduct one for the population mean.

In most cases, the null hypothesis is that the explanatory variable has no effect

on the dependent variable, i.e.

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

Fortunately, the distribution for $\hat{\beta}$ follows a t-distribution. So we can proceed exactly as we do for the population mean. We first find the t-statistic.

$$t = \frac{\hat{\beta} - \beta_{H_0}}{se(\hat{\beta})} = \frac{\hat{\beta} - 0}{se(\hat{\beta})} = \frac{\hat{\beta}}{se(\hat{\beta})}$$

Here, $se(\hat{\beta})$ is the *standard error*, which is the term for the standard deviation of the $\hat{\beta}$ distribution. We will not mention the formula for it here, as most statistical packages calculate and present the standard error along with the coefficient estimate $\hat{\beta}$.

Next, we can compare this t-statistic with the critical value for a given significance level and decide whether or not to reject the null.

Most statistical packages also provide the p-value for the test, allowing us to directly compare this with the significance level and reject the null if p-value $<$ significance level.

8.5 Interpretation

If we reject the null of no effect, then we can conclude that the explanatory variable does tell us something about the dependent variable - we can still not conclude causality unless we are sure about the independence condition.

Remember that β is the slope coefficient. Hence, it gives us the rate of change in Y with respect to X or $\frac{dY}{dX}$. In the wage-education example, we can say that people with one extra year of education have on average $\hat{\beta}$ units higher wage. Observe how the phrasing does not imply any causation.

If we can establish causation, then we can say that one extra year of education *causes* or *leads to* an increase in wage by $\hat{\beta}$.

8.6 Statistical and economic significance

The statistical significance of a regression coefficient estimate refers to whether or not we can reject the null of 0 at a given significance level. It implies that the coefficient estimate is statistically *significantly different from 0.* i.e. it is enough number of standard deviations away from zero for us to statistically reject the possibility of the null being true at a given significance level. This says nothing about the actual size of the coefficient, only about how different it is from the null of 0.

It could be the case that in the wage-education regression, the coefficient estimate is 0.01, but the standard error is small enough for it to be statistically significant at 5%. But if an extra year of education increases wage by Rs 0.01, it does not really make any difference. Hence, the result is *economically insignificant.* Economic significance deals with the size of the coefficient estimate in the real-world context in which the regression is being analysed.

8.7 Goodness of fit

Given any dataset, we can run a regression between two variables and find a best-fit line. We need some way to assess how much of the variation in the dependent variable is actually explained by the best-fit line that we have obtained. To do that, we think of the value of the dependent variable for each observation y_i as consisting of two parts - the predicted or fitted value that falls on the best-fit line, $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, and the residual $\hat{u}_i$.

$$y_i = \hat{y}_i + \hat{u}_i$$

Following this division of y_i, we also divide the variance of y_i into two parts - explained and residual. The total variance is captured by a term called the *total sum of squares* or **SST**, which is given by the sum of squared deviations of the variable from its sample mean.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The *explained sum of squares* or **SSE** is the sum of squared deviations of the fitted value from the sample mean.

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

The *sum of squared residuals* or **SSR** is self explanatory.

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

It can be shown that

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{u}_i^2$$

$$\text{or, } SST = SSE + SSR$$

Hence, we have partitioned the overall variation in the dependent variable into a part that is explained by the regression and another that is left unexplained. The ratio of the explained to the total variation or SSE/SST, is called the coefficient of determination and is represented as R^2. The value of this term for a simple linear regression is the square of the correlation coefficient.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

8.8 Causality

The ideal situation for establishing causality would be an experiment. Imagine an experiment where we heat a gas and measure its pressure. When we increase the heat, the pressure increases. When we decrease the heat, the pressure decreases. Here we can conclude that change in heat *causes* change in pressure. In this case the change in heat is *exogenous* i.e. it is being caused by factors outside the system that influence pressure, in this case by the experimenter. Hence, if we run a regression of pressure on heat, we can directly interpret the regression coefficient as the *causal effect* of heat on pressure.

In economics most often we depend on observational, as opposed to experimental, data. For example, in the wage-education case, education was not being changed exogenously. Both wage and education were being determined by individuals interacting with the labour market and other socio-economic institutions. In this case, education is *endogenous*, because factors like ability, gender, family background etc. that influence wage could also influence education. Hence, the coefficient estimate from a simple regression of wage on education cannot be interpreted as the causal effect of education on wage.

The problems with establishing causality can be broadly be thought of as being of two types.

Omitted variable bias: As we discussed earlier, if we leave out a variable that affects the dependent variable (and is hence in the error term) and also affects the explanatory variable, then the coefficient estimate of the explanatory variable is going to be biased, and will hence not be the causal effect.

The best remedy to this problem is to actually include the variable so that it is no longer omitted. Such variables that are included as additional explanatory variables are sometimes knows as *control variables*. Once included, their effect on the dependent variable is directly controlled for by their own coefficient estimate,

and the coefficient estimate on our explanatory variable of interest is no longer biased by them. In our wage-education example, if we are worried about ability causing an omitted variable bias, we could include some measure of ability, like IQ (some may dispute that this is a good measure of ability), in the regression. If on including it the coefficient estimate of education changes significantly, we can conclude that leaving out ability was causing an omitted variable bias.

Reverse causality: It could be the case that while we are looking for the effect of X on Y, it is actually Y that causes X (or both could be true). In this case too we cannot interpret the regression coefficient on X causally.

The problem of reverse causality become very important while thinking of sequential events. If **A** happened first and was followed by **B**, and we find that **A** and **B** are correlated, then we tend to conclude that **A** caused **B**. But this may not always be true, because **A** may have been caused in expectation of **B**. For example think of farmers planting seeds and rainfall. If we observe different areas, we may see than whenever farmers plant seeds, rain follows. But this does not mean that planting seeds is causing the rain. The causality is reverse, farmers are planting seeds in expectation of the rain.

As empirical economists we should be able to look at a regression and question it on these two aspects. We should be able to think if there could be any omitted variables that would cause a bias, and if there is a possibility of reverse causality. But we should also remember that if causality is not proved, it doesn't mean it does not exist. Showing a correlation is the first step towards showing causality. In the wage-education regression, if we are not able to control for ability, it does not mean that education has no causal effect on wages. It just means that the given regression does not show if there is a causal effect or not.

8.9 Multiple explanatory variables

One of the advantages of regression over correlation is that it allows us to incorporate multiple explanatory variables. We can extend the simple linear regression model to a k-variable regression.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + u$$

The regression exercise is analogous to fitting a line in a k+1-dimensional space, so it is still a linear regression. The interpretation of each $\hat{\beta}_j$ remains the same as before, as does the hypothesis testing. The additional assumption is that when we are talking about β_1 as the effect of X_1 on Y, we assume that all the other Xs are kept constant. In other words, β_j represent the ceteris paribus effect of the corresponding explanatory variable X_j on the dependent variable Y.

The condition for unbiasedness is similar.

$$E[u|X_1, X_2, ..., X_k] = 0$$

The concerns with omitted variables and reverse causality also carry over from simple regressions.

When we have the options of choosing multiple variables, we run into various issues of which variables to choose and which to leave out, what functional form to choose, what if some of the explanatory variables are highly correlated etc. We shall not tackle these issues in the current course. It is sufficient to be able to understand and interpret a multi-variable regression, and be able to question if the results represent the causal effect of the explanatory variable of interest.