Chapter 2
The Simple Linear Regression Model: Specification and Estimation

Walter R. Paczkowski
Rutgers University

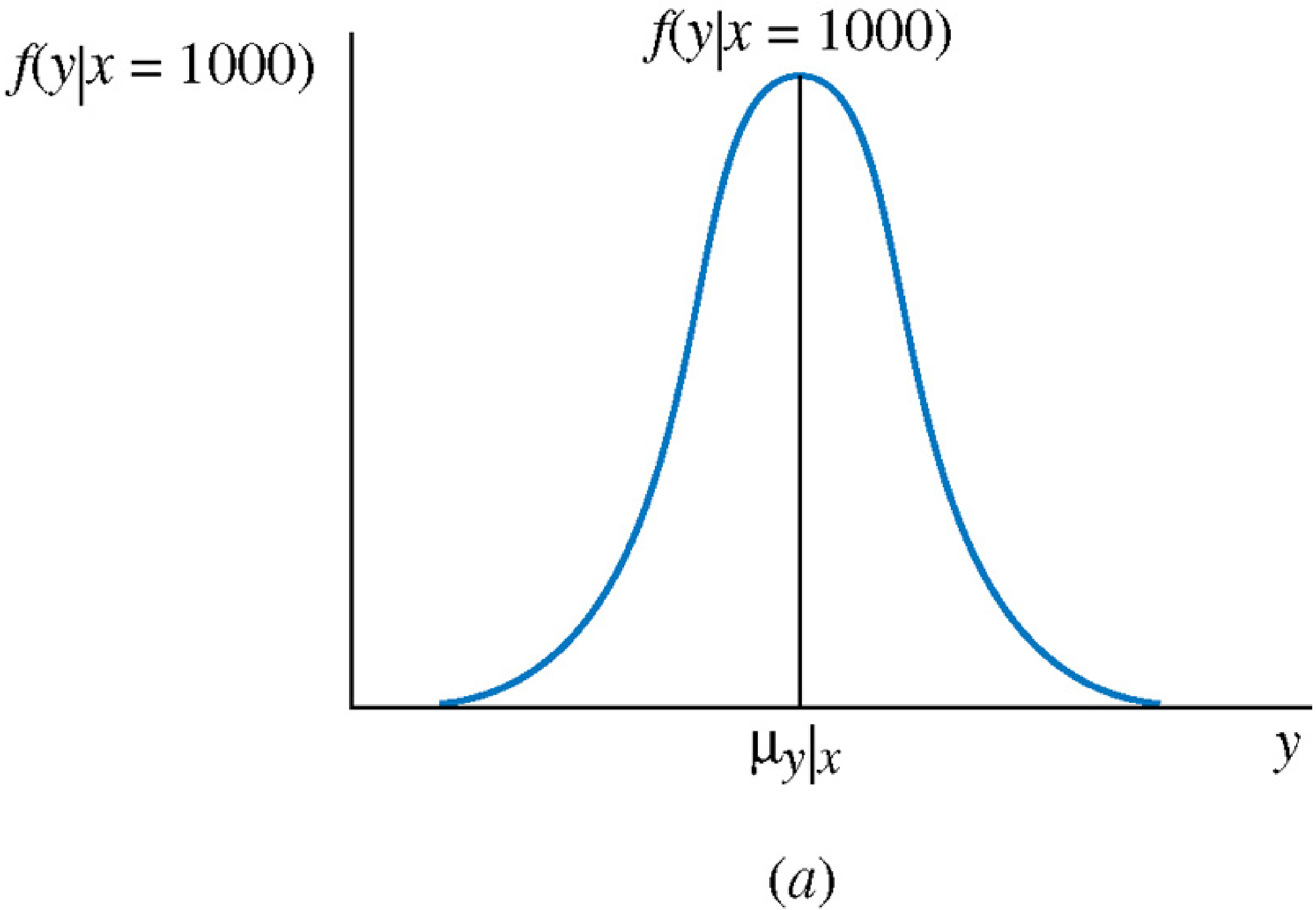
Chapter Contents

2.1
An Economic Model

- We are interested in studying the relationship between household income and expenditure of food per person.
  - Suppose that we are interested only in households with income = $1,000 per week.
  - We randomly select a sample of such households and we measure the food expenditure.
  - Denote *y* = food expenditure, then *y* is a **random variable** because the food expenditure, even if the income is the same, would vary from one household to the other.

- The *pdf* is a conditional probability density function since it is "conditional" upon an *x*
  - The conditional mean, or expected value, of y is $E(y|x)$
    - 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

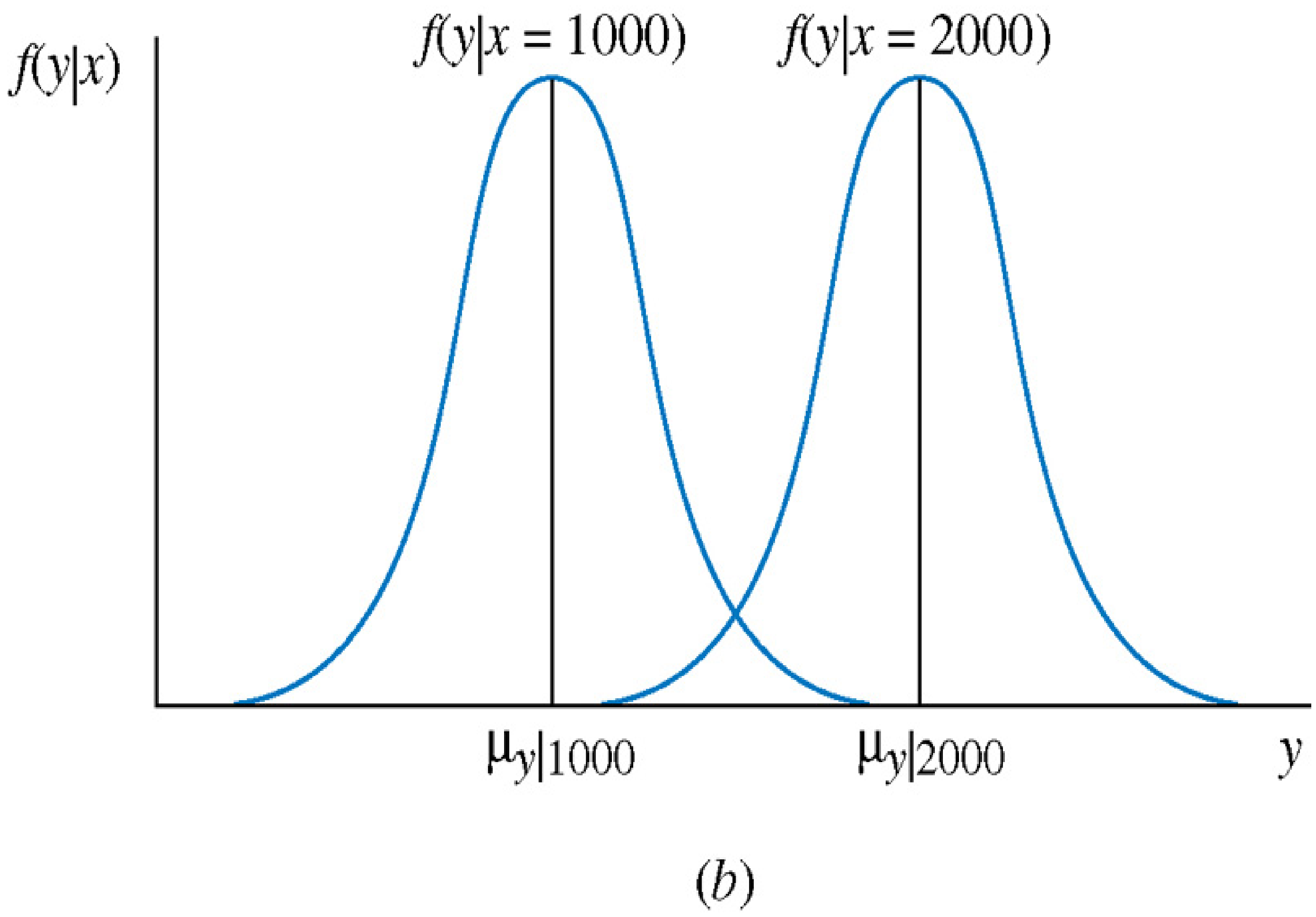
Figure 2.1a Probability distribution of food expenditure y given income x = $1000



- The conditional variance of *y* is σ^2 which measures the dispersion of *y* about its 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

- As economists we are usually more interested in studying relationships between variables
  - 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

Figure 2.1b Probability distributions of food expenditures y given incomes x = \$1000 and x = \$2000



- 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
  - This econometric model is also called a **regression model**

- The simple regression function is written as

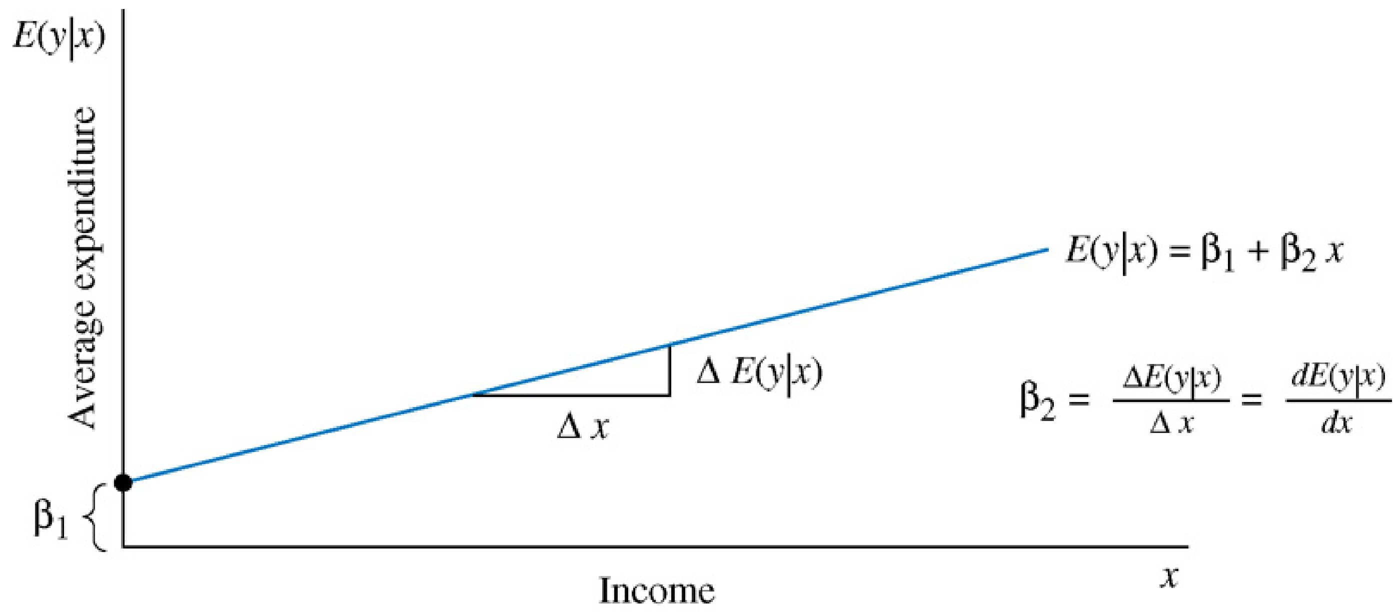
Eq. 2.1

$$E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x$$

where β_1 is the intercept and β_2 is the slope

- It is called simple regression not because it is easy, but because there is only one explanatory variable on the right-hand side of the equation

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



Eq. 2.2

- The slope of the regression line can be written as:

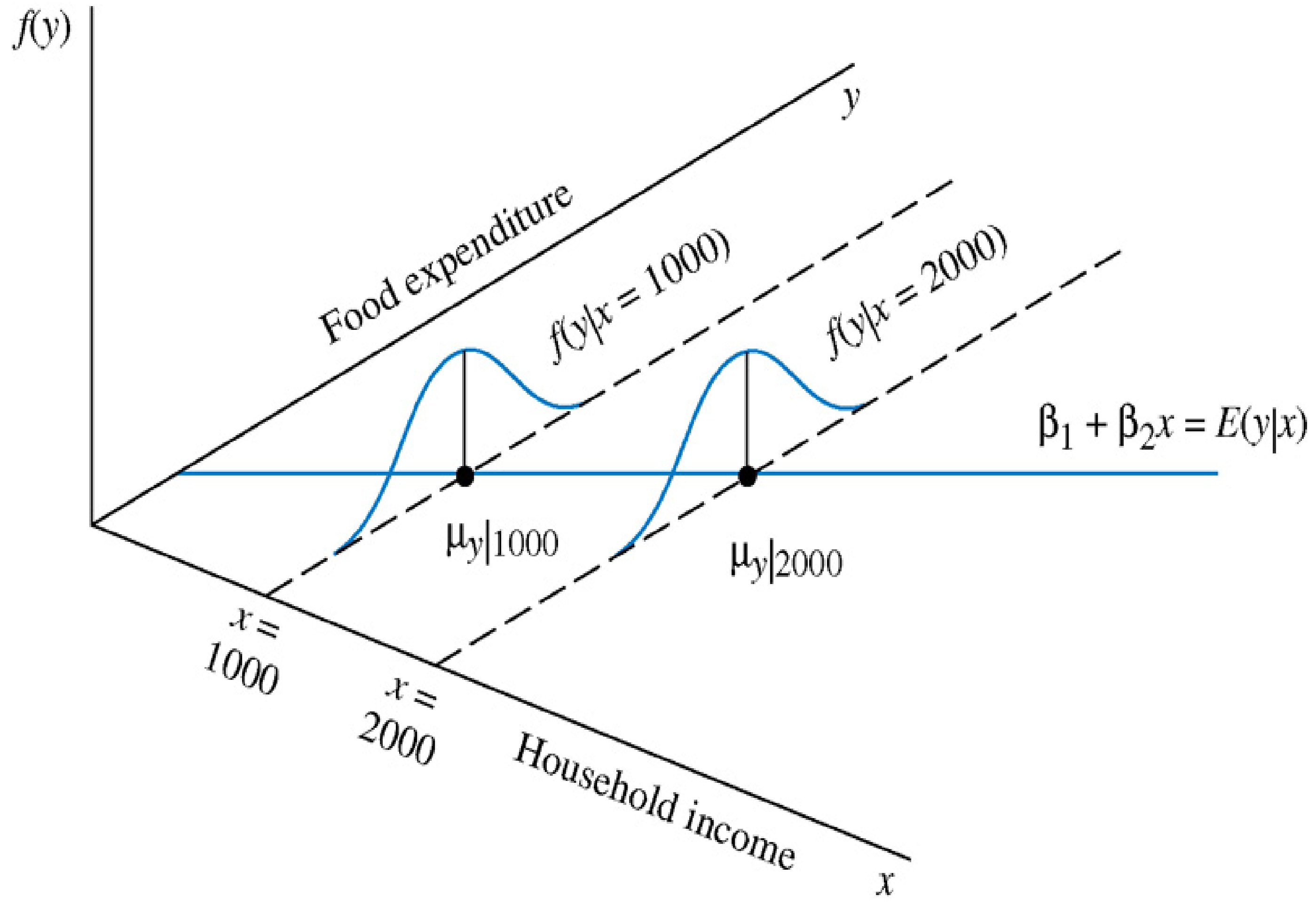
$$\beta_2 = \frac{\Delta E(y \mid x)}{\Delta x} = \frac{dE(y \mid x)}{dx}$$

where "Δ" denotes "change in" and "$dE(y|x)/dx$" denotes the derivative of the expected value of y given an x value

2.2

An Econometric Model

Figure 2.3 The probability density function for *y* at two levels of income



- There are several key assumptions underlying the simple linear regression
  - More will be added later

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - I

Assumption 1:

The mean value of *y*, for each value of *x*, is given by the *linear regression*

$$E(y \mid x) = \beta_1 + \beta_2 x$$

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - I

Assumption 2:

For each value of *x*, the values of *y* are distributed about their mean value, following probability distributions that all have the same variance

$$\text{var}(y \mid x) = \sigma^2$$

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - I

Assumption 3:

The sample values of *y* are all *uncorrelated*, and have zero *covariance*, implying that there is no linear association among them

$$\text{cov}(y_i, y_j) = 0$$

This assumption can be made stronger by assuming that the values of *y* are all statistically independent

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - I

Assumption 4:

The variable x is not random, and must take at least two different values

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - I

Assumption 5:

(*optional*) The values of *y* are *normally distributed* about their mean for each value of *x*

$$y \sim N(\beta_1 + \beta_2 x, \sigma^2)$$

2.2.1
Introducing the Error Term

- The random error term is defined as

Eq. 2.3
$$e = y - E(y|x) = y - \beta_1 - \beta_2 x$$

– Rearranging gives

Eq. 2.4
$$y = \beta_1 + \beta_2 x + e$$

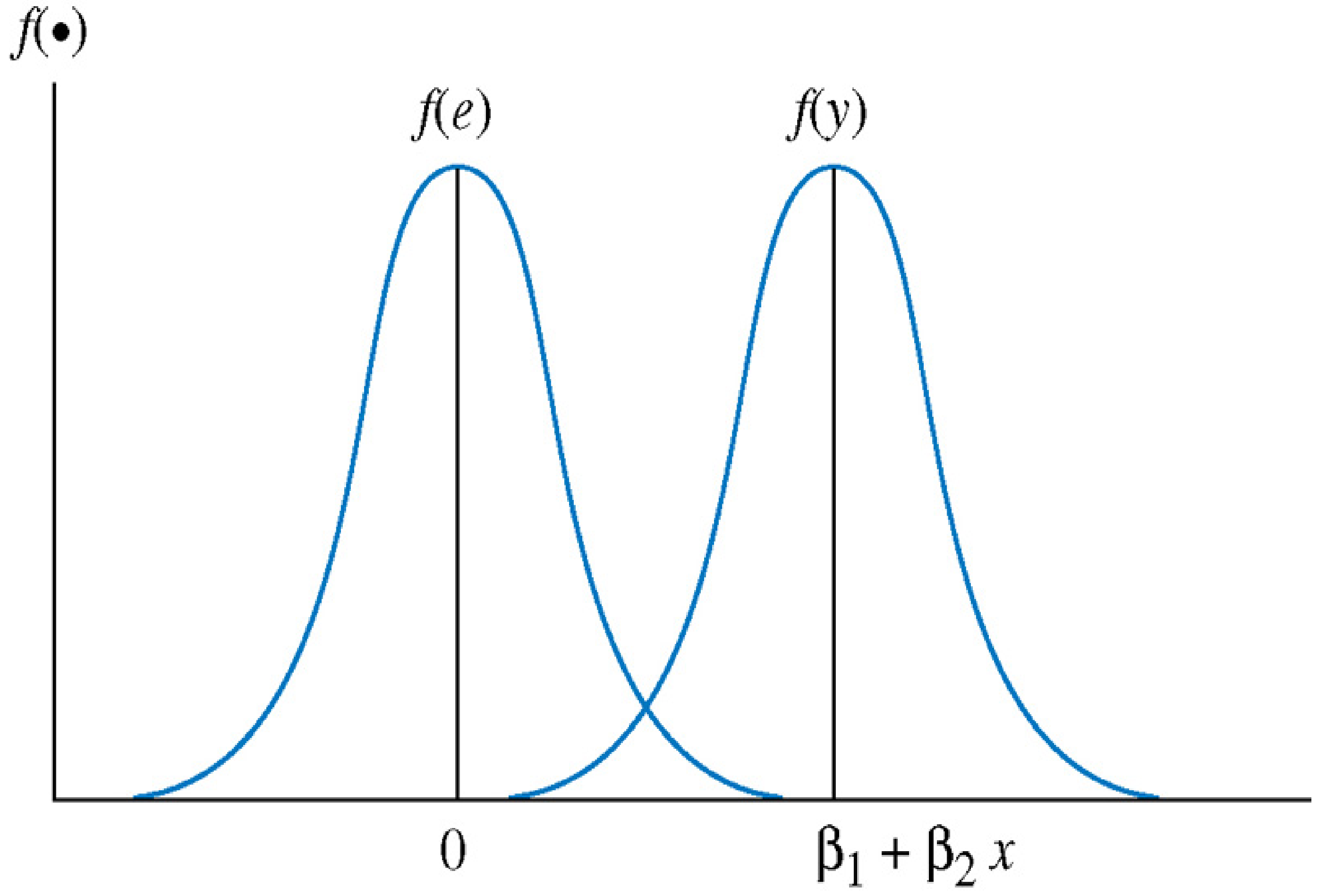
where *y* is the dependent variable and *x* is the independent variable

- The expected value of the error term, given *x*, is

$$E(e \mid x) = E(y \mid x) - \beta_1 - \beta_2 x = 0$$

The mean value of the error term, given *x*, is zero

Figure 2.4 Probability density functions for *e* and *y*



ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - II

Assumption SR1:

The value of *y*, for each value of *x*, is:

$$y = \beta_1 + \beta_2 x + e$$

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - II

Assumption SR2:

The expected value of the random error *e* is:

$$E(e) = 0$$

This is equivalent to assuming that

$$E(y) = \beta_1 + \beta_2 x$$

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - II

Assumption SR3:

The variance of the random error *e* is:

$$\text{var}(e) = \sigma^2 = \text{var}(y)$$

The random variables *y* and *e* have the same variance because they differ only by a constant.

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - II

Assumption SR4:

The covariance between any pair of random errors, e_i and e_j is:

$$\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$$

The stronger version of this assumption is that the random errors *e* are statistically independent, in which case the values of the dependent variable *y* are also statistically independent

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - II

Assumption SR5:

The variable x is not random, and must take at least two different values

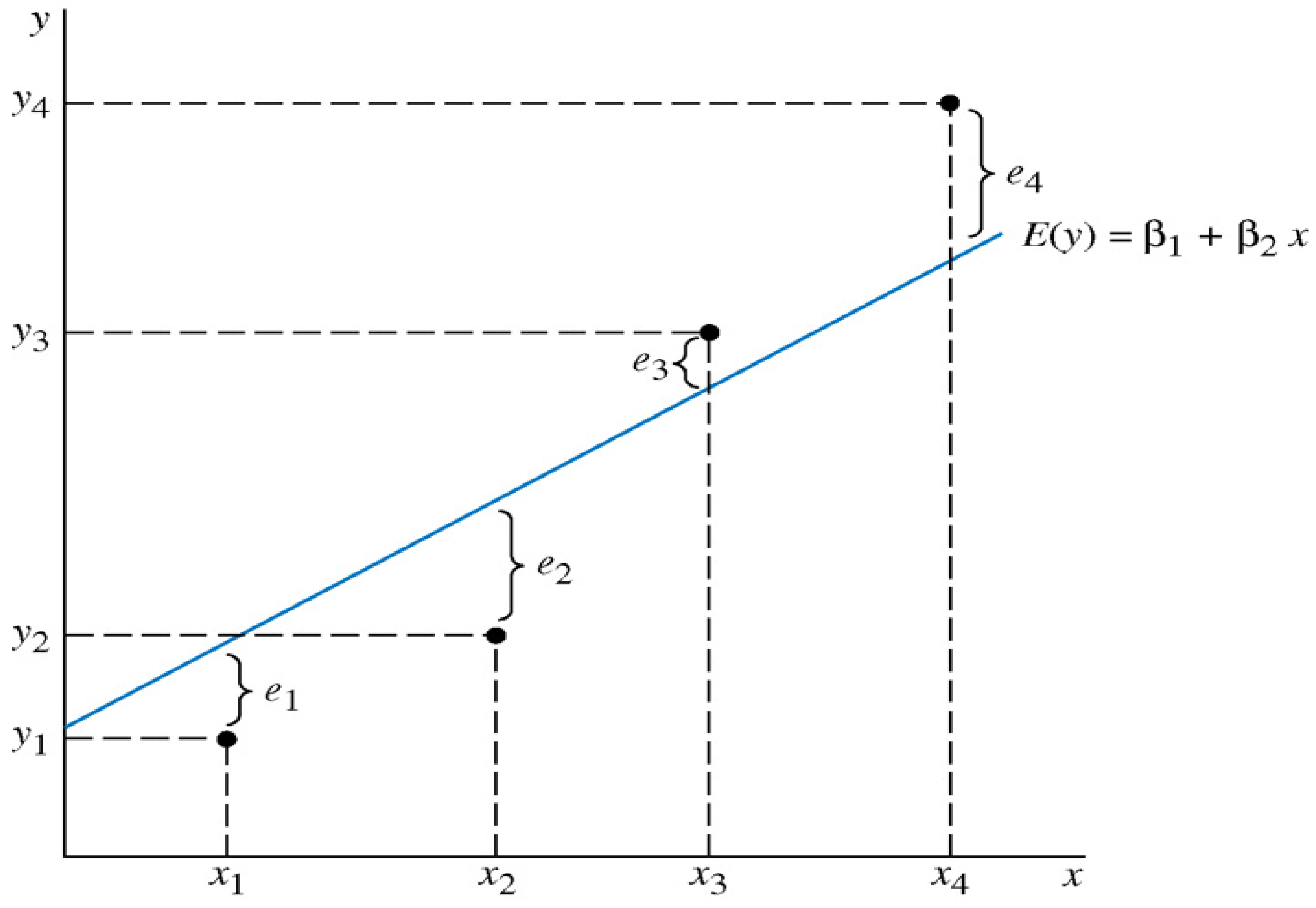
ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - II

Assumption SR6:

(*optional*) The values of *e* are *normally distributed* about their mean if the values of *y* are normally distributed, and *vice versa*

$$e \sim N(0, \sigma^2)$$

Figure 2.5 The relationship among *y*, *e* and the true regression line

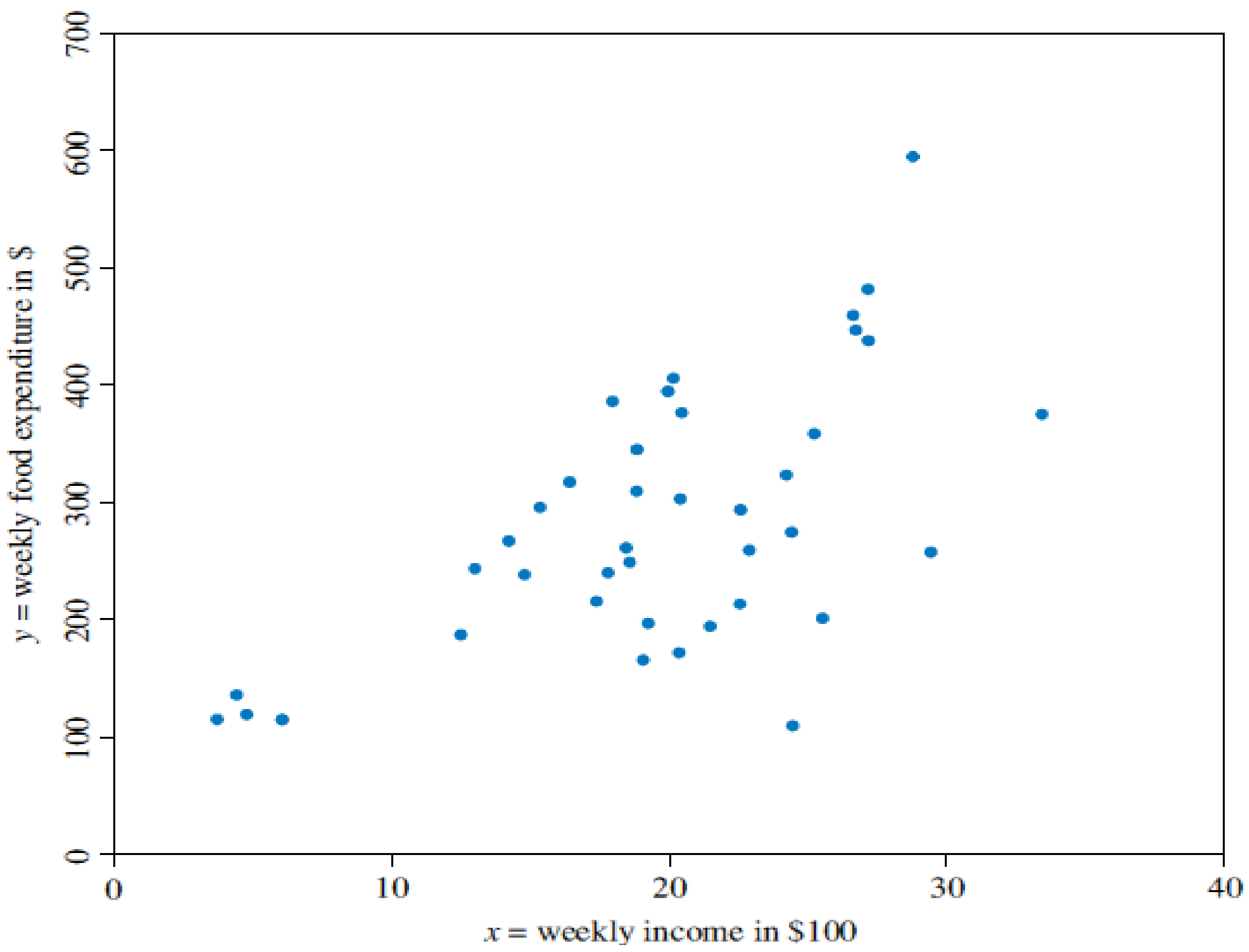


2.3
Estimating the Regression Parameters

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

Figure 2.6 Data for food expenditure example



- The fitted regression line is:

Eq. 2.5

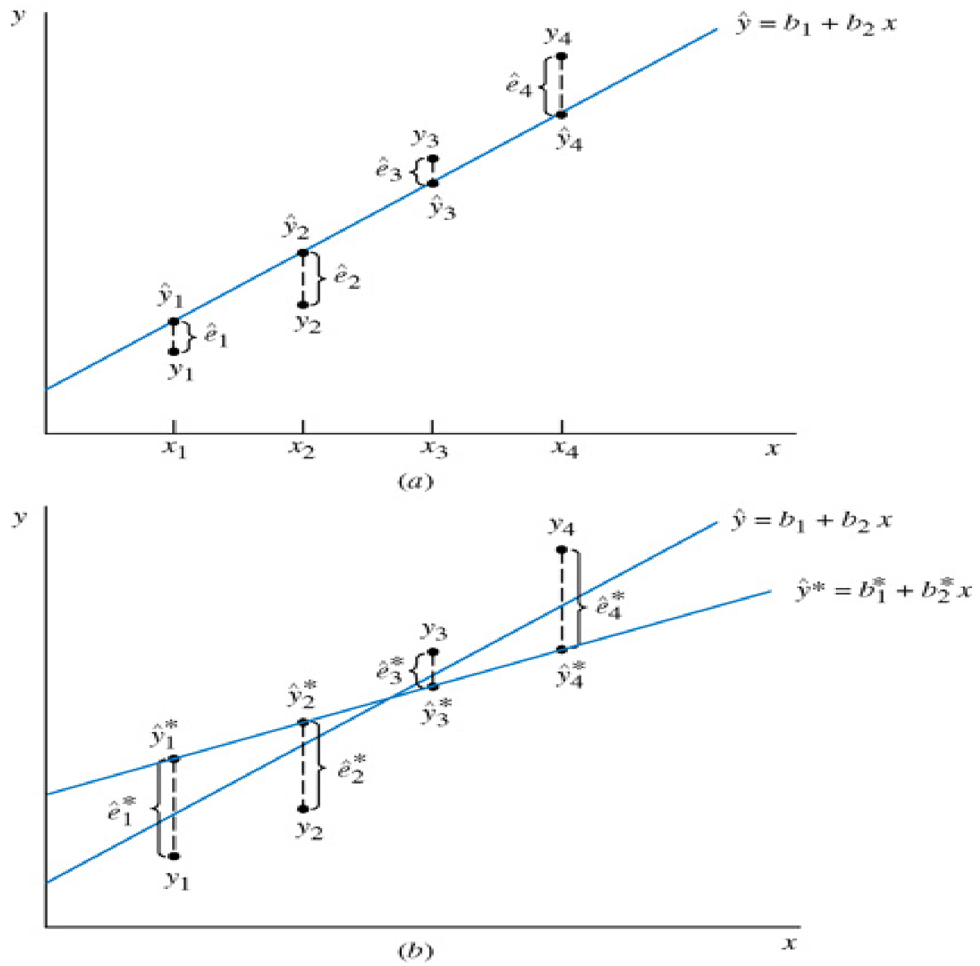
$$\hat{y}_i = b_1 + b_2 x_i$$

The least squares residual is:

Eq. 2.6

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

Figure 2.7 The relationship among y, $\hat{e}$ and the fitted regression line



- Suppose we have another fitted line:

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

The least squares line has the smaller sum of squared residuals:

$$\text{if } SSE = \sum_{i=1}^{N} \hat{e}_i^2 \text{ and } SSE^* = \sum_{i=1}^{N} \hat{e}_i^{*2} \text{ then } SSE < SSE^*$$

- Least squares estimates for the unknown parameters β_1 and β_2 are obtained my minimizing the sum of squares function:

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

THE LEAST SQUARES ESTIMATORS

Eq. 2.7

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

Eq. 2.8

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

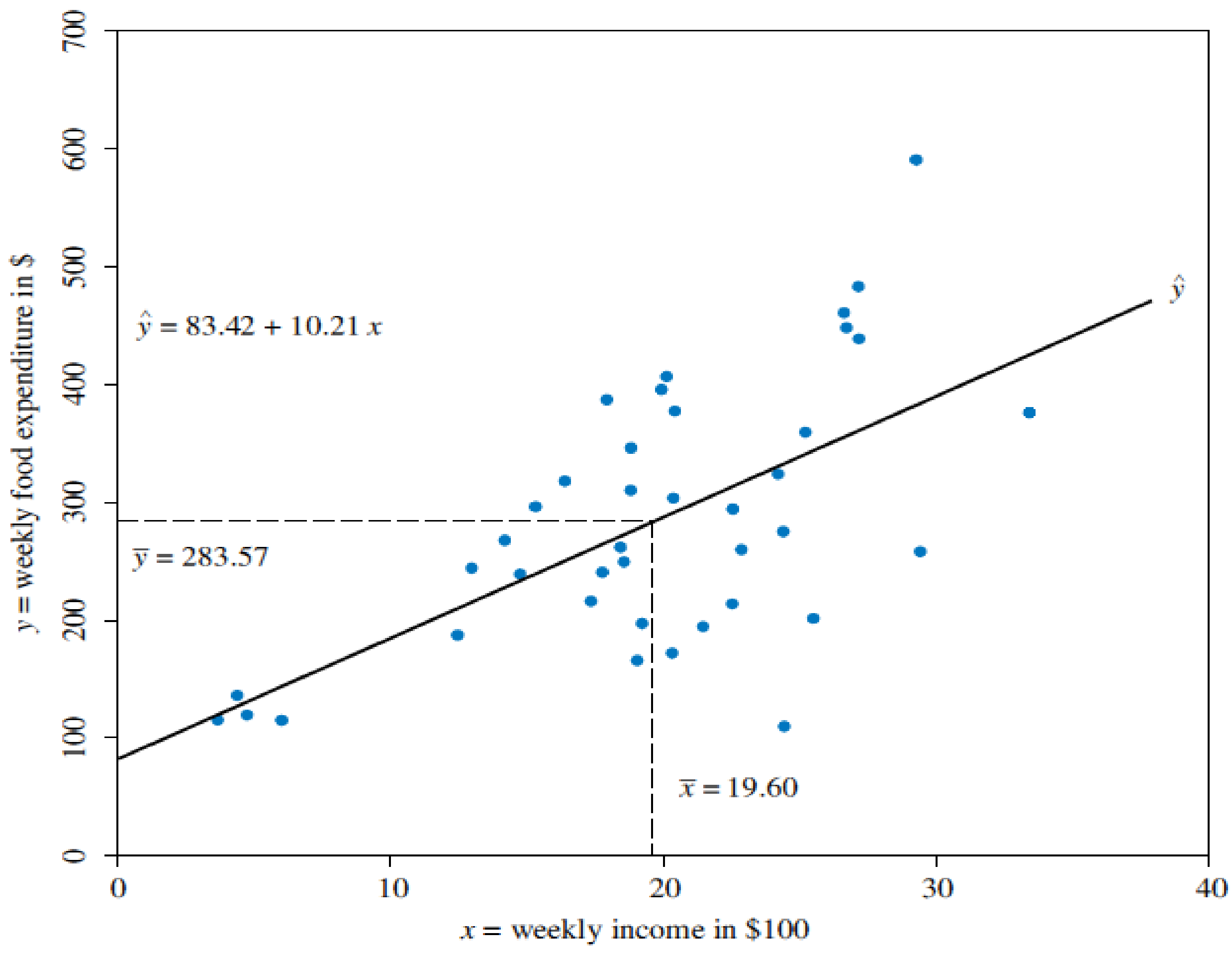
$$b_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{18671.2684}{1828.7876} = 10.2096$$

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

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

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

Figure 2.8 The fitted regression line



- The value b_2 = 10.21 is an estimate of β_2, the amount by which weekly expenditure on food per household increases when household weekly income increases by $100. Thus, we estimate that if income goes up by $100, expected weekly expenditure on food will increase by approximately $10.21
  - Strictly speaking, the intercept estimate b_1 = 83.42 is an estimate of the weekly food expenditure on food for a household with zero income

- Income elasticity is a useful way to characterize the responsiveness of consumer expenditure to changes in income. The elasticity of a variable *y* with respect to another variable *x* is:

$$\varepsilon = \frac{\text{percentage change in } y}{\text{percentage change in } x} = \frac{\Delta y}{\Delta x}\frac{x}{y}$$

In the linear economic model given by Eq. 2.1 we have shown that

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

- The elasticity of mean expenditure with respect to income is:

Eq. 2.9

$$\varepsilon = \frac{\Delta E(y)/E(y)}{\Delta x/x} = \frac{\Delta E(y)}{\Delta x}\frac{x}{E(y)} = \beta_2 \frac{x}{E(y)}$$

A frequently used alternative is to calculate the elasticity at the "point of the means" because it is a representative point on the regression line.

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

- Suppose that we wanted to predict 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:

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

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

Figure 2.9 EViews Regression Output

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		

2.4
Assessing the Least Squares Fit

- 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 are random variables, then what are their expected values, variances, covariances, and probability distributions?
    2. How do the least squares estimators compare with other procedures that might be used, and how can we compare alternative estimators?

- The estimator b_2 can be rewritten as:

Eq. 2.10

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

where

Eq. 2.11

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

It could also be write as:

Eq. 2.12

$$b_2 = \beta_2 + \sum w_i e_i$$

- We will show that if our model assumptions hold, then $E(b_2) = \beta_2$, which means that the estimator is **unbiased**. We can find the expected value of b_2 using the fact that the expected value of a sum is the sum of the expected values:

Eq. 2.13

$$\begin{aligned} E(b_2) &= E(b_2 + \sum w_i e_i) = E(\beta_2 + w_1 e_1 + w_2 e_2 + \ldots + w_N e_N) \\ &= E(\beta_2) + E(w_1 e_1) + E(w_2 e_2) + \ldots + E(w_N e_N) \\ &= E(\beta_2) + \sum E(w_i e_i) \\ &= \beta_2 + \sum w_i E(e_i) \\ &= \beta_2 \end{aligned}$$

using $E(e_i) = 0$ and $E(w_i e_i) = w_i E(e_i)$

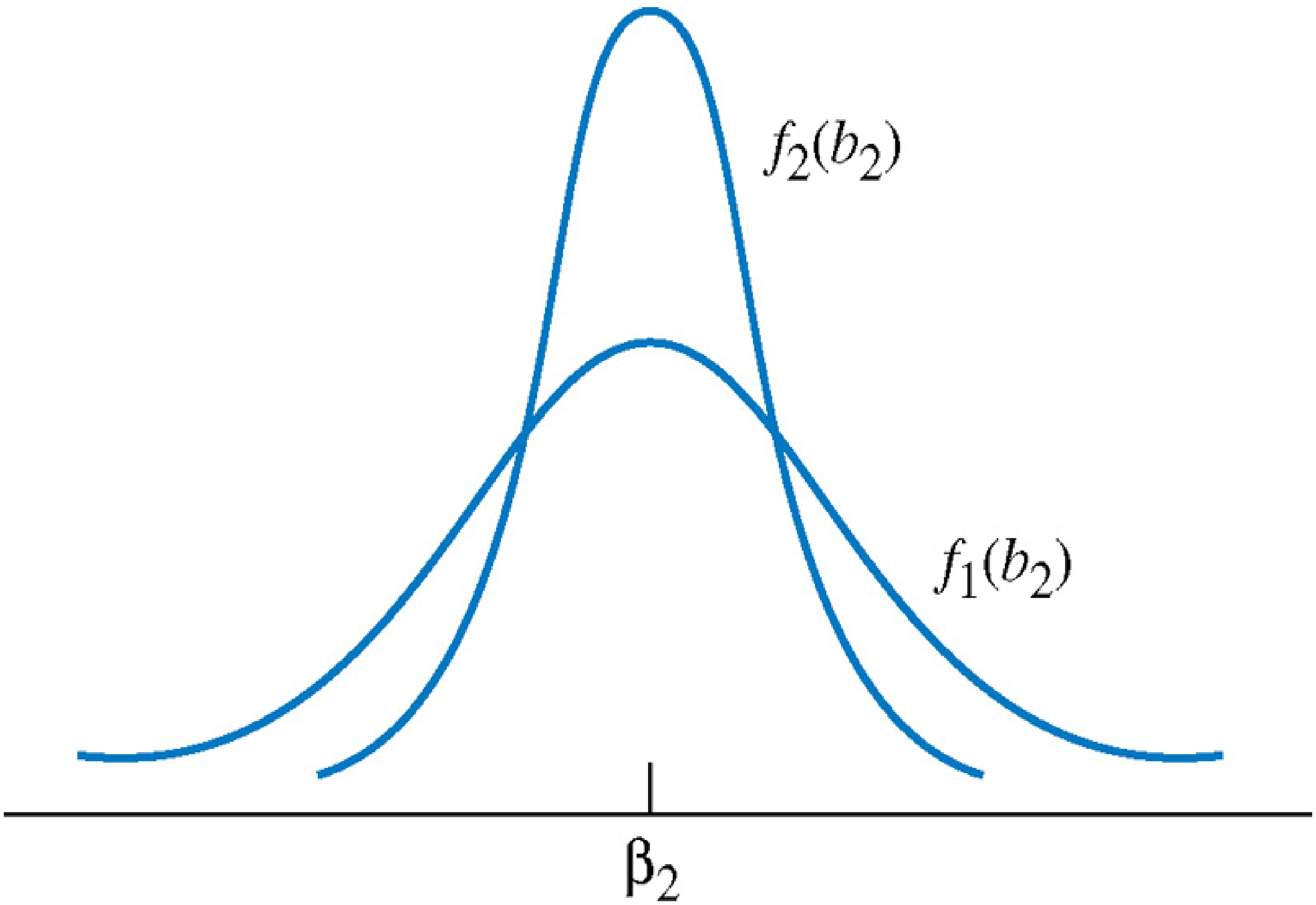
- The property of unbiasedness is about the average values of b_1 and b_2 if many samples of the same size are drawn from the same population
  - If we took the averages of estimates from many samples, these averages would approach the true parameter values b_1 and b_2
  - Unbiasedness does not say that 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

Table 2.2 Estimates from 10 Samples

Sample	b_1	b_2
1	131.69	6.48
2	57.25	10.88
3	103.91	8.14
4	46.50	11.90
5	84.23	9.29
6	26.63	13.55
7	64.21	10.93
8	79.66	9.76
9	97.30	8.05
10	95.96	7.77

Figure 2.10 Two possible probability density functions for b_2

The variance of b_2 is defined as $\text{var}(b_2) = E[b_2 - E(b_2)]^2$



- 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:

Eq. 2.14

$$\operatorname{var}(b_1) = \sigma^2 \left[\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2} \right]$$

Eq. 2.15

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

Eq. 2.16

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

MAJOR POINTS ABOUT THE VARIANCES AND COVARIANCES OF b_1 AND b_2

1. The *larger* the variance term σ^2 , the *greater* the uncertainty there is in the statistical model, and the *larger* the variances and covariance of the least squares estimators.

2. The *larger* the sum of squares, $\sum(x_i - x)^2$, the *smaller* the variances of the least squares estimators and the more *precisely* we can estimate the unknown parameters.

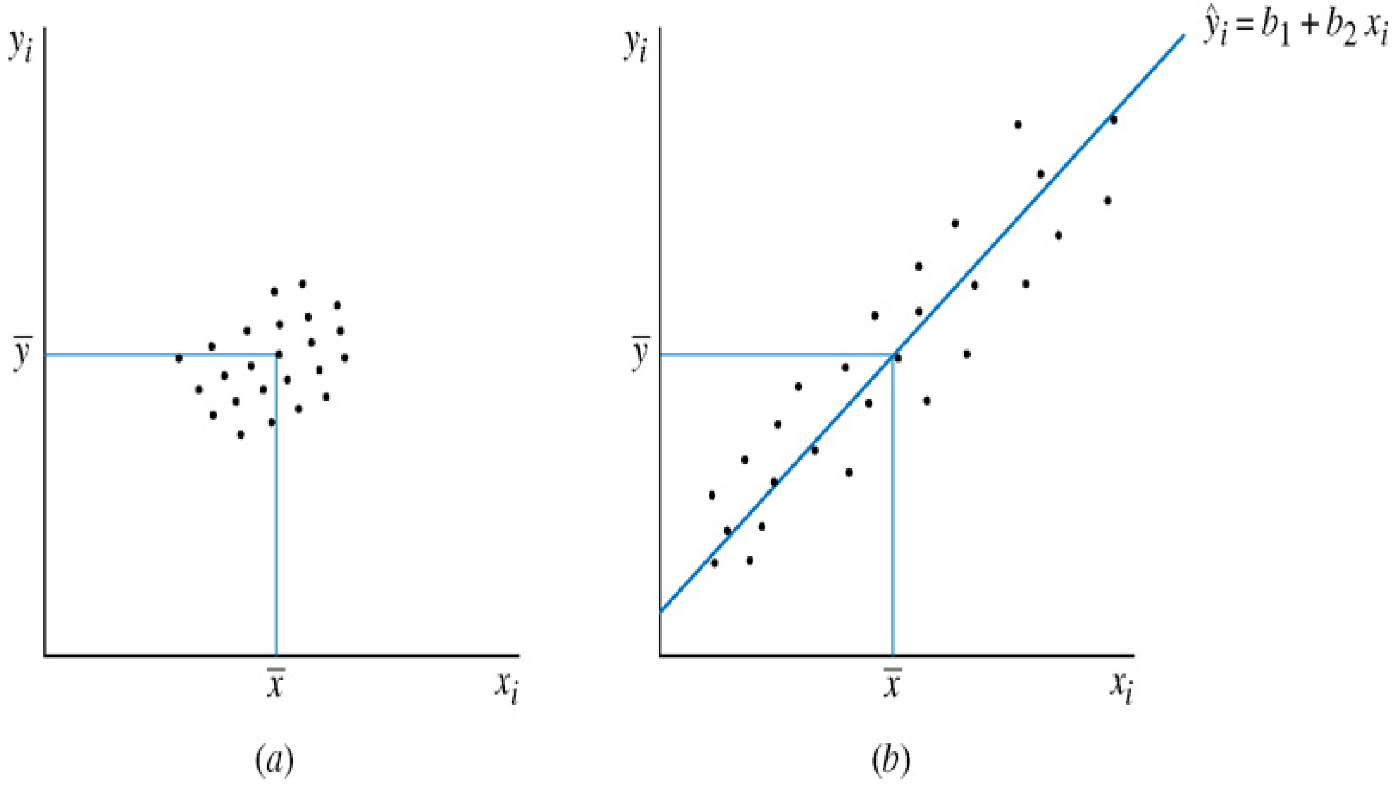
3. The larger the sample size *N*, the *smaller* the variances and covariance of the least squares estimators.

4. The larger the term $\sum x_i^2$, the larger the variance of the least squares estimator b_1.

5. The absolute magnitude of the covariance *increases* the larger in magnitude is the sample mean $\bar{x}$, and the covariance has a *sign* opposite to that of $\bar{x}$.

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

The variance of b_2 is defined as $\text{var}(b_2) = E\left[b_2 - E(b_2)\right]^2$



2.5
The Gauss-Markov Theorem

GAUSS-MARKOV THEOREM

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 b_1 and b_2. They are the **Best Linear Unbiased Estimators (BLUE)** of b_1 and b_2

MAJOR POINTS ABOUT THE GAUSS-MARKOV THEOREM

1. The estimators b_1 and b_2 are "best" when compared to similar estimators, those which 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.

MAJOR POINTS ABOUT THE GAUSS-MARKOV THEOREM

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 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.

2.6
The Probability Distributions of the Least Squares Estimators

- *If* we make the normality assumption (assumption SR6 about the error term) then the least squares estimators are normally distributed:

Eq. 2.17

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

Eq. 2.18

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

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 Eq. 2.17 and Eq. 2.18

2.7

Estimating the Variance of the Error Term

- The variance of the random error e_i is:

$$\text{var}(e_i) = \sigma^2 = E[e_i - E(e_i)]^2 = E(e_i)^2$$

if the assumption $E(e_i) = 0$ is correct.

Since the "expectation" is an average value we might consider estimating σ^2 as the average of the squared errors:

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

where the error terms are $e_i = y_i - \beta_1 - \beta_2 x_i$

- The least squares residuals are obtained by replacing the unknown parameters by their least squares estimates:

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

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

There is a simple modification that produces an unbiased estimator, and that is:

Eq. 2.19

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

so that:

$$E\left(\hat{\sigma}^2\right) = \sigma^2$$

- Replace the unknown error variance σ^2 in Eq. 2.14 – Eq. 2.16 by $\hat{\sigma}^2$ to obtain:

Eq. 2.20

$$\widehat{\operatorname{var}}(b_1) = \hat{\sigma}^2 \left[\frac{\sum x_i^2}{N \sum (x_i - x)^2} \right]$$

Eq. 2.21

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

Eq. 2.22

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

- The square roots of the estimated variances are the "standard errors" of b_1 and b_2:

Eq. 2.23

$$se(b_1) = \sqrt{\widehat{var}(b_1)}$$

Eq. 2.24

$$se(b_2) = \sqrt{\widehat{var}(b_2)}$$

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

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

- The estimated variances and covariances for a regression are arrayed in a rectangular array, or *matrix*, with variances on the diagonal and covariances in the "off-diagonal" positions.

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

- For the food expenditure data the estimated covariance matrix is:

	C	**Income**
C	1884.442	-85.90316
Income	-85.90316	4.381752

- 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.
  - 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:

$$b_2 \sim N\left(\beta_2, \operatorname{var}(b_2) = \sigma^2 \Big/ \sum \left(x_i - \bar{x}\right)^2\right)$$

- The estimator variance, var(b_2), or its square root, $\sigma_{b_2} = \sqrt{\operatorname{var}(b_2)}$ which we might call the true standard deviation of b_2, measures the sampling variation of the estimates b_2
  - 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
  - If σ_{b_2} is small relative to the parameter b_2,we know that the least squares estimate will fall near b_2 with high probability

- The question we address with the standard error is "*How much variation about their means do the estimates exhibit from sample to sample*?"

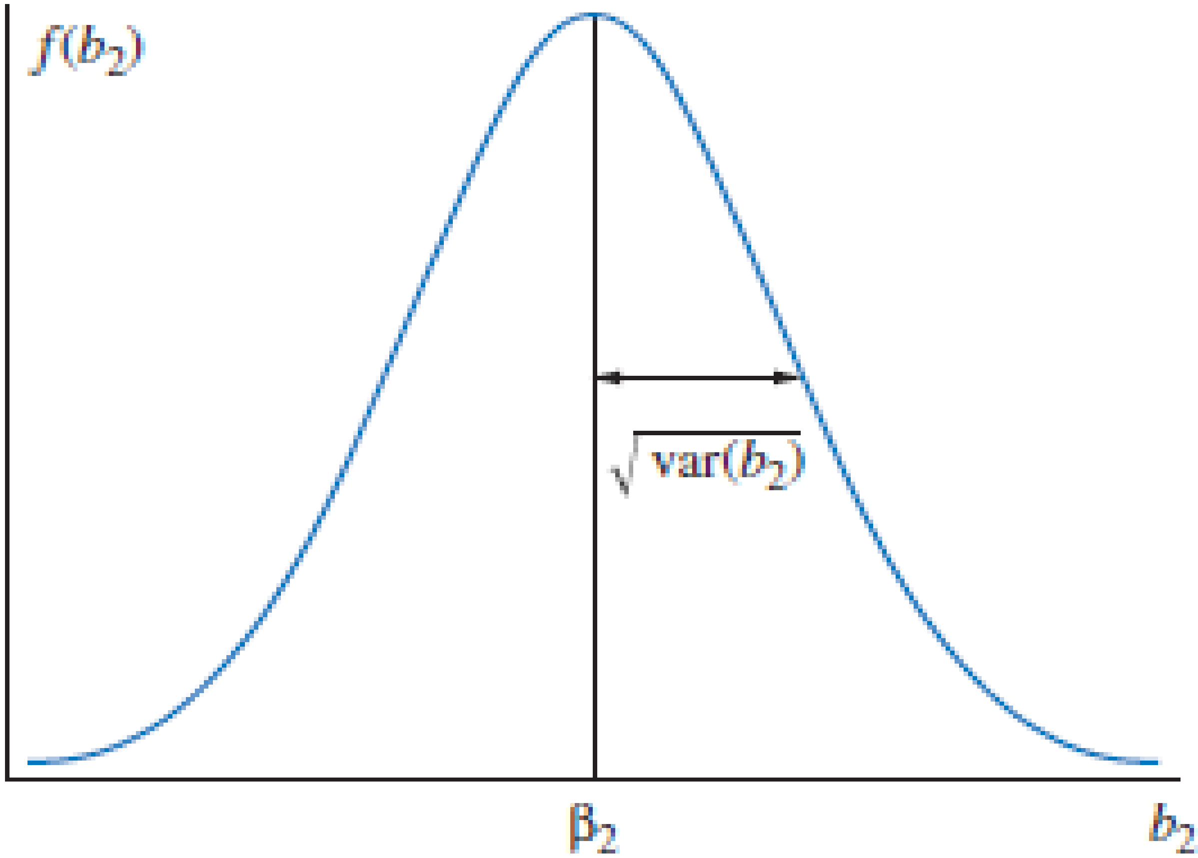
2.7.3
Interpreting the Standard Errors

- We estimate σ^2, and then estimate σ_{b_2} using:

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

  – 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

Figure 2.12 The probability density function of the least squares estimator b_2.



Key Words

Keywords

- assumptions
- asymptotic
- B.L.U.E.
- biased estimator
- degrees of freedom
- dependent variable
- deviation from the mean form
- econometric model
- economic model
- elasticity
- Gauss-Markov Theorem
- heteroskedastic
- homoskedastic
- independent variable
- least squares estimates
- least squares estimators
- least squares principle
- least squares residuals
- linear estimator
- prediction
- random error term
- regression model
- regression parameters
- repeated sampling
- sampling precision
- sampling properties
- scatter diagram
- simple linear regression function
- specification error
- unbiased estimator