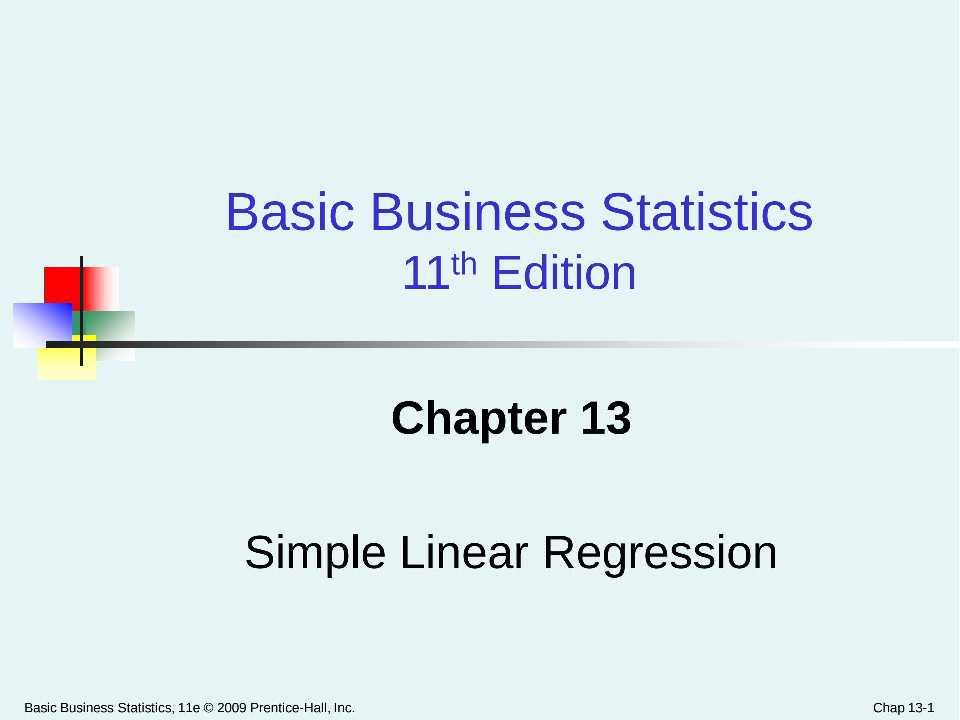
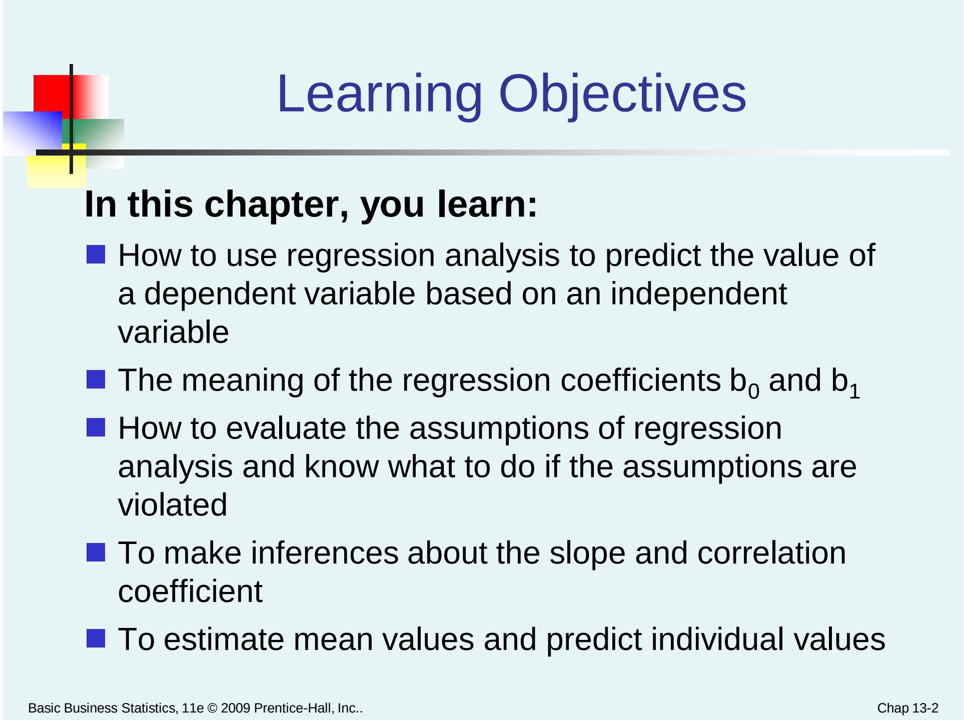


Basic Business Statistics
11th Edition

Chapter 13

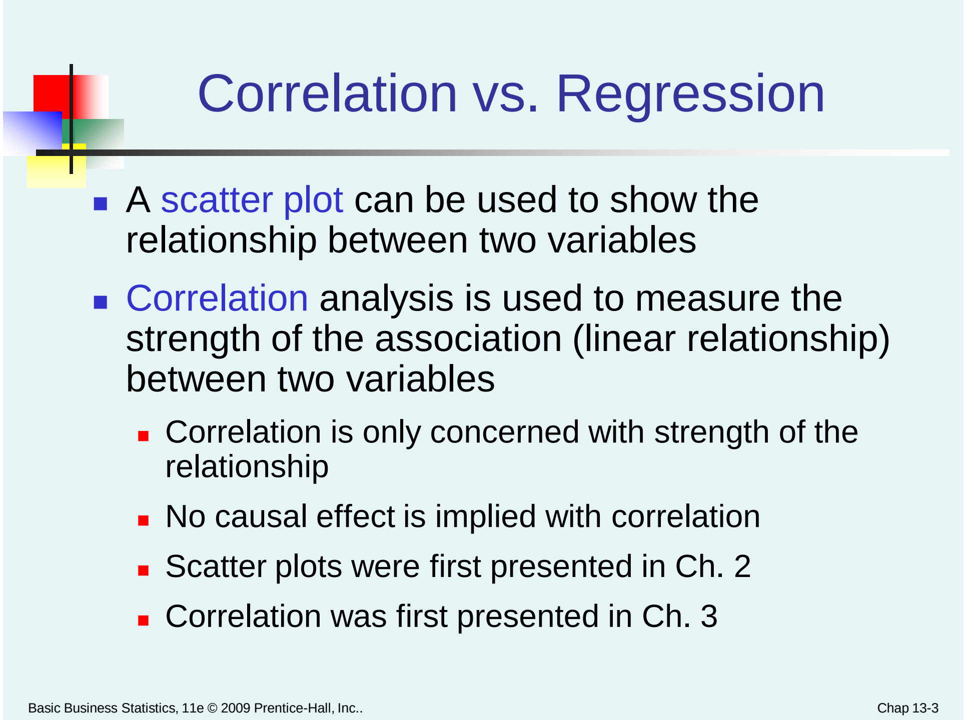
Simple Linear Regression



Learning Objectives

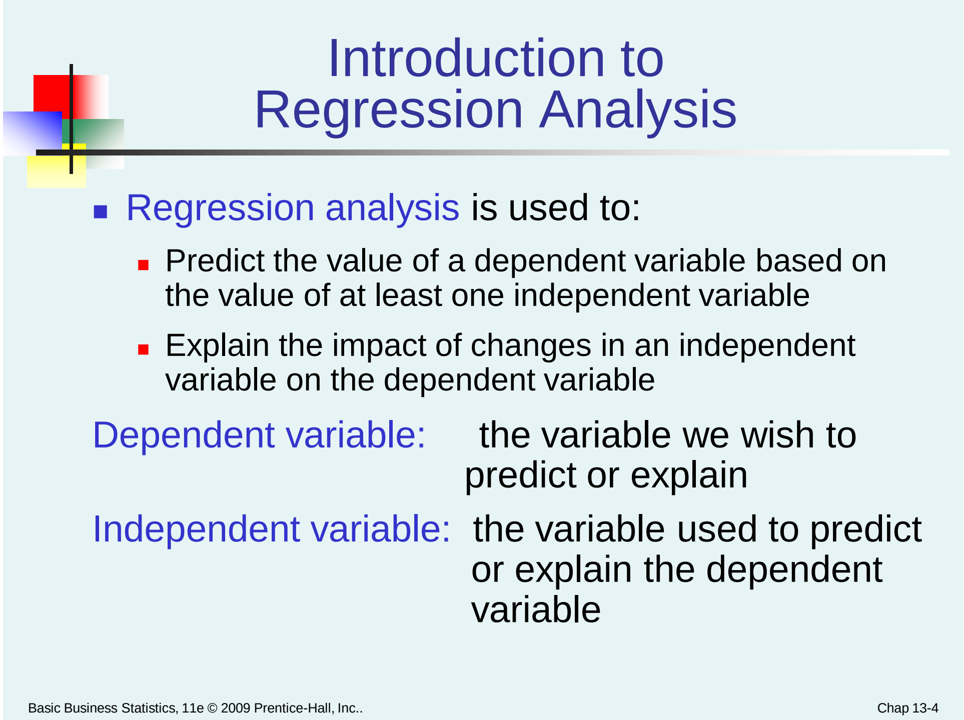
In this chapter, you learn:

- How to use regression analysis to predict the value of a dependent variable based on an independent variable
- The meaning of the regression coefficients b_0 and b_1
- How to evaluate the assumptions of regression analysis and know what to do if the assumptions are violated
- To make inferences about the slope and correlation coefficient
- To estimate mean values and predict individual values



Correlation vs. Regression

- A scatter plot can be used to show the relationship between two variables
- Correlation analysis is used to measure the strength of the association (linear relationship) between two variables
  - Correlation is only concerned with strength of the relationship
  - No causal effect is implied with correlation
  - Scatter plots were first presented in Ch. 2
  - Correlation was first presented in Ch. 3

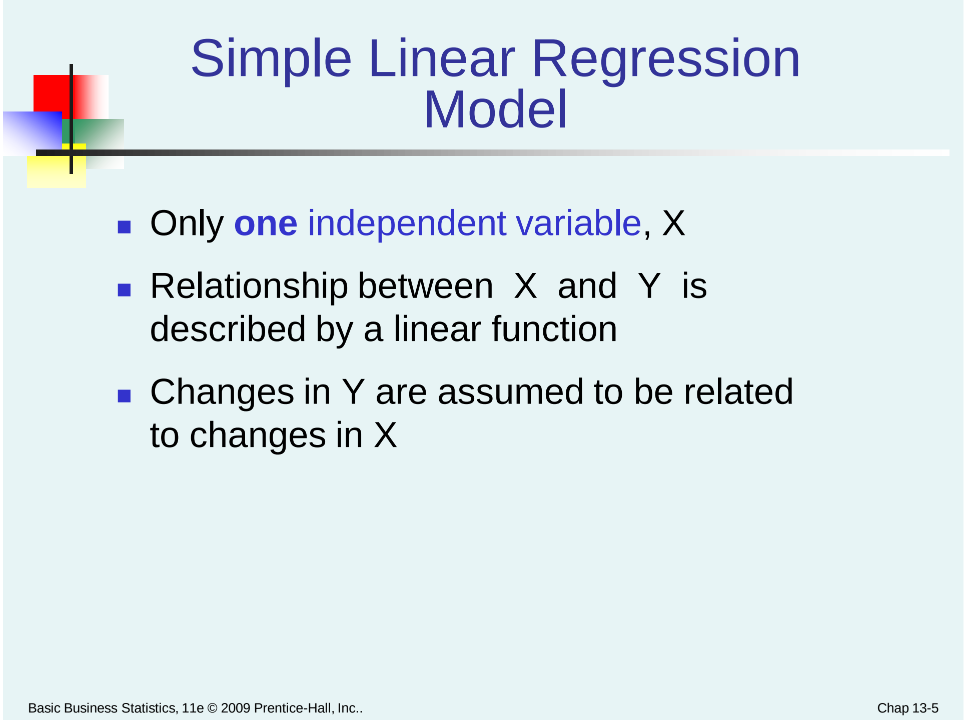


Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

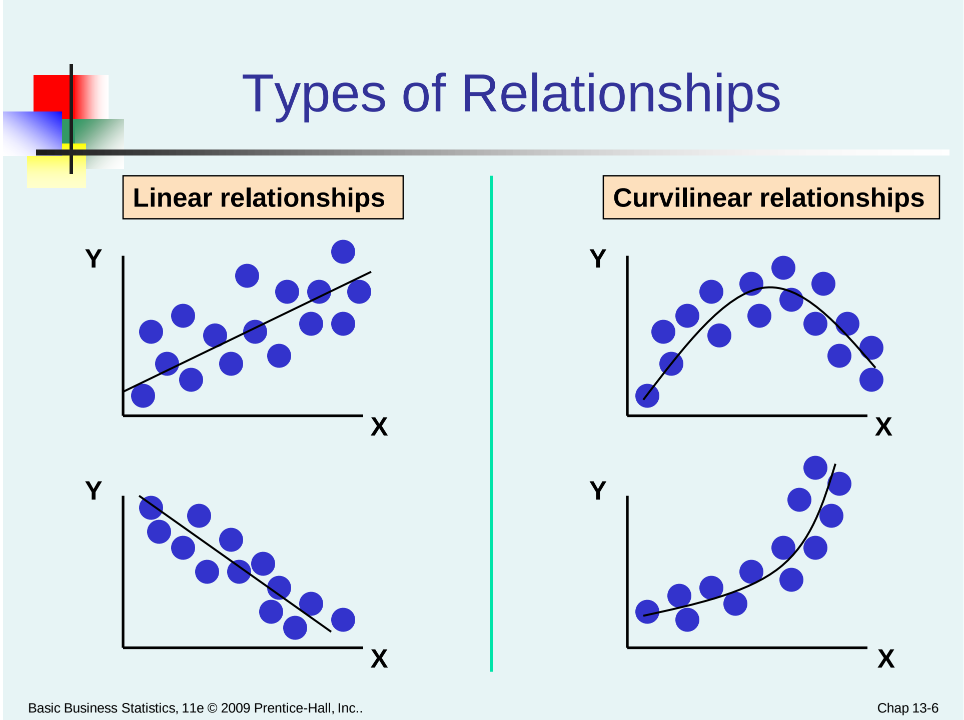
Dependent variable: the variable we wish to predict or explain

Independent variable: the variable used to predict or explain the dependent variable

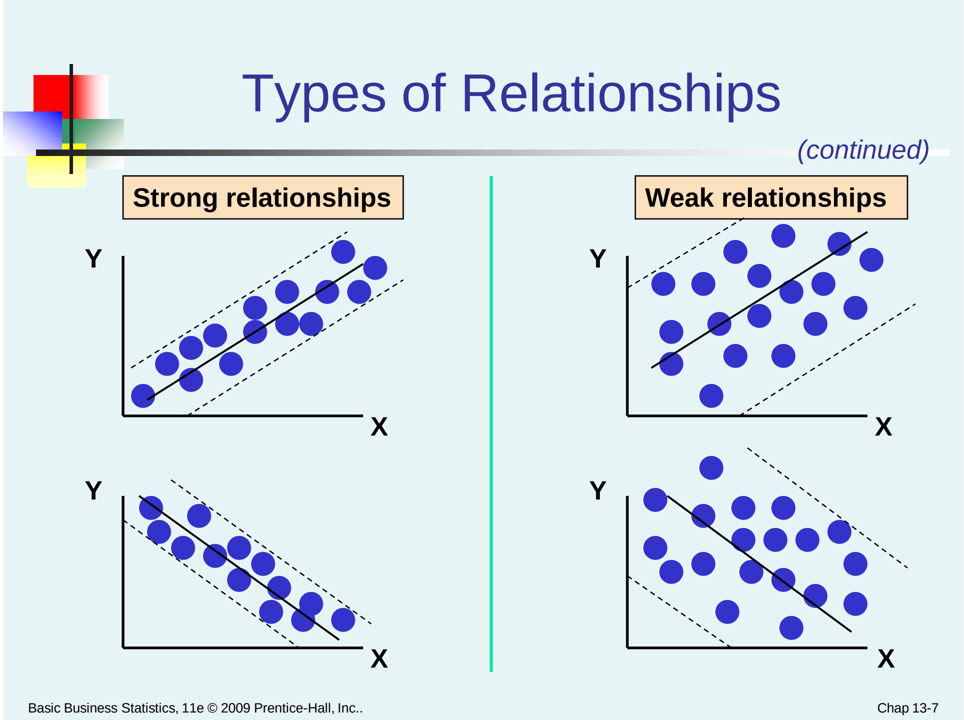


Simple Linear Regression Model

- Only **one independent variable**, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be related to changes in X



Types of Relationships



Types of Relationships

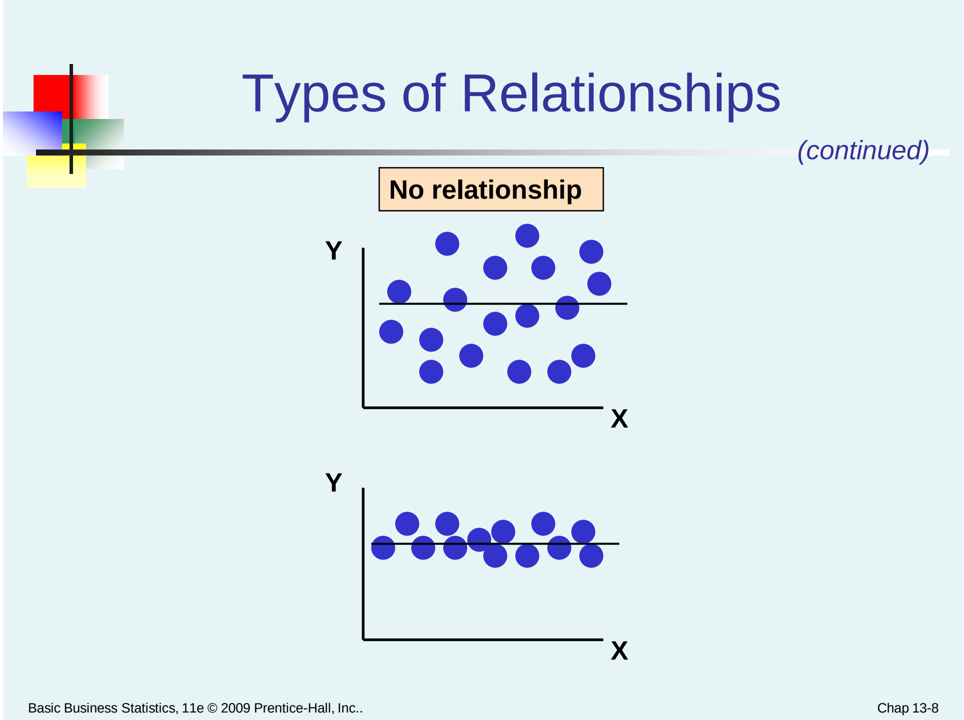
(continued)

Strong relationships

Weak relationships

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc..

Chap 13-7



Types of Relationships

(continued)

No relationship

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc..

Chap 13-8

Correlation Coefficient

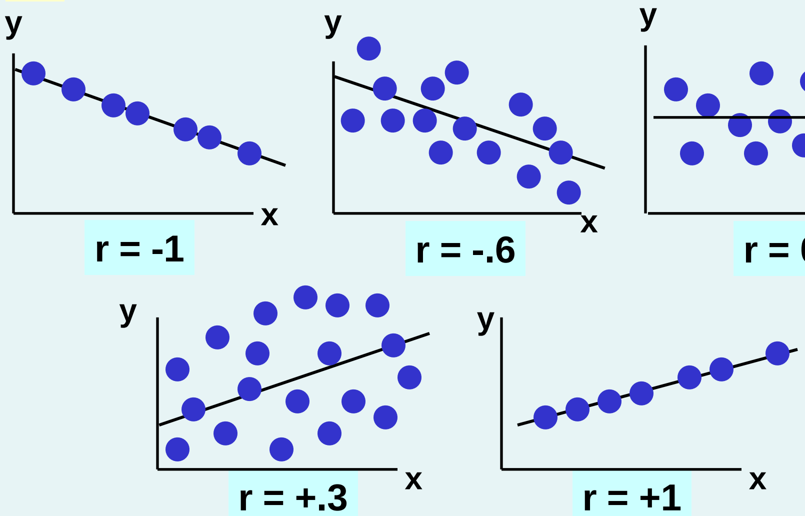
(continued)

- The population correlation coefficient ρ (rho) measures the strength of the association between the variables
- The sample correlation coefficient r is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations

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

Examples of Approximate r Values



Calculating the Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{[\sum(x-\bar{x})^2][\sum(y-\bar{y})^2]}}$$

or the algebraic equivalent:

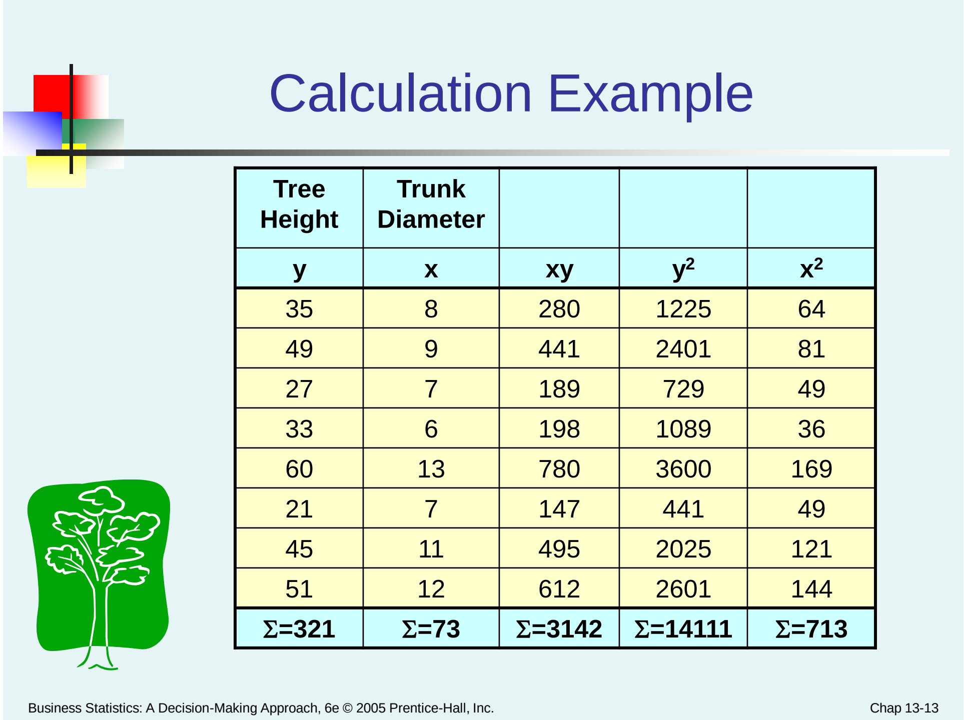
$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2)-(\sum x)^2][n(\sum y^2)-(\sum y)^2]}}$$

where:

r = Sample correlation coefficient

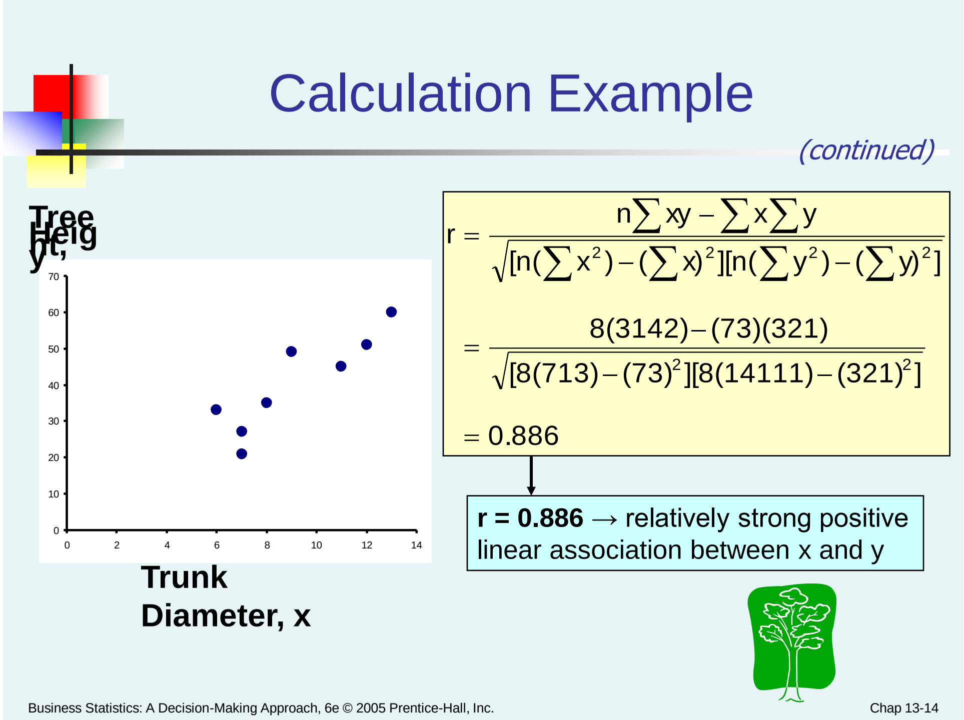
n = Sample size

x = Value of the independent



Calculation Example

Tree Height	Trunk Diameter			
y	**x**	**xy**	**y^2**	**x^2**
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
Σ=321	**Σ=73**	**Σ=3142**	**Σ=14111**	**Σ=713**



Calculation Example

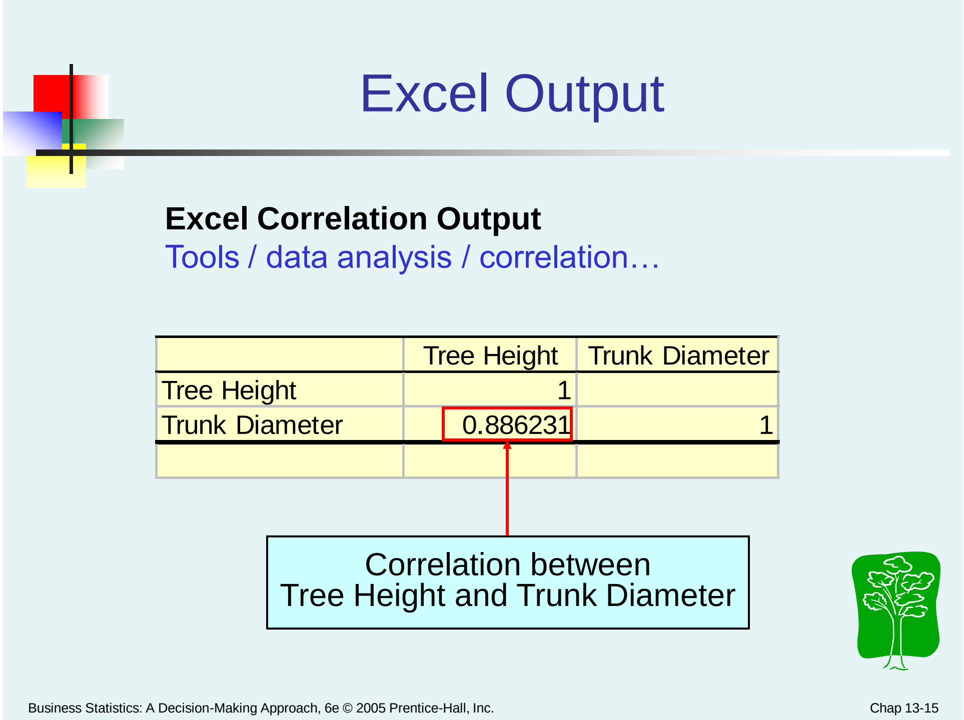
(continued)

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$

$$= 0.886$$

r = 0.886 → relatively strong positive linear association between x and y



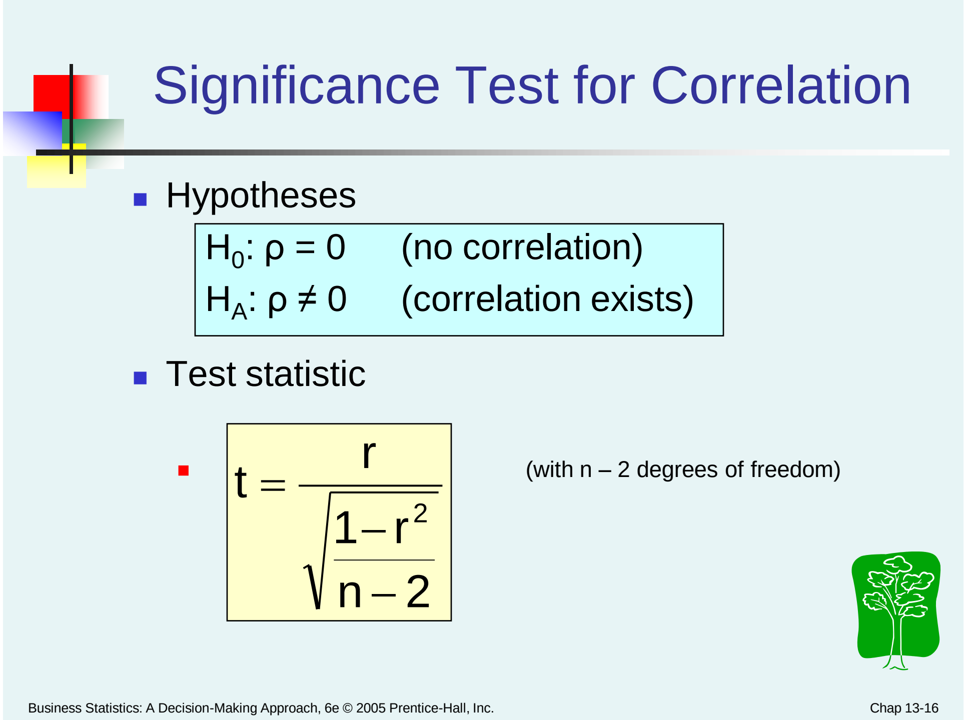
Excel Output

Excel Correlation Output

Tools / data analysis / correlation…

	Tree Height	Trunk Diameter
Tree Height	1	
Trunk Diameter	0.886231	1

Correlation between
Tree Height and Trunk Diameter



Significance Test for Correlation

- Hypotheses

H_0: $\rho = 0$ (no correlation)

H_A: $\rho \neq 0$ (correlation exists)

- Test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

(with n – 2 degrees of freedom)

Example: Produce Stores

Is there evidence of a linear relationship between tree height and trunk diameter at the .05 level of significance?

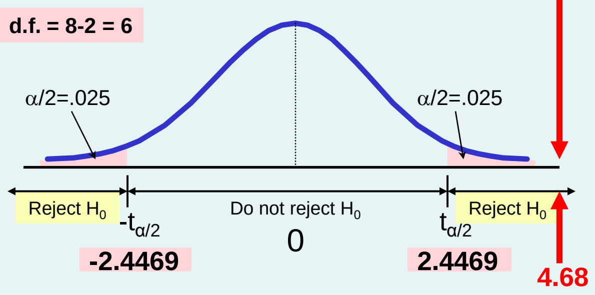
H_0: ρ = 0 (No correlation)
H_1: ρ ≠ 0 (correlation exists)

α =.05 , df = 8 - 2 = 6

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$

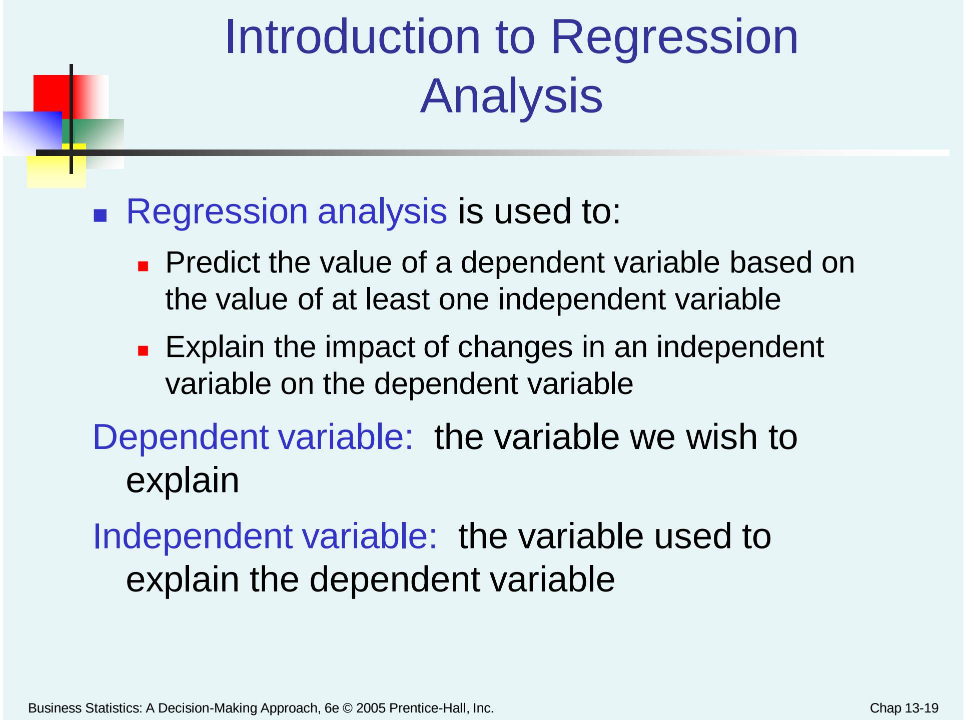
Example: Test Solution

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = \boxed{4.68}$$



Decision:
Reject H_0

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

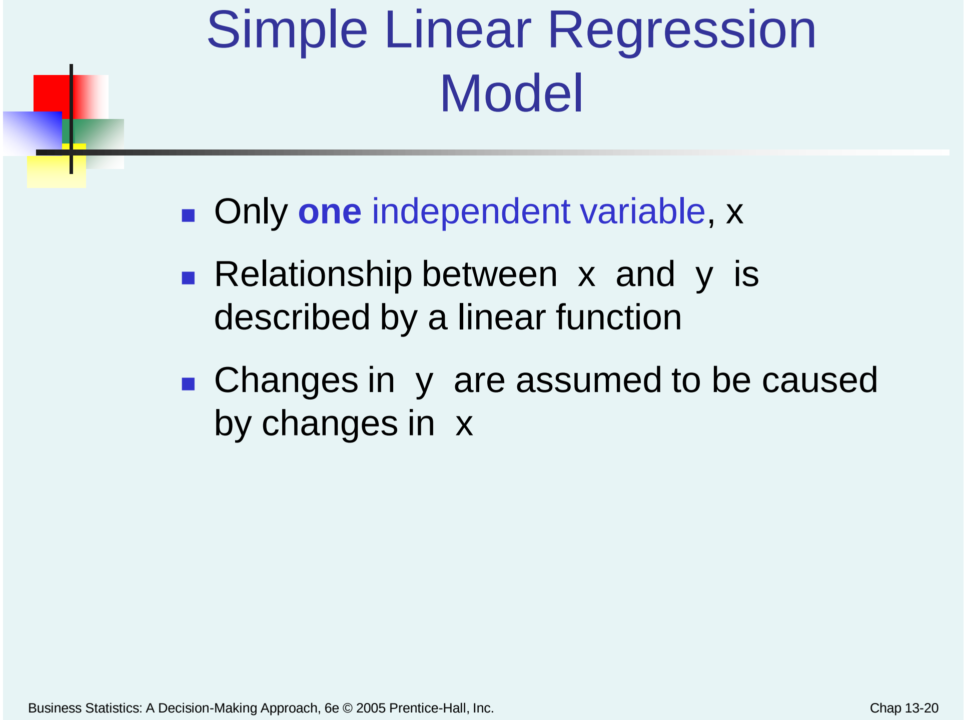


Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

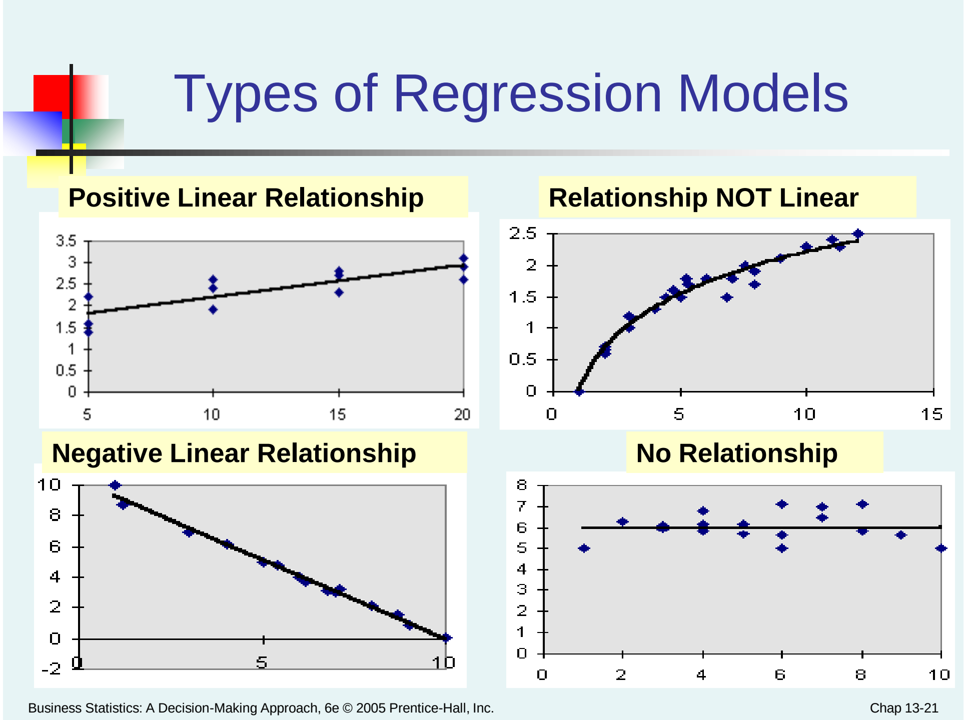
Dependent variable: the variable we wish to explain

Independent variable: the variable used to explain the dependent variable



Simple Linear Regression Model

- Only **one** independent variable, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be caused by changes in x



Types of Regression Models

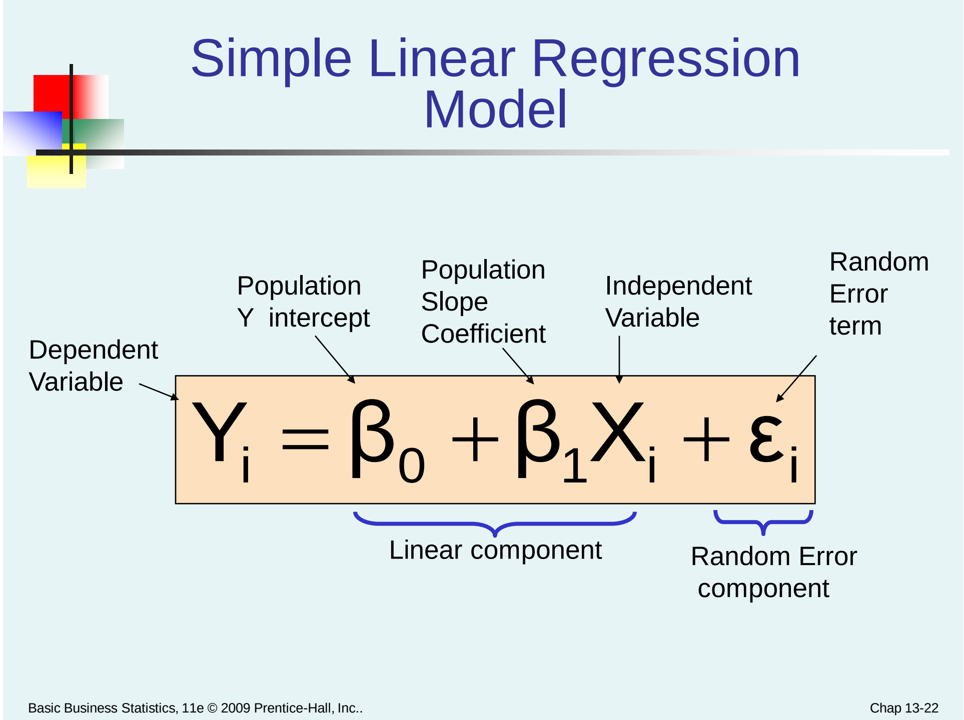
Positive Linear Relationship

Relationship NOT Linear

Negative Linear Relationship

No Relationship

Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc. Chap 13-21



Simple Linear Regression Model

Dependent Variable

Population Y intercept

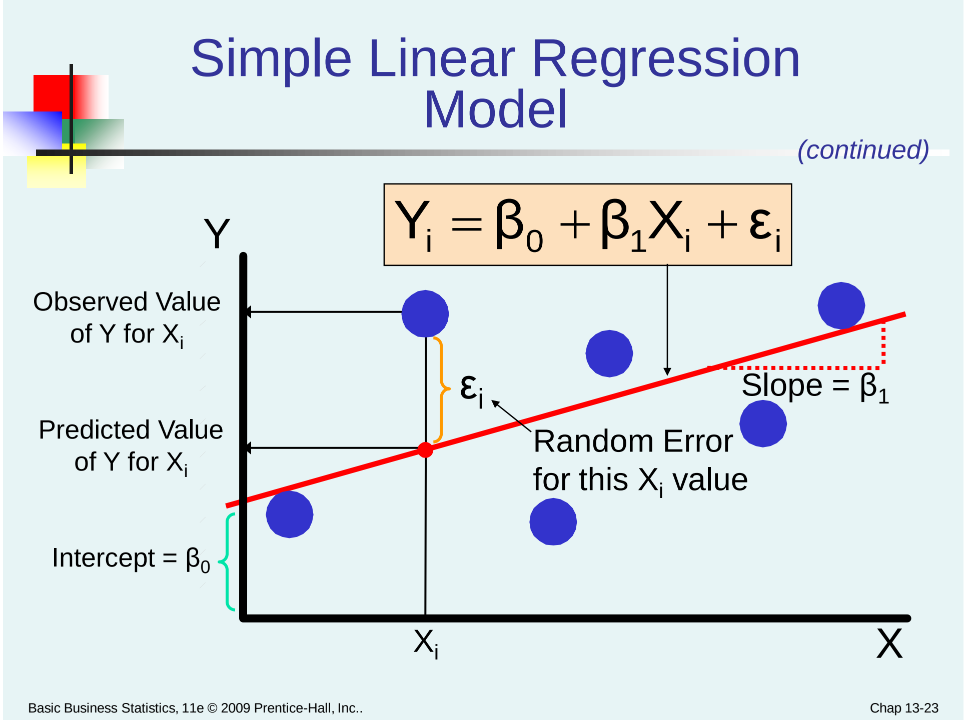
Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

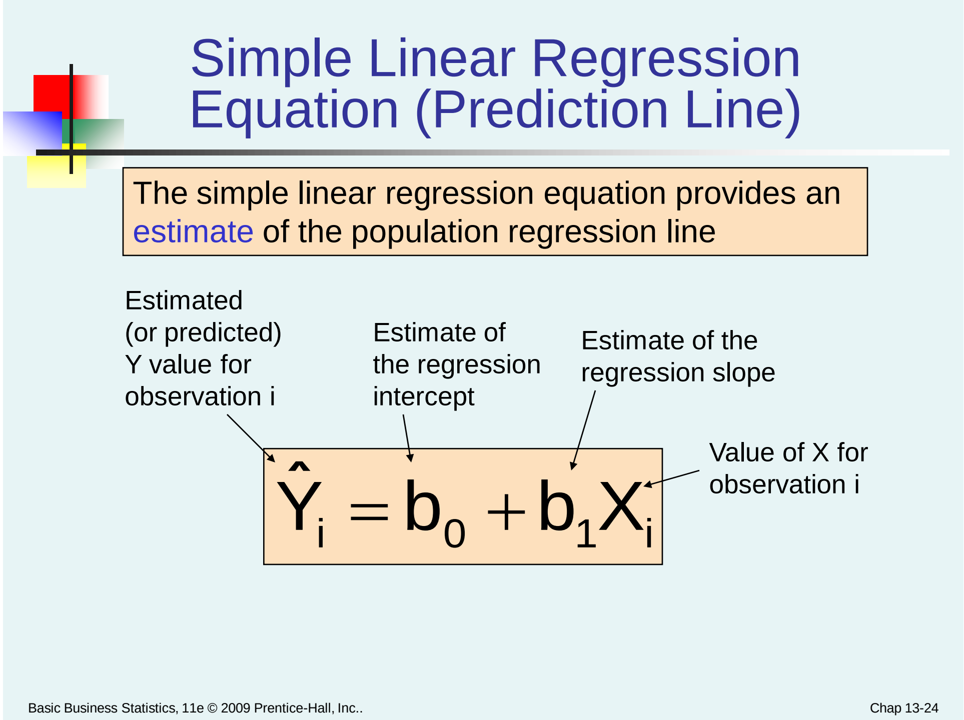
Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 13-22



Simple Linear Regression Model

(continued)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

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

Least Squares Criterion

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

$$\begin{aligned}\sum e^2 &= \sum (y - \hat{y})^2 \\ &= \sum (y - (b_0 + b_1 x))^2\end{aligned}$$

The Least Squares Equation

- The formulas for b_1 and b_0 are:

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

algebraic equivalent:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

The Least Squares Method

b_0 and b_1 are obtained by finding the values of that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$\boxed{\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2}$$

Finding the Least Squares Equation

- The coefficients b_0 and b_1 , and other regression results in this chapter, will be found using Excel or Minitab

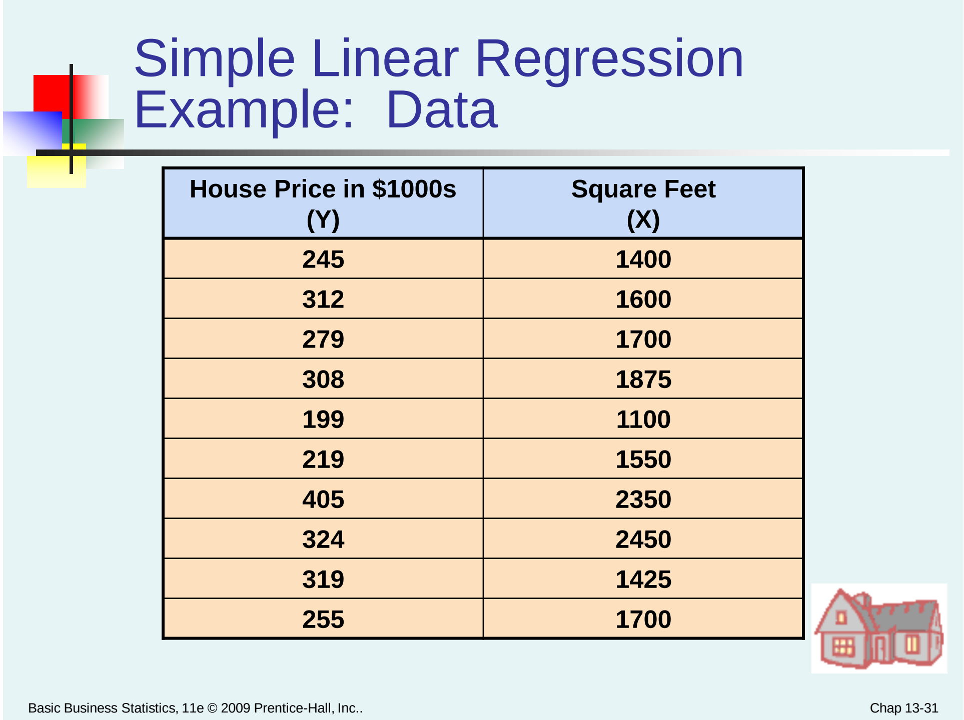
Formulas are shown in the text for those who are interested

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of Y when the value of X is zero
- b_1 is the estimated change in the average value of Y as a result of a one-unit change in X

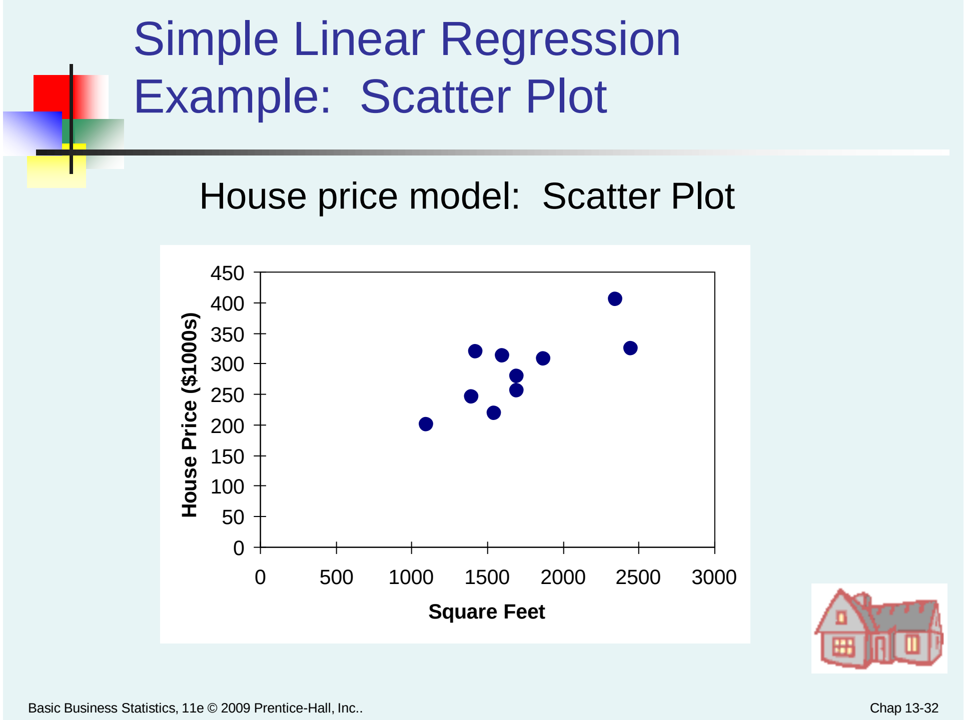
Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet



Simple Linear Regression Example: Data

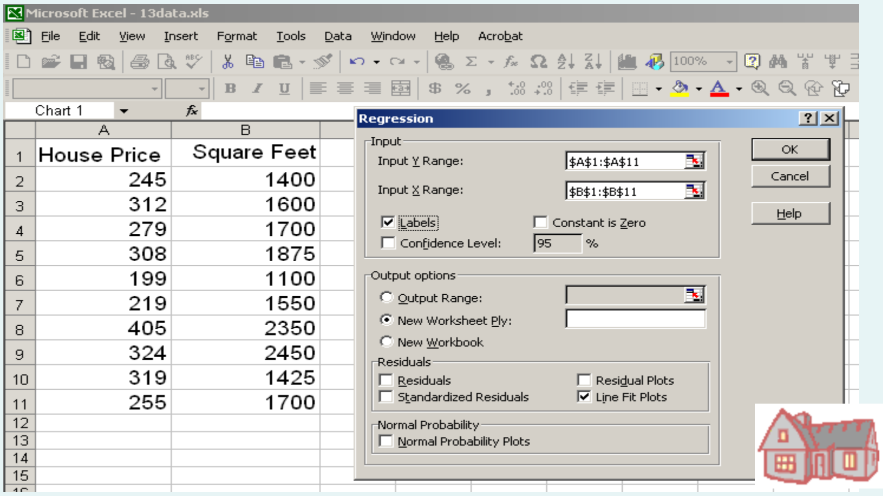
House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



Simple Linear Regression Example: Scatter Plot

House price model: Scatter Plot

Simple Linear Regression Example: Using Excel



	House Price	Square Feet
2	245	1400
3	312	1600
4	279	1700
5	308	1875
6	199	1100
7	219	1550
8	405	2350
9	324	2450
10	319	1425
11	255	1700

Input

Input Y Range: $A$1:$A$11

Input X Range: $B$1:$B$11

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☑ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

Simple Linear Regression Example: Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{house}price = 98.24833 + 0.10977\,(square feet)$$

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Simple Linear Regression Example: Minitab Output

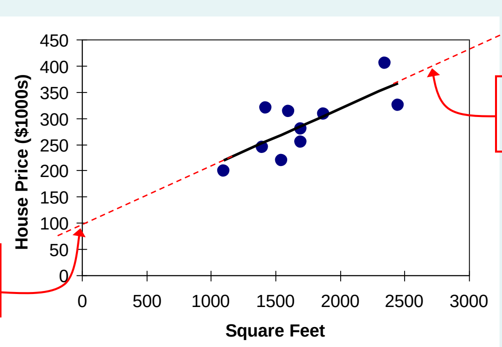
```
The regression equation is

Price = 98.2 + 0.110 Square Feet

Predictor      Coef       SE Coef      T      P
Constant       98.25      58.03        1.69  0.129
Square Feet  0.10977   0.03297      3.33  0.010

S = 41.3303   R-Sq = 58.1%   R-Sq(adj) = 52.8%

Analysis of Variance

Source              DF     SS       MS      F         P
Regression        1       18935  18935  11.08  0.010
Residual Error  8       13666    1708
Total               9        32600
```

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977 \text{ (square feet)}$$

Simple Linear Regression Example: Graphical Representation

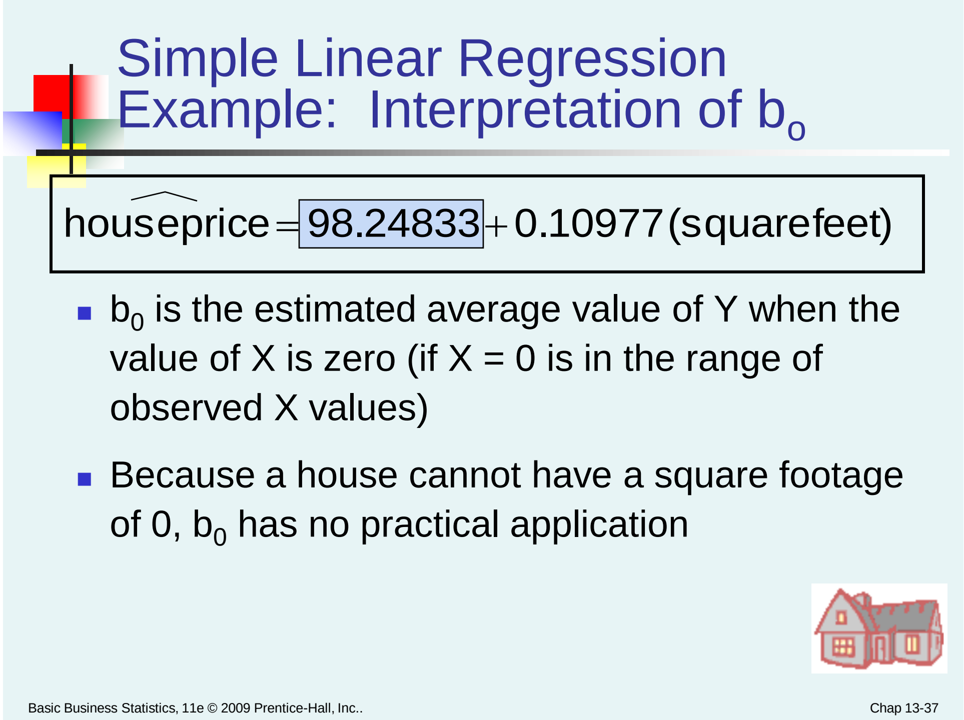
House price model: Scatter Plot and Prediction Line



Slope = 0.10977

Intercept = 98.248

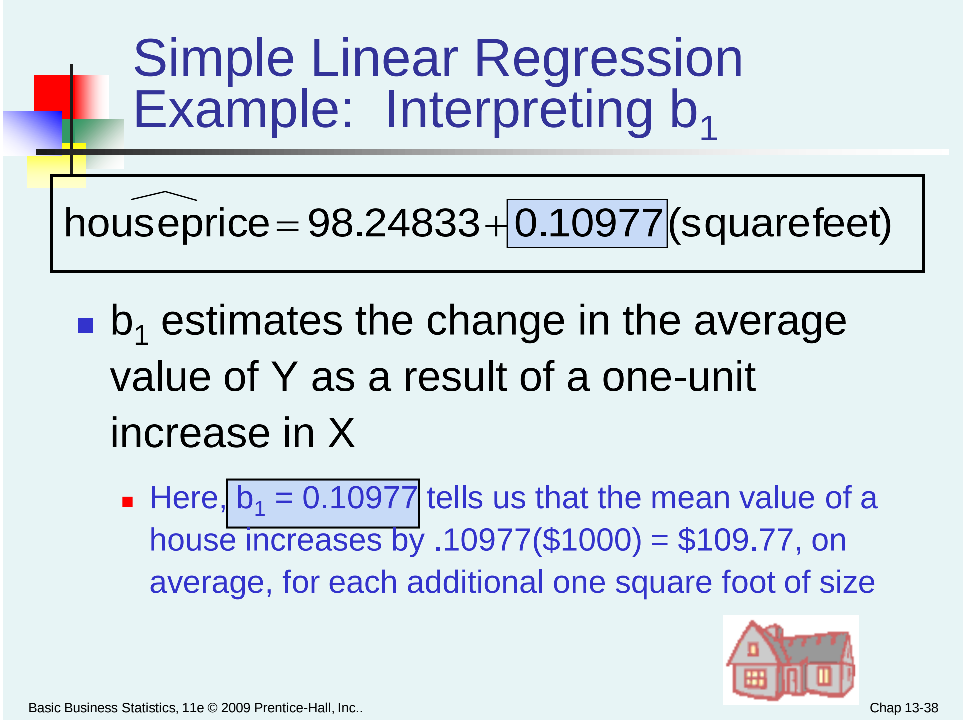
$$\widehat{\text{houseprice}} = 98.24833 + 0.10977 (\text{squarefeet})$$



Simple Linear Regression Example: Interpretation of b_o

$$\widehat{houseprice} = \boxed{98.24833} + 0.10977(squarefeet)$$

- b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
- Because a house cannot have a square footage of 0, b_0 has no practical application



Simple Linear Regression Example: Interpreting b_1

$$\widehat{houseprice} = 98.24833 + \boxed{0.10977}(squarefeet)$$

- b_1 estimates the change in the average value of Y as a result of a one-unit increase in X
  - Here, $\boxed{b_1 = 0.10977}$ tells us that the mean value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Simple Linear Regression Example: Making Predictions

Predict the price for a house with 2000 square feet:

$$\widehat{\text{houseprice}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

$$= 98.25 + 0.1098(2000)$$

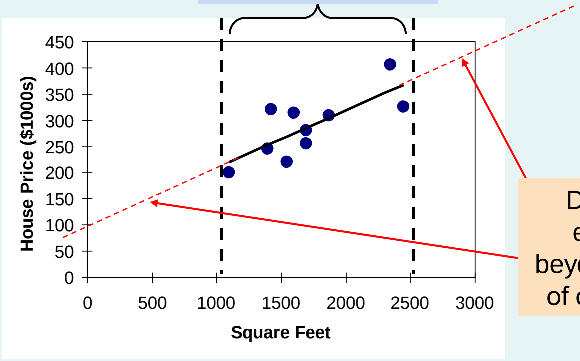
$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

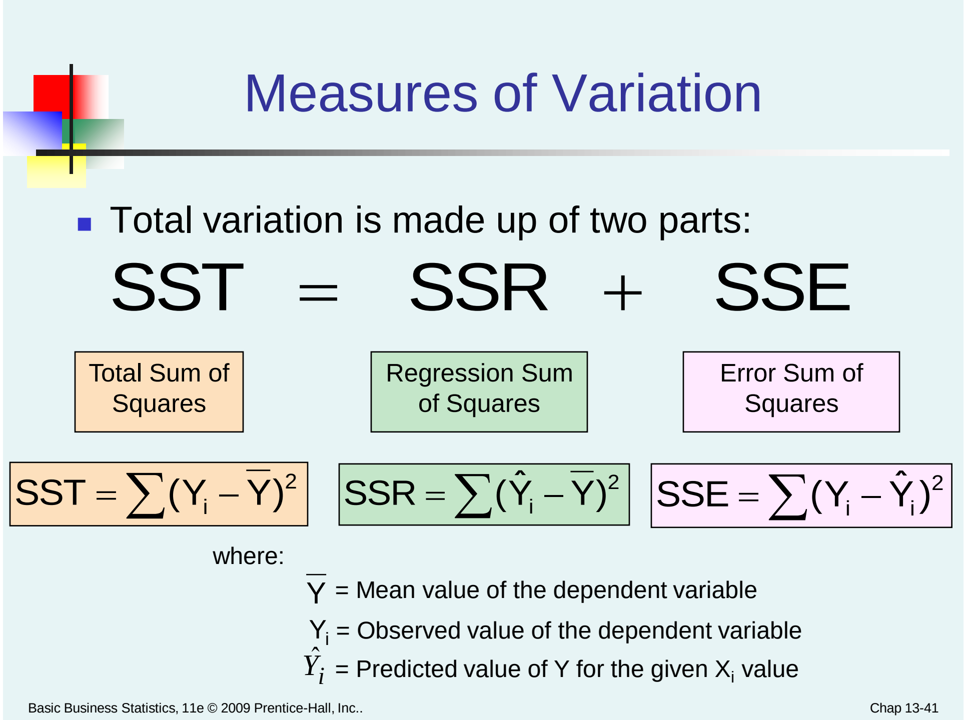
Simple Linear Regression Example: Making Predictions

- When using a regression model for prediction, only predict within the relevant range of data

Relevant range for interpolation



Do not try to extrapolate beyond the range of observed X's



Measures of Variation

- Total variation is made up of two parts:

$$SST \quad = \quad SSR \quad + \quad SSE$$

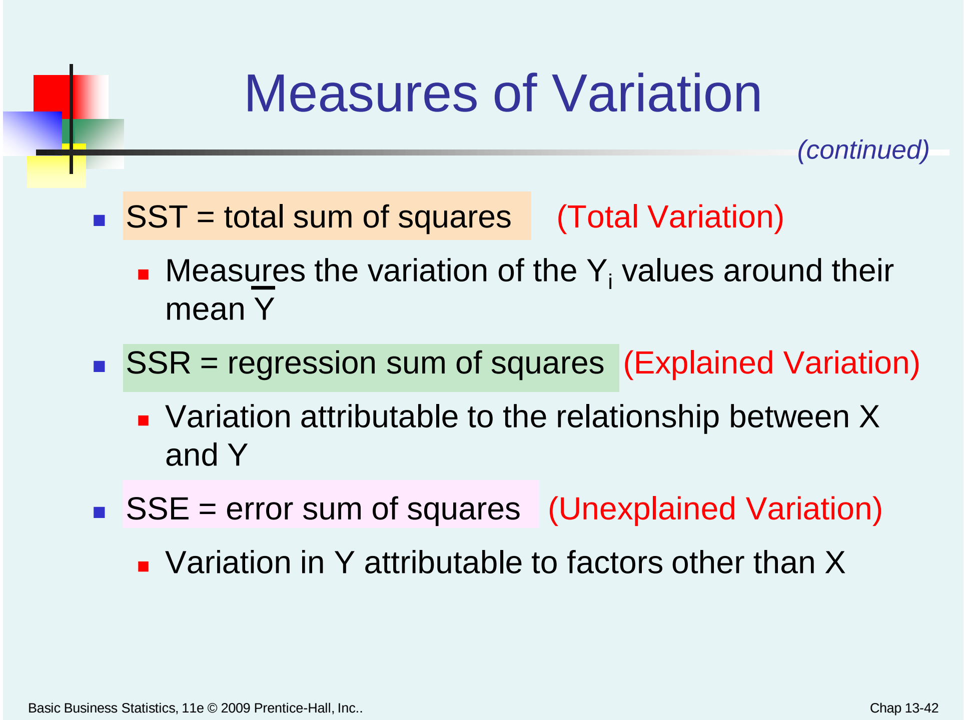
Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (Y_i - \bar{Y})^2$	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$

where:

$\bar{Y}$ = Mean value of the dependent variable

Y_i = Observed value of the dependent variable

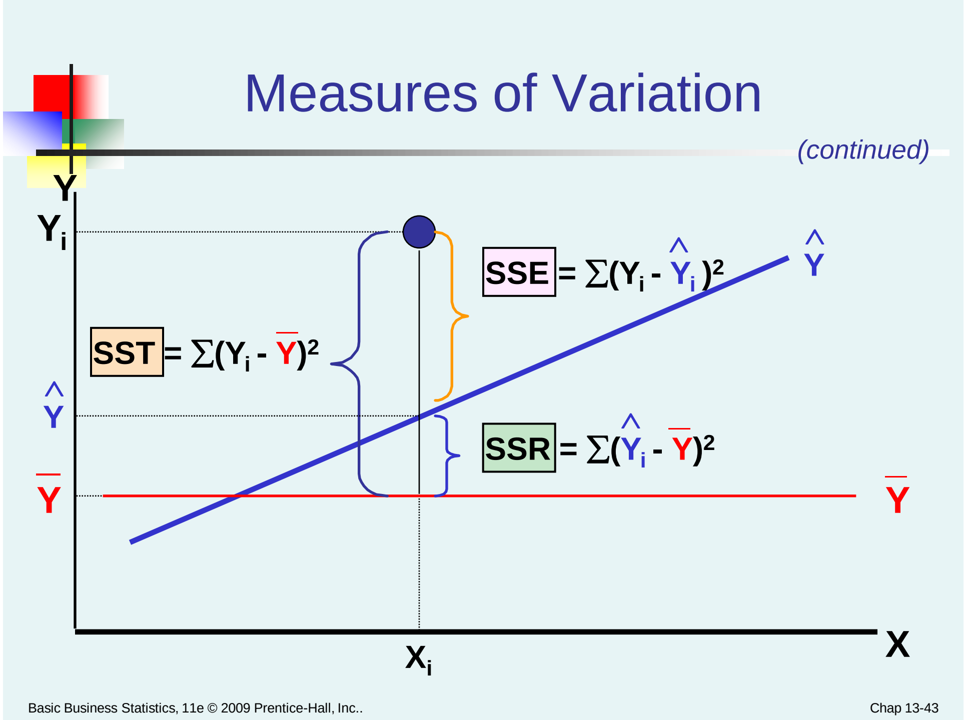
$\hat{Y}_i$ = Predicted value of Y for the given X_i value



Measures of Variation

(continued)

- SST = total sum of squares (Total Variation)
  - Measures the variation of the Y_i values around their mean $\bar{Y}$
- SSR = regression sum of squares (Explained Variation)
  - Variation attributable to the relationship between X and Y
- SSE = error sum of squares (Unexplained Variation)
  - Variation in Y attributable to factors other than X



Measures of Variation

(continued)

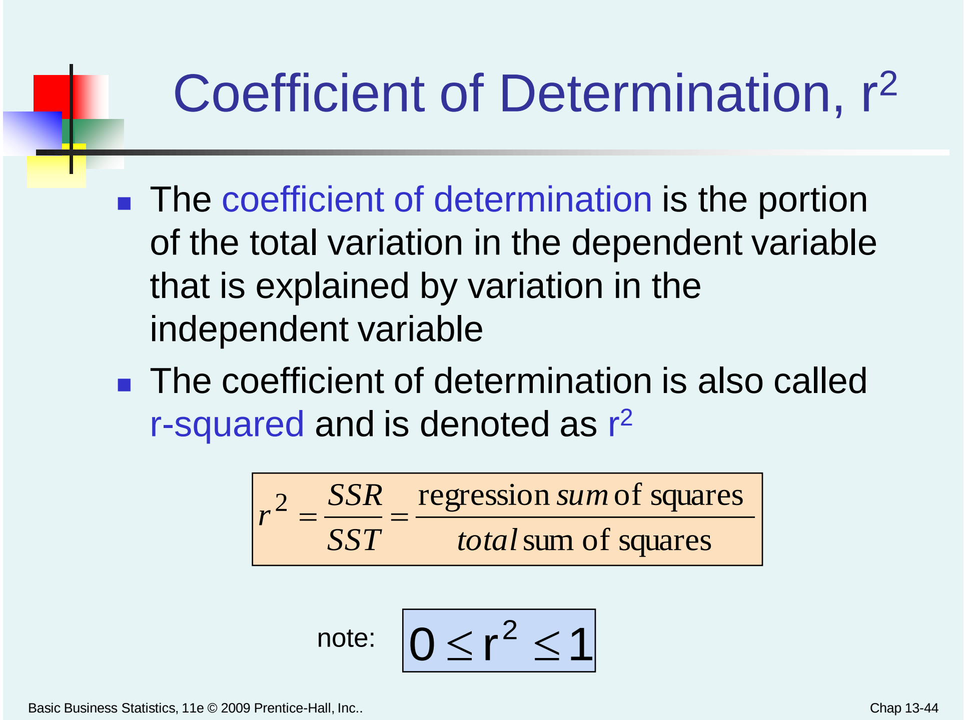
SST $= \sum(Y_i - \bar{Y})^2$

SSE $= \sum(Y_i - \hat{Y}_i)^2$

SSR $= \sum(\hat{Y}_i - \bar{Y})^2$

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc..

Chap 13-43



Coefficient of Determination, r^2

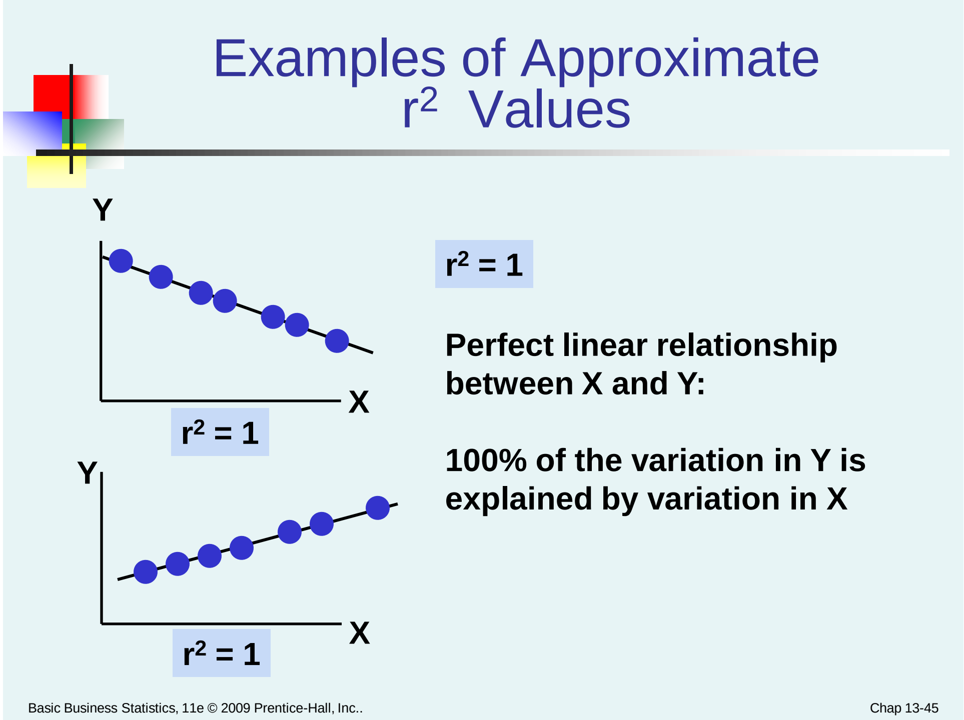
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r-squared and is denoted as r^2

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression } sum \text{ of squares}}{total \text{ sum of squares}}$$

note: $0 \le r^2 \le 1$

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc..

Chap 13-44



Examples of Approximate r^2 Values

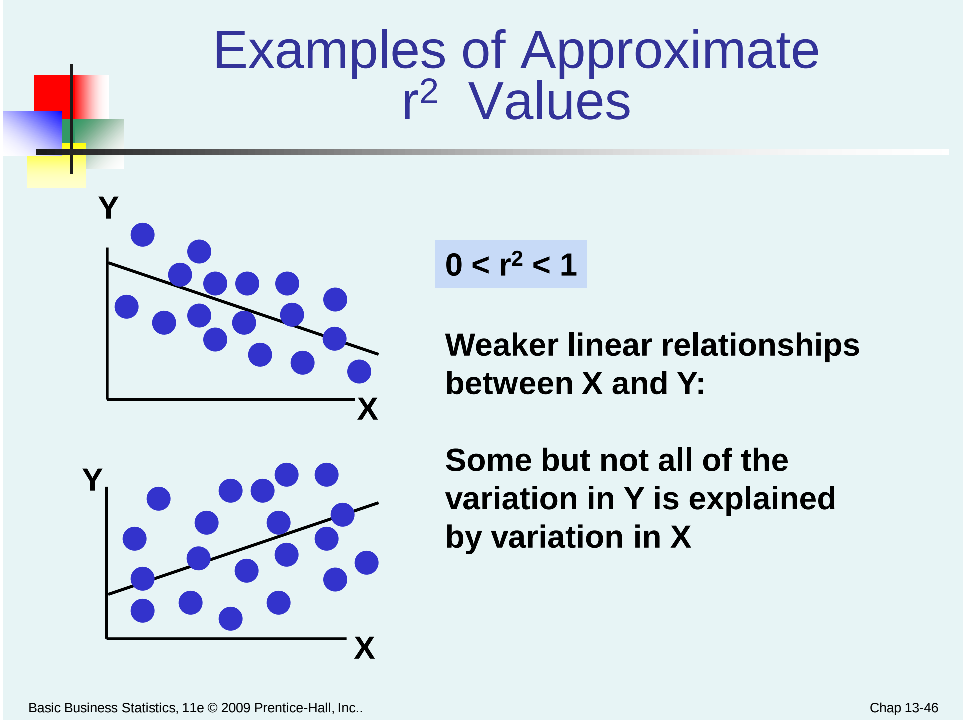
$r^2 = 1$

$r^2 = 1$

$r^2 = 1$

Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X

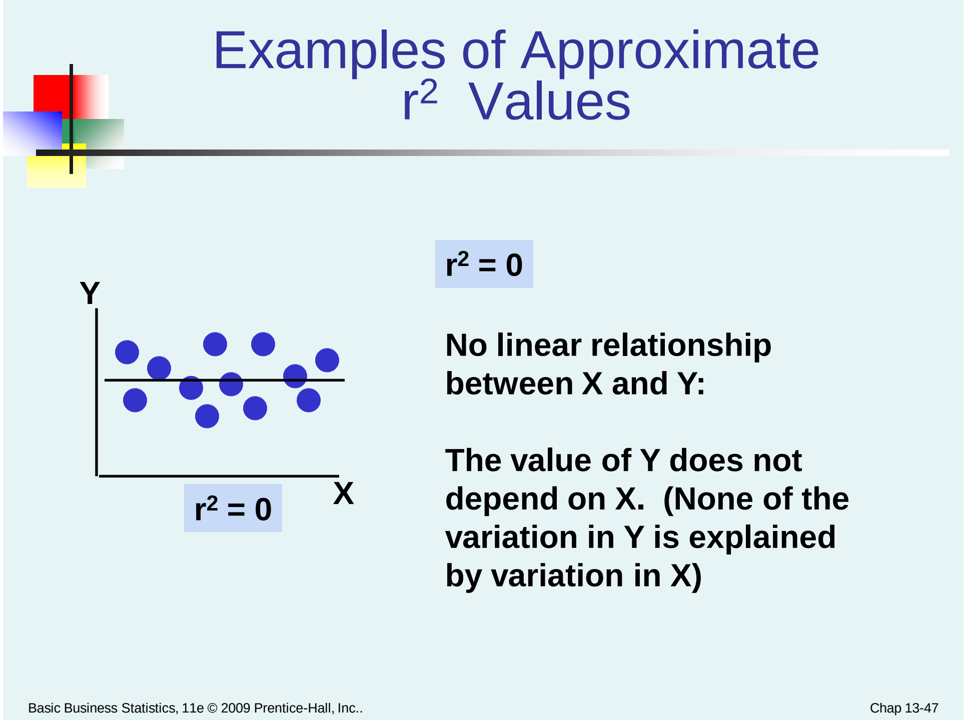


Examples of Approximate r^2 Values

$0 < r^2 < 1$

Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X

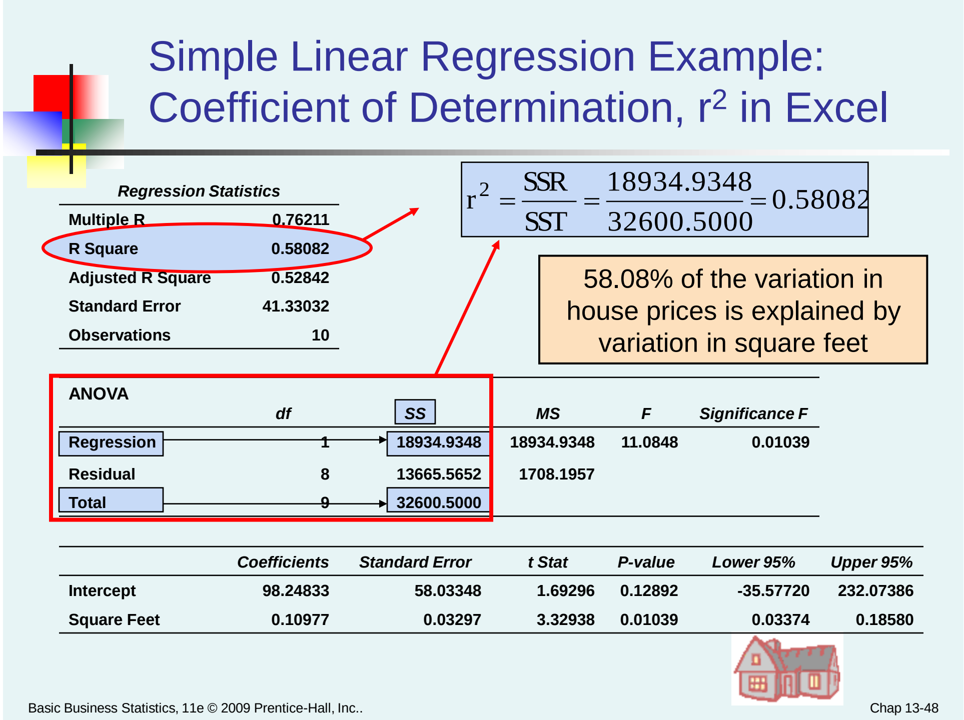


Examples of Approximate r^2 Values

$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)



Simple Linear Regression Example: Coefficient of Determination, r^2 in Excel

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Simple Linear Regression Example: Coefficient of Determination, r^2 in Minitab

```
The regression equation is

Price = 98.2 + 0.110 Square Feet

Predictor      Coef      SE Coef    T     P
Constant       98.25     58.03      1.69  0.129
Square Feet  0.10977   0.03297    3.33  0.010

S = 41.3303   R-Sq = 58.1%   R-Sq(adj) = 52.8%

Analysis of Variance

Source          DF    SS      MS      F       P
Regression      1     18935  18935  11.08  0.010
Residual Error  8     13666   1708
Total           9     32600
```

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

Standard Error of Estimate

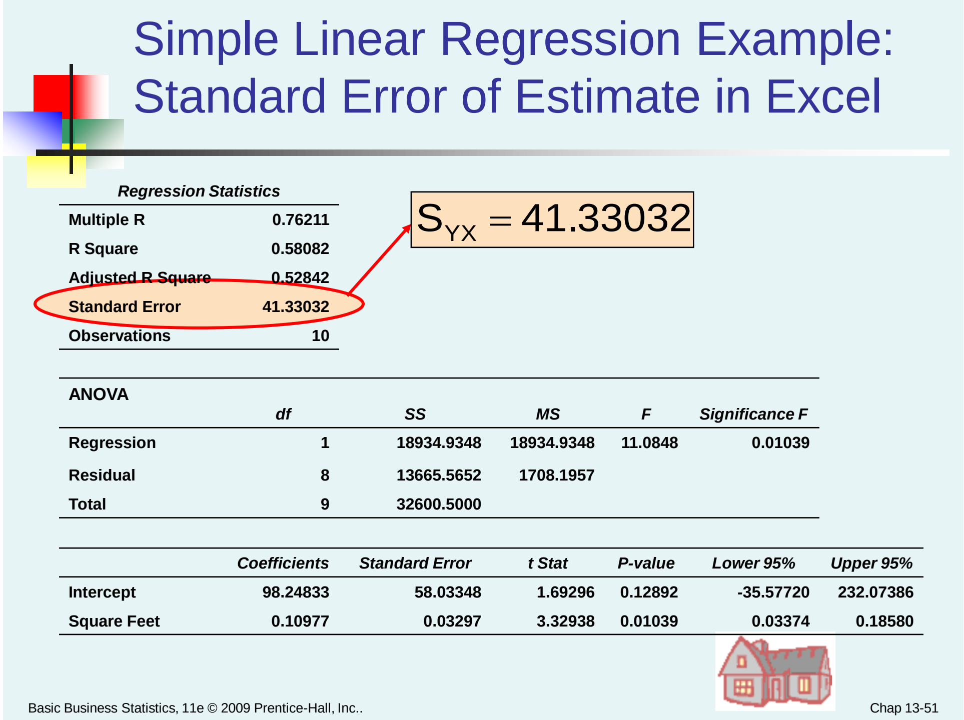
- The standard deviation of the variation of observations around the regression line is estimated by

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

Where

SSE = error sum of squares

n = sample size



Simple Linear Regression Example: Standard Error of Estimate in Excel

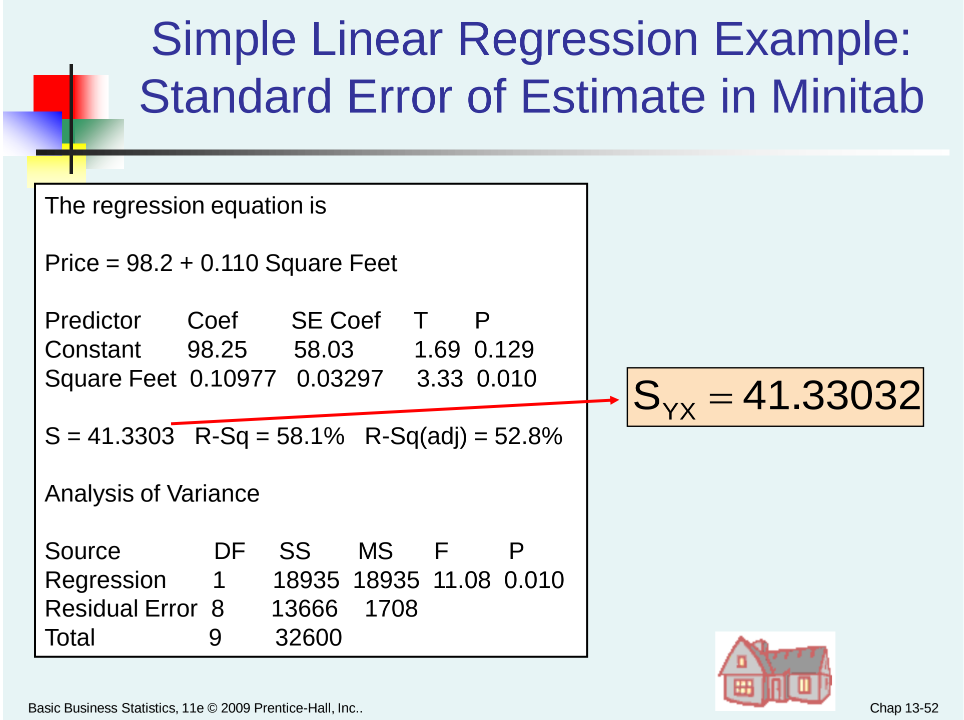
Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$S_{YX} = 41.33032$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580



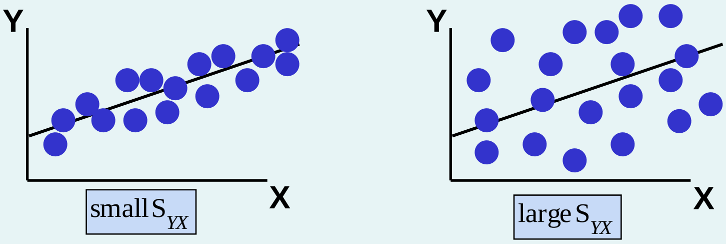
Simple Linear Regression Example: Standard Error of Estimate in Minitab

```
The regression equation is

Price = 98.2 + 0.110 Square Feet

Predictor      Coef      SE Coef     T      P
Constant       98.25     58.03       1.69  0.129
Square Feet  0.10977   0.03297     3.33  0.010

S = 41.3303   R-Sq = 58.1%   R-Sq(adj) = 52.8%

Analysis of Variance

Source          DF    SS      MS     F       P
Regression      1     18935  18935  11.08  0.010
Residual Error  8     13666   1708
Total           9     32600
```

$$S_{YX} = 41.33032$$

Comparing Standard Errors

S_{YX} is a measure of the variation of observed Y values from the regression line



The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data

i.e., S_{YX} = \$41.33K is moderately small relative to house prices in the \$200K - \$400K range

Assumptions of Regression L.I.N.E

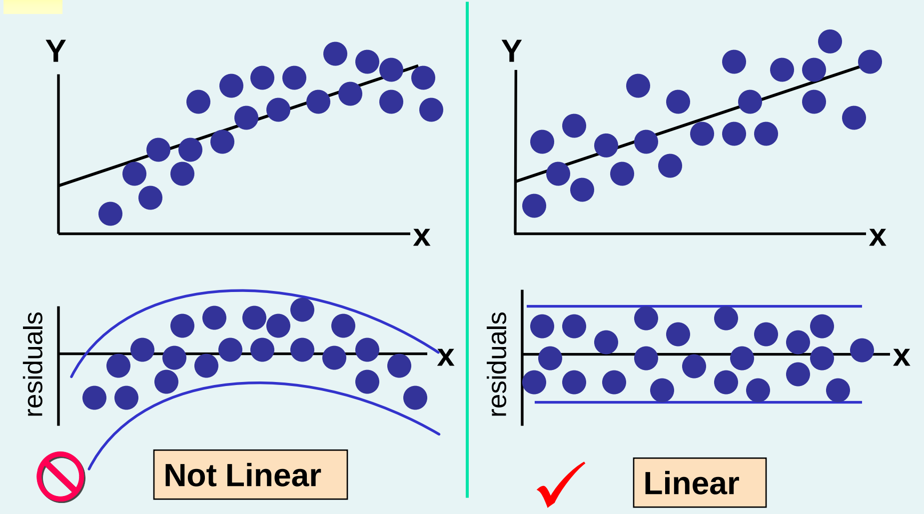
- Linearity
  - The relationship between X and Y is linear
- Independence of Errors
  - Error values are statistically independent
- Normality of Error
  - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)
  - The probability distribution of the errors has constant variance

Residual Analysis

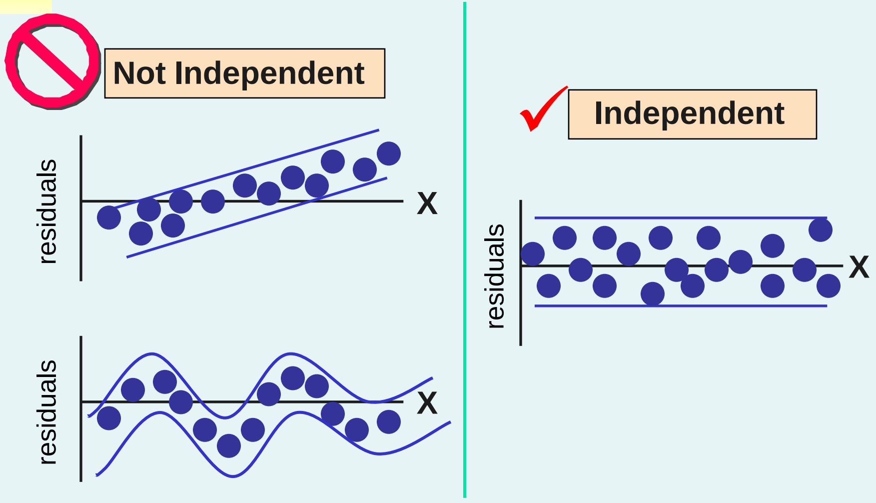
$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

Residual Analysis for Linearity

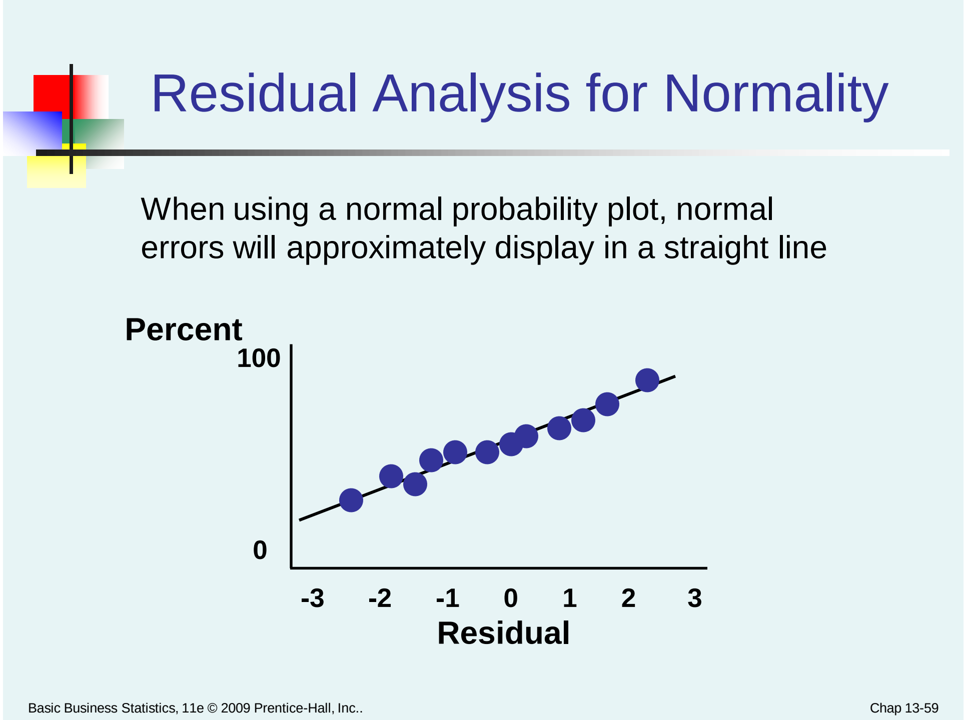


Residual Analysis for Independence



Checking for Normality

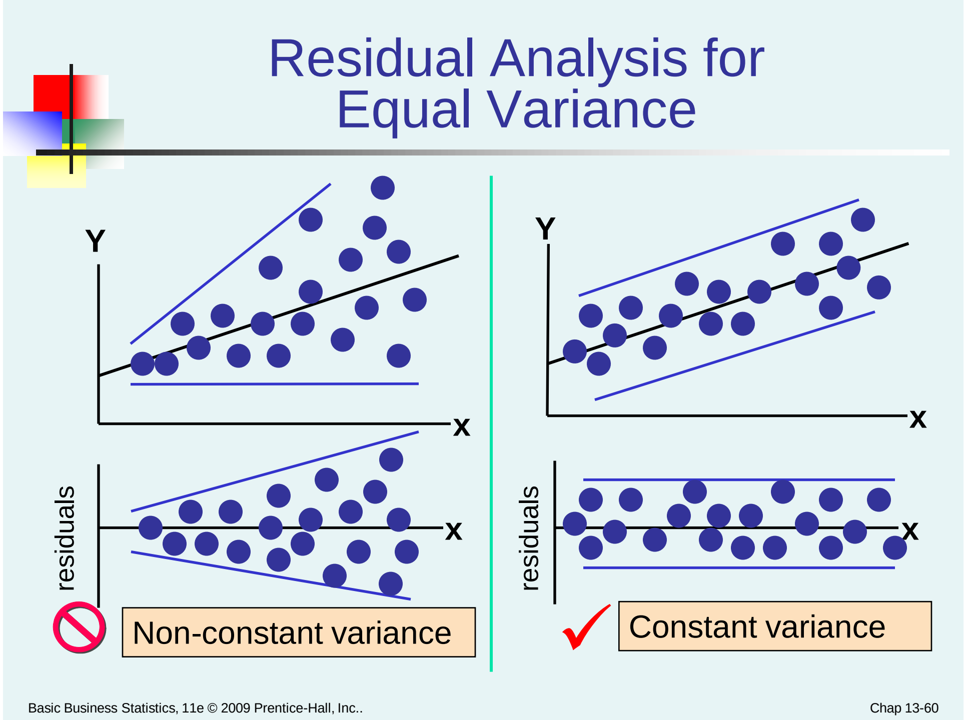
- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals



Residual Analysis for Normality

When using a normal probability plot, normal errors will approximately display in a straight line

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 13-59



Residual Analysis for Equal Variance

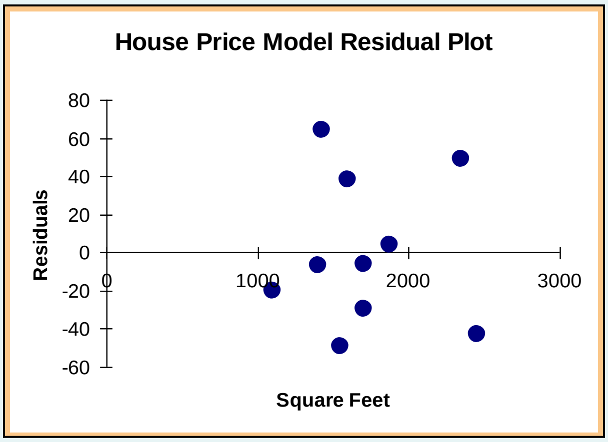
Non-constant variance

✓ Constant variance

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 13-60

Simple Linear Regression Example: Excel Residual Output

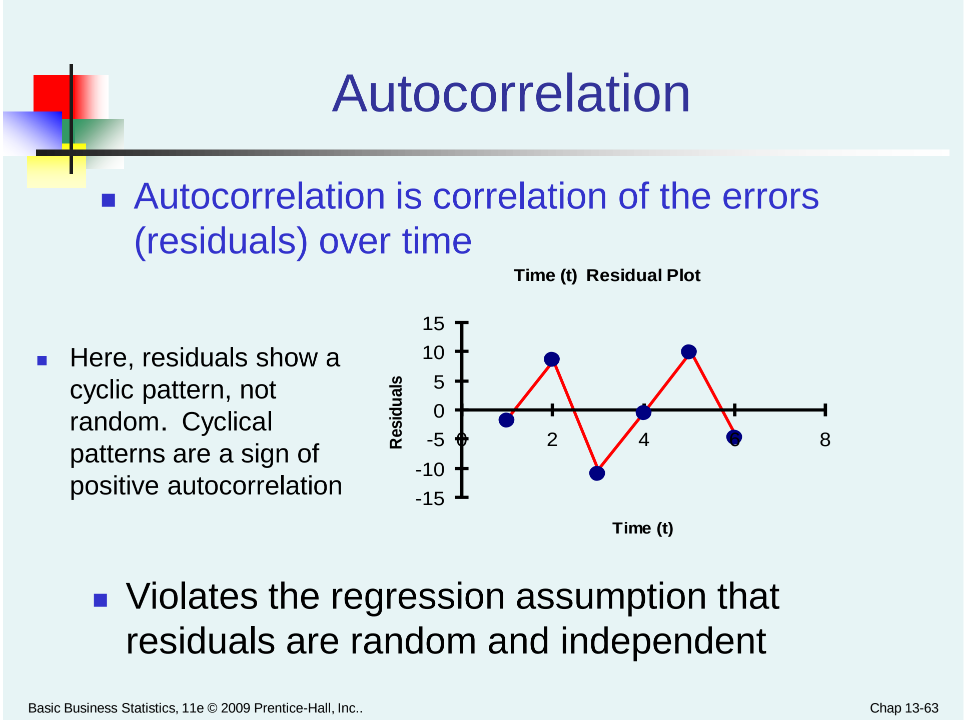
RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Does not appear to violate any regression assumptions

Measuring Autocorrelation: The Durbin-Watson Statistic

- Used when data are collected over time to detect if autocorrelation is present
- Autocorrelation exists if residuals in one time period are related to residuals in another period

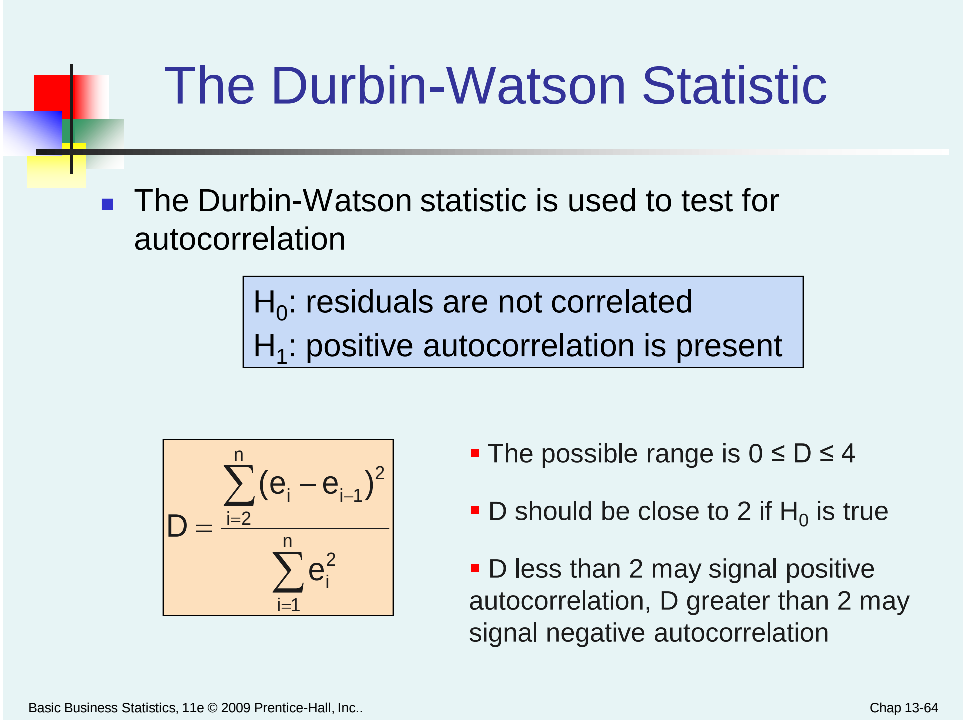


Autocorrelation

- Autocorrelation is correlation of the errors (residuals) over time

- Here, residuals show a cyclic pattern, not random. Cyclical patterns are a sign of positive autocorrelation

- Violates the regression assumption that residuals are random and independent



The Durbin-Watson Statistic

- The Durbin-Watson statistic is used to test for autocorrelation

H_0: residuals are not correlated
H_1: positive autocorrelation is present

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

- The possible range is $0 \le D \le 4$
- D should be close to 2 if H_0 is true
- D less than 2 may signal positive autocorrelation, D greater than 2 may signal negative autocorrelation

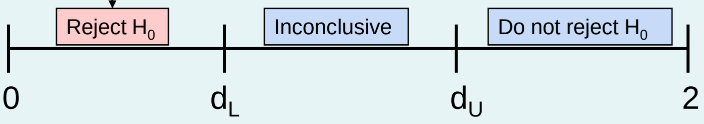
Testing for Positive Autocorrelation

H_0: positive autocorrelation does not exist

H_1: positive autocorrelation is present

- Calculate the Durbin-Watson test statistic = D
  (The Durbin-Watson Statistic can be found using Excel or Minitab)
- Find the values d_L and d_U from the Durbin-Watson table
  (for sample size **n** and number of independent variables **k**)

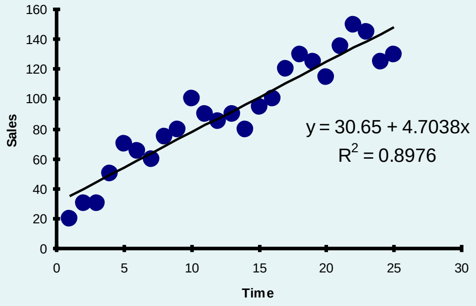
Decision rule: reject H_0 if $D < d_L$



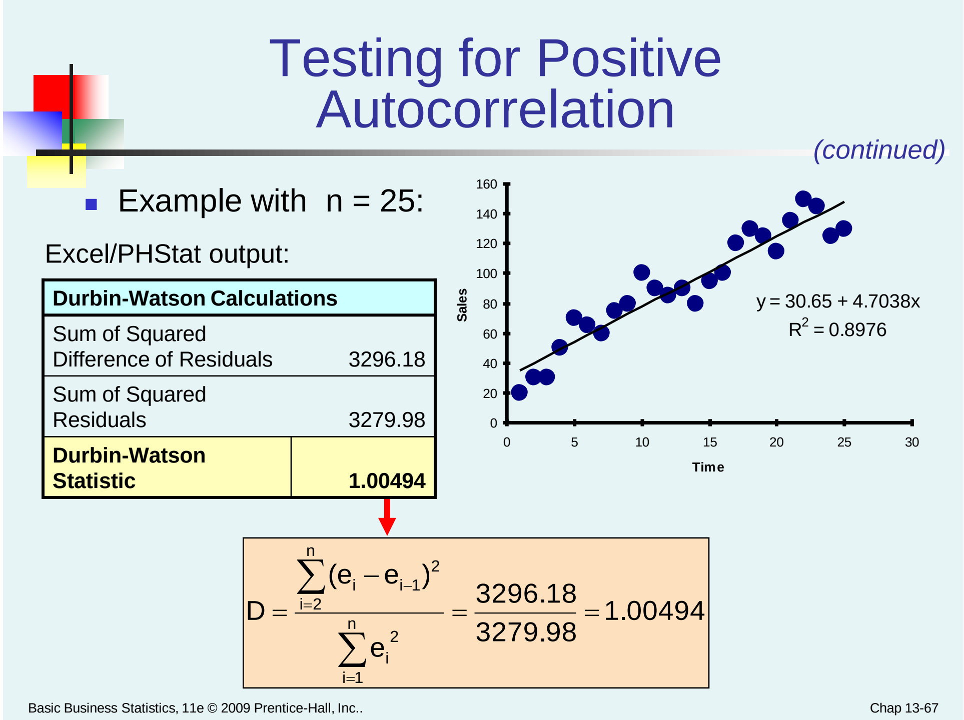
Testing for Positive Autocorrelation

(continued)

- Suppose we have the following time series data:



- Is there autocorrelation?



Testing for Positive Autocorrelation

(continued)

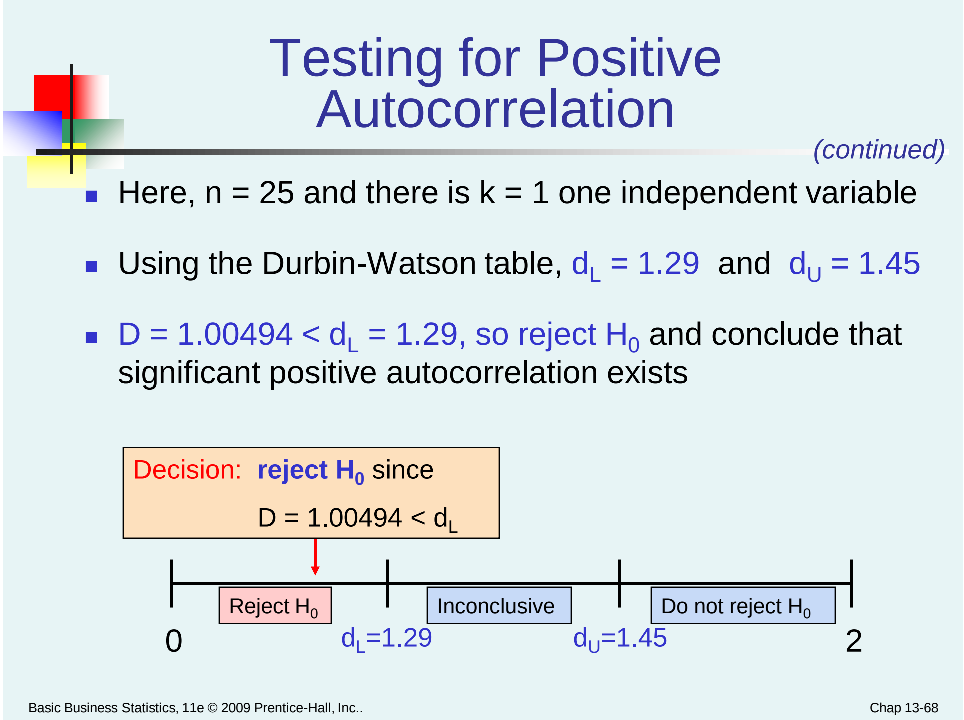
- Example with n = 25:

Excel/PHStat output:

Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	3296.18
Sum of Squared Residuals	3279.98
Durbin-Watson Statistic	**1.00494**

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{3296.18}{3279.98} = 1.00494$$

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 13-67



Testing for Positive Autocorrelation

(continued)

- Here, n = 25 and there is k = 1 one independent variable
- Using the Durbin-Watson table, $d_L = 1.29$ and $d_U = 1.45$
- $D = 1.00494 < d_L = 1.29$, so reject H_0 and conclude that significant positive autocorrelation exists

Decision: **reject H_0** since

$D = 1.00494 < d_L$

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 13-68

Inferences About the Slope

- The standard error of the regression slope coefficient (b_1) is estimated by

$$\boxed{S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \bar{X})^2}}}$$

where:

S_{b_1} = Estimate of the standard error of the slope

$S_{YX} = \sqrt{\frac{SSE}{n-2}}$ = Standard error of the estimate

Inferences About the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses
  - H_0: $\beta_1 = 0$ (no linear relationship)
  - H_1: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b1} = standard error of the slope

Inferences About the Slope: t Test Example

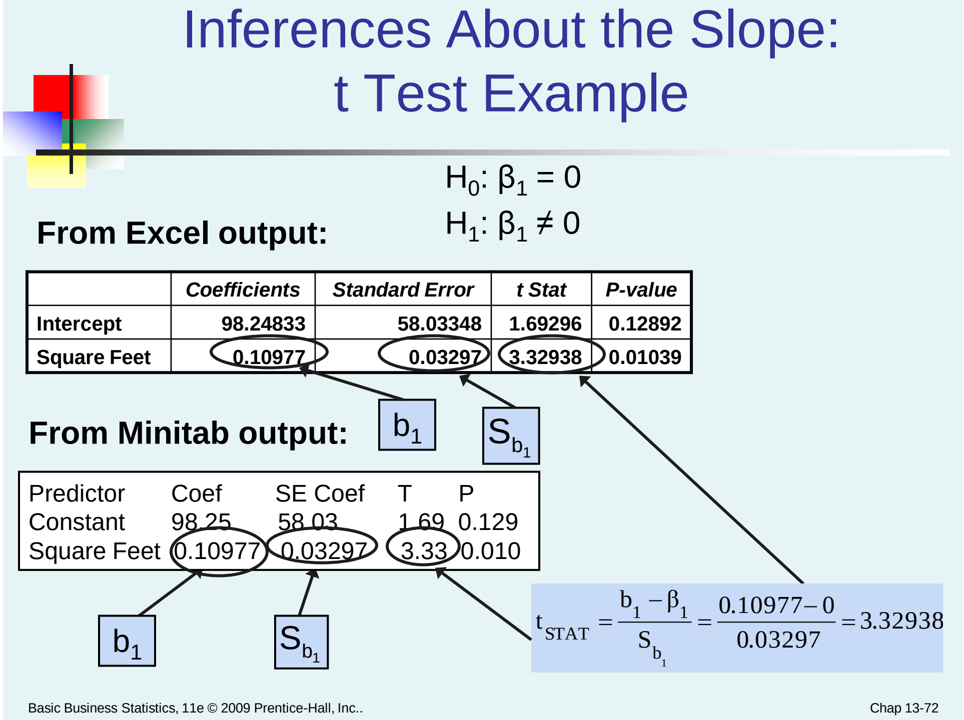
House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\text{house price} = 98.25 + 0.1098\,(\text{sq.ft.})$$

The slope of this model is 0.1098

Is there a relationship between the square footage of the house and its sales price?



Inferences About the Slope: t Test Example

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

From Excel output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

From Minitab output:

```
Predictor    Coef     SE Coef    T     P
Constant     98.25    58.03      1.69  0.129
Square Feet  0.10977  0.03297    3.33  0.010
```

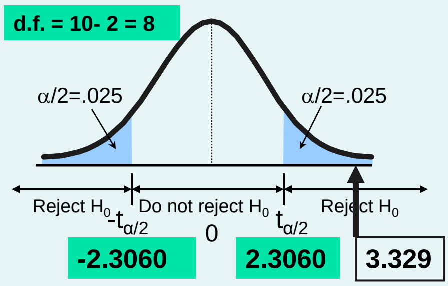
$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences About the Slope: t Test Example

Test Statistic: $t_{STAT} = 3.329$

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$



Decision: Reject H_0

There is sufficient evidence that square footage affects house price

Inferences About the Slope: t Test Example

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

From Excel output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

From Minitab output:

```
Predictor     Coef      SE Coef    T      P
Constant      98.25      58.03     1.69  0.129
Square Feet  0.10977    0.03297    3.33  0.010
```

p-value

Decision: Reject H_0, since p-value < α

There is sufficient evidence that square footage affects house price.

F Test for Significance

- F Test statistic: $$F_{STAT} = \frac{MSR}{MSE}$$

where

$$\text{MSR} = \frac{\text{SSR}}{\text{k}}$$

$$\text{MSE} = \frac{\text{SSE}}{\text{n} - \text{k} - 1}$$

where F_{STAT} follows an F distribution with k numerator and (n – k - 1) denominator degrees of freedom

(k = the number of independent variables in the regression model)

F-Test for Significance Excel Output

Regression Statistics	
Multiple R	**0.76211**
R Square	**0.58082**
Adjusted R Square	**0.52842**
Standard Error	**41.33032**
Observations	**10**

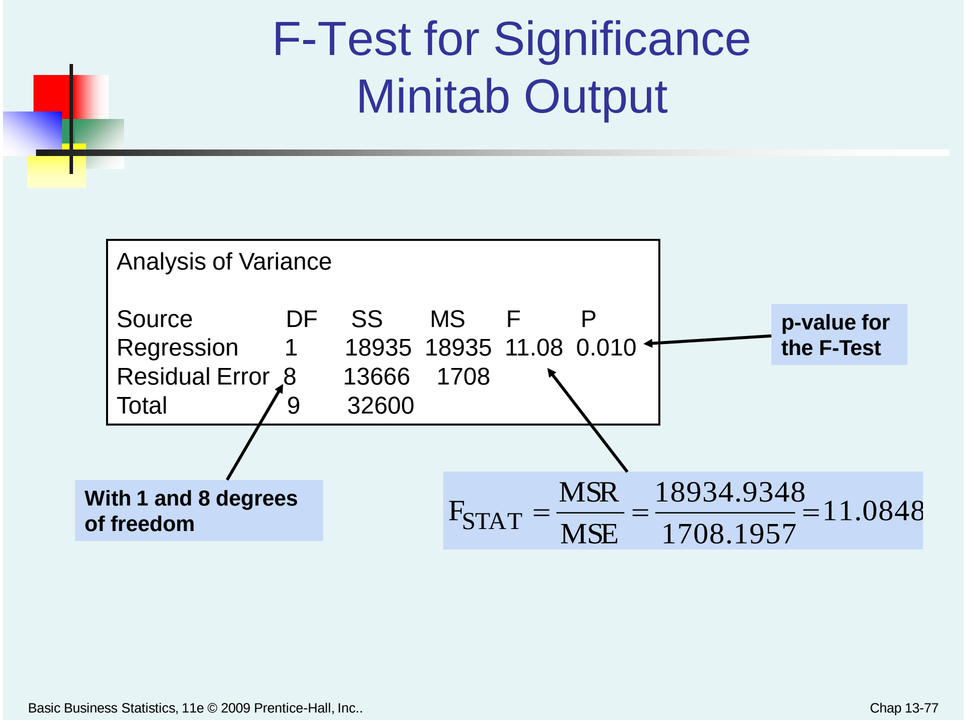
$$\text{F}_{\text{STAT}} = \frac{\text{MSR}}{\text{MSE}} = \frac{18934.9348}{1708.1957} = 11.0848$$

With 1 and 8 degrees of freedom

p-value for the F-Test

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**1**	**18934.9348**	**18934.9348**	**11.0848**	**0.01039**
Residual	**8**	**13665.5652**	**1708.1957**		
Total	**9**	**32600.5000**			



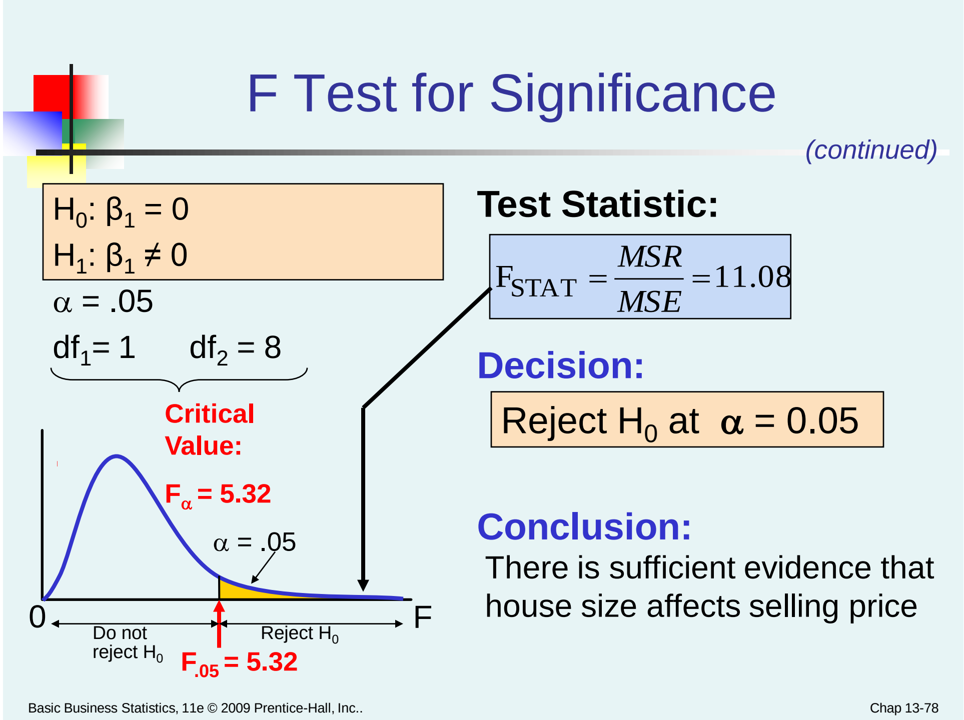
F-Test for Significance Minitab Output

```
Analysis of Variance

Source          DF    SS     MS      F      P
Regression      1    18935  18935  11.08  0.010
Residual Error  8    13666   1708
Total           9    32600
```

p-value for the F-Test

With 1 and 8 degrees of freedom

$$F_{STAT} = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$



F Test for Significance

(continued)

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

$\alpha = .05$

$\underbrace{df_1 = 1 \qquad df_2 = 8}$

Critical Value:

$F_\alpha = 5.32$

Test Statistic:

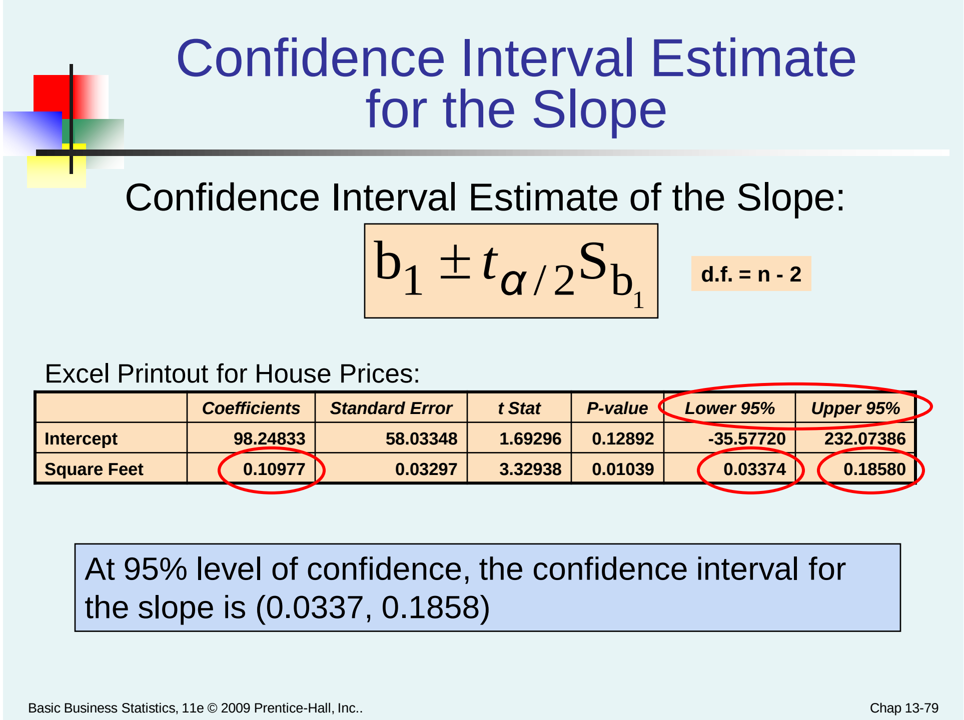
$$F_{STAT} = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price



Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

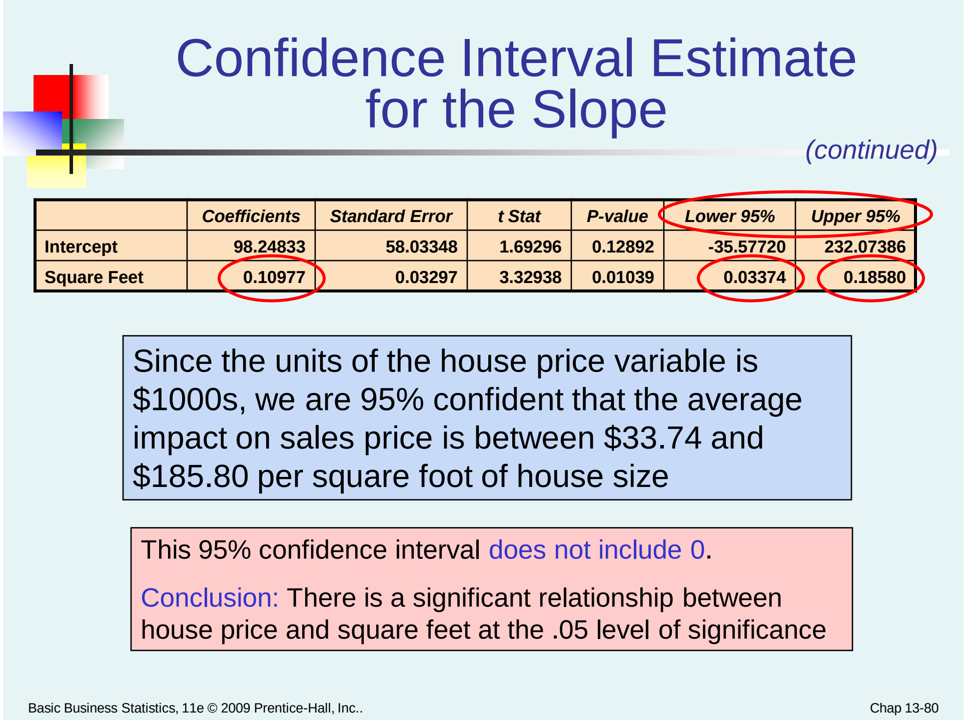
$$b_1 \pm t_{\alpha/2} S_{b_1}$$

d.f. = n - 2

Excel Printout for House Prices:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)



Confidence Interval Estimate for the Slope

(continued)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.74 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

t Test for a Correlation Coefficient

- Hypotheses

H_0: ρ = 0 (no correlation between X and Y)
H_1: ρ ≠ 0 (correlation exists)

- Test statistic

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$

(with n – 2 degrees of freedom)

where

$$r = +\sqrt{r^2} \text{ if } b_1 > 0$$

$$r = -\sqrt{r^2} \text{ if } b_1 < 0$$

t-test For A Correlation Coefficient

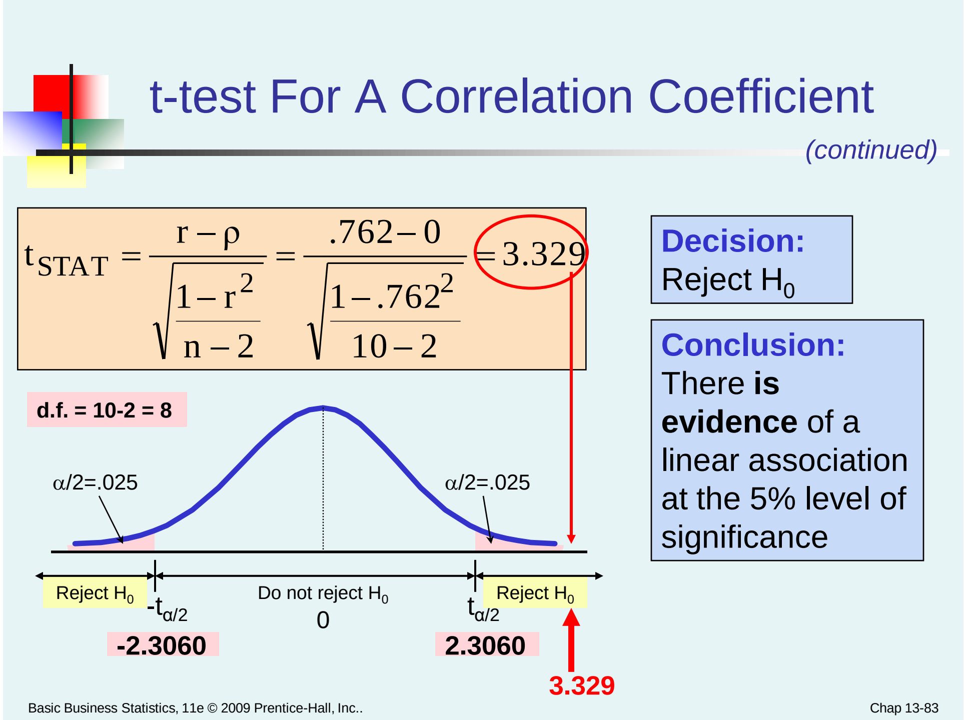
(continued)

Is there evidence of a linear relationship between square feet and house price at the .05 level of significance?

H_0: ρ = 0 (No correlation)
H_1: ρ ≠ 0 (correlation exists)

α =.05 , df = 10 - 2 = 8

$$t_{STAT} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.762-0}{\sqrt{\frac{1-.762^2}{10-2}}} = 3.329$$



t-test For A Correlation Coefficient

(continued)

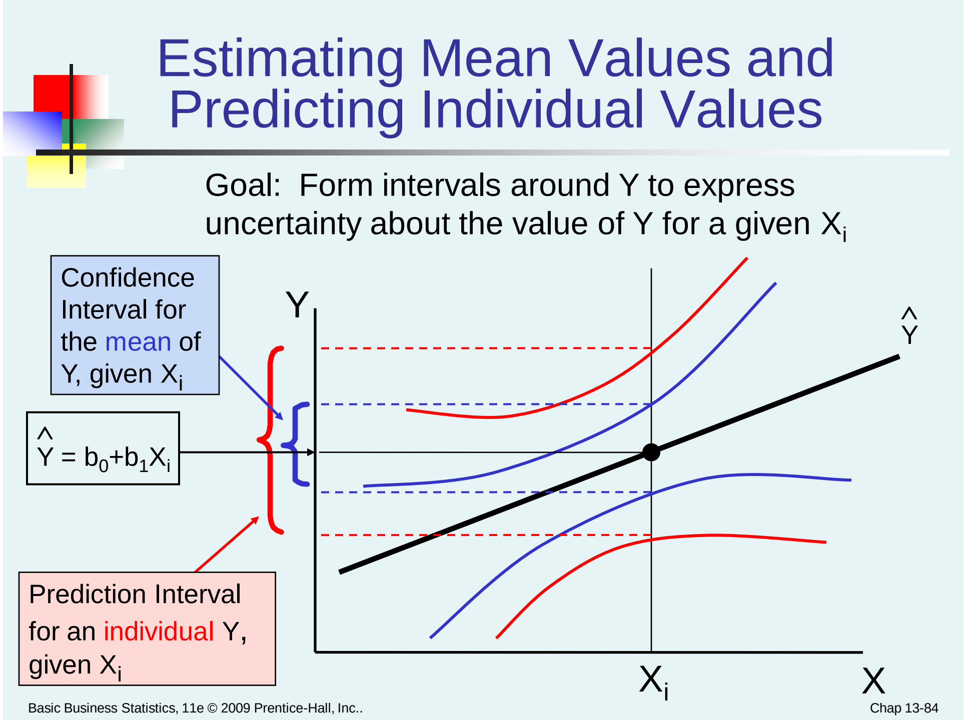
$$t_{STAT} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.762-0}{\sqrt{\frac{1-.762^2}{10-2}}} = 3.329$$

Decision:
Reject H_0

Conclusion:
There **is evidence** of a linear association at the 5% level of significance

d.f. = 10-2 = 8

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 13-83



Estimating Mean Values and Predicting Individual Values

Goal: Form intervals around Y to express uncertainty about the value of Y for a given X_i

Confidence Interval for the mean of Y, given X_i

$\hat{Y} = b_0+b_1X_i$

Prediction Interval for an individual Y, given X_i

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 13-84

Confidence Interval for the Average Y, Given X

Confidence interval estimate for the **mean value of Y** given a particular X_i

$$\text{Confidence interval for } \mu_{Y|X=X_i}:$$

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

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

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} = \frac{1}{n} + \frac{\overbrace{(X_i - \bar{X})^2}}{\sum (X_i - \bar{X})^2}$$

Prediction Interval for an Individual Y, Given X

Confidence interval estimate for an **Individual value of Y** given a particular X_i

$$\text{Confidence interval for } Y_{X=X_i}:$$

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case

Estimation of Mean Values: Example

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

Find the 95% confidence interval for the mean price of 2,000 square-foot houses

Predicted Price $\hat{Y}_i$ = 317.85 ($1,000s)

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.66 and 354.90, or from $280,660 to $354,900

Estimation of Individual Values: Example

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

Find the 95% prediction interval for an individual house with 2,000 square feet

Predicted Price $\hat{Y}_i$ = 317.85 ($1,000s)

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}} = 317.85 \pm 102.28$$

The prediction interval endpoints are 215.50 and 420.07, or from $215,500 to $420,070

Finding Confidence and Prediction Intervals in Excel

- From Excel, use

  PHStat | regression | simple linear regression …

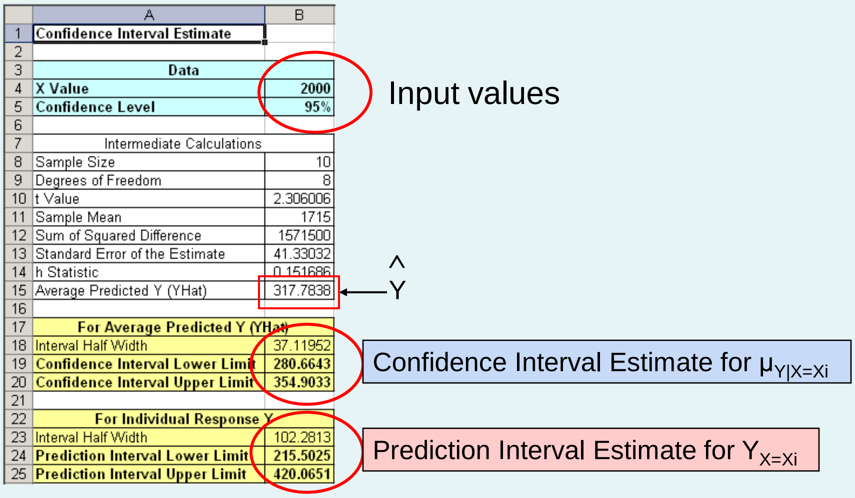
  - Check the

    “confidence and prediction interval for X=”

    box and enter the X-value and confidence level desired

Finding Confidence and Prediction Intervals in Excel

(continued)



	A	B
1	**Confidence Interval Estimate**	
2		
3	**Data**	
4	**X Value**	**2000**
5	**Confidence Level**	**95%**
6		
7	Intermediate Calculations	
8	Sample Size	10
9	Degrees of Freedom	8
10	t Value	2.306006
11	Sample Mean	1715
12	Sum of Squared Difference	1571500
13	Standard Error of the Estimate	41.33032
14	h Statistic	0.151686
15	Average Predicted Y (YHat)	317.7838
16		
17	**For Average Predicted Y (YHat)**	
18	Interval Half Width	37.11952
19	**Confidence Interval Lower Limit**	**280.6643**
20	**Confidence Interval Upper Limit**	**354.9033**
21		
22	**For Individual Response Y**	
23	Interval Half Width	102.2813
24	**Prediction Interval Lower Limit**	**215.5025**
25	**Prediction Interval Upper Limit**	**420.0651**

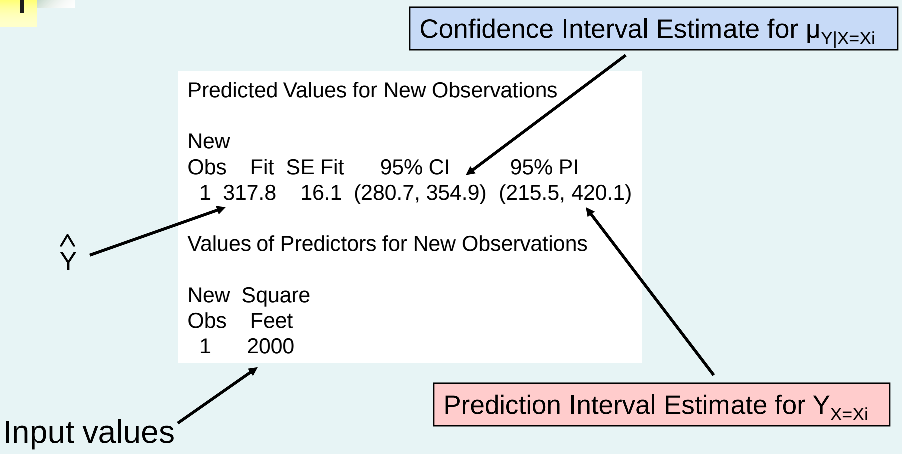
Input values

$\hat{Y}$

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

Prediction Interval Estimate for $Y_{X=Xi}$

Finding Confidence and Prediction Intervals in Minitab



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

```
Predicted Values for New Observations

New
Obs   Fit  SE Fit      95% CI          95% PI
  1  317.8    16.1  (280.7, 354.9)  (215.5, 420.1)

Values of Predictors for New Observations

New  Square
Obs    Feet
  1      2000
```

$\hat{Y}$

Input values

Prediction Interval Estimate for $Y_{X=Xi}$

Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression
- Not knowing how to evaluate the assumptions
- Not knowing the alternatives to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range

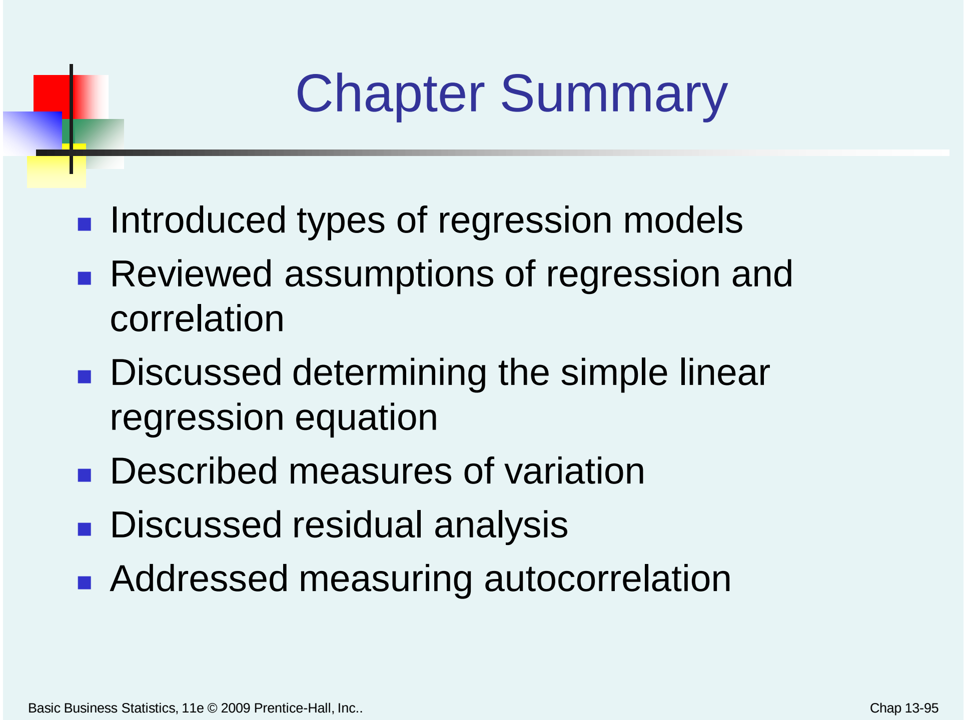
Strategies for Avoiding the Pitfalls of Regression

- Start with a scatter plot of X vs. Y to observe possible relationship
- Perform residual analysis to check the assumptions
    - Plot the residuals vs. X to check for violations of assumptions such as homoscedasticity
    - Use a histogram, stem-and-leaf display, boxplot, or normal probability plot of the residuals to uncover possible non-normality

Strategies for Avoiding the Pitfalls of Regression

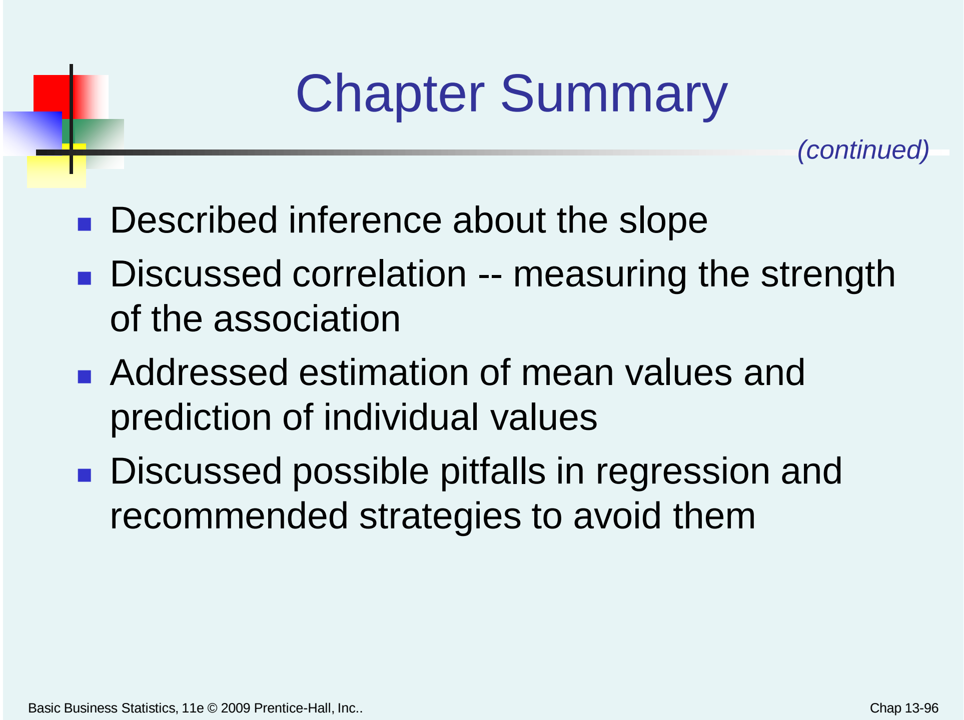
(continued)

- If there is violation of any assumption, use alternative methods or models
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals
- Avoid making predictions or forecasts outside the relevant range



Chapter Summary

- Introduced types of regression models
- Reviewed assumptions of regression and correlation
- Discussed determining the simple linear regression equation
- Described measures of variation
- 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 possible pitfalls in regression and recommended strategies to avoid them