

Chapter 1: An Overview of Regression Analysis

Lecture 2

Revised on September 25, 2007

1 The Population Regression

1.1 What is a regression?

Example (Costs of an ice-cream stand): Imagine that you are thinking of starting an ice-cream stand. You are interested in the expected cost. Even if you do not sell any ice cream, you need to pay the rent for the stand. When you sell an ice cream bar, it cost you the wholesale price of the bar. There are other costs for electricity that vary with the temperature. The expected cost increases as the quantity of ice cream bars you sell.

Example(Demand for a used CD): Imagine that you own a used music CD store and just got 24 copies of a CD by the Beatles. You are thinking

about setting the price for it. The expected value of demand for the music CD is likely to fall as you set its price higher.

These examples show that the expected value of a random variable Y may depend on the value of another variable X . In that case, we write the expected value of Y given X as $E(Y|X)$. $E(Y|X)$ is a function of X .

$$E(Y|X) = f(X) \tag{1}$$

This is called the *population regression*.

Y : The dependent variable X : The independent (or explanatory) variable

In this course, we assume that $f(X)$ is a linear function of X :

$$E(Y|X) = \beta_0 + \beta_1 X \quad (2)$$

This is a strong assumption. However, we can often make a nonlinear function fit into this by creating a new variable.

Example:

$$E(Y|X) = \beta_0 + \beta_1 X^2 \quad (3)$$

Then create a new variable

$$Z = X^2 \quad (4)$$

then

$$E(Y|X) = \beta_0 + \beta_1 Z \quad (5)$$

1.2 Regression Coefficients

In Equation (2), β_0 and β_1 are called the *coefficients*, or *regression coefficients*.

β_0 : the *constant or intercept* term

β_1 : the *slope coefficient*.

Example (Costs of an ice-cream stand-continued): Imagine that the cost is given by the rent for the stand of \$30 and the wholesale price of each bar of \$0.50. Let Y be the cost and X be the number of ice cream bars sold. Then

$$E(Y|X) = 30 + 0.5X \quad (6)$$

The constant term is 30. This means that the cost is \$30 if the number of ice cream bars sold is zero. The slope coefficient is 0.5, which means that the cost increase by \$0.5 when the number of ice cream bars sold increases by one. In Economics, \$0.5 in this example is called marginal cost. On the other hand, average cost is cost per bar, which is $(30 + 0.5X)/X = 30/X + 0.5$. In this example, average cost declines as the number of ice cream bars sold increases, while marginal cost is constant. Average cost is \$30.5 for $X = 1$, \$2 for $X = 20$, about \$1.214 for $X = 42$, and about \$1.198 for $X = 43$. If you set the price of the ice cream bar to be \$ 1.2, then the profit is only positive when the price is greater than average cost. So you need to sell at least 43 ice cream bars in order to make any profit.

It is important to note that $\beta_2 = 0.5$ means that the cost increases by \$0.5 when the number of ice cream bars sold increases by one and that this is different from average cost.

Exercise from the Demand for a used CD Example: Let Y be the expected demand for the CD and X be the price of the CD. Imagine that

$$E(Y|X) = 50 - 3X \quad (7)$$

Interpret the constant term and the slope coefficient. Then draw a diagram for this population regression.

1.3 The Stochastic Error Term

In real applications, the observed value in data is not exactly equal to $E(Y|X)$. We express this situation by adding a stochastic error term:

$$Y = \beta_0 + \beta_1 X + \epsilon. \quad (8)$$

The error term, ϵ , is a random variable and is defined by

$$\epsilon = Y - E(Y|X) \quad (9)$$

Equation (8) can be thought of having two components, the *deterministic* component and the *stochastic*, or random, component. The expression $\beta_0 + \beta_1 X$ is called the deterministic component of the regression equation. The deterministic component is the expected value of Y given X .

$$E(Y|X) = \beta_0 + \beta_1 X \quad (10)$$

The value of Y observed in the real world is unlikely to be exactly equal to the expected value given X , and the difference is the stochastic error term.

$$Y = E(Y|X) + \epsilon = \beta_0 + \beta_1 X + \epsilon. \quad (11)$$

The random component exists because of:

(1) *Omitted Variables*: Many minor influences on Y are omitted from the equation (for example, because data are unavailable).

(2) *Measurement Errors*: It is virtually impossible to avoid some sort of measurement error in the dependent variable.

(3) *Intrinsic Randomness*: Some people believe that human behavior is such that actions taken under identical circumstances will differ in a random way. The disturbance term can be thought of as representing this inherent randomness in human behavior.

1.4 Extending the Notation

Our regression notation needs to be extended to include reference to the number of observations.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i (i = 1, 2, \dots, N) \quad (12)$$

where

Y_i is the i th observation of the dependent variable

X_i is the i th observation of the independent variable

ϵ_i is the stochastic error term for the i th observation

β_0 and β_1 are the regression coefficients

N is the number of observations (the sample size).

Example (Costs of an ice-cream stand - continued) Let Y_i be costs for the i th observation, X_i be the number of the ice cream bars sold for the i th observation. Then

$$E(Y_i|X_i) = 30 + 0.5X_i. \quad (13)$$

Suppose that we have three observations for Y and X for three days on which 25, 40, and 52 bars were sold. We index observed values by subscripts.

$$\text{For } X_1 = 25, E(Y|X) = 30 + 0.5 \cdot 25 = 42.5.$$

$$\text{For } X_2 = 40, E(Y|X) = 30 + 0.5 \cdot 40 = 50.$$

$$\text{For } X_3 = 52, E(Y|X) = 30 + 0.5 \cdot 52 = 56.$$

These are summarized in the Table 1.A.

Table 1.A Expected Costs to operate an ice cream stand

| Observation | The number of ice cream bars sold | Expected costs |
|-------------|-----------------------------------|----------------|
| i | X_i | $E(Y_i X_i)$ |
| (1) | (2) | (3) |
| 1 | 25 | 42.5 |
| 2 | 40 | 50 |
| 3 | 52 | 56 |

Table 1.A gives the population regression.

Observations are usually not exactly equal to these expected values. For example, costs to operate an ice-cream stand depend on the temperature of the day for electricity, etc. In this example, instead of a realistic element such as temperature, imagine that the rent you pay for the stand changes,

depending on the results of tossing a coin twice. For HH , two dollars are discounted from the rent; for HT , a dollar is discounted; for TH , a dollar is added; and for TT , two dollars are added.

Imagine that Steve tossed a coin twice for each day, and got TT for the first day, HT for the second day, and HH for the third day. He sold 25, 40, 52 bars on the first, second, and third day, respectively. Then his costs are \$44.5, \$49, and \$54. This is summarized in Table 1.B.

Table 1.B Data of costs to operate Steve's ice cream stand

| Observation | The number of ice cream bars sold | Observed costs |
|-------------|-----------------------------------|----------------|
| i | X_i | Y_i |
| (1) | (2) | (3) |
| 1 | 25 | 44.5 |
| 2 | 40 | 49 |
| 3 | 52 | 54 |

Imagine that Jane tossed a coin twice for each day, and got HT for the first day, TT for the second day, and TH for the third day. She also sold 25, 40, 52 bars on the first, second, and third day, respectively. Then her costs are \$41.5, \$52, and \$57. This is summarized in Table 1.C.

Table 1.C Data of costs to operate Jane's ice cream stand

| Observation | The number of ice cream bars sold | Observed costs |
|-------------|-----------------------------------|----------------|
| i | X_i | Y_i |
| (1) | (2) | (3) |
| 1 | 25 | 41.5 |
| 2 | 40 | 52 |
| 3 | 52 | 57 |

1.5 The Estimated Regression

The regression coefficients β_0 and β_1 need to be estimated from a data set (or a sample).

Suppose that we estimate β_0 and β_1 by the estimator $\hat{\beta}_0$ and the estimator $\hat{\beta}_1$, respectively from a sample of (X, Y) , $\{(Y_1, X_1), (Y_2, X_2), \dots, (Y_N, X_N)\}$. Tables 1.A and 1.B give examples of two samples. Then

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (14)$$

is called the *estimated (or sample) regression equation* \hat{Y}_i is called the *estimated or fitted value*.

There are many possible estimators for the regression coefficients. The most popular estimator is the one we will learn in Chapter 2 called *Ordinary Least Squares (OLS)* estimator.

By applying OLS, I obtained $\hat{\beta}_0 = 35.52$ and $\hat{\beta}_1 = 0.35$ from the data in Table 1.B. We write

$$\hat{Y}_i = 35.52 + 0.35X_i. \quad (15)$$

for this data set.

By applying OLS, I obtained $\hat{\beta}_0 = 27.55$ and $\hat{\beta}_1 = 0.58$ from the data in Table 1.C. We write

$$\hat{Y}_i = 27.55 + 0.58X_i. \quad (16)$$

for this data set.

In this example, we know the population regression is

$$E(Y_i|X_i) = 30 + 0.5X_i. \quad (17)$$

Thus, the population regression and the estimated regression are not the same, and can be very different. The true regression coefficients and the estimated coefficients are not the same, and can be very different. The

expected value given X and the estimated value are not the same, and can be very different.

For each observation i , Equation (14) does not fit perfectly. The difference between Y_i and \hat{Y}_i is called the *residual* and denoted by e_i in this course:

$$e_i = Y_i - \hat{Y}_i \quad (18)$$

Note the distinction between the residual and the stochastic error term

$$\epsilon_i = Y_i - E(Y_i|X_i). \quad (19)$$

We write

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \quad (20)$$

Using $\hat{\beta}_0 = 35.52$ and $\hat{\beta}_1 = 0.35$, I computed the residual and stochastic error term from the data in Table 1.B.

Table 1.D The Residual and Error in Steve's Data

| Raw Data | | | Calculations | | | |
|----------|-------|-------|--------------|-------------------------|--------------|---------------------------------|
| i | X_i | Y_i | \hat{Y}_i | $e_i = Y_i - \hat{Y}_i$ | $E(Y_i X_i)$ | $\epsilon_i = Y_i - E(Y_i X_i)$ |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 1 | 25 | 44.5 | 44.27 | 0.23 | 42.5 | 2 |
| 2 | 40 | 49 | 49.52 | -0.52 | 50 | -1 |
| 3 | 52 | 54 | 53.72 | 0.28 | 56 | -2 |

The residual, e_i , is the sample counterpart of ϵ_i . However, they are not the same, and can be very different.