The Simple Linear Regression Model: Specification and Estimation

Chapter 2

Prepared by Vera Tabakova, East Carolina University

Chapter 2:
The Simple Regression Model

- 2.1 An Economic Model
- 2.2 An Econometric Model
- 2.3 Estimating the Regression Parameters
- 2.4 Assessing the Least Squares Estimators
- 2.5 The Gauss-Markov Theorem
- 2.6 The Probability Distributions of the Least Squares Estimators
- 2.7 Estimating the Variance of the Error Term

2.1 An Economic Model

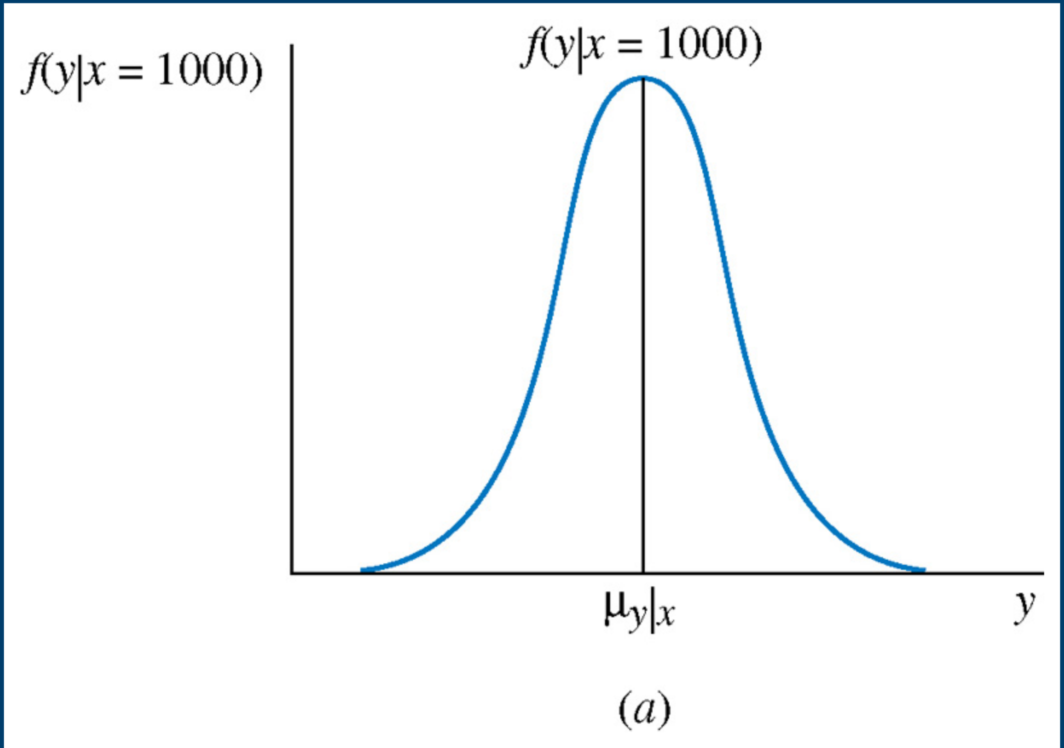


Figure **2.1a** Probability distribution of food expenditure *y* given income *x* = $1000

2.1 An Economic Model

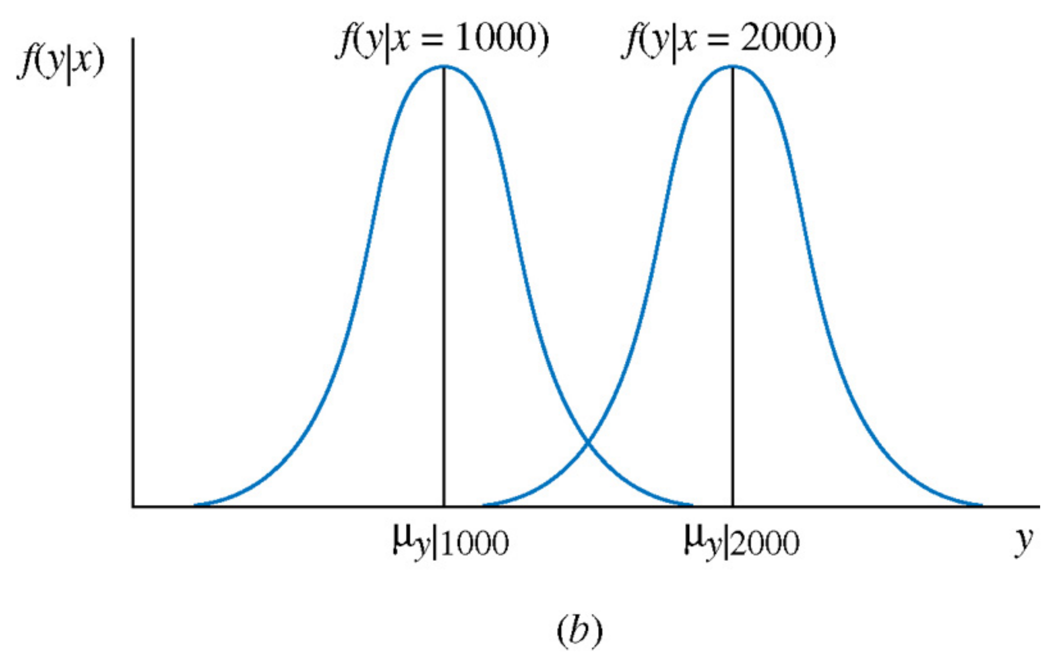


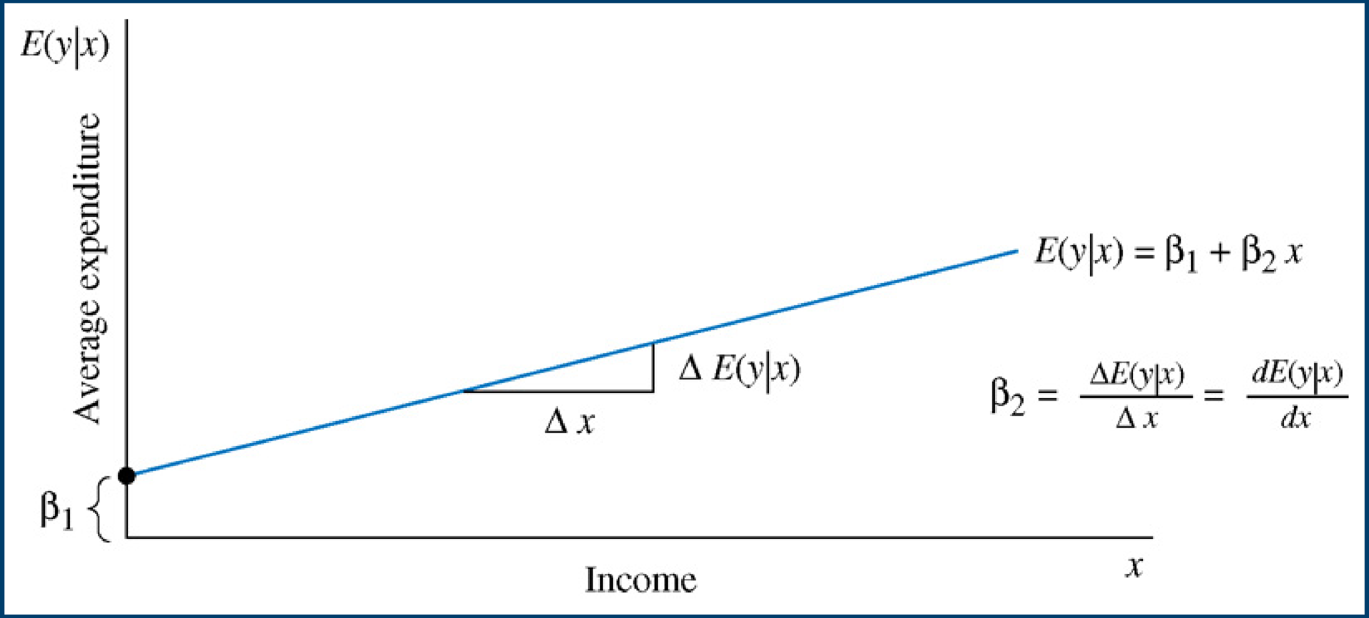
Figure **2.1b** Probability distributions of food expenditures *y* given incomes *x* = \$1000 and *x* = \$2000

2.1 An Economic Model

- The simple regression function

$$E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x \tag{2.1}$$

2.1 An Economic Model



$E(y|x)$
Average expenditure
$E(y|x) = \beta_1 + \beta_2 x$
$\Delta E(y|x)$
Δx
$\beta_2 = \frac{\Delta E(y|x)}{\Delta x} = \frac{dE(y|x)}{dx}$
β_1
Income
x

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

2.1 An Economic Model

- Slope of regression line

$$\beta_2 = \frac{\Delta E(y|x)}{\Delta x} = \frac{dE(y|x)}{dx} \qquad (2.2)$$

"Δ" denotes "change in"

2.2 An Econometric Model

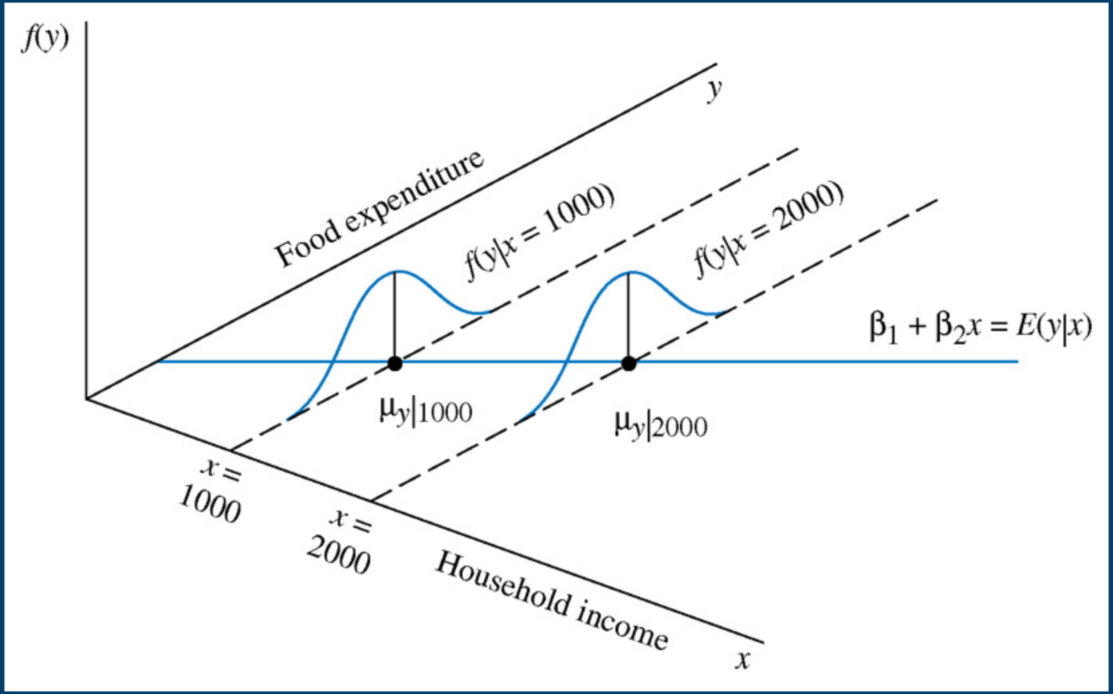


Figure **2.3** The probability density function for *y* at two levels of income

2.2 An Econometric Model

Assumptions of the Simple Linear Regression Model – I

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

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

2.2 An Econometric Model

Assumptions of the Simple Linear Regression Model – I

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

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

2.2 An Econometric Model

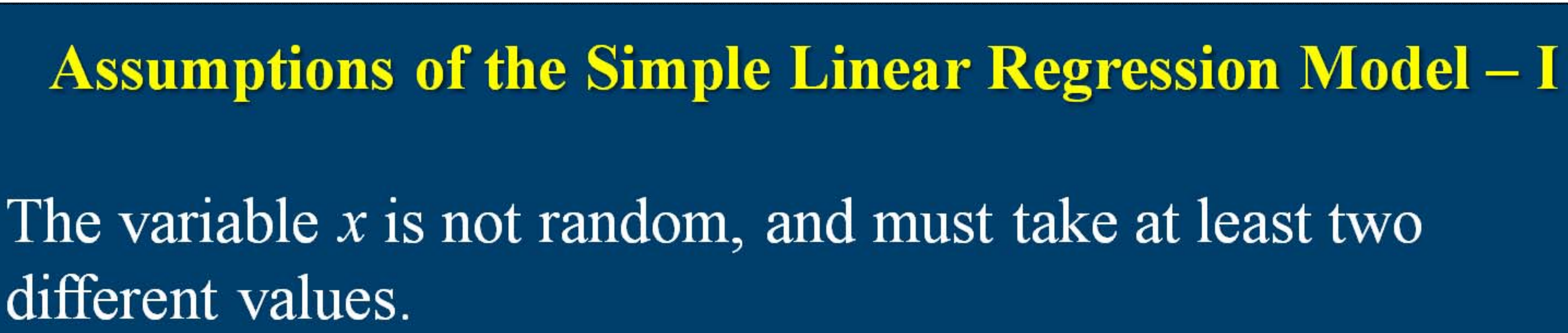
Assumptions of the Simple Linear Regression Model – I

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

$$\operatorname{cov}\left(y_i, y_j\right) = 0$$

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

2.2 An Econometric Model



Assumptions of the Simple Linear Regression Model – I

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

2.2 An Econometric Model

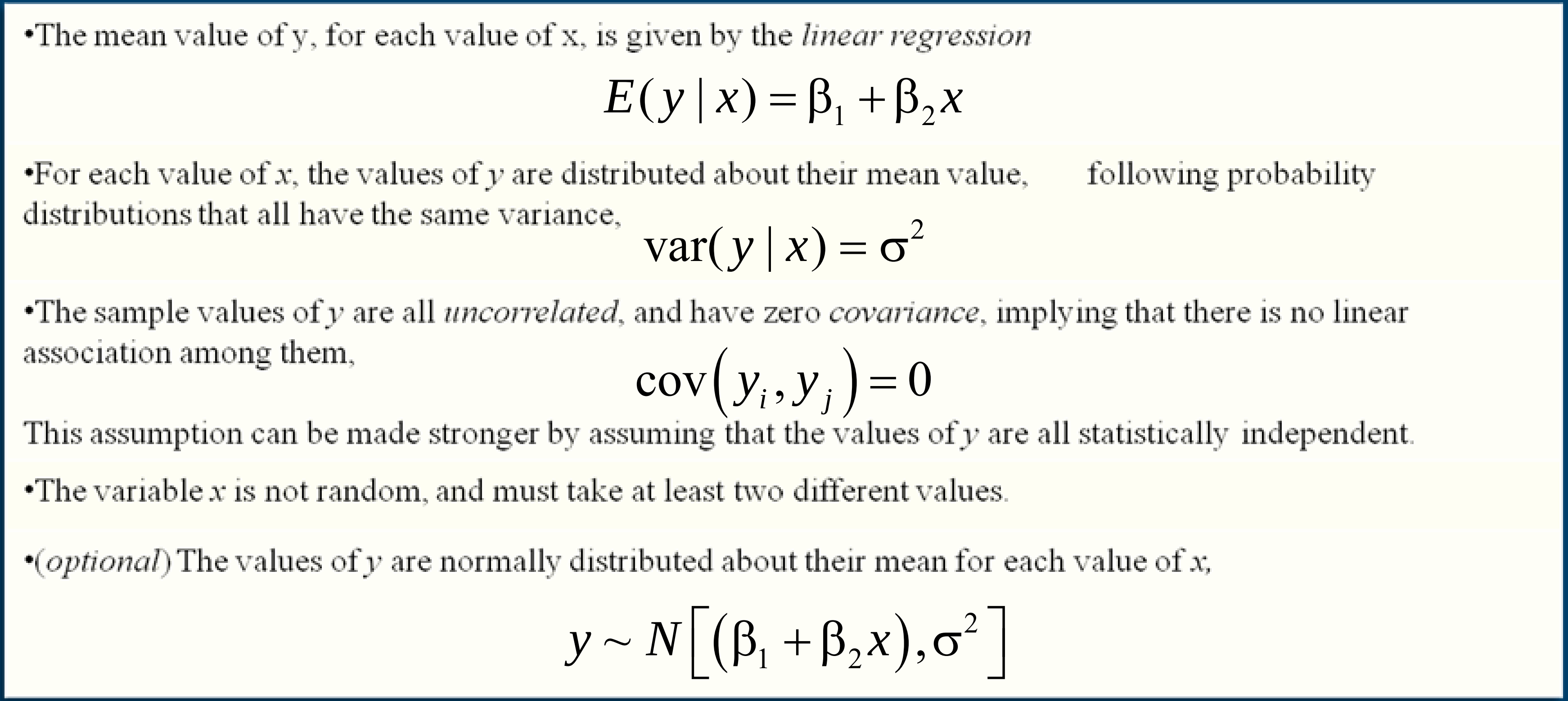
Assumptions of the Simple Linear Regression Model – I

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

$$y \square N\left[\beta_1 + \beta_2 x,\ \sigma^2\right]$$

2.2 An Econometric Model

Assumptions of the Simple Linear Regression Model - I



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

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

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

$$\operatorname{var}(y\,|\,x)=\sigma^2$$

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

$$\operatorname{cov}\left(y_i,y_j\right)=0$$

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

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

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

$$y\sim N\left[\left(\beta_1+\beta_2 x\right),\sigma^2\right]$$

2.2 An Econometric Model

- 2.2.1 Introducing the Error Term
  - The random error term is defined as

$$e = y - E(y \mid x) = y - \beta_1 - \beta_2 x \qquad (2.3)$$

  - Rearranging gives

$$y = \beta_1 + \beta_2 x + e \qquad (2.4)$$

y is dependent variable; *x* is independent variable

2.2 An Econometric Model

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

$$E\left(e\,|\,x\right)=E\left(y\,|\,x\right)-\beta_1-\beta_2 x=0$$

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

2.2 An Econometric Model

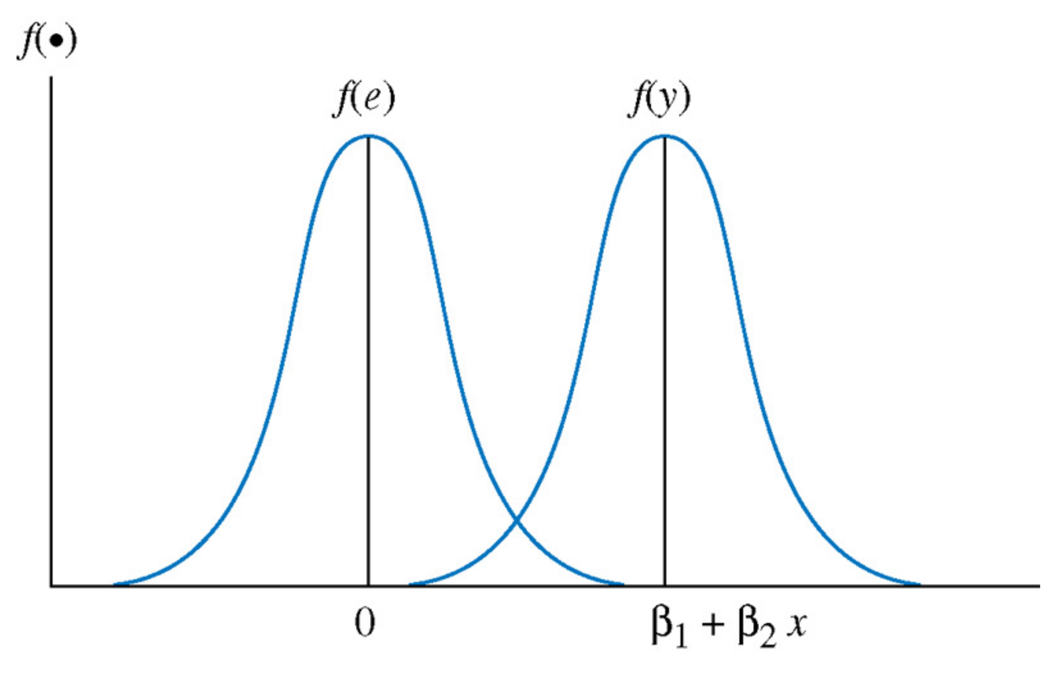


Figure **2.4** Probability density functions for *e* and *y*

2.2 An Econometric Model

Assumptions of the Simple Linear Regression Model – II

SR1. The value of *y*, for each value of *x*, is

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

2.2 An Econometric Model

Assumptions of the Simple Linear Regression Model – II

SR2. The expected value of the random error *e* is

$$E(e) = 0$$

Which is equivalent to assuming that

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

2.2 An Econometric Model

Assumptions of the Simple Linear Regression Model – II

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.

2.2 An Econometric Model

Assumptions of the Simple Linear Regression Model – II

SR4. The covariance between any pair of random errors, e_i and e_j is

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

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.

2.2 An Econometric Model

Assumptions of the Simple Linear Regression Model – II

SR5. The variable x is not random, and must take at least two different values.

2.2 An Econometric Model

Assumptions of the Simple Linear Regression Model – II

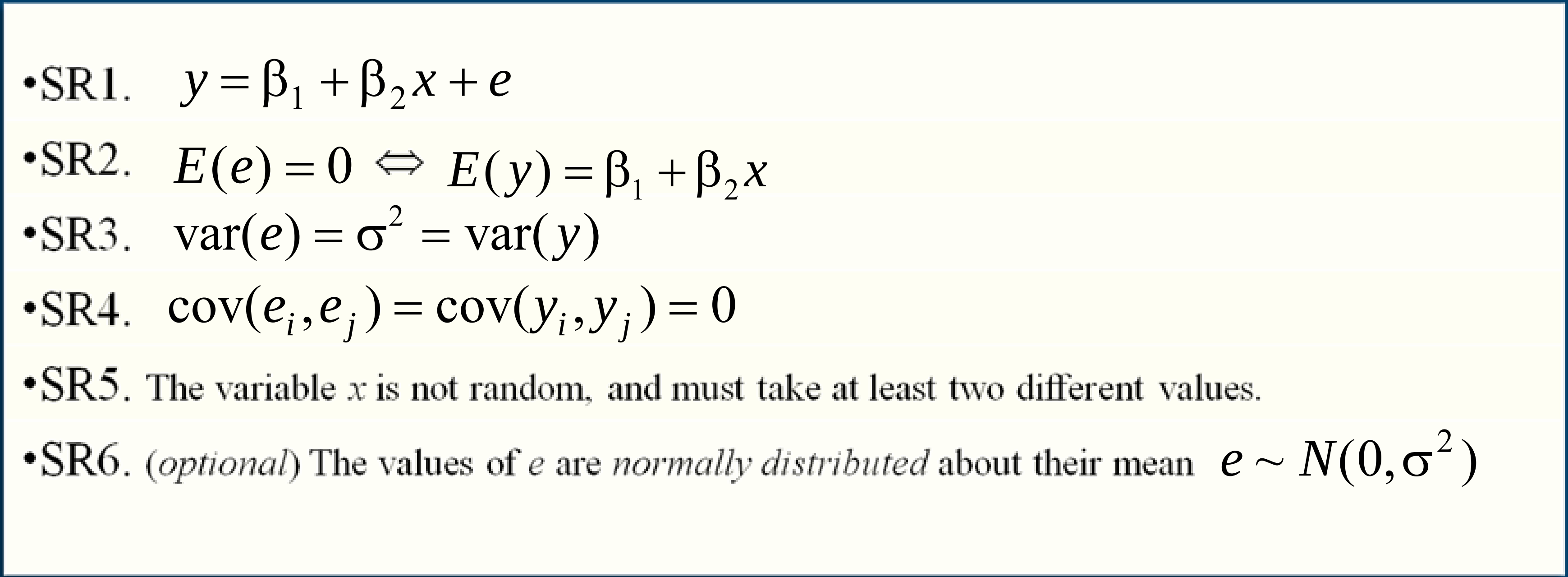
SR6. (*optional*) The values of e are *normally distributed* about their mean

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

if the values of *y* are normally distributed, and *vice versa*.

2.2 An Econometric Model

Assumptions of the Simple Linear Regression Model - II



- SR1. $y = \beta_1 + \beta_2 x + e$
- SR2. $E(e) = 0 \Leftrightarrow E(y) = \beta_1 + \beta_2 x$
- SR3. $\text{var}(e) = \sigma^2 = \text{var}(y)$
- SR4. $\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$
- SR5. The variable x is not random, and must take at least two different values.
- SR6. (*optional*) The values of e are *normally distributed* about their mean $e \sim N(0, \sigma^2)$

2.2 An Econometric Model

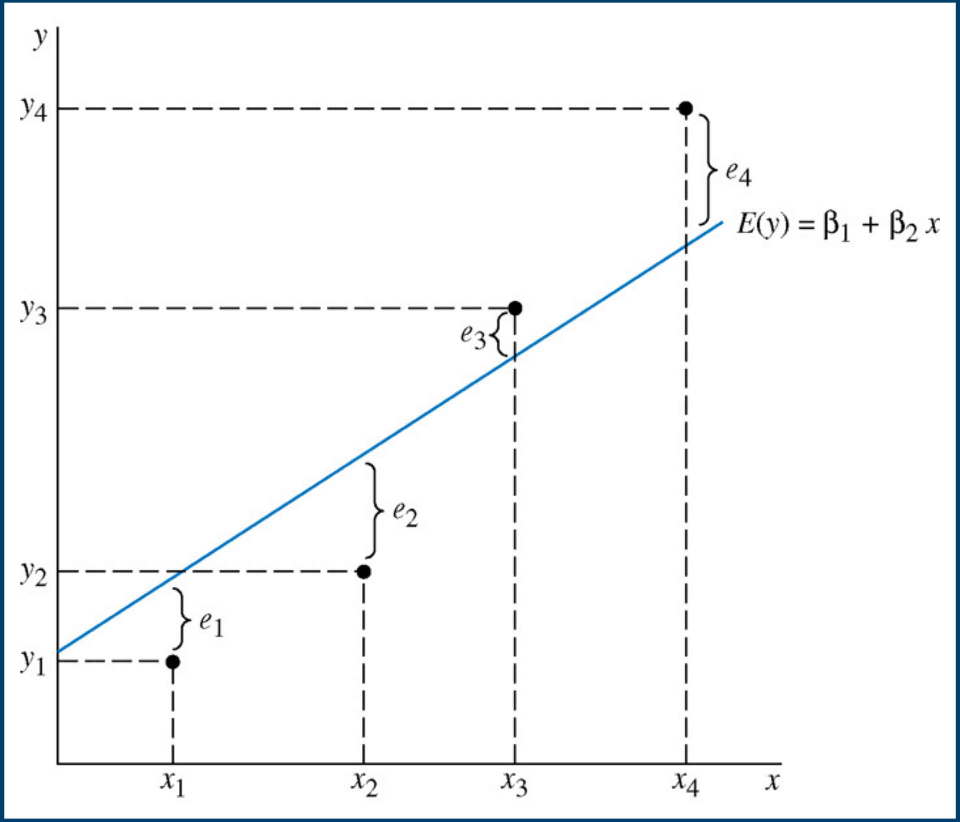


Figure **2.5** The relationship among *y*, *e* and the true regression line

2.3 Estimating The Regression Parameters

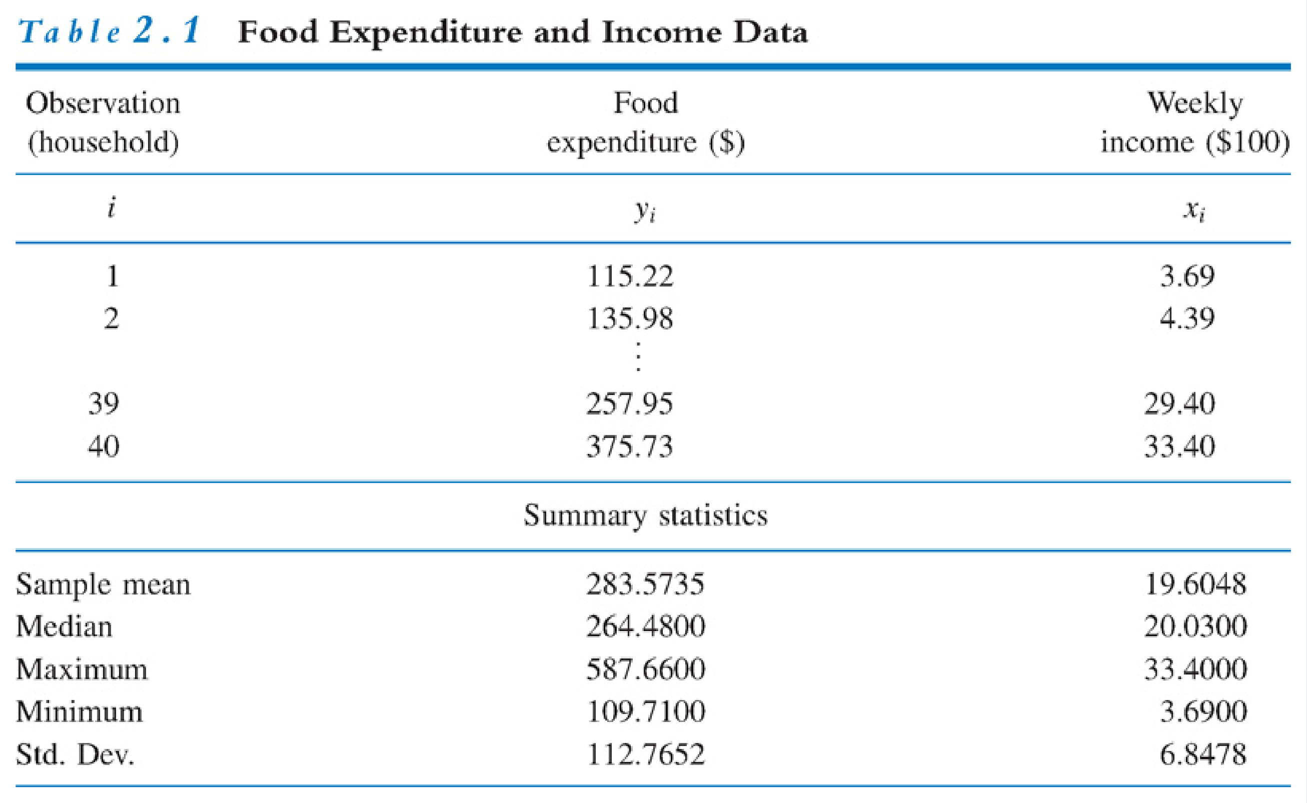


Table 2.1 Food Expenditure and Income Data

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.	112.7652	6.8478

2.3 Estimating The Regression Parameters

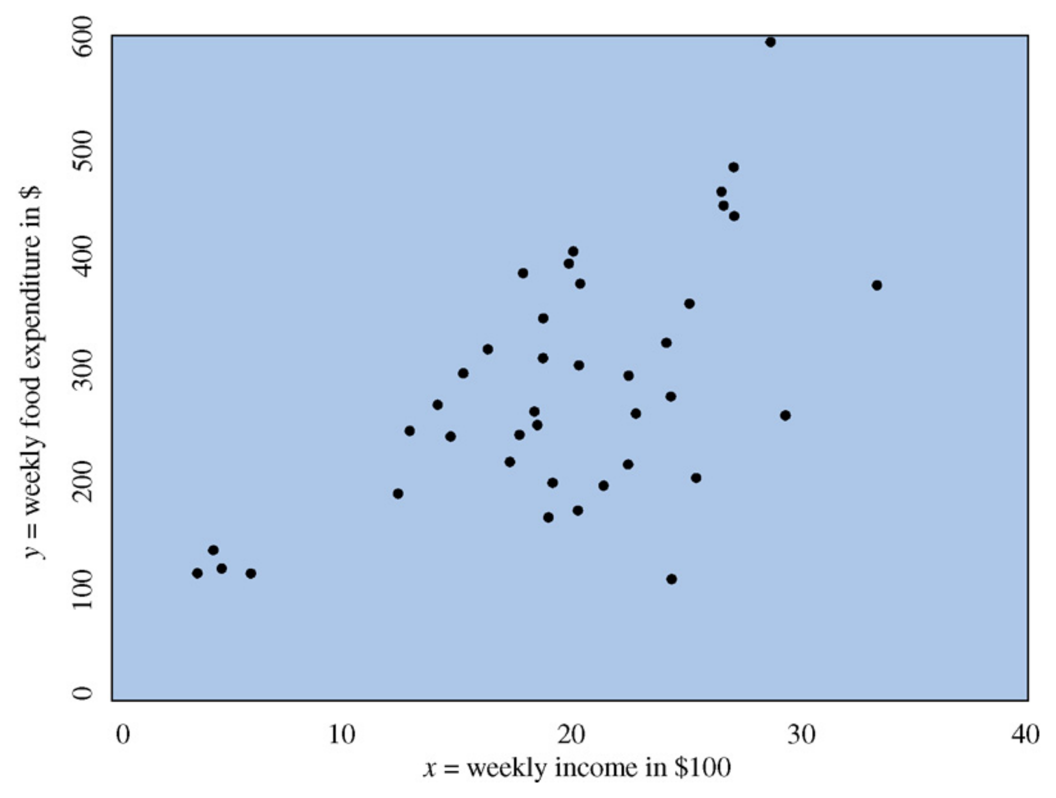


Figure **2.6** Data for food expenditure example

2.3 Estimating The Regression Parameters

- 2.3.1 The Least Squares Principle
  - The fitted regression line is

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

  - The least squares residual

$$e\text{垐}_i = y_i - y_i = y_i - b_1 - b_2 x_i \tag{2.6}$$

2.3 Estimating The Regression Parameters

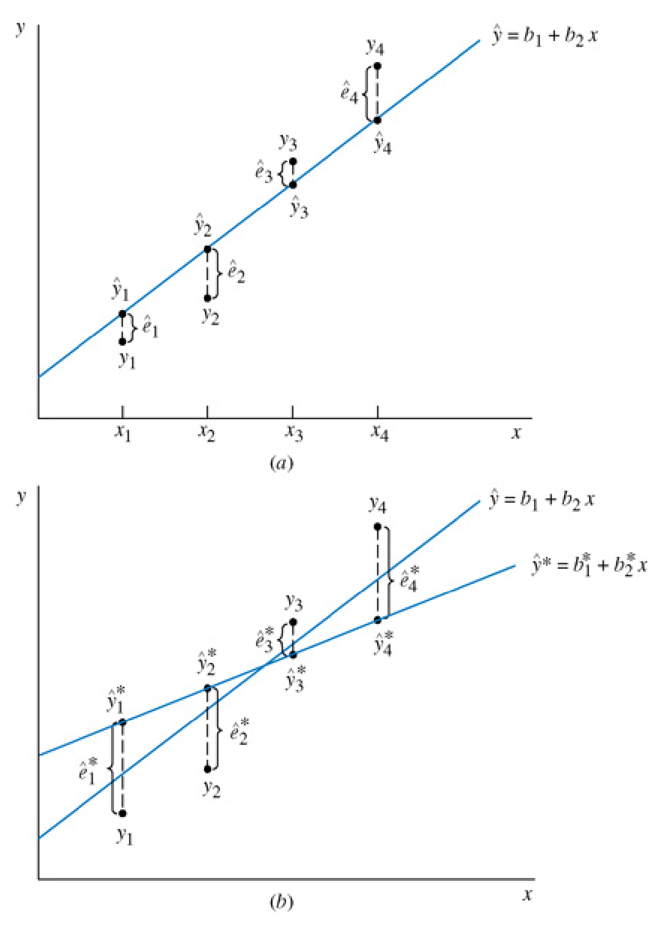


Figure **2.7** The relationship among *y*, ê and the fitted regression line

2.3 Estimating The Regression Parameters

- Any other fitted line

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

- Least squares line has smaller sum of squared residuals

$$\text{if } SSE = \sum_{i=1}^{N} e_i \text{垐} \text{ and } SSE^* = \sum_{i=1}^{N} e_i^{*2} \text{ then } SSE < SSE^*$$

2.3 Estimating The Regression Parameters

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

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

2.3 Estimating The Regression Parameters

- The Least Squares Estimators

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

$$b_1 = \bar{y} - b_2 \bar{x} \qquad (2.8)$$

2.3 Estimating The Regression Parameters

- 2.3.2 Estimates for the Food Expenditure Function

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

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

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

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

2.3 Estimating The Regression Parameters

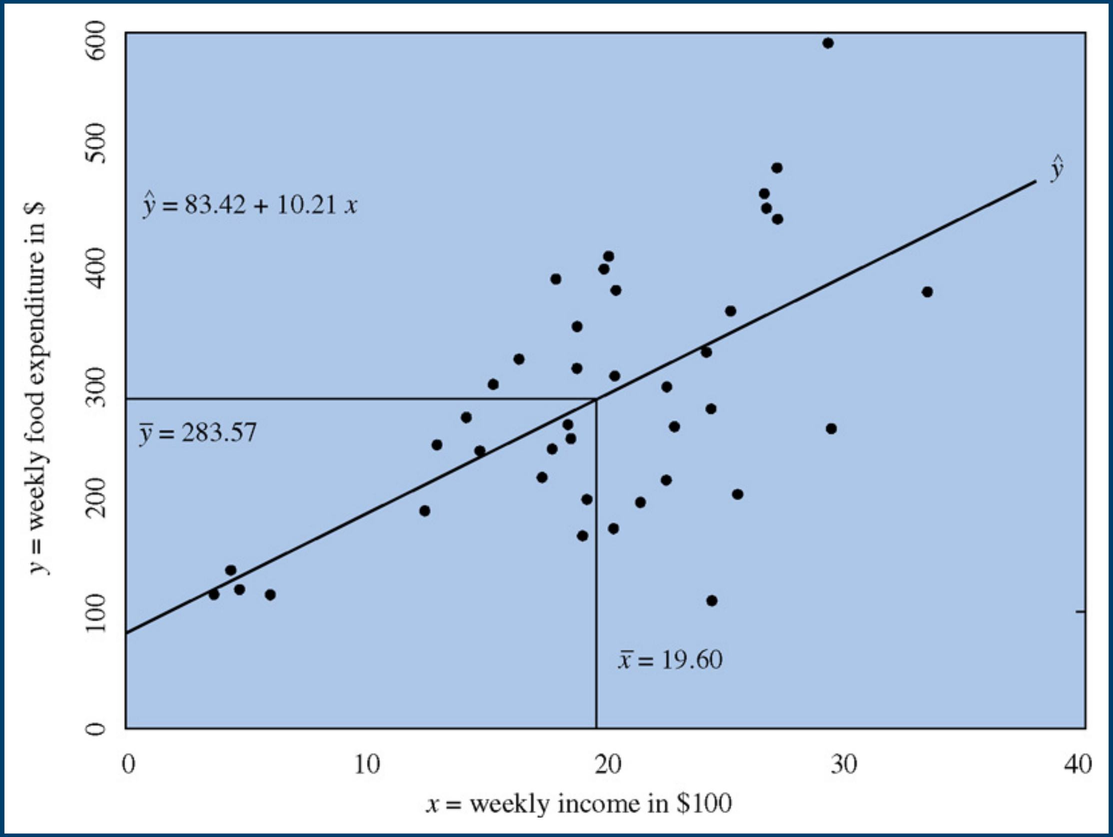


Figure **2.8** The fitted regression line

2.3 Estimating The Regression Parameters

- 2.3.3 Interpreting the Estimates

  - 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.

2.3 Estimating The Regression Parameters

- 2.3.3a Elasticities
  - 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

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

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

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

2.3 Estimating The Regression Parameters

- The elasticity of mean expenditure with respect to income is

$$\varepsilon = \frac{\Delta E(y)/E(y)}{\Delta x/x} = \frac{\Delta E(y)}{\Delta x} \cdot \frac{x}{E(y)} = \beta_2 \cdot \frac{x}{E(y)} \qquad (2.9)$$

- 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.

$$\hat{\varepsilon} = b_2 \frac{\overline{x}}{\overline{y}} = 10.21 \times \frac{19.60}{283.57} = .71$$

2.3 Estimating The Regression Parameters

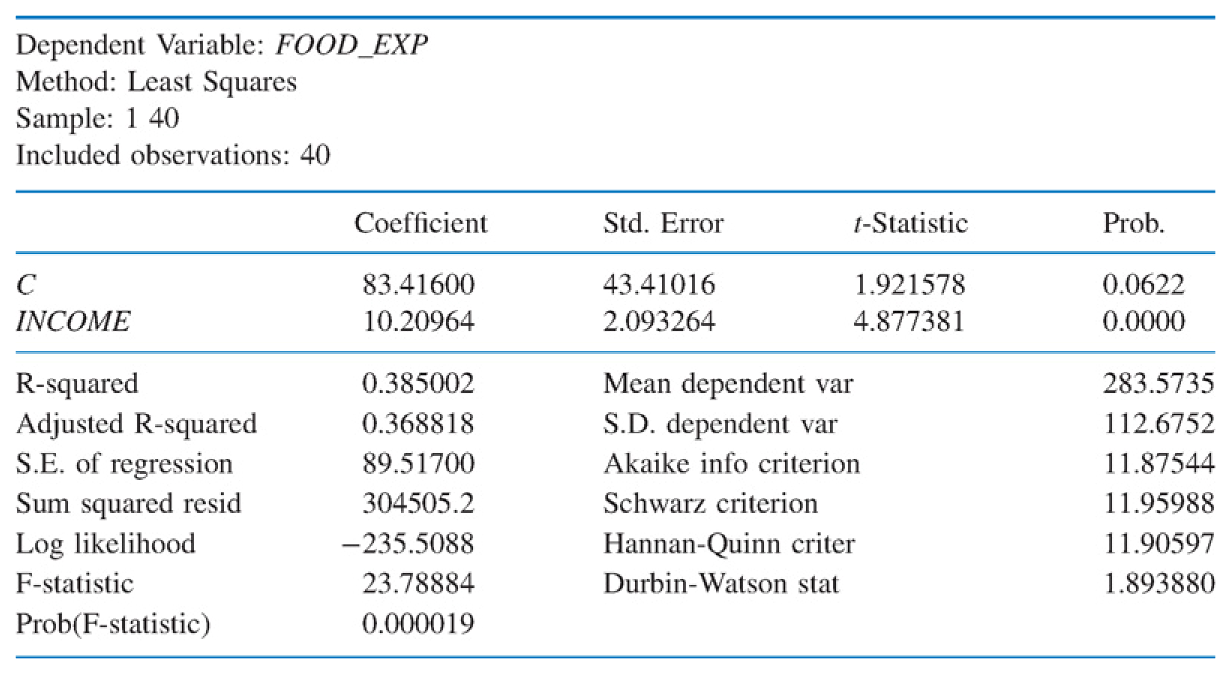
- 2.3.3b Prediction
  - Suppose that we wanted to predict weekly food expenditure for a household with a weekly income of $2000. This prediction is carried out by substituting *x* = 20 into our estimated equation to obtain

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

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

2.3 Estimating The Regression Parameters

- 2.3.3c Examining Computer 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
```

Figure **2.9** EViews Regression Output

2.3 Estimating The Regression Parameters

- 2.3.4 Other Economic Models
  - The “log-log” model

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

$$\frac{d[\ln(y)]}{dx} = \frac{1}{y} \cdot \frac{dy}{dx}$$

$$\frac{d[\beta_1 + \beta_2 \ln(x)]}{dx} = \frac{1}{x} \cdot \beta_2$$

$$\beta_2 = \frac{dy}{dx} \cdot \frac{x}{y}$$

2.4 Assessing the Least Squares Estimators

- 2.4.1 The estimator b_2

$$b_2 = \sum_{i=1}^{N} w_i y_i \qquad (2.10)$$

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

$$b_2 = \beta_2 + \sum w_i e_i \qquad (2.12)$$

2.4 Assessing the Least Squares Estimators

- 2.4.2 The Expected Values of b_1 and b_2
- We will show that if our model assumptions hold, then $E\left(b_2\right)=\beta_2$, 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 expected values

$$\begin{aligned} E(b_2) &= E\left(\beta_2+\sum w_i e_i\right)=E\left(\beta_2+w_1e_1+w_2e_2+\cdots+w_Ne_N\right) \\ &= E\left(\beta_2\right)+E\left(w_1e_1\right)+E\left(w_2e_2\right)+\cdots+E\left(w_Ne_N\right) \\ &= E(\beta_2)+\sum E(w_ie_i) \\ &= \beta_2+\sum w_iE(e_i)=\beta_2 \end{aligned} \qquad (2.13)$$

using $E\left(w_ie_i\right)=w_iE\left(e_i\right)$ and $E(e_i)=0$

2.4 Assessing the Least Squares Estimators

2.4.3 Repeated Sampling

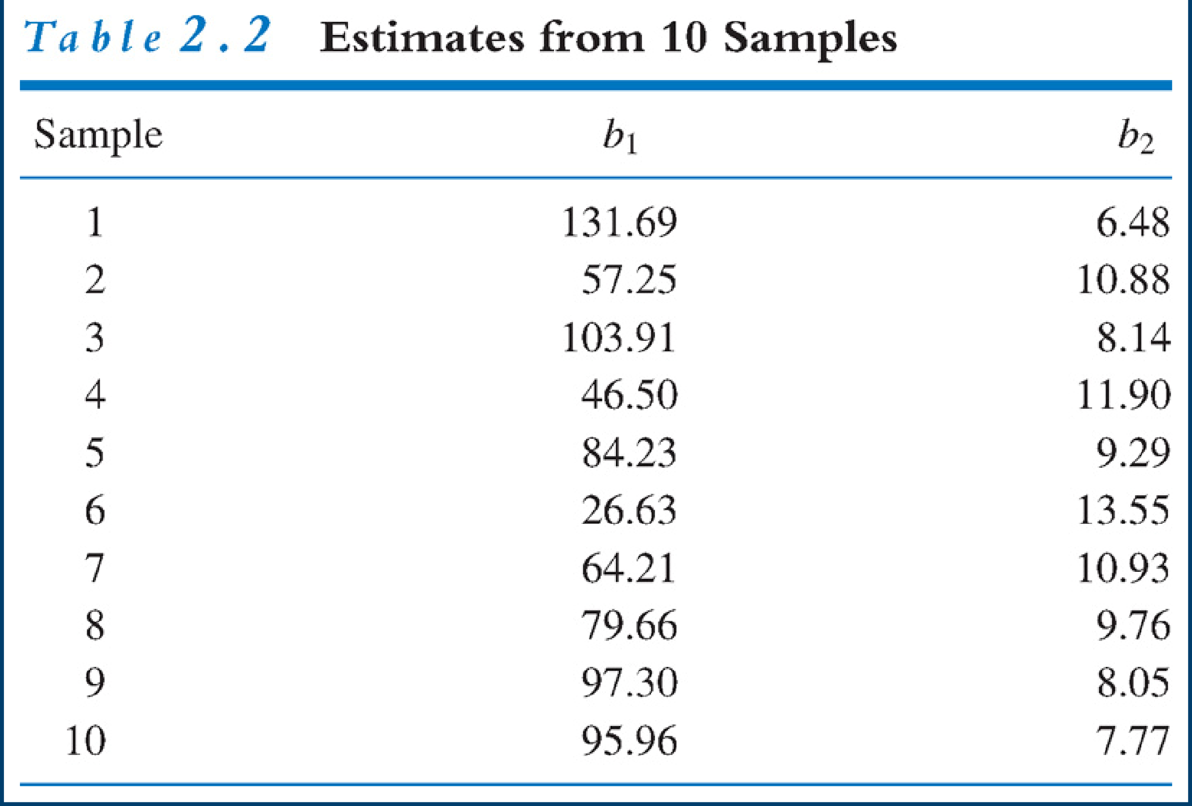


Table 2.2 Estimates from 10 Samples

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

2.4 Assessing the Least Squares Estimators

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

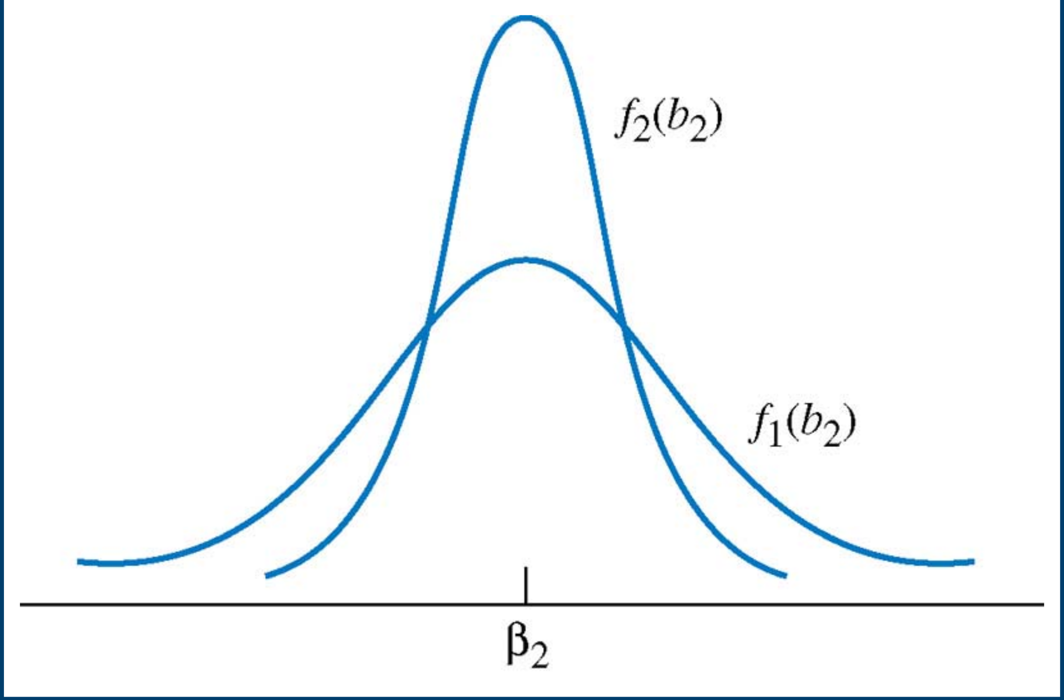


Figure 2.10 Two possible probability density functions for b_2

2.4 Assessing the Least Squares Estimators

- 2.4.4 The Variances and Covariances of b_1 and 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:

$$\operatorname{var}(b_1) = \sigma^2 \left[\frac{\sum x_i^2}{N \sum (x_i - \overline{x})^2} \right] \quad (2.14)$$

$$\operatorname{var}(b_2) = \frac{\sigma^2}{\sum (x_i - \overline{x})^2} \quad (2.15)$$

$$\operatorname{cov}(b_1, b_2) = \sigma^2 \left[\frac{-\overline{x}}{\sum (x_i - \overline{x})^2} \right] \quad (2.16)$$

2.4 Assessing the Least Squares Estimators

- 2.4.4 The Variances and Covariances of b_1 and b_2
- 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.
- The *larger* the sum of squares, $\sum(x_i - \overline{x})^2$, the *smaller* the variances of the least squares estimators and the more *precisely* we can estimate the unknown parameters.
- The larger the sample size *N*, the *smaller* the variances and covariance of the least squares estimators.
- The larger this term $\sum x_i^2$ is, the larger the variance of the least squares estimator b_1.
- The absolute magnitude of the covariance *increases* the larger in magnitude is the sample mean $\overline{x}$, and the covariance has a *sign* opposite to that of $\overline{x}$.

2.4 Assessing the Least Squares Estimators

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

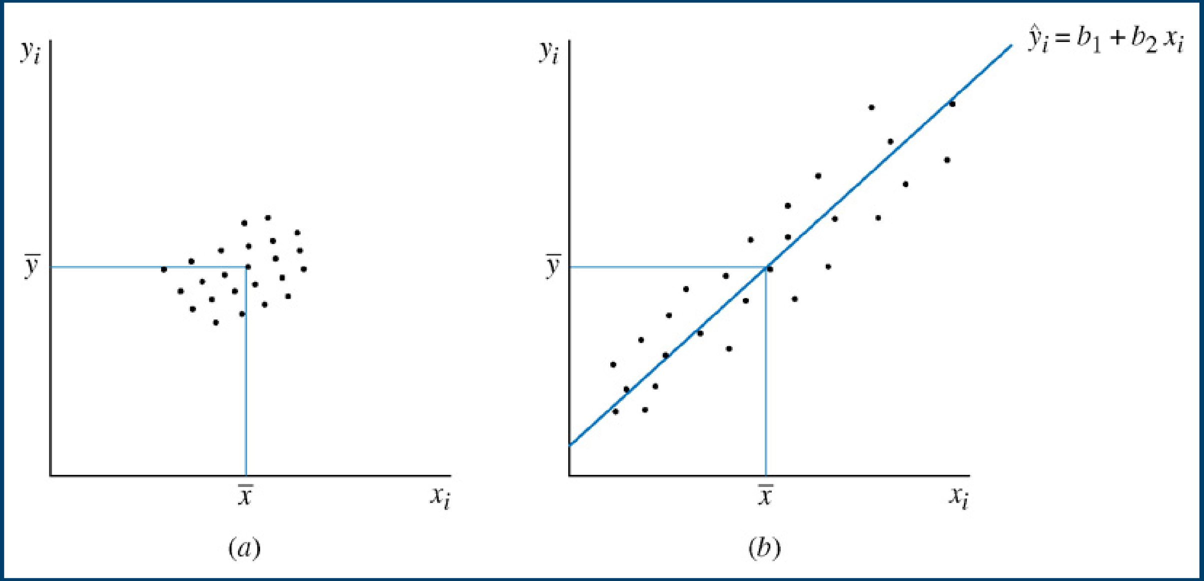
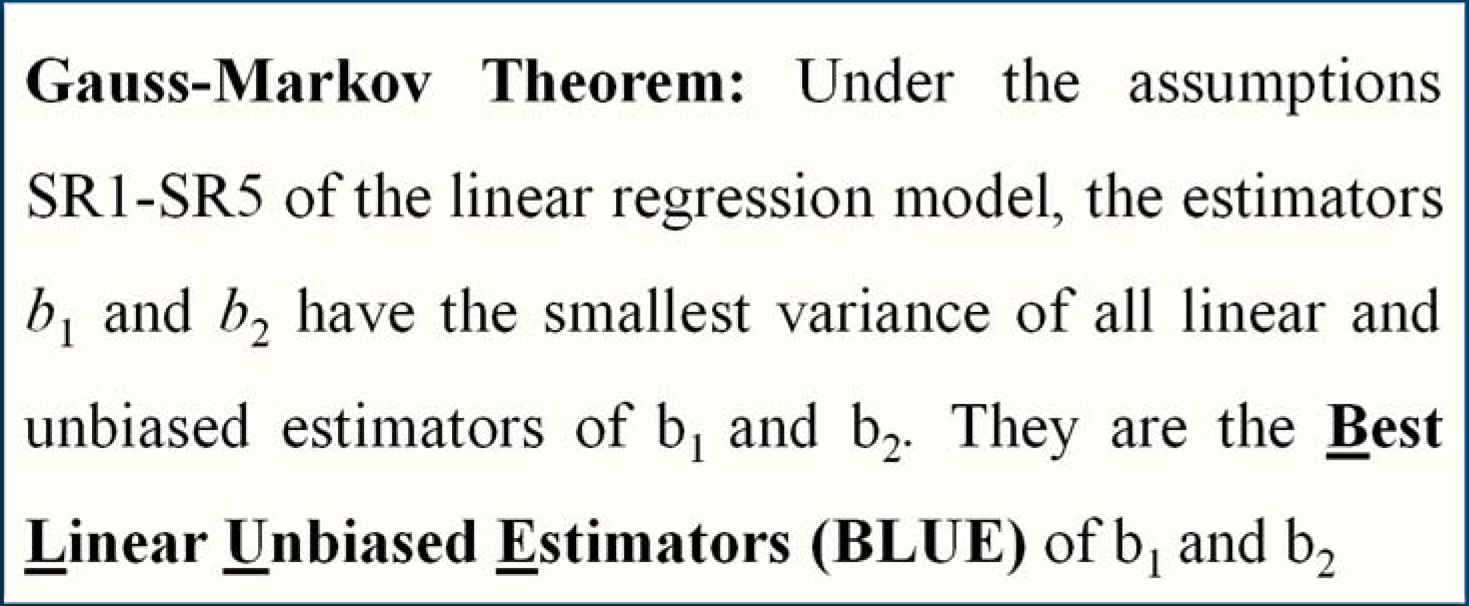


Figure 2.11 The influence of variation in the explanatory variable *x* on precision of estimation (a) Low *x* variation, low precision (b) High *x* variation, high precision

2.5 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

2.5 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.

2.5 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.

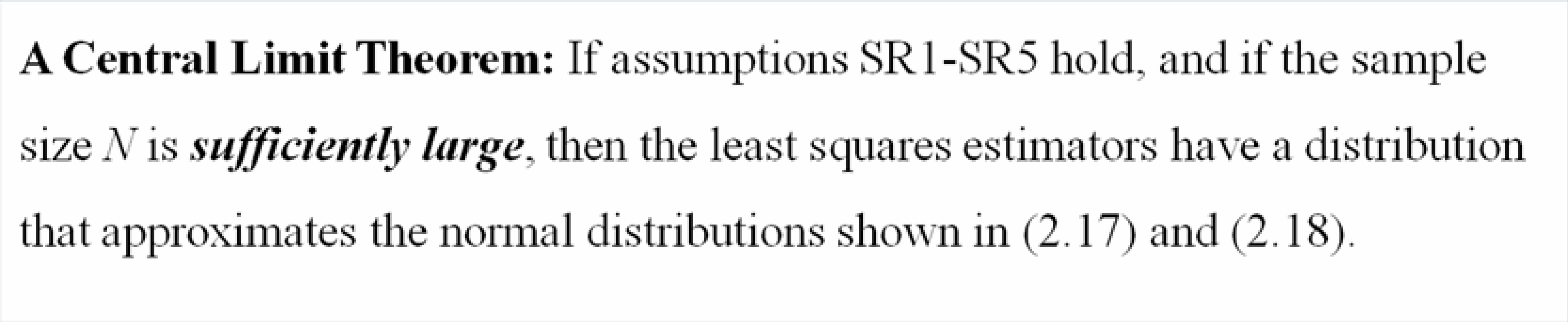
6. The Gauss-Markov theorem applies to the least squares estimators. It *does not* apply to the least squares *estimates* from a single sample.

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

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

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



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).

2.7 Estimating the Variance of the Error Term

The variance of the random error e_i is

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

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,

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

Recall that the random errors are

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

2.7 Estimating the Variance of the Error Term

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

$$e_i\text{垐} = y_i - y_i = y_i - b_1 - b_2 x_i$$

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

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

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

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

2.7.1 Estimating the Variances and Covariances of the Least Squares Estimators

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

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

$$\widehat{\operatorname{var}}\left(b_2\right) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} \qquad (2.21)$$

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

2.7.1 Estimating the Variances and Covariances of the Least Squares Estimators

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

$$\text{se}(b_1) = \sqrt{\overset{\square}{\text{var}}(b_1)} \tag{2.23}$$

$$\text{se}(b_2) = \sqrt{\overset{\square}{\text{var}}(b_2)} \tag{2.24}$$

2.7.2 Calculations for the Food Expenditure Data

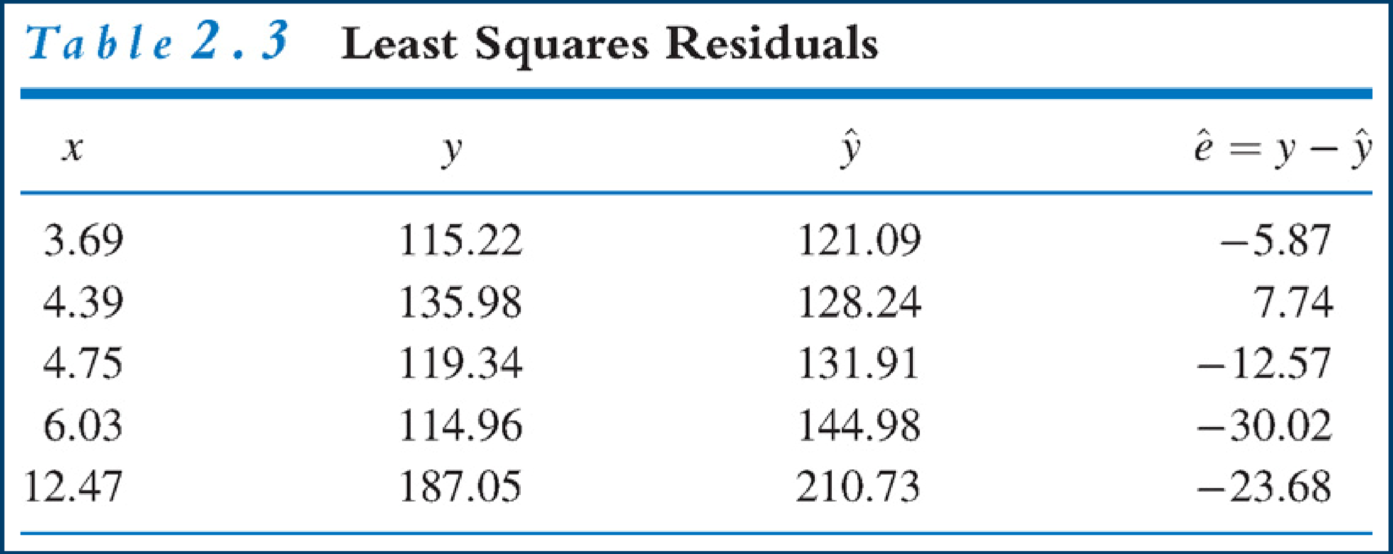


Table 2.3 **Least Squares Residuals**

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

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

2.7.2 Calculations for the Food Expenditure Data

- 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.

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

2.7.2 Calculations for the Food Expenditure Data

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

	C	*INCOME*
C	1884.442	-85.90316
INCOME	-85.90316	4.381752

2.7.2 Calculations for the Food Expenditure Data

$$\widehat{\text{var}}\left(b_1\right)=1884.442$$

$$\widehat{\text{var}}\left(b_2\right)=4.381752$$

$$\widehat{\text{cov}}\left(b_1,b_2\right)=-85.90316$$

$$\text{se}\left(b_1\right)=\sqrt{\widehat{\text{var}}\left(b_1\right)}=\sqrt{1884.442}=43.410$$

$$\text{se}\left(b_2\right)=\sqrt{\widehat{\text{var}}\left(b_2\right)}=\sqrt{4.381752}=2.093$$

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

Chapter 2 Appendices

- Appendix 2A Derivation of the least squares estimates
- Appendix 2B Deviation from the mean form of b_2
- Appendix 2C b_2 is a linear estimator
- Appendix 2D Derivation of Theoretical Expression for b_2
- Appendix 2E Deriving the variance of b_2
- Appendix 2F Proof of the Gauss-Markov Theorem

Appendix 2A
Derivation of the least squares estimates

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

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

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

Appendix 2A
Derivation of the least squares estimates

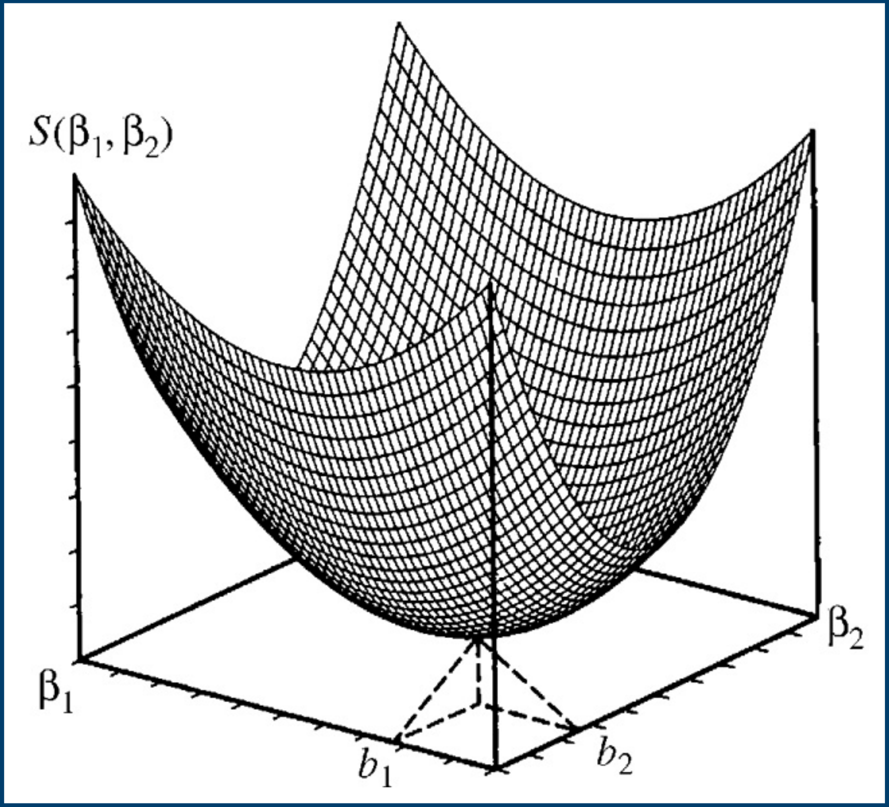


Figure **2A.1** The sum of squares function and the minimizing values b_1 and b_2

Appendix 2A
Derivation of the least squares estimates

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

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

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

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

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

Appendix 2B
Deviation From The Mean Form of b_2

$$\sum (x_i - \overline{x})^2 = \sum x_i^2 - 2\overline{x}\sum x_i + N\,\overline{x}^2 = \sum x_i^2 - 2\overline{x}\left(N\frac{1}{N}\sum x_i \right) + N\,\overline{x}^2$$

$$= \sum x_i^2 - 2N\,\overline{x}^2 + N\,\overline{x}^2 = \sum x_i^2 - N\,\overline{x}^2 \tag{2B.1}$$

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

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

Appendix 2B
Deviation From The Mean Form of b_2

We can rewrite b_2 in deviation from the mean form as:

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

Appendix 2C
b_2 is a Linear Estimator

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

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

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

Appendix 2D
Derivation of Theoretical Expression for b_2

To obtain (2.12) replace y_i in (2.11) by $y_i = \beta_1 + \beta_2 x_i + e_i$ and simplify:

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

$$= \beta_1 \sum w_i + \beta_2 \sum w_i x_i + \sum w_i e_i$$

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

Appendix 2D
Derivation of Theoretical Expression for *b*$_2$

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

$$\sum w_i x_i = 1$$

$$\beta_2 \sum w_i x_i = \beta_2$$

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

Appendix 2D
Derivation of Theoretical Expression for *b*$_2$

$$\sum \left(x_i - \overline{x}\right)^2 = \sum \left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right)$$

$$= \sum \left(x_i - \overline{x}\right)x_i - \overline{x}\sum \left(x_i - \overline{x}\right)$$

$$= \sum \left(x_i - \overline{x}\right)x_i$$

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

Appendix 2E
Deriving the Variance of b_2

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

$$\operatorname{var}(b_2) = E\left[b_2 - E(b_2)\right]^2$$

Appendix 2E
Deriving the Variance of *b*$_2$

$$\operatorname{var}(b_2) = E\left[\beta_2 + \sum w_i e_i - \beta_2\right]^2$$

$$= E\left[\sum w_i e_i\right]^2$$

$$= E\left[\sum w_i^2 e_i^2 + 2\sum\sum_{i \neq j} w_i w_j e_i e_j\right] \qquad \text{[square of bracketed term]}$$

$$= \sum w_i^2 E\left(e_i^2\right) + 2\sum\sum_{i \neq j} w_i w_j E\left(e_i e_j\right) \qquad \text{[because } w_i \text{ not random]}$$

$$= \sigma^2 \sum w_i^2$$

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

Appendix 2E
Deriving the Variance of *b*$_2$

$$\sigma^2 = \operatorname{var}(e_i) = E\left[e_i - E(e_i)\right]^2 = E\left[e_i - 0\right]^2 = E\left(e_i^2\right)$$

$$\operatorname{cov}(e_i, e_j) = E\Big[\big(e_i - E(e_i)\big)\big(e_j - E(e_j)\big)\Big] = E\left(e_i e_j\right) = 0$$

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

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

Appendix 2E
Deriving the Variance of b_2

$$\operatorname{var}(b_2) = \operatorname{var}\left(\beta_2 + \sum w_i e_i\right) \qquad \text{[since } \beta_2 \text{ is a constant]}$$

$$= \sum w_i^2 \operatorname{var}(e_i) + \sum_{i \neq j}\sum w_i w_j \operatorname{cov}\left(e_i, e_j\right) \qquad \text{[generalizing the variance rule]}$$

$$= \sum w_i^2 \operatorname{var}(e_i) \qquad \text{[using } \operatorname{cov}\left(e_i, e_j\right) = 0\text{]}$$

$$= \sigma^2 \sum w_i^2 \qquad \text{[using } \operatorname{var}(e_i) = \sigma^2\text{]}$$

$$= \frac{\sigma^2}{\sum (x_i - \overline{x})^2}$$

Appendix 2F
Proof of the Gauss-Markov Theorem

- Let $b_2^* = \sum k_i y_i$ be any other linear estimator of β$_2$.
- Suppose that $k_i = w_i + c_i$.

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

Appendix 2F
Proof of the Gauss-Markov Theorem

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

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

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

Appendix 2F
Proof of the Gauss-Markov Theorem

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

$$\operatorname{var}\left(b_2^* \right) = \operatorname{var}\left[\beta_2 + \sum \left(w_i + c_i \right) e_i \right] = \sum \left(w_i + c_i \right)^2 \operatorname{var}\left(e_i \right)$$

$$= \sigma^2 \sum \left(w_i + c_i \right)^2 = \sigma^2 \sum w_i^2 + \sigma^2 \sum c_i^2$$

$$= \operatorname{var}\left(b_2 \right) + \sigma^2 \sum c_i^2$$

$$\geq \operatorname{var}\left(b_2 \right)$$