Business Statistics: A First Course

(3rd Edition)

Chapter 10

Simple Linear Regression

Chapter Topics

- Types of Regression Models
- Determining the Simple Linear Regression Equation
- Measures of Variation
- Assumptions of Regression and Correlation
- Residual Analysis
- Measuring Autocorrelation
- Inferences about the Slope

Chapter Topics

(continued)

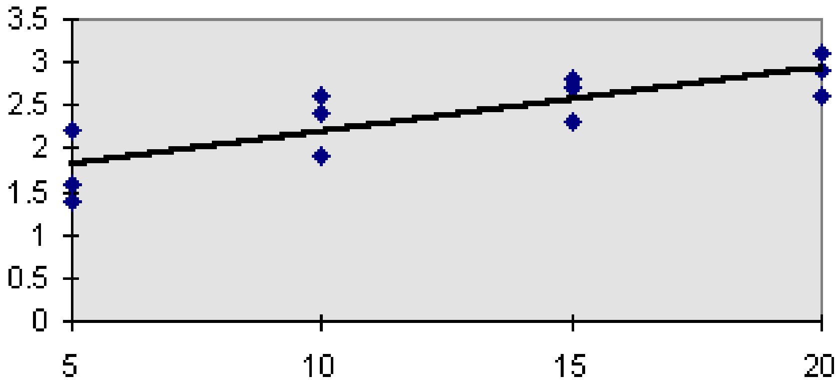
- Correlation - Measuring the Strength of the Association
- Estimation of Mean Values and Prediction of Individual Values
- Pitfalls in Regression and Ethical Issues

Purpose of Regression Analysis

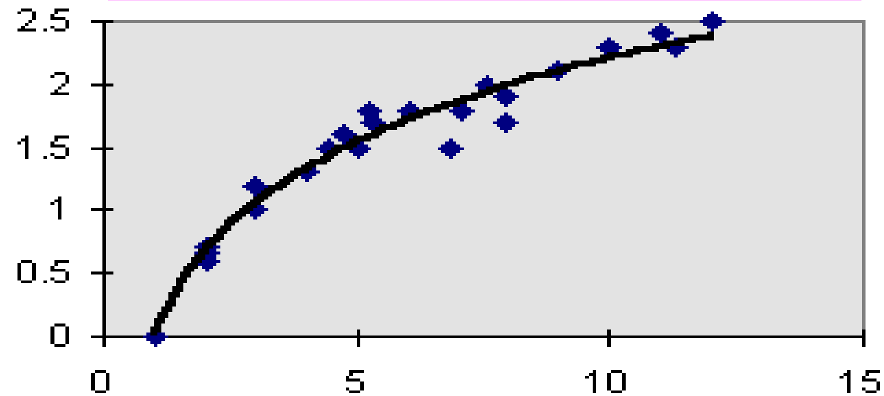
- Regression Analysis is Used Primarily to Model Causality and Provide Prediction
  - Predict the values of a dependent (response) variable based on values of at least one independent (explanatory) variable
  - Explain the effect of the independent variables on the dependent variable

Types of Regression Models

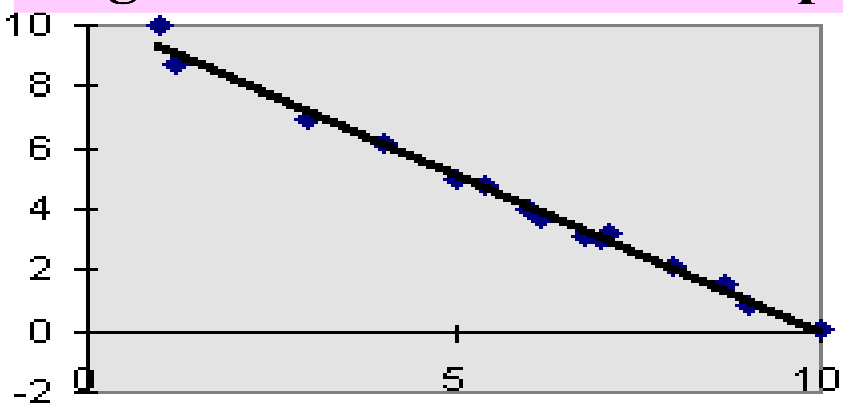
Positive Linear Relationship



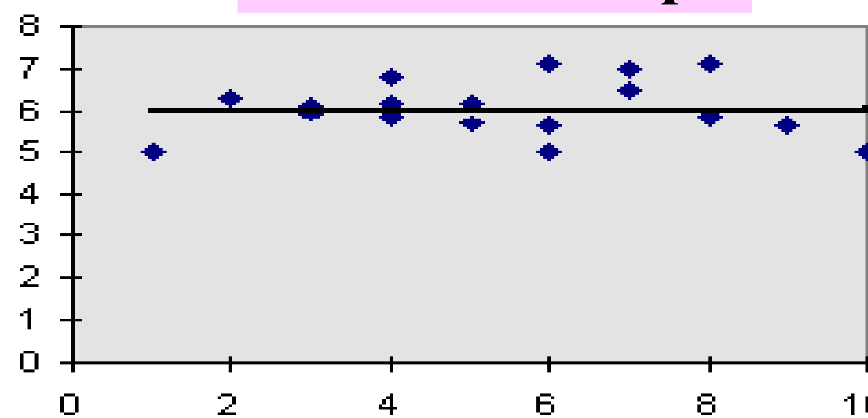
Relationship NOT Linear



Negative Linear Relationship



No Relationship



Simple Linear Regression Model

- Relationship Between Variables is Described by a Linear Function
- The Change of One Variable Causes the Other Variable to Change
- A Dependency of One Variable on the Other

Simple Linear Regression Model

(continued)

Population regression line is a straight line that describes the dependence of the average value (conditional mean) of one variable on the other

Population *Y* intercept

Population Slope Coefficient

Random Error

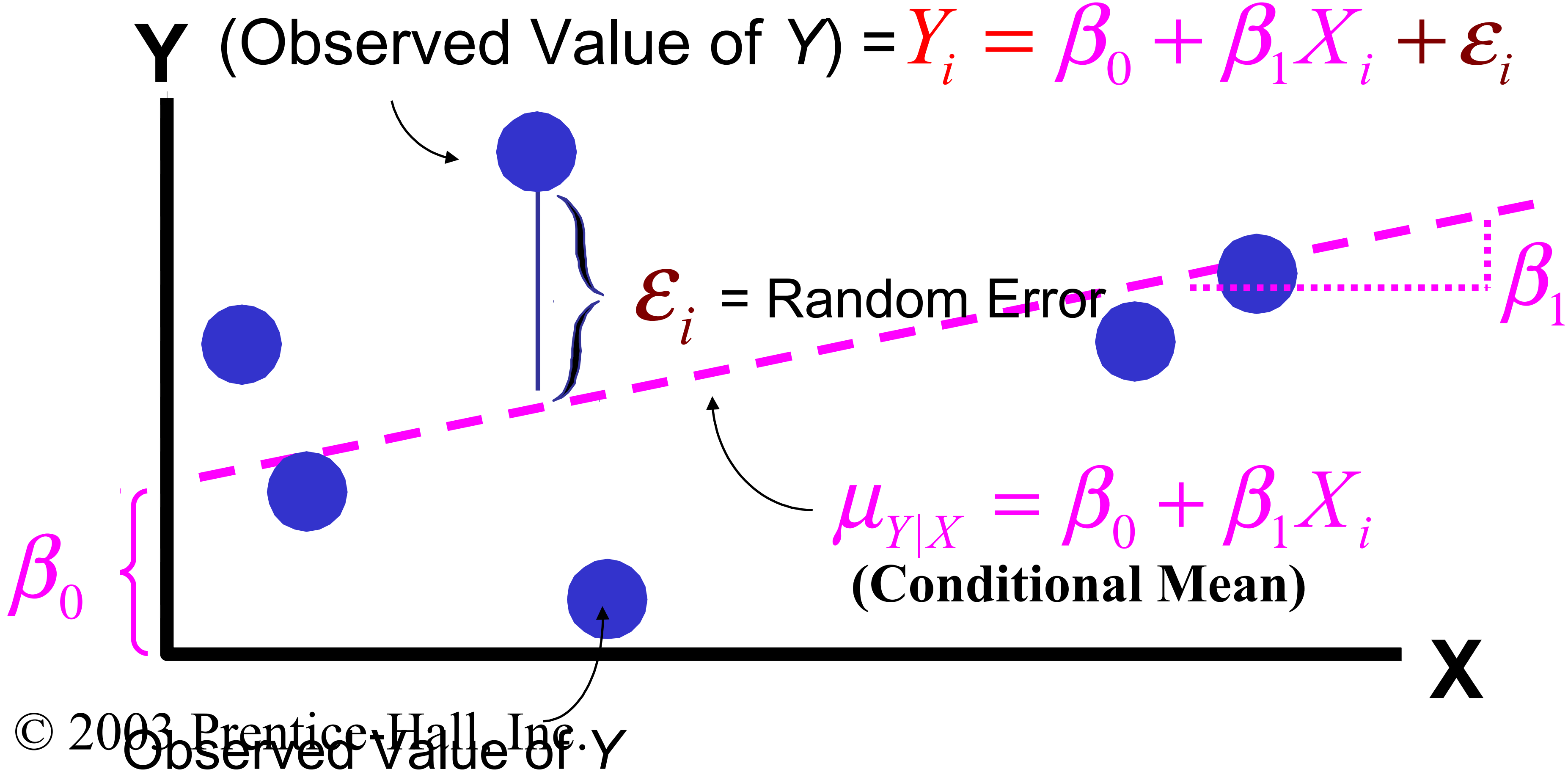
$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\mu_{Y|X}} + \varepsilon_i$$

Dependent (Response) Variable

Population Regression Line (conditional mean)

Independent (Explanatory) Variable

Simple Linear Regression Model
(continued)



Linear Regression Equation

Sample regression line provides an *estimate* of the population regression line as well as a predicted value of Y

Sample Y Intercept

Sample Slope Coefficient

$$Y_i = \underbrace{b_0 + b_1 X_i}_{} + e_i$$

Residual

$$\hat{Y} = b_0 + b_1 X =$$ **Simple Regression Equation (Fitted Regression Line, Predicted Value)**

Linear Regression Equation

(continued)

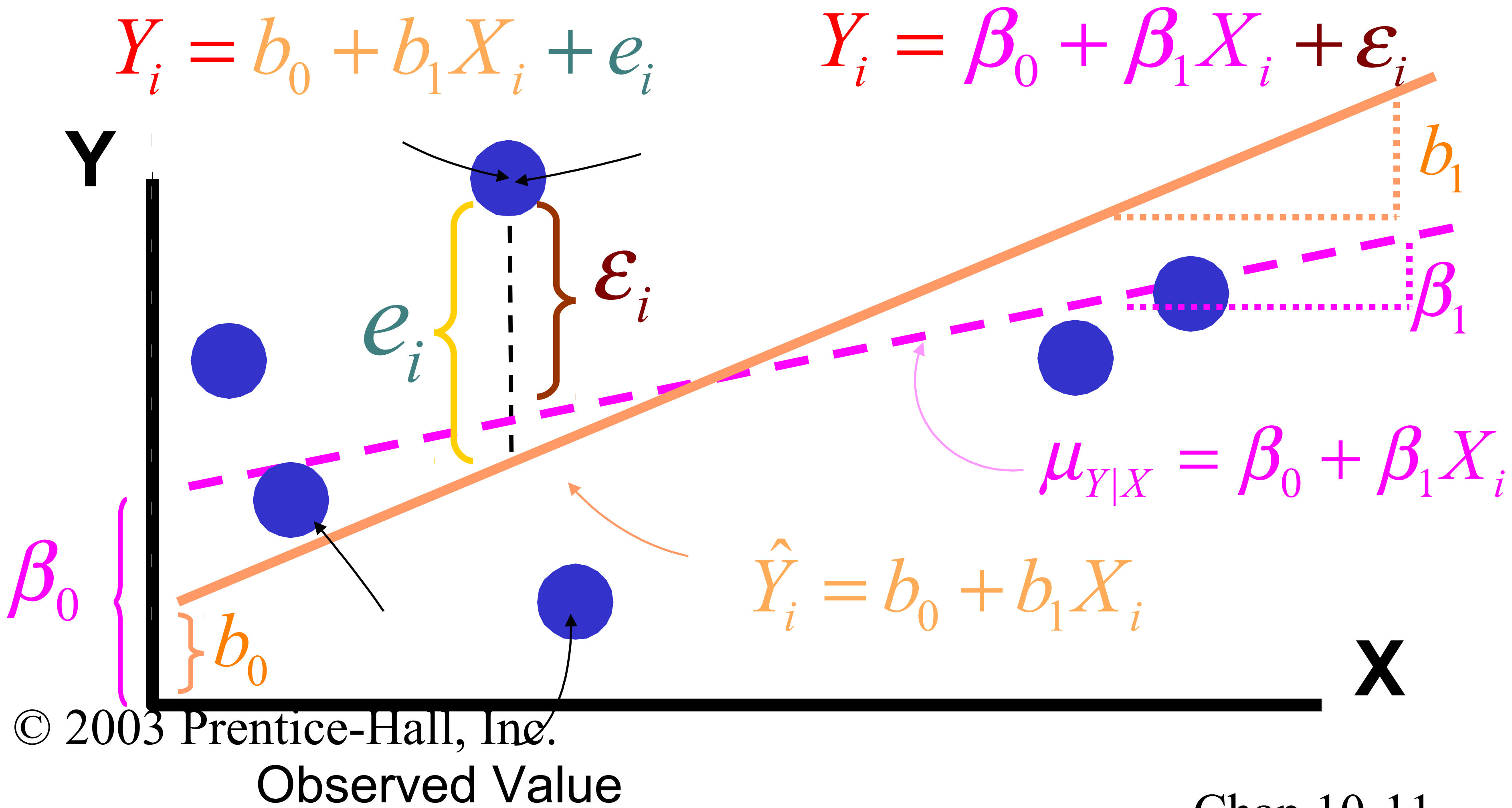
- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimizes the sum of the squared residuals

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2$$

- b_0 provides an *estimate* of β_0
- b_1 provides and *estimate* of β_1

Linear Regression Equation

(continued)



Interpretation of the Slope and Intercept

- $\beta_0 = \mu_{Y|X=0}$ is the average value of Y when the value of X is zero.

- $\beta_1 = \dfrac{\Delta \mu_{Y|X}}{\Delta X}$ measures the change in the average value of Y as a result of a one-unit change in X.

Interpretation of the Slope and Intercept *(continued)*

- $b_0 = \hat{\mu}_{Y|X=0}$ is the *estimated* average value of Y when the value of X is zero.

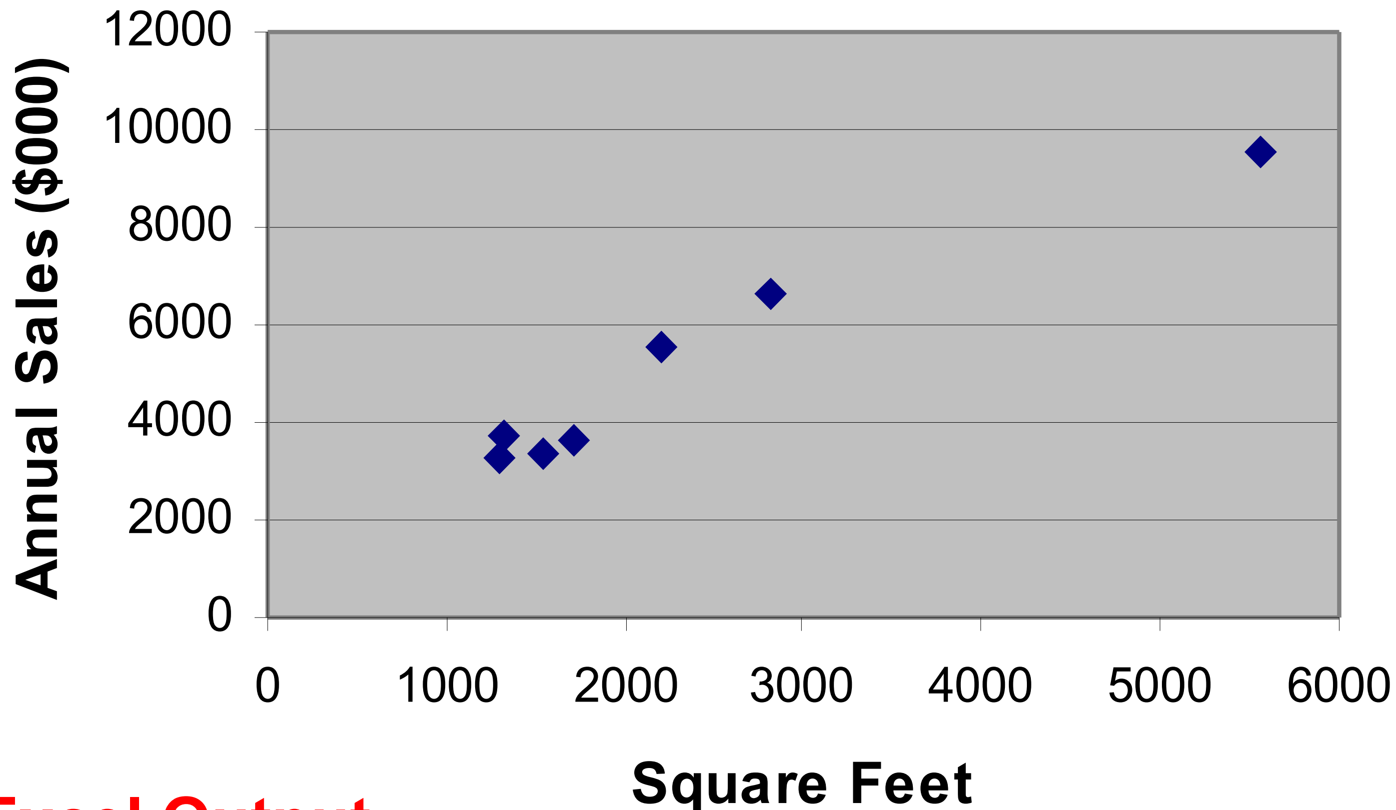
- $b_1 = \dfrac{\Delta\hat{\mu}_{Y|X}}{\Delta X}$ is the *estimated* change in the average value of Y as a result of a one-unit change in X.

Simple Linear Regression: Example

You wish to examine the linear dependency of the annual sales of produce stores on their sizes in square footage. Sample data for 7 stores were obtained. Find the equation of the straight line that fits the data best.

Store	Square Feet	Annual Sales ($1000)
1	1,726	3,681
2	1,542	3,395
3	2,816	6,653
4	5,555	9,543
5	1,292	3,318
6	2,208	5,563
7	1,313	3,760

Scatter Diagram: Example



Excel Output

Simple Linear Regression Equation: Example

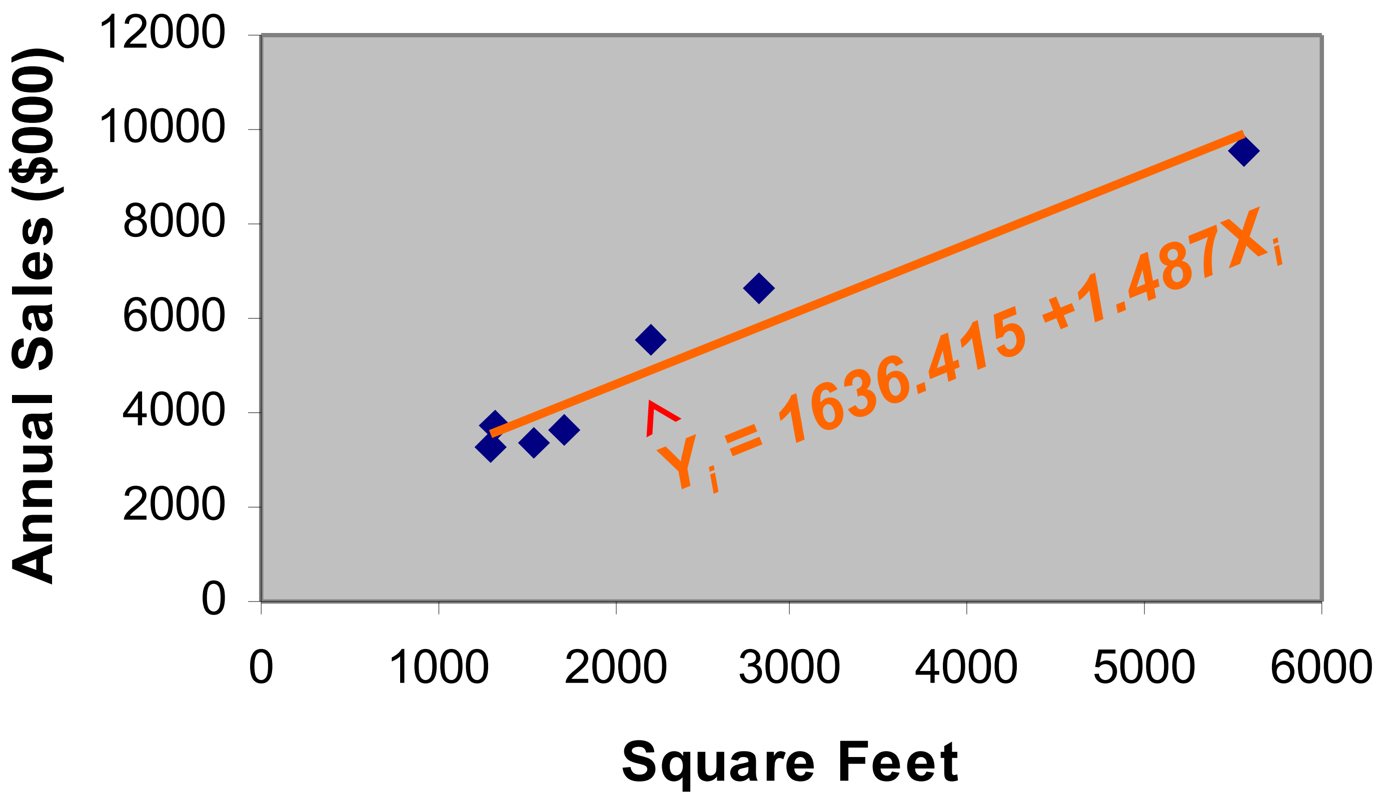
$$\hat{Y}_i = b_0 + b_1 X_i$$

$$= 1636.415 + 1.487 X_i$$

From Excel Printout:

	Coefficients
Intercept	1636.414726
X Variable	1.486633657

Graph of the Simple Linear Regression Equation: Example



Interpretation of Results: Example

$$\hat{Y}_i = 1636.415 + 1.487 X_i$$

The slope of 1.487 means that each increase of one unit in X, we predict the average of Y to increase by an estimated 1.487 units.

The equation estimates that for each increase of 1 square foot in the size of the store, the expected annual sales are predicted to increase by $1487.

Simple Linear Regression in PHStat

- In Excel, use PHStat | Regression | Simple Linear Regression …
- EXCEL Spreadsheet of Regression Sales on Footage

Microsoft Excel Worksheet

Measures of Variation: The Sum of Squares

SST	=	***SSR***	+	***SSE***
Total Sample Variability	=	**Explained Variability**	+	**Unexplained Variability**

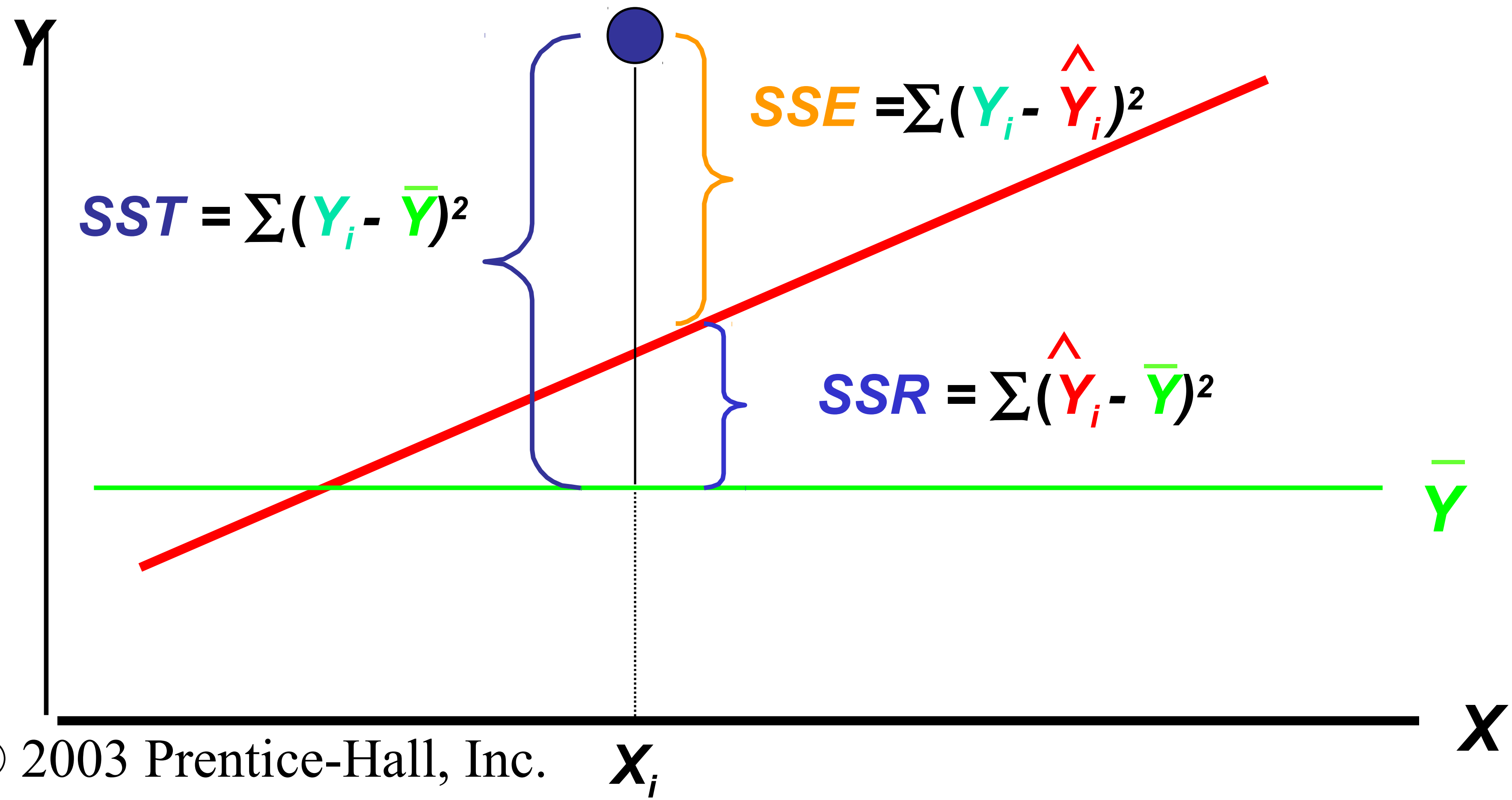
Measures of Variation: The Sum of Squares

(continued)

- SST = Total Sum of Squares
  - Measures the variation of the Y_i values around their mean, $\bar{Y}$
- SSR = Regression Sum of Squares
  - Explained variation attributable to the relationship between X and Y
- SSE = Error Sum of Squares
  - Variation attributable to factors other than the relationship between X and Y

Measures of Variation: The Sum of Squares

(continued)



The ANOVA Table in Excel

ANOVA					
	df	SS	MS	F	Significance F
Regression	k	SSR	MSR =SSR/k	MSR/MSE	P-value of the F Test
Residuals	n-k-1	SSE	MSE =SSE/(n-k-1)		
Total	n-1	SST			

Measures of Variation
The Sum of Squares: Example

Excel Output for Produce Stores

Degrees of freedom

ANOVA					
	df	***SS***	*MS*	*F*	*Significance F*
Regression	1	30380456.12	30380456	81.17909	0.000281201
Residual	5	1871199.595	374239.92		
Total	6	32251655.71			

Regression (explained) df

Error (residual) df

Total df

SSR

SSE

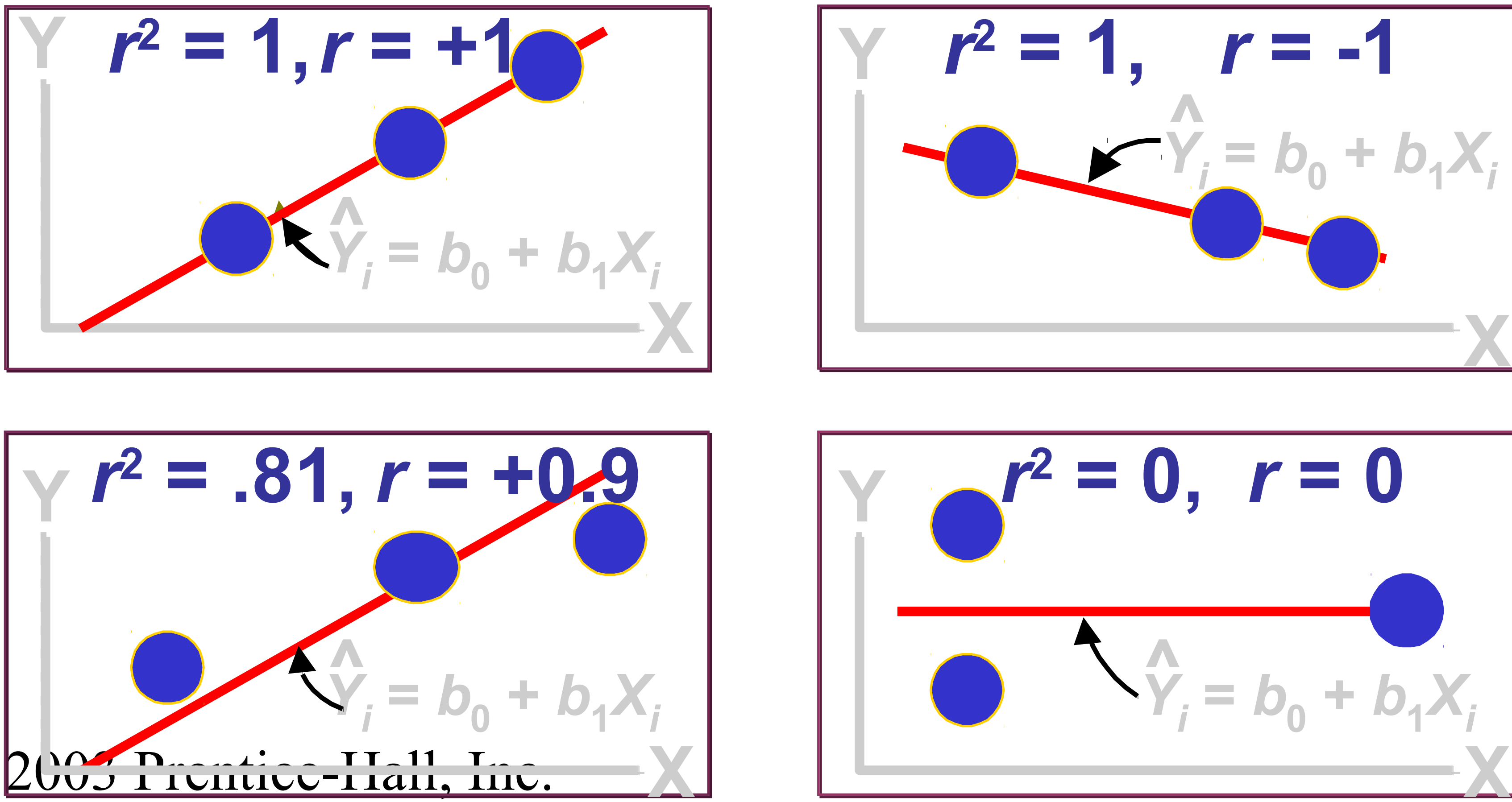
SST

The Coefficient of Determination

- $$r^2 = \frac{SSR}{SST} = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}}$$

- Measures the proportion of variation in Y that is explained by the independent variable X in the regression model

Coefficients of Determination (r^2) and Correlation (r)



Standard Error of Estimate

- $$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y - \hat{Y}_i\right)^2}{n-2}}$$

- The standard deviation of the variation of observations around the regression equation

Measures of Variation: Produce Store Example

Excel Output for Produce Stores

Regression Statistics	
Multiple R	0.9705572
R Square	0.94198129
Adjusted R Square	0.93037754
Standard Error	611.751517
Observations	7

$r^2 = .94$

S_{yx}

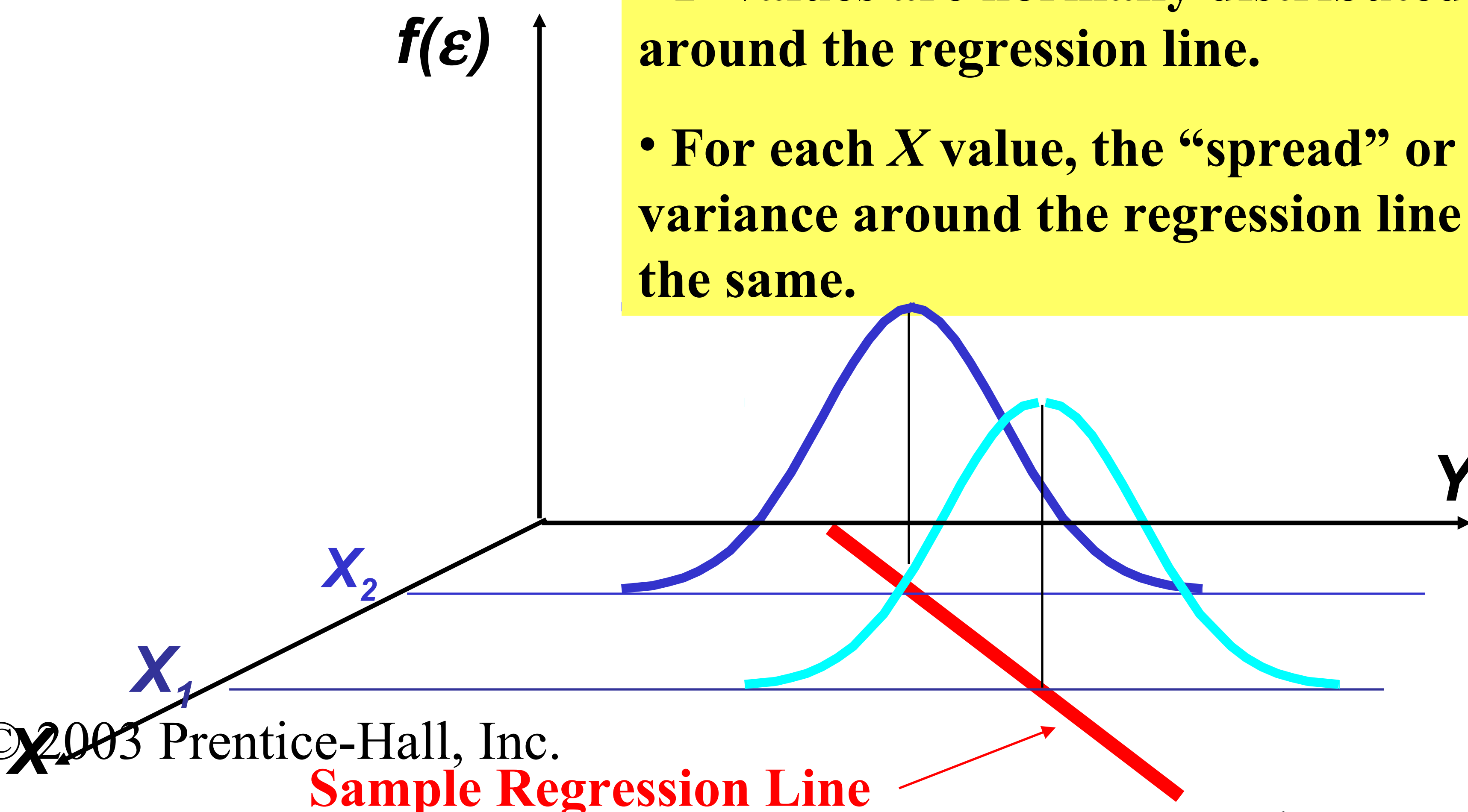
94% of the variation in annual sales can be explained by the variability in the size of the store as measured by square footage

Linear Regression Assumptions

- Normality
  - Y values are normally distributed for each *X*
  - Probability distribution of error is normal
- 2. Homoscedasticity (Constant Variance)
- 3. Independence of Errors

Variation of Errors Around the Regression Line

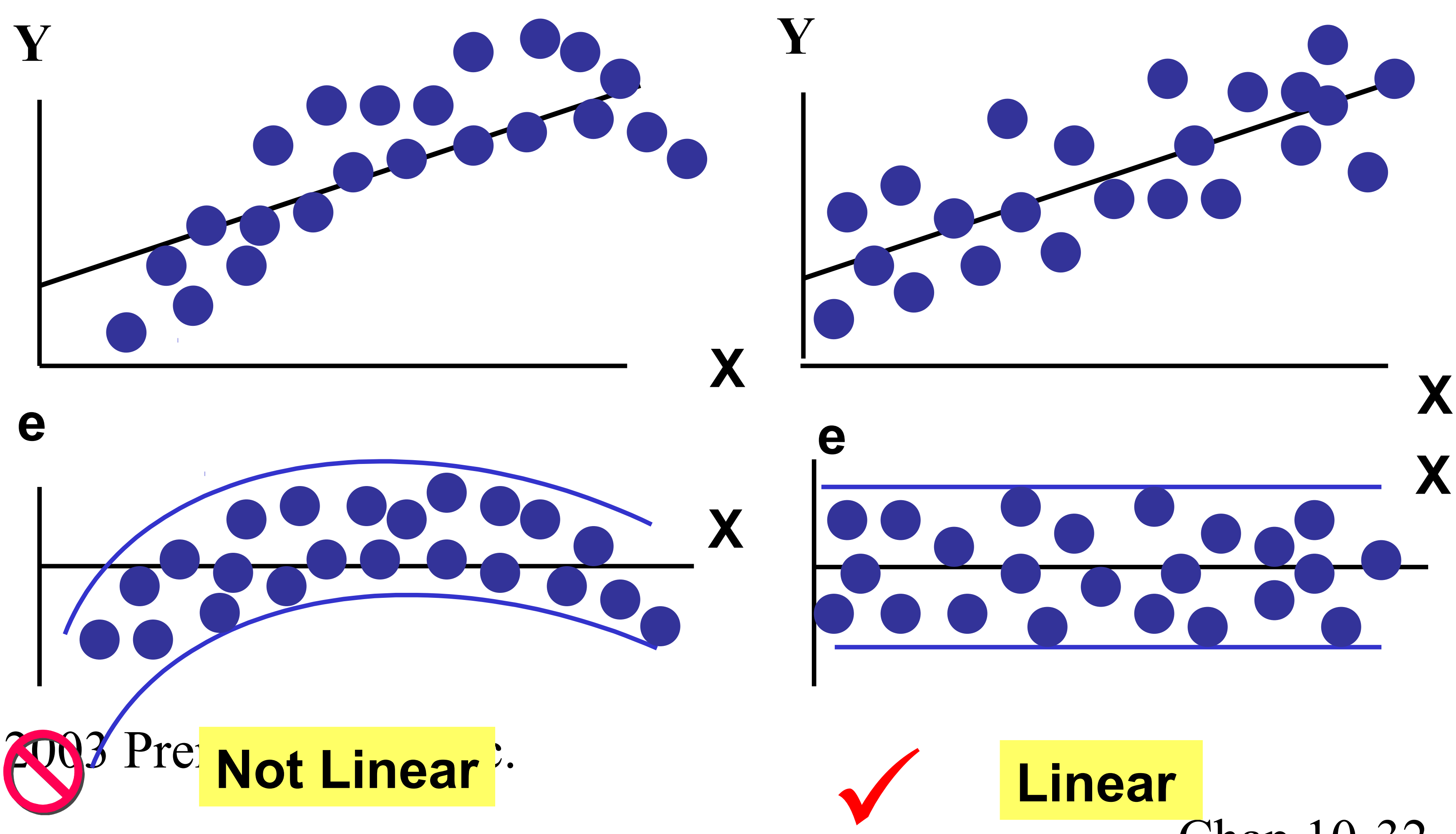
- Y values are normally distributed around the regression line.
- For each X value, the "spread" or variance around the regression line is the same.



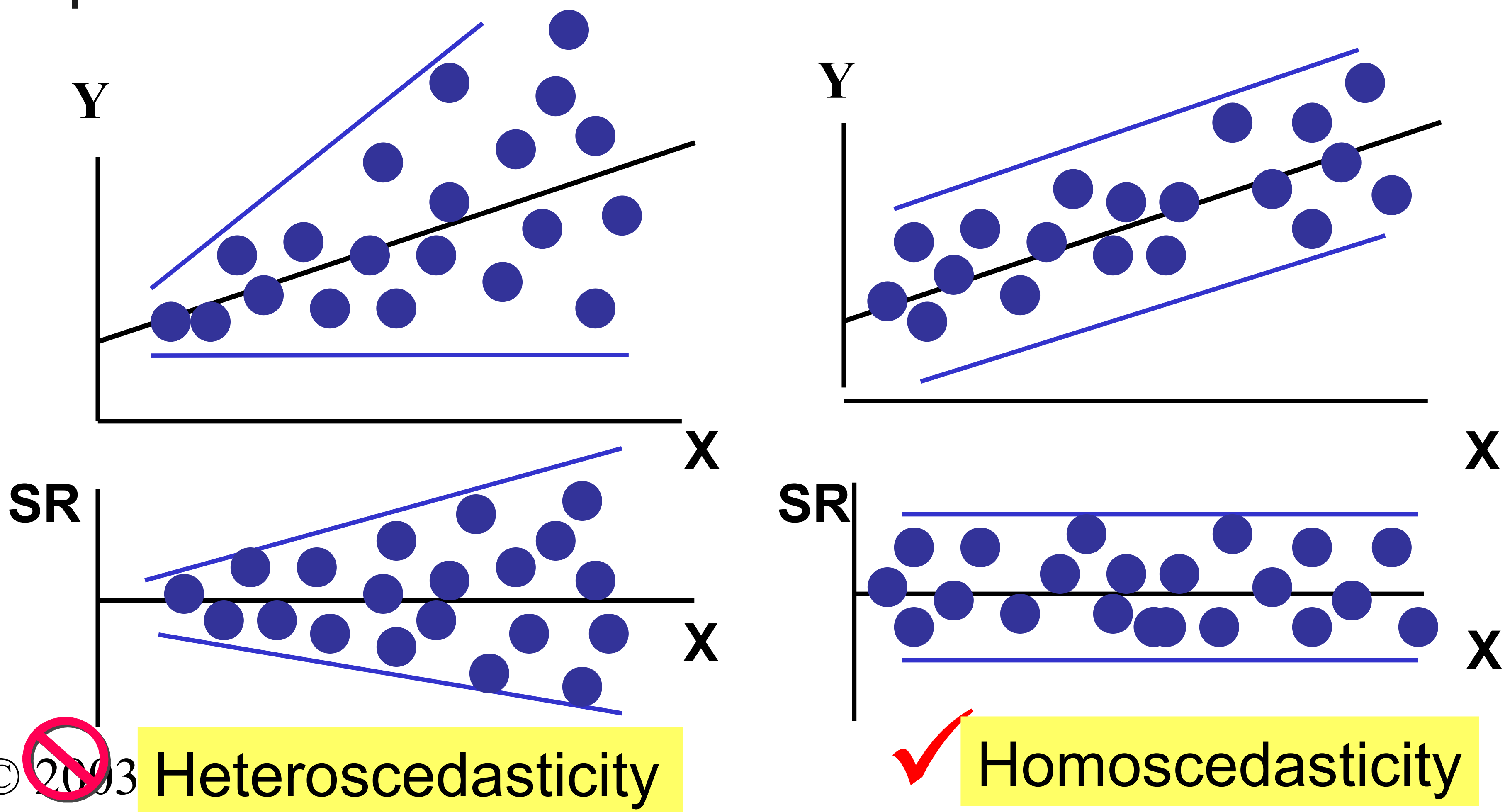
Residual Analysis

- Purposes
  - Examine linearity
  - Evaluate violations of assumptions
- Graphical Analysis of Residuals
  - Plot residuals vs. X and time

Residual Analysis for Linearity



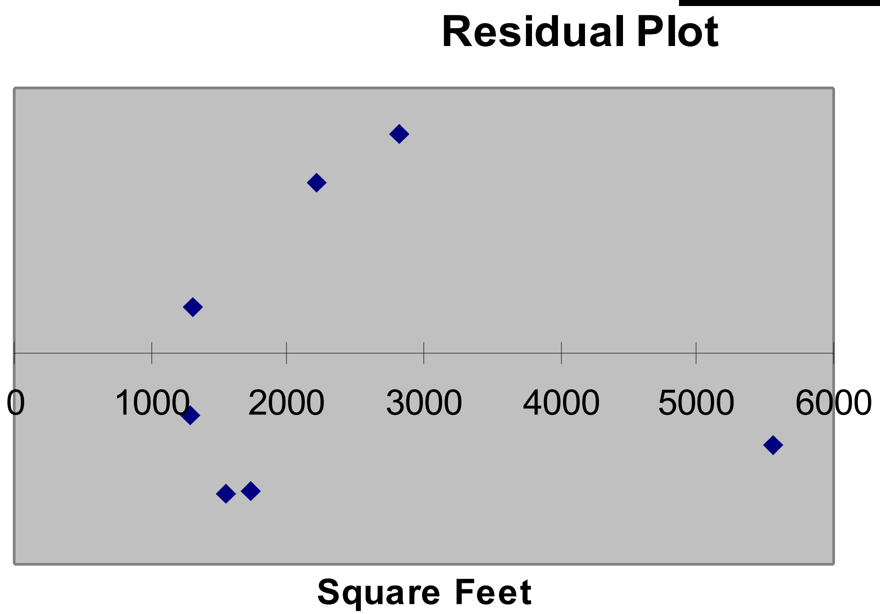
Residual Analysis for Homoscedasticity



Residual Analysis:Excel Output for Produce Stores Example

Excel Output

Observation	*Predicted Y*	*Residuals*
1	4202.344417	-521.3444173
2	3928.803824	-533.8038245
3	5822.775103	830.2248971
4	9894.664688	-351.6646882
5	3557.14541	-239.1454103
6	4918.90184	644.0981603
7	3588.364717	171.6352829



Residual Analysis for Independence

- The Durbin-Watson Statistic
  - Used when data is collected over time to detect autocorrelation (residuals in one time period are related to residuals in another period)
  - Measures violation of independence assumption

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Should be close to 2.

If not, examine the model for autocorrelation.

Durbin-Watson Statistic in PHStat

- PHStat | Regression | Simple Linear Regression …
  - Check the box for Durbin-Watson Statistic

Obtaining the Critical Values of Durbin-Watson Statistic

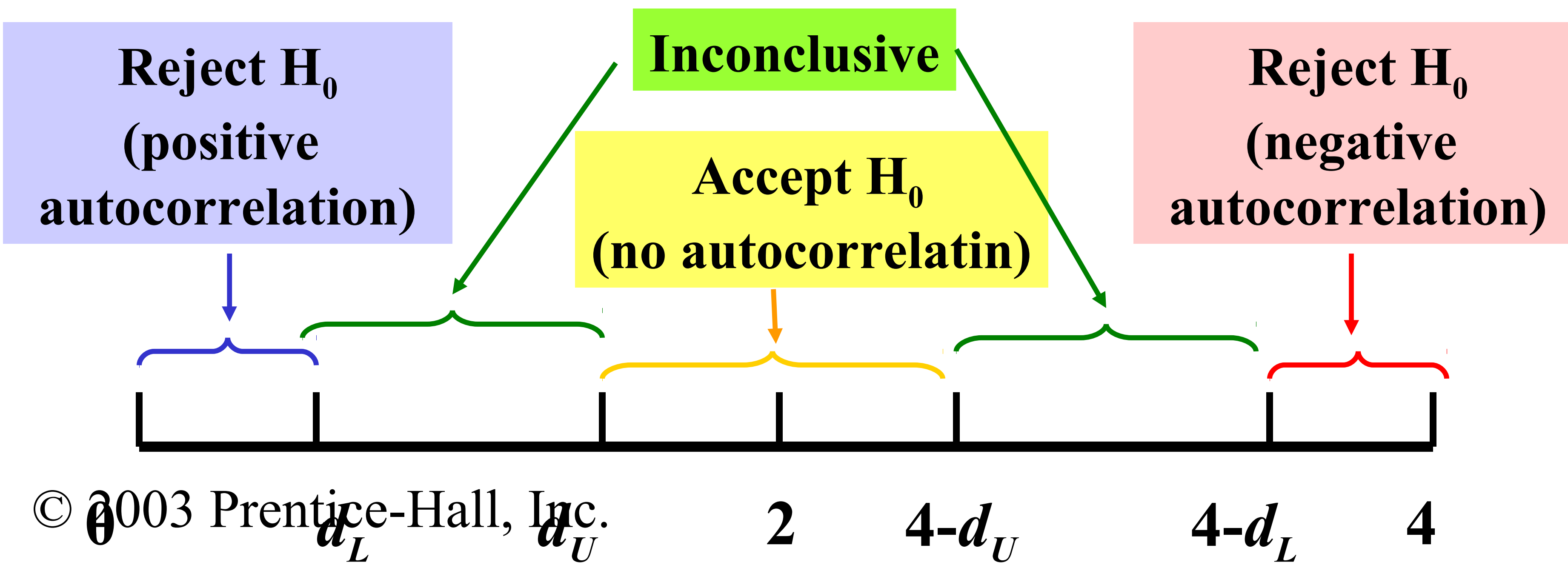
Table 13.4 Finding critical values of Durbin-Watson Statistic

$\alpha = .05$				
	k=1		***k=2***	
n	d_L	d_U	d_L	d_U
15	**1.08**	**1.36**	**.95**	**1.54**
16	**1.10**	**1.37**	**.98**	**1.54**

Using the Durbin-Watson Statistic

H_0: No autocorrelation (error terms are independent)

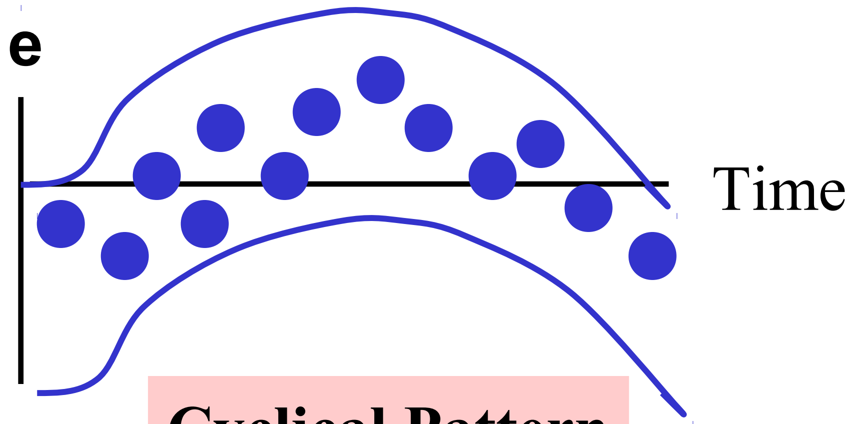
H_1: There is autocorrelation (error terms are not independent)



Residual Analysis for Independence

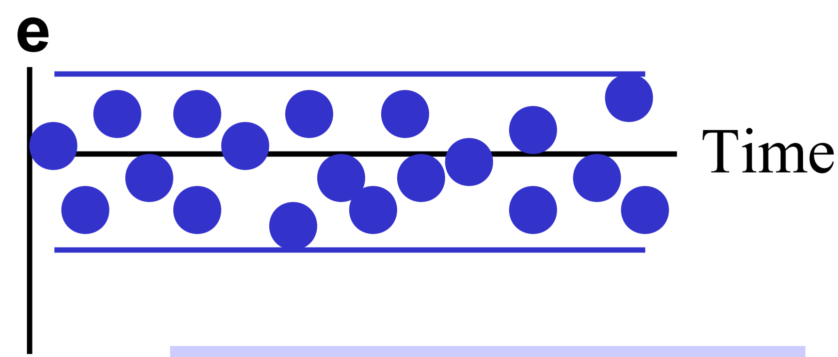
Graphical Approach

Not Independent



Cyclical Pattern

✓ **Independent**



No Particular Pattern

Residual Is Plotted Against Time to Detect Any Autocorrelation

Inference about the Slope: *t* Test

- *t* Test for a Population Slope
  - Is there a linear dependency of *Y* on *X* ?
- Null and Alternative Hypotheses
  - H_0: $\beta_1 = 0$ (No Linear Dependency)
  - H_1: $\beta_1 \neq 0$ (Linear Dependency)
- Test Statistic
  - $$t = \frac{b_1 - \beta_1}{S_{b_1}} \text{ where } S_{b_1} = \frac{S_{YX}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$
  - $d.f. = n - 2$

Example: Produce Store

Data for 7 Stores:

Store	Square Feet	Annual Sales ($000)
1	1,726	3,681
2	1,542	3,395
3	2,816	6,653
4	5,555	9,543
5	1,292	3,318
6	2,208	5,563
7	1,313	3,760

Estimated Regression Equation:

$$\hat{Y} = 1636.415 + 1.487 X_i$$

The slope of this model is 1.487.

Does Square Footage Affect Annual Sales?

Inferences about the Slope: t Test Example

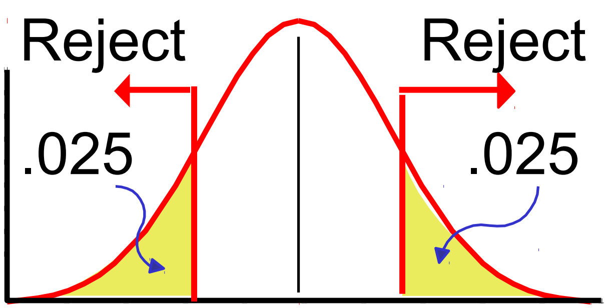
H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

$\alpha = .05$

df = 7 - 2 = 5

Critical Value(s):



-2.5706 0 2.5706 t

Test Statistic:

From Excel Printout b_1 S_{b1} t

	Coefficients	Standard Error	t Stat	P-value
Intercept	1636.4147	451.4953	3.6244	0.01515
Footage	1.4866	0.1650	9.0099	0.00028

Decision:

Reject H_0

Conclusion:

There is evidence that square footage affects annual sales.

Inferences about the Slope: Confidence Interval Example

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{n-2} S_{b_1}$$

Excel Printout for Produce Stores

	Lower 95%	*Upper 95%*
Intercept	475.810926	2797.01853
X Variable	1.06249037	1.91077694

At 95% level of confidence the confidence interval for the slope is (1.062, 1.911). Does not include 0.

Conclusion: There is a significant linear dependency of annual sales on the size of the store.

Inferences about the Slope: *F* Test

- F Test for a Population Slope
  - Is there a linear dependency of *Y* on *X* ?
- Null and Alternative Hypotheses
  - H_0: $\beta_1 = 0$ (No Linear Dependency)
  - H_1: $\beta_1 \neq 0$ (Linear Dependency)
- Test Statistic
  - $$F = \frac{\frac{SSR}{1}}{\frac{SSE}{(n-2)}}$$
  - Numerator *d.f.=1*, denominator *d.f.=n-2*

Relationship between a t Test and an F Test

- Null and Alternative Hypotheses
  - H_0: $\beta_1 = 0$ (No Linear Dependency)
  - H_1: $\beta_1 \neq 0$ (Linear Dependency)

$$\left(t_{n-2} \right)^2 = F_{1,n-2}$$

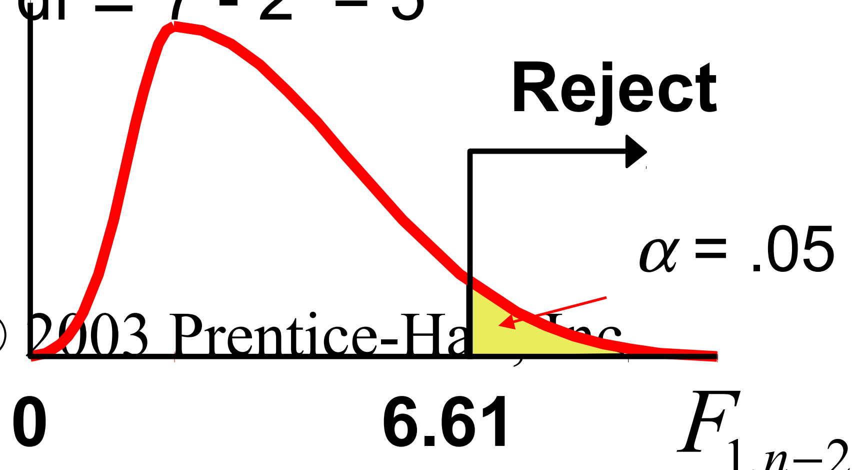
Inferences about the Slope: *F* Test Example

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$
$\alpha = .05$
numerator
df = 1
denominator
df = 7 - 2 = 5

Test Statistic:
From Excel Printout

ANOVA

	df	SS	MS	F	Significance F
Regression	1	30380456.12	30380456.12	81.179	0.000281
Residual	5	1871199.595	374239.919		
Total	6	32251655.71			



Decision: Reject H_0

Conclusion:

There is evidence that square footage affects annual sales.

Purpose of Correlation Analysis

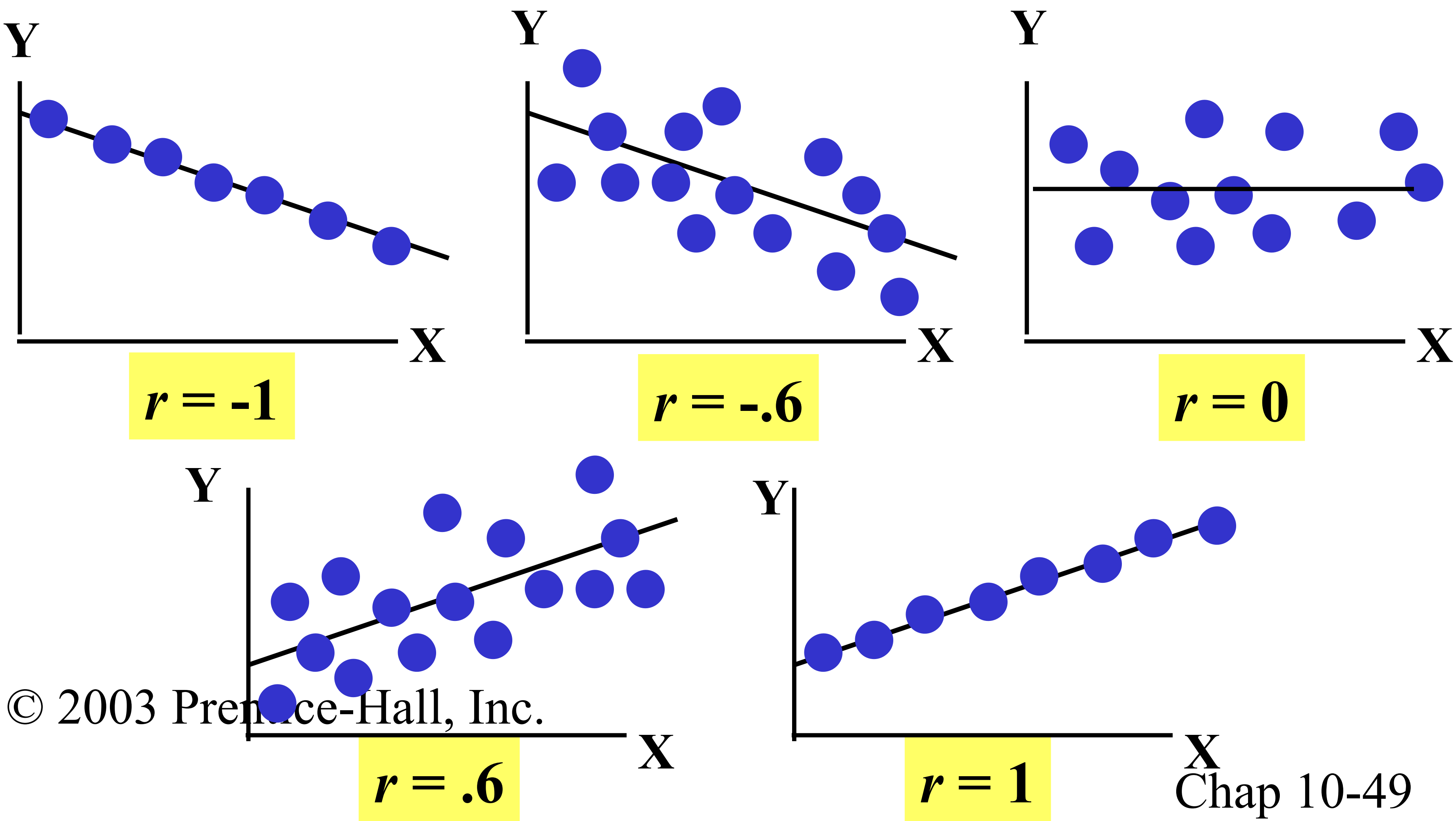
- Correlation Analysis is Used to Measure Strength of Association (Linear Relationship) Between 2 Numerical Variables
  - Only Strength of the Relationship is Concerned
  - No Causal Effect is Implied

Purpose of Correlation Analysis

(continued)

- Population Correlation Coefficient ρ (Rho) is Used to Measure the Strength between the Variables
- Sample Correlation Coefficient r is an Estimate of ρ and is Used to Measure the Strength of the Linear Relationship in the Sample Observations

Sample of Observations from Various *r* Values



Features of ρ and r

- Unit Free
- Range between -1 and 1
- The Closer to -1, the Stronger the Negative Linear Relationship
- The Closer to 1, the Stronger the Positive Linear Relationship
- The Closer to 0, the Weaker the Linear Relationship

t Test for Correlation

- Hypotheses
  - H_0: $\rho = 0$ (No Correlation)
  - H_1: $\rho \neq 0$ (Correlation)
- Test Statistic
  - $$t = \frac{r - \rho}{\sqrt{\dfrac{1 - r^2}{n - 2}}} \quad \text{where}$$

$$r = \sqrt{r^2} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2}}$$

Example: Produce Stores

Is there any evidence of linear relationship between Annual Sales of a store and its Square Footage at .05 level of significance?

From Excel Printout

r

Regression Statistics	
Multiple R	0.9705572
R Square	0.94198129
Adjusted R Square	0.93037754
Standard Error	611.751517
Observations	7

H_0: $\rho = 0$ (No association)

H_1: $\rho \neq 0$ (Association)

$\alpha = .05$

df $= 7 - 2 = 5$

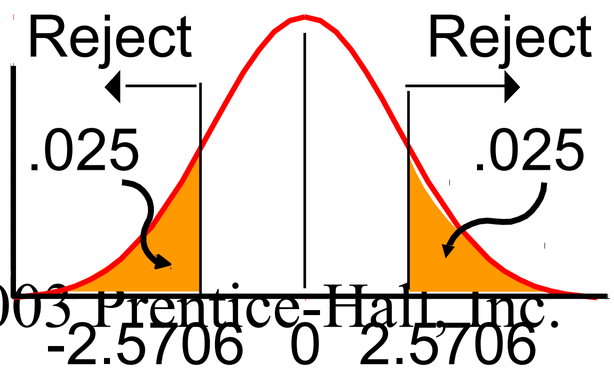
Example: Produce Stores Solution

$$t = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.9706}{\sqrt{\frac{1-.9420}{5}}} = 9.0099$$

Decision:
Reject H_0

Conclusion:
There is evidence of a linear relationship at 5% level of significance

Critical Value(s):



The value of the t statistic is exactly the same as the t statistic value for test on the slope coefficient

Estimation of Mean Values

Confidence Interval Estimate for $\mu_{Y|X=X_i}$:

The Mean of Y given a particular X_i

Size of interval vary according to distance away from mean, $\bar{X}$

Standard error of the estimate

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

t value from table with df=n-2

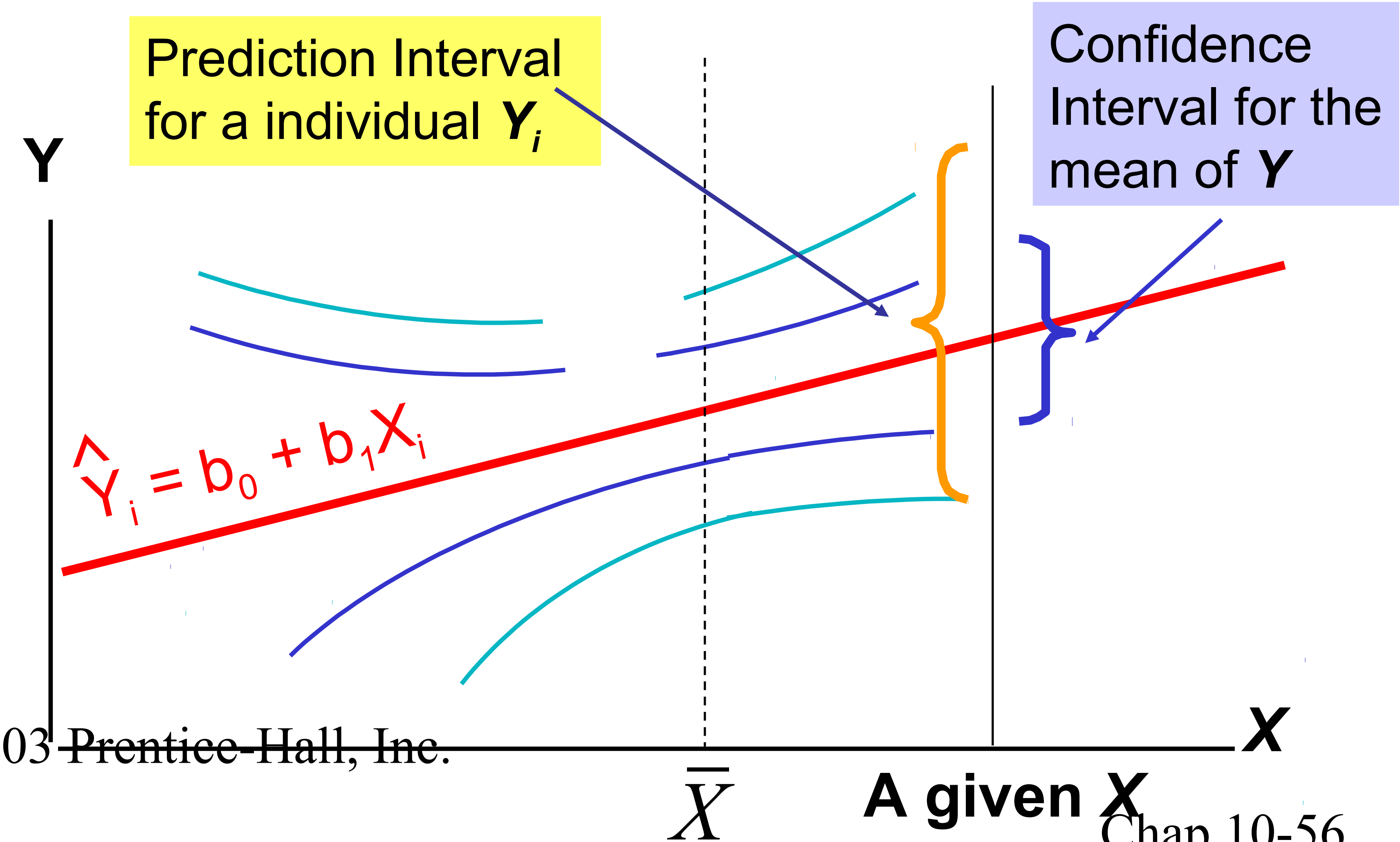
Prediction of Individual Values

Prediction Interval for Individual Response Y_i at a Particular X_i

Addition of 1 increases width of interval from that for the mean of Y

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Interval Estimates for Different Values of X



Example: Produce Stores

Data for 7 Stores:

Store	Square Feet	Annual Sales ($000)
1	1,726	3,681
2	1,542	3,395
3	2,816	6,653
4	5,555	9,543
5	1,292	3,318
6	2,208	5,563
7	1,313	3,760

Consider a store with 2000 square feet.

Regression Equation Obtained:

$$\hat{Y} = 1636.415 + 1.487 X_i$$

Estimation of Mean Values: Example

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

Find the 95% confidence interval for the average annual sales for stores of 2,000 square feet

Predicted Sales $\hat{Y} = 1636.415 + 1.487 X_i = 4610.45 \left(\$000 \right)$

$\bar{X} = 2350.29 \qquad S_{YX} = 611.75 \qquad t_{n-2} = t_5 = 2.5706$

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}} = 4610.45 \pm 612.66$$

Prediction Interval for Y : Example

Prediction Interval for Individual $Y_{X=X_i}$

Find the 95% prediction interval for annual sales of one particular store of 2,000 square feet

Predicted Sales) $\hat{Y} = 1636.415 + 1.487 X_i = 4610.45 (\$000)$

$\bar{X} = 2350.29 \quad S_{YX} = 611.75 \quad t_{n-2} = t_5 = 2.5706$

$$\hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}} = 4610.45 \pm 1687.68$$

Estimation of Mean Values and Prediction of Individual Values in PHStat

- In Excel, use PHStat | Regression | Simple Linear Regression …
  - Check the "Confidence and Prediction Interval for X=" box
- EXCEL Spreadsheet of Regression Sales on Footage

Microsoft Excel Worksheet

Pitfalls of Regression Analysis

- Lacking an Awareness of the Assumptions Underlining Least-squares Regression
- Not Knowing How to Evaluate the Assumptions
- Not Knowing What the Alternatives to Least-squares Regression are if a Particular Assumption is Violated
- Using a Regression Model Without Knowledge of the Subject Matter

Strategy for Avoiding the Pitfalls of Regression

- Start with a scatter plot of X on Y to observe possible relationship
- Perform residual analysis to check the assumptions
- Use a histogram, stem-and-leaf display, box-and-whisker plot, or normal probability plot of the residuals to uncover possible non-normality

Strategy for Avoiding the Pitfalls of Regression

(continued)

- If there is violation of any assumption, use alternative methods (e.g., least absolute deviation regression or least median of squares regression) to least-squares regression or alternative least-squares models (e.g., curvilinear or multiple regression)
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals

Chapter Summary

- Introduced Types of Regression Models
- Discussed Determining the Simple Linear Regression Equation
- Described Measures of Variation
- Addressed Assumptions of Regression and Correlation
- Discussed Residual Analysis
- Addressed Measuring Autocorrelation

Chapter Summary

(continued)

- Described Inference about the Slope
- Discussed Correlation - Measuring the Strength of the Association
- Addressed Estimation of Mean Values and Prediction of Individual Values
- Discussed Pitfalls in Regression and Ethical Issues