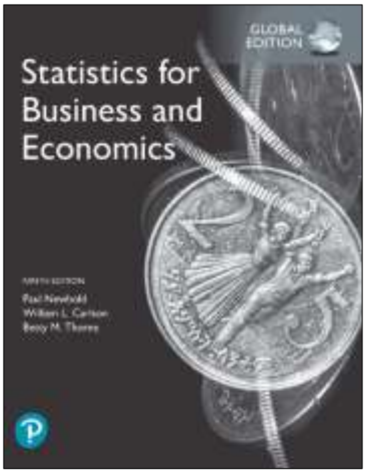
Statistics for Business and Economics

Ninth Edition, Global Edition



Chapter 11

Simple Regression

Chapter Goals (1 of 2)

After completing this chapter, you should be able to:

- Explain the simple linear regression model
- Obtain and interpret the simple linear regression equation for a set of data
- Describe R^2 as a measure of explanatory power of the regression model
- Understand the assumptions behind regression analysis
- Explain measures of variation and determine whether the independent variable is significant

Chapter Goals (2 of 2)

After completing this chapter, you should be able to:

- Calculate and interpret confidence intervals for the regression coefficients
- Use a regression equation for prediction
- Form forecast intervals around an estimated *Y* value for a given *X*
- Use graphical analysis to recognize potential problems in regression analysis
- Explain the correlation coefficient and perform a hypothesis test for zero population correlation

Section 11.1 Overview of Linear Models

- An equation can be fit to show the best linear relationship between two variables:

$$Y = \beta_0 + \beta_1 X$$

Where *Y* is the dependent variable and
X is the independent variable
β_0 is the *Y*-intercept
β_1 is the slope

Least Squares Regression

- Estimates for coefficients β_0 and β_1 are found using a Least Squares Regression technique
- The least-squares regression line, based on sample data, is

$$\hat{y} = b_0 + b_1 x$$

- Where b_1 is the slope of the line and b_0 is the y-intercept:

$$b_1 = \frac{Cov(x, y)}{s_x^2} = r\left(\frac{s_y}{s_x}\right) \qquad b_0 = \bar{y} - b_1\bar{x}$$

 Slide - 5

5

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
(also called the endogenous variable)

Independent variable: the variable used to explain the dependent variable
(also called the exogenous variable)

 Slide - 6

6

Section 11.2 Linear Regression Model

- The relationship between X and Y is described by a linear function
- Changes in Y are assumed to be influenced by changes in X
- Linear regression population equation model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

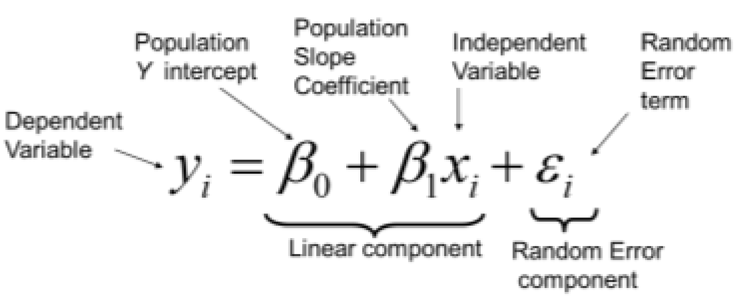
- Where β_0 and β_1 are the population model coefficients and ε is a random error term.

 Slide - 7

7

Simple Linear Regression Model (1 of 2)

The population regression model:



 Slide - 8

8

Linear Regression Assumptions

- The true relationship form is linear (*Y* is a linear function of *X*, plus random error)
- The error terms, ε_i are independent of the *x* values
- The error terms are random variables with mean 0 and constant variance, σ^2
  (the uniform variance property is called homoscedasticity)

$$E[\varepsilon_i] = 0 \text{ and } E[\varepsilon_i^2] = \sigma^2 \text{ for } (i = 1, \ldots, n)$$

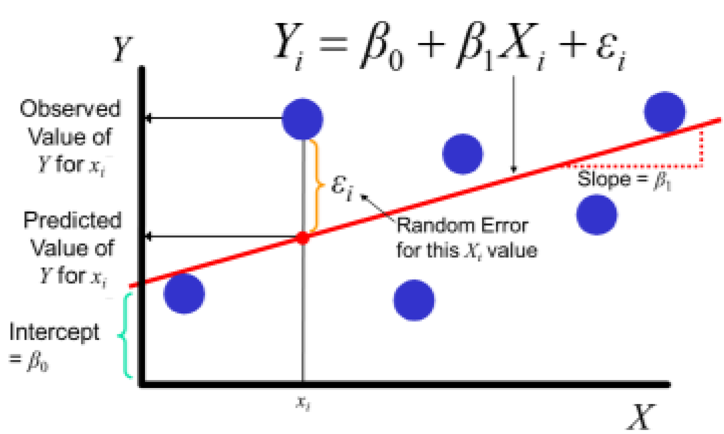
- The random error terms ε_i, are not correlated with one another, so that

$$E[\varepsilon_i \varepsilon_j] = 0 \text{ for all } i \neq j$$

 Slide - 9

9

Simple Linear Regression Model (2 of 2)



 Slide - 10

10

Simple Linear Regression Equation

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*

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

The individual random error terms e_i have a mean of zero

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

11

Section 11.3 Least Squares Coefficient Estimators (1 of 2)

- 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 (errors), SSE:

$$\begin{aligned} \min \text{ SSE } &= \min \sum_{i=1}^{n} e_i^2 \\ &= \min \sum (y_i - \hat{y}_i)^2 \\ &= \min \sum [y_i - (b_0 + b_1 x_i)]^2 \end{aligned}$$

Differential calculus is used to obtain the coefficient estimators b_0 and b_1 that minimize SSE

12

Least Squares Coefficient Estimators (2 of 2)

- The slope coefficient estimator is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{s_x^2} = r\frac{s_y}{s_x}$$

- And the constant or *y*-intercept is

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

- The regression line always goes through the mean $\bar{x}, \bar{y}$

Computer Computation of Regression Coefficients

- The coefficients b_0 and b_1, and other regression results in this chapter, will be found using a computer
  - Hand calculations are tedious
  - Statistical routines are built into Excel
  - Other statistical analysis software can be used

Interpretation of the Slope and the Intercept

- 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)

- b_1 is the estimated change in the average value of *y* as a result of a one-unit change in *x*

15

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

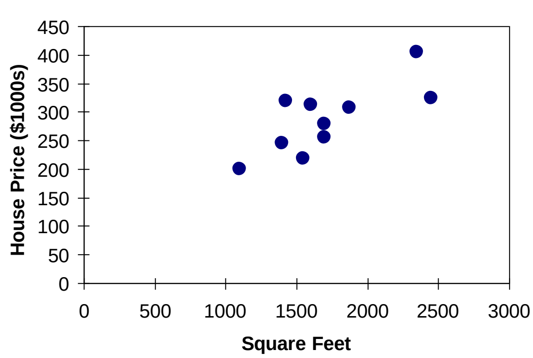
16

Sample Data for House Price Model

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

Graphical Presentation (1 of 2)

- House price model: scatter plot

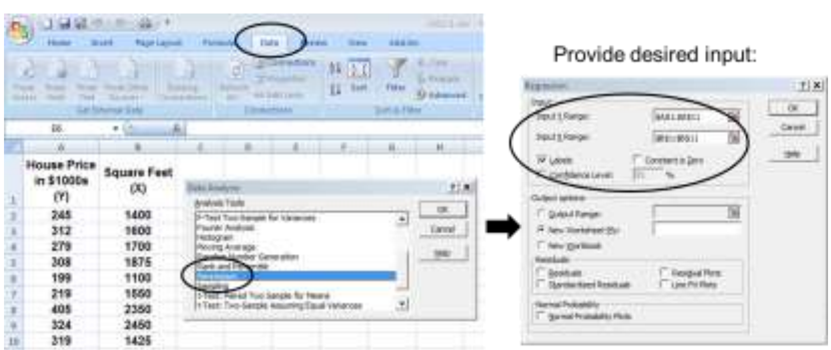


Regression Using Excel (1 of 2)

- Excel will be used to generate the coefficients and measures of goodness of fit for regression
  - Data / Data Analysis / Regression

Regression Using Excel (2 of 2)

- Data / Data Analysis / Regression



Excel Output (1 of 6)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.762113713					
5	R Square	0.580817312					
6	Adjusted R Square	0.528419476					
7	Standard Error	41.33032365					
8	Observations	10					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	18934.9348	18934.9348	11.0848	0.01039	
13	Residual	8	13665.5652	1708.1957			
14	Total	9	32600.5				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	98.24833	58.03348	1.69296	0.12892	-35.57711	232.07377
18	Square Feet (X)	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

 Slide - 21

21

Excel Output (2 of 6)

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:

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

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			

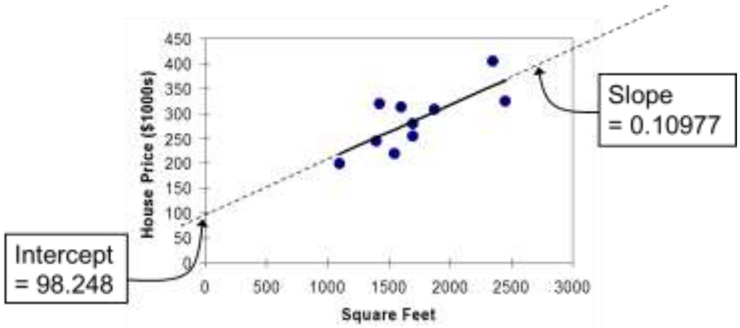
	Coefficients	Standard Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

 Slide - 22

22

Graphical Presentation (2 of 2)

- House price model: scatter plot and regression line



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

 Slide - 23

23

Interpretation of the Intercept, *b* Sub 0

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

- 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)
  - Here, no houses had 0 square feet, so $\boxed{b_0 = 98.24833}$ just indicates that, for houses within the range of sizes observed, \$98,248.33 is the portion of the house price not explained by square feet

 Slide - 24

24

Interpretation of the Slope Coefficient, *b* Sub 1

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

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

25

Section 11.4 Explanatory Power of a Linear Regression Equation

- Total variation is made up of two parts:

$$\text{SST} = \text{SSR} + \text{SSE}$$

Total Sum of Squares | Regression Sum of Squares | Error (residual) Sum of Squares

$$\text{SST} = \sum (y_i - \bar{y})^2 \quad \text{SSR} = \sum (\hat{y}_i - \bar{y})^2 \quad \text{SSE} = \sum (y_i - \hat{y}_i)^2$$

where:

$\bar{y}$ = Average value of the dependent variable
y_i = Observed values of the dependent variable
$\hat{y}_i$ = Predicted value of *y* for the given x_i value

26

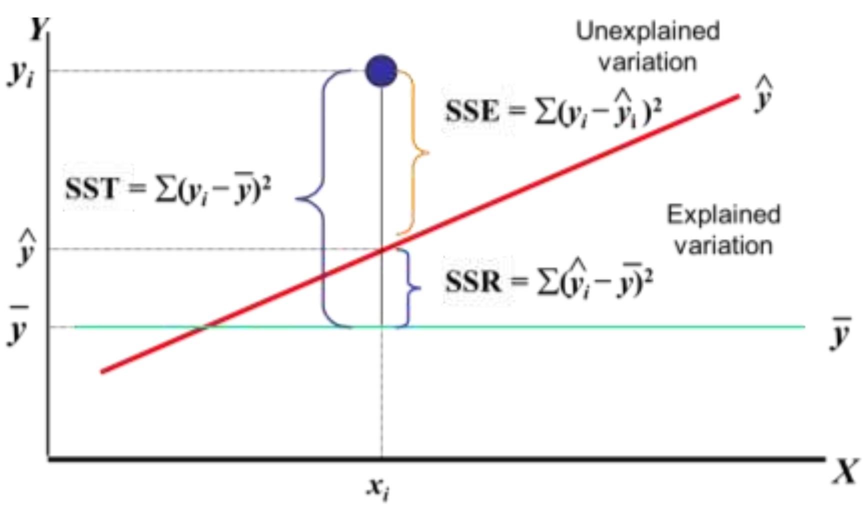
Analysis of Variance (1 of 2)

- 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 linear relationship between *x* and *y*
- SSE = error sum of squares
  - Variation attributable to factors other than the linear relationship between *x* and *y*

 Slide - 27

27

Analysis of Variance (2 of 2)



 Slide - 28

28

Coefficient of Determination, *R* Squared

- 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

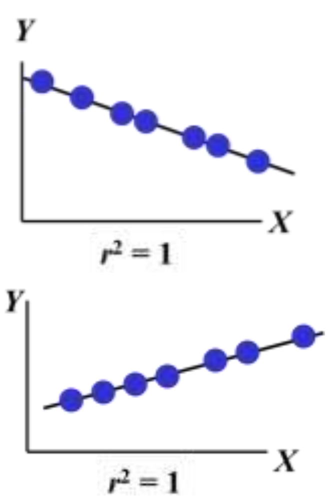
$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

note: $0 \leq R^2 \leq 1$

 Slide - 29

29

Examples of Approximate *r* Squared Values (1 of 3)



$r^2 = 1$

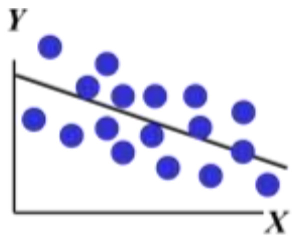
Perfect linear relationship between *X* and *Y*:

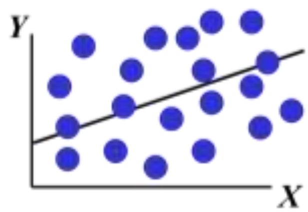
100% of the variation in *Y* is explained by variation in *X*

 Slide - 30

30

Examples of Approximate *r* Squared Values (2 of 3)



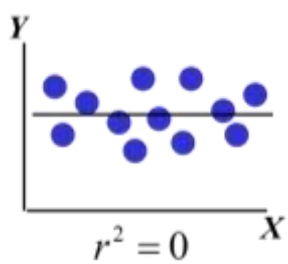


$$0 < r^2 < 1$$

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* Squared Values (3 of 3)



$$r^2 = 0$$

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*)

Excel Output (3 of 6)

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

33

Correlation and *R* Squared

- The coefficient of determination, R^2, for a simple regression is equal to the simple correlation squared

$$R^2 = r^2$$

34

Estimation of Model Error Variance

- An estimator for the variance of the population model error is

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\text{SSE}}{n-2}$$

- Division by $n - 2$ instead of $n - 1$ is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one

$s_e = \sqrt{s_e^2}$ is called the standard error of the estimate

Excel Output (4 of 6)

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

$s_e = 41.33032$

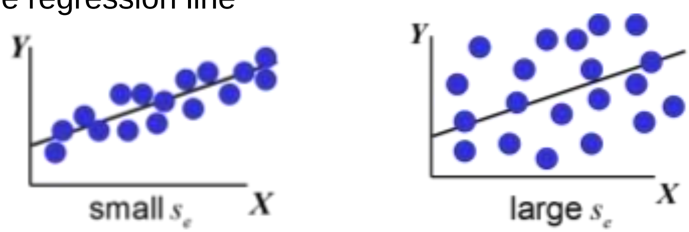
ANOVA

	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

Comparing Standard Errors

s_e is a measure of the variation of observed *y* values from the regression line



The magnitude of s_e should always be judged relative to the size of the *y* values in the sample data

i.e., $s_e = \$41.33K$ is moderately small relative to house prices in the $\$200 - \$300K$ range

Section 11.5 Statistical Inference: Hypothesis Tests and Confidence Intervals

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

$$s_{b_1}^2 = \frac{s_e^2}{\sum (x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x^2}$$

where:

s_{b_1} = Estimate of the standard error of the least squares slope

$s_e = \sqrt{\dfrac{\text{SSE}}{n-2}}$ = Standard error of the estimate

Excel Output (5 of 6)

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

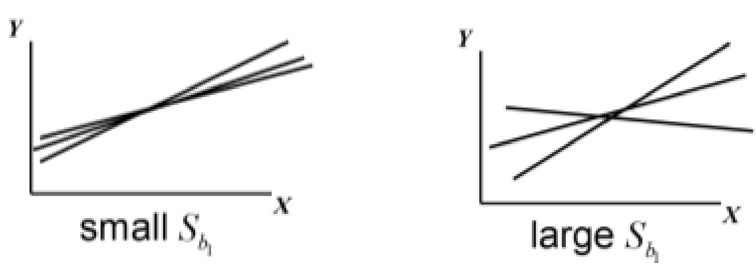
$$S_{b_1} = 0.03297$$

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

Comparing Standard Errors of the Slope

S_{b_1} is a measure of the variation in the slope of regression lines from different possible samples



Inference About the Slope: *t* Test (1 of 2)

- *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

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$\text{d.f.} = n - 2$$

where:

$b_1 =$ regression slope coefficient

$\beta_1 =$ hypothesized slope

$s_{b_1} =$ standard error of the slope

Pearson Slide - 41

41

Inference About the Slope: *t* Test (2 of 2)

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:

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

The slope of this model is 0.1098

Does square footage of the house significantly affect its sales price?

Pearson Slide - 42

42

Inferences About the Slope: *t* Test Example (1 of 3)

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

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

(b_1 = 0.10977, s_{b_1} = 0.03297)

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences About the Slope: *t* Test Example (2 of 3)

Test Statistic: ***t* = 3.329**

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

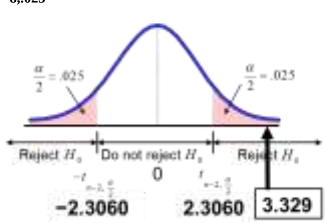
d.f. = 10 − 2 = 8

$t_{8,.025} = 2.3060$

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

(b_1 = 0.10977, s_{b_1} = 0.03297, t = 3.32938)



Decision:

Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Inferences About the Slope: *t* Test Example (3 of 3)

P-value = **0.01039**

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

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

This is a two-tail test, so the *p*-value is

$P(t > 3.329) + P(t < -3.329)$

$= 0.01039$

(for 8 d.f.)

Decision: *P*-value $< \alpha$ so Reject H_0

Conclusion:
There is sufficient evidence that square footage affects house price

 Slide - 45

45

Confidence Interval Estimate for the Slope (1 of 2)

Confidence Interval Estimate of the Slope:

$$b_1 - t_{n-2,\frac{\alpha}{2}} s_{b_1} < \beta_1 < b_1 + t_{n-2,\frac{\alpha}{2}} s_{b_1}$$

$$\textbf{d.f.} = n - 2$$

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)

 Slide - 46

46

Confidence Interval Estimate for the Slope (2 of 2)

	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 $33.70 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

Hypothesis Test for Population Slope Using the *F* Distribution (1 of 2)

- *F* Test statistic:

$$F = \frac{\text{MSR}}{\text{MSE}}$$

where

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

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

where *F* 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)

Hypothesis Test for Population Slope Using the *F* Distribution (2 of 2)

- An alternate test for the hypothesis that the slope is zero:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- Use the *F* statistic

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}}{s_e^2}$$

- The decision rule is

reject H_0 if $F \geq F_{1,n-2,\alpha}$

Excel Output (6 of 6)

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

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

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			

	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

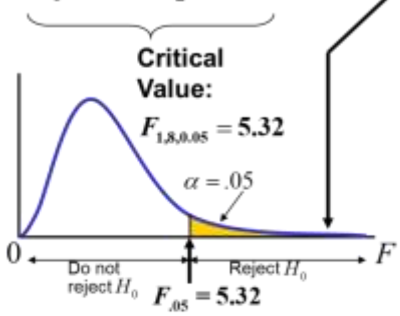
F-Test for Significance

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\alpha = .05$

$df_1 = 1 \quad df_2 = 8$



Test Statistic:

$$F = \frac{\text{MSR}}{\text{MSE}} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

 Slide - 51

51

Section 11.6 Prediction

- The regression equation can be used to predict a value for *y*, given a particular *x*
- For a specified value, x_{n+1}, the predicted value is

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$

 Slide - 52

52

Predictions Using Regression Analysis

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$
$$= 98.25 + 0.1098(2000)$$
$$= 317.85$$

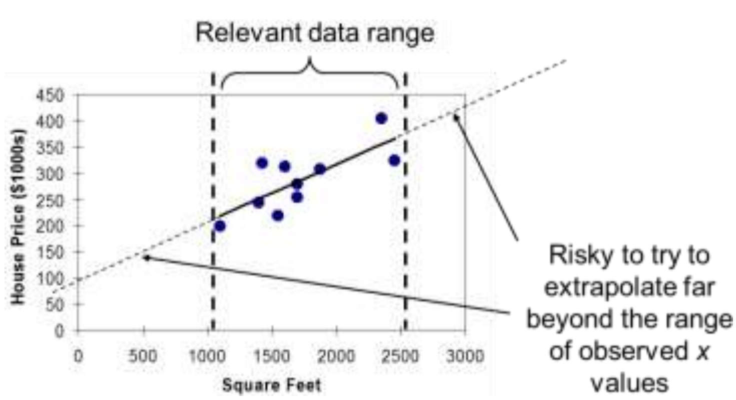
The predicted price for a house with 2000 square feet is $317.85(\$1{,}000\text{s}) = \$317{,}850$

Pearson Slide - 53

53

Relevant Data Range

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

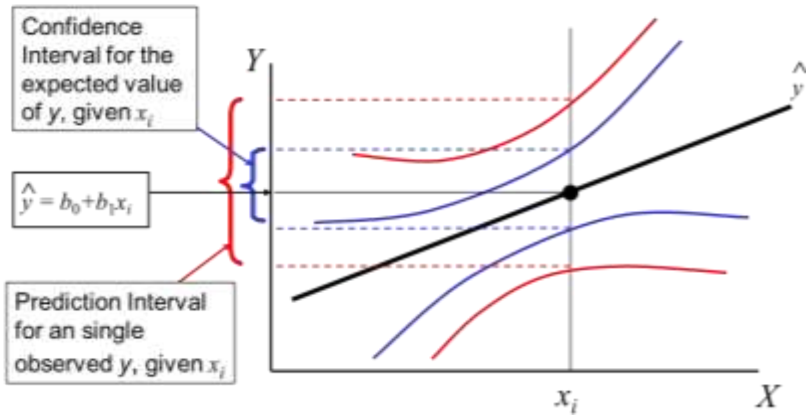


Pearson Slide - 54

54

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



 Slide - 55

55

Confidence Interval for the Average *Y*, Given *X*

Confidence interval estimate for the **expected value of *y*** given a particular x_i

Confidence interval for $E(Y_{n+1} | X_{n+1})$:

$$\hat{y}_{n+1} \pm t_{n-2,\frac{\alpha}{2}} s_e \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

Notice that the formula involves the term $(x_{n+1} - \bar{x})^2$

so the size of interval varies according to the distance x_{n+1} is from the mean, $\bar{x}$

 Slide - 56

56

Prediction Interval for an Individual *Y*, Given *X*

Confidence interval estimate for an **actual observed value of *y*** given a particular x_i

Confidence interval for $\hat{y}_{n+1}$:

$$\hat{y}_{n+1} \pm t_{n-2,\frac{\alpha}{2}} s_e \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

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

 Slide - 57

57

Example: Confidence Interval for the Average *Y*, Given *X* (1 of 2)

Confidence Interval Estimate for $E(Y_{n+1} | X_{n+1})$

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

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

$$\hat{y}_{n+1} \pm t_{n-2,\frac{\alpha}{2}} s_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.73 and 354.97, or from $280,730 to $354,970

 Slide - 58

58

Example: Confidence Interval for the Average *Y*, Given *X* (2 of 2)

Confidence Interval Estimate for $\hat{y}_{n+1}$

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

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

$$\hat{y}_{n+1} \pm t_{n-1,\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = 317.85 \pm 102.28$$

The confidence interval endpoints are 215.57 and 420.13, or from \$215,570 to \$420,130

Section 11.7 Correlation Analysis (1 of 2)

- Correlation analysis is used to measure 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
  - Correlation was first presented in Chapter 4

Section 11.7 Correlation Analysis (2 of 2)

- The population correlation coefficient is denoted ρ (the Greek letter rho)
- The sample correlation coefficient is

$$r = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

 Slide - 61

61

Test for Zero Population Correlation

- To test the null hypothesis of no linear association,

$$H_0 : \rho = 0$$

the test statistic follows the Student's *t* distribution with (n − 2) degrees of freedom:

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

 Slide - 62

62

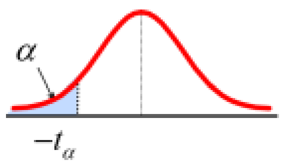
Decision Rules

Hypothesis Test for Correlation

Lower-tail test:

$H_0: \rho \geq 0$

$H_1: \rho < 0$

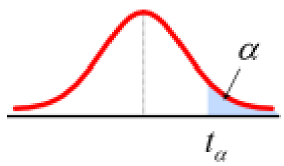


Reject H_0 if $t < -t_{n-2,\alpha}$

Upper-tail test:

$H_0: \rho \leq 0$

$H_1: \rho > 0$

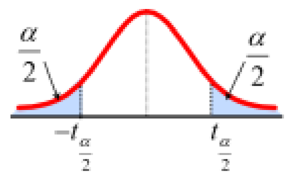


Reject H_0 if $t > t_{n-2,\alpha}$

Two-tail test:

$H_0: \rho = 0$

$H_1: \rho \neq 0$



Reject H_0 if $t < -t_{n-2,\frac{\alpha}{2}}$

or $t > t_{n-2,\frac{\alpha}{2}}$

Where $t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$ has $n-2$ d.f.

 Slide - 63

63

Section 11.8 Beta Measure of Financial Risk

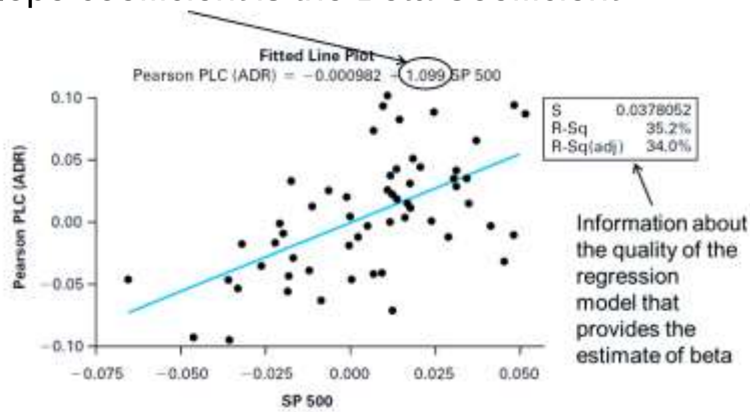
- A Beta Coefficient is a measure of how the returns of a particular firm respond to the returns of a broad stock index (such as the S&P 500)
- For a specific firm, the Beta Coefficient is the slope coefficient from a regression of the firm's returns compared to the overall market returns over some specified time period

 Slide - 64

64

Beta Coefficient Example

- Slope coefficient is the Beta Coefficient



Regression Analysis: Pearson PLC (ADR) versus SP 500

Section 11.9 Graphical Analysis

- The linear regression model is based on minimizing the sum of squared errors
- If outliers exist, their potentially large squared errors may have a strong influence on the fitted regression line
- Be sure to examine your data graphically for outliers and extreme points
- Decide, based on your model and logic, whether the extreme points should remain or be removed

Chapter Summary

- Introduced the linear regression model
- Reviewed correlation and the assumptions of linear regression
- Discussed estimating the simple linear regression coefficients
- Described measures of variation
- Described inference about the slope
- Addressed estimation of mean values and prediction of individual values