Eco 213: Basic Data Analysis and Econometrics

Lecture 1: Regression Analysis
January 5, 2019

Dr. Garima Malik
Department of Economics
Shiv Nadar University

What is Econometrics?

- Econometrics literally means **"economic measurement"**
- It is the **quantitative measurement** and **analysis** of actual economic and business phenomena—and so involves:
  - economic theory
  - Statistics
  - Math
  - observation/data collection

What is Econometrics? (cont.)

- Three major uses of econometrics:
  - Describing economic reality
  - Testing hypotheses about economic theory
  - Forecasting future economic activity
- So econometrics is all about **questions:** the researcher first asks questions and then uses econometrics to answer them

What is Econometrics?

- Development of statistical techniques to uncover economic relationships in data, to test economic theories, evaluating and implementing government and business policy.

WHAT IS ECONOMETRICS?

- Econometrics means ñeconomic measurementò. the scope of econometrics is much broader, as can be seen from the following definitions:
- ñConsists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical resultsò (Gerhard 1968).
- ñEconometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomenaò (Goldberger 1964).
- ñEconometrics is concerned with the empirical determination of economic lawsò (Theil 1971).

WHY ECONOMETRICS IS A SEPARATE DISCIPLINE?

- The subject deserves to be studied in its own right for the following reasons:
- Economic theory makes statements or hypotheses that are mostly qualitative in nature (the law of demand), the law does not provide any numerical measure of the relationship. This is the job of the econometrician.
- The main concern of mathematical economics is to express economic theory in mathematical form without regard to measurability or empirical verification of the theory. Econometrics is mainly interested in the empirical verification of economic theory.
- Economic statistics is mainly concerned with collecting, processing, and presenting economic data in the form of charts and tables. It does not go any further. The one who does that is the econometrician.

METHODOLOGY OF ECONOMETRICS

- Broadly speaking, traditional econometric methodology proceeds along the following lines:

1. Statement of theory or hypothesis.
2. Specification of the mathematical model of the theory
3. Specification of the statistical, or econometric, model
4. Collecting the data
5. Estimation of the parameters of the econometric model
6. Hypothesis testing
7. Forecasting or prediction
8. Using the model for control or policy purposes.

- To illustrate the preceding steps, let us consider the well-known Keynesian theory of consumption.

1. Statement of Theory or Hypothesis

- Keynes states that on average, consumers increase their consumption as their income increases, but not as much as the increase in their income (MPC < 1).

2. Specification of the Mathematical Model of Consumption (single-equation model)

$$Y = \beta_1 + \beta_2 X \qquad 0 < \beta_2 < 1 \qquad (I.3.1)$$

Y = *consumption expenditure and (*dependent variable)

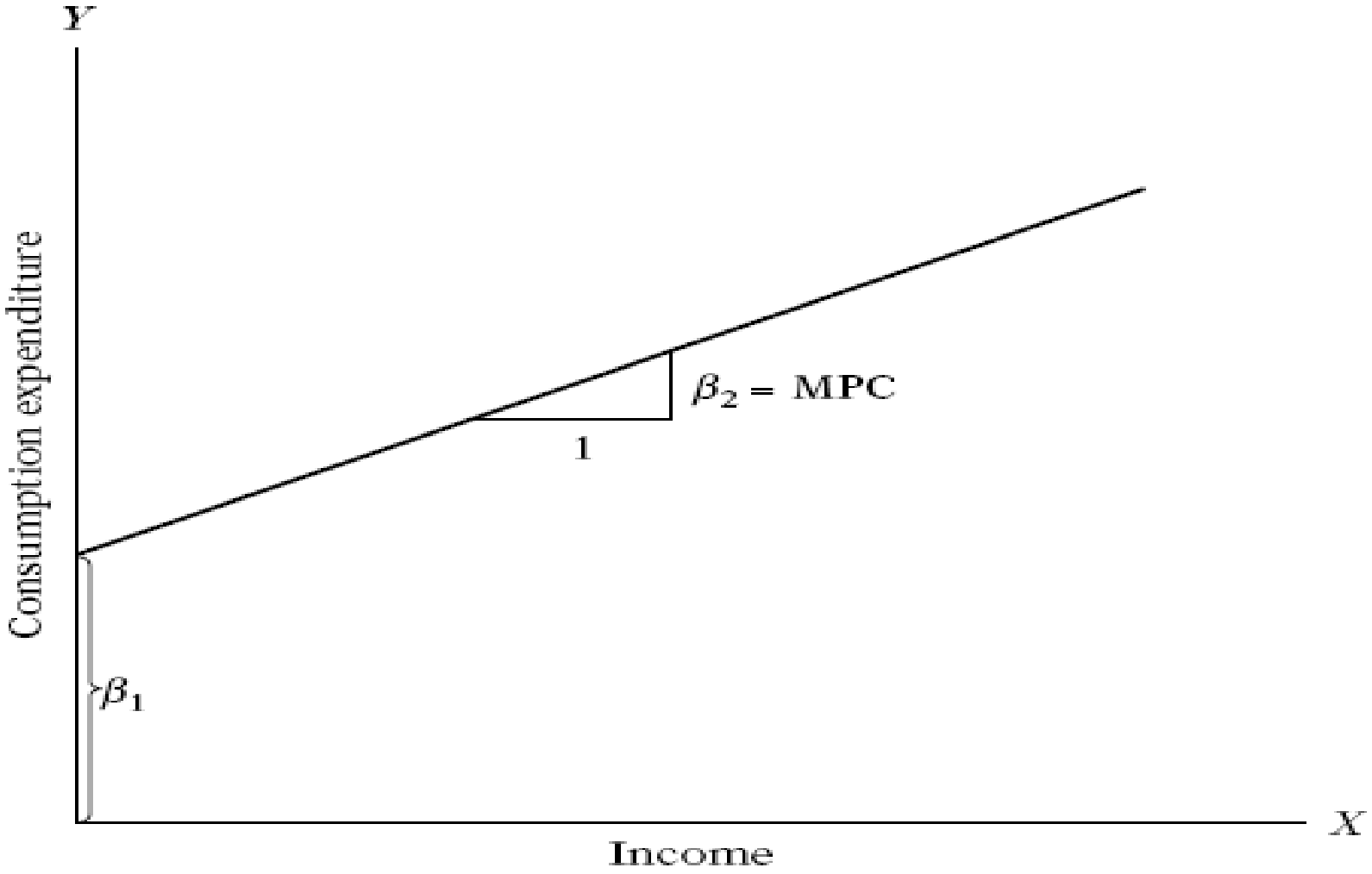
X = income, (independent, or explanatory variable)

β_1 = the intercept

β_2 = the slope coefficient

- The slope coefficient β_2 *measures the MPC.*

- Geometrically,



3. Specification of the Econometric Model of Consumption

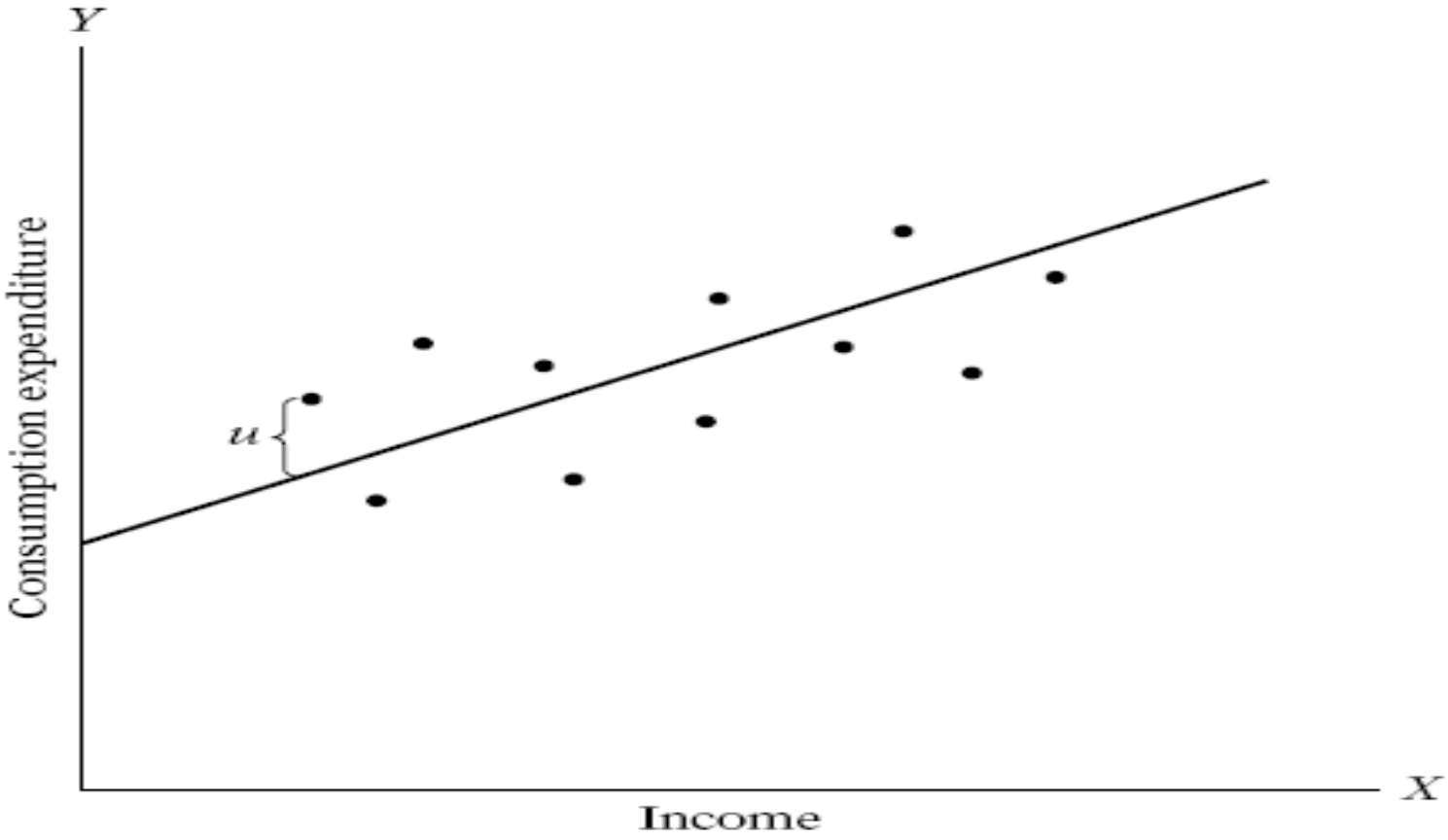
- The relationships between economic variables are generally *inexact*. In addition to income, other variables affect consumption expenditure. For example, *size of family, ages of the members in the family, family religion*, etc., are likely to exert some influence on consumption.

- To allow for the *inexact* relationships between economic variables, (I.3.1) is modified as follows:

- $Y = \beta_1 + \beta_2 X + u$ *(I.3.2)*

- where u, known as *the disturbance, or error, term,* is a random (stochastic) variable that has well-defined *probabilistic properties*. The disturbance term *u may well represent all those factors that affect consumption but are not taken into account* explicitly.

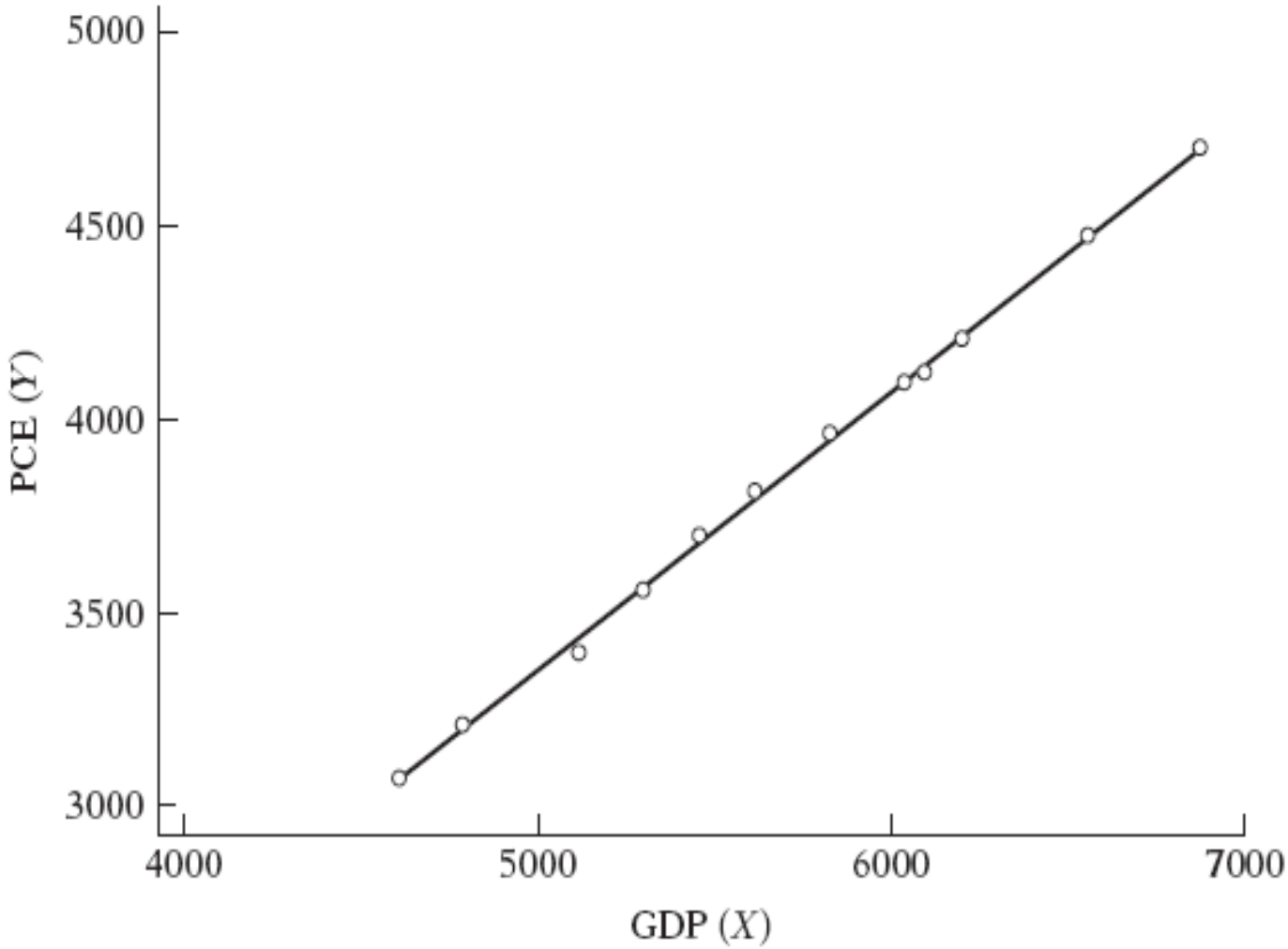
- (I.3.2) is an example of a linear regression model, i.e., it hypothesizes that *Y is linearly related to X, but that the relationship between the two is not exact*; it is subject to individual variation. The econometric model of (I.3.2) can be depicted as shown in Figure I.2.



4. Obtaining Data

- To obtain the numerical values of *β1 and β2, we need data.* Look at Table I.1, which relate to the *personal consumption expenditure* (PCE) and the *gross domestic product (GDP)*. The data are in "real" terms.

Year	*Y*	*X*
1982	3081.5	4620.3
1983	3240.6	4803.7
1984	3407.6	5140.1
1985	3566.5	5323.5
1986	3708.7	5487.7
1987	3822.3	5649.5
1988	3972.7	5865.2
1989	4064.6	6062.0
1990	4132.2	6136.3
1991	4105.8	6079.4
1992	4219.8	6244.4
1993	4343.6	6389.6
1994	4486.0	6610.7
1995	4595.3	6742.1
1996	4714.1	6928.4


5000
4500
4000
3500
3000
PCE (Y)
4000
5000
6000
7000
GDP (X)

5. Estimation of the Econometric Model

- *Regression analysis* is the main tool used to obtain the estimates. Using this technique and the data given in Table I.1, we obtain the following estimates of β_1 *and* β_2*, namely,* -184.08 *and* 0.7064. Thus, the estimated consumption function is:

- $\hat{Y} = -184.08 + 0.7064X_i$ *(I.3.3)*

- *The estimated* regression line is shown in Figure I.3. The regression line fits the data quite well. The slope coefficient (i.e., the MPC) was about 0.70, an increase in real income of 1 dollar led, *on average, to an increase of about 70 cents in real consumption*.

6. Hypothesis Testing

- That is to find out whether the estimates obtained in, Eq. (I.3.3) are in accord *with the expectations of the theory that is being tested*. Keynes expected the *MPC to be positive but less than 1*. In our example we found the MPC to be about 0.70. But before we accept this finding as confirmation of Keynesian consumption theory, we must enquire whether this estimate is sufficiently below unity. In other words, is *0.70 statistically less than 1? If it is, it may support* Keynes' theory.
- Such confirmation or refutation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as statistical inference (hypothesis testing).

7. Forecasting or Prediction

- To illustrate, suppose we want to predict the mean consumption expenditure for 1997. The GDP value for 1997 was 7269.8 billion dollars consumption would be:

$$\hat{Y}_{1997} = -184.0779 + 0.7064\,(7269.8) = 4951.3 \qquad \text{(I.3.4)}$$

- The *actual value* of the consumption expenditure reported in 1997 was 4913.5 billion dollars. The estimated model (I.3.3) thus over-predicted the actual consumption expenditure by about 37.82 billion dollars. We could say the *forecast error* is about 37.8 billion dollars, which is about 0.76 percent of the actual GDP value for 1997.
- Now suppose the government decides to propose a reduction in the income tax. What will be the effect of such a policy on income and thereby on consumption expenditure and ultimately on employment?

- Suppose that, as a result of the proposed policy change, investment expenditure increases. What will be the effect on the economy? As macroeconomic theory shows, the change in income following, a dollarôs worth of change in investment expenditure is given by the income multiplier *M,* which is defined as:

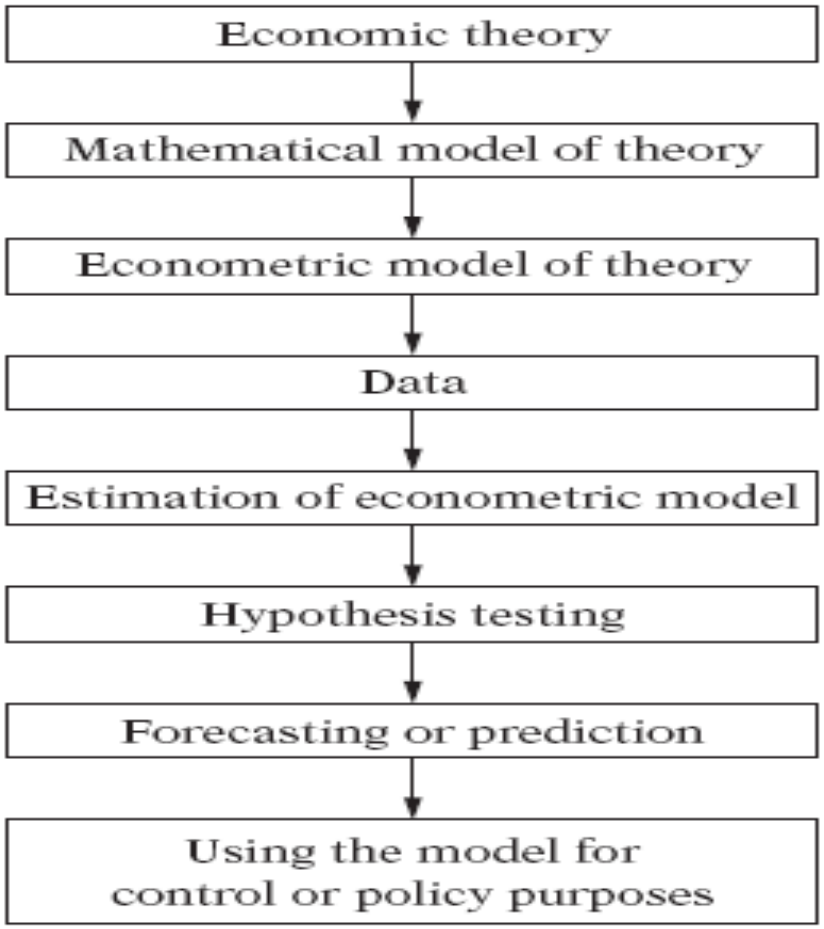
- $M = 1/(1 - \text{MPC})$ (I.3.5)

- The multiplier is about *M = 3.33. That is, an increase (decrease) of a dollar in investment will eventually* lead to more than a threefold increase (decrease) in income; note that it takes time for the multiplier to work.
- The critical value in this computation is *MPC*. Thus, a quantitative estimate *of MPC provides valuable information* for policy purposes. Knowing MPC, one can predict the future course of income, consumption expenditure, and employment following a change in the governmentôs fiscal policies.

8. Use of the Model for Control or Policy Purposes

- Suppose we have the estimated consumption function given in (I.3.3). Suppose further the government believes that consumer expenditure of about 4900 will keep the *unemployment rate at its current level of about 4.2%*. What level of income will guarantee the target amount of consumption expenditure?
- If the regression results given in (I.3.3) seem reasonable, simple arithmetic will show that:
- $4900 = \bar{\imath}\ 184.0779 + 0.7064X$ *(I.3.6)*
- which gives *X = 7197, approximately. That is, an income level of about* 7197 (billion) dollars, given an MPC of about 0.70, will produce an expenditure of about 4900 billion dollars. As these calculations suggest, an estimated model may be used for control, or policy, purposes. By appropriate fiscal and monetary policy mix, the government can manipulate the *control variable X to produce the desired level of the target variable Y.*

Classical econometric modeling.



Choosing among Competing Models

- When a governmental agency (e.g., the U.S. Department of Commerce) collects economic data, such as that shown in Table I.1, it does not *necessarily have any economic theory in mind*. How then does one know that the data really support the Keynesian theory of consumption? Is it because the Keynesian consumption function (i.e., the regression line) shown in Figure I.3 is extremely close to the actual data points? Is it possible that another consumption model (theory) might equally fit the data as well? For example, Milton Friedman has developed a model of consumption, called the *permanent income hypothesis*. *Robert Hall has also developed a model of* consumption, called the *life-cycle permanent income hypothesis*. *Could one* or both of these models also fit the data in Table I.1?
- In short, the question facing a researcher in practice is how to choose among competing hypotheses or models of a given phenomenon, such as the consumption–income relationship.

- The eight-step classical econometric methodology discussed above is neutral in the sense that it can be used to test any of these rival hypotheses. Is it possible to develop a methodology that is comprehensive enough to include competing hypotheses? This is an involved and controversial topic.

Example

- Consider the general and purely theoretical relationship:

$$Q = f(P, P_s, Yd) \qquad (1.1)$$

- Econometrics allows this general and purely theoretical relationship to become explicit:

$$Q = 27.7 - 0.11P + 0.03P_s + 0.23Yd \qquad (1.2)$$

What is Regression Analysis?

- **Economic theory** can give us the **direction** of a change, e.g. the change in the demand for dvd's following a price decrease (or price increase)
- But what if we want to know not just **"how?"** but also **"how much?"**
- Then we need:
  - A sample of data
  - A way to estimate such a relationship
    - one of the most frequently ones used is **regression analysis**

What is Regression Analysis? (cont.)

- Formally, regression analysis is a **statistical technique** that attempts to "explain" movements in one variable, the **dependent** variable, as a function of movements in a set of other variables, the **independent** (or **explanatory**) variables, through the quantification of a single equation

Example

- Return to the example from before:

$$Q = f(P, P_s, Yd) \qquad (1.1)$$

- Here, Q is the **dependent** variable and P, P_s, Yd are the **independent** variables
- Don't be confused by the words **dependent** and **independent**, however
- A statistically significant regression result does **not necessarily** imply **causality**
- We also need:
  - Economic theory
  - Common sense

Single-Equation Linear Models

- The simplest example is:

$$Y = \beta_0 + \beta_1 X \qquad (1.3)$$

- The βs are denoted "**coefficients**"
  - β_0 is the "**constant**" or "**intercept**" term
  - β_1 is the "**slope coefficient**": the amount that Y will change when X increases by one unit; for a linear model, β_1 is constant over the entire function

Figure 1.1
Graphical Representation of the Coefficients of the Regression Line

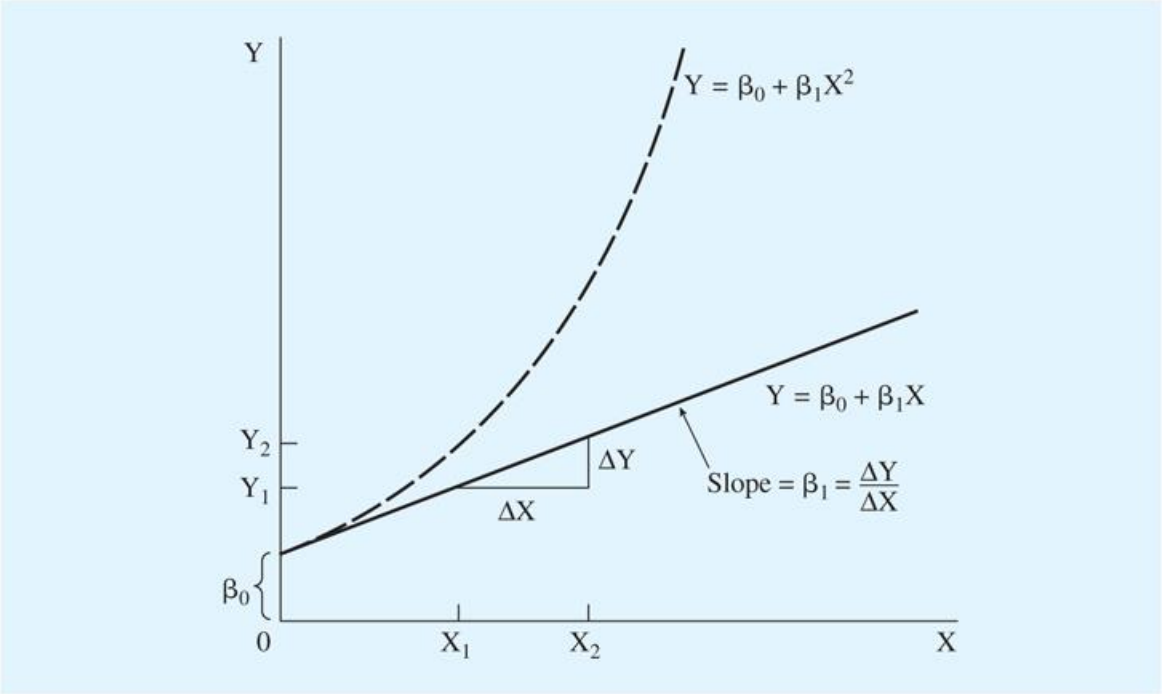


Figure 1.1 Graphical Representation of the Coefficients of the Regression Line

The graph of the equation $Y = \beta_0 + \beta_1 X$ is linear with a constant slope equal to $\beta_1 = \Delta Y/\Delta X$. The graph of the equation $Y = \beta_0 + \beta_1 X^2$, on the other hand, is nonlinear with an increasing slope (if $\beta_1 > 0$).

Single-Equation Linear Models (cont.)

- Application of linear regression techniques requires that the equation be **linear—**
- By contrast, the equation

$$Y = \beta_0 + \beta_1 X^2 \qquad (1.4)$$

  is **not linear**
- What to do? First define

$$Z = X^2 \qquad (1.5)$$

- Substituting into (1.4) yields:

$$Y = \beta_0 + \beta_1 Z \qquad (1.6)$$

- This redefined equation is now **linear** (in the coefficients β_0 and β_1 **and** in the variables Y and Z)

Single-Equation Linear Models (cont.)

- Is (1.3) a complete description of origins of variation in Y?
- No, at least four sources of variation in Y other than the variation in the included Xs:
  - Other potentially important explanatory variables may be missing (e.g., X_2 and X_3)
  - Measurement error
  - Incorrect functional form
  - Purely random and totally unpredictable occurrences
- Inclusion of a "**stochastic error term**" (u) effectively "takes care" of all these other sources of variation in Y that are NOT captured by X, so that (1.3) becomes:

$$Y = \beta_0 + \beta_1 X + u \qquad (1.7)$$

Single-Equation Linear Models (cont.)

- Two components in (1.7):
  - **deterministic** component ($\beta_0 + \beta_1 X$)
  - **stochastic/random** component (u)
- Why **"deterministic"**?
  - Indicates the value of Y that is determined by a given value of X (which is assumed to be non-stochastic)
  - Alternatively, the det. comp. can be thought of as the **expected value** of Y **given** X—namely E(Y|X)—i.e. the **mean (or average)** value of the Ys associated with a **particular value** of X
  - This is also denoted the **conditional expectation** (that is, **expectation** of Y **conditional** on X)

Example: Aggregate Consumption Function

- Aggregate consumption as a function of aggregate income may be lower (or higher) than it would otherwise have been due to:
  - consumer uncertainty—hard (impossible?) to measure, i.e. is an omitted variable
  - Observed consumption may be different from actual consumption due to measurement error
  - The "true" consumption function may be nonlinear but a linear one is estimated (see **Figure 1.2** for a graphical illustration)
  - Human behavior always contains some element(s) of pure chance; unpredictable, i.e. random events may increase or decrease consumption at any given time
- Whenever one or more of these factors are at play, the observed Y will differ from the Y predicted from the deterministic part, $b_0 + b_1X$

Figure 1.2
Errors Caused by Using a Linear Functional Form to Model a Nonlinear Relationship

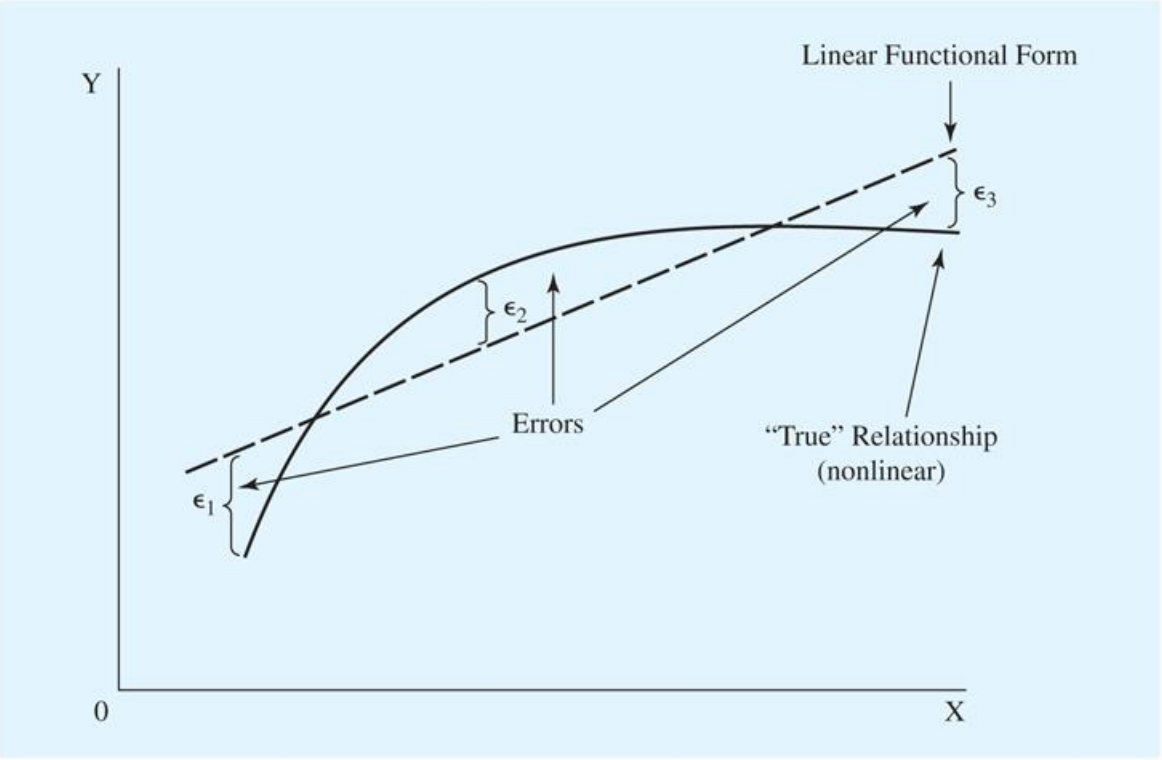


Figure 1.2 Errors Caused by Using a Linear Functional Form to Model a Nonlinear Relationship

Extending the Notation

- Include reference to the number of observations
  - Single-equation linear case:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (i = 1,2,\ldots,N) \qquad (1.10)$$

    - So there are really N equations, one for each observation
    - the coefficients, β_0 and β_1, are **the same**
    - the values of Y, X, and u **differ** across observations

Extending the Notation (cont.)

- The general case: **multivariate** regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad (i = 1,2,\ldots,N) \quad (1.11)$$

- Each of the slope coefficients gives the impact of a one-unit increase in the corresponding X variable on Y, holding the other included independent variables constant (i.e., **ceteris paribus**)
- As an (implicit) consequence of this, the impact of variables that are **not included** in the regression are **not held constant**

Example: Wage Regression

- Let wages (WAGE) depend on:
  - years of work experience (EXP)
  - years of education (EDU)
  - gender of the worker (GEND: 1 if male, 0 if female)
- Substituting into equation (1.11) yields:

$$WAGE_i = \beta_0 + \beta_1 EXP_i + \beta_2 EDU_i + \beta_3 GEND_i + u_i \quad (1.12)$$

Indexing Conventions

- Subscript "i" for data on individuals (so called **"cross section"** data)
- Subscript "t" for **time series** data (e.g., series of years, months, or days—daily exchange rates, for example)
- Subscript "it" when we have **both** (for example, **"panel data"**)

The Estimated Regression Equation

- The regression equation considered so far is the **"true"—but unknown—theoretical** regression equation
- Instead of "true," might think about this as the **population** regression vs. the **sample/estimated** regression
- How do we obtain the empirical counterpart of the theoretical regression model (1.14)?
- It has to be **estimated**
- The empirical counterpart is:

- The signs on top of the estimates are denoted "hat," so that we have "Y-hat," for example

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

The Estimated Regression Equation (cont.)

- For each sample we get a **different** set of estimated regression coefficients
- $\hat{Y}$ is the **estimated** value of Y_i (i.e. the dependent variable for observation i); similarly it is the **prediction** of $E(Y_i|X_i)$ from the regression equation
- The **closer** $\hat{Y}$ is to the **observed** value of Y_i, the better is the **"fit"** of the equation
- Similarly, the **smaller** is the estimated error term, e_i, often denoted the **"residual,"** the better is the fit

The Estimated Regression Equation (cont.)

- This can also be seen from the fact that

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

- Note difference with the error term, ϵ_i, given as

$$\epsilon_i = Y_i - E(Y_i|X_i)$$

- This all comes together in Figure 1.3

Figure 1.3
True and Estimated Regression Lines

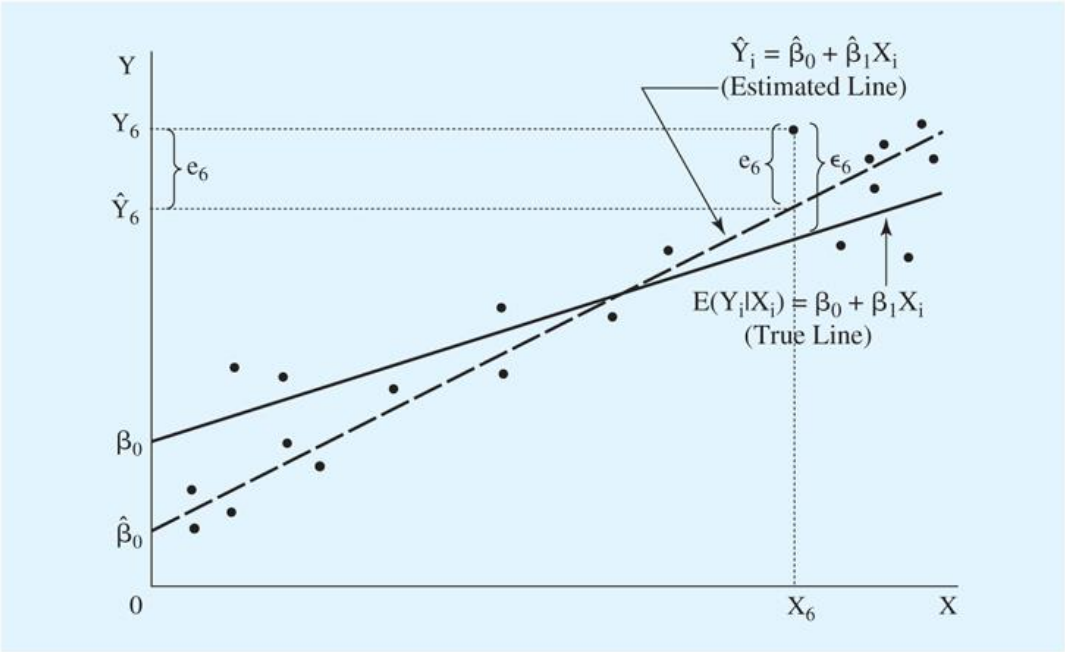


Figure 1.3 True and Estimated Regression Lines

The true relationship between X and Y (the solid line) typically cannot be observed, but the estimated regression line (the dashed line) can. The difference between an observed data point (for example, i = 6) and the true line is the value of the stochastic error term (ϵ_6). The difference between the observed Y_6 and the estimated value from the regression line ($\hat{Y}_6$) is the value of the residual for this observation, e_6.

Example: Using Regression to Explain Housing prices

- Houses are not homogenous products, like corn or gold, that have generally known market prices
- So, how to appraise a house against a given asking price?
- Yes, it's true: many real estate appraisers actually use regression analysis for this!
- Consider specific case: Suppose the asking price was $230,000

Example: Using Regression to Explain Housing prices (cont.)

- Is this fair / too much /too little?
- Depends on size of house (higher size, higher price)
- So, collect cross-sectional data on prices (in thousands of $) and sizes (in square feet) for, say, 43 houses
- Then say this yields the following estimated regression line:

(1.23)

$$\widehat{PRICE}_i = 40.0 + 0.138\,SIZE_i$$

Figure 1.5 A Cross-Sectional Model of Housing Prices

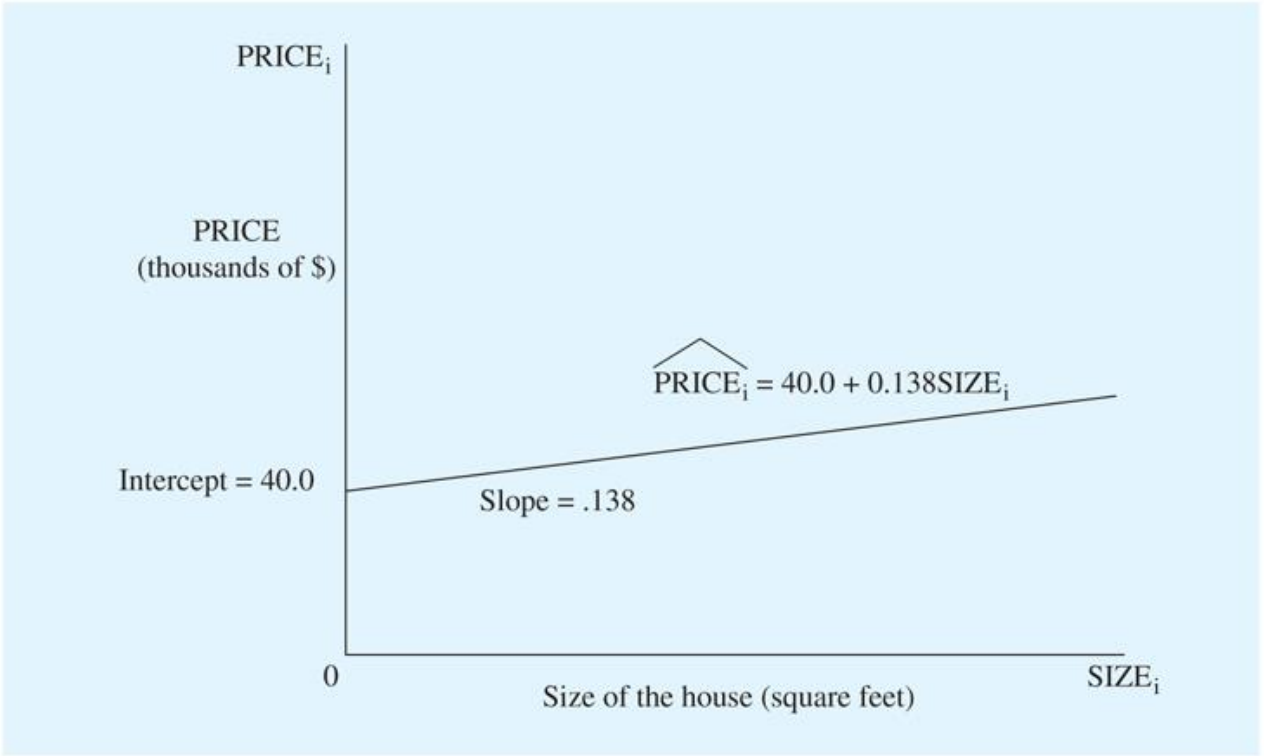


Figure 1.5 A Cross-Sectional Model of Housing Prices

A regression equation that has the price of a house in Southern California as a function of the size of that house has an intercept of 40.0 and a slope of 0.138, using Equation 1.23.

Example: Using Regression to Explain Housing prices (cont.)

- Note that the **interpretation** of the **intercept** term is problematic in this case
- The literal interpretation of the intercept here is the price of a house with a size of **zero** square feet…

Example: Using Regression to Explain Housing prices (cont.)

- How to use the estimated regression line / estimated regression coefficients to answer the question?
  - Just plug the particular size of the house, you are interested in (here, 1,600 square feet) into (1.23)
  - Alternatively, read off the estimated price using Figure 1.5
- Either way, we get an estimated price of $260.8 (thousand, remember!)
- So, in terms of our original question, it's a good deal—go ahead and purchase
- Note that we simplified a lot in this example by assuming that **only** size matters for housing prices