CHAPTER 2

The Simple Linear Regression Model

LEARNING OBJECTIVES

Remark

Learning Objectives and *Keywords* sections will appear at the beginning of each chapter. We urge you to think about and possibly write out answers to the questions, and make sure you recognize and can define the keywords. If you are unsure about the questions or answers, consult your instructor. When examples are requested in *Learning Objectives* sections, you should think of examples *not* in the book.

Based on the material in this chapter you should be able to

1. Explain the difference between an estimator and an estimate, and why the least squares estimators are random variables, and why least squares estimates are not.
2. Discuss the interpretation of the slope and intercept parameters of the simple regression model, and sketch the graph of an estimated equation.
3. Explain the theoretical decomposition of an observable variable y into its systematic and random components, and show this decomposition graphically.
4. Discuss and explain each of the assumptions of the simple linear regression model.
5. Explain how the least squares principle is used to fit a line through a scatter plot of data. Be able to define the least squares residual and the least squares fitted value of the dependent variable and show them on a graph.
6. Define the elasticity of y with respect to x and explain its computation in the simple linear regression model when y and x are not transformed in any way, and when y and/or x have been transformed to model a nonlinear relationship.
7. Explain the meaning of the statement ''If regression model assumptions SR1–SR5 hold, then the least squares estimator b_2 is unbiased.'' In particular, what exactly does ''unbiased'' mean? Why is b_2 biased if an important variable has been omitted from the model?
8. Explain the meaning of the phrase ''sampling variability.''
9. Explain how the factors σ^2, $\sum(x_i - \bar{x})^2$, and N affect the precision with which we can estimate the unknown parameter β_2.

10. State and explain the Gauss–Markov theorem.
11. Use the least squares estimator to estimate nonlinear relationships and interpret the results.
12. Explain the difference between an explanatory variable that is fixed in repeated samples and an explanatory variable that is random.
13. Explain the term ''random sampling.''

KEYWORDS

assumptions
asymptotic
biased estimator
BLUE
degrees of freedom
dependent variable
deviation from the mean form
econometric model
economic model
elasticity
exogenous variable
Gauss–Markov theorem
heteroskedastic
homoskedastic
independent variable
indicator variable
least squares estimates
least squares estimators
least squares principle
linear estimator
log-linear model
nonlinear relationship
prediction
quadratic model
random error term
random-*x*
regression model
regression parameters
repeated sampling
sampling precision
sampling properties
scatter diagram
simple linear regression analysis
simple linear regression function
specification error
strictly exogenous
unbiased estimator

Economic theory suggests many relationships between economic variables. In microeconomics, you considered demand and supply models in which the quantities demanded and supplied of a good depend on its price. You considered "production functions" and "total product curves" that explained the amount of a good produced as a function of the amount of an input, such as labor, that is used. In macroeconomics, you specified "investment functions" to explain that the amount of aggregate investment in the economy depends on the interest rate and "consumption functions" that related aggregate consumption to the level of disposable income.

Each of these models involves a relationship between economic variables. In this chapter, we consider how to use a sample of economic data to quantify such relationships. As economists, we are interested in questions such as the following: If one variable (e.g., the price of a good) changes in a certain way, *by how much* will another variable (the quantity demanded or supplied) change? Also, given that we know the value of one variable, can we *forecast* or *predict* the corresponding value of another? We will answer these questions by using a **regression model**. Like all models, the regression model is based on **assumptions**. In this chapter, we hope to be very clear about these assumptions, as they are the conditions under which the analysis in subsequent chapters is appropriate.

2.1 An Economic Model

In order to develop the ideas of regression models, we are going to use a simple, but important, economic example. Suppose that we are interested in studying the relationship between household income and expenditure on food. Consider the "experiment" of randomly selecting households from a particular population. The population might consist of households within a particular city, state, province, or country. For the present, suppose that we are interested only in households with an income of $1000 per week. In this experiment, we randomly select a number of households from this population and interview them. We ask the question "How much did you spend per person on food last week?" Weekly food expenditure, which we denote as y, is a *random variable* since the value is unknown to us until a household is selected and the question is asked and answered.

Remark

In the Probability Primer and Appendices B and C, we distinguished random variables from their values by using uppercase (Y) letters for random variables and lowercase (y) letters for their values. We *will not* make this distinction any longer because it leads to complicated notation. We will use lowercase letters, like "y," to denote random variables as well as their values, and we will make the interpretation clear in the surrounding text.

The continuous random variable y has a probability density function (which we will abbreviate as *pdf*) that describes the probabilities of obtaining various food expenditure values. *If you are rusty or uncertain about probability concepts, see the Probability Primer and Appendix B at the end of this book for a comprehensive review*. The amount spent on food per person will vary from one household to another for a variety of reasons: some households will be devoted to gourmet food, some will contain teenagers, some will contain senior citizens, some will be vegetarian, and some will eat at restaurants more frequently. All of these factors and many others, including random, impulsive buying, will cause weekly expenditures on food to vary from one household to another, even if they all have the same income. The *pdf* $f(y)$ describes how expenditures are "distributed" over the population and might look like Figure 2.1.

The *pdf* in Figure 2.1a is actually a conditional *pdf* since it is "conditional" upon household income. If x = weekly household income = \$1000, then the conditional *pdf* is $f(y|x = \$1000)$. The *conditional mean*, or *expected value*, of y is $E(y|x = \$1000) = \mu_{y|x}$ and is our population's mean weekly food expenditure per person.

Remark

The expected value of a random variable is called its "mean" value, which is really a contraction of *population mean*, the center of the probability distribution of the random variable. This is *not* the same as the *sample mean*, which is the arithmetic average of numerical values. Keep the distinction between these two usages of the term "mean" in mind.

The *conditional variance* of y is $\text{var}(y|x = \$1000) = \sigma^2$, which measures the dispersion of household expenditures y about their mean $\mu_{y|x}$. The parameters $\mu_{y|x}$ and σ^2, if they were known, would give us some valuable information about the population we are considering. If we knew these parameters, and if we knew that the conditional distribution $f(y|x = \$1000)$ was *normal*,

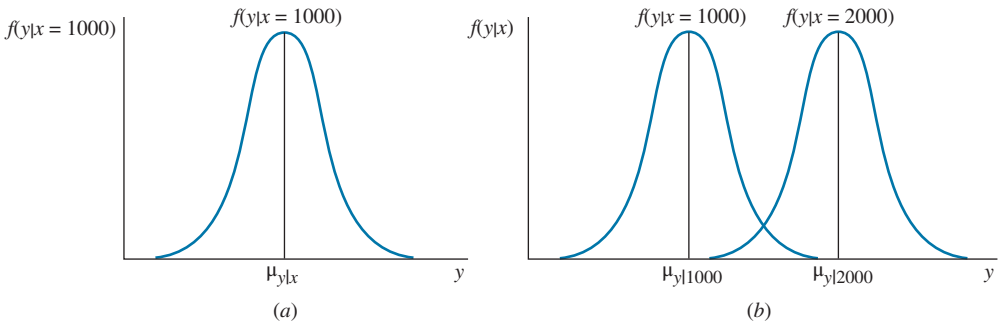


FIGURE 2.1 **(*a*) Probability distribution $f(y|x = 1000)$ of food expenditure y given income $x = \$1000$. (*b*) Probability distributions of food expenditure y given incomes $x = \$1000$ and $x = \$2000$.**

$N(\mu_{y|x}, \sigma^2)$, then we could calculate probabilities that y falls in specific intervals using properties of the normal distribution. That is, we could compute the proportion of the household population that spends between \$50 and \$75 per person on food, given \$1000 per week income.

As economists, we are usually interested in studying relationships between variables, in this case the relationship between y = weekly food expenditure per person and x = weekly household income. Economic theory tells us that expenditure on economic goods depends on income. Consequently, we call y the "**dependent variable**" and x the "**independent**" or "**explanatory**" variable. In econometrics, we recognize that real-world expenditures are random variables, and we want to use data to learn about the relationship.

An econometric analysis of the expenditure relationship can provide answers to some important questions, such as: If weekly income goes up by \$100, **how much** will average weekly food expenditures rise? Or, could weekly food expenditures fall as income rises? How much would we predict the weekly per person expenditure on food to be for a household with an income of \$2000 per week? The answers to such questions provide valuable information for decision makers.

> *Using … per person food spending information … one can determine the similarities and disparities in the spending habits of households of differing sizes, races, incomes, geographic areas, and other socioeconomic and demographic features. This information is valuable for assessing existing market conditions, product distribution patterns, consumer buying habits, and consumer living conditions. Combined with demographic and income projections, this information may be used to anticipate consumption trends. The information may also be used to develop typical market baskets of food for special population groups, such as the elderly. These market baskets may, in turn, be used to develop price indices tailored to the consumption patterns of these population groups*. [*Blisard, Noel, Food Spending in American Households, 1997–1998, Electronic Report from the Economic Research Service, U.S. Department of Agriculture, Statistical Bulletin Number 972, June 2001*]

From a business perspective, if we are managers of a supermarket chain (or restaurant, or health food store, etc.), we must consider long-range plans. If economic forecasters are predicting that local income will increase over the next few years, then we must decide whether, and how much, to expand our facilities to serve our customers. Or, if we plan to open franchises in high-income and low-income neighborhoods, then forecasts of expenditures on food per person, along with neighborhood demographic information, give an indication of how large the stores in those areas should be.

In order to investigate the relationship between expenditure and income, we must build an **economic model** and then a corresponding **econometric model** that forms the basis for a quantitative or *empirical* economic analysis. In our food expenditure example, economic theory suggests that average weekly per person household expenditure on food, represented mathematically by the conditional mean $E(y|x) = \mu_{y|x}$, depends on household income x. If we consider households with different levels of income, we expect the average expenditure on food to change. In Figure 2.1b, we show the *pdfs* of food expenditure for two different levels of weekly income, \$1000 and \$2000. Each conditional *pdf* $f(y|x)$ shows that expenditures will be distributed about a mean value $\mu_{y|x}$, but the mean expenditure by households with higher income is larger than the mean expenditure by lower income households.

In order to use data, we must now specify an *econometric model* that describes how the data on household income and food expenditure are obtained and that guides the econometric analysis.

2.2 An Econometric Model

Given the economic reasoning in the previous section, and to quantify the relationship between food expenditure and income, we must progress from the ideas in Figure 2.1, to an **econometric model**. First, suppose a three-person household has an unwavering rule that each week they

spend \$80 and then also spend 10 cents of each dollar of income received on food. Let y = weekly household food expenditure (\$) and let x = weekly household income (\$). Algebraically their rule is $y = 80 + 0.10x$. Knowing this relationship we calculate that in a week in which the household income is \$1000, the household will spend \$180 on food. If weekly income increases by \$100 to \$1100, then food expenditure increases to \$190. These are **predictions** of food expenditure given income. **Predicting** the value of one variable given the value of another, or others, is one of the primary uses of regression analysis.

A second primary use of regression analysis is to attribute, or relate, changes in one variable to changes in another variable. To that end, let "Δ" denote "change in" in the usual algebraic way. A change in income of \$100 means that $\Delta x = 100$. Because of the spending rule $y = 80 + 0.10x$ the change in food expenditure is $\Delta y = 0.10\Delta x = 0.10(100) = 10$. An increase in income of \$100 leads to, or causes, a \$10 increase in food expenditure. Geometrically, the rule is a line with "y-intercept" 80 and slope $\Delta y/\Delta x = 0.10$. An economist might say that the household "marginal propensity to spend on food is 0.10," which means that from each additional dollar of income 10 cents is spent on food. Alternatively, in a kind of economist shorthand, the "marginal effect of income on food expenditure is 0.10." Much of economic and econometric analysis is an attempt to measure a **causal relationship** between two economic variables. Claiming **causality** here, that is, changing income leads to a change in food expenditure, is quite clear given the household's expenditure rule. It is not always so straightforward.

In reality, many other factors may affect household expenditure on food; the ages and sexes of the household members, their physical size, whether they do physical labor or have desk jobs, whether there is a party following the big game, whether it is an urban or rural household, whether household members are vegetarians or into a paleo-diet, as well as other taste and preference factors ("I really like truffles") and impulse shopping ("Wow those peaches look good!"). Lots of factors. Let e = *everything else* affecting food expenditure other than income. Furthermore, even if a household has a rule, strict or otherwise, we do not know it. To account for these realities, we suppose that the household's food expenditure decision is based on the equation

$$y = \beta_1 + \beta_2 x + e \tag{2.1}$$

In addition to y and x, equation (2.1) contains two unknown **parameters**, β_1 and β_2, instead of "80" and "0.10," and an **error term** e, which represents all those other factors (*everything else*) affecting weekly household food expenditure.

Imagine that we can perform an experiment on the household. Let's increase the household's income by \$100 per week and hold other things constant. Holding other things constant, or holding all else (*everything else*) equal, is the *ceteris paribus* assumption discussed extensively in economic principles courses. Let $\Delta x = 100$ denote the change in household income. Assuming everything else affecting household food expenditure, e, is held constant means that $\Delta e = 0$. The effect of the change in income is $\Delta y = \beta_2 \Delta x + \Delta e = \beta_2 \Delta x = \beta_2 \times 100$. The change in weekly food expenditure $\Delta y = \beta_2 \times 100$ is explained by, or caused by, the change in income. The unknown parameter β_2, the marginal propensity to spend on food from income, tells us the proportion of the increase in income used for food purchases; it answers the "how much" question "How much will food expenditure change given a change in income, holding all else constant?"

The experiment in the previous paragraph is not feasible. We can give a household an extra \$100 income, but we cannot hold all else constant. The simple calculation of the marginal effect of an increase in income on food expenditure $\Delta y = \beta_2 \times 100$ is not possible. However, we can shed light on this "how much" question by using **regression analysis** to estimate β_2. Regression analysis is a statistical method that uses data to explore relationships between variables. A **simple linear regression analysis** examines the relationship between a y-variable and one x-variable. It is said to be "simple" not because it is easy, but because there is only one x-variable. The y-variable is called the dependent variable, the outcome variable, the explained variable, the left-hand-side variable, or the regressand. In our example, the dependent variable is

y = weekly household expenditure on food. The variable x = weekly household income is called the independent variable, the explanatory variable, the right-hand-side variable, or the regressor. Equation (2.1) is the **simple linear regression** model.

All models are abstractions from reality and working with models requires assumptions. The same is true for the regression model. The first assumption of the simple linear regression model is that relationship (2.1) holds for the members of the population under consideration. For example, define the population to be three-person households in a given geographic region, say southern Australia. The unknowns β_1 and β_2 are called **population parameters**. We assert the behavioral rule $y = \beta_1 + \beta_2 x + e$ holds for all households in the population. Each week food expenditure equals β_1, plus a proportion β_2 of income, plus other factors, e.

The field of statistics was developed because, in general, populations are large, and it is impossible (or impossibly costly) to examine every population member. The population of three-person households in a given geographic region, even if it is only a medium-sized city, is too large to survey individually. Statistical and econometric methodology examines and analyzes a **sample of data** from the population. After analyzing the data, we make **statistical inferences**. These are conclusions or judgments about a population based on the data analysis. Great care must be taken when drawing inferences. The inferences are conclusions about the particular population from which the data were collected. Data on households from southern Australia may, or may not, be useful for making inferences, drawing conclusions, about households from the southern United States. Do Melbourne, Australia, households have the same food spending patterns as households in New Orleans, Louisiana? That might be an interesting research topic. If not, then we may not be able to draw valid conclusions about New Orleans household behavior from the sample of Australian data.

2.2.1 Data Generating Process

The sample of data, and how the data are actually obtained, is crucially important for subsequent inferences. The exact mechanisms for collecting a sample of data are very discipline specific (e.g., agronomy is different from economics) and beyond the scope of this book.[1] For the household food expenditure example, let us assume that we can obtain a sample at a point in time [these are **cross-sectional data**] consisting of N data pairs that are **randomly** selected from the population. Let (y_i, x_i) denote the ith data pair, $i = 1, \ldots, N$. The variables y_i and x_i are **random variables**, because their values are not known until they are observed. Randomly selecting households makes the first observation pair (y_1, x_1) statistically independent of all other data pairs, and each observation pair (y_i, x_i) is **statistically independent** of every other data pair, (y_j, x_j), where $i \neq j$. We further assume that the random variables y_i and x_i have a joint $pdf\, f(y_i, x_i)$ that describes their distribution of values. We often do not know the exact nature of the joint distribution (such as bivariate normal; see Probability Primer, Section P.7.1), but all pairs drawn from the same population are assumed to follow the same joint pdf, and, thus, the data pairs are not only statistically independent but are also **identically distributed** (abbreviated **i.i.d.** or *iid*). Data pairs that are *iid* are said to be a **random sample**.

If our first assumption is true, that the behavioral rule $y = \beta_1 + \beta_2 x + e$ holds for all households in the population, then restating (2.1) for each (y_i, x_i) data pair

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N \tag{2.1}$$

This is sometimes called the **data generating process** (**DGP**) because we assume that the observable data follow this relationship.

[1] See, for example, Paul S. Levy and Stanley Lemeshow (2008) *Sampling of Populations: Methods and Applications, 4th Edition*, Hoboken, NJ: John Wiley and Sons, Inc.

2.2.2 The Random Error and Strict Exogeneity

The second assumption of the simple regression model (2.1) concerns the "everything else" term e. The variables (y_i, x_i) are random variables because we do not know what values they take until a particular household is chosen and they are observed. The error term e_i is also a random variable. All the other factors affecting food expenditure except income will be different for each population household if for no other reason that everyone's tastes and preferences are different. Unlike food expenditure and income, the **random error term** e_i is **not observable**; it is **unobservable**. We cannot measure tastes and preferences in any direct way, just as we cannot directly measure the economic "utility" derived from eating a slice of cake. The second regression assumption is that the x-variable, income, cannot be used to predict the value of e_i, the effect of the collection of all other factors affecting the food expenditure by the ith household. Given an income value x_i for the ith household, the *best* (optimal) predictor[2] of the random error e_i is the conditional expectation, or conditional mean, $E(e_i|x_i)$. The assumption that x_i cannot be used to predict e_i is equivalent to saying $E(e_i|x_i) = 0$. That is, given a household's income we cannot do any better than predicting that the random error is zero; the effects of all other factors on food expenditure average out, in a very specific way, to zero. We will discuss other situations in which this might or might not be true in Section 2.10. For now, recall from the Probability Primer, Section P.6.5, that $E(e_i|x_i) = 0$ has two implications. The first is $E(e_i|x_i) = 0 \Longrightarrow E(e_i) = 0$; if the conditional expected value of the random error is zero, then the **unconditional expectation** of the random error is also zero. In the population, the average effect of all the omitted factors summarized by the random error term is zero.

The second implication is $E(e_i|x_i) = 0 \Longrightarrow \text{cov}(e_i, x_i) = 0$. If the conditional expected value of the random error is zero, then e_i, the random error for the ith observation, has covariance zero and correlation zero, with the corresponding observation x_i. In our example, the random component e_i, representing all factors affecting food expenditure except income for the ith household, is uncorrelated with income for that household. You might wonder how that could possibly be shown to be true. After all, e_i is unobservable. The answer is that it is very hard work. You must convince yourself and your audience that anything that might have been omitted from the model is not correlated with x_i. The primary tool is economic reasoning: your own intellectual experiments (i.e., thinking), reading literature on the topic and discussions with colleagues or classmates. And we really can't prove that $E(e_i|x_i) = 0$ is true with absolute certainty in most economic models.

We noted that $E(e_i|x_i) = 0$ has two implications. If either of the implications is **not** true, then $E(e_i|x_i) = 0$ is not true, that is,

$$E(e_i|x_i) \neq 0 \text{ if (i) } E(e_i) \neq 0 \text{ or if (ii) } \text{cov}(e_i, x_i) \neq 0$$

In the first case, if the population average of the random errors e_i is not zero, then $E(e_i|x_i) \neq 0$. In a certain sense, we will be able to work around the case when $E(e_i) \neq 0$, say if $E(e_i) = 3$, as you will see below. The second implication of $E(e_i|x_i) = 0$ is that $\text{cov}(e_i, x_i) = 0$; the random error for the ith observation has zero covariance and correlation with the ith observation on the explanatory variable. If $\text{cov}(e_i, x_i) = 0$, the explanatory variable x is said to be **exogenous**, providing our first assumption that the pairs (y_i, x_i) are *iid* holds. When x is exogenous, regression analysis can be used successfully to estimate β_1 and β_2. To differentiate the weaker condition $\text{cov}(e_i, x_i) = 0$, simple **exogeneity**, from the stronger condition $E(e_i|x_i) = 0$, we say that x is **strictly exogenous** if $E(e_i|x_i) = 0$. If $\text{cov}(e_i, x_i) \neq 0$, then x is said to be **endogenous**. When x is endogenous, it is more difficult, sometimes much more difficult, to carry out statistical inference. A great deal will be said about exogeneity and strict exogeneity in the remainder of this book.

[2] You will learn about optimal prediction in Appendix 4C.

EXAMPLE 2.1 | A Failure of the Exogeneity Assumption

Consider a regression model exploring the relationship between a working person's wage and their years of education, using a random sample of data. The simple regression model is $WAGE_i = \beta_1 + \beta_2 EDUC_i + e_i$, where $WAGE_i$ is the hourly wage rate of the ith randomly selected person and $EDUC_i$ is their years of education. The pairs $(WAGE_i, EDUC_i)$ from the random sample are assumed to be *iid*. In this model, the random error e_i accounts for all those factors other than $EDUC_i$ that affect a person's wage rate. What might some of those factors be? Ability, intelligence, perseverance, and industriousness are all important characteristics of an employee and likely to influence their wage rate. Are any of these factors which are bundled into e_i likely to be correlated with $EDUC_i$? A few moments reflection will lead you to say "yes." It is very plausible that those with higher education have higher ability, intelligence, perseverance, and industriousness. Thus, there is a strong argument that $EDUC_i$ is an endogenous regressor in this regression and that the strict exogeneity assumption fails.

2.2.3 The Regression Function

The importance of the strict exogeneity assumption is the following. If the strict exogeneity assumption $E(e_i|x_i) = 0$ is true, then the conditional expectation of y_i given x_i is

$$E(y_i|x_i) = \beta_1 + \beta_2 x_i + E(e_i|x_i) = \beta_1 + \beta_2 x_i, \quad i = 1, \ldots, N \tag{2.2}$$

The conditional expectation $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ in (2.2) is called the **regression function**, or **population regression function**. It says that in the population the average value of the dependent variable for the ith observation, conditional on x_i, is given by $\beta_1 + \beta_2 x_i$. It also says that given a change in x, Δx, the resulting change in $E(y_i|x_i)$ is $\beta_2 \Delta x$ **holding all else constant**, in the sense that given x_i the average of the random errors is zero, and any change in x is not correlated with any corresponding change in the random error e. In this case, we can say that a change in x leads to, or **causes**, a change in the expected (population average) value of y given x_i, $E(y_i|x_i)$.

The regression function in (2.2) is shown in Figure 2.2, with y-intercept $\beta_1 = E(y_i|x_i = 0)$ and slope

$$\beta_2 = \frac{\Delta E(y_i|x_i)}{\Delta x_i} = \frac{dE(y_i|x_i)}{dx_i} \tag{2.3}$$

where Δ denotes "change in" and $dE(y|x)/dx$ denotes the "derivative" of $E(y|x)$ with respect to x. We will not use derivatives to any great extent in this book, and if you are not too familiar with the concept you can think of "d" as a stylized version of Δ and go on. See Appendix A.3 for a discussion of derivatives.

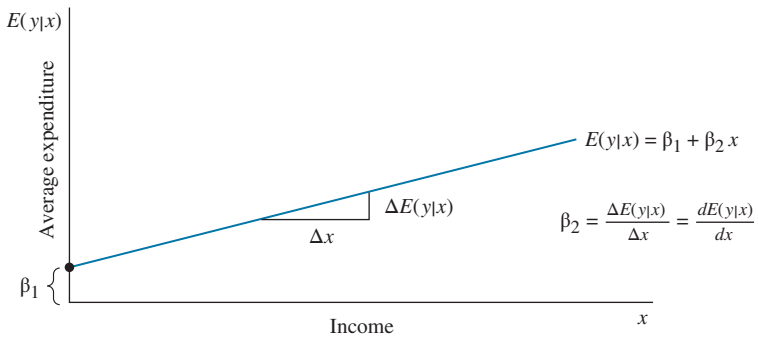


FIGURE 2.2 **The economic model: a linear relationship between average per person food expenditure and income.**

EXAMPLE 2.2 | Strict Exogeneity in the Household Food Expenditure Model

The strict exogeneity assumption is that the average of *everything else* affecting the food expenditure of the ith household, given the income of the ith household, is zero. Could this be true? One test of this possibility is the question "Using the income of the ith household, can we predict the value of e_i, the combined influence of all factors affecting food expenditure other than income?" If the answer is yes, then strict exogeneity fails. If not, then $E(e_i|x_i) = 0$ *may be* a plausible assumption. And if it is, then equation (2.1) can be interpreted as a causal model, and β_2 can be thought of as the marginal effect of income on expected (average) household food expenditure, holding all else constant, as shown in equation (2.3). If $E(e_i|x_i) \neq 0$ then x_i can be used to predict a nonzero value for e_i, which in turn will affect the value of y_i. In this case, β_2 will not capture all the effects of an income change, and the model cannot be interpreted as causal.

Another important consequence of the assumption of strict exogeneity is that it allows us to think of the econometric model as decomposing the dependent variable into two components: one that varies systematically as the values of the independent variable change and another that is random "noise." That is, the econometric model $y_i = \beta_1 + \beta_2 x_i + e_i$ can be broken into two parts: $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ and the random error, e_i. Thus

$$y_i = \beta_1 + \beta_2 x_i + e_i = E(y_i|x_i) + e_i$$

The values of the dependent variable y_i vary systematically due to variation in the conditional mean $E(y_i|x_i) = \beta_1 + \beta_2 x_i$, as the value of the explanatory variable changes, and the values of the dependent variable y_i vary randomly due to e_i. The conditional *pdf*s of e and y are identical except for their location, as shown in Figure 2.3. Two values of food expenditure y_1 and y_2 for households with $x = \$1000$ of weekly income are shown in Figure 2.4 relative to their conditional mean. There will be variation in household expenditures on food from one household to another because of variations in tastes and preferences, and everything else. Some will spend more than the average value for households with the same income, and some will spend less. If we knew β_1 and β_2, then we could compute the conditional mean expenditure $E(y_i|x = 1000) = \beta_1 + \beta_2(1000)$ and also the value of the random errors e_1 and e_2. We never know β_1 and β_2 so we can never compute e_1 and e_2. What we are assuming, however, is that at each level of income x the average value of all that is represented by the random error is zero.

2.2.4 Random Error Variation

We have made the assumption that the conditional expectation of the random error term is zero, $E(e_i|x_i) = 0$. For the random error term we are interested in both its conditional mean,

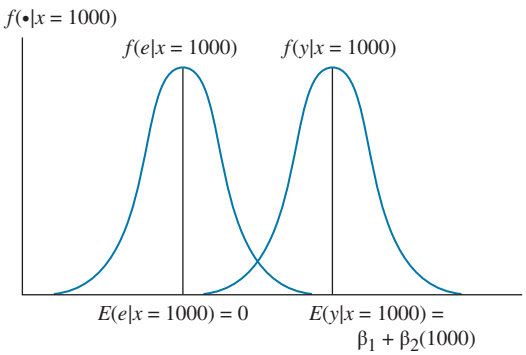


FIGURE 2.3 **Conditional probability densities for e and y.**

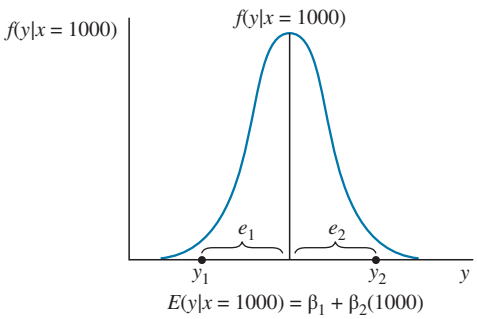


FIGURE 2.4 **The random error.**

or expected value, and its variance. Ideally, the **conditional variance** of the random error is constant.

$$\text{var}(e_i|x_i) = \sigma^2 \tag{2.4}$$

This is the **homoskedasticity** (also spelled homoscedasticity) assumption. At each x_i the variation of the random error component is the same. Assuming the population relationship $y_i = \beta_1 + \beta_2 x_i + e_i$ the conditional variance of the dependent variable is

$$\text{var}(y_i|x_i) = \text{var}(\beta_1 + \beta_2 x_i + e_i|x_i) = \text{var}(e_i|x_i) = \sigma^2$$

The simplification works because by conditioning on x_i we are treating it as if it is known and therefore not random. Given x_i the component $\beta_1 + \beta_2 x_i$ is not random, so the variance rule (P.14) applies.

This was an explicit assumption in Figure 2.1(b) where the *pdf*s $f(y|x = 1000)$ and $f(y|x = 2000)$ have the same variance, σ^2. If strict exogeneity holds, then the regression function is $E(y_i|x_i) = \beta_1 + \beta_2 x_i$, as shown in Figure 2.2. The conditional distributions $f(y|x = 1000)$ and $f(y|x = 2000)$ are placed along the conditional mean function in Figure 2.5. In the household expenditure example, the idea is that for a particular level of household income x, the values of household food expenditure will vary randomly about the conditional mean due to the assumption that at each x the average value of the random error e is zero. Consequently, at each level of income, household food expenditures are centered on the regression function. The conditional homoskedasticity assumption implies that at each level of income the variation in food

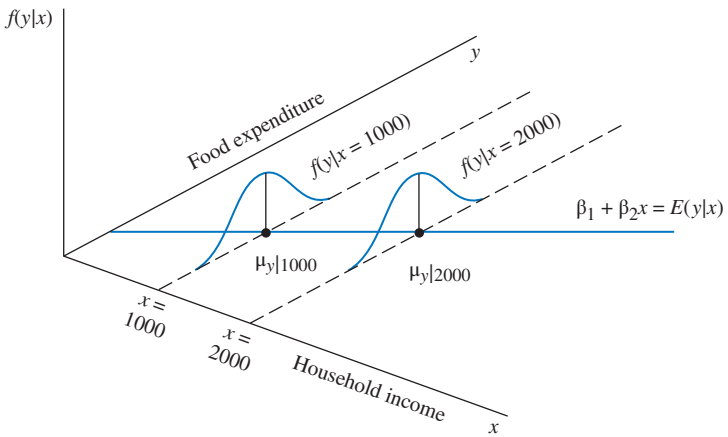


FIGURE 2.5 **The conditional probability density functions for y, food expenditure, at two levels of income.**

expenditure about its mean is the same. That means that at each and every level of income we are *equally* uncertain about how far food expenditures might fall from their mean value, $E(y_i|x_i) = \beta_1 + \beta_2 x_i$. Furthermore, this uncertainty does not depend on income or anything else. If this assumption is violated, and $\text{var}(e_i|x_i) \neq \sigma^2$, then the random errors are said to be **heteroskedastic**.

2.2.5 Variation in x

In a regression analysis, one of the objectives is to estimate $\beta_2 = \Delta E(y_i|x_i)/\Delta x_i$. If we are to hope that a sample of data can be used to estimate the effects of changes in x, then we must observe some different values of the explanatory variable x in the sample. Intuitively, if we collect data **only** on households with income \$1000, we will not be able to measure the effect of changing income on the average value of food expenditure. Recall from elementary geometry that "it takes two points to determine a line." The minimum number of x-values in a sample of data that will allow us to proceed is two. You will find out in Section 2.4.4 that in fact the more different values of x, and the more variation they exhibit, the better our regression analysis will be.

2.2.6 Error Normality

In the discussion surrounding Figure 2.1, we explicitly made the assumption that food expenditures, given income, were normally distributed. In Figures 2.3–2.5, we implicitly made the assumption of conditionally normally distributed errors and dependent variable by drawing classically bell-shaped curves. It is not at all necessary for the random errors to be conditionally normal in order for regression analysis to "work." However, as you will discover in Chapter 3, when samples are small, it is advantageous for statistical inferences that the random errors, and dependent variable y, given each x-value, are normally distributed. The normal distribution has a long and interesting history,[3] as a little Internet searching will reveal. One argument for assuming regression errors are normally distributed is that they represent a collection of many different factors. The **Central Limit Theorem**, see Appendix C.3.4, says roughly that collections of many random factors tend toward having a normal distribution. In the context of the food expenditure model, if we consider that the random errors reflect tastes and preferences, it is entirely plausible that the random errors at each income level are normally distributed. When the assumption of conditionally normal errors is made, we write $e_i|x_i \sim N(0, \sigma^2)$ and also then $y_i|x_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$. It is a very strong assumption when it is made, and as mentioned it is not strictly speaking necessary, so we call it an *optional* assumption.

2.2.7 Generalizing the Exogeneity Assumption

So far we have assumed that the data pairs (y_i, x_i) have been drawn from a random sample and are *iid*. What happens if the sample values of the explanatory variable are correlated? And how might that happen?

A lack of independence occurs naturally when using financial or macroeconomic **time-series** data. Suppose we observe the monthly report on new housing starts, y_t, and the current 30-year fixed mortgage rate, x_t, and we postulate the model $y_t = \beta_1 + \beta_2 x_t + e_t$. The data (y_t, x_t) can be described as macroeconomic **time-series** data. In contrast to cross-section data where we have observations on a number of units (say households or firms or persons or countries) at a given point in time, with time-series data we have observations over time on a number of variables. It is customary to use a "t" subscript to denote time-series data and to use T to denote the sample size. In the data pairs (y_t, x_t), $t = 1, \ldots, T$, both y_t and x_t are random because we do not know the values

[3] For example, Stephen M. Stigler (1990) *The History of Statistics: The Measurement of Uncertainty, Reprint Edition*, Belknap Press, 73–76.

until they are observed. Furthermore, each of the data series is likely to be correlated across time. For example, the monthly fixed mortgage rate is likely to change slowly over time making the rate at time t correlated with the rate at time $t-1$. The assumption that the pairs (y_t, x_t) represent random *iid* draws from a probability distribution is not realistic. When considering the exogeneity assumption for this case, we need to be concerned not just with possible correlation between x_t and e_t, but also with possible correlation between e_t and every other value of the explanatory variable, namely, x_s, $s = 1, 2, \ldots, T$. If x_s is correlated with x_t, then it is possible that x_s (say, the mortgage rate in one month) may have an impact on y_t (say, housing starts in the next month). Since it is x_t, not x_s that appears in the equation $y_t = \beta_1 + \beta_2 x_t + e_t$, the effect of x_s will be included in e_t, implying $E(e_t|x_s) \neq 0$. We could use x_s to help predict the value of e_t. This possibility is ruled out when the pairs (y_t, x_t) are assumed to be independent. That is, independence of the pairs (y_t, x_t) **and** the assumption $E(e_t|x_t) = 0$ imply $E(e_t|x_s) = 0$ for all $s = 1, 2, \ldots, T$.

To extend the strict exogeneity assumption to models where the values of x are correlated, we need to assume $E(e_t|x_s) = 0$ for all $(t, s) = 1, 2, \ldots, T$. This means that we cannot predict the random error at time t, e_t, using any of the values of the explanatory variable. Or, in terms of our earlier notation, $E(e_i|x_j) = 0$ for all $(i, j) = 1, 2, \ldots, N$. To write this assumption in a more convenient form, we introduce the notation $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. That is, we are using $\mathbf{x}$ to denote all sample observations on the explanatory variable. Then, a more general way of writing the strict exogeneity assumption is $E(e_i|\mathbf{x}) = 0, i = 1, 2, \ldots, N$. From this assumption, we can also write $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$ for $i = 1, 2, \ldots, N$. This assumption is discussed further in the context of alternative types of data in Section 2.10 and in Chapter 9. The assumption $E(e_i|\mathbf{x}) = 0$, $i = 1, 2, \ldots, N$, is a weaker assumption than assuming $E(e_i|x_i) = 0$ **and** that the pairs (y_i, x_i) are independent, and it enables us to derive a number of results for cases where different observations on x may be correlated as well as for the case where they are independent.

2.2.8 Error Correlation

In addition to possible correlations between a random error for one household (e_i) or one time period (e_t) being correlated with the value of an explanatory variable for another household (x_j) or time period (x_s), it is possible that there are correlations between the random error terms.

With cross-sectional data, data on households, individuals, or firms collected at one point in time, there may be a lack of statistical independence between random errors for individuals who are **spatially** connected. That is, suppose that we collect observations on two (or more) individuals who live in the same neighborhood. It is very plausible that there are similarities among people who live in a particular neighborhood. Neighbors can be expected to have similar incomes if the homes in a neighborhood are homogenous. Some suburban neighborhoods are popular because of green space and schools for young children, meaning households may have members similar in ages and interests. We might add a spatial component s to the error and say that the random errors $e_i(s)$ and $e_j(s)$ for the ith and jth households are possibly correlated because of their common location. Within a larger sample of data, there may be **clusters** of observations with correlated errors because of the spatial component.

In a time-series context, your author is writing these pages on the tenth anniversary of Hurricane Katrina, which devastated the U.S. Gulf Coast and the city of New Orleans, Louisiana, in particular. The impact of that shock did not just happen and then go away. The effect of that huge random event had an effect on housing and financial markets during August 2005, and also in September, October, and so on, to this day. Consequently, the random errors in the population relationship $y_t = \beta_1 + \beta_2 x_t + e_t$ are correlated over time, so that $\text{cov}(e_t, e_{t+1}) \neq 0$, $\text{cov}(e_t, e_{t+2}) \neq 0$, and so on. This is called **serial correlation**, or **autocorrelation**, in econometrics.

The starting point in regression analysis is to assume that there is no error correlation. In time-series models, we start by assuming $\text{cov}(e_t, e_s|\mathbf{x}) = 0$ for $t \neq s$, and for cross-sectional data we start by assuming $\text{cov}(e_i, e_j|\mathbf{x}) = 0$ for $i \neq j$. We will cope with failure of these assumptions in Chapter 9.

2.2.9 Summarizing the Assumptions

We summarize the starting assumptions of the simple regression model in a very general way. In our summary we use subscripts i and j but the assumptions are general, and apply equally to time-series data. If these assumptions hold, then regression analysis can successfully estimate the unknown population parameters β_1 and β_2 and we can claim that $\beta_2 = \Delta E(y_i|x_i)/\Delta x_i = dE(y_i|x_i)/dx_i$ measures a causal effect. We begin our study of regression analysis and econometrics making these strong assumptions about the DGP. For future reference, the assumptions are named SR1–SR6, "SR" denoting "simple regression."

Econometrics is in large part devoted to handling data and models for which these assumptions **may not** hold, leading to modifications of usual methods for estimating β_1 and β_2, testing hypotheses, and predicting outcomes. In Chapters 2 and 3, we study the simple regression model under these, or similar, strong assumptions. In Chapter 4, we introduce modeling issues and diagnostic testing. In Chapter 5, we extend our model to **multiple regression analysis** with more than one explanatory variable. In Chapter 6, we treat modeling issues concerning the multiple regression model, and starting in Chapter 8 we address situations in which SR1–SR6 are violated in one way or another.

Assumptions of the Simple Linear Regression Model

SR1: Econometric Model All data pairs (y_i, x_i) collected from a population satisfy the relationship

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

SR2: Strict Exogeneity The conditional expected value of the random error e_i is zero. If $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, then

$$E(e_i|\mathbf{x}) = 0$$

If strict exogeneity holds, then the population regression function is

$$E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i, \quad i = 1, \ldots, N$$

and

$$y_i = E(y_i|\mathbf{x}) + e_i, \quad i = 1, \ldots, N$$

SR3: Conditional Homoskedasticity The conditional variance of the random error is constant.

$$\text{var}(e_i|\mathbf{x}) = \sigma^2$$

SR4: Conditionally Uncorrelated Errors The conditional covariance of random errors e_i and e_j is zero.

$$\text{cov}(e_i, e_j|\mathbf{x}) = 0 \quad \text{for} \quad i \neq j$$

SR5: Explanatory Variable Must Vary In a sample of data, x_i must take at least two different values.

SR6: Error Normality (*optional*) The conditional distribution of the random errors is normal.

$$e_i|\mathbf{x} \sim N(0, \sigma^2)$$

The random error e and the dependent variable y are both random variables, and as we have shown, the properties of one variable can be determined from the properties of the other. There is, however, one interesting difference between them: y is "observable" and e is "unobservable."

If the **regression parameters** β_1 and β_2 were *known*, then for any values y_i and x_i we could calculate $e_i = y_i - (\beta_1 + \beta_2 x_i)$. This is illustrated in Figure 2.4. Knowing the regression function $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$ we could separate y_i into its systematic and random parts. However, β_1 and β_2 *are never known*, and it is impossible to calculate e_i.

What comprises the error term e? The random error e represents all factors affecting y other than x, or what we have called *everything else*. These factors cause individual observations y_i to differ from the conditional mean value $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$. In the food expenditure example, what factors can result in a difference between household expenditure per person y_i and its conditional mean $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$?

1. We have included income as the only explanatory variable in this model. Any *other* economic factors that affect expenditures on food are "collected" in the error term. Naturally, in any economic model, we want to include all the important and relevant explanatory variables in the model, so the error term e is a "storage bin" for unobservable and/or unimportant factors affecting household expenditures on food. As such, it adds noise that masks the relationship between x and y.
2. The error term e captures any approximation error that arises because the *linear* functional form we have assumed may be only an approximation to reality.
3. The error term captures any elements of random behavior that may be present in each individual. Knowing all the variables that influence a household's food expenditure might not be enough to perfectly predict expenditure. Unpredictable human behavior is also contained in e.

If we have omitted some important factor, or made any other serious **specification error**, then assumption SR2 $E(e_i|\mathbf{x}) = 0$ will be violated, which will have serious consequences.

2.3 Estimating the Regression Parameters

EXAMPLE 2.3 | Food Expenditure Model Data

The economic and econometric models we developed in the previous section are the basis for using a sample of data to *estimate* the intercept and slope parameters, β_1 and β_2. For illustration we examine typical data on household food expenditure and weekly income from a random sample of 40 households. Representative observations and summary statistics are given in Table 2.1. We control for household size by considering only three-person households. The values of y are weekly food expenditures for a three-person household, in dollars. Instead of measuring income in dollars, we measure it in units of $100, because a $1 increase in income has a numerically small effect on food expenditure. Consequently, for the first household, the reported income is $369 per week with weekly food expenditure of $115.22. For the 40th household, weekly income is $3340 and weekly food expenditure is $375.73. The complete data set of observations is in the data file *food*.

TABLE 2.1 Food Expenditure and Income Data

Observation (household)	Food Expenditure ($)	Weekly Income ($100)
i	y_i	x_i
1	115.22	3.69
2	135.98	4.39
	⋮	
39	257.95	29.40
40	375.73	33.40
	Summary Statistics	
Sample mean	283.5735	19.6048
Median	264.4800	20.0300
Maximum	587.6600	33.4000
Minimum	109.7100	3.6900
Std. dev.	112.6752	6.8478

Remark

In this book, data files are referenced with a descriptive name in italics such as *food*. The actual files which are located at the book websites www.wiley.com/college/hill and www.principlesofeconometrics.com come in various formats and have an extension that denotes the format, for example, *food.dat*, *food.wf1*, *food.dta*, and so on. The corresponding data definition file is *food.def*.

We assume that the expenditure data in Table 2.1 satisfy the assumptions SR1–SR5. That is, we assume that the regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ describes a population relationship and that the random error has conditional expected value zero. This implies that the conditional expected value of household food expenditure is a linear function of income. The conditional variance of y, which is the same as that of the random error e, is assumed constant, implying that we are equally uncertain about the relationship between y and x for all observations. Given **x** the values of y for different households are assumed uncorrelated with each other.

Given this theoretical model for explaining the sample observations on household food expenditure, the problem now is how to use the sample information in Table 2.1, specific values of y_i and x_i, to estimate the unknown regression parameters β_1 and β_2. These parameters represent the unknown intercept and slope coefficients for the food expenditure–income relationship. If we represent the 40 data points as (y_i, x_i), $i = 1, \ldots, N = 40$, and plot them, we obtain the **scatter diagram** in Figure 2.6.

Remark

It will be our notational convention to use i subscripts for cross-sectional data observations, with the number of sample observations being N. For time-series data observations, we use the subscript t and label the total number of observations T. In purely algebraic or generic situations, we may use one or the other.

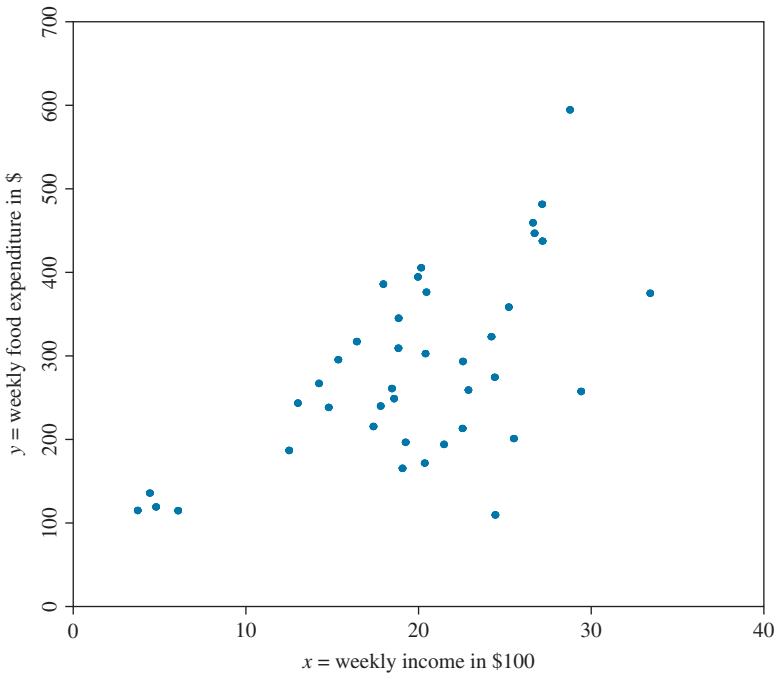


FIGURE 2.6 **Data for the food expenditure example.**

Our problem is to estimate the location of the mean expenditure line $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$. We would expect this line to be somewhere in the middle of all the data points since it represents population mean, or average, behavior. To estimate β_1 and β_2, we could simply draw a freehand line through the middle of the data and then measure the slope and intercept with a ruler. The problem with this method is that different people would draw different lines, and the lack of a formal criterion makes it difficult to assess the accuracy of the method. Another method is to draw a line from the expenditure at the smallest income level, observation $i = 1$, to the expenditure at the largest income level, $i = 40$. This approach does provide a formal rule. However, it may not be a very good rule because it ignores information on the exact position of the remaining 38 observations. It would be better if we could devise a rule that uses all the information from all the data points.

2.3.1 The Least Squares Principle

To estimate β_1 and β_2 we want a rule, or formula, that tells us how to make use of the sample observations. Many rules are possible, but the one that we will use is based on the **least squares principle**. This principle asserts that to fit a line to the data values we should make the sum of the squares of the vertical distances from each point to the line as small as possible. The distances are squared to prevent large positive distances from being canceled by large negative distances. This rule is arbitrary, but very effective, and is simply one way to describe a line that runs through the middle of the data. The intercept and slope of this line, the line that best fits the data using the least squares principle, are b_1 and b_2, the **least squares estimates** of β_1 and β_2. The fitted line itself is then

$$\hat{y}_i = b_1 + b_2 x_i \tag{2.5}$$

The vertical distances from each point to the fitted line are the **least squares residuals**. They are given by

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i \tag{2.6}$$

These residuals are depicted in Figure 2.7a.

Now suppose we fit another line, *any other line*, to the data. Denote the new line as

$$\hat{y}_i^* = b_1^* + b_2^* x_i$$

where b_1^* and b_2^* are any other intercept and slope values. The residuals for this line, $\hat{e}_i^* = y_i - \hat{y}_i^*$, are shown in Figure 2.7b. The least squares estimates b_1 and b_2 have the property that the sum of their squared residuals is *less than* the sum of squared residuals for *any* other line. That is, if

$$SSE = \sum_{i=1}^{N} \hat{e}_i^2$$

is the sum of squared least squares residuals from (2.6) and

$$SSE^* = \sum_{i=1}^{N} \hat{e}_i^{*2} = \sum_{i=1}^{N} \left(y_i - \hat{y}_i^*\right)^2$$

is the sum of squared residuals based on any other estimates, then

$$SSE < SSE^*$$

no matter how the other line might be drawn through the data. The least squares principle says that the estimates b_1 and b_2 of β_1 and β_2 are the ones to use, since the line using them as intercept and slope fits the data best.

The problem is to find b_1 and b_2 in a convenient way. Given the sample observations on y and x, we want to find values for the unknown parameters β_1 and β_2 that minimize the "sum of squares" function

$$S(\beta_1, \beta_2) = \sum_{i=1}^{N} \left(y_i - \beta_1 - \beta_2 x_i\right)^2$$

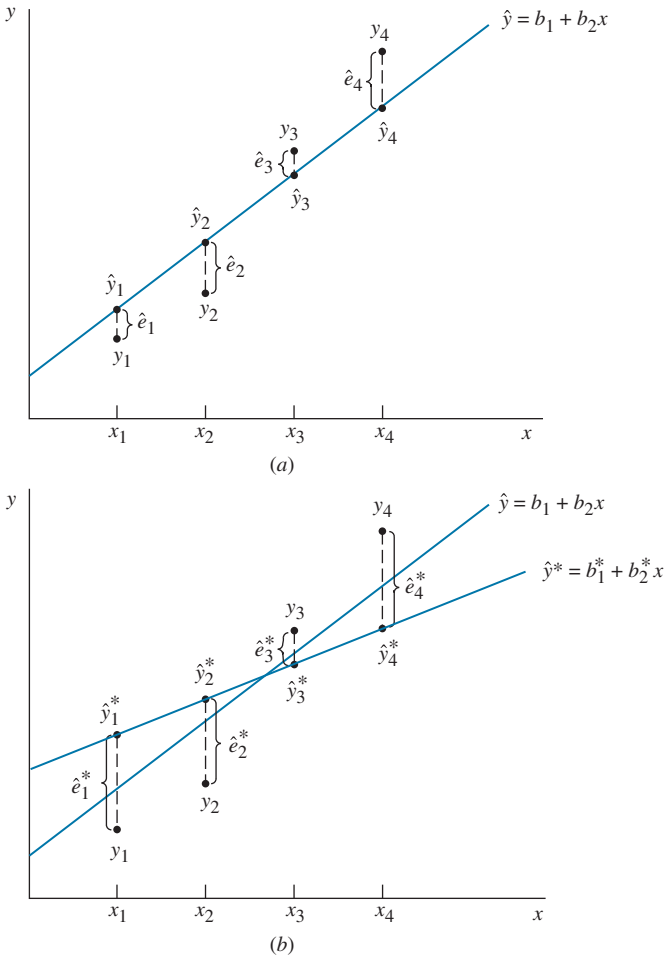


FIGURE 2.7 **(*a*) The relationship among y, $\hat{e}$, and the fitted regression line. (*b*) The residuals from another fitted line.**

This is a straightforward calculus problem, the details of which are given in Appendix 2A. The formulas for the least squares estimates of β_1 and β_2 that give the minimum of the sum of squared residuals are

The Ordinary Least Squares (OLS) Estimators

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{2.7}$$

$$b_1 = \bar{y} - b_2\bar{x} \tag{2.8}$$

where $\bar{y} = \sum y_i/N$ and $\bar{x} = \sum x_i/N$ are the sample means of the observations on y and x.

We will call the estimators b_1 and b_2, given in equations (2.7) and (2.8), the **ordinary least squares estimators**. "Ordinary least squares" is abbreviated as **OLS**. These least squares estimators are called "ordinary," despite the fact that they are extraordinary, because these estimators are used day in and day out in many fields of research in a routine way, and to distinguish them

from other methods called **generalized least squares**, and **weighted least squares**, and **two-stage least squares**, all of which are introduced later in this book.

The formula for b_2 reveals why we had to assume [SR5] that in the sample x_i must take at least two different values. If $x_i = 5$, for example, for all observations, then b_2 in (2.7) is mathematically undefined and does not exist since its numerator and denominator are zero!

If we plug the sample values y_i and x_i into (2.7) and (2.8), then we obtain the least squares *estimates* of the intercept and slope parameters β_1 and β_2. It is interesting, however, and very important, that the formulas for b_1 and b_2 are perfectly general and can be used no matter what the sample values turn out to be. This should ring a bell. When the formulas for b_1 and b_2 are taken to be rules that are used whatever the sample data turn out to be, then b_1 and b_2 are random variables. When actual sample values are substituted into the formulas, we obtain numbers that are the observed values of random variables. To distinguish these two cases, we call the rules or general formulas for b_1 and b_2 the **least squares estimators**. We call the numbers obtained when the formulas are used with a particular sample **least squares estimates**.

- Least squares *estimators* are general formulas and are *random variables*.
- Least squares *estimates* are numbers that we obtain by applying the general formulas to the observed data.

The distinction between *estimators* and *estimates* is a fundamental concept that is essential to understand everything in the rest of this book.

EXAMPLE 2.4a | Estimates for the Food Expenditure Function

Using the least squares estimators (2.7) and (2.8), we can obtain the least squares estimates for the intercept and slope parameters β_1 and β_2 in the food expenditure example using the data in Table 2.1. From (2.7), we have

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{18671.2684}{1828.7876} = 10.2096$$

and from (2.8)

$$b_1 = \bar{y} - b_2\bar{x} = 283.5735 - (10.2096)(19.6048) = 83.4160$$

A convenient way to report the values for b_1 and b_2 is to write out the *estimated* or *fitted* regression line, with the estimates rounded appropriately:

$$\hat{y}_i = 83.42 + 10.21x_i$$

This line is graphed in Figure 2.8. The line's slope is 10.21, and its intercept, where it crosses the vertical axis, is 83.42. The least squares fitted line passes through the middle of the data in a very precise way, since one of the characteristics of the fitted line based on the least squares parameter estimates is that it passes through the point defined by the sample means, $(\bar{x}, \bar{y}) = (19.6048, 283.5735)$. This follows directly from rewriting (2.8) as $\bar{y} = b_1 + b_2\bar{x}$. Thus, the "point of the means" is a useful reference value in regression analysis.

Interpreting the Estimates

Once obtained, the least squares estimates are interpreted in the context of the economic model under consideration. The value $b_2 = 10.21$ is an estimate of β_2. Recall that x, weekly household income, is measured in \$100 units. The regression slope β_2 is the amount by which expected weekly expenditure on food per household increases when household weekly income increases by \$100. Thus, we estimate that if weekly household income goes up by \$100, expected weekly expenditure on food will increase by approximately \$10.21, holding all else constant. A supermarket executive with information on likely changes in the income and the number of households in an area could estimate that it will sell \$10.21 more per typical household per week for every \$100 increase in income. This is a very valuable piece of information for long-run planning.

Strictly speaking, the intercept estimate $b_1 = 83.42$ is an estimate of the expected weekly food expenditure for a household with zero income. In most economic models we must be very careful when interpreting the estimated intercept. The problem is that we often do not have any data points near $x = 0$, something that is true for the food expenditure data shown in Figure 2.8. If we have no observations in the region where income is zero, then our estimated relationship may not be a good approximation to reality in that region. So, although our estimated model suggests that a household with zero income is expected to spend \$83.42 per week on food, it might be risky to take this estimate literally. This is an issue that you should consider in each economic model that you estimate.

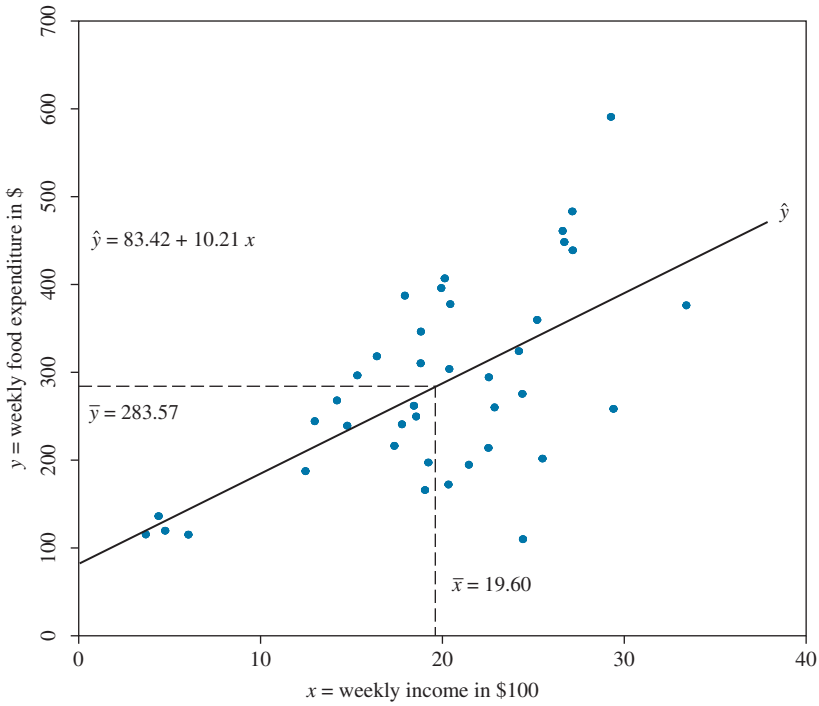


FIGURE 2.8 **The fitted regression.**

Elasticities Income elasticity is a useful way to characterize the responsiveness of consumer expenditure to changes in income. See Appendix A.2.2 for a discussion of elasticity calculations in a linear relationship. The **elasticity** of a variable y with respect to another variable x is

$$\varepsilon = \frac{\text{percentage change in } y}{\text{percentage change in } x} = \frac{100(\Delta y/y)}{100(\Delta x/x)} = \frac{\Delta y}{\Delta x}\cdot\frac{x}{y}$$

In the linear economic model given by (2.1), we have shown that

$$\beta_2 = \frac{\Delta E(y|\mathbf{x})}{\Delta x}$$

so the elasticity of mean expenditure with respect to income is

$$\varepsilon = \frac{\Delta E(y|\mathbf{x})}{\Delta x}\cdot\frac{x}{E(y|\mathbf{x})} = \beta_2\cdot\frac{x}{E(y|\mathbf{x})} \tag{2.9}$$

EXAMPLE 2.4b | Using the Estimates

To estimate this elasticity we replace β_2 by $b_2 = 10.21$. We must also replace "x" and "$E(y|\mathbf{x})$" by something, since in a linear model the elasticity is different on each point on the regression line. Most commonly, the elasticity is calculated at the "point of the means" $(\bar{x}, \bar{y}) = (19.60, 283.57)$ because it is a representative point on the regression line. If we calculate the income elasticity at the point of the means, we obtain

$$\hat{\varepsilon} = b_2\frac{\bar{x}}{\bar{y}} = 10.21 \times \frac{19.60}{283.57} = 0.71$$

This *estimated* income elasticity takes its usual interpretation. We estimate that a 1% increase in weekly household income will lead to a 0.71% increase in expected weekly household expenditure on food, when x and y take their sample mean values, $(\bar{x}, \bar{y}) = (19.60, 283.57)$. Since the estimated income elasticity is less than one, we would classify food as a "necessity" rather than a "luxury," which is consistent with what we would expect for an average household.

Prediction

The estimated equation can also be used for prediction or forecasting purposes. Suppose that we wanted to predict average weekly food expenditure for a household with a weekly income of $2000. This prediction is carried out by substituting $x = 20$ into our estimated equation to obtain

$$\hat{y}_i = 83.42 + 10.21x_i = 83.42 + 10.21(20) = 287.61$$

We *predict* that a household with a weekly income of $2000 will on average spend $287.61 per week on food.

Computer Output

Many different software packages can compute least squares estimates. Every software package's regression output looks different and uses different terminology to describe the output. Despite these differences, the various outputs provide the same basic information, which you should be able to locate and interpret. The matter is complicated somewhat by the fact that the packages also report various numbers whose meaning you may not know. For example, using the food expenditure data, the output from the software package EViews is shown in Figure 2.9.

In the EViews output, the parameter estimates are in the "Coefficient" column, with names "*C*," for constant term (the estimate b_1) and *INCOME* (the estimate b_2). Software programs typically name the estimates with the name of the variable as assigned in the computer program (we named our variable *INCOME*) and an abbreviation for "constant." The estimates that we report in the text are rounded to two significant digits. The other numbers that you can recognize at this time are $SSE = \sum \hat{e}_i^2 = 304505.2$, which is called "Sum squared resid," and the sample mean of y, $\bar{y} = \sum y_i/N = 283.5735$, which is called "Mean dependent var."

We leave discussion of the rest of the output until later.

Dependent Variable: *FOOD_EXP*
Method: Least Squares
Sample: 1 40
Included observations: 40

	Coefficient	Std. Error	*t*-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000
R-squared	0.385002	Mean dependent var		283.5735
Adjusted R-squared	0.368818	S.D. dependent var		112.6752
S.E. of regression	89.51700	Akaike info criterion		11.87544
Sum squared resid	304505.2	Schwarz criterion		11.95988
Log likelihood	−235.5088	Hannan-Quinn criter		11.90597
F-statistic	23.78884	Durbin-Watson stat		1.893880
Prob(F-statistic)	0.000019			

FIGURE 2.9 **EViews regression output.**

2.3.2 Other Economic Models

We have used the household expenditure on food versus income relationship as an example to introduce the ideas of simple regression. The simple regression model can be applied to estimate

the parameters of many relationships in economics, business, and the social sciences. The applications of regression analysis are fascinating and useful. For example,

- If the hourly wage rate of electricians rises by 5%, how much will new house prices increase?
- If the cigarette tax increases by $1, how much additional revenue will be generated in the state of Louisiana?
- If the central banking authority raises interest rates by one-half a percentage point, how much will consumer borrowing fall within six months? How much will it fall within one year? What will happen to the unemployment rate in the months following the increase?
- If we increase funding on preschool education programs in 2018, what will be the effect on high school graduation rates in 2033? What will be the effect on the crime rate by juveniles in 2028 and subsequent years?

The range of applications spans economics and finance, as well as most disciplines in the social and physical sciences. Any time you ask **how much** a change in one variable will affect another variable, regression analysis is a potential tool.

Similarly, any time you wish to **predict** the value of one variable given the value of another then least squares regression is a tool to consider.

2.4 Assessing the Least Squares Estimators

Using the food expenditure data, we have estimated the parameters of the regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ using the least squares formulas in (2.7) and (2.8). We obtained the least squares estimates $b_1 = 83.42$ and $b_2 = 10.21$. It is natural, but, as we shall argue, misguided, to ask the question "How good are these estimates?" This question is not answerable. We will never know the true values of the population parameters β_1 or β_2, so we cannot say how close $b_1 = 83.42$ and $b_2 = 10.21$ are to the true values. The least squares estimates are numbers that may or may not be close to the true parameter values, and we will never know.

Rather than asking about the quality of the estimates we will take a step back and examine the quality of the least squares estimation procedure. The motivation for this approach is this: if we were to collect another sample of data, by choosing another set of 40 households to survey, we would have obtained *different* estimates b_1 and b_2, even if we had carefully selected households with the same incomes as in the initial sample. This **sampling variation** is unavoidable. Different samples will yield different estimates because household food expenditures, y_i, $i = 1, \ldots, 40$, are random variables. Their values are not known until the sample is collected. Consequently, when viewed as an estimation procedure, b_1 and b_2 are also random variables, because their values depend on the random variable y. In this context, we call b_1 and b_2 the **least squares estimators**.

We can investigate the properties of the estimators b_1 and b_2, which are called their **sampling properties**, and deal with the following important questions:

1. If the least squares estimators b_1 and b_2 are random variables, then what are their expected values, variances, covariances, and probability distributions?
2. The least squares principle is only *one* way of using the data to obtain estimates of β_1 and β_2. How do the least squares estimators compare with other procedures that might be used, and how can we compare alternative estimators? For example, is there another estimator that has a higher probability of producing an estimate that is close to β_2?

We examine these questions in two steps to make things easier. In the first step, we investigate the properties of the least squares estimators conditional on the values of the explanatory variable in the sample. That is, conditional on **x**. Making the analysis conditional on **x** is equivalent to saying that, when we consider all possible samples, the household income values in the sample stay the

same from one sample to the next; only the random errors and food expenditure values change. This assumption is clearly not realistic but it simplifies the analysis. By conditioning on **x**, we are holding it constant, or fixed, meaning that we can treat the x-values as "not random."

In the second step, considered in Section 2.10, we return to the random sampling assumption and recognize that (y_i, x_i) data pairs are random, and randomly selecting households from a population leads to food expenditures and incomes that are random. However, even in this case and treating **x** as random, we will discover that most of our conclusions that treated **x** as nonrandom remain the same.

In either case, whether we make the analysis conditional on **x** or make the analysis general by treating **x** as random, the answers to the questions above depend critically on whether the assumptions SR1–SR5 are satisfied. In later chapters, we will discuss how to check whether the assumptions we make hold in a specific application, and what we might do if one or more assumptions are shown not to hold.

Remark

We will summarize the properties of the least squares estimators in the next several sections. "Proofs" of important results appear in the appendices to this chapter. In many ways, it is good to see these concepts in the context of a simpler problem before tackling them in the regression model. Appendix C covers the topics in this chapter, and the next, in the familiar and algebraically easier problem of estimating the mean of a population.

2.4.1 The Estimator b_2

Formulas (2.7) and (2.8) are used to compute the least squares estimates b_1 and b_2. However, they are not well suited for examining theoretical properties of the estimators. In this section, we rewrite the formula for b_2 to facilitate its analysis. In (2.7), b_2 is given by

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

This is called the **deviation from the mean form** of the estimator because the data have their sample means subtracted. Using assumption SR1 and a bit of algebra (Appendix 2C), we can write b_2 as a **linear estimator**,

$$b_2 = \sum_{i=1}^{N} w_i y_i \tag{2.10}$$

where

$$w_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2} \tag{2.11}$$

The term w_i depends only on **x**. Because we are conditioning our analysis on **x**, the term w_i is treated as if it is a constant. We remind you that conditioning on **x** is equivalent to treating **x** as given, as in a controlled, repeatable experiment.

Any estimator that is a weighted average of y_i's, as in (2.10), is called a **linear estimator**. This is an important classification that we will speak more of later. Then, with yet more algebra (Appendix 2D), we can express b_2 in a theoretically convenient way,

$$b_2 = \beta_2 + \sum w_i e_i \tag{2.12}$$

where e_i is the random error in the linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i$. This formula is not useful for computations, because it depends on β_2, which we do not know, and on the e_i's,

which are unobservable. However, for understanding the sampling properties of the least squares estimator, (2.12) is very useful.

2.4.2 The Expected Values of b_1 and b_2

The OLS estimator b_2 is a random variable since its value is unknown until a sample is collected. What we will show is that if our model assumptions hold, then $E(b_2|\mathbf{x}) = \beta_2$; that is, given $\mathbf{x}$ the expected value of b_2 is equal to the true parameter β_2. When the expected value of *any* estimator of a parameter equals the true parameter value, then that estimator is **unbiased**. Since $E(b_2|\mathbf{x}) = \beta_2$, the least squares estimator b_2 given $\mathbf{x}$ is an unbiased estimator of β_2. In Section 2.10, we will show that the least squares estimator b_2 is **unconditionally unbiased** also, $E(b_2) = \beta_2$. The intuitive meaning of unbiasedness comes from the sampling interpretation of mathematical expectation. Recognize that one sample of size N is just one of many samples that we could have been selected. If the formula for b_2 is used to estimate β_2 in each of those possible samples, then, if our assumptions are valid, the average value of the estimates b_2 obtained from all possible samples will be β_2.

We will show that this result is true so that we can illustrate the part played by the assumptions of the linear regression model. In (2.12), what parts are random? The parameter β_2 is not random. It is a population parameter we are trying to estimate. Conditional on $\mathbf{x}$ we can treat x_i as if it is not random. Then, conditional on $\mathbf{x}$, w_i is not random either, as it depends only on the values of x_i. The only random factors in (2.12) are the random error terms e_i. We can find the conditional expected value of b_2 using the fact that the expected value of a sum is the sum of the expected values:

$$\begin{aligned} E(b_2|\mathbf{x}) &= E(\beta_2 + \sum w_i e_i|\mathbf{x}) = E(\beta_2 + w_1 e_1 + w_2 e_2 + \cdots + w_N e_N|\mathbf{x}) \\ &= E(\beta_2) + E(w_1 e_1|\mathbf{x}) + E(w_2 e_2|\mathbf{x}) + \cdots + E(w_N e_N|\mathbf{x}) \\ &= \beta_2 + \sum E(w_i e_i|\mathbf{x}) \\ &= \beta_2 + \sum w_i E(e_i|\mathbf{x}) = \beta_2 \end{aligned} \tag{2.13}$$

The rules of expected values are fully discussed in the Probability Primer, Section P.5, and Appendix B.1.1. In the last line of (2.13), we use two assumptions. First, $E(w_i e_i|\mathbf{x}) = w_i E(e_i|\mathbf{x})$ because conditional on $\mathbf{x}$ the terms w_i are not random, and constants can be factored out of expected values. Second, we have relied on the assumption that $E(e_i|\mathbf{x}) = 0$. Actually, if $E(e_i|\mathbf{x}) = c$, where c is any constant value, such as 3, then $E(b_2|\mathbf{x}) = \beta_2$. Given $\mathbf{x}$, the OLS estimator b_2 is an **unbiased estimator** of the regression parameter β_2. On the other hand, if $E(e_i|\mathbf{x}) \neq 0$ and it depends on $\mathbf{x}$ in some way, then b_2 is a **biased estimator** of β_2. One leading case in which the assumption $E(e_i|\mathbf{x}) = 0$ fails is due to **omitted variables**. Recall that e_i contains everything else affecting y_i other than x_i. If we have omitted anything that is important and that is correlated with $\mathbf{x}$ then we would expect that $E(e_i|\mathbf{x}) \neq 0$ and $E(b_2|\mathbf{x}) \neq \beta_2$. In Chapter 6 we discuss this **omitted variables bias**. Here we have shown that conditional on $\mathbf{x}$, and under SR1–SR5, the least squares estimator is linear and unbiased. In Section 2.10, we show that $E(b_2) = \beta_2$ without conditioning on $\mathbf{x}$.

The unbiasedness of the estimator b_2 is an important sampling property. On average, over all possible samples from the population, the least squares estimator is "correct," on average, and this is one desirable property of an estimator. This statistical property by itself does not mean that b_2 is a good estimator of β_2, but it is part of the story. The unbiasedness property is related to what happens in all possible samples of data from the same population. The fact that b_2 is unbiased does not imply *anything* about what might happen *in just one sample*. An individual estimate (a number) b_2 may be near to, or far from, β_2. Since β_2 is *never* known we will never know, given

one sample, whether our estimate is "close" to β_2 or not. Thus, the estimate $b_2 = 10.21$ may be close to β_2 or not.

The least squares estimator b_1 of β_1 is also an unbiased estimator, and $E(b_1|\mathbf{x}) = \beta_1$ if the model assumptions hold.

2.4.3 Sampling Variation

To illustrate how the concept of unbiased estimation relates to sampling variation, we present in Table 2.2 least squares estimates of the food expenditure model from 10 hypothetical random samples (data file *table2_2*) of size $N = 40$ from the same population with the same incomes as the households given in Table 2.1. Note the variability of the least squares parameter estimates from sample to sample. This **sampling variation** is due to the fact that we obtain 40 *different* households in each sample, and their weekly food expenditure varies randomly.

The property of unbiasedness is about the *average* values of b_1 and b_2 if used in all possible samples of the same size drawn from the same population. The average value of b_1 in these 10 samples is $\bar{b}_1 = 96.11$. The average value of b_2 is $\bar{b}_2 = 8.70$. If we took the averages of estimates from more samples, these averages would approach the true parameter values β_1 and β_2. Unbiasedness does not say that an estimate from any one sample is close to the true parameter value, and thus we cannot say that an *estimate* is unbiased. We can say that the least squares estimation procedure (or the least squares estimator) is unbiased.

2.4.4 The Variances and Covariance of b_1 and b_2

Table 2.2 shows that the least squares estimates of β_1 and β_2 vary from sample to sample. Understanding this variability is a key to assessing the reliability and sampling precision of an estimator. We now obtain the variances and covariance of the estimators b_1 and b_2. Before presenting the expressions for the variances and covariance, let us consider why they are important to know. The variance of the random variable b_2 is the average of the squared distances between the possible values of the random variable and its mean, which we now know is $E(b_2|\mathbf{x}) = \beta_2$. The conditional variance of b_2 is defined as

$$\text{var}(b_2|\mathbf{x}) = E\left\{\left[b_2 - E(b_2|\mathbf{x})\right]^2 \middle| \mathbf{x}\right\}$$

TABLE 2.2 Estimates from 10 Hypothetical Samples

Sample	b_1	b_2
1	93.64	8.24
2	91.62	8.90
3	126.76	6.59
4	55.98	11.23
5	87.26	9.14
6	122.55	6.80
7	91.95	9.84
8	72.48	10.50
9	90.34	8.75
10	128.55	6.99

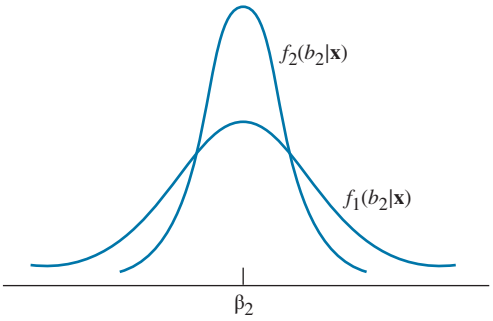


FIGURE 2.10 **Two possible probability density functions for b_2.**

It measures the spread of the probability distribution of b_2. In Figure 2.10 are graphs of two possible probability distributions of b_2, $f_1(b_2|\mathbf{x})$ and $f_2(b_2|\mathbf{x})$, that have the same mean value but different variances.

The *pdf* $f_2(b_2|\mathbf{x})$ has a smaller variance than $f_1(b_2|\mathbf{x})$. Given a choice, we are interested in estimator precision and would prefer that b_2 have the *pdf* $f_2(b_2|\mathbf{x})$, rather than $f_1(b_2|\mathbf{x})$. With the distribution $f_2(b_2|\mathbf{x})$, the probability is more concentrated around the true parameter value β_2 giving, relative to $f_1(b_2|\mathbf{x})$, a higher probability of getting an estimate that is close to β_2. Remember, getting an estimate close to β_2 is a primary objective of regression analysis.

The variance of an estimator measures the *precision* of the estimator in the sense that it tells us how much the estimates can vary from sample to sample. Consequently, we often refer to the **sampling variance** or **sampling precision** of an estimator. The smaller the variance of an estimator is, the greater the sampling precision of that estimator. One estimator is more precise than another estimator if its sampling variance is less than that of the other estimator.

We will now present and discuss the conditional variances and covariance of b_1 and b_2. Appendix 2E contains the derivation of the variance of the least squares estimator b_2. If the regression model assumptions SR1–SR5 are correct (assumption SR6 is not required), then the variances and covariance of b_1 and b_2 are

$$\text{var}(b_1|\mathbf{x}) = \sigma^2 \left[\frac{\sum x_i^2}{N\sum (x_i - \bar{x})^2} \right] \tag{2.14}$$

$$\text{var}(b_2|\mathbf{x}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{2.15}$$

$$\text{cov}(b_1, b_2|\mathbf{x}) = \sigma^2 \left[\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right] \tag{2.16}$$

At the beginning of this section we said that for unbiased estimators, smaller variances are better than larger variances. Let us consider the factors that affect the variances and covariance in (2.14)–(2.16).

1. The variance of the random error term, σ^2, appears in each of the expressions. It reflects the dispersion of the values y about their expected value $E(y|\mathbf{x})$. The greater the variance σ^2, the greater is that dispersion, and the greater is the uncertainty about where the values of y fall relative to their conditional mean $E(y|\mathbf{x})$. When σ^2 is larger, the information we have about β_1 and β_2 is less precise. In Figure 2.5, the variance is reflected in the spread of the probability distributions $f(y|x)$. The *larger* the variance term σ^2, the *greater is* the uncertainty in the statistical model, and the *larger* the variances and covariance of the least squares estimators.

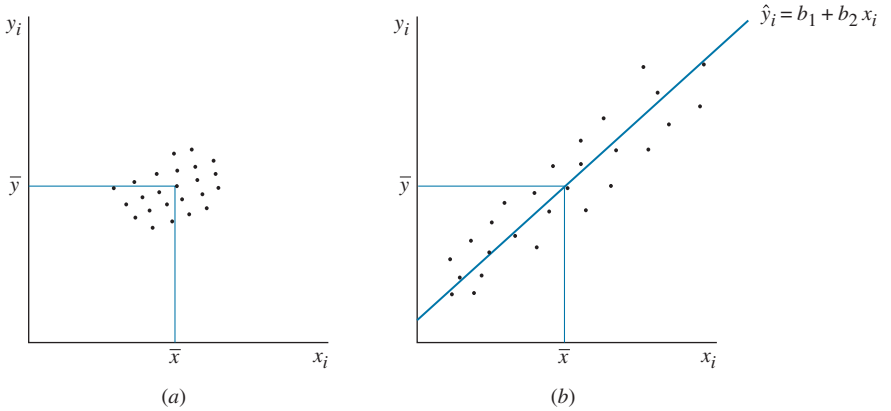


FIGURE 2.11 **The influence of variation in the explanatory variable x on precision of estimation: (*a*) low x variation, low precision: (*b*) high x variation, high precision.**

2. The sum of squares of the values of x about their sample mean, $\sum(x_i - \bar{x})^2$, appears in each of the variances and in the covariance. This expression measures how *spread out* about their mean are the sample values of the independent or explanatory variable x. The more they are spread out, the larger the sum of squares. The less they are spread out, the smaller the sum of squares. You may recognize this sum of squares as the numerator of the sample variance of the x-values. See Appendix C.4. The *larger* the sum of squares, $\sum(x_i - \bar{x})^2$, the *smaller* the conditional variances of the least squares estimators and the more *precisely* we can estimate the unknown parameters. The intuition behind this is demonstrated in Figure 2.11. Panel (b) is a data scatter in which the values of x are widely spread out along the x-axis. In panel (a), the data are "bunched." Which data scatter would you prefer given the task of fitting a line by hand? Pretty clearly, the data in panel (b) do a better job of determining where the least squares line must fall, because they are more spread out along the x-axis.
3. The larger the sample size N, the *smaller* the variances and covariance of the least squares estimators; it is better to have *more* sample data than *less*. The sample size N appears in each of the variances and covariance because each of the sums consists of N terms. Also, N appears explicitly in $\text{var}(b_1|\mathbf{x})$. The sum of squares term $\sum(x_i - \bar{x})^2$ gets larger as N increases because each of the terms in the sum is positive or zero (being zero if x happens to equal its sample mean value for an observation). Consequently, as N gets larger, both $\text{var}(b_2|\mathbf{x})$ and $\text{cov}(b_1, b_2|\mathbf{x})$ get smaller, since the sum of squares appears in their denominator. The sums in the numerator and denominator of $\text{var}(b_1|\mathbf{x})$ both get larger as N gets larger and offset one another, leaving the N in the denominator as the dominant term, ensuring that $\text{var}(b_1|\mathbf{x})$ also gets smaller as N gets larger.
4. The term $\sum x_i^2$ appears in $\text{var}(b_1|\mathbf{x})$. The larger this term is, the larger the variance of the least squares estimator b_1. Why is this so? Recall that the intercept parameter β_1 is the expected value of y given that $x = 0$. The farther our data are from $x = 0$, the more difficult it is to interpret β_1, as in the food expenditure example, and the more difficult it is to accurately estimate β_1. The term $\sum x_i^2$ measures the squared distance of the data from the origin, $x = 0$. If the values of x are near zero, then $\sum x_i^2$ will be small, and this will reduce $\text{var}(b_1|\mathbf{x})$. But if the values of x are large in magnitude, either positive or negative, the term $\sum x_i^2$ will be large and $\text{var}(b_1)$ will be larger, other things being equal.
5. The sample mean of the x-values appears in $\text{cov}(b_1, b_2|\mathbf{x})$. The absolute magnitude of the covariance *increases* with an increase in magnitude of the sample mean $\bar{x}$, and the covariance has a *sign* opposite to that of $\bar{x}$. The reasoning here can be seen from Figure 2.11. In panel (b)

the least squares fitted line must pass through the point of the means. Given a fitted line through the data, imagine the effect of increasing the estimated slope b_2. Since the line must pass through the point of the means, the effect must be to lower the point where the line hits the vertical axis, implying a reduced intercept estimate b_1. Thus, when the sample mean is positive, as shown in Figure 2.11, there is a negative covariance between the least squares estimators of the slope and intercept.

2.5 The Gauss–Markov Theorem

What can we say about the least squares estimators b_1 and b_2 so far?

- The estimators are perfectly general. Formulas (2.7) and (2.8) can be used to estimate the unknown parameters β_1 and β_2 in the simple linear regression model, no matter what the data turn out to be. Consequently, viewed in this way, the least squares estimators b_1 and b_2 are random variables.
- The least squares estimators are *linear* estimators, as defined in (2.10). Both b_1 and b_2 can be written as weighted averages of the y_i values.
- If assumptions SR1–SR5 hold, then the least squares estimators are conditionally *unbiased*. This means that $E(b_1|\mathbf{x}) = \beta_1$ and $E(b_2|\mathbf{x}) = \beta_2$.
- Given $\mathbf{x}$ we have expressions for the variances of b_1 and b_2 and their covariance. Furthermore, we have argued that for any unbiased estimator, having a smaller variance is better, as this implies we have a higher chance of obtaining an estimate close to the true parameter value.

Now we will state and discuss the famous **Gauss–Markov theorem**, which is proven in Appendix 2F.

Gauss–Markov Theorem:

Given $\mathbf{x}$ and under the assumptions SR1–SR5 of the linear regression model, the estimators b_1 and b_2 have the smallest variance of all linear and unbiased estimators of β_1 and β_2. They are the **best linear unbiased estimators (BLUE)** of β_1 and β_2.

Let us clarify what the Gauss–Markov theorem does, and does not, say.

1. The estimators b_1 and b_2 are "best" when compared to similar estimators, those that are linear and unbiased. The theorem does *not* say that b_1 and b_2 are the best of all *possible* estimators.
2. The estimators b_1 and b_2 are best within their class because they have the minimum variance. When comparing two linear and unbiased estimators, we *always* want to use the one with the smaller variance, since that estimation rule gives us the higher probability of obtaining an estimate that is close to the true parameter value.
3. In order for the Gauss–Markov theorem to hold, assumptions SR1–SR5 must be true. If any of these assumptions are *not* true, then b_1 and b_2 are *not* the best linear unbiased estimators of β_1 and β_2.
4. The Gauss–Markov theorem does *not* depend on the assumption of normality (assumption SR6).
5. In the simple linear regression model, if we want to use a linear and unbiased estimator, then we have to do no more searching. The estimators b_1 and b_2 are the ones to use. This explains

why we are studying these estimators (we would not have you study *bad* estimation rules, would we?) and why they are so widely used in research, not only in economics but in all social and physical sciences as well.

6. The Gauss–Markov theorem applies to the least squares estimators. It *does not* apply to the least squares *estimates* from a single sample.

The results we have presented so far treat **x** as given. In Section 2.10 we show that the Gauss–Markov theorem also holds in general, and it does not depend on a specific **x**.

2.6 The Probability Distributions of the Least Squares Estimators

The properties of the least squares estimators that we have developed so far do not depend in any way on the normality assumption SR6. If we also make this assumption, that the random errors e_i are normally distributed, with mean zero and variance σ^2, then the conditional probability distributions of the least squares estimators are also normal. This conclusion is obtained in two steps. First, given **x** and based on assumption SR1, if e_i is normal then so is y_i. Second, the least squares estimators are linear estimators of the form $b_2 = \sum w_i y_i$. Given **x** this weighted sum of normal random variables is also normally distributed. Consequently, *if* we make the normality assumption (assumption SR6 about the error term), and treat **x** as given, then the least squares estimators are normally distributed:

$$b_1|\mathbf{x} \sim N\left(\beta_1, \frac{\sigma^2 \sum x_i^2}{N\sum(x_i - \bar{x})^2}\right) \tag{2.17}$$

$$b_2|\mathbf{x} \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right) \tag{2.18}$$

As you will see in Chapter 3, the normality of the least squares estimators is of great importance in many aspects of statistical inference.

What if the errors are not normally distributed? Can we say anything about the probability distribution of the least squares estimators? The answer is, sometimes, yes.

A Central Limit Theorem:

If assumptions SR1–SR5 hold, and if the sample size N is **sufficiently large**, then the least squares estimators have a distribution that approximates the normal distributions shown in (2.17) and (2.18).

The million-dollar question is "How large is sufficiently large?" The answer is that there is no specific number. The reason for this vague and unsatisfying answer is that "how large" depends on many factors, such as what the distributions of the random errors look like (are they smooth? symmetric? skewed?) and what the x_i values are like. In the simple regression model, some would say that $N = 30$ is sufficiently large. Others would say that $N = 50$ would be a more reasonable number. The bottom line is, however, that these are rules of thumb and that the meaning of "sufficiently large" will change from problem to problem. Nevertheless, for better or worse, this *large sample*, or **asymptotic**, result is frequently invoked in regression analysis. This important result is an application of a central limit theorem, like the one discussed in Appendix C.3.4. If you are not familiar with this important theorem, you may want to review it now.

2.7 Estimating the Variance of the Error Term

The variance of the random error term, σ^2, is the one unknown parameter of the simple linear regression model that remains to be estimated.

The conditional variance of the random error e_i is

$$\text{var}(e_i|\mathbf{x}) = \sigma^2 = E\left\{\left[e_i - E(e_i|\mathbf{x})\right]^2 \middle| \mathbf{x}\right\} = E(e_i^2|\mathbf{x})$$

if the assumption $E(e_i|\mathbf{x}) = 0$ is correct. Since the "expectation" is an average value, we might consider estimating σ^2 as the average of the squared errors

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{N}$$

This formula is unfortunately of no use since the random errors e_i are *unobservable*! However, although the random errors themselves are unknown, we do have an analog to them—namely, the least squares residuals. Recall that the random errors are

$$e_i = y_i - \beta_1 - \beta_2 x_i$$

From (2.6) the least squares residuals are obtained by replacing the unknown parameters by their least squares estimates:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$$

It seems reasonable to replace the random errors e_i by their analogs, the least squares residuals, so that

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N}$$

This estimator, though quite satisfactory in large samples, is a *biased* estimator of σ^2. But there is a simple modification that produces an unbiased estimator:

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} \tag{2.19}$$

The 2 that is subtracted in the denominator is the number of *regression parameters* (β_1, β_2) in the model, and this subtraction makes the estimator $\hat{\sigma}^2$ unbiased, so that $E\left(\hat{\sigma}^2|\mathbf{x}\right) = \sigma^2$.

2.7.1 Estimating the Variances and Covariance of the Least Squares Estimators

Having an unbiased estimator of the error variance means we can *estimate* the conditional variances of the least squares estimators b_1 and b_2 and the covariance between them. Replace the unknown error variance σ^2 in (2.14)–(2.16) by $\hat{\sigma}^2$ to obtain

$$\widehat{\text{var}}(b_1|\mathbf{x}) = \hat{\sigma}^2 \left[\frac{\sum x_i^2}{N\sum(x_i - \bar{x})^2}\right] \tag{2.20}$$

$$\widehat{\text{var}}(b_2|\mathbf{x}) = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2} \tag{2.21}$$

$$\widehat{\text{cov}}(b_1, b_2|\mathbf{x}) = \hat{\sigma}^2 \left[\frac{-\bar{x}}{\sum(x_i - \bar{x})^2}\right] \tag{2.22}$$

The square roots of the estimated variances are the "standard errors" of b_1 and b_2. These quantities are used in hypothesis testing and confidence intervals. They are denoted as $\text{se}(b_1)$ and $\text{se}(b_2)$

$$\text{se}(b_1) = \sqrt{\widehat{\text{var}}(b_1|\mathbf{x})} \tag{2.23}$$

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}}(b_2|\mathbf{x})} \tag{2.24}$$

EXAMPLE 2.5 | Calculations for the Food Expenditure Data

Let us make some calculations using the food expenditure data. The least squares estimates of the parameters in the food expenditure model are shown in Figure 2.9. First, we will compute the least squares residuals from (2.6) and use them to calculate the estimate of the error variance in (2.19). In Table 2.3 are the least squares residuals for the first five households in Table 2.1.

TABLE 2.3 Least Squares Residuals

x	y	$\hat{y}$	$\hat{e} = y - \hat{y}$
3.69	115.22	121.09	−5.87
4.39	135.98	128.24	7.74
4.75	119.34	131.91	−12.57
6.03	114.96	144.98	−30.02
12.47	187.05	210.73	−23.68

Recall that we have estimated that for the food expenditure data the fitted least squares regression line is $\hat{y} = 83.42 + 10.21x$. For each observation, we compute the least squares residual $\hat{e}_i = y_i - \hat{y}_i$. Using the residuals for all $N = 40$ observations, we estimate the error variance to be

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = \frac{304505.2}{38} = 8013.29$$

The numerator, 304505.2, is the sum of squared least squares residuals, reported as "Sum squared resid" in Figure 2.9. The denominator is the number of sample observations, $N = 40$, minus the number of estimated regression parameters, 2; the quantity $N - 2 = 38$ is often called the **degrees of freedom** for reasons that will be explained in Chapter 3. In Figure 2.9, the value $\hat{\sigma}^2$ is not reported. Instead, EViews software reports $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{8013.29} = 89.517$, labeled "S.E. of regression," which stands for "standard error of the regression."

It is typical for software not to report the estimated variances and covariance unless requested. However, all software packages automatically report the standard errors. For example, in the EViews output shown in Figure 2.9 the column labeled "Std. Error" contains $\text{se}(b_1) = 43.410$ and $\text{se}(b_2) = 2.093$. The entry called "S.D. dependent var" is the sample standard deviation of y, that is, $\left[\sum(y_i - \bar{y})^2/(N-1)\right]^{1/2} = 112.6752$.

The full set of estimated variances and covariances for a regression is usually obtained by a simple computer command, or option, depending on the software being used. They are arrayed in a rectangular array, or matrix, with variances on the diagonal and covariances in the "off-diagonal" positions.

$$\begin{bmatrix} \widehat{\text{var}}(b_1|\mathbf{x}) & \widehat{\text{cov}}(b_1, b_2|\mathbf{x}) \\ \widehat{\text{cov}}(b_1, b_2|\mathbf{x}) & \widehat{\text{var}}(b_2|\mathbf{x}) \end{bmatrix}$$

For the food expenditure data, the estimated covariance matrix of the least squares estimators is

	C	*INCOME*
C	1884.442	−85.90316
INCOME	−85.90316	4.381752

where C stands for the "constant term," which is the estimated intercept parameter in the regression, or b_1; similarly, the software reports the variable name *INCOME* for the column relating to the estimated slope b_2. Thus

$$\widehat{\text{var}}(b_1|\mathbf{x}) = 1884.442, \quad \widehat{\text{var}}(b_2|\mathbf{x}) = 4.381752,$$
$$\widehat{\text{cov}}(b_1, b_2|\mathbf{x}) = -85.90316$$

The standard errors are

$$\text{se}(b_1) = \sqrt{\widehat{\text{var}}(b_1|\mathbf{x})} = \sqrt{1884.442} = 43.410$$

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}}(b_2|\mathbf{x})} = \sqrt{4.381752} = 2.093$$

These values will be used extensively in Chapter 3.

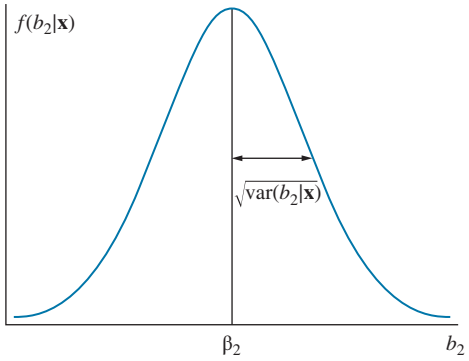


FIGURE 2.12 **The conditional probability density function of the least squares estimator b_2.**

2.7.2 Interpreting the Standard Errors

The standard errors of b_1 and b_2 are measures of the **sampling variability** of the least squares estimates b_1 and b_2 in **repeated samples**. As illustrated in Table 2.2, when we collect different samples of data, the parameter estimates change from sample to sample. The estimators b_1 and b_2 are general formulas that are used whatever the sample data turn out to be. That is, the estimators are random variables. As such, they have probability distributions, means, and variances. In particular, if assumption SR6 holds, and the random error terms e_i are normally distributed, then $b_2|\mathbf{x} \sim N\left(\beta_2, \text{var}(b_2|\mathbf{x}) = \sigma^2/\sum(x_i - \bar{x})^2\right)$. This *pdf* $f(b_2|\mathbf{x})$ is shown in Figure 2.12.

The estimator variance, $\text{var}(b_2|\mathbf{x})$, or, its square root $\sigma_{b_2} = \sqrt{\text{var}(b_2|\mathbf{x})}$, which we might call the true standard deviation of b_2, measures the sampling variation of the estimates b_2 and determines the width of the *pdf* in Figure 2.12. The bigger σ_{b_2} is the more variation in the least squares estimates b_2 we see from sample to sample. If σ_{b_2} is large, then the estimates might change a great deal from sample to sample. The parameter σ_{b_2} would be a valuable number to know, because if it were large relative to the parameter β_2 we would know that the least squares estimator is not precise, and the estimate that we obtain may be far from the true value β_2 that we are trying to estimate. On the other hand, if σ_{b_2} is small relative to the parameter β_2, we know that the least squares estimate will fall near β_2 with high probability. Recall that for the normal distribution, 99.9% of values fall within the range of three standard deviations from the mean, so that 99.9% of the least squares estimates will fall in the range $\beta_2 - 3\sigma_{b_2}$ to $\beta_2 + 3\sigma_{b_2}$.

To put this in another context, in Table 2.2 we report estimates from 10 samples of data. We noted in Section 2.4.3 that the average values of those estimates are $\bar{b}_1 = 96.11$ and $\bar{b}_2 = 8.70$. The question we address with the standard error is "How much variation about their means do the estimates exhibit from sample to sample?" For those 10 samples, the sample standard deviations are std. dev.$(b_1) = 23.61$ and std. dev.$(b_2) = 1.58$. What we would **really like** is the values of the standard deviations for a **very large** number of samples. Then we would know how much variation the least squares estimates exhibit from sample to sample. Unfortunately, we do not have a large number of samples, and because we do not know the true value of the variance of the error term σ^2 we cannot know the true value of σ_{b_2}.

Then what do we do? We estimate σ^2, and then estimate σ_{b_2} using

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}}(b_2|\mathbf{x})} = \sqrt{\frac{\hat{\sigma}^2}{\Sigma(x_i - \bar{x})^2}}$$

The standard error of b_2 is thus an estimate of what the standard deviation of many estimates b_2 would be in a very large number of samples and is an indicator of the width of the *pdf* of b_2 shown in Figure 2.12. Using our one sample of data, *food*, the standard error of b_2 is 2.093, as shown in

the computer output in Figure 2.9. This value is reasonably close to std. dev. $(b_2) = 1.58$ from the 10 samples in Table 2.2. To put this to a further test, in Appendix 2H, we perform a simulation experiment, called a **Monte Carlo experiment**, in which we create many artificial samples to demonstrate the properties of the least squares estimator and how well $\text{se}(b_2)$ reflects the true sampling variation in the estimates.

2.8 Estimating Nonlinear Relationships

The world is not linear. Economic variables are not always related by straight-line relationships; in fact, many economic relationships are represented by curved lines and are said to display **curvilinear** forms. Fortunately, the simple linear regression model $y = \beta_1 + \beta_2 x + e$ is much more flexible than it looks at first glance, because the variables y and x can be transformations, involving logarithms, squares, cubes, or reciprocals, of the basic economic variables, or they can be **indicator variables** that take only the values zero and one. Including these possibilities means the simple linear regression model can be used to account for **nonlinear relationships** between variables.[4]

Nonlinear relationships can sometimes be anticipated. Consider a model from real estate economics in which the price (*PRICE*) of a house is related to the house size measured in square feet (*SQFT*). As a starting point, we might consider the linear relationship

$$PRICE = \beta_1 + \beta_2 SQFT + e \tag{2.25}$$

In this model, β_2 measures the increase in expected price given an additional square foot of living area. In the linear specification, the expected price per additional square foot is constant. However, it may be reasonable to assume that larger and more expensive homes have a higher value for an additional square foot of living area than smaller, less expensive homes. How can we build this idea into our model? We will illustrate the use of two approaches: first, a **quadratic** equation in which the explanatory variable is $SQFT^2$; and second, a **log-linear** equation in which the dependent variable is ln(*PRICE*). In each case, we will find that the slope of the relationship between *PRICE* and *SQFT* is not constant, but changes from point to point.

2.8.1 Quadratic Functions

The quadratic function $y = a + bx^2$ is a parabola.[5] The y-intercept is a. The shape of the curve is determined by b; if $b > 0$, then the curve is U-shaped; and if $b < 0$, then the curve has an inverted-U shape. The slope of the function is given by the derivative[6] $dy/dx = 2bx$, which changes as x changes. The elasticity or the percentage change in y given a 1% change in x is $\varepsilon = slope \times x/y = 2bx^2/y$. If a and b are greater than zero, the curve resembles Figure 2.13.

2.8.2 Using a Quadratic Model

A **quadratic model** for house prices includes the **squared** value of *SQFT*, giving

$$PRICE = \alpha_1 + \alpha_2 SQFT^2 + e \tag{2.26}$$

This is a simple regression model, $y = \alpha_1 + \alpha_2 x + e$, with $y = PRICE$ and $x = SQFT^2$. Here, we switch from using β to denote the parameters to using α, because the parameters of (2.26) are not comparable to the parameters of (2.25). In (2.25) β_2 is a slope, but α_2 is not a slope. Because

[4]The term linear in "linear regression" means that the parameters are not transformed in any way. In a linear regression model, the parameters must not be raised to powers or transformed, so expressions like $\beta_1\beta_2$ or $\beta_2^{\beta_1}$ are not permitted.

[5]This is a special case of the more general quadratic function $y = a + bx + cx^2$.

[6]See Appendix A.3.1, Derivative Rules 1–5.

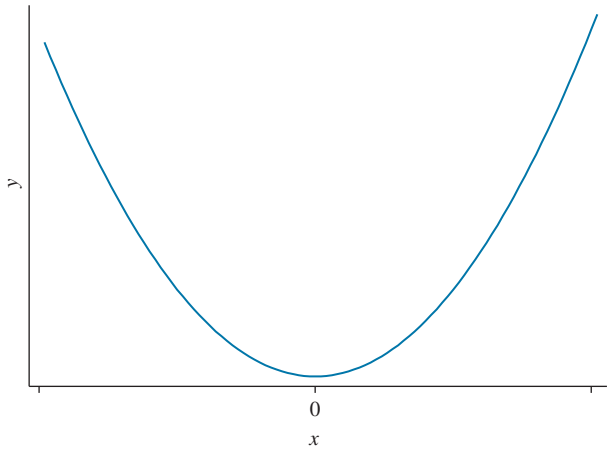


FIGURE 2.13 **A quadratic function, $y = a + bx^2$.**

$SQFT > 0$, the house price model will resemble the right side of the curve in Figure 2.13. Using ^ to denote estimated values, the least squares estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$, of α_1 and α_2, are calculated using the estimators in (2.7) and (2.8), just as earlier. The fitted equation is $\widehat{PRICE} = \hat{\alpha}_1 + \hat{\alpha}_2 SQFT^2$. It has slope

$$\frac{d\left(\widehat{PRICE}\right)}{dSQFT} = 2\hat{\alpha}_2 SQFT \tag{2.27}$$

If $\hat{\alpha}_2 > 0$, then larger houses will have larger slope, and a larger estimated price per additional square foot.

EXAMPLE 2.6 | Baton Rouge House Data

The data file *br* contains data on 1080 houses sold in Baton Rouge, Louisiana, during mid-2005. Using these data, the estimated quadratic equation is $\widehat{PRICE} = 55776.56 + 0.0154SQFT^2$. The data scatter and fitted quadratic relationship are shown in Figure 2.14.

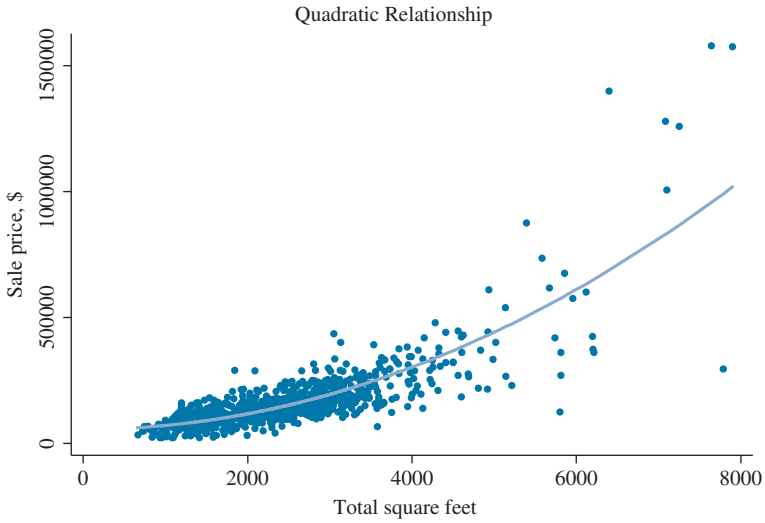


FIGURE 2.14 **A fitted quadratic relationship.**

The estimated slope is $\widehat{slope} = 2(0.0154)\,SQFT$ (estimated price per additional square foot), which for a 2000-square-foot house is \$61.69, for a 4000-square-foot house is \$123.37, and for a 6000-square-foot house is \$185.05. The elasticity of house price with respect to house size is the percentage increase in estimated price given a 1% increase in house size. Like the slope, the elasticity changes at each point. In our example

$$\hat{\varepsilon} = \widehat{slope} \times \frac{SQFT}{PRICE} = \left(2\hat{\alpha}_2 SQFT\right) \times \frac{SQFT}{PRICE}$$

To compute an estimate, we must select values for *SQFT* and *PRICE* on the fitted relationship. That is, we choose a value for *SQFT* and choose for price the corresponding fitted value $\widehat{PRICE}$. For houses of 2000, 4000, and 6000 square feet, the estimated elasticities are 1.05 [using $\widehat{PRICE}$ = \$117,461.77], 1.63 [using $\widehat{PRICE}$ = \$302,517.39], and 1.82 [using $\widehat{PRICE}$ = \$610,943.42], respectively. For a 2000-square-foot house, we estimate that a 1% increase in house size will increase price by 1.05%.

2.8.3 A Log-Linear Function

The log-linear equation $\ln(y) = a + bx$ has a logarithmic term on the left-hand side of the equation and an untransformed (linear) variable on the right-hand side. Both its slope and elasticity change at each point and are the same sign as b. Using the antilogarithm, we see that $\exp[\ln(y)] = y = \exp(a + bx)$, so that the log-linear function is an exponential function. The function requires $y > 0$. The slope[7] at any point is $dy/dx = \exp(a + bx) \times b = by$, which for $b > 0$ means that the marginal effect increases for larger values of y. An economist might say that this function is increasing at an increasing rate, as shown in Figure 2.15.

The elasticity, the percentage change in y given a 1% increase in x, at a point on this curve is $\varepsilon = slope \times x/y = bx$.

Using the slope expression, we can solve for a **semi-elasticity**, which tells us the percentage change in y given a one-unit increase in x. Divide both sides of the slope dy/dx by y, then multiply by 100 to obtain

$$\eta = \frac{100(dy/y)}{dx} = 100b \tag{2.28}$$

In this expression, the numerator $100(dy/y)$ is the percentage change in y; dx represents the change in x. If $dx = 1$, then a one-unit change in x leads to a $100b$ percentage change in y. This interpretation can sometimes be quite handy.

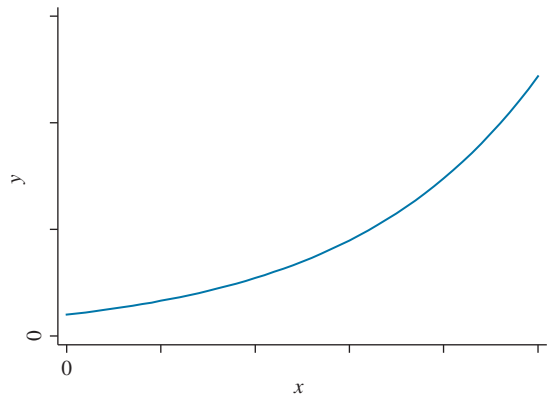


FIGURE 2.15 **A log-linear function.**

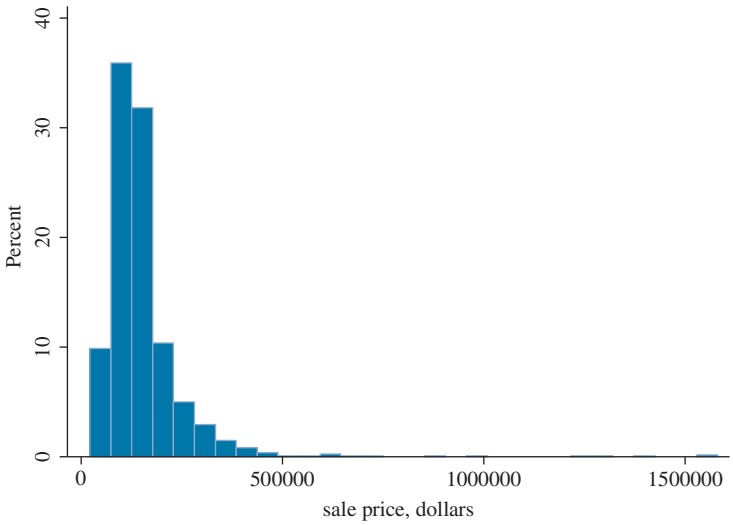
[7] See Appendix A.3.1, Derivative Rule 7.

2.8.4 Using a Log-Linear Model

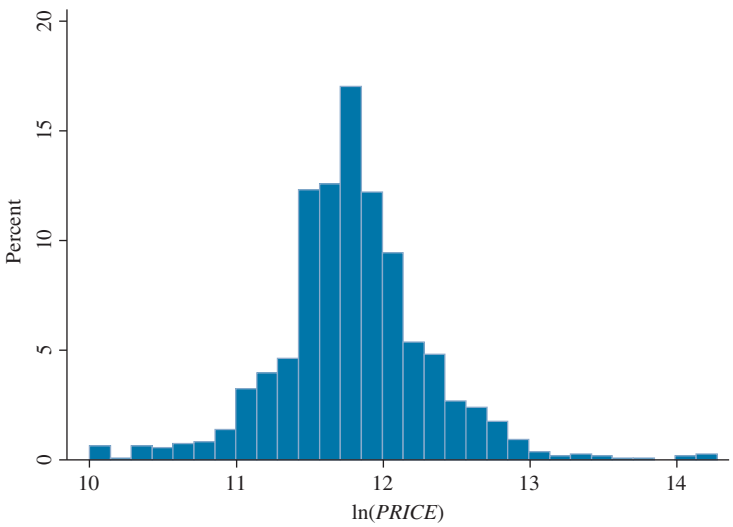
The use of logarithms is very common in economic modeling. The **log-linear model** uses the logarithm of a variable as the dependent variable, and an independent, explanatory variable, that is not transformed, such as[8]

$$\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e \tag{2.29}$$

What effects does this have? First, the logarithmic transformation can regularize data that is skewed with a long tail to the right. In Figure 2.16(a), we show the histogram of *PRICE* and in Figure 2.16(b) the histogram of ln(*PRICE*). The median house price in this sample is $130,000,



(*a*)



(*b*)

FIGURE 2.16 **(*a*) Histogram of *PRICE*. (*b*) Histogram of ln(*PRICE*).**

[8]Once again we use different symbols for the parameters of this model, γ_1 and γ_2, as a reminder that these parameters are not directly comparable to β's in (2.25) or α's in (2.26).

and 95% of house prices are below $315,000, but there are 24 houses out of the 1080 with prices above $500,000, and an extreme value of $1,580,000. The extremely skewed distribution of *PRICE* becomes more symmetric, if not bell-shaped, after taking the logarithm. Many economic variables, including prices, incomes, and wages, have skewed distributions, and the use of logarithms in models for such variables is common.

Second, using a log-linear model allows us to fit regression curves like that shown in Figure 2.15.

EXAMPLE 2.7 | Baton Rouge House Data, Log-Linear Model

Using the Baton Rouge data, the fitted log-linear model is

$$\widehat{\ln(PRICE)} = 10.8386 + 0.0004113\ SQFT$$

To obtain predicted price, take the antilogarithm,[9] which is the exponential function

$$\widehat{PRICE} = \exp\left[\widehat{\ln(PRICE)}\right] = \exp(10.8386 + 0.0004113\ SQFT)$$

The fitted value of *PRICE* is shown in Figure 2.17.

The slope of the log-linear model is

$$\frac{d\left(\widehat{PRICE}\right)}{dSQFT} = \hat{\gamma}_2 \widehat{PRICE} = 0.0004113\ \widehat{PRICE}$$

For a house with a predicted *PRICE* of $100,000, the estimated increase in *PRICE* for an additional square foot of house area is $41.13, and for a house with a predicted *PRICE* of $500,000, the estimated increase in *PRICE* for an additional square foot of house area is $205.63. The estimated elasticity is $\hat{\varepsilon} = \hat{\gamma}_2\ SQFT = 0.0004113\ SQFT$. For a house with 2000 square feet, the estimated elasticity is 0.823: a 1% increase in house size is estimated to increase selling price by 0.823%. For a house with 4000 square feet, the estimated elasticity is 1.645: a 1% increase in house size is estimated to increase selling price by 1.645%. Using the "semi-elasticity" defined in equation (2.28), we can say that, for a one-square-foot increase in size, we estimate a price increase of 0.04%. Or, perhaps more usefully, we estimate that a 100-square-foot increase will increase price by approximately 4%.

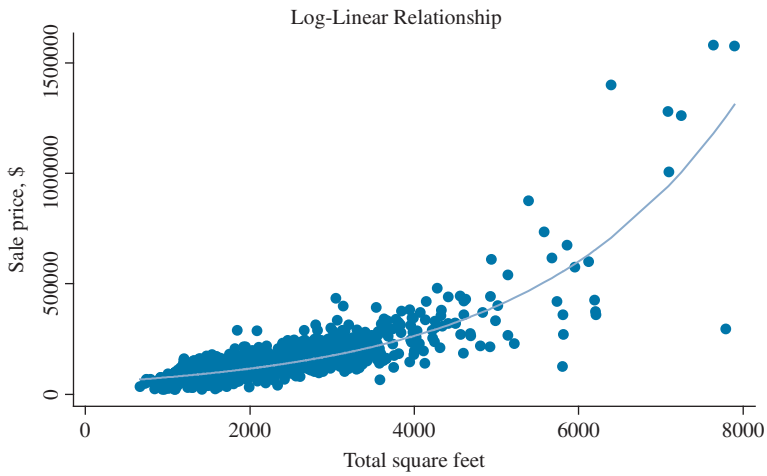


FIGURE 2.17 **The fitted log-linear model.**

[9] In Chapter 4 we present an improved predictor for this model.

2.8.5 Choosing a Functional Form

For the Baton Rouge house price data, should we use the quadratic functional form or the log-linear functional form? This is not an easy question. Economic theory tells us that house price should be related to the size of the house, and perhaps that larger, more expensive homes have a higher price per additional square foot of living area. But economic theory does not tell us what the exact algebraic form of the relationship should be. We should do our best to choose a functional form that is consistent with economic theory, that fits the data well, and that is such that the assumptions of the regression model are satisfied. In real-world problems, it is sometimes difficult to achieve all these goals. Furthermore, we will never truly know the correct functional relationship, no matter how many years we study econometrics. The truth is out there, but we will never know it. In applications of econometrics, we must simply do the best we can to choose a satisfactory functional form. At this point, we mention one dimension of the problem used for evaluating models with the same dependent variable. By comparing the sum of squared residuals (*SSE*) of alternative models, or, equivalently, $\hat{\sigma}^2$ or $\hat{\sigma}$, we can choose the model that is a better fit to the data. Smaller values of these quantities mean a smaller sum of squared residuals and a better model fit. This comparison is **not** valid for comparing models with dependent variables y and $\ln(y)$, or when other aspects of the models are different. We study the choice among functions like these further in Chapter 4.

2.9 Regression with Indicator Variables

An indicator variable is a binary variable that takes the values zero or one; it is used to represent a nonquantitative characteristic, such as gender, race, or location. For example, in the data file *utown.dot* we have a sample of 1,000 observations on house prices (*PRICE*, in thousands of dollars) in two neighborhoods. One neighborhood is near a major university and called University Town. Another similar neighborhood, called Golden Oaks, is a few miles away from the university. The indicator variable of interest is

$$UTOWN = \begin{cases} 1 & \text{house is in University Town} \\ 0 & \text{house is in Golden Oaks} \end{cases}$$

The histograms of the prices in these two neighborhoods, shown in Figure 2.18, are revealing. The mean of the distribution of house prices in University Town appears to be larger than the mean of the distribution of house prices from Golden Oaks. The sample mean of the 519 house prices in University Town is 277.2416 thousand dollars, whereas the sample mean of the 481 Golden Oaks houses is 215.7325 thousand dollars.

If we include *UTOWN* in a regression model as an explanatory variable, what do we have? The simple regression model is

$$PRICE = \beta_1 + \beta_2 UTOWN + e$$

If the regression assumptions SR1–SR5 hold, then the least squares estimators in (2.7) and (2.8) can be used to estimate the unknown parameters β_1 and β_2.

When an indicator variable is used in a regression, it is important to write out the regression function for the different values of the indicator variable.

$$E(PRICE|UTOWN) = \beta_1 + \beta_2 UTOWN = \begin{cases} \beta_1 + \beta_2 & \text{if } UTOWN = 1 \\ \beta_1 & \text{if } UTOWN = 0 \end{cases}$$

In this case, we find that the "regression function" reduces to a model that implies that the population mean house prices in the two subdivisions are different. The parameter β_2 is not a slope

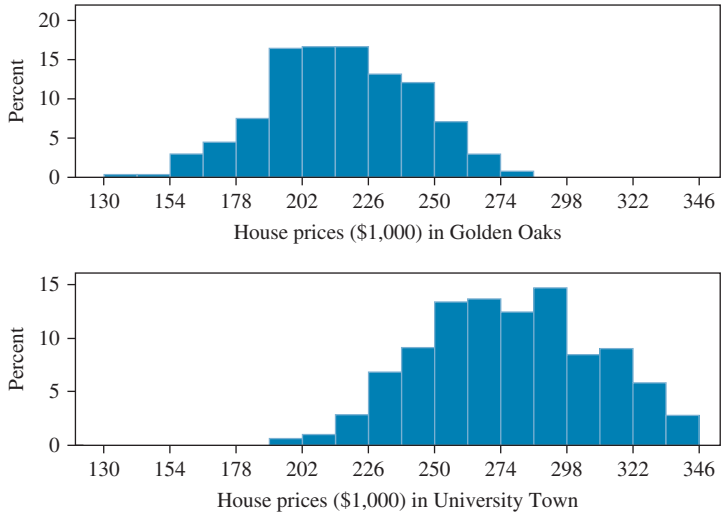


FIGURE 2.18 **Distributions of house prices.**

in this model. Here β_2 is the difference between the population means for house prices in the two neighborhoods. The expected price in University Town is $\beta_1 + \beta_2$, and the expected price in Golden Oaks is β_1. In our model, there are no factors other than location affecting price, and the indicator variable splits the observations into two populations.

The estimated regression is

$$\widehat{PRICE} = b_1 + b_2 UTOWN = 215.7325 + 61.5091\, UTOWN$$

$$= \begin{cases} 277.2416 & \text{if } UTOWN = 1 \\ 215.7325 & \text{if } UTOWN = 0 \end{cases}$$

We see that the estimated price for the houses in University Town is \$277,241.60, which is also the sample mean of the house prices in University Town. The estimated price for houses outside University Town is \$215,732.50, which is the sample mean of house prices in Golden Oaks.

In the regression model approach, we estimate the regression intercept β_1, which is the expected price for houses in Golden Oaks, where $UTOWN = 0$, and the parameter β_2, which is the difference between the population means for house prices in the two neighborhoods. The least squares estimators b_1 and b_2 in this indicator variable regression can be shown to be

$$b_1 = \overline{PRICE}_{\text{Golden Oaks}}$$

$$b_2 = \overline{PRICE}_{\text{University Town}} - \overline{PRICE}_{\text{Golden Oaks}}$$

where $\overline{PRICE}_{\text{Golden Oaks}}$ is the sample mean (average) price of houses in Golden Oaks and $\overline{PRICE}_{\text{University Town}}$ is the sample mean price of houses from University Town.

In the simple regression model, an indicator variable on the right-hand side gives us a way to estimate the differences between population means. This is a common problem in statistics, and the direct approach using samples means is discussed in Appendix C.7.2. Indicator variables are used in regression analysis very frequently in many creative ways. See Chapter 7 for a full discussion.

2.10 The Independent Variable[10]

Earlier in this chapter we specified a number of assumptions for the simple regression model and then used these assumptions to derive some properties of the least squares estimators of the coefficients in the model. In the household food expenditure example, we assumed a DGP where pairs (y_i, x_i) are randomly drawn from some population. We then went on to make a strict exogeneity assumption $E(e_i|\mathbf{x}) = 0$ to accommodate other types of DGPs. Using this and other assumptions, we derived properties of the least squares estimator conditional on the sample values $\mathbf{x}$. In this section, we say more about different possible DGPs, explore their implications for the assumptions of the simple regression model, and investigate how the properties of the least squares estimator change, if at all, when we no longer condition on $\mathbf{x}$.

Our regression model $y = \beta_1 + \beta_2 x + e$ has five components, three of which are unobservable: β_1, β_2, and e. The two observable components are y the random outcome, or dependent variable, and x the explanatory, independent variable. Is this explanatory variable random or not and why does it matter? We address these questions in this section.

How do we obtain values for the observable pair of variables (y, x)? In an experimental DGP, a scientist under carefully controlled conditions specifies the values of x, performs an experiment, and observes the outcomes y. For example, an agronomist might vary the number of pounds of pesticide spread per acre of cropland and observe the resulting yield. In this case, the independent variable, pounds of pesticide, is in fact an *independent* factor and not random. It is fixed. It is not affected by random influences and the treatment can be replicated time and time again. Laboratory and other controlled experiments can claim that the values of the independent variable are fixed. In the world of economics and business, there are few examples of laboratory and controlled experiments.[11] One exception is retail sales. Merchants display the prices of goods and services and observe consumer purchases. The merchant controls the prices, store displays, advertising and the shopping environment. In this case, we can argue that x, the price of a product in a retail store, is fixed and not random; it is *given*. When x is fixed and not random, the idea of repeated experimental trials makes intuitive sense. The sampling properties of the least squares estimators are a summary of how the estimators perform under a series of controlled experiments with fixed values for the independent variables. We have shown that the least squares estimator is the best linear unbiased estimator, given $\mathbf{x}$, and we have variance equations (2.14) and (2.15) that describe how much variation the estimates exhibit from sample to sample.

In the next three sections, we treat cases in which x-values are random. Each of these cases represents a different type of DGP. We start with the strongest assumption about random-x and then look at weaker cases.

2.10.1 Random and Independent x

Suppose our agronomist takes another strategy, using a random number between 0 and 100 to determine the amount of pesticide applied to a given acre of land. In this case, x is random, as its value is unknown until it is randomly selected. Why might a scientist use this approach? Well, no one could imply that such an experiment was rigged to produce a particular outcome. It is a "fair" experiment because the scientist keeps "hands off" the controls. What are the sampling properties of the least squares estimator in this setting? Is the least squares estimator the best, linear unbiased estimator in this case?

[10] This section contains a more advanced discussion of the assumptions of the simple regression model.

[11] Economists understand the benefits of controlled experiments. The field of experimental economics has grown tremendously in the past 20 years. Also, there have been some social experiments. One example is Tennessee's Project STAR that examined the consequences on school children of having small classes rather than larger ones. This example is explored further in Chapter 7.5.

In order to answer these questions, we make explicit that x is *statistically independent* of the error term e. The assumptions for the **independent random-x model** (IRX) are as follows:

Assumptions of the Independent Random-x Linear Regression Model

IRX1: The observable variables y and x are related by $y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, \ldots, N$, where β_1 and β_2 are unknown population parameters and e_i is a random error term.

IRX2: The random error has mean zero, $E(e_i) = 0$.

IRX3: The random error has constant variance, $\text{var}(e_i) = \sigma^2$.

IRX4: The random errors e_i and e_j for any two observations are uncorrelated, $\text{cov}(e_i, e_j) = 0$.

IRX5: The random errors $e_1, e_2, \ldots, e_N$ are statistically independent of $x_1, \ldots, x_N$, and x_i takes at least two different values.

IRX6: $e_i \sim N(0, \sigma^2)$.

Compare the assumptions IRX2, IRX3, and IRX4 with the initial assumptions about the simple regression model, SR2, SR3, and SR4. You will note that conditioning on $\mathbf{x}$ has disappeared. The reason is because when x-values and random errors e are statistically independent $E(e_i|x_j) = E(e_i) = 0$, $\text{var}(e_i|x_j) = \text{var}(e_i) = \sigma^2$ and $\text{cov}(e_i, e_j|\mathbf{x}) = \text{cov}(e_i, e_j) = 0$. Refer back to the Probability Primer Sections P.6.1 and P.6.2 for a discussion of why conditioning has no effect on the expected value and variance of statistically independent random variables. Also, it is extremely important to recognize that "i" and "j" simply represent different data observations that may be cross-sectional data or time-series data. What we say applies to both types of data.

The least squares estimators b_1 and b_2 are the best linear unbiased estimators of β_1 and β_2 if assumptions IRX1–IRX5 hold. These results are derived in Appendix 2G.2. The one apparent change is that an "expected value" appears in the formulas for the estimator variances. For example,

$$\text{var}(b_2) = \sigma^2 E\left[\frac{1}{\sum(x_i - \bar{x})^2}\right]$$

We must take the expected value of the term involving x. In practice, this actually changes nothing, because we estimate the variance in the usual way.

$$\widehat{\text{var}}(b_2) = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}$$

The estimator of the error variance remains $\hat{\sigma}^2 = \sum \hat{e}_i^2/(N-2)$ and all the usual interpretations remain the same. Thus, the computational aspects of least squares regression do not change. What has changed is our understanding of the DGP. Furthermore, if IRX6 holds then, conditional on $\mathbf{x}$, the least squares estimators have normal distributions.[12]

As we will see in Chapter 3, procedures for inference, namely interval estimators and hypothesis tests, will work in this independent random-x model the same way as in a fixed-x model. And, thanks to the central limit theorem, cited in Section 2.6, it will still be true that in *large samples* the least squares estimator has an approximate normal distribution whether x is fixed or random. This will be explored further in Chapter 5.

[12] If we do not condition on $\mathbf{x}$, no longer treating it as fixed and given, the exact distribution of the least squares estimator is not normal and is in fact unknown. Equation (2.12) shows that b_2 is a complicated combination of x's and random errors, e. Even if we know the distributions of x and e the product of random variables w_i and e_i has an unknown distribution.

2.10.2 Random and Strictly Exogenous x

Statistical independence between x_i and e_j, for all values of i and j (which may denote time-series or cross-sectional observations) is a very strong assumption and most likely only suitable in experimental situations. A weaker assumption is that the explanatory variable x is **strictly exogenous**. The phrases "strictly exogenous" and "strict exogeneity" refer to a particular technical, statistical assumption. You have no doubt heard the term **exogenous** before in your principles of economics classes. For example, in a supply and demand model, we know that the equilibrium price and quantity in a competitive market are jointly determined by the forces of supply and demand. Price and quantity are **endogenous** variables that are determined within the equilibrium system. However, we know that consumer income affects the demand equation. If income increases, the demand for a normal good increases. Income is not determined within the equilibrium system that determines equilibrium price and quantity; it is determined outside this market and is said to be **exogenous**. The exogenous variable income affects market demand, but market demand does not affect consumer income. In regression analysis models, the independent, explanatory variable x is also termed an **exogenous variable** because its variation affects the outcome variable y, but there is no reverse causality; changes in y have no effect on x.

Because interrelationships among economic variables and forces can be complex, we wish to be very precise about exogenous explanatory variables. The independent variable x is **strictly exogenous** if $E(e_i|x_j) = 0$ for all values of i and j, or equivalently, $E(e_i|x_1, x_2, \ldots, x_N) = E(e_i|\mathbf{x}) = 0$. This is exactly assumption SR2. If $i = 3$, for example, then $E(e_3|x_1) = 0$, and $E(e_3|x_3) = 0$, and $E(e_3|x_7) = 0$. The conditional expectation of the ith error term e_i is zero given *any and all* x_j. If it will help you remember them, relabel SR1–SR6 as SEX1–SEX6, where SEX stands for "strictly exogenous-x." Let the phrase "simple regression is sexy" remind you that ***S***trictly ***E***xogenous-***X*** is the baseline regression assumption.

What are the properties of the least squares estimator under the assumption of strict exogeneity? They are the same as in the case of statistical independence between all x_j and e_i. The least squares estimators are the best linear unbiased estimators of the regression parameters. These results are proved in Appendix 2G.3. This is a nice finding because while still strong, strict exogeneity is less strong than assuming x and e are statistically independent. Furthermore, if the errors are normally distributed, then the least squares estimator $b_2|\mathbf{x}$ has a normal distribution.

The Implications of Strict Exogeneity

Strict exogeneity implies quite a bit. If x is strictly exogenous, then the least squares estimator works the way we want it to and no fancier or more difficult estimators are required. Life is simple. If, on the other hand, strict exogeneity does not hold, then econometric analysis becomes more complicated, which, unfortunately, is often the case. How can we tell if the technical, statistical assumption called "strict exogeneity" holds? The only sure way is to perform a controlled experiment in which x is fixed in repeated samples or chosen randomly as described in Section 2.10.1. For most economic analyses, such experiments are impossible or too expensive.

Are there perhaps some statistical tests that can be used to check for strict exogeneity? The answer is yes, but using statistics it is much easier to determine if something is probably false rather than to argue that it is true. The common practice is to check that the implications of strict exogeneity are true. If these implications don't seem to be true, either based on economic logic or statistical tests, then we will conclude that strict exogeneity does not hold and deal with the consequences, making life more difficult. The two direct implications of strict exogeneity, $E(e_i|x_1, x_2, \ldots, x_N) = E(e_i|\mathbf{x}) = 0$, derived in Appendix 2G.1, are as follows:

Implication 1: $E(e_i) = 0$. The "average" of all factors omitted from the regression model is zero.

Implication 2: $\text{cov}(x_i, e_j) = 0$. There is no correlation between the omitted factors associated with observation j and the value of the explanatory variable for observation i.

If x satisfies the strict exogeneity condition, then $E(e_i) = 0$ and $\text{cov}(x_i, e_j) = 0$. If either of these implications is *not true*, then x is *not strictly exogenous*.

Can we check Implication 1: $E(e_i) = 0$? Is the average of all omitted factors equal to zero? In practice, this usually reduces to the question "Have I omitted anything important from the model?" If you have it is likely to be because you didn't know it was important (weak economic theory) or because, while you know it is an important factor (such as an individual's average lifetime income or an individual's perseverance in the face of adversity), it cannot be easily or well measured. In any event, omitted variables damage the least squares estimator only when Implication 2 is violated. Consequently, Implication 2 draws the most attention.

Can we check Implication 2: $\text{cov}(x_i, e_j) = 0$? Yes, we can, and we show some statistical tests in Chapter 10. However, logical arguments, and thought experiments, should always come before any statistical tests. In some cases, we can anticipate the failure of strict exogeneity, as the following examples in models using *time-series* data illustrate. In these cases, we usually index the observations using the subscript t, so that x_t is the value of the explanatory variable at time t and e_s is the value of the random error in time period s. In this context, strict exogeneity would be expressed as $E(e_s|x_t) = 0$ for all s and t. The zero covariance implication of strict exogeneity is $\text{cov}(x_t, e_s) = 0$.

Example 1. Suppose that x_t represents a policy variable, perhaps public spending on roads and bridges in month or quarter t. If the area is "shocked" by a hurricane, tornado, or other natural disaster at time s, then some time later ($t > s$) we may very well expect public spending on roads and bridges to increase, not only for one time period but perhaps for many. Then, $\text{cov}(x_{t=s+1}, e_s) \neq 0$, $\text{cov}(x_{t=s+2}, e_s) \neq 0$, and so on. Strict exogeneity fails in this case because the shock to the error term, the natural disaster, is correlated with a subsequent change in the explanatory variable, public spending, implying $E(e_s|x_t) \neq 0$.

Example 2. Suppose the quarterly sales by a firm are related to its advertising expenditures. We might write $SALES_t = \beta_1 + \beta_2 ADVERT_t + e_t$. However, advertising expenditures at time t may depend on sales revenues in the same quarter during the previous year, at time $t - 4$. That is, $ADVERT_t = f(SALES_{t-4})$. Because $SALES_{t-4} = \beta_1 + \beta_2 ADVERT_{t-4} + e_{t-4}$, it follows that there will be a correlation, and covariance, between $ADVERT_t$ and e_{t-4}. Therefore, the strict exogeneity condition fails, and $E(e_{t-4}|ADVERT_t) \neq 0$. Note the similarities between this example and the first. The effect of a past error e_s is carried forward to affect a future value of the explanatory variable, x_t, $t > s$.

Example 3. Let U_t represent the unemployment rate in quarter t, and we suppose that it is affected by the governmental expenditures, G_t. The regression might be specified as $U_t = \beta_1 + \beta_2 G_t + e_t$. However, we might imagine that the unemployment rate in this quarter is affected by government spending in previous quarters, such as G_{t-1}. Because G_{t-1} is not included in the model specification, it makes up a portion of the error term, $e_t = f(G_{t-1})$. Furthermore, we expect that there is a strong positive correlation and covariance between government spending this quarter and in previous quarters, so that $\text{cov}(G_t, G_{t-1}) > 0$. This means that we can anticipate a correlation between the error term in time t and previous levels of government spending, so that $\text{cov}(e_t, G_{t-1}) \neq 0$. Therefore, $\text{cov}(e_t|G_t) \neq 0$ and the strict exogeneity assumption fails.

The implications of a failure of the strict exogeniety assumption for least squares estimation, and the introduction of weaker assumptions to accommodate situations like those in Examples 1–3, are considered in Chapters 5, 9, and 10.

2.10.3 Random Sampling

The food expenditure example we have carried through this chapter is another case in which the DGP leads to an x that is random. We randomly sampled a population and selected 40 households. These are cross-sectional data observations. For each household, we recorded their food expenditure (y_i) and income (x_i). Because both of these variables' values are unknown to us until they are observed, both the outcome variable y and the explanatory variable x are random.

The same questions are relevant. What are the sampling properties of the least squares estimator in this case? Is the least squares estimator the best, linear unbiased estimator?

Such survey data is collected by **random sampling** from a population. Survey methodology is an important area of statistics. Public opinion surveys, market research surveys, government surveys, and censuses are all examples of collecting survey data. Several important ones are carried out by the U.S. Bureau of Labor Statistics (BLS).[13] The idea is to collect data pairs (y_i, x_i) in such a way that the ith pair [the "Smith" household] is statistically independent of the jth pair [the "Jones" household]. This ensures that x_j is statistically independent of e_i if $i \neq j$. Then, the strict exogeneity assumption reduces to concern about a possible relationship between x_i and e_i. If the conditional expectation $E(e_i|x_i) = 0$, then x is strictly exogenous, and the implications are $E(e_i) = 0$ and $\text{cov}(x_i, e_i) = 0$. Note also that if we assume that the data pairs are independent, then we no longer need make the separate assumption that the errors are uncorrelated.

What are the properties of the least squares estimator under these assumptions? They are the same as in the cases of statistical independence between all x_j and e_i (Section 2.10.1) and strict exogeneity in the general sense (Section 2.10.2). The least squares estimators are the best linear unbiased estimators of the regression parameters and conditional on **x** they have a normal distribution if SR6 (or IRX6) holds.

One final idea associated with random sampling is that the data pairs, (y_i, x_i), $i = 1, \ldots, N$, have the same joint *pdf*, $f(y, x)$. In this case, the data pairs are independent and identically distributed, *iid*. In statistics, the phrase **random sample** implies that the data are ***iid***. This is a reasonable assumption if all the data pairs are collected from the same population.

When discussing examples of the implications of strict exogeneity, we showed how the strict exogeneity assumption can be violated when using time-series data if there is correlation between e_s and a future or past value x_t ($t \neq s$). For an example of how strict exogeneity fails with random sampling of cross-sectional data, we need an example of where e_i is correlated with a value x_i corresponding to the same ith observation.

Assumptions of the Simple Linear Regression Model Under Random Sampling

RS1: The observable variables y and x are related by $y_i = \beta_1 + \beta_2 x_i + e_i$, $i = 1, \ldots, N$, where β_1 and β_2 are unknown population parameters and e_i is a random error term.

RS2: The data pairs (y_i, x_i) are statistically independent of all other data pairs and have the same joint distribution $f(y_i, x_i)$. They are independent and identically distributed.

RS3: $E(e_i|x_i) = 0$ for $i = 1, \ldots, N$; x is strictly exogenous.

RS4: The random error has constant conditional variance, $\text{var}(e_i|x_i) = \sigma^2$.

RS5: x_i takes at least two different values.

RS6: $e_i \sim N(0, \sigma^2)$.

Example 4. Suppose that x_i is a measure of the quantity of inputs used in a production process by a randomly chosen firm in an equation designed to explain a firm's production costs. The error term e_i may contain unmeasured features associated with the ability of the firm's managers. It is possible that more able managers are able to use fewer inputs in the production process, so we might expect $\text{cov}(x_i, e_i) < 0$. In this case, strict exogeneity fails. The ith firm's input usage is correlated with unmeasured characteristics of firm managers contained in the ith error, e_i. A firm's input usage is not strictly exogenous, and in econometric terms, it is said to be **endogenous**. Explanatory variables are *endogenous* if they are correlated with the error term.

[13] http://www.bls.gov/nls/home.htm

2.11 Exercises

2.11.1 Problems

2.1 Consider the following five observations. You are to do all the parts of this exercise using only a calculator.

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
3	4				
2	2				
1	3				
–1	1				
0	0				
$\sum x_i =$	$\sum y_i =$	$\sum(x_i-\bar{x}) =$	$\sum(x_i-\bar{x})^2 =$	$\sum(y_i-\bar{y}) =$	$\sum(x_i-\bar{x})(y_i-\bar{y}) =$

a. Complete the entries in the table. Put the sums in the last row. What are the sample means $\bar{x}$ and $\bar{y}$?

b. Calculate b_1 and b_2 using (2.7) and (2.8) and state their interpretation.

c. Compute $\sum_{i=1}^{5} x_i^2, \sum_{i=1}^{5} x_i y_i$. Using these numerical values, show that $\sum(x_i-\bar{x})^2 = \sum x_i^2 - N\bar{x}^2$ and $\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum x_i y_i - N\bar{x}\bar{y}$.

d. Use the least squares estimates from part (b) to compute the fitted values of y, and complete the remainder of the table below. Put the sums in the last row.
Calculate the sample variance of y, $s_y^2 = \sum_{i=1}^{N}(y_i-\bar{y})^2/(N-1)$, the sample variance of x, $s_x^2 = \sum_{i=1}^{N}(x_i-\bar{x})^2/(N-1)$, the sample covariance between x and y, $s_{xy} = \sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x})/(N-1)$, the sample correlation between x and y, $r_{xy} = s_{xy}/(s_x s_y)$ and the coefficient of variation of x, $CV_x = 100(s_x/\bar{x})$. What is the median, 50th percentile, of x?

x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$x_i\hat{e}_i$
3	4				
2	2				
1	3				
–1	1				
0	0				
$\sum x_i =$	$\sum y_i =$	$\sum \hat{y}_i =$	$\sum \hat{e}_i =$	$\sum \hat{e}_i^2 =$	$\sum x_i\hat{e}_i =$

e. On graph paper, plot the data points and sketch the fitted regression line $\hat{y}_i = b_1 + b_2 x_i$.

f. On the sketch in part (e), locate the point of the means $(\bar{x}, \bar{y})$. Does your fitted line pass through that point? If not, go back to the drawing board, literally.

g. Show that for these numerical values $\bar{y} = b_1 + b_2\bar{x}$.

h. Show that for these numerical values $\bar{\hat{y}} = \bar{y}$, where $\bar{\hat{y}} = \sum \hat{y}_i/N$.

i. Compute $\hat{\sigma}^2$.

j. Compute $\widehat{\text{var}}(b_2|\mathbf{x})$ and $\text{se}(b_2)$.

2.2 A household has weekly income of \$2000. The mean weekly expenditure for households with this income is $E(y|x = \$2000) = \mu_{y|x=\$2000} = \$220$, and expenditures exhibit variance $\text{var}(y|x = \$2{,}000) = \sigma^2_{y|x=\$2{,}000} = \$121$.

a. Assuming that weekly food expenditures are normally distributed, find the probability that a household with this income spends between \$200 and \$215 on food in a week. Include a sketch with your solution.

b. Find the probability that a household with this income spends more than $250 on food in a week. Include a sketch with your solution.

c. Find the probability in part (a) if the variance of weekly expenditures is $\text{var}(y|x = \$2{,}000) = \sigma^2_{y|x=\$2{,}000} = 144$.

d. Find the probability in part (b) if the variance of weekly expenditures is $\text{var}(y|x = \$2{,}000) = \sigma^2_{y|x=\$2{,}000} = 144$.

2.3 Graph the following observations of x and y on graph paper.

TABLE 2.4 Exercise 2.3 Data

x	1	2	3	4	5	6
y	6	4	11	9	13	17

a. Using a ruler, draw a line that fits through the data. Measure the slope and intercept of the line you have drawn.

b. Use formulas (2.7) and (2.8) to compute, using only a hand calculator, the least squares estimates of the slope and the intercept. Plot this line on your graph.

c. Obtain the sample means $\bar{y} = \sum y_i/N$ and $\bar{x} = \sum x_i/N$. Obtain the predicted value of y for $x = \bar{x}$ and plot it on your graph. What do you observe about this predicted value?

d. Using the least squares estimates from (b), compute the least squares residuals $\hat{e}_i$.

e. Find their sum, $\sum \hat{e}_i$, and their sum of squared values, $\sum \hat{e}_i^2$.

f. Calculate $\sum x_i \hat{e}_i$.

2.4 We have defined the simple linear regression model to be $y = \beta_1 + \beta_2 x + e$. Suppose, however, that we knew, for a fact, that $\beta_1 = 0$.

a. What does the linear regression model look like, algebraically, if $\beta_1 = 0$?

b. What does the linear regression model look like, graphically, if $\beta_1 = 0$?

c. If $\beta_1 = 0$, the least squares "sum of squares" function becomes $S(\beta_2) = \sum_{i=1}^{N} (y_i - \beta_2 x_i)^2$. Using the data in Table 2.4 from Exercise 2.3, plot the value of the sum of squares function for enough values of β_2 for you to locate the approximate minimum. What is the significance of the value of β_2 that minimizes $S(\beta_2)$? [*Hint*: Your computations will be simplified if you algebraically expand $S(\beta_2) = \sum_{i=1}^{N} (y_i - \beta_2 x_i)^2$ by squaring the term in parentheses and carrying through the summation operator.]

d. Using calculus, show that the formula for the least squares estimate of β_2 in this model is $b_2 = \sum x_i y_i / \sum x_i^2$. Use this result to compute b_2 and compare this value with the value you obtained geometrically.

e. Using the estimate obtained with the formula in (d), plot the fitted (estimated) regression function. On the graph locate the point $(\bar{x}, \bar{y})$. What do you observe?

f. Using the estimate obtained with the formula in (d), obtain the least squares residuals, $\hat{e}_i = y_i - b_2 x_i$. Find their sum.

g. Calculate $\sum x_i \hat{e}_i$.

2.5 A small business hires a consultant to predict the value of weekly sales of their product if their weekly advertising is increased to $2000 per week. The consultant takes a record of how much the firm spent on advertising per week and the corresponding weekly sales over the past six months. The consultant writes, "Over the past six months the average weekly expenditure on advertising has been $1500 and average weekly sales have been $10,000. Based on the results of a simple linear regression, I predict sales will be $12,000 if $2000 per week is spent on advertising."

a. What is the estimated simple regression used by the consultant to make this prediction?

b. Sketch a graph of the estimated regression line. Locate the average weekly values on the graph.

2.6 A soda vendor at Louisiana State University football games observes that the warmer the temperature at game time the greater the number of sodas that are sold. Based on 32 home games covering five years, the vendor estimates the relationship between soda sales and temperature to be $\hat{y} = -240 + 20x$, where y = the number of sodas she sells and x = temperature in degrees Fahrenheit.

a. Interpret the estimated slope and intercept. Do the estimates make sense? Why or why not?

b. On a day when the temperature at game time is forecast to be 80°F, predict how many sodas the vendor will sell.
c. Below what temperature are the predicted sales zero?
d. Sketch a graph of the estimated regression line.

2.7 We have 2008 data on y = income per capita (in thousands of dollars) and x = percentage of the population with a bachelor's degree or more for the 50 U.S. states plus the District of Columbia, a total of $N = 51$ observations. We have results from a simple linear regression of y on x.

a. The estimated error variance is $\hat{\sigma}^2 = 14.24134$. What is the sum of squared least squares residuals?
b. The estimated variance of b_2 is 0.009165. What is the standard error of b_2? What is the value of $\sum(x_i - \bar{x})^2$?
c. The estimated slope is $b_2 = 1.02896$. Interpret this result.
d. Using $\bar{x} = 27.35686$ and $\bar{y} = 39.66886$, calculate the estimate of the intercept.
e. Given the results in (b) and (d), what is $\sum x_i^2$?
f. For the state of Georgia, the value of $y = 34.893$ and $x = 27.5$. Compute the least squares residual, using the information in parts (c) and (d).

2.8 Professor I.M. Mean likes to use averages. When fitting a regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ using the $N = 6$ observations in Table 2.4 from Exercise 2.3, (y_i, x_i), Professor Mean calculates the sample means (averages) of (y_i, x_i) for the first three and second three observations in the data $\left(\bar{y}_1 = \sum_{i=1}^{3} y_i/3,\ \bar{x}_1 = \sum_{i=1}^{3} x_i/3\right)$ and $\left(\bar{y}_2 = \sum_{i=4}^{6} y_i/3,\ \bar{x}_2 = \sum_{i=4}^{6} x_i/3\right)$. Then Dr. Mean's estimator of the slope is $\hat{\beta}_{2,mean} = (\bar{y}_2 - \bar{y}_1)/(\bar{x}_2 - \bar{x}_1)$ and the Dr. Mean intercept estimator is $\hat{\beta}_{1,mean} = \bar{y} - \hat{\beta}_{2,mean}\bar{x}$, where $(\bar{y}, \bar{x})$ are the sample means using all the data. You may use a spreadsheet or other software to carry out tedious calculations.

a. Calculate $\hat{\beta}_{1,mean}$ and $\hat{\beta}_{2,mean}$. Plot the data, and the fitted line $\hat{y}_{i,mean} = \hat{\beta}_{1,mean} + \hat{\beta}_{2,mean} x_i$.
b. Calculate the residuals $\hat{e}_{i,mean} = y_i - \hat{y}_{i,mean} = y_i - \left(\hat{\beta}_{1,mean} + \hat{\beta}_{2,mean} x_i\right)$. Find $\sum_{i=1}^{6} \hat{e}_{i,mean}$, and $\sum_{i=1}^{6} x_i \hat{e}_{i,mean}$.
c. Compare the results in (b) to the corresponding values based on the least squares regression estimates. See Exercise 2.3.
d. Compute $\sum_{i=1}^{6} \hat{e}_{i,mean}^2$. Is this value larger or smaller than the sum of squared least squares residuals in Exercise 2.3(d)?

2.9 Professor I.M. Mean likes to use averages. When fitting a regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ using the $N = 6$ observations in Table 2.4 from Exercise 2.3, (y_i, x_i), Professor Mean calculates the sample means (averages) of (y_i, x_i) for the first three and second three observations in the data $\left(\bar{y}_1 = \sum_{i=1}^{3} y_i/3,\ \bar{x}_1 = \sum_{i=1}^{3} x_i/3\right)$ and $\left(\bar{y}_2 = \sum_{i=4}^{6} y_i/3,\ \bar{x}_2 = \sum_{i=4}^{6} x_i/3\right)$. Then Dr. Mean's estimator of the slope is $\hat{\beta}_{2,mean} = (\bar{y}_2 - \bar{y}_1)/(\bar{x}_2 - \bar{x}_1)$.

a. Assuming assumptions SR1–SR6 hold, show that, conditional on $\mathbf{x} = (x_1, \ldots, x_6)$, Dr. Mean's estimator is unbiased, $E\left(\hat{\beta}_{2,mean}|\mathbf{x}\right) = \beta_2$.
b. Assuming assumptions SR1–SR6 hold, show that $E\left(\hat{\beta}_{2,mean}\right) = \beta_2$.
c. Assuming assumptions SR1–SR6 hold, find the theoretical expression for $\text{var}\left(\hat{\beta}_{2,mean}|\mathbf{x}\right)$. Is this variance larger or smaller than the variance of the least squares estimator $\text{var}(b_2|\mathbf{x})$? Explain.

2.10 Consider fitting a regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ using the $N = 6$ observations in Table 2.4 from Exercise 2.3, (y_i, x_i). Suppose that based on a theoretical argument we **know** that $\beta_2 = 0$.

a. What does the regression model look like, algebraically, if $\beta_2 = 0$?
b. What does the regression model look like, graphically, if $\beta_2 = 0$?
c. If $\beta_2 = 0$ the sum of squares function becomes $S(\beta_1) = \sum_{i=1}^{N} (y_i - \beta_1)^2$. Using the data in Table 2.4, plot the sum of squares function for enough values of β_1 so that you can locate the approximate minimum. What is this value? [*Hint*: Your calculations will be easier if you square the term in parentheses and carry through the summation operator.]
d. Using calculus, show that the formula for the least squares estimate of β_1 in this model is $\hat{\beta}_1 = \left(\sum_{i=1}^{N} y_i\right)/N$.
e. Using the data in Table 2.4 and the result in part (d), compute an estimate of β_1. How does this value compare to the value you found in part (c)?

f. Using the data in Table 2.4, calculate the sum of squared residuals $S\left(\hat{\beta}_1\right) = \sum_{i=1}^{N}\left(y_i - \hat{\beta}_1\right)^2$. Is this sum of squared residuals larger or smaller than the sum of squared residuals $S\left(b_1, b_2\right) = \sum_{i=1}^{N}\left(y_i - b_1 - b_2 x_i\right)^2$ using the least squares estimates? [See Exercise 2.3 (d).]

2.11 Let y = expenditure ($) on food away from home per household member per month in the past quarter and x = monthly household income (in hundreds of dollars) during the past year.

a. Using 2013 data from three-person households ($N = 2334$), we obtain least squares estimates $\hat{y} = 13.77 + 0.52x$. Interpret the estimated slope and intercept from this relation.

b. Predict the expenditures on food away from home for a household with $2000 a month income.

c. Calculate the elasticity of expenditure on food away from home with respect to income when household income is $2000 per month. [*Hint*: Elasticity must be calculated for a point on the fitted regression.]

d. We estimate the log-linear model to be $\widehat{\ln(y)} = 3.14 + 0.007x$. What is the estimated elasticity of expenditure on food away from home with respect to income, if household income is $2000 per month?

e. For the log-linear model in part (d), calculate $\hat{y} = \exp(3.14 + 0.007x)$ when $x = 20$ and when $x = 30$. Evaluate the slope of the relation between y and x, dy/dx, for each of these $\hat{y}$ values. Based on these calculations for the log-linear model, is expenditure on food away from home increasing with respect to income at an increasing or decreasing rate?

f. When estimating the log-linear model in part (d), the number of observations used in the regression falls to $N = 2005$. How many households in the sample reported no expenditures on food away from home in the past quarter?

2.12 Let y = expenditure ($) on food away from home per household member per month in the past quarter and $x = 1$ if the household includes a member with an advanced degree, a Master's, or Ph.D./Professional degree, and $x = 0$ otherwise.

a. Using 2013 data from three-person households ($N = 2334$), we obtain least squares estimates $\hat{y} = 44.96 + 30.41x$. Interpret the coefficient of x and the intercept from this relation.

b. What is the per person sample mean of food expenditures away from home for a household including someone with an advanced degree?

c. What is the per person sample mean of food expenditures away from home for a household that does not include someone with an advanced degree?

2.13 Using 2011 data on 141 U.S. public research universities, we examine the relationship between academic cost per student, *ACA* (real total academic cost per student in thousands of dollars) and full-time enrollment *FTESTU* (in thousands of students).

a. The least squares fitted relation is $\widehat{ACA} = 14.656 + 0.266FTESTU$. What is the economic interpretation of the estimated parameters? Why isn't the intercept zero?

b. In 2011 Louisiana State University (LSU) had a full-time student enrollment of 27,950. Using the fitted related in part (a), compute the predicted value of *ACA*.

c. The actual value of *ACA* for LSU that year was 21.403. Calculate the least squares residual for LSU? Does the model overpredict or underpredict *ACA* for LSU?

d. The sample mean (average) full-time enrollment in U.S. public research universities in 2011 was 22,845.77. What was the sample mean of academic cost per student?

2.14 Consider the regression model $WAGE = \beta_1 + \beta_2 EDUC + e$, where *WAGE* is hourly wage rate in U.S. 2013 dollars and *EDUC* is years of education, or schooling. The regression model is estimated twice using the least squares estimator, once using individuals from an urban area, and again for individuals in a rural area.

$$\text{Urban} \quad \underset{(\text{se})}{\widehat{WAGE} = -10.76 + 2.46\,EDUC,} \quad \underset{(2.27)\ (0.16)}{N = 986}$$

$$\text{Rural} \quad \underset{(\text{se})}{\widehat{WAGE} = -4.88 + 1.80\,EDUC,} \quad \underset{(3.29)\ (0.24)}{N = 214}$$

a. Using the estimated rural regression, compute the elasticity of wages with respect to education at the "point of the means." The sample mean of *WAGE* is $19.74.

b. The sample mean of *EDUC* in the urban area is 13.68 years. Using the estimated urban regression, compute the standard error of the elasticity of wages with respect to education at the "point of the means." Assume that the mean values are "givens" and not random.

c. What is the predicted wage for an individual with 12 years of education in each area? With 16 years of education?

2.15 Professor E.Z. Stuff has decided that the least squares estimator is too much trouble. Noting that two points determine a line, Dr. Stuff chooses two points from a sample of size N and draws a line between them, calling the slope of this line the EZ estimator of β_2 in the simple regression model. Algebraically, if the two points are (x_1, y_1) and (x_2, y_2), the EZ estimation rule is

$$b_{EZ} = \frac{y_2 - y_1}{x_2 - x_1}$$

Assuming that all the assumptions of the simple regression model hold:

a. Show that b_{EZ} is a "linear" estimator.

b. Show that b_{EZ} is an unbiased estimator.

c. Find the conditional variance of b_{EZ}.

d. Find the conditional probability distribution of b_{EZ}.

e. Convince Professor Stuff that the EZ estimator is not as good as the least squares estimator. No proof is required here.

2.11.2 Computer Exercises

2.16 The capital asset pricing model (CAPM) is an important model in the field of finance. It explains variations in the rate of return on a security as a function of the rate of return on a portfolio consisting of all publicly traded stocks, which is called the *market* portfolio. Generally, the rate of return on any investment is measured relative to its opportunity cost, which is the return on a risk-free asset. The resulting difference is called the *risk premium*, since it is the reward or punishment for making a risky investment. The CAPM says that the risk premium on security j is *proportional* to the risk premium on the market portfolio. That is,

$$r_j - r_f = \beta_j(r_m - r_f)$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, r_m is the return on the market portfolio, and β_j is the jth security's "*beta*" value. A stock's *beta* is important to investors since it reveals the stock's volatility. It measures the sensitivity of security j's return to variation in the whole stock market. As such, values of *beta* less than one indicate that the stock is "defensive" since its variation is less than the market's. A *beta* greater than one indicates an "aggressive stock." Investors usually want an estimate of a stock's *beta* before purchasing it. The CAPM model shown above is the "economic model" in this case. The "econometric model" is obtained by including an intercept in the model (even though theory says it should be zero) and an error term

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f) + e_j$$

a. Explain why the econometric model above is a simple regression model like those discussed in this chapter.

b. In the data file *capm5* are data on the monthly returns of six firms (GE, IBM, Ford, Microsoft, Disney, and Exxon-Mobil), the rate of return on the market portfolio (*MKT*), and the rate of return on the risk-free asset (*RISKFREE*). The 180 observations cover January 1998 to December 2012. Estimate the CAPM model for each firm, and comment on their estimated *beta* values. Which firm appears most aggressive? Which firm appears most defensive?

c. Finance theory says that the intercept parameter α_j should be zero. Does this seem correct given your estimates? For the Microsoft stock, plot the fitted regression line along with the data scatter.

d. Estimate the model for each firm under the assumption that $\alpha_j = 0$. Do the estimates of the *beta* values change much?

2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

a. Plot house price against house size in a scatter diagram.

b. Estimate the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$. Interpret the estimates. Draw a sketch of the fitted line.

c. Estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.

d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

f. For the regressions in (b) and (c), compute the least squares residuals and plot them against *SQFT*. Do any of our assumptions appear violated?

g. One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (*SSE*) from the models in (b) and (c). Which model has a lower *SSE*? How does having a lower *SSE* indicate a "better-fitting" model?

2.18 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

a. Create histograms for *PRICE* and ln(*PRICE*). Are the distributions skewed or symmetrical?

b. Estimate the log-linear regression model $\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$. Interpret the OLS estimates, $\hat{\gamma}_1$ and $\hat{\gamma}_2$. Graph the fitted *PRICE*, $\widehat{PRICE} = \exp\left(\hat{\gamma}_1 + \hat{\gamma}_2 SQFT\right)$, against *SQFT*, and sketch the tangent line to the curve for a house with 2000 square feet of living area. What is the slope of the tangent line?

c. Compute the least squares residuals from the model in (b) and plot them against *SQFT*. Do any of our assumptions appear violated?

d. Calculate summary statistics for *PRICE* and *SQFT* for homes close to Louisiana State University (*CLOSE* = 1) and for homes not close to the university (*CLOSE* = 0). What differences and/or similarities do you observe?

e. Estimate the log-linear regression model $\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$ for homes close to Louisiana State University (*CLOSE* = 1) and for homes not close to the university (*CLOSE* = 0). Interpret the estimated coefficient of *SQFT* in each sample's regression.

f. Are the regression results in part (b) valid if the differences you observe in part (e) are substantial? Think in particular about whether SR1 is satisfied.

2.19 The data file *stockton5_small* contains observations on 1200 houses sold in Stockton, California, during 1996–1998. [Note: the data file *stockton5* includes 2610 observations.] Scale the variable *SPRICE* to units of $1000, by dividing it by 1000.

a. Plot house selling price *SPRICE* against house living area for all houses in the sample.

b. Estimate the regression model $SPRICE = \beta_1 + \beta_2 LIVAREA + e$ for all the houses in the sample. Interpret the estimates. Draw a sketch of the fitted line.

c. Estimate the quadratic model $SPRICE = \alpha_1 + \alpha_2 LIVAREA^2 + e$ for all the houses in the sample. What is the marginal effect of an additional 100 square feet of living area for a home with 1500 square feet of living area.

d. In the same graph, plot the fitted lines from the linear and quadratic models. Which seems to fit the data better? Compare the sum of squared residuals (*SSE*) for the two models. Which is smaller?

e. If the quadratic model is in fact "true," what can we say about the results and interpretations we obtain for the linear relationship in part (b)?

2.20 The data file *stockton5_small* contains observations on 1200 houses sold in Stockton, California, during 1996–1998. [Note: The data file *stockton5* includes 2610 observations.]. Scale the variable *SPRICE* to units of $1000, by dividing it by 1000.

a. Estimate the regression model $SPRICE = \beta_1 + \beta_2 LIVAREA + e$ using only houses that are on large lots. Repeat the estimation for houses that are not on large lots. Finally, estimate the regression using data on both large and small lots. Interpret the estimates. How do the estimates compare?

b. Estimate the regression model $SPRICE = \alpha_1 + \alpha_2 LIVAREA^2 + e$ using only houses that are on large lots. Repeat the estimation for houses that are not on large lots. Interpret the estimates. How do the estimates compare?

c. Estimate a linear regression $SPRICE = \eta_1 + \eta_2 LGELOT + e$ with dependent variable *SPRICE* and independent variable the indicator *LGELOT*, which identifies houses on larger lots. Interpret these results.

d. If the estimates in part (a) and/or part (b) differ substantially for the large lot and small lot subsamples, will assumption SR1 be satisfied in the model that pools all the observations together? If not, why not? Do the results in (c) offer any information about the potential validity of SR1?

2.21 The data file *stockton5_small* contains observations on 1200 houses sold in Stockton, California, during 1996–1998. [Note: the data file *stockton5* includes 2610 observations.] Scale the variable *SPRICE* to units of $1000, by dividing it by 1000.

a. Estimate the linear model $SPRICE = \delta_1 + \delta_2 AGE + e$. Interpret the estimated coefficients. Predict the selling price of a house that is 30 years old.

b. Using the results in part (a), plot house selling price against *AGE* and show the fitted regression line. Based on the plot, does the model fit the data well? Explain.

c. Estimate the log-linear model $\ln(SPRICE) = \theta_1 + \theta_2 AGE + e$. Interpret the estimated slope coefficient.

d. Using the results in part (c), compute $\widehat{SPRICE} = \exp\left(\hat{\theta}_1 + \hat{\theta}_2 AGE\right)$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the OLS estimates. Plot $\widehat{SPRICE}$ against *AGE* (connecting the dots) and *SPRICE* vs. *AGE* in the same graph.

e. Predict the selling price of a house that is 30 years old using $\widehat{SPRICE} = \exp\left(\hat{\theta}_1 + \hat{\theta}_2 AGE\right)$.

f. Based on the plots and visual fit of the estimated regression lines, which of the two models in (a) or (c) would you prefer? Explain. For each model calculate $\sum_{i=1}^{1200}\left(SPRICE - \widehat{SPRICE}\right)^2$. Is this at all useful in making a comparison between the models? If so, how?

2.22 A longitudinal experiment was conducted in Tennessee beginning in 1985 and ending in 1989. A single cohort of students was followed from kindergarten through third grade. In the experiment children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes are contained in the data file *star5_small*. [Note: The data file *star5* contains more observations and variables.]

a. Using children who are in either a regular-sized class or a small class, estimate the regression model explaining students' combined aptitude scores as a function of class size, $TOTALSCORE_i = \beta_1 + \beta_2 SMALL_i + e_i$. Interpret the estimates. Based on this regression result, what do you conclude about the effect of class size on learning?

b. Repeat part (a) using dependent variables *READSCORE* and *MATHSCORE*. Do you observe any differences?

c. Using children who are in either a regular-sized class or a regular-sized class with a teacher aide, estimate the regression model explaining students' combined aptitude scores as a function of the presence of a teacher aide, $TOTALSCORE = \gamma_1 + \gamma_2 AIDE + e$. Interpret the estimates. Based on this regression result, what do you conclude about the effect on learning of adding a teacher aide to the classroom?

d. Repeat part (c) using dependent variables *READSCORE* and *MATHSCORE*. Do you observe any differences?

2.23 Professor Ray C. Fair has for a number of years built and updated models that explain and predict the U.S. presidential elections. Visit his website at https://fairmodel.econ.yale.edu/vote2016/index2.htm. See in particular his paper entitled "Presidential and Congressional Vote-Share Equations: November 2010 Update." The basic premise of the model is that the Democratic Party's share of the two-party [Democratic and Republican] popular vote is affected by a number of factors relating to the economy, and variables relating to the politics, such as how long the incumbent party has been in power, and whether the President is running for reelection. Fair's data, 26 observations for the election years from 1916 to 2016, are in the data file *fair5*. The dependent variable is *VOTE* = percentage share of the popular vote won by the Democratic Party. Consider the effect of economic growth on *VOTE*. If Democrats are the incumbent party (*INCUMB* = 1) then economic growth, the growth rate in real per capita GDP in the first three quarters of the election year (annual rate), should enhance their chances of winning. On the other hand, if the Republicans are the incumbent party (*INCUMB* = −1), growth will diminish the Democrats' chances of winning. Consequently, we define the explanatory variable *GROWTH* = *INCUMB* × growth rate.

a. Using the data for 1916–2012, plot a scatter diagram of *VOTE* against *GROWTH*. Does there appear to be a positive association?

b. Estimate the regression $VOTE = \beta_1 + \beta_2 GROWTH + e$ by least squares using the data from 1916 to 2012. Report and discuss the estimation result. Plot the fitted line on the scatter diagram from (a).

c. Using the model estimated in (b), predict the 2016 value of *VOTE* based on the actual 2016 value for *GROWTH*. How does the predicted vote for 2016 compare to the actual result?

d. Economy wide inflation may spell doom for the incumbent party in an election. The variable $INFLAT = INCUMB \times$ inflation rate, where the inflation rate is the growth in prices over the first 15 quarters of an administration. Using the data from 1916 to 2012, plot *VOTE* against *INFLAT*.

e. Using the data from 1916 to 2012, report and discuss the estimation results for the model $VOTE = \alpha_1 + \alpha_2 INFLAT + e$.

f. Using the model estimated in (e), predict the 2016 value of *VOTE* based on the actual 2012 value for *INFLAT*. How does the predicted vote for 2016 compare to the actual result?

2.24 Using data on the "Ashcan School"[14] we have an opportunity to study the market for art. What factors determine the value of a work of art? Use the data in *ashcan_small*. [Note: The file *ashcan* contains more variables.] For this exercise, use data only on works that sold ($SOLD = 1$).

a. Using data on works that sold, construct a histogram for *RHAMMER* and compute summary statistics. What are the mean and median prices for the artwork sold? What are the 25th and 75th percentiles?

b. Using data on works that sold, construct a histogram for ln(*RHAMMER*). Describe the shape of this histogram as compared to that in part (a).

c. Plot ln(*RHAMMER*) against the age of the painting at the time of its sale, $YEARS\_OLD = DATE\_AUCTN - CREATION$. Include in the plot the least squares fitted line. What patterns do you observe?

d. Use data on works that sold, estimate the regression $\ln(RHAMMER) = \beta_1 + \beta_2 YEAR\_SOLD + e$. Interpret the estimated coefficient of *YEARS_OLD*.

e. *DREC* is an indicator variable equaling 1 if the work was sold during a recession. Using data on works that sold, estimate the regression $\ln(RHAMMER) = \alpha_1 + \alpha_2 DREC + e$. Interpret the estimated coefficient of *DREC*.

2.25 Consumer expenditure data from 2013 are contained in the file *cex5_small*. [Note: *cex5* is a larger version with more observations and variables.] Data are on three-person households consisting of a husband and wife, plus one other member, with incomes between \$1000 per month to \$20,000 per month. *FOODAWAY* is past quarter's food away from home expenditure per month per person, in dollars, and *INCOME* is household monthly income during past year, in \$100 units.

a. Construct a histogram of *FOODAWAY* and its summary statistics. What are the mean and median values? What are the 25th and 75th percentiles?

b. What are the mean and median values of *FOODAWAY* for households including a member with an advanced degree? With a college degree member? With no advanced or college degree member?

c. Construct a histogram of ln(*FOODAWAY*) and its summary statistics. Explain why *FOODAWAY* and ln(*FOODAWAY*) have different numbers of observations.

d. Estimate the linear regression $\ln(FOODAWAY) = \beta_1 + \beta_2 INCOME + e$. Interpret the estimated slope.

e. Plot ln(*FOODAWAY*) against *INCOME*, and include the fitted line from part (d).

f. Calculate the least squares residuals from the estimation in part (d). Plot them vs. *INCOME*. Do you find any unusual patterns, or do they seem completely random?

2.26 Consumer expenditure data from 2013 are contained in the file *cex5_small*. [Note: *cex5* is a larger version with more observations and variables.] Data are on three-person households consisting of a husband and wife, plus one other member, with incomes between \$1000 per month to \$20,000 per month. *FOODAWAY* is past quarter's food away from home expenditure per month per person, in dollars, and *INCOME* is household monthly income during past year, in \$100 units.

a. Estimate the linear regression $FOODAWAY = \beta_1 + \beta_2 INCOME + e$. Interpret the estimated slope.

b. Calculate the least squares residuals from the estimation in part (b). Plot them vs. *INCOME*. Do you find any unusual patterns, or do they seem completely random?

c. Estimate the linear regression $FOODAWAY = \alpha_1 + \alpha_2 ADVANCED + e$. Interpret the estimated coefficient of *ADVANCED*.

d. What are the sample means of *FOODAWAY* for households including a member with an advanced degree? With no advanced degree member? How do these values relate to the regression in part (c)?

[14] Robert B. Ekelund, Jr., John D. Jackson, and Robert D. Tollison "Are Art Auction Estimates Biased" published in *Southern Economic Journal*, 80(2), 2013, 454–465; also http://en.wikipedia.org/wiki/Ashcan_School

2.27 The owners of a motel discovered that a defective product was used in its construction. It took seven months to correct the defects during which 14 rooms in the 100-unit motel were taken out of service for one month at a time. For this exercise use the data file *motel*.

a. Graph $y = MOTEL\_PCT$, percentage motel occupancy, against $x = 100RELPRICE$, which is the percentage of the competitor's price per room charged by the motel in question. Describe the relationship between the variables based on the graph. Is there a positive association, an inverse association, or no association?

b. Consider the linear regression $MOTEL\_PCT_t = \beta_1 + \beta_2 100RELPRICE_t + e_t$. What sign do you predict for the slope coefficient? Why? Does the sign of the estimated slope agree with your expectation?

c. Calculate the least squares residuals from the regression in (b). Plot the residuals against $TIME = 1, \ldots, 25$ (month 1 = March 2003, … , month 25 = March 2005). On the graph indicate residuals when $TIME = 17, 18, \ldots, 23$. These are the months of repair. Does the model overpredict or underpredict the motel's occupancy rates for those months?

d. Estimate the linear regression $MOTEL\_PCT_t = \alpha_1 + \alpha_2 REPAIR_t + e_t$, where $REPAIR_t = 1$ for months when repairs were occurring and $REPAIR_t = 0$ otherwise. What was the motel's mean occupancy rate when there were no repairs being made? What was the motel's mean occupancy rate when repairs were being made?

2.28 How much does education affect wage rates? The data file *cps5_small* contains 1200 observations on hourly wage rates, education, and other variables from the 2013 Current Population Survey (CPS). [Note: *cps5* is a larger version.]

a. Obtain the summary statistics and histograms for the variables *WAGE* and *EDUC*. Discuss the data characteristics.

b. Estimate the linear regression $WAGE = \beta_1 + \beta_2 EDUC + e$ and discuss the results.

c. Calculate the least squares residuals and plot them against *EDUC*. Are any patterns evident? If assumptions SR1–SR5 hold, should any patterns be evident in the least squares residuals?

d. Estimate separate regressions for males, females, blacks, and whites. Compare the results.

e. Estimate the quadratic regression $WAGE = \alpha_1 + \alpha_2 EDUC^2 + e$ and discuss the results. Estimate the marginal effect of another year of education on wage for a person with 12 years of education and for a person with 16 years of education. Compare these values to the estimated marginal effect of education from the linear regression in part (b).

f. Plot the fitted linear model from part (b) and the fitted values from the quadratic model from part (e) in the same graph with the data on *WAGE* and *EDUC*. Which model appears to fit the data better?

2.29 How much does education affect wage rates? The data file *cps5_small* contains 1200 observations on hourly wage rates, education, and other variables from the 2013 Current Population Survey (CPS). [Note: *cps5* is a larger version with more observations and variables.]

a. Create the variable $LWAGE = \ln(WAGE)$. Construct a histogram and calculate detailed summary statistics. Does the histogram appear bell shaped and normally distributed? A normal distribution is symmetrical with no skewness, *skewness* = 0. The tails of the normal distribution have a certain "thickness." A measure of the tail thickness is *kurtosis*, discussed in Appendix C.4.2. For a normal distribution, the *kurtosis* = 3, discussed in Appendix C.7.4. How close are the measures of *skewness* and *kurtosis* for *LWAGE* to 0 and 3, respectively?

b. Obtain the OLS estimates from the log-linear regression model $\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$ and interpret the estimated value of β_2.

c. Obtain the predicted wage, $\widehat{WAGE} = \exp(b_1 + b_2 EDUC)$, for a person with 12 years of education and for a person with 16 years of education.

d. What is the marginal effect of additional education for a person with 12 years of education and for a person with 16 years of education? [*Hint*: This is the slope of the fitted model at those two points.]

e. Plot the fitted values $\widehat{WAGE} = \exp(b_1 + b_2 EDUC)$ versus *EDUC* in a graph. Also include in the graph the fitted linear relationship. Based on the graph, which model seems to fit the data better, the linear or log-linear model?

f. Using the fitted values from the log-linear model, compute $\sum\left(WAGE - \widehat{WAGE}\right)^2$. Compare this value to the sum of squared residuals from the estimated linear relationship. Using this as a basis of comparison, which model fits the data better?

2.30 In this exercise, we consider the amounts that are borrowed for single family home purchases in Las Vegas, Nevada, during 2010. Use the data file *vegas5_small* for this exercise.

- **a.** Compute summary statistics for *AMOUNT*, *FICO, RATE*, and *TERM30*. What is the sample average amount borrowed? What *FICO* score corresponds to the 90th percentile? What is the median interest rate paid, and what percent of the mortgages were for 30-year terms?
- **b.** Construct histograms for *AMOUNT*, ln(*AMOUNT*), *FICO*, and *RATE*. Are the empirical distributions symmetrical? Do they have one peak (unimodal) or two peaks (bimodal)?
- **c.** Estimate regressions for dependent variables *AMOUNT* and ln(*AMOUNT*) against the independent variable *FICO*. For each regression, interpret the coefficient of *FICO*.
- **d.** Estimate regressions for dependent variables *AMOUNT* and ln(*AMOUNT*) against the independent variable *RATE*. For each regression, interpret the coefficient of *RATE*.
- **e.** Estimate a regression with dependent variable *AMOUNT* and explanatory variable *TERM30*. Obtain the summary statistics for *AMOUNT* for transactions with 30-year loans and for those transactions when the term was not 30 years. Explain the regression results in terms of the summary statistics you have calculated.

Appendix 2A Derivation of the Least Squares Estimates

Given the sample observations on y and x, we want to find values for the unknown parameters β_1 and β_2 that minimize the "sum of squares" function

$$S(\beta_1, \beta_2) = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2 \tag{2A.1}$$

Since the points (y_i, x_i) have been observed, the sum of squares function S depends only on the unknown parameters β_1 and β_2. This function, which is a quadratic in terms of the unknown parameters β_1 and β_2, is a "bowl-shaped surface" like the one depicted in Figure 2A.1.

Our task is to find, out of all the possible values β_1 and β_2, the point (b_1, b_2) at which the sum of squares function S is a minimum. This minimization problem is a common one in calculus, and the minimizing point is at the "bottom of the bowl."

Those of you familiar with calculus and "partial differentiation" can verify that the partial derivatives of S with respect to β_1 and β_2 are

$$\frac{\partial S}{\partial \beta_1} = 2N\beta_1 - 2\sum y_i + 2\left(\sum x_i\right)\beta_2$$

$$\frac{\partial S}{\partial \beta_2} = 2\left(\sum x_i^2\right)\beta_2 - 2\sum x_i y_i + 2\left(\sum x_i\right)\beta_1 \tag{2A.2}$$

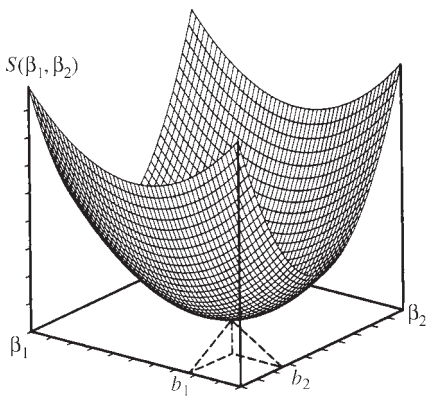


FIGURE 2A.1 **The sum of squares function and the minimizing values b_1 and b_2.**

These derivatives are equations of the slope of the bowl-like surface in the directions of the axes. Intuitively, the "bottom of the bowl" occurs where the slope of the bowl, in the direction of each axis, $\partial S/\partial\beta_1$ and $\partial S/\partial\beta_2$, is zero.

Algebraically, to obtain the point (b_1, b_2), we set (2A.2) to zero and replace β_1 and β_2 by b_1 and b_2, respectively, to obtain

$$2\left[\sum y_i - Nb_1 - \left(\sum x_i\right)b_2\right] = 0$$

$$2\left[\sum x_i y_i - \left(\sum x_i\right)b_1 - \left(\sum x_i^2\right)b_2\right] = 0$$

Simplifying these gives equations usually known as the **normal equations**:

$$Nb_1 + \left(\sum x_i\right)b_2 = \sum y_i \tag{2A.3}$$

$$\left(\sum x_i\right)b_1 + \left(\sum x_i^2\right)b_2 = \sum x_i y_i \tag{2A.4}$$

These two equations have two unknowns b_1 and b_2. We can find the least squares estimates by solving these two linear equations for b_1 and b_2. To solve for b_2, multiply (2A.3) by $\sum x_i$, multiply (2A.4) by N, then subtract the first equation from the second, and then isolate b_2 on the left-hand side.

$$b_2 = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{2A.5}$$

This formula for b_2 is in terms of data sums, cross-products, and squares. The deviation from the mean form of the estimator is derived in Appendix 2B.

To solve for b_1, given b_2, divide both sides of (2A.3) by N and rearrange.

Appendix 2B Deviation from the Mean Form of b_2

The first step in the conversion of the formula for b_2 into (2.7) is to use some tricks involving summation signs. The first useful fact is that

$$\begin{aligned}\sum\left(x_i - \bar{x}\right)^2 &= \sum x_i^2 - 2\bar{x}\sum x_i + N\bar{x}^2 = \sum x_i^2 - 2\bar{x}\left(N\tfrac{1}{N}\sum x_i\right) + N\bar{x}^2 \\ &= \sum x_i^2 - 2N\bar{x}^2 + N\bar{x}^2 = \sum x_i^2 - N\bar{x}^2\end{aligned} \tag{2B.1}$$

Should you ever have to calculate $\sum\left(x_i - \bar{x}\right)^2$, using the shortcut formula $\sum\left(x_i - \bar{x}\right)^2 = \sum x_i^2 - N\bar{x}^2$ is usually much easier. Then

$$\sum\left(x_i - \bar{x}\right)^2 = \sum x_i^2 - N\bar{x}^2 = \sum x_i^2 - \bar{x}\sum x_i = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N} \tag{2B.2}$$

To obtain this result, we have used the fact that $\bar{x} = \sum x_i/N$, so $\sum x_i = N\bar{x}$.

The second useful fact is similar to the first, and it is

$$\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \sum x_i y_i - N\bar{x}\bar{y} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{N} \tag{2B.3}$$

This result is proven in a similar manner.

If the numerator and denominator of b_2 in (2A.5) are divided by N, then using (2B.1)–(2B.3), we can rewrite b_2 in *deviation from the mean form* as

$$b_2 = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum\left(x_i - \bar{x}\right)^2}$$

This formula for b_2 is one that you should remember, as we will use it time and time again in the next few chapters.

Appendix 2C b_2 Is a Linear Estimator

In order to derive (2.10), we make a further simplification using another property of sums. The sum of any variable about its average is zero; that is,

$$\sum(x_i - \bar{x}) = 0$$

Then, the formula for b_2 becomes

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})y_i - \bar{y}\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$
$$= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \sum\left[\frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right] y_i = \sum w_i y_i$$

where w_i is given in (2.11).

Appendix 2D Derivation of Theoretical Expression for b_2

To obtain (2.12) replace y_i in (2.10) by $y_i = \beta_1 + \beta_2 x_i + e_i$ and simplify:

$$b_2 = \sum w_i y_i = \sum w_i(\beta_1 + \beta_2 x_i + e_i)$$
$$= \beta_1 \sum w_i + \beta_2 \sum w_i x_i + \sum w_i e_i$$
$$= \beta_2 + \sum w_i e_i$$

We used two more summation tricks to simplify this. First, $\sum w_i = 0$; this eliminates the term $\beta_1 \sum w_i$. Secondly, $\sum w_i x_i = 1$, so $\beta_2 \sum w_i x_i = \beta_2$, and (2.10) simplifies to (2.12).

The term $\sum w_i = 0$ because

$$\sum w_i = \sum\left[\frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right] = \frac{1}{\sum(x_i - \bar{x})^2}\sum(x_i - \bar{x}) = 0$$

where in the last step we used the fact that $\sum(x_i - \bar{x}) = 0$.

To show that $\sum w_i x_i = 1$ we again use $\sum(x_i - \bar{x}) = 0$. Another expression for $\sum(x_i - \bar{x})^2$ is

$$\sum(x_i - \bar{x})^2 = \sum(x_i - \bar{x})(x_i - \bar{x})$$
$$= \sum(x_i - \bar{x})x_i - \bar{x}\sum(x_i - \bar{x})$$
$$= \sum(x_i - \bar{x})x_i$$

Consequently,

$$\sum w_i x_i = \frac{\sum(x_i - \bar{x})x_i}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})x_i}{\sum(x_i - \bar{x})x_i} = 1$$

Appendix 2E Deriving the Conditional Variance of b_2

The starting point is equation (2.12), $b_2 = \beta_2 + \sum w_i e_i$. The least squares estimator is a random variable whose conditional variance is defined to be

$$\text{var}(b_2|\mathbf{x}) = E\left\{\left[b_2 - E(b_2|\mathbf{x})\right]^2 \middle| \mathbf{x}\right\}$$

Substituting in (2.12) and using the conditional unbiasedness of the least squares estimator, $E(b_2|\mathbf{x}) = \beta_2$, we have

$$\begin{aligned}
\text{var}(b_2|\mathbf{x}) &= E\left\{ \left[\beta_2 + \sum w_i e_i - \beta_2\right]^2 \middle| \mathbf{x} \right\} \\
&= E\left\{ \left[\sum w_i e_i\right]^2 \middle| \mathbf{x} \right\} \\
&= E\left\{ \left[\sum w_i^2 e_i^2 + \sum\sum_{i \neq j} w_i w_j e_i e_j \right] \middle| \mathbf{x} \right\} \qquad \text{[square of bracketed term]} \\
&= E\left\{ \left[\sum w_i^2 e_i^2\right] \middle| \mathbf{x} \right\} + E\left\{ \left[\sum\sum_{i \neq j} w_i w_j e_i e_j \right] \middle| \mathbf{x} \right\} \\
&= \sum w_i^2 E(e_i^2|\mathbf{x}) + \sum\sum_{i \neq j} w_i w_j E(e_i e_j|\mathbf{x}) \qquad \text{[because } w_i \text{ not random given } \mathbf{x}\text{]} \\
&= \sigma^2 \sum w_i^2 \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

The next to last line is obtained by using two assumptions: First,

$$\sigma^2 = \text{var}(e_i|\mathbf{x}) = E\left\{ \left[e_i - E(e_i|\mathbf{x})\right]^2 \middle| \mathbf{x} \right\} = E\left[\left(e_i - 0\right)^2 \middle| \mathbf{x} \right] = E(e_i^2|\mathbf{x})$$

Second, $\text{cov}(e_i, e_j|\mathbf{x}) = E\left\{ \left[e_i - E(e_i|\mathbf{x})\right]\left[e_j - E(e_j|\mathbf{x})\right] \middle| \mathbf{x} \right\} = E(e_i e_j|\mathbf{x}) = 0$. Then, the very last step uses the fact that

$$\sum w_i^2 = \sum \left[\frac{(x_i - \bar{x})^2}{\left\{ \sum (x_i - \bar{x})^2 \right\}^2} \right] = \frac{\sum (x_i - \bar{x})^2}{\left\{ \sum (x_i - \bar{x})^2 \right\}^2} = \frac{1}{\sum (x_i - \bar{x})^2}$$

Alternatively, we can employ the rule for finding the variance of a sum. If X and Y are random variables, and a and b are constants, then

$$\text{var}(aX + bY) = a^2 \text{ var}(X) + b^2 \text{ var}(Y) + 2ab\text{cov}(X, Y)$$

Appendix B.4 reviews all the basic properties of random variables. In the second line below we use this rule extended to more than two random variables. Then,

$$\begin{aligned}
\text{var}(b_2|\mathbf{x}) &= \text{var}\left[\left(\beta_2 + \sum w_i e_i\right) \middle| \mathbf{x} \right] && \text{[since } \beta_2 \text{ is a constant]} \\
&= \sum w_i^2 \text{var}(e_i|\mathbf{x}) + \sum\sum_{i \neq j} w_i w_j \text{cov}(e_i, e_j|\mathbf{x}) && \text{[generalizing the variance rule]} \\
&= \sum w_i^2 \text{var}(e_i|\mathbf{x}) && \text{[using } \text{cov}(e_i, e_j|\mathbf{x}) = 0\text{]} \\
&= \sigma^2 \sum w_i^2 && \text{[using } \text{var}(e_i|\mathbf{x}) = \sigma^2\text{]} \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

Carefully note that the derivation of the variance expression for b_2 depends on assumptions SR3 and SR4. If the $\text{cov}(e_i, e_j|\mathbf{x}) \neq 0$, then we cannot drop out all those terms in the double summation. If $\text{var}(e_i|\mathbf{x}) \neq \sigma^2$ for all observations, then σ^2 cannot be factored out of the summation. If either of these assumptions fails to hold, then the conditional variance $\text{var}(b_2|\mathbf{x})$ is *something else*, and is not given by (2.15). The same is true for the conditional variance of b_1 and the conditional covariance between b_1 and b_2.

Appendix 2F Proof of the Gauss–Markov Theorem

We will prove the Gauss–Markov theorem for the least squares estimator b_2 of β_2. Our goal is to show that in the class of linear and unbiased estimators the estimator b_2 has the smallest variance. Let $b_2^* = \sum k_i y_i$ (where k_i are constants) be any other linear estimator of β_2. To make comparison to the least squares estimator b_2 easier, suppose that $k_i = w_i + c_i$, where c_i is another constant and w_i is given in (2.11). While this is tricky, it is legal, since for any k_i that someone might choose we can find c_i. Into this new estimator, substitute y_i and simplify, using the properties of w_i in Appendix 2D.

$$\begin{aligned} b_2^* &= \sum k_i y_i = \sum (w_i + c_i) y_i = \sum (w_i + c_i)(\beta_1 + \beta_2 x_i + e_i) \\ &= \sum (w_i + c_i)\beta_1 + \sum (w_i + c_i)\beta_2 x_i + \sum (w_i + c_i) e_i \\ &= \beta_1 \sum w_i + \beta_1 \sum c_i + \beta_2 \sum w_i x_i + \beta_2 \sum c_i x_i + \sum (w_i + c_i) e_i \\ &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i + \sum (w_i + c_i) e_i \end{aligned} \tag{2F.1}$$

since $\sum w_i = 0$ and $\sum w_i x_i = 1$.

Take the mathematical expectation of the last line in (2F.1), using the properties of expectation and the assumption that $E(e_i|\mathbf{x}) = 0$:

$$\begin{aligned} E(b_2^*|\mathbf{x}) &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i + \sum (w_i + c_i) E(e_i|\mathbf{x}) \\ &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i \end{aligned} \tag{2F.2}$$

In order for the linear estimator $b_2^* = \sum k_i y_i$ to be unbiased, it must be true that

$$\sum c_i = 0 \quad \text{and} \quad \sum c_i x_i = 0 \tag{2F.3}$$

These conditions must hold in order for $b_2^* = \sum k_i y_i$ to be in the class of *linear* and *unbiased estimators*. So we will assume that conditions (2F.3) hold and use them to simplify expression (2F.1):

$$b_2^* = \sum k_i y_i = \beta_2 + \sum (w_i + c_i) e_i \tag{2F.4}$$

We can now find the variance of the linear unbiased estimator b_2^* following the steps in Appendix 2E and using the additional fact that

$$\sum c_i w_i = \sum \left[\frac{c_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] = \frac{1}{\sum (x_i - \bar{x})^2} \sum c_i x_i - \frac{\bar{x}}{\sum (x_i - \bar{x})^2} \sum c_i = 0$$

Use the properties of variance to obtain

$$\begin{aligned} \text{var}(b_2^*|\mathbf{x}) &= \text{var}\left\{ \left[\beta_2 + \sum (w_i + c_i) e_i \right] |\mathbf{x} \right\} = \sum (w_i + c_i)^2 \text{var}(e_i|\mathbf{x}) \\ &= \sigma^2 \sum (w_i + c_i)^2 = \sigma^2 \sum w_i^2 + \sigma^2 \sum c_i^2 \\ &= \text{var}(b_2|\mathbf{x}) + \sigma^2 \sum c_i^2 \\ &\geq \text{var}(b_2|\mathbf{x}) \end{aligned}$$

The last line follows since $\sum c_i^2 \geq 0$ and establishes that for the family of linear and unbiased estimators b_2^*, each of the alternative estimators has variance that is greater than or equal to that of the least squares estimator b_2. The *only* time that $\text{var}(b_2^*) = \text{var}(b_2)$ is when all the $c_i = 0$, in which case $b_2^* = b_2$. Thus, there is no *other linear and unbiased estimator of* β_2 that is better than b_2, which proves the Gauss–Markov theorem.

Appendix 2G Proofs of Results Introduced in Section 2.10

2G.1 The Implications of Strict Exogeneity

First, if x is strictly exogenous, then the unconditional expected value of the error term e_i is zero. To show this, we use the law of iterated expectations

$$E(e_i) = E_{x_j}\left[E(e_i|x_j)\right] = E_{x_j}(0) = 0$$

Second, the covariance between X and Y can be calculated as $\text{cov}(X, Y) = E_X\left[(X - \mu_x)\,E(Y|X)\right]$, as discussed in Probability Primer Section P.6.5. Using this result, we obtain

$$\text{cov}(x_j, e_i) = E_{x_j}\left\{\left[x_j - E(x_j)\right]E(e_i|x_j)\right\} = E_{x_j}\left\{\left[x_j - E(x_j)\right]0\right\} = 0$$

If x is strictly exogenous, then the covariance between x_j and e_i is zero for all values of i and j. Recall that zero covariance means "no linear association" but not statistical independence. Thus, strict exogeneity rules out any covariance, any linear association, between any x_j and any e_i. The covariance between x_j and e_i can be rewritten as a simpler expectation using the facts that $E(e_i) = 0$ and $E(x_j)$ is not random

$$\begin{aligned}\text{cov}(x_j, e_i) &= E\left\{\left[x_j - E(x_j)\right]\left[e_i - E(e_i)\right]\right\} = E\left\{\left[x_j - E(x_j)\right]e_i\right\} = E(x_j e_i) - E\left[E(x_j)\,e_i\right] \\ &= E(x_j e_i) - E(x_j)\,E(e_i) = E(x_j e_i)\end{aligned}$$

Strict exogeneity implies $E(x_j e_i) = 0$ for all x_j and e_i.

Using the covariance decomposition we can show yet more. Let $g(x_j)$ be a function of x_j. Then

$$\begin{aligned}\text{cov}\left[g(x_j), e_i\right] &= E_{x_j}\left\{\left[g(x_j) - E\left(g(x_j)\right)\right]E(e_i|x_j)\right\} = E_{x_j}\left\{\left[g(x_j) - E\left(g(x_j)\right)\right]0\right\} = 0 \\ &= E\left[g(x_j)e_i\right]\end{aligned}$$

If x is strictly exogenous, then the covariance between a function of x_j [like x_j^2 or $\ln(x_j)$] and e_i is zero for all values of i and j. Thus, strict exogeneity rules out any covariance, any linear association, between a function of x_j and any e_i.

2G.2 The Random and Independent *x* Case

In Section 2.10.1 we considered the case in which x-values are random but statistically independent of the random error e. In this appendix, we show the algebra behind our conclusions. Consider b_2 the least squares estimator of the slope parameter β_2. b_2 is a linear estimator and as shown in (2.10) $b_2 = \sum_{i=1}^N w_i y_i$, where $w_i = (x_i - \bar{x})/\sum_{i=1}^N (x_i - \bar{x})^2$. Notice that $w_i = g(x_1, \ldots, x_N)$ is a function of all the random x_i values and it is random. For notational ease, let $\mathbf{x}$ represent $x_1, \ldots, x_N$ so $w_i = g(x_1, \ldots, x_N) = g(\mathbf{x})$. Because IRX5 makes clear that x_i is random and is statistically independent of the random error e_i for all values of i and j, then $w_i = g(\mathbf{x})$ is statistically independent of each random error e_i. Substituting $y_i = \beta_1 + \beta_2 x_i + e_i$, we obtain $b_2 = \beta_2 + \sum w_i e_i$ and, using the fact $E(w_i e_i) = E(w_i)E(e_i)$ because of independence, we have

$$E(b_2) = \beta_2 + \sum E(w_i e_i) = \beta_2 + \sum E(w_i)E(e_i) = \beta_2 + \sum E(w_i)\,0 = \beta_2$$

In the case in which x is random but statistically independent of the error terms, the least squares estimator is unconditionally unbiased.

The derivation of the variance of the least squares estimator changes in a similar way:

$$\begin{aligned}\operatorname{var}(b_2) &= E\left[(b_2-\beta_2)^2\right] = E\left[(\beta_2+\sum w_i e_i-\beta_2)^2\right] = E\left[\left(\sum w_i e_i\right)^2\right] \\ &= E\left(\sum w_i^2 e_i^2 + \sum\sum_{i\neq j} w_i w_j e_i e_j\right) \\ &= \sum E(w_i^2)E(e_i^2) + \sum\sum_{i\neq j} E(w_i w_j)\,E(e_i e_j) \\ &= \sigma^2\sum E(w_i^2) = \sigma^2 E\left(\sum w_i^2\right) = \sigma^2 E\left[\frac{1}{\sum(x_i-\bar{x})^2}\right]\end{aligned}$$

In the third line we used the statistical independence of w_i and each random error e_i twice. In the fourth line we used the fact that the expected value of a sum is the sum of the expected values, and finally that $\sum w_i^2$ is known, as shown in Appendix 2E.

The usual estimator of the error variance is $\hat{\sigma}^2 = \sum \hat{e}_i^2/(N-2)$ and conditional on $\mathbf{x}$ this estimator is unbiased, $E\left(\hat{\sigma}^2|\mathbf{x}\right) = \sigma^2$. The proof is messy and not shown. This is a conditional expectation saying *given* $x_1, \ldots, x_N$ the estimator $\hat{\sigma}^2$ is unbiased. Now we use the law of iterated expectations from the Probability Primer Section P.6.3:

$$E\left(\hat{\sigma}^2\right) = E_{\mathbf{x}}\left[E\left(\hat{\sigma}^2|\mathbf{x}\right)\right] = E_{\mathbf{x}}\left[\sigma^2\right] = \sigma^2$$

where $E_{\mathbf{x}}(\,)$ means the expected value treating $\mathbf{x}$ as random. Because the conditional expectation $E\left(\hat{\sigma}^2|\mathbf{x}\right) = \sigma^2$ is a constant that does not depend on $\mathbf{x}$, its expectation treating $\mathbf{x}$ as random is also a constant, σ^2. So, in the case in which x is random and independent of the error, $\hat{\sigma}^2$ is conditionally *and* unconditionally unbiased.

The variance of the least squares estimator is

$$\operatorname{var}(b_2) = \sigma^2 E_{\mathbf{x}}\left[\frac{1}{\sum(x_i-\bar{x})^2}\right]$$

The usual variance estimator from (2.21) is

$$\widehat{\operatorname{var}}(b_2|\mathbf{x}) = \hat{\sigma}^2\frac{1}{\sum(x_i-\bar{x})^2}$$

It is an unbiased estimator of $\operatorname{var}(b_2)$ conditional on $\mathbf{x}$. Using the law of iterated expectations, we have

$$E_{\mathbf{x}}\left\{E\left[\widehat{\operatorname{var}}(b_2|\mathbf{x})\right]\right\} = E_{\mathbf{x}}\left\{\sigma^2\frac{1}{\sum(x_i-\bar{x})^2}\,\middle|\,\mathbf{x}\right\} = \sigma^2 E_{\mathbf{x}}\left[\frac{1}{\sum(x_i-\bar{x})^2}\right] = \operatorname{var}(b_2)$$

Thus, the usual estimator of $\operatorname{var}(b_2)$ is unbiased.

What about the Gauss–Markov theorem? It says, for fixed x, or *given* $\mathbf{x}$, $\operatorname{var}(b_2|\mathbf{x})$, is less than the variance $\operatorname{var}(b_2^*|\mathbf{x})$ of any other linear and unbiased estimator b_2^*. That is,

$$\operatorname{var}(b_2|\mathbf{x}) < \operatorname{var}(b_2^*|\mathbf{x})$$

Using the variance decomposition $\operatorname{var}(b_2) = \operatorname{var}_{\mathbf{x}}\left[E(b_2|\mathbf{x})\right] + E_{\mathbf{x}}\left[\operatorname{var}(b_2|\mathbf{x})\right] = E_{\mathbf{x}}\left[\operatorname{var}(b_2|\mathbf{x})\right]$ because $\operatorname{var}_{\mathbf{x}}\left[E(b_2|\mathbf{x})\right] = \operatorname{var}_{\mathbf{x}}(\beta_2) = 0$. Similarly, $\operatorname{var}(b_2^*) = E_{\mathbf{x}}\left[\operatorname{var}(b_2^*|\mathbf{x})\right]$. Then

$$\operatorname{var}(b_2) = E_{\mathbf{x}}\left[\operatorname{var}(b_2|\mathbf{x})\right] < \operatorname{var}(b_2^*) = E_{\mathbf{x}}\left[\operatorname{var}(b_2^*|\mathbf{x})\right]$$

The logic of the argument is that if $\operatorname{var}(b_2|\mathbf{x})$ is less than the variance of any other estimator $\operatorname{var}(b_2^*|\mathbf{x})$ for any given $\mathbf{x}$, it must also be true for all $\mathbf{x}$, and will remain true if we average over all possible $\mathbf{x}$, by taking the expected value treating $\mathbf{x}$ as random, $E_{\mathbf{x}}(\,)$.

Finally, what about normality? If IRX6 holds, $e_i \sim N(0, \sigma^2)$, then what is the probability distribution of the least squares estimator? We have used the fact that $b_2 = \beta_2 + \sum w_i e_i$. If w_i is constant, then we can assert that the least squares estimator has a normal distribution because linear combinations of normal random variables are normal. However, in the random-x case, even though x is independent of e, the distributions of $w_i e_i$ are not normal. The function $w_i = g(x_1, \ldots, x_N)$ has an unknown probability distribution and its product with the normally distributed e_i results in an unknown distribution. What we can say is that $b_2|\mathbf{x}$ is normal, since conditioning on $x_1, \ldots, x_N$ means that they are treated as given, or fixed.

2G.3 The Random and Strictly Exogenous x Case

In Section 2.10.2 we examine the consequences of an assumption that is weaker than the statistical independence of x and e. There we assert that even with the weaker assumption called "strict exogeneity" the properties of the least squares estimator are unchanged, and here we give the proof. The least squares estimator of the slope parameter, b_2, is a linear estimator and as shown in (2.10) $b_2 = \sum_{i=1}^{N} w_i y_i$, where $w_i = (x_i - \bar{x}) / \sum_{i=1}^{N} (x_i - \bar{x})^2$. Notice that $w_i = g(x_1, \ldots, x_N)$ is a function of all the random x_i values and it is random. Substituting $y_i = \beta_1 + \beta_2 x_i + e_i$, we obtain $b_2 = \beta_2 + \sum w_i e_i$. The strict exogeneity assumption says $E(e_i|x_j) = 0$ for all values of i and j, or equivalently, $E(e_i|\mathbf{x}) = 0$. Using the law of iterated expectations, we show that b_2 is a conditionally unbiased estimator. First, find the conditional expectation of b_2 given $\mathbf{x}$,

$$E(b_2|\mathbf{x}) = \beta_2 + \sum E(w_i e_i|\mathbf{x}) = \beta_2 + \sum w_i E(e_i|\mathbf{x}) = \beta_2 + \sum w_i 0 = \beta_2$$

Conditional on $\mathbf{x}$, which is equivalent to assuming $\mathbf{x}$ is given, the function $w_i = g(x_1, \ldots, x_N)$ is treated like a constant and is factored out in the third equality. Applying the law of iterated expectations, we find

$$E(b_2) = E_{\mathbf{x}}\left[E(b_2|\mathbf{x})\right] = E_{\mathbf{x}}(\beta_2) = \beta_2$$

The notation $E_{\mathbf{x}}(\ )$ means take the expected value treating $\mathbf{x}$ as random. In this case, that is not difficult because β_2 is a constant, nonrandom parameter. The least squares estimator is unbiased, both conditional on $\mathbf{x}$ and unconditionally, under strict exogeneity.

The derivation of the variance of the least squares estimator changes in a similar way. First find the variance of b_2 given $\mathbf{x}$.

$$\begin{aligned}
\text{var}(b_2|\mathbf{x}) &= E\left[\left(b_2 - E(b_2|\mathbf{x})\right)^2 \middle| \mathbf{x}\right] = E\left[(\beta_2 + \sum w_i e_i - \beta_2)^2 \middle| \mathbf{x}\right] = E\left[\left(\sum w_i e_i\right)^2 \middle| \mathbf{x}\right] \\
&= E\left[\left(\sum w_i^2 e_i^2 + \sum\sum_{i \neq j} w_i w_j e_i e_j\right) \middle| \mathbf{x}\right] = \sum w_i^2 E(e_i^2|\mathbf{x}) + \sum\sum_{i \neq j} w_i w_j E(e_i e_j|\mathbf{x}) \\
&= \sigma^2 \sum w_i^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

The variance of b_2 given $\mathbf{x}$ is exactly the same as when $\mathbf{x}$ was assumed random and statistically independent of the random errors. Now find the variance of b_2 using the variance decomposition from the Probability Primer equation (P.29). For two random variables X and Y,

$$\text{var}(Y) = \text{var}_X[E(Y|X)] + E_X[\text{var}(Y|X)]$$

Letting $Y = b_2$ and $X = \mathbf{x}$, we have

$$\text{var}(b_2) = \text{var}_{\mathbf{x}}\left[E(b_2|\mathbf{x})\right] + E_{\mathbf{x}}\left[\text{var}(b_2|\mathbf{x})\right] = \text{var}_{\mathbf{x}}(\beta_2) + E_{\mathbf{x}}\left[\frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right] = \sigma^2 E_{\mathbf{x}}\left[\frac{1}{\sum (x_i - \bar{x})^2}\right]$$

since $\text{var}_{\mathbf{x}}(\beta_2) = 0$. This is exactly the same result as in the case in which x_j and e_i are statistically independent.

2G.4 Random Sampling

In the case of random sampling, data pairs (y_i, x_i) are *iid*, and the strict exogeneity assumption reduces to $E(e_i|x_i) = 0$. The results in the previous section hold in exactly the same way because it is still true that $E(e_i|\mathbf{x}) = 0$.

Appendix 2H Monte Carlo Simulation

The statistical properties of the least squares estimators are well known if the assumptions in Section 2.1 hold. In fact, we know that the least squares estimators are the best linear unbiased estimators of the regression parameters under these assumptions. And if the random errors are normal, then we know that, given **x**, the estimators themselves have normal distributions in **repeated experimental trials**. The meaning of "repeated trials" is difficult to grasp. **Monte Carlo** simulation experiments use random number generators to replicate the random way that data are obtained. In Monte Carlo simulations, we specify a **data generation process** and create samples of artificial data. Then, we "try out" estimation methods on the data we have created. We create **many** samples of size N and examine the **repeated sampling properties** of the estimators. In this way, we can study how statistical procedures behave under ideal, as well as not so ideal, conditions. This is important because economic, business, and social science data are not always (indeed, not usually) as nice as the assumptions we make.

The DGP for the simple linear regression model is given by

$$y_i = E(y_i|x_i) + e_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

Each value of the dependent variable y_i is obtained, or generated, by adding a random error e_i to the regression function $E(y_i|x_i)$. To simulate values of y_i, we create values for the systematic portion of the regression relationship $E(y_i|x_i)$ and add to it the random error e_i. This is analogous to a physical experiment in which variable factors are set at fixed levels and the experiment run. The outcome is different in each experimental trial because of random uncontrolled errors.

2H.1 The Regression Function

The regression function $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ is the systematic portion of the regression relationship. To create these values we must select the following:

1. *A sample size N.* From the discussion in Section 2.4.4, we know that the larger the sample size is, the greater is the precision of estimation of the least squares estimators b_1 and b_2. Following the numerical examples in the book, we choose $N = 40$. This is not a large sample, but assuming SR1–SR5 are true, the least squares estimators' properties hold for any sample of size $N > 2$ in the simple regression model. In more complex situations, varying the sample size to see how estimators perform is an important ingredient of the simulation.
2. *We must choose x_i values.* For simplicity, we initially assume values of the explanatory variable that are fixed in repeated experimental trials. Following the depiction in Figure 2.1,[15] we set the values $x_1, x_2, \ldots, x_{20} = 10$ and $x_{21}, x_{22}, \ldots, x_{40} = 20$, using the chapter assumption that x is measured in hundreds. Does it matter how we choose the x_i values? Yes, it does. The variances and covariances of the least squares estimators depend on the variation in x_i, $\sum(x_i - \bar{x})^2$, how far the values are from 0, as measured by $\sum x_i^2$, and on the sample mean $\bar{x}$. Thus, if the values x_i change, the precision of estimation of the least squares estimators will change.

[15] This design is used in Briand, G. & Hill, R. C. (2013). Teaching Basic Econometric Concepts using Monte Carlo Simulations in Excel, *International Review of Economics Education*, 12(1), 60–79.

3. *We must choose* β_1 *and* β_2. Interestingly, for the least squares estimator under assumptions SR1–SR5, the actual magnitudes of these parameters do not matter a great deal. The estimator variances and covariances do not depend on them. The difference between the least squares estimator and the true parameter value, $E(b_2) - \beta_2$ given in (2.13), does not depend on the magnitude of β_2, only on the x_i values and the random errors e_i. To roughly parallel the regression results we obtained in Figure 2.10, we set $\beta_1 = 100$ and $\beta_2 = 10$.

Given the values above we can create $N = 40$ values $E(y_i|x_i) = \beta_1 + \beta_2 x_i$. These values are

$$E(y_i|x_i = 10) = 100 + 10x_i = 100 + 10 \times 10 = 200, \qquad i = 1, \ldots, 20$$
$$E(y_i|x_i = 20) = 100 + 10x_i = 100 + 10 \times 20 = 300, \qquad i = 21, \ldots, 40$$

2H.2 The Random Error

To be consistent with assumptions SR2–SR4, the random errors should have mean zero, constant variance $\text{var}(e_i|x_i) = \sigma^2$ and be uncorrelated with one another, so that $\text{cov}(e_i, e_j|\mathbf{x}) = 0$. Researchers in the field of numerical analysis have studied how to simulate random numbers from a variety of probability distributions, such as the normal distribution. Of course, the computer-generated numbers cannot be truly random, because they are generated by a computer code. The random numbers created by computer software are "pseudorandom," in that they behave like random numbers. The numbers created will begin to recycle after about 2^{19937} values are drawn, using the so-called Mersenne Twister algorithm. Each software vendor uses its own version of a random number generator. Consequently, you should not expect to obtain exactly the same numbers that we have, and your replication will produce slightly different results, even though the major conclusions will be the same. See Appendix B.4 for a discussion of how random numbers are created.

Following assumption SR6, we assume the random error terms have a normal distribution with mean zero and a homoskedastic variance $\text{var}(e_i|x_i) = \sigma^2$. The variance σ^2 affects the precision of estimation through the variances and covariances of the least squares estimators in (2.14)–(2.16). The bigger the value of σ^2, the bigger the variances and covariances of the least squares estimators, and the more spread out the probability distribution of the estimators, as shown in Figure 2.11. We choose $\text{var}(e_i|x_i) = \sigma^2 = 2500$, which also means that $\text{var}(y_i|x_i) = \sigma^2 = 2500$.

2H.3 Theoretically True Values

Using the values above, we plot the theoretically true *pdfs* for y_i in Figure 2H.1. The solid curve on the left is $N(200, 2500 = 50^2)$. The first 20 simulated observations will follow this *pdf*. The dashed curve on the right is $N(300, 2500 = 50^2)$, which is the *pdf* for the second 20 observations.

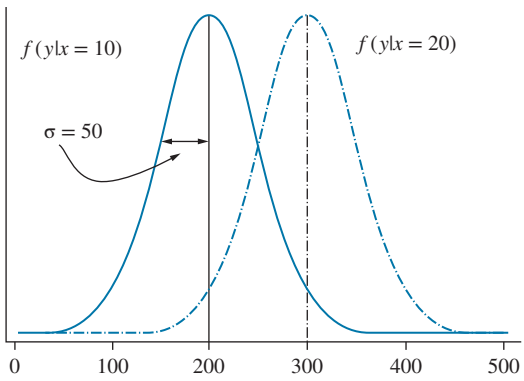


FIGURE 2H.1 **The true *pdfs* of the data.**

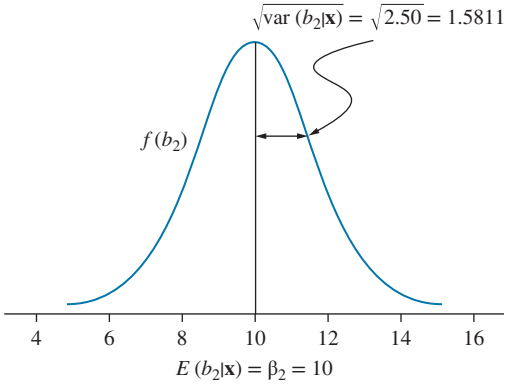


FIGURE 2H.2 **The true pdf of the estimator b_2.**

Given the parameter $\sigma^2 = 2500$ and the x_i values, we can compute the true conditional variances of the estimators:

$$\text{var}(b_1|\mathbf{x}) = \sigma^2\left[\frac{\sum x_i^2}{N\sum(x_i-\bar{x})^2}\right] = 2500\left[\frac{10000}{40\times 1000}\right] = 625$$

$$\text{var}(b_2|\mathbf{x}) = \frac{\sigma^2}{\sum(x_i-\bar{x})^2} = \frac{2500}{1000} = 2.50$$

$$\text{cov}(b_1, b_2|\mathbf{x}) = \sigma^2\left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right] = 2500\left[\frac{-15}{1000}\right] = -37.50$$

The true standard deviation of b_2 is $\sqrt{\text{var}(b_2|\mathbf{x})} = \sqrt{2.50} = 1.5811$. The true *pdf* of $b_2|\mathbf{x}$ is $N(\beta_2 = 10, \text{var}(b_2|\mathbf{x}) = 2.5)$. Using the cumulative probabilities for the standard normal distribution in Statistical Table 1, we find that 98% of values from a normal distribution fall within 2.33 standard deviations of the mean. Applying this rule to the estimates b_2, we have

$$\beta_2 \pm 2.33 \times \sqrt{\text{var}(b_2|\mathbf{x})} = 10 \pm 2.33 \times 1.5811 = [6.316, 13.684]$$

We expect almost all values of b_2 (98% of them) to fall in the range 6.32–13.68. The plot of the true *pdf* of the estimator b_2 is shown in Figure 2H.2.

2H.4 Creating a Sample of Data

Most software will automatically create random values, z_i, from the standard normal distribution, $N(0, 1)$. To obtain a random value from a $N(0, \sigma^2)$ distribution, we multiply z_i by the standard deviation σ. That is, $e_i = \sigma \times z_i$. Given values z_i from the standard normal distribution, we obtain the $N = 40$ sample values from the chosen DGP as

$$y_i = E(y_i|x_i = 10) + e_i = 200 + 50 \times z_i \qquad i = 1, \ldots, 20$$
$$y_i = E(y_i|x_i = 20) + e_i = 300 + 50 \times z_i \qquad i = 21, \ldots, 40$$

One sample of data is in the data file *mc1_fixed_x*. Using these values, we obtain the least squares estimates. It is convenient to display the coefficient estimates and standard errors together, with the standard error reported below the coefficients:

$$\hat{y} = 127.2055 + 8.7325x$$
$$\text{(se)} \quad (23.3262) \quad (1.4753)$$

The estimate $\hat{\sigma} = 46.6525$. The estimated variances and covariance of b_1 and b_2 are $\widehat{\text{var}}(b_1) = 544.1133$, $\widehat{\text{var}}(b_2) = 2.1765$, and $\widehat{\text{cov}}(b_1, b_2) = -32.6468$.

For this one sample, the parameter estimates are reasonably near their true values. However, what happens in one sample does not prove anything. The repeated sampling properties of the least squares estimators are about what happens in many samples of data, from the same DGP.

2H.5 Monte Carlo Objectives

What do we hope to achieve with a Monte Carlo experiment? After the Monte Carlo experiment, we will have many least squares estimates. If we obtain $M = 10,000$ samples, we will have 10,000 estimates $b_{1,1}, \ldots, b_{1,M}$, 10,000 estimates $b_{2,1}, \ldots, b_{2,M}$, and 10,000 estimates $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2$.

- We would like to verify that under SR1–SR5 the least squares estimators are unbiased. The estimator b_2 is unbiased if $E(b_2) = \beta_2$. Since an expected value is an average in many repeated experimental trials, we should observe that the average value of all the slope estimates, $\bar{b}_2 = \sum_{m=1}^{M} b_{2m}/M$, is close to $\beta_2 = 10$.
- We would like to verify that under SR1–SR5 the least squares estimators have sampling variances given by (2.14) and (2.16). The estimator variances measure the sampling variation in the estimates. The sampling variation of the estimates in the Monte Carlo simulation can be measured by their sample variance. For example, the sample variance of the estimates $b_{2,1}, \ldots, b_{2,M}$ is $s_{b_2}^2 = \sum_{m=1}^{M}\left(b_{2,m} - \bar{b}_2\right)^2/(M-1)$. This value should be close to $\text{var}(b_2) = 2.50$, and the standard deviation s_{b_2} should be close to the true standard deviation of the regression estimates 1.5811.
- We would like to verify that the estimator of the error variance (2.19) is an unbiased estimator of $\sigma^2 = 2500$, or that $\hat{\sigma}^2 = \sum_{m=1}^{M} \hat{\sigma}_m^2/M$ is close to the true value.
- Because we have assumed the random errors are normal, SR6, we expect the least squares estimates to have a normal distribution.

2H.6 Monte Carlo Results

The numerical results of the Monte Carlo experiment are shown in Table 2H.1. The averages (or "Sample Means") of the 10,000 Monte Carlo estimates are close to their true values.

For example, the average of the slope estimates is $\bar{b}_2 = \sum_{m=1}^{M} b_{2,m}/M = 10.0130$ compared to the true value $\beta_2 = 10$. The sample variance of the estimates $s_{b_2}^2 = \sum_{m=1}^{M}\left(b_{2,m} - \bar{b}_2\right)^2/(M-1) = 2.4691$ compared to the true value $\text{var}(b_2) = 2.50$. The standard deviation of the estimates is $s_{b_2} = 1.5713$ compared to the true standard deviation $\sqrt{\text{var}(b_2)} = \sqrt{2.50} = 1.5811$. The theoretical 1st and 99th percentiles of b_2 are [6.316, 13.684], which is reflected by the estimates [6.3268, 13.6576].

As for the normality of the estimates, we see from the histogram in Figure 2H.3 that the actual values follow the superimposed normal distribution very closely.[16]

TABLE 2H.1 **Summary of 10,000 Monte Carlo Samples**

	Mean	Variance	Std. Dev.	Minimum	Maximum	1st Pct.	99th Pct.
b_1 (100)	99.7463	613.4323	24.7676	12.1000	185.5361	42.2239	156.5996
b_2 (10)	10.0130	2.4691	1.5713	4.5881	16.5293	6.3268	13.6576
$\hat{\sigma}^2$ (2500)	2490.67	329964.7	574.4256	976.447	5078.383	1366.225	4035.681

[16] A normal distribution is symmetrical with no skewness, and for the estimates b_2 the skewness is −0.0027. A normal distribution has kurtosis of three, and for the estimates b_2 the kurtosis is 3.02. The Jarque–Bera test statistic that combines skewness and kurtosis measures is 0.1848 yielding a p-value of 0.91, meaning that we fail to reject the normality. See Appendix C.7.4 for a discussion of the Jarque–Bera test.

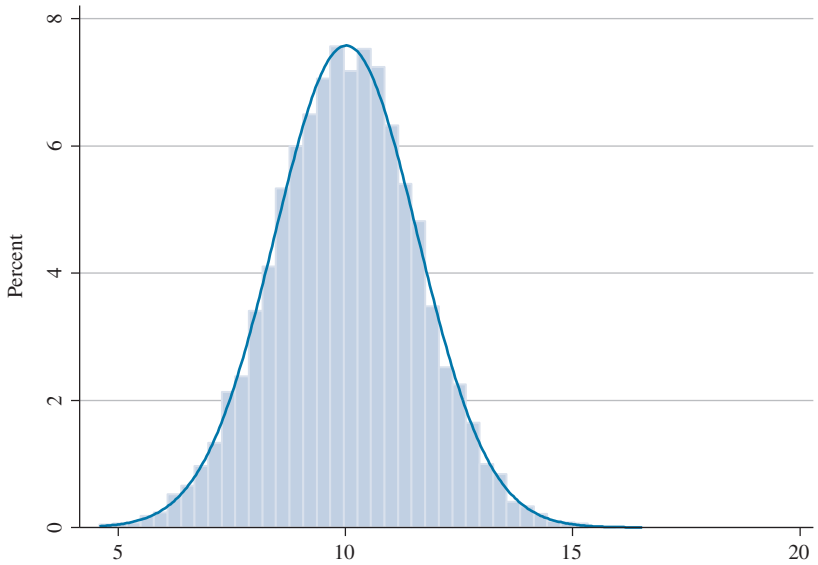


FIGURE 2H.3 **The sampling distribution of b_2 in 10,000 Monte Carlo samples when x is fixed in repeated trials.**

If you are replicating these results, some suggested exercises are as follows:

1. Test if mean of $\bar{b}_2$ is equal to β_2 using the test described in Appendix C.6.1.
2. Calculate the percentage of estimates falling in a given interval, such as between 8 and 9, and compare it to the probability based on the normal distribution.

2H.7 Random-x Monte Carlo Results

We used the "fixed-x" framework in the simulation results above. In each Monte Carlo sample, the x-values were $x_i = 10$ for the first 20 observations and $x_i = 20$ for the next 20 observations. Now we modify the experiment to the random-x case. The data generating equation remains $y_i = 100 + 10x_i + e_i$ with the random errors having a normal distribution with mean zero and standard deviation 50, $e_i \sim N(0, 50^2 = 2500)$. We randomly choose x-values from a normal distribution with mean $\mu_x = 15$ and standard deviation $\sigma_x = 1.6$, so $x \sim N(15, 1.6^2 = 2.56)$. We chose $\sigma_x = 1.6$ so that 99.73% of the **random-*x*** values fall between 10.2 and 19.8, which is similar in spirit to the fixed-x simulation in the previous section.

One sample of data is in the file *mc1_random_x*. Using these values, we obtain the least squares estimates and standard errors

$$\hat{y} = 116.7410 + 9.7628x$$
$$\text{(se)} \quad (84.7107) \quad (5.5248)$$

and the estimate $\hat{\sigma} = 51.3349$. The estimates are close to the true values.

The numerical results of the Monte Carlo experiment are shown in Table 2H.2. The averages (or "Sample Means") of the 10,000 Monte Carlo estimates are close to their true values.

For example, the average of the slope estimates is $\bar{b}_2 = \sum_{m=1}^{M} b_{2,m}/M = 10.0313$ compared to the true value $\beta_2 = 10$. In the random-x case, the true variance of the least squares estimator is

$$\text{var}(b_2) = \sigma^2 E\left[\frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right] = \frac{\sigma^2}{(N-3)\sigma_x^2} = \frac{2500}{(37)(2.56)} = 26.3936$$

Calculating the variance we use a special property resulting from the normality of x. When x is normally distributed $N(\mu_x, \sigma_x^2)$ the unbiased estimator of σ_x^2 is $s_x^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2/(N-1)$.

TABLE 2H.2 Summary of 10,000 Random-x Monte Carlo Samples

	Mean	Var.	Std. Dev.	Min.	Max.	1st Pct.	99th Pct.
b_1 (100)	99.4344	6091.4412	78.0477	−196.8826	405.8328	−83.1178	283.8266
b_2 (10)	10.0313	26.8503	5.1817	−10.4358	29.3168	−2.2196	22.3479
$\widehat{\text{var}}(b_2)$ (26.3936)	26.5223	78.9348	8.8845	7.8710	91.1388	11.8325	54.0177
$\hat{\sigma}^2$ (2500)	2498.4332	332622.6	576.7344	809.474	5028.047	1366.957	4056.279

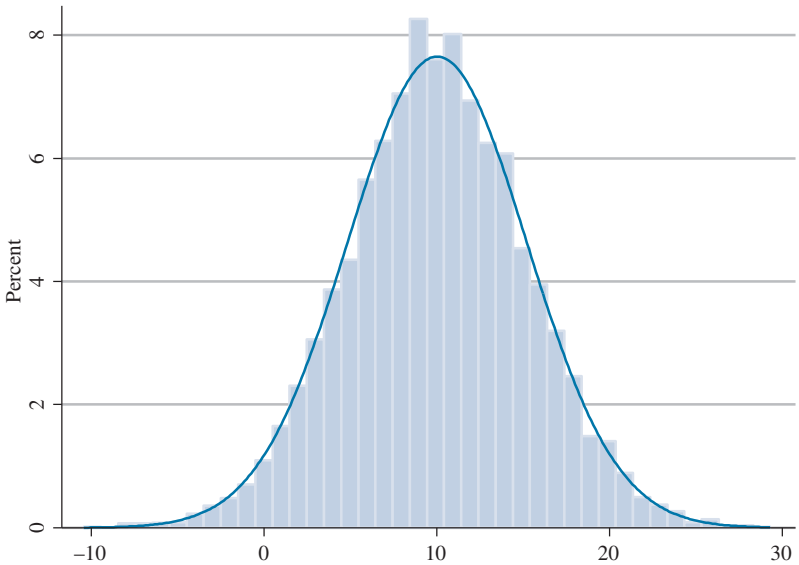


FIGURE 2H.4 The sampling distribution of b_2 in 10,000 Monte Carlo samples when x is random in repeated trials.

In Appendix C.7.1 we use the fact that $(N-1)s_x^2/\sigma_x^2 \sim \chi^2_{(N-1)}$. This implies that $V = \sum_{i=1}^{N}(x_i - \bar{x})^2 \sim \sigma_x^2 \chi^2_{(N-1)}$. Using the properties of the inverse chi-square distribution $E(1/V) = E\left[1/\sum_{i=1}^{N}(x_i - \bar{x})^2\right] = 1/\left[(N-3)\sigma_x^2\right]$.[17] Note that the Monte Carlo mean of the estimated $\text{var}(b_2)$ is 26.5223, confirming that $\widehat{\text{var}}(b_2) = 2500/[37(2.56)] = 26.3936$ is an unbiased estimator even in the random-x case.

Recall, however, that in the random-x case the distribution of the least squares estimator b_2 is not normal. The histogram of the 10,000 Monte Carlo estimates is shown in Figure 2H.4. It is symmetrical but there are too many central values, and the peak is too high. Statistically we can reject that this distribution is normal.[18]

If you are replicating these results, some suggested exercises are as follows:

1. Test if mean of $\bar{b}_2$ is equal to β_2 using the test described in Appendix C.6.1.
2. Calculate the percentage of estimates falling in a given interval, such as between 8 and 9, and compare it with the probability based on the normal distribution.

[17] See Appendix B.3.6 and Appendix C.7.1 for the theory behind this result.

[18] A normal distribution is symmetrical with no skewness, and for the estimates b_2 the skewness is −0.001. A normal distribution has kurtosis of three, and for the estimates b_2 the kurtosis is 3.14. The Jarque–Bera test statistic that combines skewness and kurtosis measures is 8.32 yielding a p-value of 0.016, meaning that we reject the hypothesis of normality at the 5% level of significance. See Appendix C.7.4 for a discussion of the Jarque–Bera test.