

Hồi quy tuyến tính một biến (Simple linear regression)

Đặng Thanh Hải (Ph.D)

School of Engineering and Technology, VNUH

Email: hai.dang@vnu.edu.vn

Estimation of Random Variables

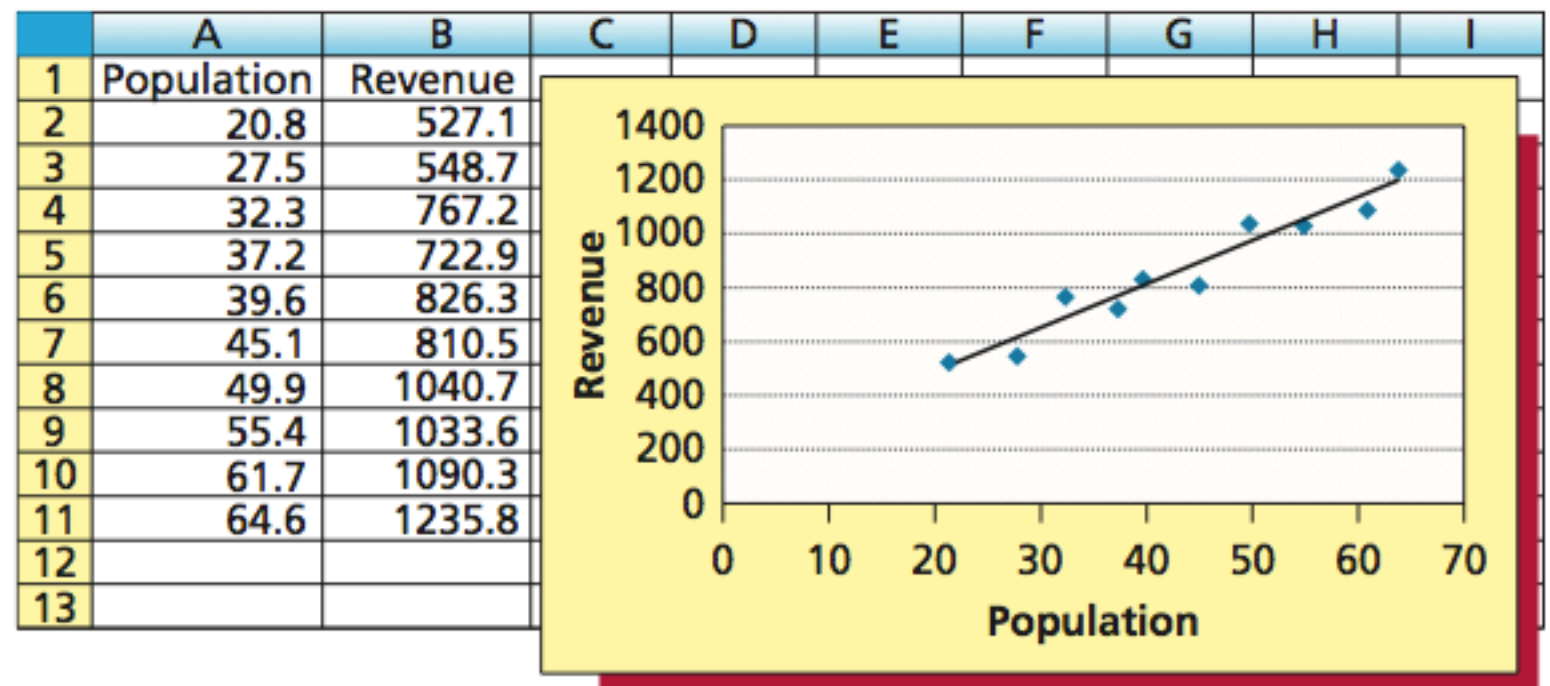
- ◆ Very often, we want to estimate the value of an inaccessible random variable Y (**dependent or response variable**) in terms of the observation of an accessible random variable X (**independent or predictor variable**)

TABLE 14.1 The Tasty Sub Shop Revenue Data

DS TastySub1

Restaurant	Population Size, x (Thousands of Residents)	Yearly Revenue, y (Thousands of Dollars)
1	20.8	527.1
2	27.5	548.7
3	32.3	767.2
4	37.2	722.9
5	39.6	826.3
6	45.1	810.5
7	49.9	1040.7
8	55.4	1033.6
9	61.7	1090.3
10	64.6	1235.8

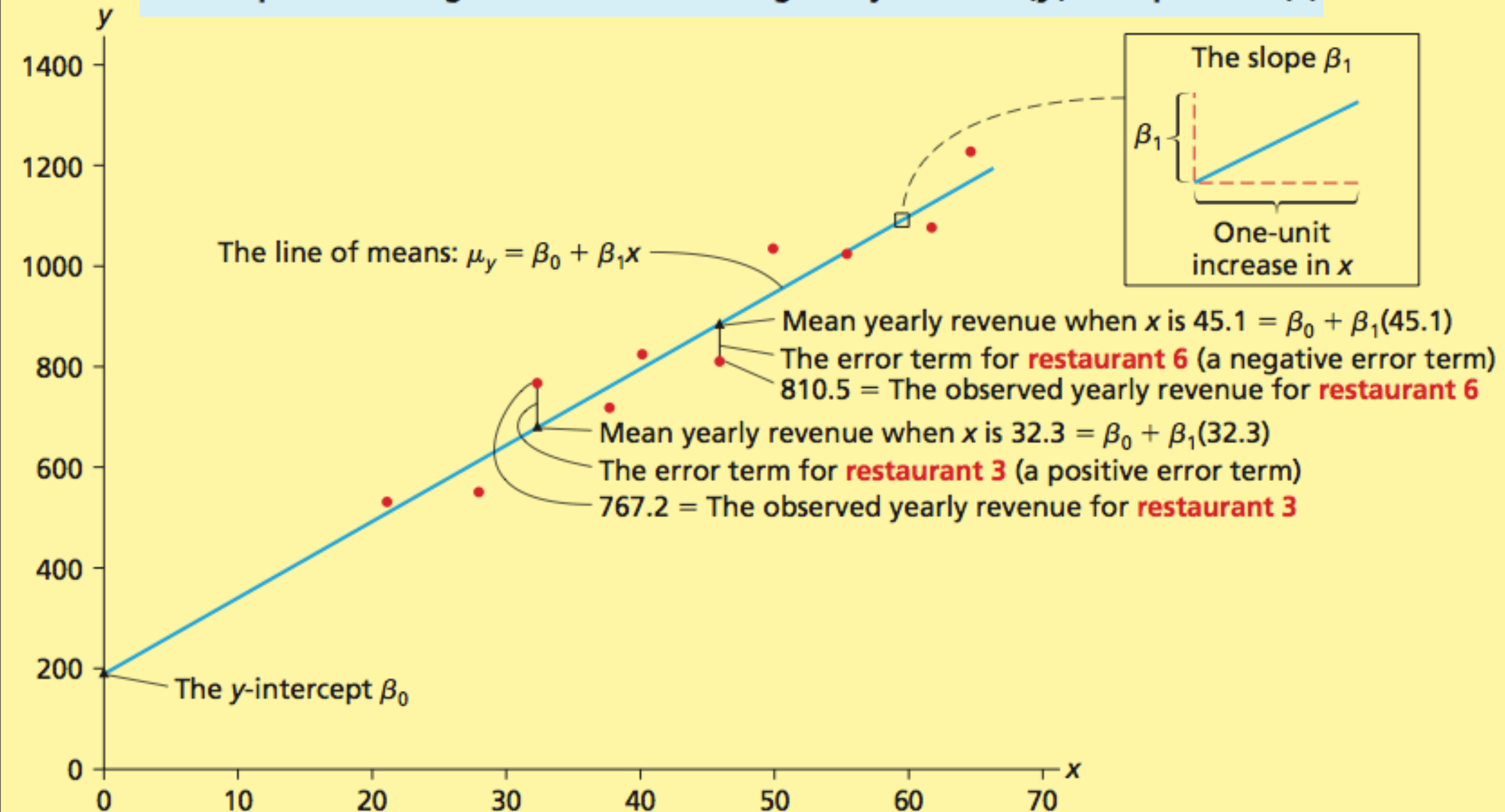
FIGURE 14.1 Excel Output of a Scatter Plot of y versus x



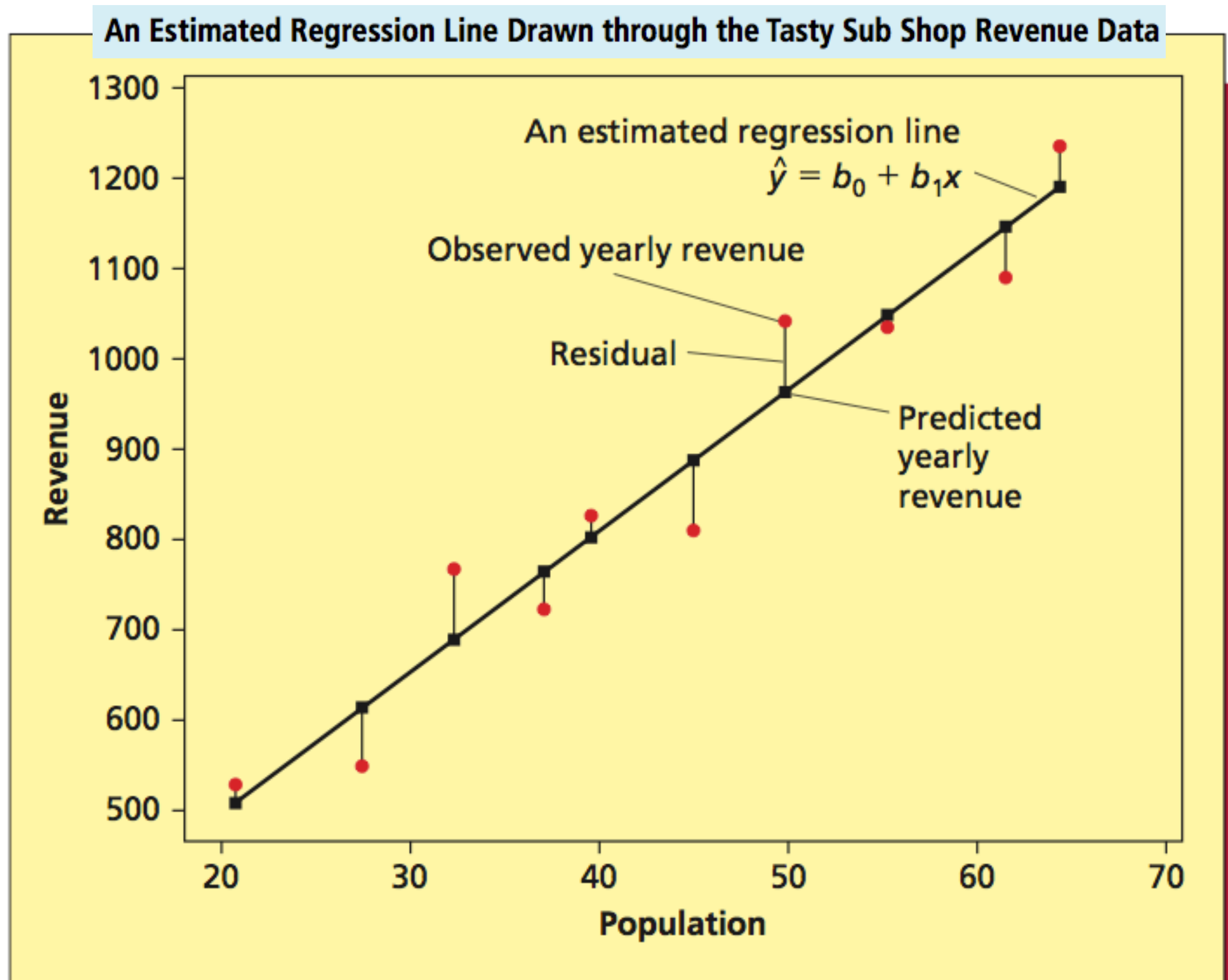
Mô hình hồi quy tuyến tính đơn giản

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The Simple Linear Regression Model Relating Yearly Revenue (y) to Population (x)



Xây dựng mô hình hồi quy tuyến tính đơn giản



Xây dựng mô hình hồi quy tuyến tính đơn giản

- ◆ The least squares point estimates (Các ước lượng điểm bình phương tối thiểu)
- ✓ Predicted value of the dependent variable y (Giá trị dự đoán của biến phụ thuộc y)

$$\hat{y}_i = b_0 + b_1 x_i$$

- ✓ Minimize sum of squared residuals (Cực tiểu hoá tổng các lỗi bình phương)

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Xây dựng mô hình hồi quy tuyến tính đơn giản

The Least Squares Point Estimates

For the simple linear regression model:

- 1** The least squares point estimate of the slope β_1 is

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{where}$$

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} \quad \text{and} \quad SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

- 2** The least squares point estimate of the y-intercept β_0 is

$$b_0 = \bar{y} - b_1 \bar{x} \quad \text{where}$$

$$\bar{y} = \frac{\sum y_i}{n} \quad \text{and} \quad \bar{x} = \frac{\sum x_i}{n}$$

Here n is the number of observations (an observation is an observed value of x and its corresponding value of y).

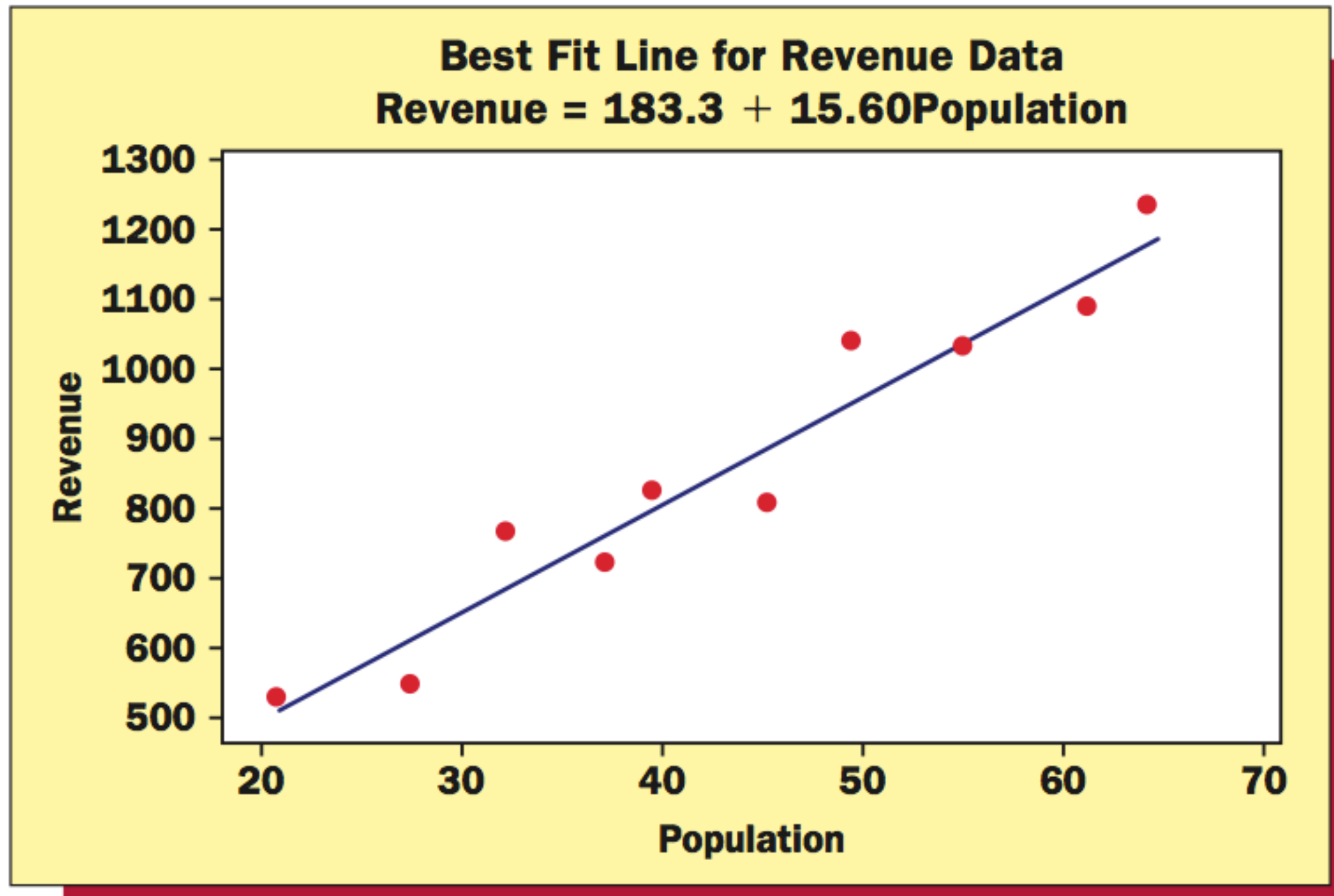
Xây dựng mô hình hồi quy tuyến tính đơn giản

◆ The Tasty Sub Shop Case

y_i	x_i	x_i^2	$x_i y_i$
527.1	20.8	$(20.8)^2 = 432.64$	$(20.8)(527.1) = 10963.68$
548.7	27.5	$(27.5)^2 = 756.25$	$(27.5)(548.7) = 15089.25$
767.2	32.3	$(32.3)^2 = 1,043.29$	$(32.3)(767.2) = 24780.56$
722.9	37.2	$(37.2)^2 = 1,383.84$	$(37.2)(722.9) = 26891.88$
826.3	39.6	$(39.6)^2 = 1,568.16$	$(39.6)(826.3) = 32721.48$
810.5	45.1	$(45.1)^2 = 2,034.01$	$(45.1)(810.5) = 36553.55$
1040.7	49.9	$(49.9)^2 = 2,490.01$	$(49.9)(1040.7) = 51930.93$
1033.6	55.4	$(55.4)^2 = 3,069.16$	$(55.4)(1033.6) = 57261.44$
1090.3	61.7	$(61.7)^2 = 3,806.89$	$(61.7)(1090.3) = 67271.51$
1235.8	64.6	$(64.6)^2 = 4,173.16$	$(64.6)(1235.8) = 79832.68$
<hr/> $\sum y_i = 8603.1$	<hr/> $\sum x_i = 434.1$	<hr/> $\sum x_i^2 = 20,757.41$	<hr/> $\sum x_i y_i = 403,296.96$

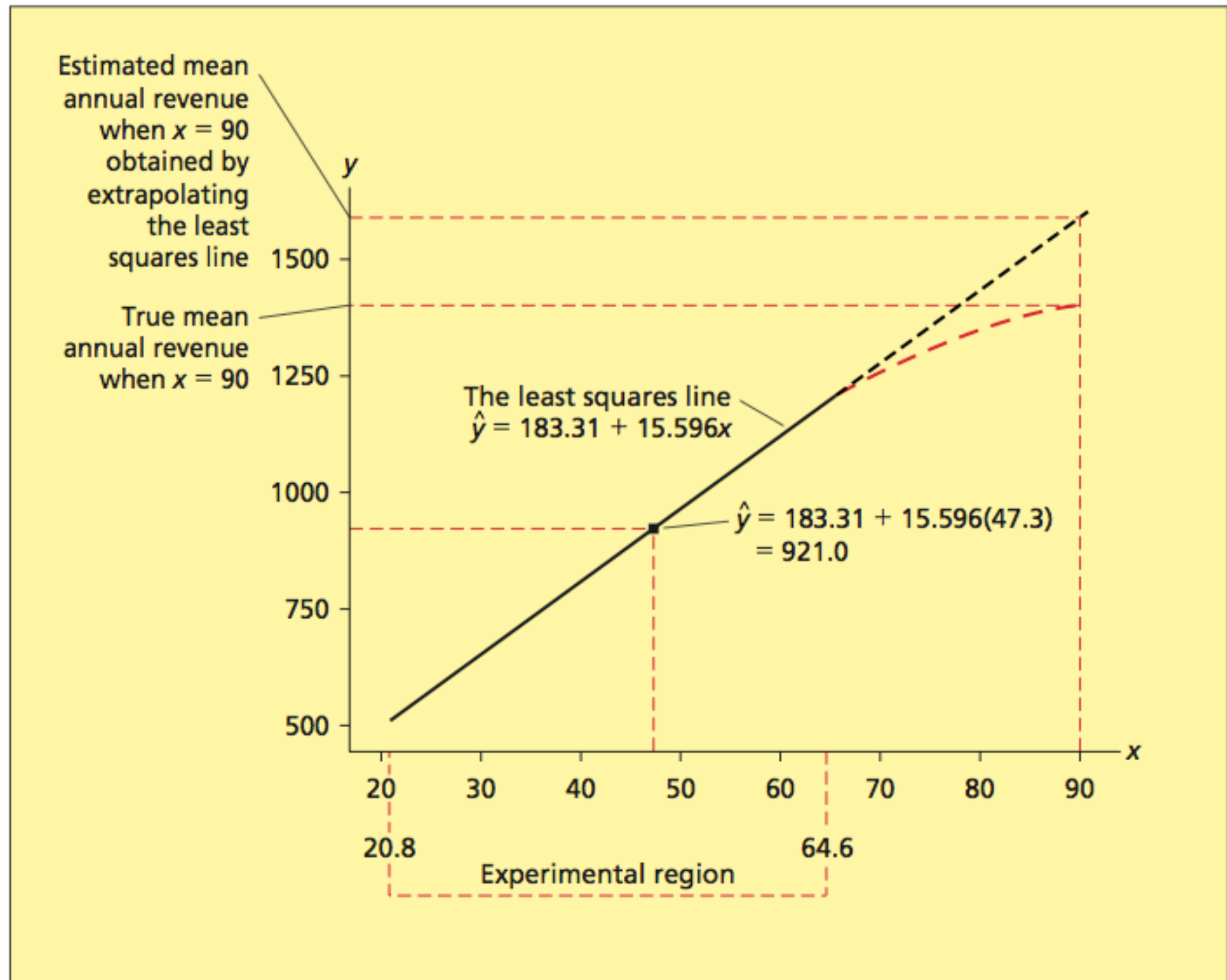
Xây dựng mô hình hồi quy tuyến tính đơn giản

◆ The Tasty Sub Shop Case



Sử dụng mô hình hồi quy tuyến tính đơn giản

◆ Point Estimation and Point Prediction, and the Danger of Extrapolation



Sử dụng mô hình hồi quy tuyến tính đơn giản

◆ Point Estimation and Point Prediction, and the Danger of Extrapolation

Let b_0 and b_1 be the least squares point estimates of the y -intercept β_0 and the slope β_1 in the simple linear regression model, and suppose that x_0 , a specified value of the independent variable x , is inside the experimental region. Then

$$\hat{y} = b_0 + b_1x_0$$

- 1** is the **point estimate** of the **mean value** of the **dependent variable** when the value of the independent variable is x_0 .
- 2** is the **point prediction** of an **individual value** of the **dependent variable** when the value of the independent variable is x_0 . Here we predict the error term to be 0.

Sử dụng mô hình hồi quy tuyến tính đơn giản

♦ Các yêu cầu để có thể kiểm định giả thuyết và xây dựng khoảng tin cậy khi sử dụng mô hình

1 At any given value of x , the population of potential error term values has a **mean equal to 0**.

2 Constant Variance Assumption

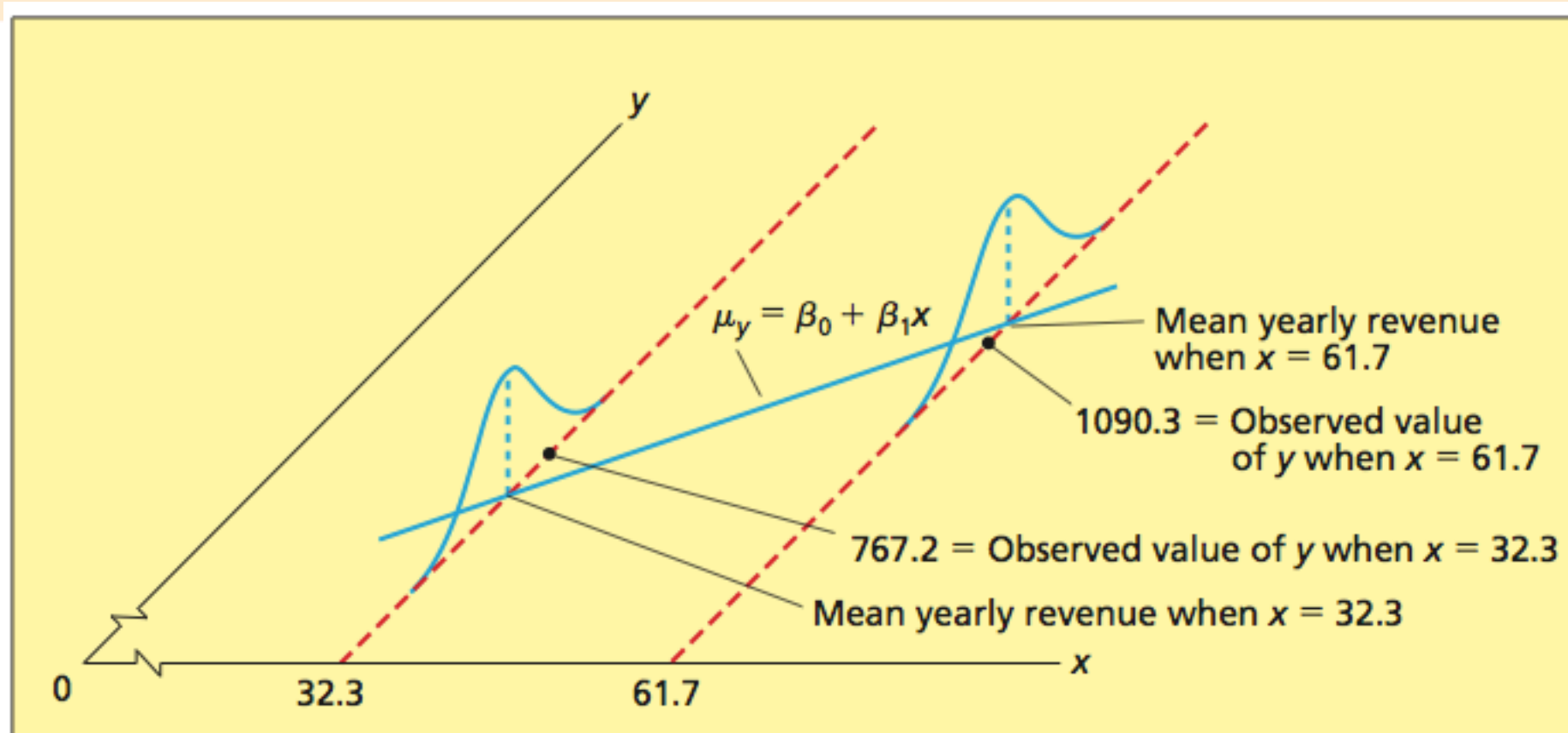
At any given value of x , the population of potential error term values has a variance that does not depend on the value of x . That is, the different populations of potential error term values corresponding to different values of x have **equal variances**. We denote the **constant variance** as σ^2 .

3 Normality Assumption

At any given value of x , the population of potential error term values has a **normal distribution**.

4 Independence Assumption

Any one value of the error term ε is **statistically independent** of any other value of ε . That is, the value of the error term ε corresponding to an observed value of y is statistically independent of the value of the error term corresponding to any other observed value of y .



Sử dụng mô hình hồi quy tuyệt tính đơn giản

- ♦ Lỗi kỳ vọng bình phương (**mean square error**) và Lỗi chuẩn (**standard error**) của mô hình

If the regression assumptions are satisfied and SSE is the sum of squared residuals:

1 The point estimate of σ^2 is the **mean square error**

$$s^2 = \frac{SSE}{n - 2}$$

2 The point estimate of σ is the **standard error**

$$s = \sqrt{\frac{SSE}{n - 2}}$$

Trong đó:

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$n - 2$ là **số bậc tự do** gắn với SSE .

Sử dụng mô hình hồi quy tuyệt tính đơn giản

- ◆ Mô hình chỉ thực sự có ý nghĩa khi thực sự (về mặt thống kê) tồn tại mối quan hệ giữa y và x .
 - ✓ Kiểm định mức ý nghĩa (thống kê) của hệ số góc và hệ số chặn (Testing the Significance of the Slope and y-Intercept)
 - Kiểm định: $H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$
- ◆ Khi giả thuyết cho mô hình hồi quy tuyến tính đúng
 - ✓ Tất cả giá trị b_1 phân phối chuẩn với kỳ vọng β_1 và độ lệch chuẩn $\sigma_{b_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$
 - Do lỗi chuẩn s là ước lượng điểm của σ , nên ước lượng điểm của σ_{b_1} là:
$$s_{b_1} = \frac{s}{\sqrt{SS_{xx}}} \quad (\text{Sai số chuẩn của ước lượng } b_1)$$

Sử dụng mô hình hồi quy tuyệt tính đơn giản

- ◆ Do đó, khi giả thuyết cho mô hình hồi quy tuyến tính đúng

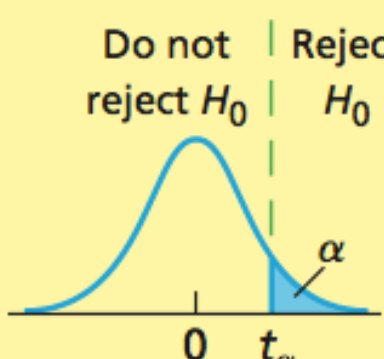
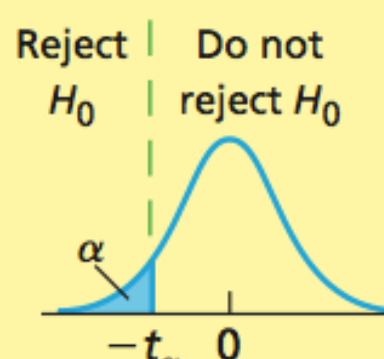
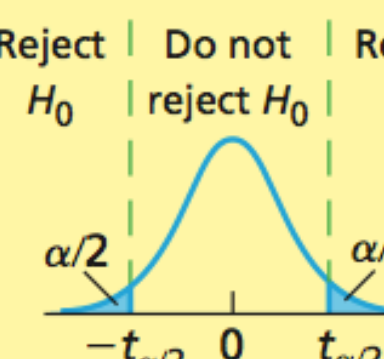
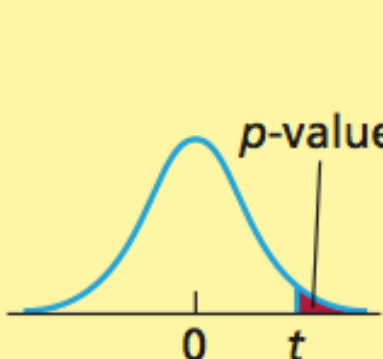
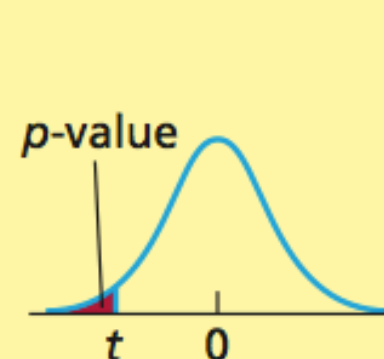
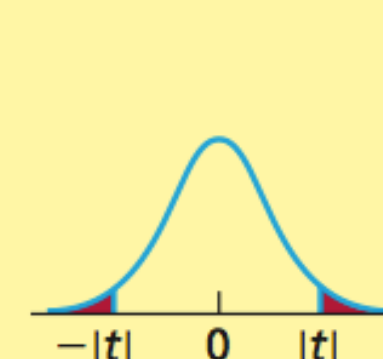
Các giá trị $\frac{b_1 - \beta_1}{s_{b_1}}$ tuân theo **phân bố t** với $n-2$ bậc tự do

- ◆ Do đó, khi $H_0: \beta_1 = 0$ đúng:

Sử dụng thống kê: $t = \frac{b_1}{s_{b_1}}$, tuân theo **phân bố t** với $n-2$ bậc tự do

Kiểm định mối quan hệ hồi quy tuyến tính

Null Hypothesis	$H_0: \beta_1 = 0$	Test Statistic	$t = \frac{b_1}{s_{b_1}}$ where $s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$	Assumptions	The regression assumptions
------------------------	--------------------	-----------------------	--	--------------------	----------------------------

Critical Value Rule			p-Value (Reject H_0 if p-Value $< \alpha$)		
$H_a: \beta_1 > 0$	$H_a: \beta_1 < 0$	$H_a: \beta_1 \neq 0$	$H_a: \beta_1 > 0$	$H_a: \beta_1 < 0$	$H_a: \beta_1 \neq 0$
 <p>Do not reject H_0 Reject H_0</p> <p>Reject H_0 if $t > t_\alpha$</p>	 <p>Reject H_0 Do not reject H_0</p> <p>Reject H_0 if $t < -t_\alpha$</p>	 <p>Reject H_0 Do not reject H_0 Reject H_0</p> <p>Reject H_0 if $t > t_{\alpha/2}$—that is, $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$</p>	 <p>p-value</p> <p>p-value = area to the right of t</p>	 <p>p-value</p> <p>p-value = area to the left of t</p>	 <p>p-value = twice the area to the right of t</p>

Here $t_{\alpha/2}$, t_α , and all p -values are based on $n - 2$ degrees of freedom. If we can reject $H_0: \beta_1 = 0$ at a given value of α , then we conclude that the slope (or, equivalently, the regression relationship) is significant at the α level.

A Confidence Interval for the Slope

If the regression assumptions hold, a $100(1 - \alpha)$ percent confidence interval for the true slope β_1 is $[b_1 \pm t_{\alpha/2} s_{b_1}]$. Here $t_{\alpha/2}$ is based on $n - 2$ degrees of freedom.

Kiểm định mức ý nghĩa (thống kê) của hệ số chặn y

$$H_0: \beta_0 = 0$$

$$H_a: \beta_0 \neq 0.$$

Sử dụng thống kê:

$$t = \frac{b_0}{s_{b_0}} \quad \text{where} \quad s_{b_0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$$

t tuân theo **phân bố t** với $n-2$ bậc tự do

A Confidence Interval and a Prediction Interval

If the regression assumptions hold,

- 1 A $100(1 - \alpha)$ percent confidence interval for the mean value of y when x equals x_0 is

$$\left[\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} \right]$$

- 2 A $100(1 - \alpha)$ percent prediction interval for an individual value of y when x equals x_0 is

$$\left[\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} \right]$$

Here, $t_{\alpha/2}$ is based on $(n - 2)$ degrees of freedom.

Tính hiệu quả của mô hình

- ◆ Được đo thông qua “hệ số quyết định đơn giản” (simple coefficient of determination) r^2

For the simple linear regression model

- 1** Total variation = $\sum (y_i - \bar{y})^2$
- 2** Explained variation = $\sum (\hat{y}_i - \bar{y})^2$
- 3** Unexplained variation = $\sum (y_i - \hat{y}_i)^2$
- 4** Total variation = Explained variation + Unexplained variation

5 The simple coefficient of determination is

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

- 6** r^2 is the proportion of the total variation in the n observed values of the dependent variable that is explained by the simple linear regression model.

◆ Hệ số tương quan

The simple correlation coefficient between y and x , denoted by r , is

$$r = +\sqrt{r^2} \text{ if } b_1 \text{ is positive} \quad \text{and} \quad r = -\sqrt{r^2} \text{ if } b_1 \text{ is negative}$$

where b_1 is the slope of the least squares line relating y to x . This correlation coefficient measures the strength of the linear relationship between y and x .

Kiểm định F-Test cho mô hình

- ♦ Một cách khác để kiểm định độ ý nghĩa của mô hình hồi quy tuyến tính

Suppose that the regression assumptions hold, and define the **overall F statistic** to be

$$F(\text{model}) = \frac{\text{Explained variation}}{(\text{Unexplained variation})/(n - 2)}$$

Also define the p -value related to $F(\text{model})$ to be the area under the curve of the F distribution (having 1 numerator and $n - 2$ denominator degrees of freedom) to the right of $F(\text{model})$ —see Figure 14.20(b).


We can reject $H_0: \beta_1 = 0$ in favor of $H_a: \beta_1 \neq 0$ at level of significance α if either of the following equivalent conditions holds:

- 1 $F(\text{model}) > F_\alpha$
- 2 $p\text{-value} < \alpha$

Here the point F_α is based on 1 numerator and $n - 2$ denominator degrees of freedom.

Hồi quy đa biến (Multiple linear regression)

A case study

TABLE 15.1 The Tasty Sub Shop Revenue Data 

Restaurant	Population Size, x_1 (Thousands of Residents)	Business Rating, x_2	Yearly Revenue, y (Thousands of Dollars)
1	20.8	3	527.1
2	27.5	2	548.7
3	32.3	6	767.2
4	37.2	5	722.9
5	39.6	8	826.3
6	45.1	3	810.5
7	49.9	9	1040.5
8	55.4	5	1033.6
9	61.7	4	1090.3
10	64.6	7	1235.8

FIGURE 15.1 Plot of y (Yearly Revenue) versus x_1 (Population Size)

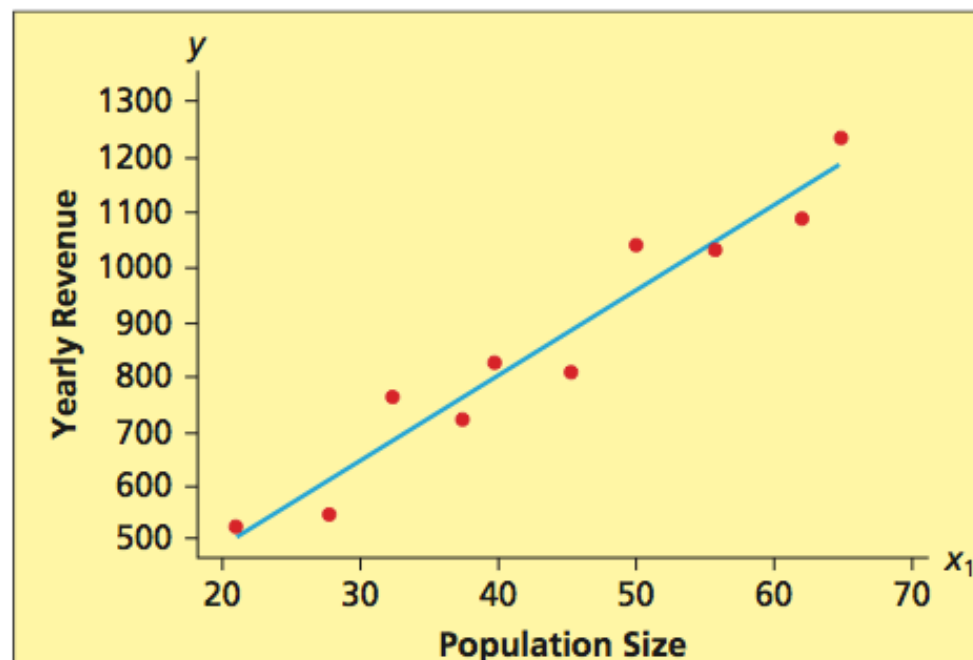
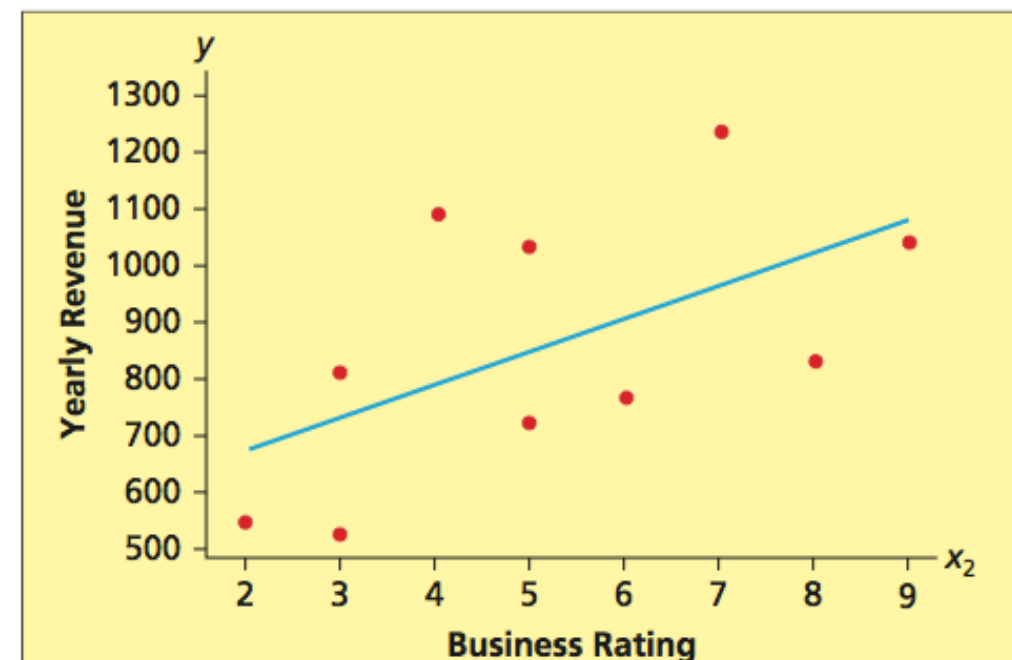
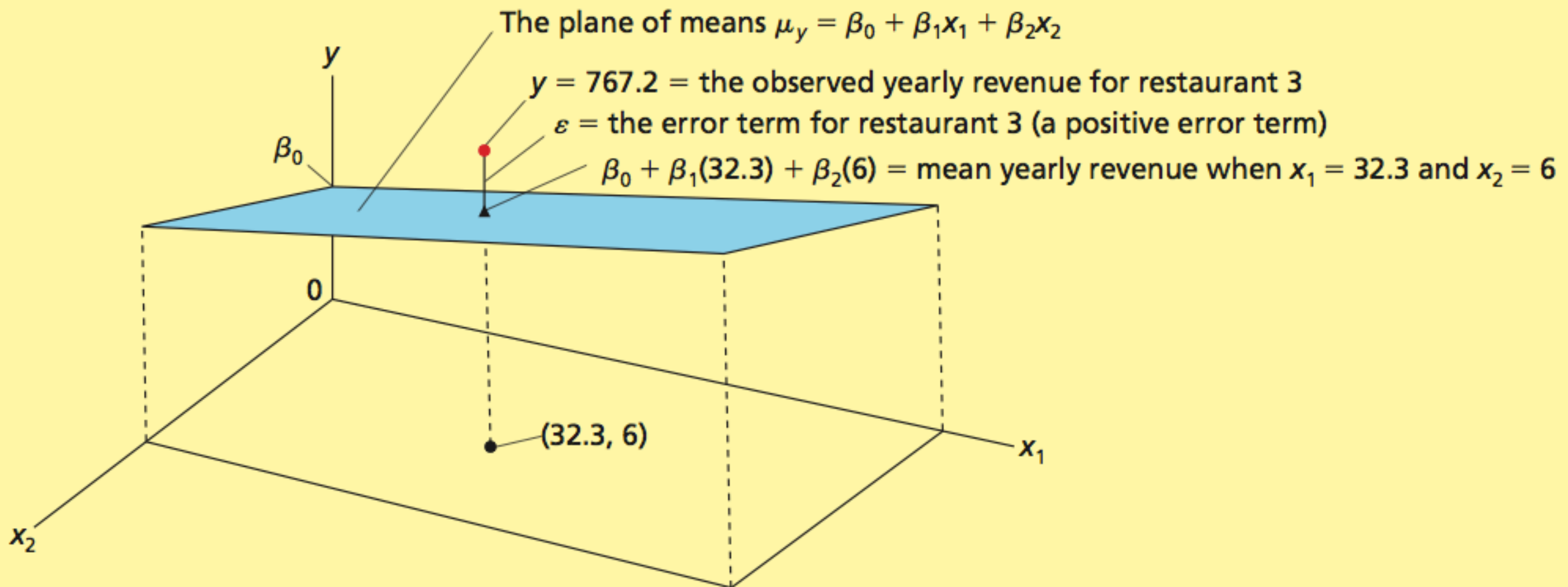


FIGURE 15.2 Plot of y (Yearly Revenue) versus x_2 (Business Rating)



A case study

A Geometrical Interpretation of the Regression Model Relating y to x_1 and x_2



A case study

TABLE 15.2 The Point Predictions and Residuals Using the Least Squares Point Estimates, $b_0 = 125.29$, $b_1 = 14.1996$, and $b_2 = 22.811$

Restaurant	Population Size, x_1 (Thousands of Residents)	Business Rating, x_2	Yearly Revenue, y (Thousands of Dollars)	Predicted Yearly Revenue $\hat{y} = 125.29 + 14.1996x_1$ $+ 22.811x_2$	Residual, $y - \hat{y}$
1	20.8	3	527.1	489.07	38.03
2	27.5	2	548.7	561.40	-12.70
3	32.3	6	767.2	720.80	46.40
4	37.2	5	722.9	767.57	-44.67
5	39.6	8	826.3	870.08	-43.78
6	45.1	3	810.5	834.12	-23.62
7	49.9	9	1040.7	1039.15	1.55
8	55.4	5	1033.6	1026.00	7.60
9	61.7	4	1090.3	1092.65	-2.35
10	64.6	7	1235.8	1202.26	33.54

$$SSE = (38.03)^2 + (-12.70)^2 + \cdots + (33.54)^2 = 9420.8$$

The multiple regression model relating y to x_1, x_2, \dots, x_k is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Here

- $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is the mean value of the dependent variable y when the values of the independent variables are x_1, x_2, \dots, x_k .
- $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are (unknown) **regression parameters** relating the mean value of y to x_1, x_2, \dots, x_k .
- ε is an **error term** that describes the effects on y of all factors other than the values of the independent variables x_1, x_2, \dots, x_k .

Hồi quy tuyến tính đa biến

Assumptions for the Multiple Regression Model

- 1** At any given combination of values of x_1, x_2, \dots, x_k , the population of potential error term values has a mean equal to 0.
- 2** **Constant variance assumption:** At any given combination of values of x_1, x_2, \dots, x_k , the population of potential error term values has a variance that does not depend on the combination of values of x_1, x_2, \dots, x_k . That is, the different populations of potential error term values corresponding to different combinations of values of x_1, x_2, \dots, x_k have equal variances. We denote the constant variance as σ^2 .
- 3** **Normality assumption:** At any given combination of values of x_1, x_2, \dots, x_k , the population of potential error term values has a **normal distribution**.
- 4** **Independence assumption:** Any one value of the error term ε is **statistically independent** of any other value of ε . That is, the value of the error term ε corresponding to an observed value of y is statistically independent of the error term corresponding to any other observed value of y .

Hồi quy tuyến tính đa biến

- ◆ Lỗi kỳ vọng bình phương (mean square error) và Lỗi chuẩn (standard error) của mô hình

The Mean Square Error and the Standard Error

Suppose that the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

utilizes k independent variables and thus has $(k + 1)$ parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Then, if the regression assumptions are satisfied, and if SSE denotes the sum of squared residuals for the model:

- 1 A point estimate of σ^2 is the **mean square error**

$$s^2 = \frac{SSE}{n - (k + 1)}$$

- 2 A point estimate of σ is the **standard error**

$$s = \sqrt{\frac{SSE}{n - (k + 1)}}$$

Tính hiệu quả của mô hình

- ♦ Được đo thông qua “hệ số quyết định đơn giản” (simple coefficient of determination) r^2

The Mean Square Error and the Standard Error

Suppose that the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

utilizes k independent variables and thus has $(k + 1)$ parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Then, if the regression assumptions are satisfied, and if SSE denotes the sum of squared residuals for the model:

- 1 A point estimate of σ^2 is the **mean square error**

$$s^2 = \frac{SSE}{n - (k + 1)}$$

- 2 A point estimate of σ is the **standard error**

$$s = \sqrt{\frac{SSE}{n - (k + 1)}}$$

Adjusted R^2

The **adjusted multiple coefficient of determination (adjusted R^2)** is

$$\bar{R}^2 = \left(R^2 - \frac{k}{n - 1} \right) \left(\frac{n - 1}{n - (k + 1)} \right)$$

where R^2 is the multiple coefficient of determination, n is the number of observations, and k is the number of independent variables in the model under consideration.

Kiểm định F-test tổng thể cho mô hình

An F -Test for the Multiple Regression Model

Suppose that the regression assumptions hold and that the multiple regression model has $(k + 1)$ parameters, and consider testing

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

versus

H_a : At least one of $\beta_1, \beta_2, \dots, \beta_k$ does not equal 0.

We define the **overall F statistic** to be

$$F(\text{model}) = \frac{(\text{Explained variation})/k}{(\text{Unexplained variation})/[n - (k + 1)]}$$

Also define the p -value related to $F(\text{model})$ to be the area under the curve of the F distribution (having k and $[n - (k + 1)]$ degrees of freedom) to the right of $F(\text{model})$. Then, we can reject H_0 in favor of H_a at level of significance α if either of the following equivalent conditions holds:

1 $F(\text{model}) > F_\alpha$

2 $p\text{-value} < \alpha$

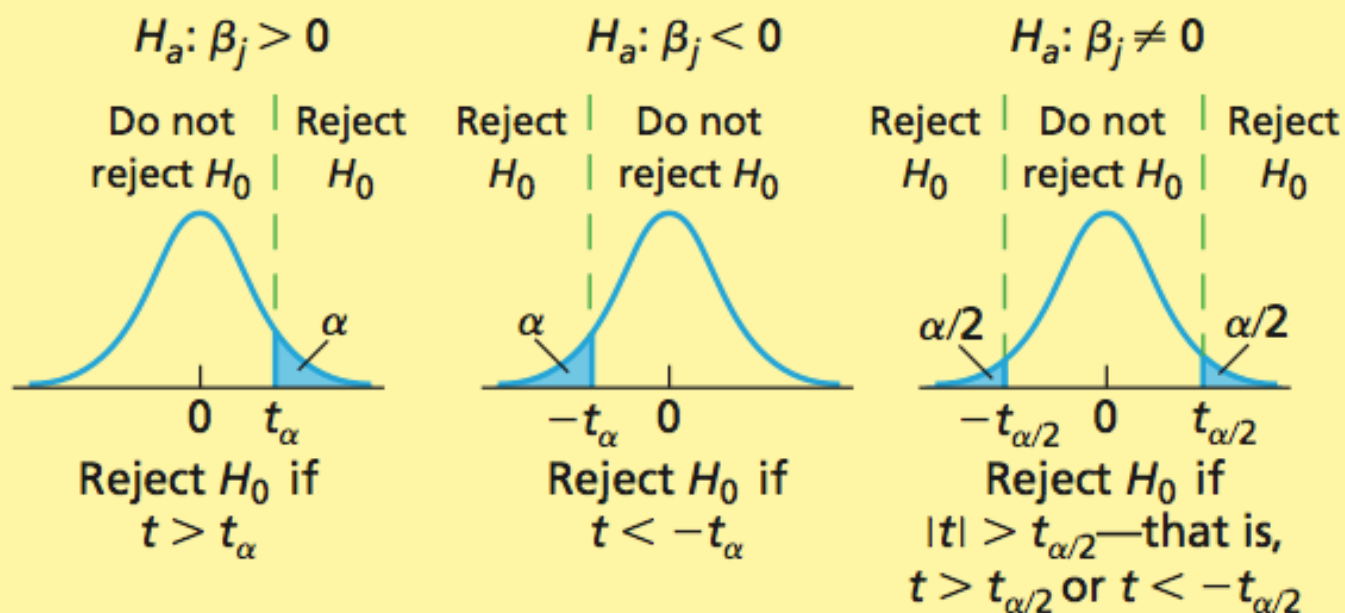
Here the point F_α is based on k numerator and $n - (k + 1)$ denominator degrees of freedom.

Kiểm định mức ý nghĩa của từng biến

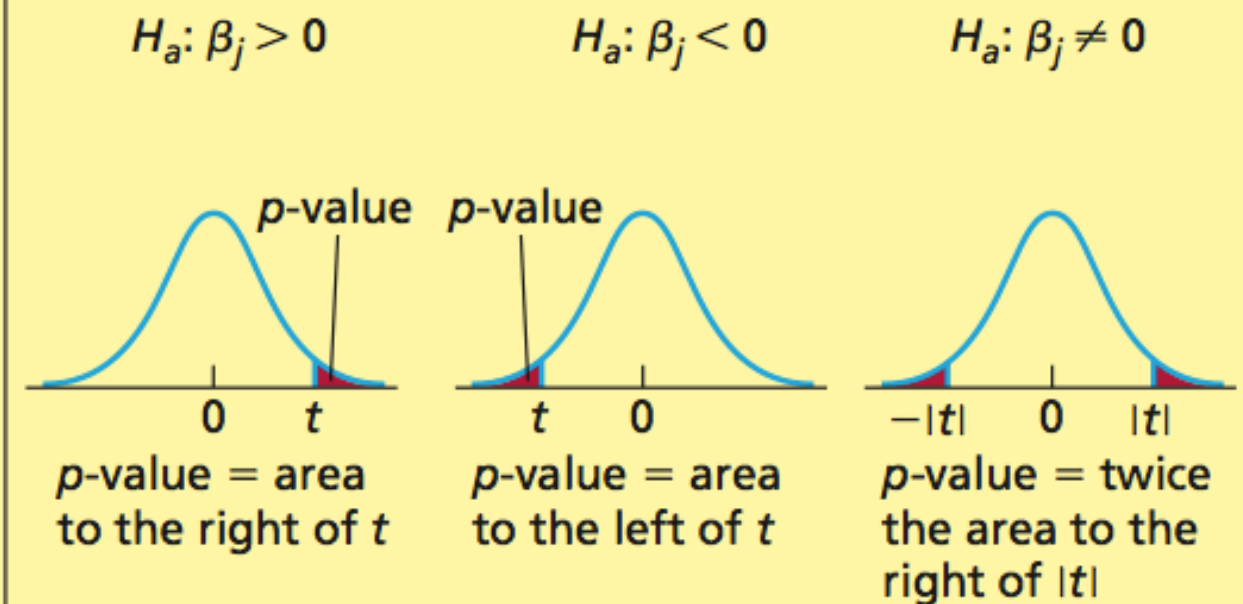
Testing the Significance of the Independent Variable x_j

Null Hypothesis	$H_0: \beta_j = 0$	Test Statistic	$t = \frac{b_j}{s_{b_j}}$	$df = n - (k + 1)$	Assumptions	The regression assumptions
------------------------	--------------------	-----------------------	---------------------------	--------------------	--------------------	----------------------------

Critical Value Rule



p -Value (Reject H_0 if p -Value $< \alpha$)



A Confidence Interval for the Regression Parameter β_j

If the regression assumptions hold, a $100(1 - \alpha)$ percent confidence interval for β_j is

$$[b_j \pm t_{\alpha/2} s_{b_j}]$$

Here $t_{\alpha/2}$ is based on $n - (k + 1)$ degrees of freedom.

A Confidence Interval and a Prediction Interval

If the regression assumptions hold,

- 1 A $100(1 - \alpha)$ percent confidence interval for the mean value of y when the values of the independent variables are x_1, x_2, \dots, x_k is

$$[\hat{y} \pm t_{\alpha/2} s \sqrt{\text{distance value}}]$$

- 2 A $100(1 - \alpha)$ percent prediction interval for an individual value of y when the values of the independent variables are x_1, x_2, \dots, x_k is

$$[\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \text{distance value}}]$$

Here $t_{\alpha/2}$ is based on $n - (k + 1)$ degrees of freedom and s is the standard error (see page 566). Furthermore, the formula for the **distance value** (also sometimes called the **leverage value**) involves matrix algebra and is given in Bowerman, O'Connell, and Koehler (2005). In practice, we can obtain the distance value from the outputs of statistical software packages (such as MINITAB and an Excel add-in).

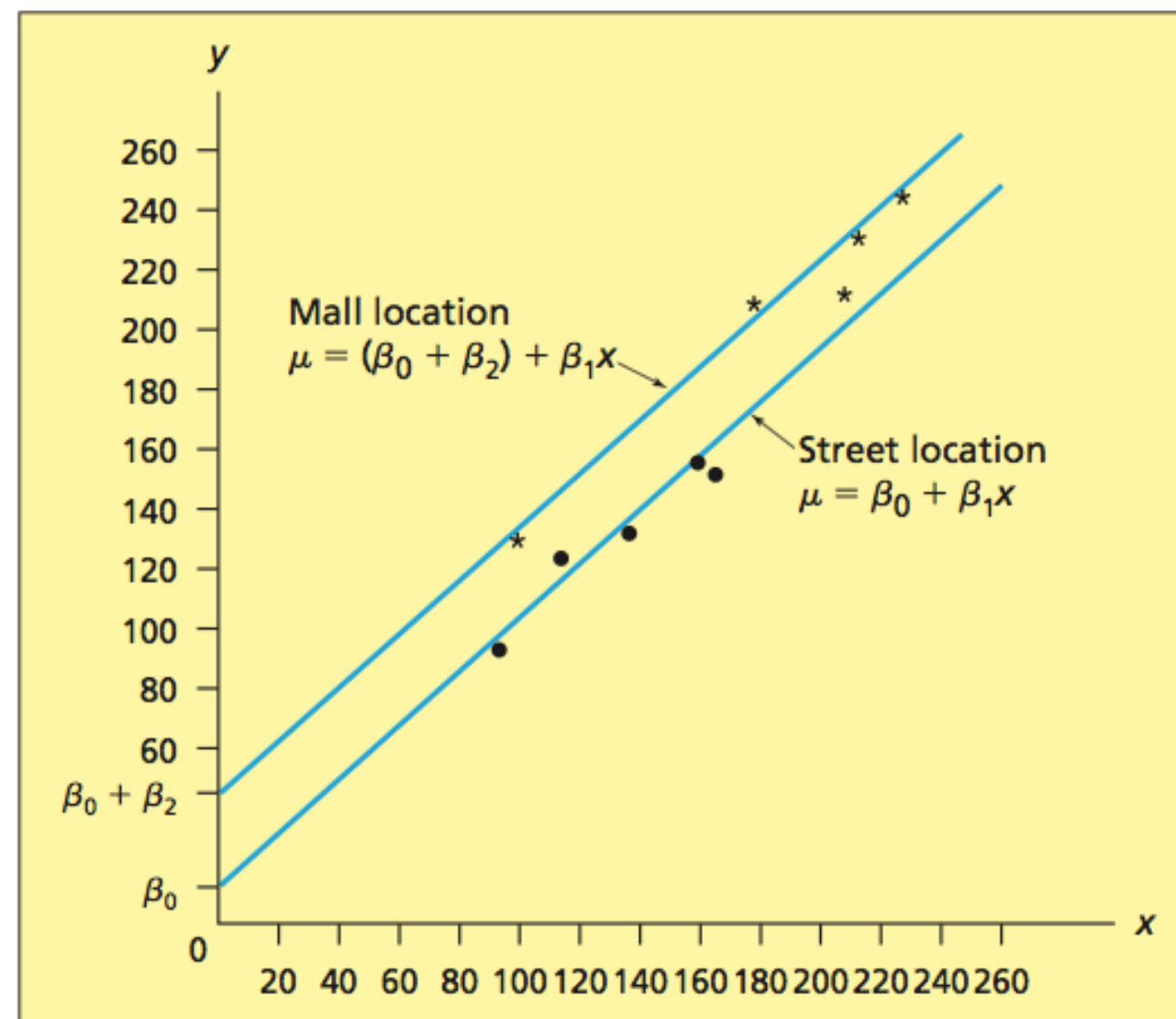
Mô hình các biến định tính

◆ Sử dụng biến giả

TABLE 15.8 The Electronics World Sales Volume Data
DS Electronics1

Store	Number of Households, x	Location	Sales Volume, y
1	161	Street	157.27
2	99	Street	93.28
3	135	Street	136.81
4	120	Street	123.79
5	164	Street	153.51
6	221	Mall	241.74
7	179	Mall	201.54
8	204	Mall	206.71
9	214	Mall	229.78
10	101	Mall	135.22

FIGURE 15.12 Plot of the Sales Volume Data and a Geometrical Interpretation of the Model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \varepsilon$

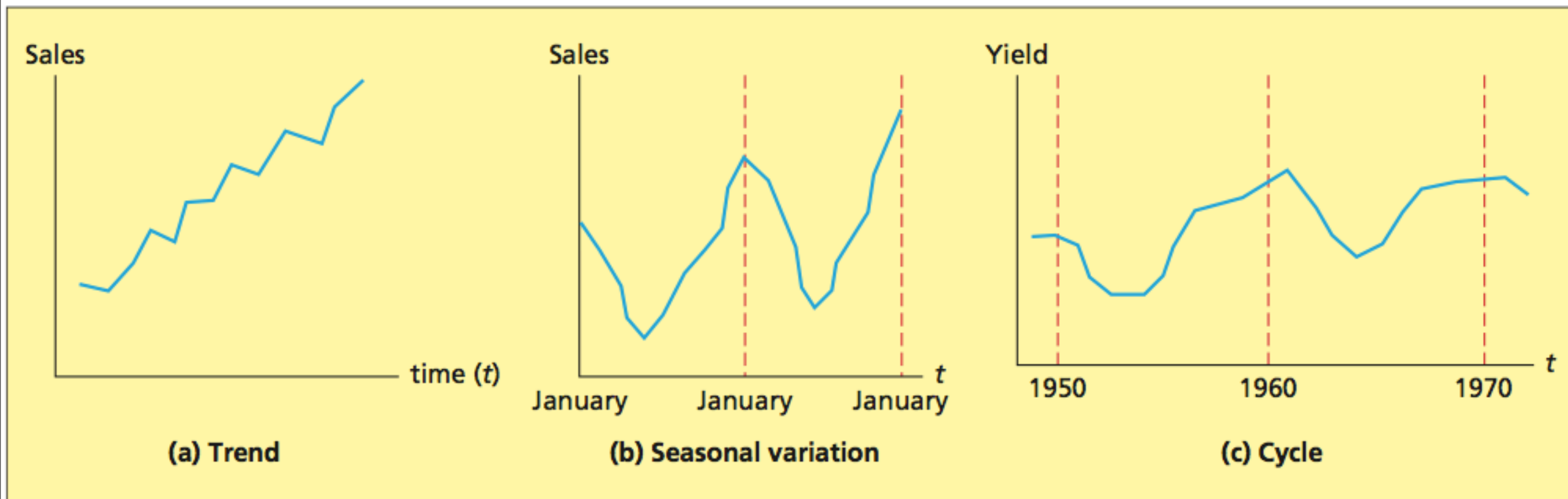


Dữ liệu dòng thời gian (Time series data)

Các thành phần của dữ liệu dòng thời gian

- ◆ 4 components of time series data
 - ◆ **Trend, Seasonal variation, Cycle**
 - ◆ **Irregular fluctuations** are erratic time series movements that follow no recognizable or regular pattern.

FIGURE 16.1 Time Series Exhibiting Trend, Seasonal, and Cyclical Components



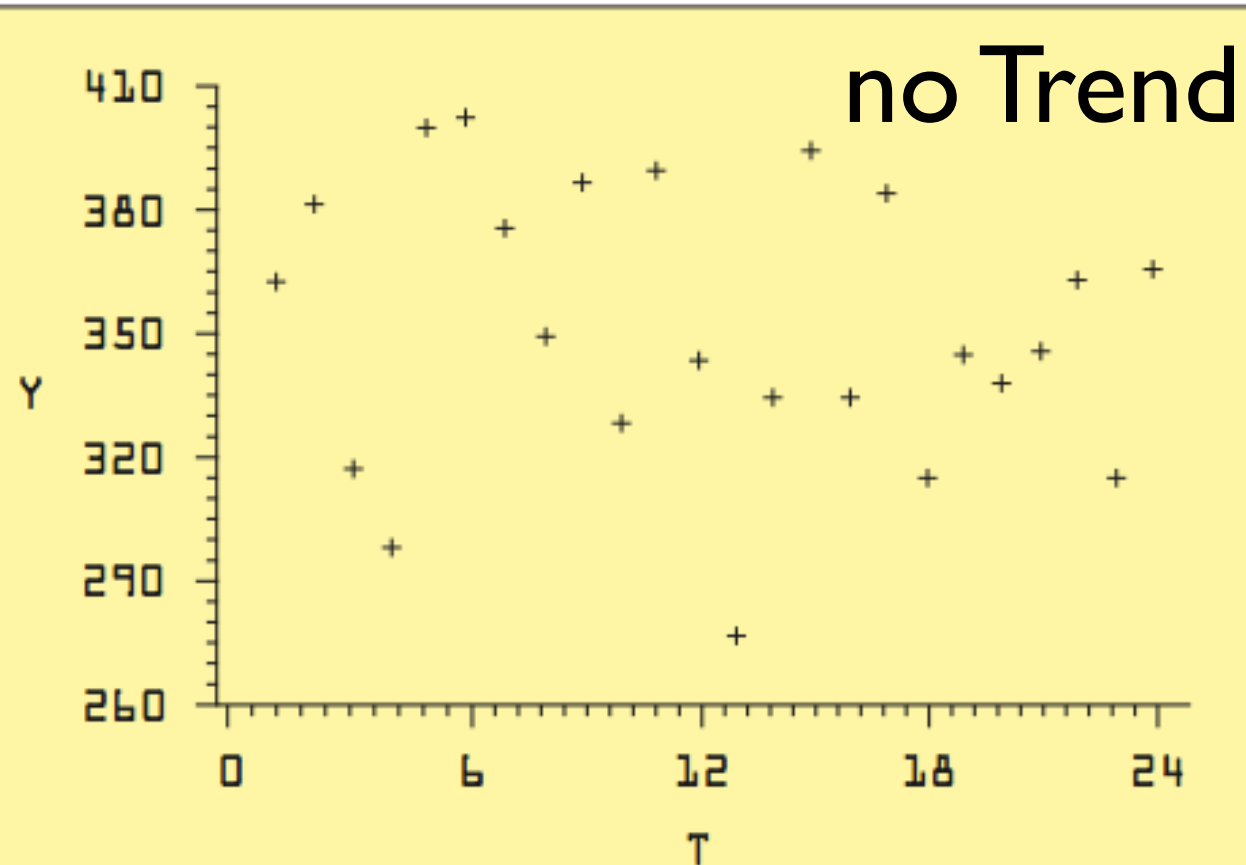
Methods for time series data

- ◆ The time series components remain essentially constant over time
 - ✓ **Series regression models**
 - ✓ **Multiplicative decomposition**
- ◆ The time series components might be changing slowly over time
 - ✓ **Exponential smoothing**
- ◆ The time series components might be changing fairly quickly over time
 - ✓ **Box–Jenkins methodology**

Time Series Regression

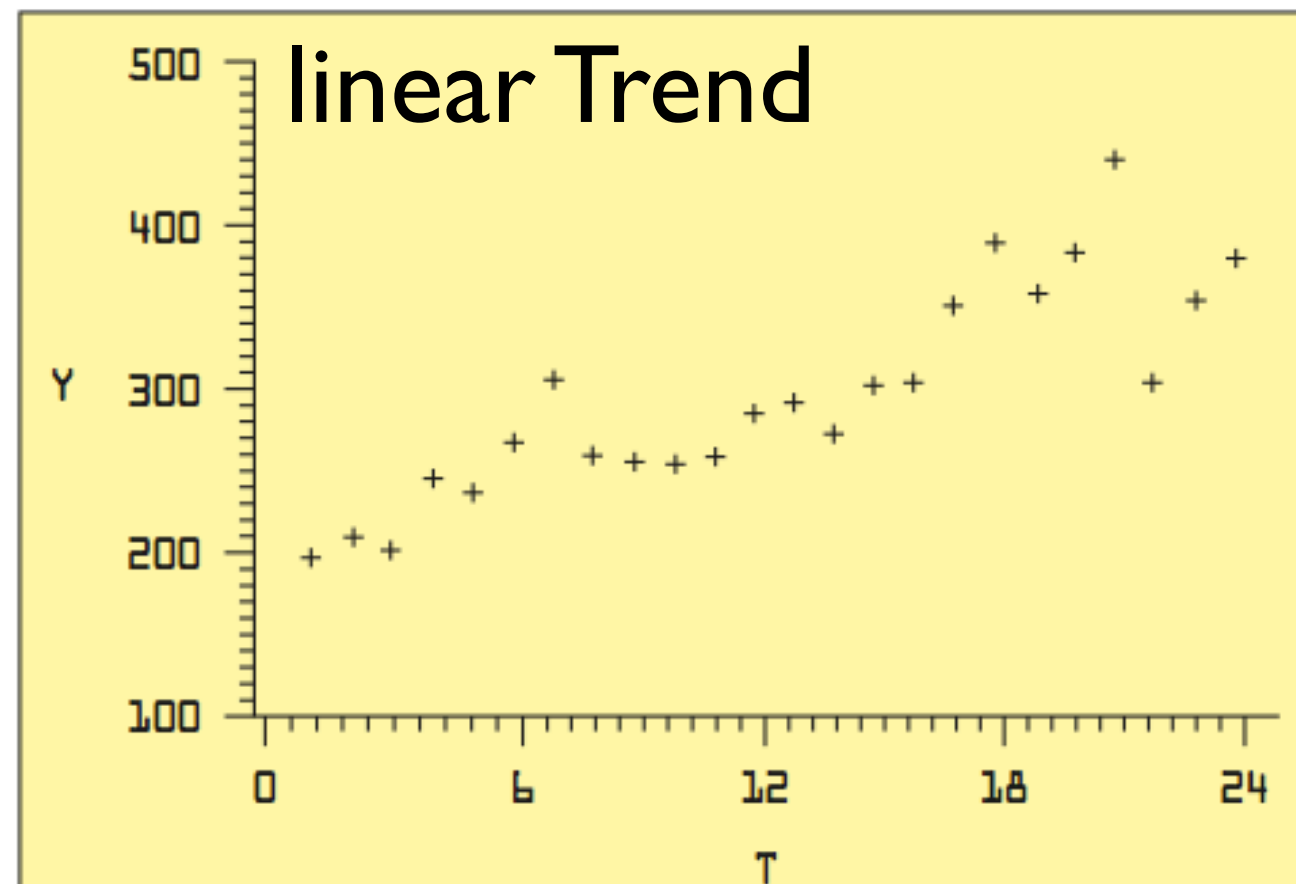
◆ Modeling trend components

FIGURE 16.2 Plot of Cod Catch versus Time



$$y_t = \beta_0 + \varepsilon_t$$


FIGURE 16.3 Plot of Calculator Sales versus Time



$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

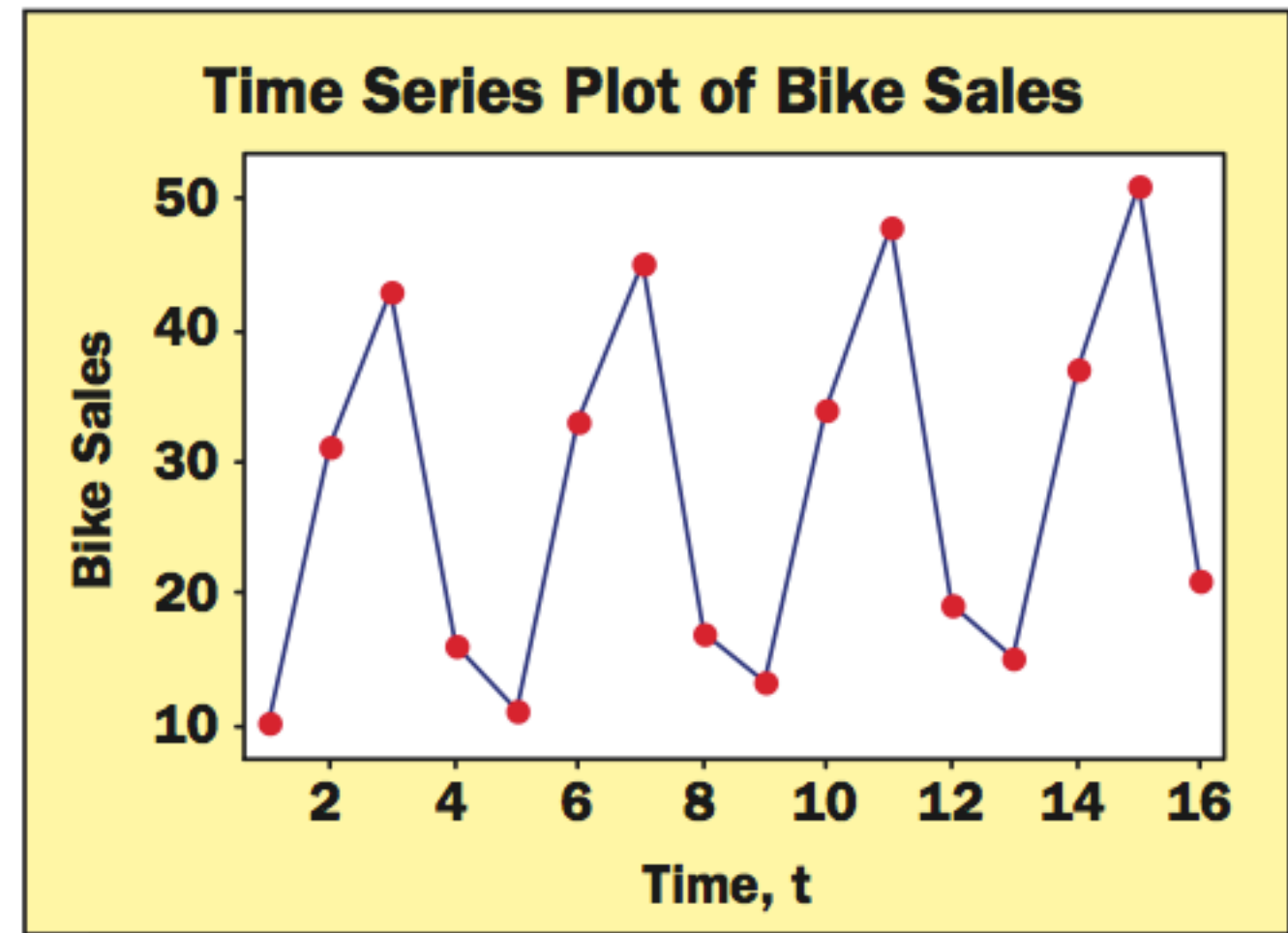
Time Series Regression

◆ Modeling seasonal components

TABLE 16.3 Quarterly Sales of the TRK-50 Mountain Bike  BikeSales

Year	Quarter	t	Sales, y_t
1	1 (Winter)	1	10
	2 (Spring)	2	31
	3 (Summer)	3	43
	4 (Fall)	4	16
2	1	5	11
	2	6	33
	3	7	45
	4	8	17
3	1	9	13
	2	10	34
	3	11	48
	4	12	19
4	1	13	15
	2	14	37
	3	15	51
	4	16	21

FIGURE 16.5 Time Series Plot of TRK-50 Bike Sales



$$y_t = \beta_0 + \beta_1 t + \beta_{Q2} Q_2 + \beta_{Q3} Q_3 + \beta_{Q4} Q_4 + \varepsilon_t$$

Time Series Regression

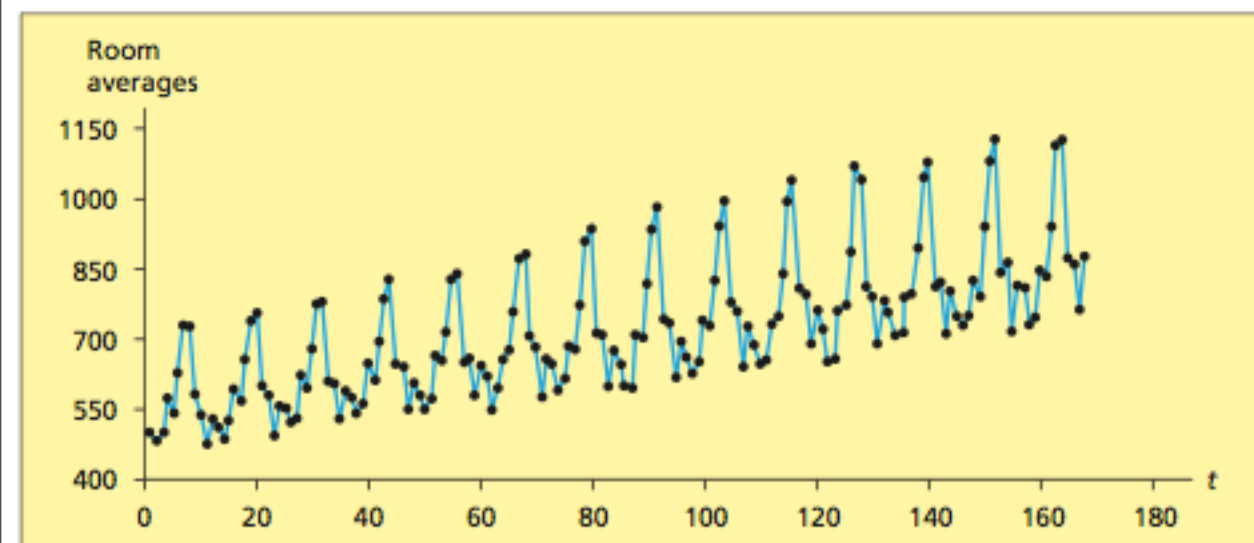
◆ Constant seasonal variation

- ✓ The magnitude of the seasonal swing does not depend on the level of the time series

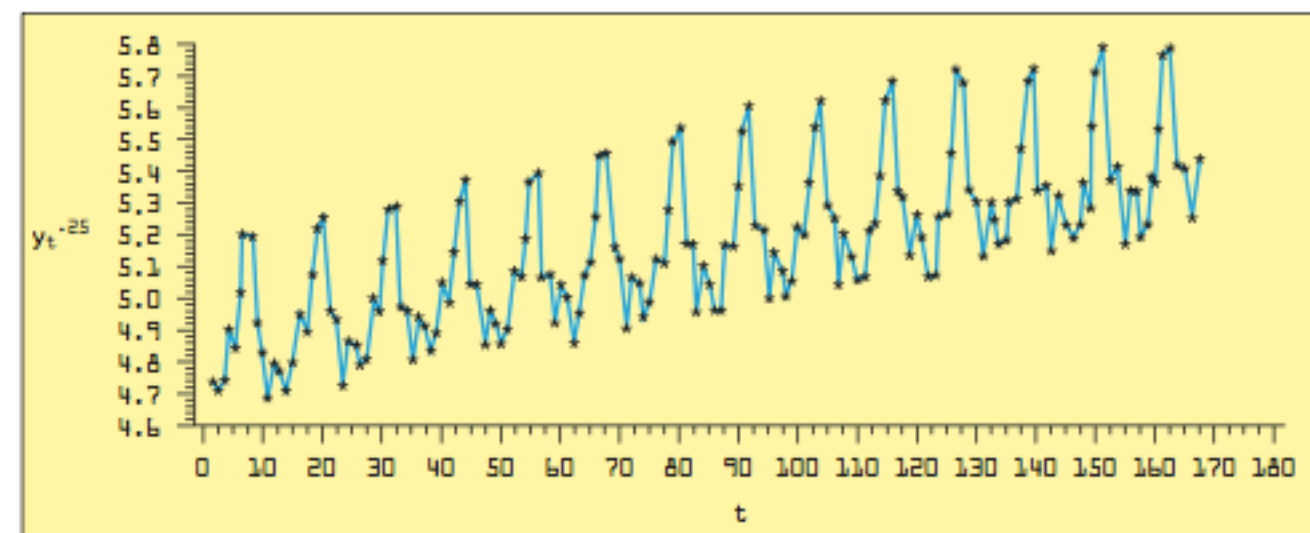
◆ Increasing seasonal variation

- ✓ The magnitude of the seasonal swing increases as the level of the time series increases.
- ✓ First use a **fractional power transformation**: taking the square roots, quartic roots, and natural logarithms

(a) Plot of the monthly hotel room averages versus time



(b) Plot of the quartic roots of the monthly hotel averages versus time

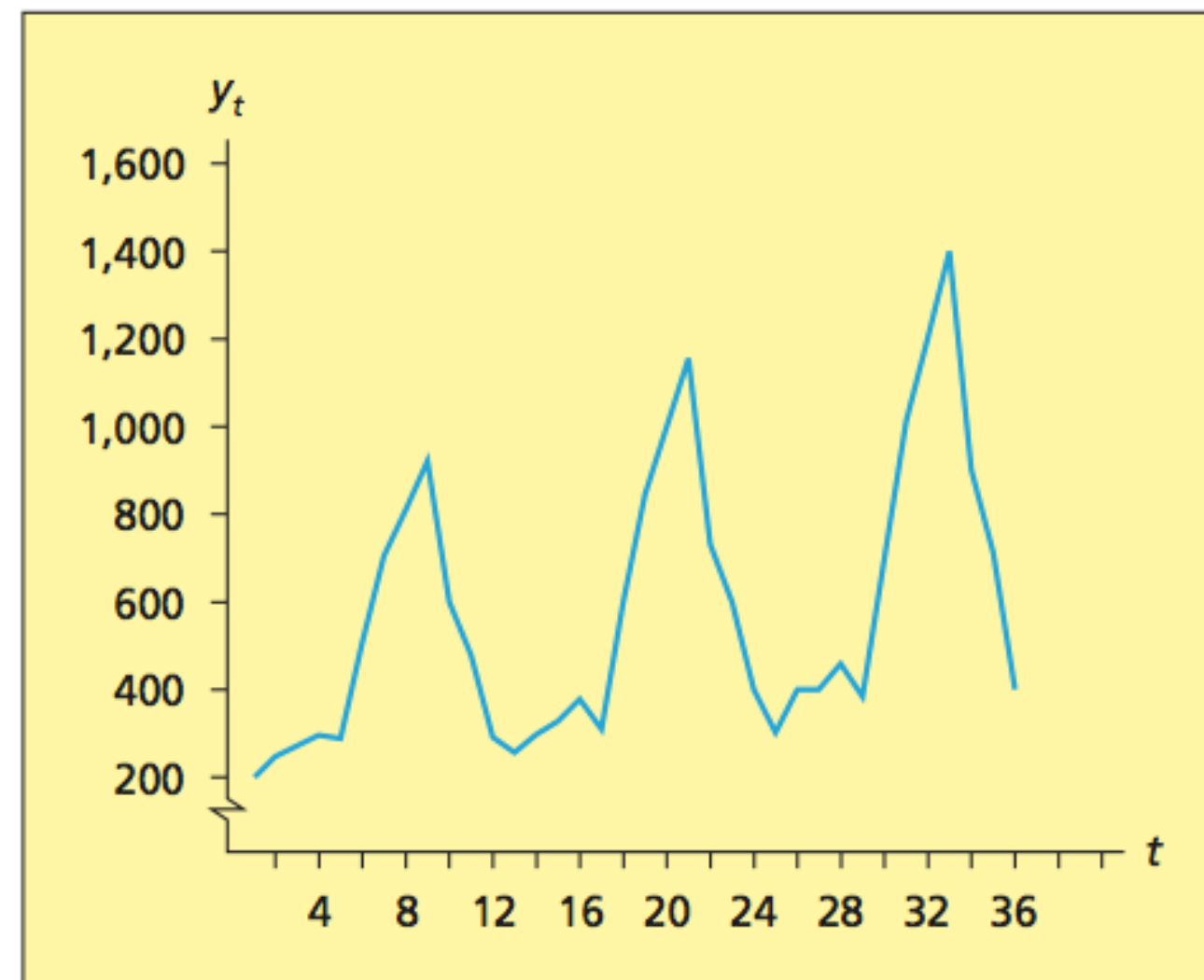


Multiplicative Decomposition

TABLE 16.9 Monthly Sales of Tasty Cola (in Hundreds of Cases) 

Year	Month	t	Sales, y_t	Year	Month	t	Sales, y_t
1	1 (Jan.)	1	189	2	7	19	831
	2 (Feb.)	2	229		8	20	960
	3 (Mar.)	3	249		9	21	1,152
	4 (Apr.)	4	289		10	22	759
	5 (May)	5	260		11	23	607
	6 (June)	6	431		12	24	371
	7 (July)	7	660	3	1	25	298
	8 (Aug.)	8	777		2	26	378
	9 (Sept.)	9	915		3	27	373
	10 (Oct.)	10	613		4	28	443
	11 (Nov.)	11	485		5	29	374
	12 (Dec.)	12	277		6	30	660
2	1	13	244		7	31	1,004
	2	14	296		8	32	1,153
	3	15	319		9	33	1,388
	4	16	370		10	34	904
	5	17	313		11	35	715
	6	18	556		12	36	441

FIGURE 16.10 Time Series Plot of the Tasty Cola Sales Data



Multiplicative Decomposition

- ◆ Decompose the time series into its **trend**, **seasonal**, **cyclical**, and **irregular components**

$$y_t = TR_t \times SN_t \times CL_t \times IR_t$$

Here **TR_t** , **SN_t** , **CL_t** , and **IR_t** represent the trend, seasonal, cyclical, and irregular components of the time series in time period t .

TABLE 16.10 Tasty Cola Sales and the Multiplicative Decomposition Method

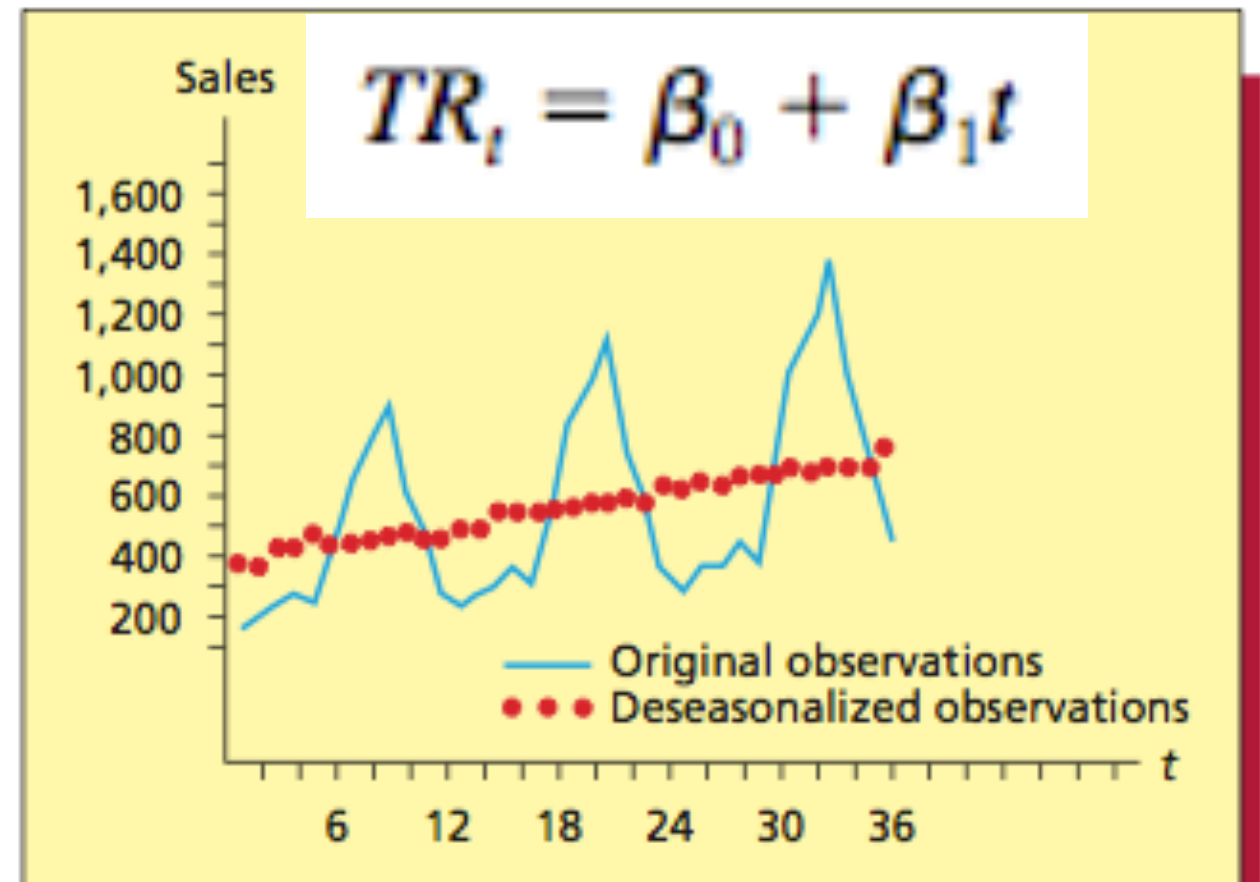
t Time Period	y _t Tasty Cola Sales	First Step: 12-Period Moving Average	tr _t × cl _t : Centered Moving Average	sn _t × ir _t : y _t tr _t × cl _t	sn _t : Table 16.11	d _t : y _t sn _t	tr _t : 380.163 +9.489t	tr _t × sn _t : Multiply tr _t by sn _t	cl _t × ir _t : y _t tr _t × sn _t	cl _t : 3-Period Moving Average	ir _t : cl _t × ir _t cl _t
1 (Jan)	189				.493	383.37	389.652	192.10	.9839		
2	229				.596	384.23	399.141	237.89	.9626	.9902	.9721
3	249				.595	418.49	408.630	243.13	1.0241	1.0010	1.0231
4	289				.680	425	418.119	284.32	1.0165	1.0396	.9778
5	260				.564	460.99	427.608	241.17	1.0781	1.0315	1.0452
6	431				.986	437.12	437.097	430.98	1.0000	1.0285	.9723
7	660	447.833	450.125	1.466	1.467	449.9	446.586	655.14	1.0074	1.0046	1.0028
8	777	452.417	455.2085	1.707	1.693	458.95	456.075	772.13	1.0063	1.0004	1.0059
9	915	458	460.9165	1.985	1.990	459.79	465.564	926.47	.9876	.9937	.9939
10	613	563.833	467.208	1.312	1.307	469.01	475.053	620.89	.9873	.9825	1.0049
11	485	470.583	472.7915	1.026	1.029	471.33	489.542	498.59	.9727	.9648	1.0082
12	277	475	480.2085	.577	.600	461.67	494.031	296.42	.9345	.9634	.9700
13 (Jan)	244	485.417	492.542	.495	.493	494.97	503.520	248.24	.9829	.9618	1.0219
14	296	499.667	507.292	.583	.596	496.64	513.009	305.75	.9681	.9924	.9755
15	319	514.917	524.792	.608	.595	536.13	522.498	310.89	1.0261	1.0057	1.0203
16	370	534.667	540.75	.684	.680	544.12	531.987	361.75	1.0228	1.0246	.9982
17	313	546.833	551.9165	.567	.564	554.97	541.476	305.39	1.0249	1.0237	1.0012
18	556	557	560.9165	.991	.986	563.89	550.965	543.25	1.0235	1.0197	1.0037
19	831	564.833	567.083	1.465	1.467	566.46	560.454	822.19	1.0107	1.0097	1.0010
20	960	569.333	572.75	1.676	1.693	567.04	569.943	964.91	.9949	1.0016	.9933
21	1,152	576.167	578.417	1.992	1.990	578.89	579.432	1,153.07	.9991	.9934	1.0057
22	759	580.667	583.7085	1.300	1.307	580.72	588.921	769.72	.9861	.9903	.9958
23	607	586.75	589.2915	1.030	1.029	589.89	598.410	615.76	.9858	.9964	.9894
24	371	591.833	596.1665	.622	.600	618.33	607.899	364.74	1.0172	.9940	1.0233
25 (Jan)	298	600.5	607.7085	.490	.493	604.46	617.388	304.37	.9791	1.0027	.9765
26	378	614.917	622.9585	.607	.596	634.23	626.877	373.62	1.0117	.9920	1.0199
27	373	631	640.8335	.582	.595	626.89	636.366	378.64	.9851	1.0018	.9833
28	443	650.667	656.7085	.675	.680	651.47	645.855	439.18	1.0087	1.0030	1.0057
29	374	662.75	667.25	.561	.564	663.12	655.344	369.61	1.0119	1.0091	1.0028
30	660	671.75	674.6665	.978	.986	669.37	664.833	655.53	1.0068	1.0112	.9956
31	1,004	677.583			1.467	684.39	674.322	989.23	1.0149	1.0059	1.0089
32	1,153				1.693	681.04	683.811	1,157.69	.9959	1.0053	.9906
33	1,388				1.990	697.49	693.300	1,379.67	1.0060	.9954	1.0106
34	904				1.307	691.66	702.789	918.55	.9842	.9886	.9955
35	715				1.029	694.85	712.278	732.93	.9755	.9927	.9827
36	441				.600	735	721.707	433.06	1.0183		

Multiplicative Decomposition

TABLE 16.11 Estimation of the Seasonal Factors

		$sn_t \times ir_t = y_t / (tr_t \times cl_t)$		$sn_t = 1.0008758(\bar{sn}_t)$	
		Year 1	Year 2	\bar{sn}_t	
1	Jan.	.495	.490	.4925	.493
2	Feb.	.583	.607	.595	.596
3	Mar.	.608	.582	.595	.595
4	Apr.	.684	.675	.6795	.680
5	May	.567	.561	.564	.564
6	June	.991	.978	.9845	.986
7	July	1.466	1.465	1.4655	1.467
8	Aug.	1.707	1.676	1.6915	1.693
9	Sep.	1.985	1.992	1.9885	1.990
10	Oct.	1.312	1.300	1.306	1.307
11	Nov.	1.026	1.030	1.028	1.029
12	Dec.	.577	.622	.5995	.600

FIGURE 16.11 Plot of Tasty Cola Sales and Deseasonalized Sales



- ◆ If there is no pattern in the irregular component, we predict IR_t to equal 1. Therefore:

$$\hat{y}_t = tr_t \times sn_t \times cl_t$$

Simple Exponential Smoothing

- ◆ Forecasts of future time series values are made each period for succeeding periods

- 1 Suppose that the time series y_1, \dots, y_n is described by the equation

$$y_t = \beta_0 + \varepsilon_t$$

where the average level β_0 of the process may be slowly changing over time. Then the estimate S_T of β_0 made in time period T is given by the **smoothing equation**

$$S_T = \alpha y_T + (1 - \alpha)S_{T-1}$$

where α is a smoothing constant between 0 and

1 and S_{T-1} is the estimate of β_0 made in time period $T-1$.

- 2 A point forecast made in time period T for any future value of the time series is S_T .

- 3 If we observe y_{T+1} in time period $T+1$, we can update S_T to S_{T+1} by using the equation

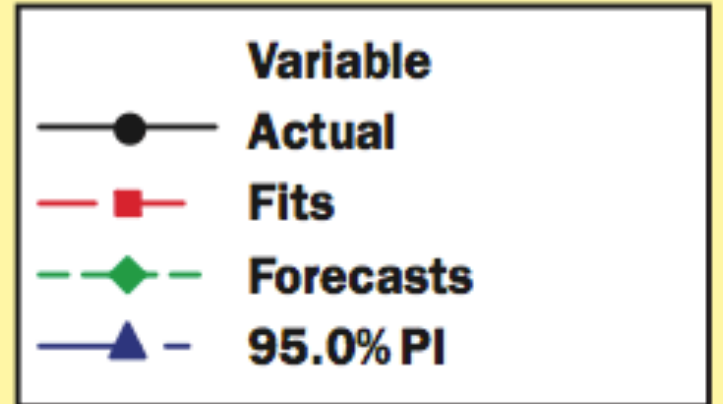
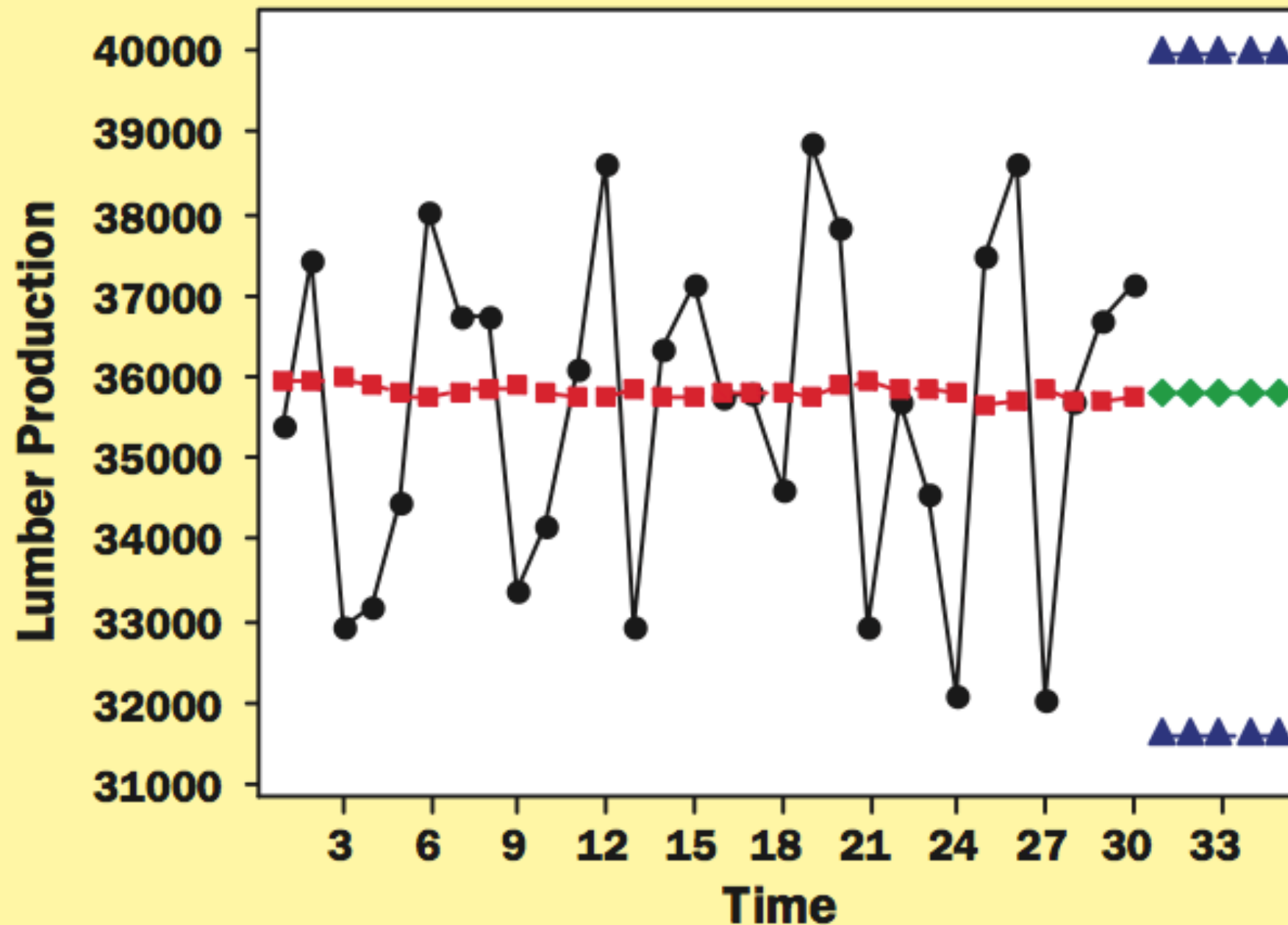
$$S_{T+1} = \alpha y_{T+1} + (1 - \alpha)S_T$$

and a point forecast made in time period $T+1$ for any future value of the time series is S_{T+1} .

- ◆ Experience has shown that, in general, it is reasonable to calculate **initial estimates** in exponential smoothing procedures using the mean of half of the historical data.

Simple Exponential Smoothing

Single Exponential Smoothing for Lumber Production



Smoothing Constant
Alpha 0.0361553

Accuracy Measures
MAPE 5
MAD 1712
MSD 4201951

Period	Forecast	Lower	Upper
31	35782.6	31588.9	39976.3

Holt–Winters' double exponential smoothing

◆ The model $y_t = \beta_0 + \beta_1 t + \varepsilon_t$

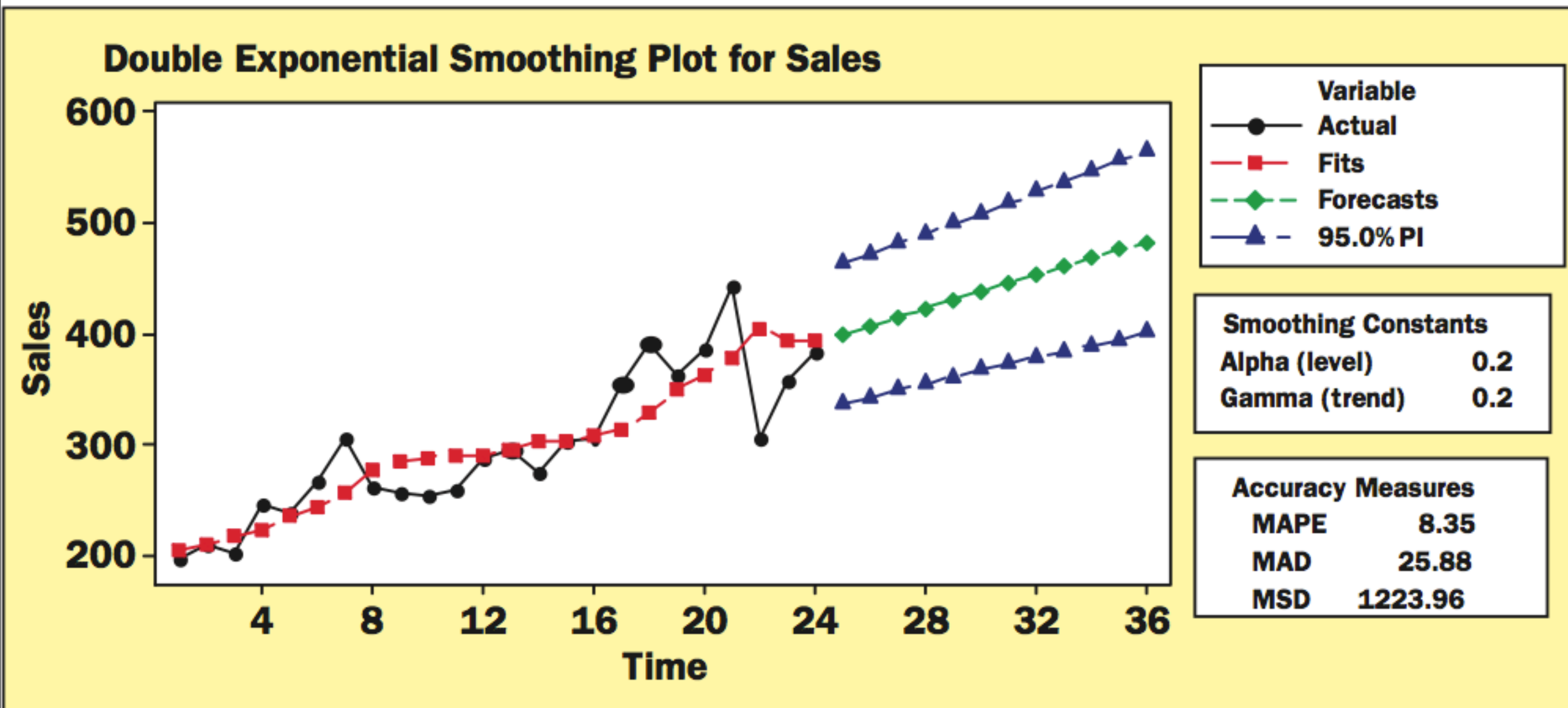
ℓ_{T-1} denote the estimate of $\beta_0 + \beta_1(T - 1)$ in period T-1

b_{T-1} denote the estimate of β_1 in period T-1

$$\ell_T = \alpha y_T + (1 - \alpha) [\ell_{T-1} + b_{T-1}]$$

$$b_T = \gamma [\ell_T - \ell_{T-1}] + (1 - \gamma) b_{T-1}$$

Holt–Winters' double exponential smoothing



Multiplicative Winters' method

◆ The model $y_t = (\beta_0 + \beta_1 t) \times SN_t + \varepsilon_t$

ℓ_{T-1} denote the estimate of $\beta_0 + \beta_1(T-1)$ in period T-1

b_{T-1} denote the estimate of β_1 in period T-1

sn_{T-L} denote the “most recent” estimate of the seasonal factor for the season corresponding to time period T; where L denotes the number of seasons in a year

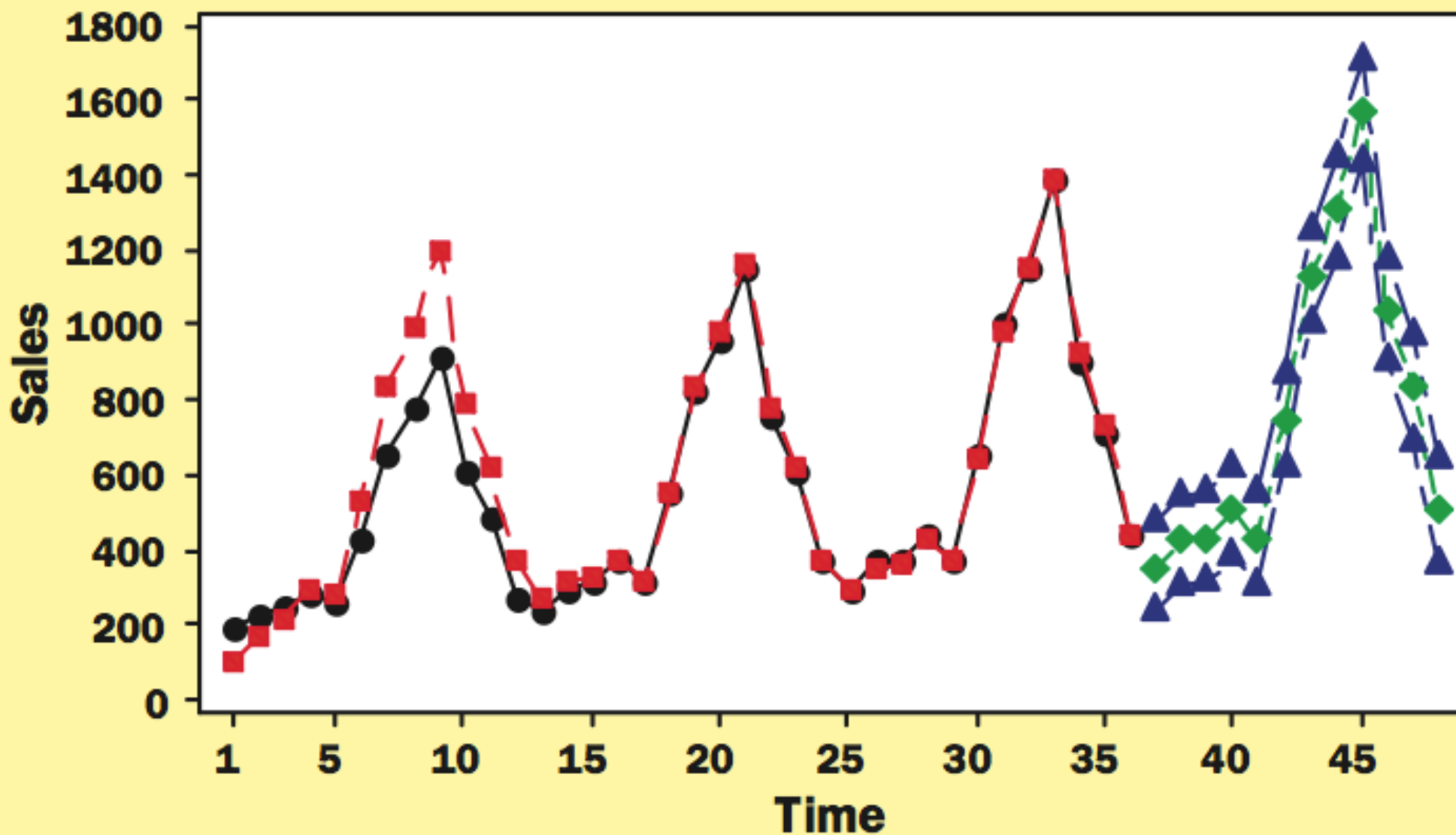
$$\ell_T = \alpha \frac{y_T}{sn_{T-L}} + (1 - \alpha)[\ell_{T-1} + b_{T-1}]$$

$$b_T = \gamma[\ell_T - \ell_{T-1}] + (1 - \gamma)b_{T-1}$$

$$sn_T = \delta \frac{y_T}{\ell_T} + (1 - \delta)sn_{T-L}$$

Multiplicative Winters' method

Winters' Method Plot for Sales
Multiplicative Method



Variable	
—●—	Actual
—■—	Fits
- -◆- -	Forecasts
- -▲- -	95.0% PI

Smoothing Constants	
Alpha (level)	0.2
Gamma (trend)	0.2
Delta (seasonal)	0.2

Accuracy Measures	
MAPE	9.80
MAD	46.93
MSD	6812.61

Index Numbers

- ◆ To compare a value of a time series relative to another value of the time series.
 - ✓ Must describe the time series
 - **Series decomposition** can be employed
 - **Index numbers** can be used

A **simple index** is obtained by dividing the current value of a time series by the value of the time series in the base time period and by multiplying this ratio by 100. That is, if y_t denotes the current value and if y_0 denotes the value in the base time period, then the **simple index number** is

$$\frac{y_t}{y_0} \times 100$$