

Research Methods

Linear Regression and Correlation Analysis

Learning outcomes

By the end of this Session, you will be able to:

- Calculate and interpret the correlation between two variables.
- Conduct a test of hypothesis to determine whether the coefficient of correlation in the population is zero.
- Calculate and interpret the simple linear regression equation for a set of data.
- Understand the assumptions behind regression analysis.

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

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

Correlation Analysis

Correlation analysis is used to measure the strength of the association (linear relationship) between two variables

- Positive – as x increases, y increases
 - » As advertising spend increases, sales increase
- Negative – as x increases, y decreases
 - » As expenses increase, net income decreases

Correlation is only concerned with strength of the relationship

No causal effect is implied with correlation

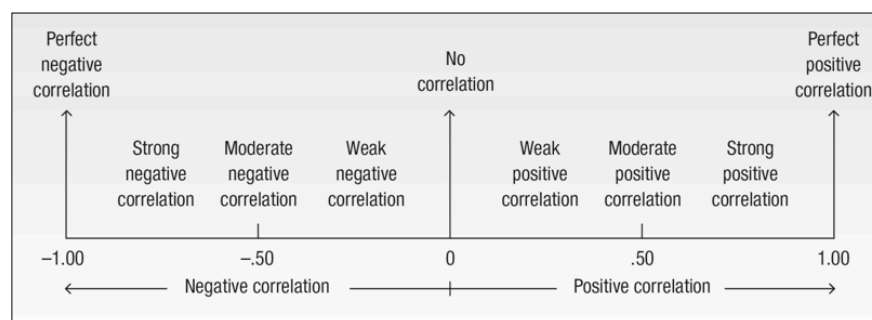
A spurious correlation is when a correlation exists between two seemingly unrelated variables

- the correlation between company sales and the number of children employees have.

Correlation Coefficient

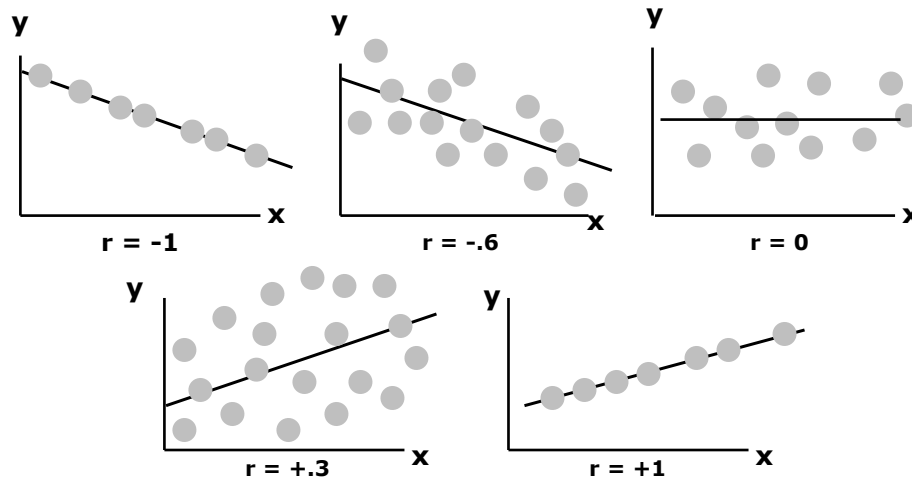
- The sample correlation coefficient r is a measure of the strength of the linear relationship between two variables, based on sample observations
 - Unit free
 - Range between -1 and 1
 - The closer to -1, the stronger the negative linear relationship
 - The closer to 1, the stronger the positive linear relationship
 - The closer to 0, the weaker the linear relationship
 - +1 or -1 are perfect correlations where all data points fall on a straight line

Correlation Coefficient: Interpretation



Statistical Techniques in Business & Economics, ©The McGraw-Hill Companies, Inc. 2008

Examples of Approximate r Values



Business Statistics: A Decision-Making Approach, 8e©Pearson Education, Inc. 2011

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 variable
- y = Value of the dependent variable

Business Statistics: A Decision-Making Approach, 8e©Pearson Education, Inc. 2011

Significance Test for Correlation

$H_0: \rho = 0$ (the correlation in the population is 0)

$H_1: \rho \neq 0$ (the correlation in the population is not 0)

Reject H_0 if:

$$t > t_{\alpha/2, n-2} \text{ or } t < -t_{\alpha/2, n-2}$$

- Test statistic

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

(with $n - 2$ degrees of freedom)

We lose 1 degree of freedom for each sample mean

Testing the Significance of the Correlation Coefficient: Example

Suppose a sales manager wants to test for a correlation between number of sales calls made and items sold in a population of phone salespeople:

$H_0: \rho = 0$ (the correlation in the population is 0)

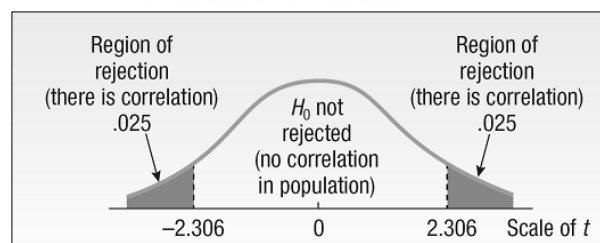
$H_1: \rho \neq 0$ (the correlation in the population is not 0)

Reject H_0 if:

$$t > t_{\alpha/2, n-2} \text{ or } t < -t_{\alpha/2, n-2}$$

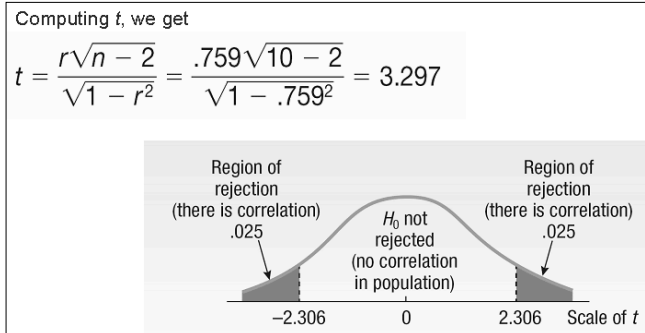
$$t > t_{0.025, 8} \text{ or } t < -t_{0.025, 8}$$

$$t > 2.306 \text{ or } t < -2.306$$



Statistical Techniques in Business & Economics, ©The McGraw-Hill Companies, Inc. 2008

Testing the Significance of the Correlation Coefficient: Example, Cont.



The computed t (3.297) is within the rejection region, therefore, we will reject H_0 . This means the correlation in the population is not zero. From a practical standpoint, it indicates to the sales manager that there is correlation with respect to the number of sales calls made and the number of items sold in the population of salespeople.

Statistical Techniques in Business & Economics, ©The McGraw-Hill Companies, Inc. 2008

Simple Linear Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

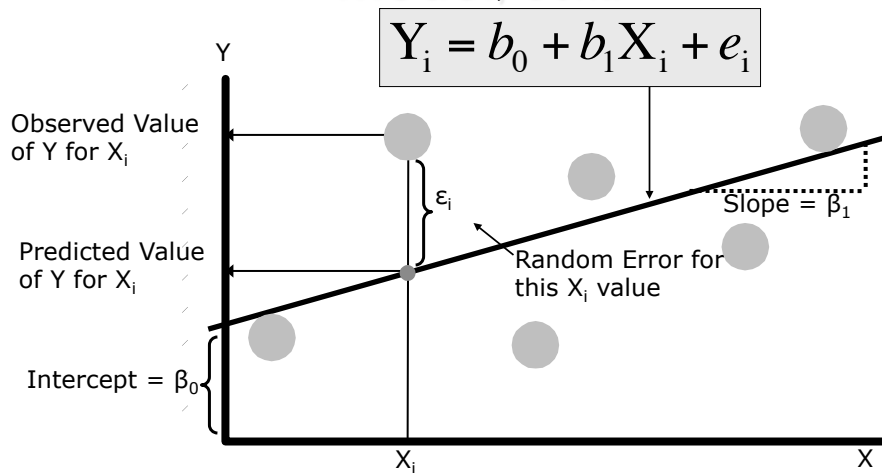
$$Y_i = b_0 + b_1 X_i + e_i$$

Linear component

Random Error component

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

Simple Linear Regression Model, Cont.



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

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

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

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

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

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

Measures of Variation

- Total variation is made up of two parts:

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

Total Sum
of Squares

Regression Sum
of Squares

Error Sum of
Squares

$$\boxed{SST = \sum (Y_i - \bar{Y})^2} \quad \boxed{SSR = \sum (\hat{Y}_i - \bar{Y})^2} \quad \boxed{SSE = \sum (Y_i - \hat{Y}_i)^2}$$

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

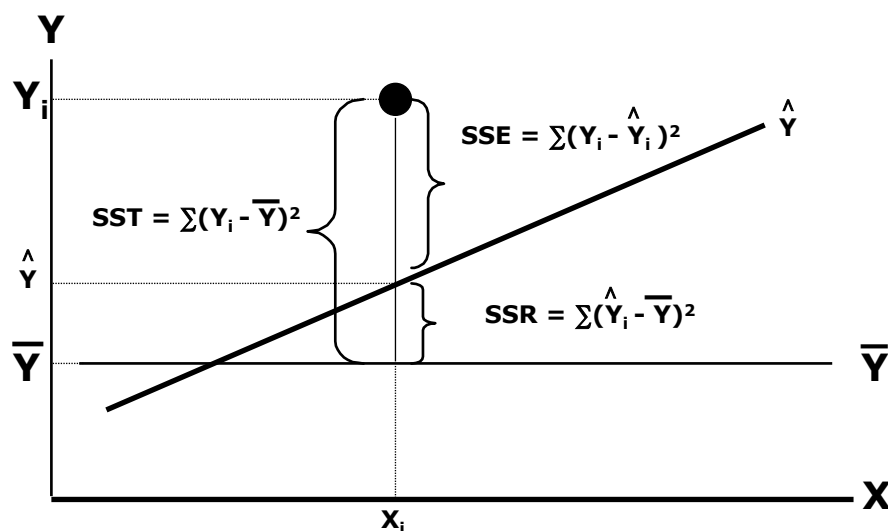
Basic Business Statistics, 11e © Prentice-Hall, Inc. 2009

Measures of Variation, Cont.

- 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

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

Measures of Variation



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

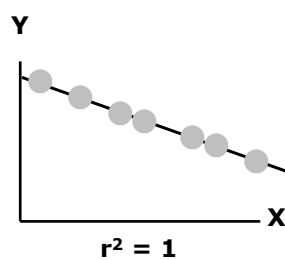
$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

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

Chap 13-19

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

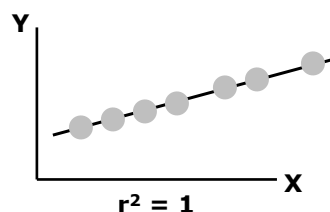
Examples of Approximate r^2 Values



$$r^2 = 1$$

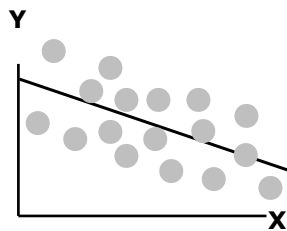
**Perfect linear relationship
between X and Y:**

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



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

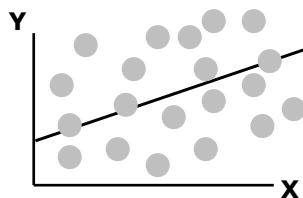
Examples of Approximate r^2 Values, Cont.



$$0 < r^2 < 1$$

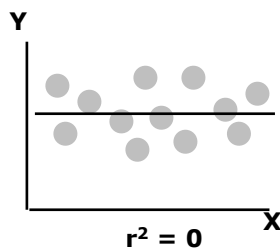
Weaker linear relationships between X and Y:

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



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

Examples of Approximate r^2 Values, Cont.



$$r^2 = 0$$

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)

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

The Least Squares Method

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

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

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

Simple Linear Regression: Example

- An 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



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

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



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

Simple Linear Regression Example: Using Excel

Microsoft Excel - 13data.xls

File Edit View Insert Format Tools Data Window Help Acrobat

Chart 1

	A	B
1	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
12		
13		
14		
15		

Regression

Input:

Input Y Range:

Input X Range:

☒ Labels ☐ Constant is Zero

☐ Confidence Level: %

Output options:

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals:

☐ Residuals ☐ Standardized Residuals ☐ Residual Plots ☒ Line Fit Plots

Normal Probability:

☐ Normal Probability Plots

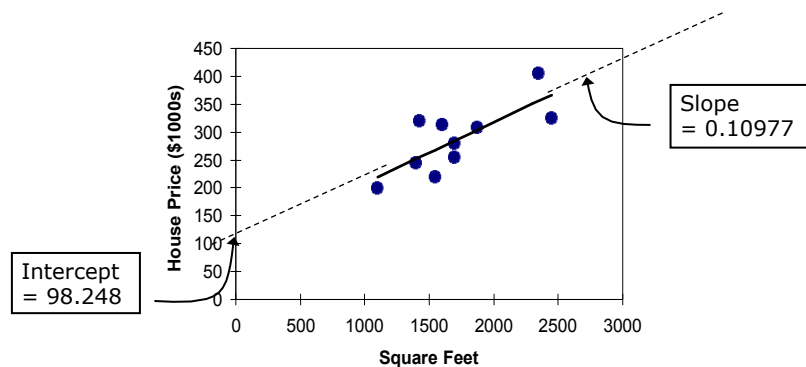
OK Cancel Help



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

Simple Linear Regression Example: Graphical Representation

House price model: Scatter Plot and Prediction Line



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



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

Simple Linear Regression Example: Interpretation of b_0

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

- 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)
- Because a house cannot have a square footage of 0, b_0 has no practical application



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

Simple Linear Regression Example: Interpreting b_1

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

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



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

Simple Linear Regression Example: Making Predictions

Predict the price for a house with 2000 square feet:

$$\begin{aligned}\widehat{\text{house price}} &= 98.25 + 0.1098 (\text{sq.ft.}) \\ &= 98.25 + 0.1098(2000) \\ &= 317.85\end{aligned}$$

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

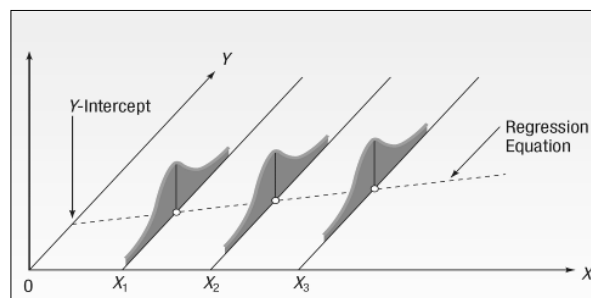


Basic Business Statistics, 11e © Prentice-Hall, Inc. 2009

Assumptions Underlying Linear Regression

For each value of X , there is a group of Y values, and these

- Y values are *normally distributed*. The *means* of these normal distributions of Y values all lie on the straight line of regression.
- The *standard deviations* of these normal distributions are equal.
- The Y values are statistically independent. This means that in the selection of a sample, the Y values chosen for a particular X value do not depend on the Y values for any other X values.



Statistical Techniques in Business & Economics, ©The McGraw-Hill Companies, Inc. 2008