

The Regression Model: Simple Regression

- Regression analysis is a technique for estimating the values of the coefficients in a model of an economic process. First we will discuss simple regression involving two variables, later we will apply the technique to more involved models.
- Simple regression is appropriate when we believe that the values of one variable are systematically determined by the values of just one other variable.
- We are going to study economic processes. The input and output of the process are going to be observable, but the actual operation of the process is not.
- We are going to theorize the following: the economic process is the data generating process of n observations, each having a value for the variable X , which is input into the process that produces the variable Y as an output. We collect this information, but there are some random factors that we do not observe.

- Notice that we assume that X is the only variable that affects Y , any other variable plays no direct role in determining Y . Also, the values of X are taken as given, the economic process does not say anything about how the values of X came to be. Our only interest is in how a value of Y relates to a value of X .
- Now we are going to be more concrete in defining the specification that represents the process, we are going to use the *simple regression model*:

$$Y_i = \beta_0 + \beta_1 X_i + u_i. \quad (1)$$

In this equation, Y is called the *dependent variable*, and X is the *independent variable*, also called exogenous variable. β_0 and β_1 are parameters that have fixed values and are called the *coefficients* of the regression model. The term u is called the *disturbance*.

- The disturbance is a random term that does not depend on X . This term is meant to represent chance factors affecting the determination of Y . Also, minor things we cannot observe that affect Y will also enter in the disturbance term. Also, just measurement error. This term can be positive or negative, large or small, but on average we think it will be close to zero in a given data set.

- Consider the process by which families determine their savings. We will hypothesize that it only depends on their income. We take income as given. If we write this in the regression framework, β_0 represents the saving of families with no income (negative?), and β_1 is the increase in saving that would result from a unit increase in family income. In economic terms, β_1 is the marginal propensity to save.
- Based on the simple regression model we can decompose each value of Y_i into a systematic component $\beta_0 + \beta_1 X_i$ and a random component u_i . The value of the systematic component is called the *expected value* of Y for each observation:

$$E[Y_i] = \beta_0 + \beta_1 X_i, \quad (2)$$

given the definition of the regression process, it is the value that Y_i would take if the disturbance were equal to zero. We can then rewrite the relationship between X and Y as

$$Y_i = E[Y_i] + u_i. \quad (3)$$

Then we can actually graph the systematic part of the relationship between X and Y , as the line

$$E[Y] = \beta_0 + \beta_1 X, \quad (4)$$

which is called the *true regression line*.

- The value of Y_i for a single observation can be decomposed vertically into the distance up to a point on the true regression line and the distance from that point to the observation.

Estimation of the Model

- Once we believe that a given economic process is correctly described by the simple regression model, we face the issue that β_0 and β_1 are unknown. They are not written anywhere and cannot be directly measured. They can only be estimated out of the data on Y and X , which comes from the economic process we are analyzing.
- Now we start only with the data points, the points in the graph, we theorize there is a true regression line underlying the process but we do not observe it. But we can use the data to estimate the parameters of the line.
- Estimating the parameters β_0 and β_1 will be carried out by fitting an actual line through the data. The intercept of the line will be $\hat{\beta}_0$ and estimate of β_0 . The slope of the actual line will be $\hat{\beta}_1$, and it serves as our estimate of β_1 .
- Suppose the $\hat{\beta}_0$ and the $\hat{\beta}_1$ we end up correspond to the line

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X, \quad (5)$$

this is the *estimated regression line*, or the *fitted regression line*, or the *sample regression line*. For any data point (X_i, Y_i) this line decomposes the total value of Y_i into two parts.

- The first part, \hat{Y}_i , is the *fitted value* or *predicted value* for Y

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad (6)$$

which is the height up to a point on the estimated regression line, above X_i .

- The second part, e_i is called the *residual* or *error of fit*:

$$e_i = Y_i - \hat{Y}_i, \quad (7)$$

which is the vertical distance from the line to the data point. From this it is clear that

$$Y_i = \hat{Y}_i + e_i, \quad (8)$$

the residual can be positive or negative.

- Now, it all comes down to how do we determine $\hat{\beta}_0$ and $\hat{\beta}_1$. We want to find the line that best fits the data we have. There are many ways of doing that, but the standard techniques is to minimize some function of the sum of the residuals, the distances from the points to the estimated regression line.
- The most commonly used approach is what we call *ordinary least squares* (OLS). With this method, the criterion for being the best fit is that the line must make the *sum of the squared residuals* (SSR) as small as possible:

$$OLS \text{ criterion : minimize } SSR = \sum_{i=1}^n e_i^2. \quad (9)$$

Based on this criterion we can develop mathematical rules or formulae for calculating $\hat{\beta}_0$, and $\hat{\beta}_1$.

- Let's make this more formal. Suppose we have n observations on Y and X . We draw a sample regression line through this data, and associated with particular values of $\hat{\beta}_0$, and $\hat{\beta}_1$, are a set of residuals, e_i that are determined by

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i. \quad (10)$$

Then consider the sum of the squared residuals around a fitted line:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2. \quad (11)$$

In a particular data set Y_i and X_i are specific numbers, and what we want to find are the values of $\hat{\beta}_0$, and $\hat{\beta}_1$ that minimize this expression.

- A derivation using calculus leads us to the following expressions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (12)$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (13)$$

These are known as the *OLS estimators* of β_1 and β_0 . When we use these estimators we can be confident of getting the best fitting line.

- Notice that in order for the formula of $\hat{\beta}_1$ to make sense not all the values of X can be the same, otherwise the denominator would be zero. Notice too that the sum of residuals is also zero for this fitted line.

A note on interpretation

- In order to interpret regression results in the way we are going to, we have to be willing to assume that all the data come from the economic process that we specify. Then the regression fit to the data can be interpreted as an estimate of the true regression underlying the data.
- Imagine we actually estimate the unknown parameters of a relationship between family income and family savings with data in thousands of dollars, and the estimated regression is the following:

$$\hat{Y}_i = -0.0386 + 0.0863X_i. \quad (14)$$

These numbers are just examples, do not represent a serious study. Here $\hat{\beta}_0 = -0.0386$ and $\hat{\beta}_1 = 0.0863$ are the estimates of the parameters β_0 and β_1 of the behavioral relationship.

- The estimated slope tells us that if a family's income were to increase by, say, \$1,000 ($\Delta X = 1000$) its predicted saving would increase by 0.0863 thousand dollars (\$86.3). If the increase in income is around \$2,500 the increase in saving would be around \$216. Notice that the intercept of the equation plays no role here.

- The estimated intercept says that if a family income were to be zero, its predicted saving would be -0.0386 thousand dollars (or -\$38.6). This could sound odd, but what it means is that saving can be negative in a period, meaning that the family would consume part of its wealth.
- We can also use the estimated regression line to make predictions of saving for a family with some specific income. So if a family has income equal to \$ 8,000 then

$$\hat{Y}_i = -0.0386 + (0.0863)(8.0) = 0.6518 \quad (15)$$

in thousands of dollars, so about \$652. Notice that what we are doing is finding the height of the estimated regression line corresponding to the specific value of X .