Chapter 13
Introduction to Linear Regression and Correlation Analysis

Chapter Goals

To understand the methods for displaying and describing relationship among variables

Methods for Studying Relationships

- Graphical
  - Scatterplots
  - Line plots
  - 3-D plots
- Models
  - Linear regression
  - Correlations
  - Frequency tables

Two Quantitative Variables

The *response variable*, also called the *dependent variable*, is the variable we want to predict, and is usually denoted by *y*.

The *explanatory variable*, also called the *independent variable*, is the variable that attempts to explain the response, and is denoted by *x*.

YDI 7.1

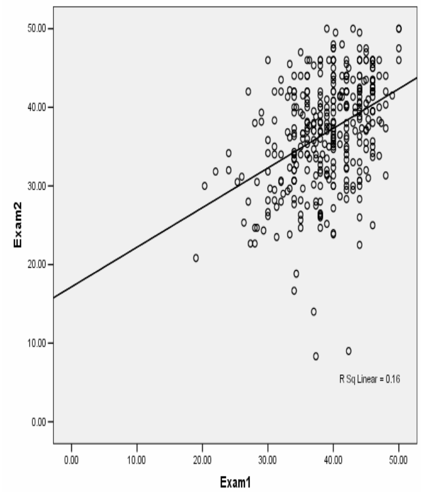
Response (*y*)	Explanatory (*x*)
Height of son	
Weight	

Scatter Plots and Correlation

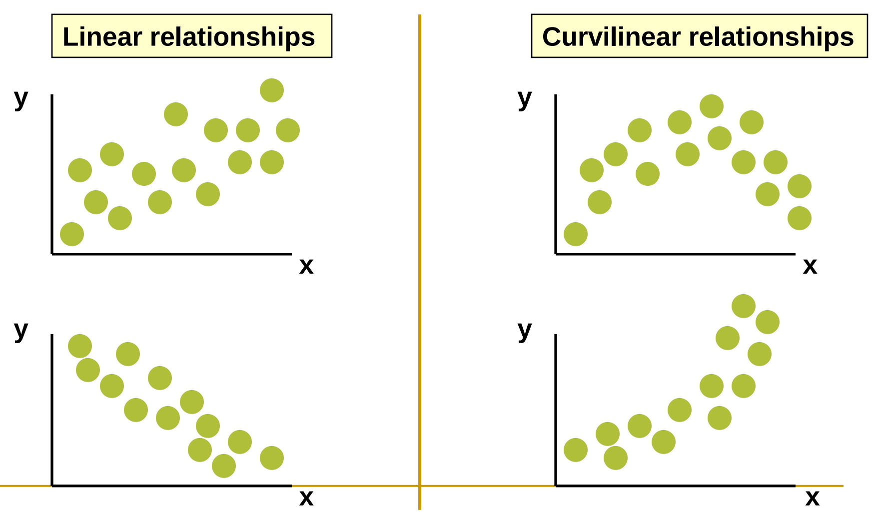
- A scatter plot (or scatter diagram) is used to show the relationship between two variables
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Only concerned with strength of the relationship
  - No causal effect is implied

Example

- The following graph shows the scatterplot of Exam 1 score (x) and Exam 2 score (y) for 354 students in a class. Is there a relationship?

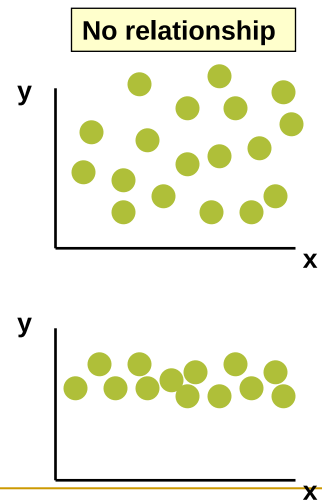


Scatter Plot Examples



Scatter Plot Examples

(continued)



Fall 2006 – Fundamentals of Business Statistics 9

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

Fall 2006 – Fundamentals of Business Statistics 10

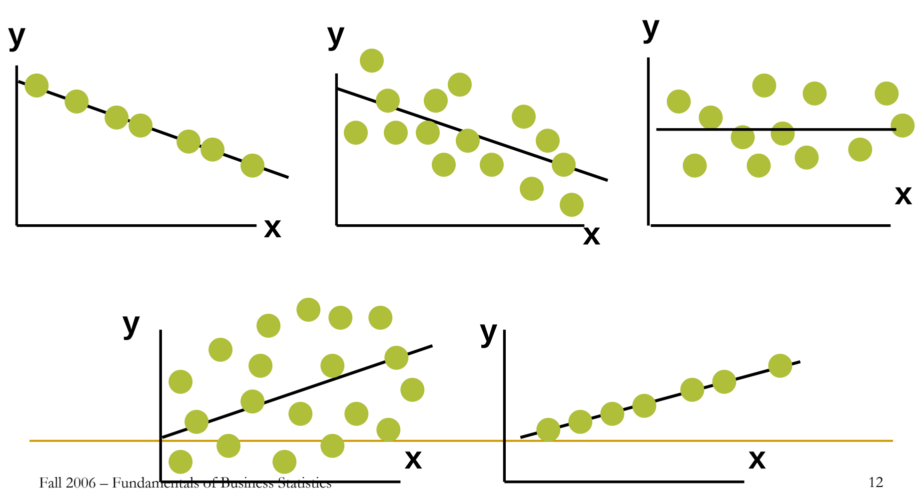
Features of ρ and r

- 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

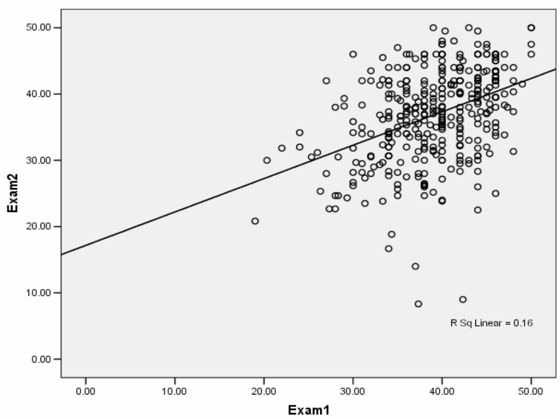
Examples of Approximate r Values

Tag with appropriate value:

-1, -.6, 0, +.3, 1



Earlier Example



Correlations

		Exam1	Exam2
Exam1	Pearson Correlation	1	.400**
	Sig. (2-tailed)		.000
	N	366	351
Exam2	Pearson Correlation	.400**	1
	Sig. (2-tailed)	.000	
	N	351	356

**. Correlation is significant at the 0.01 level

YDI 7.3

What kind of relationship would you expect in the following situations:

- age (in years) of a car, and its price.
- number of calories consumed per day and weight.
- height and IQ of a person.

YDI 7.4

Identify the two variables that vary and decide which should be the independent variable and which should be the dependent variable. Sketch a graph that you think best represents the relationship between the two variables.

1. The size of a persons vocabulary over his or her lifetime.
2. The distance from the ceiling to the tip of the minute hand of a clock hung on the wall.

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

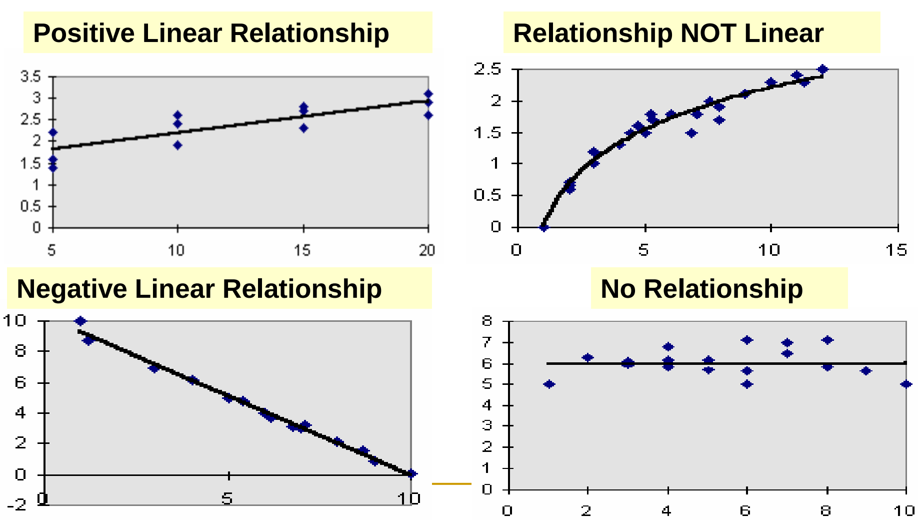
Dependent variable: the variable we wish to explain

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



Population Linear Regression

The population regression model:

Dependent Variable; Population y intercept; Population Slope Coefficient; Independent Variable; Random Error term, or residual

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

Linear component ($\beta_0 + \beta_1 x$); Random Error component (ε)

Fall 2006 – Fundamentals of Business Statistics 19

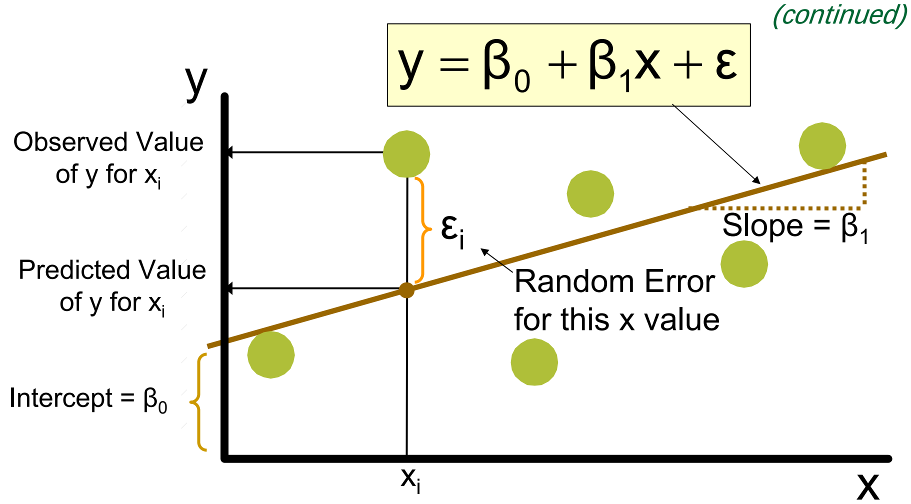
Linear Regression Assumptions

- Error values (ε) are statistically independent
- Error values are normally distributed for any given value of x
- The probability distribution of the errors is normal
- The probability distribution of the errors has constant variance
- The underlying relationship between the x variable and the y variable is linear

Fall 2006 – Fundamentals of Business Statistics 20

Population Linear Regression

(continued)



Fall 2006 – Fundamentals of Business Statistics 21

Estimated Regression Model

The sample regression line provides an estimate of the population regression line

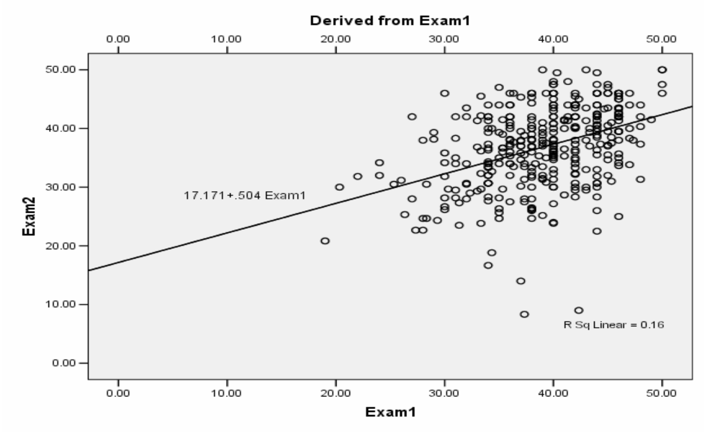
$$\hat{y}_i = b_0 + b_1 x$$

- Estimated (or predicted) y value: $\hat{y}_i$
- Estimate of the regression intercept: b_0
- Estimate of the regression slope: b_1
- Independent variable: x

The individual random error terms e_i have a mean of zero

Fall 2006 – Fundamentals of Business Statistics 22

Earlier Example



Residual

A ***residual*** *is the difference between the observed response y and the predicted response ŷ. Thus, for each pair of observations* (x_i, y_i), *the* i^{th} *residual is*
$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x)$

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

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

Fall 2006 – Fundamentals of Business Statistics 25

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

Fall 2006 – Fundamentals of Business Statistics 26

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

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

Fall 2006 – Fundamentals of Business Statistics 27

Finding the Least Squares Equation

- The coefficients b_0 and b_1 will usually be found using computer software, such as Excel, Minitab, or SPSS.
- Other regression measures will also be computed as part of computer-based regression analysis

Fall 2006 – Fundamentals of Business Statistics 28

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) = squar

Fall 2006 – Fundamentals of Business Statistics 29

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

Fall 2006 – Fundamentals of Business Statistics 30

SPSS Output

The regression equation is:

$$\widehat{\text{house price}} = 98.248 + 0.110\,(\text{square feet})$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.762[a]	.581	.528	41.33032

a. Predictors: (Constant), Square Feet

Coefficients[a]

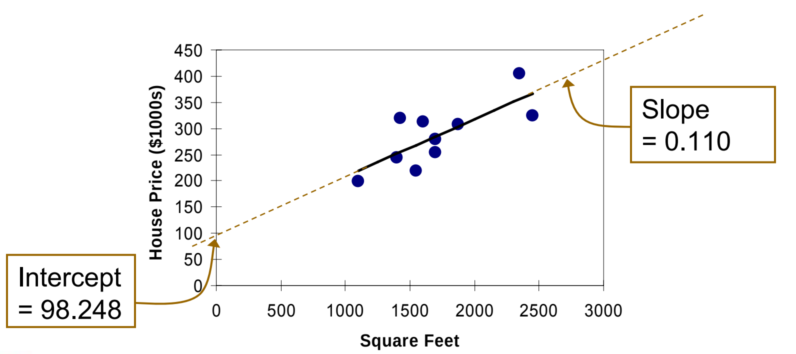
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	98.248	58.033		1.693	.129
	Square Feet	.110	.033	.762	3.329	.010

a. Dependent Variable: House Price

Fall 2006 – Fundamentals of Business Statistics 31

Graphical Presentation

- House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.248 + 0.110\,(\text{square feet})$$

Fall 2006 – Fundamentals of Business Statistics 32

Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = 98.248 + 0.110\,(\text{square feet})$$

- 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 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 fee

Fall 2006 – Fundamentals of Business Statistics 33

Interpretation of the Slope Coefficient, b_1

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

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, 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

Fall 2006 – Fundamentals of Business Statistics 34

Least Squares Regression Properties

- The sum of the residuals from the least squares regression line is 0 ($\sum(y-\hat{y})=0$)
- The sum of the squared residuals is a minimum (minimized $\sum(y-\hat{y})^2$)
- The simple regression line always passes through the mean of the y variable and the mean of the x variable
- The least squares coefficients are unbiased estimates of β_0 and β_1

Fall 2006 – Fundamentals of Business Statistics 35

YDI 7.6

The growth of children from early childhood through adolescence generally follows a linear pattern. Data on the heights of female Americans during childhood, from four to nine years old, were compiled and the least squares regression line was obtained as $\hat{y} = 32 + 2.4x$ where $\hat{y}$ is the predicted height in inches, and x is age in years.

- Interpret the value of the estimated slope $b_1 = 2.4$.
- Would interpretation of the value of the estimated y-intercept, $b_0 = 32$, make sense here?
- What would you predict the height to be for a female American at 8 years old?
- What would you predict the height to be for a female American at 25 years old? How does the quality of this answer compare to the previous question?

Fall 2006 – Fundamentals of Business Statistics 36

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

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

Coefficient of Determination, R^2

(continued)

Note: In the single independent variable case, the coefficient of determination is

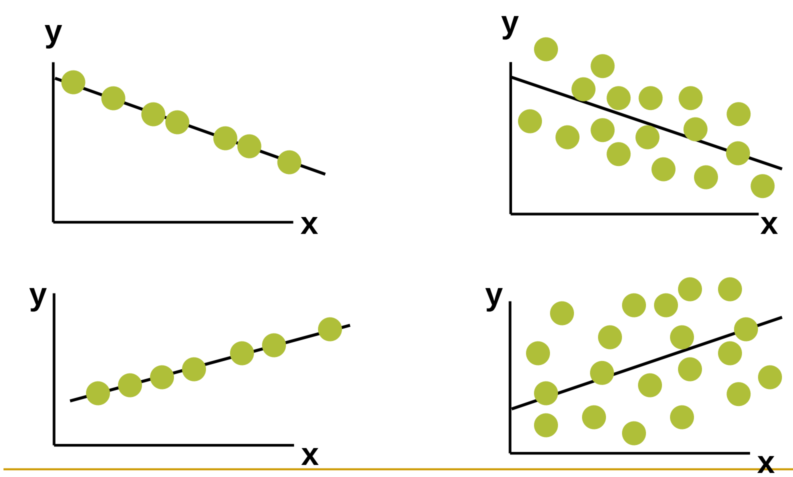
$$R^2 = r^2$$

where:

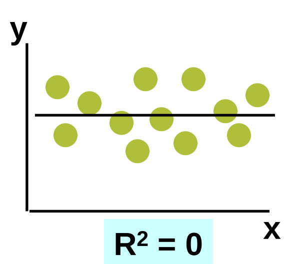
R^2 = Coefficient of determination

r = Simple correlation coefficient

Examples of Approximate R^2 Values



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)

SPSS Output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.762[a]	.581	.528	41.33032

a. Predictors: (Constant), Square Feet

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18934.935	1	18934.935	11.085	.010[a]
	Residual	13665.565	8	1708.196		
	Total	32600.500	9			

a. Predictors: (Constant), Square Feet

b. Dependent Variable: House Price

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	98.248	58.033		1.693	.129
	Square Feet	.110	.033	.762	3.329	.010

a. Dependent Variable: House Price

Fall 2006 – Fundamentals of Business Statistics 41

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is called the *standard error of estima* s_{ε}
- The standard error of the regression slope coefficient (b_1) is given by s_{b1}

Fall 2006 – Fundamentals of Business Statistics 42

SPSS Output

$s_\varepsilon = 41.33032$

$s_{b_1} = 0.03297$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.762[a]	.581	.528	41.33032

a. Predictors: (Constant), Square Feet

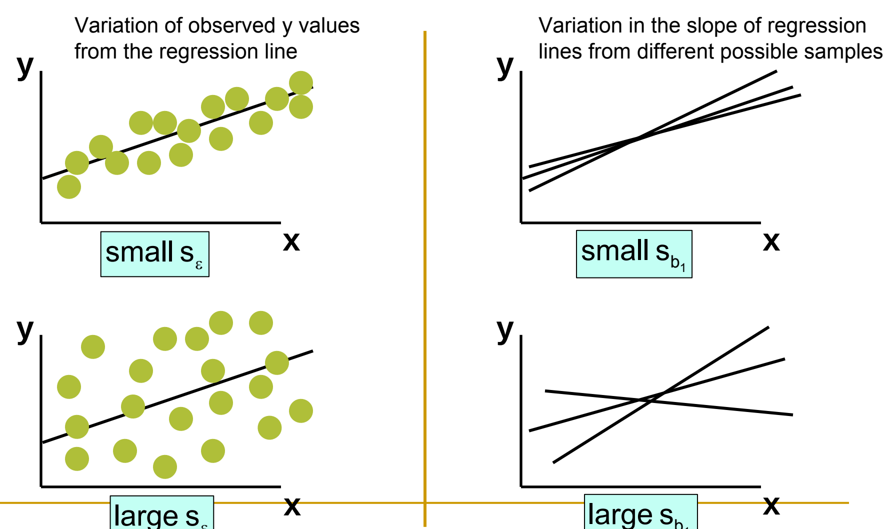
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	98.248	58.033		1.693	.129
	Square Feet	.110	.033	.762	3.329	.010

a. Dependent Variable: House Price

Fall 2006 – Fundamentals of Business Statistics 43

Comparing Standard Errors



Fall 2006 – Fundamentals of Business Statistics 44

Inference 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
  - $$t = \frac{b_1 - \beta_1}{s_{b_1}}$$
  - $$d.f. = n - 2$$

where:

b_1 = Sample regression slope coefficient

β_1 = Hypothesized slope

s_{b1} = Estimator of the standard error of the slope

Fall 2006 – Fundamentals of Business Statistics 45

Inference about the Slope: t Test

(continued)

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 affect its sales price?

Fall 2006 – Fundamentals of Business Statistics 46

Inferences about the Slope: t Test Example

Test Statistic: t = 3.329

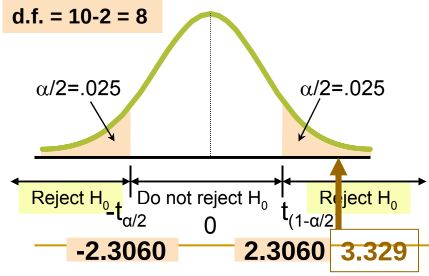
H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$

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 = 3.32938)

d.f. = 10-2 = 8



Decision:
Reject H_0

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

Fall 2006 – Fundamentals of Business Statistics 47

Regression Analysis for Description

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{(1-\alpha/2)} s_{b_1}$$

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)

Fall 2006 – Fundamentals of Business Statistics 48

Regression Analysis for Description

	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

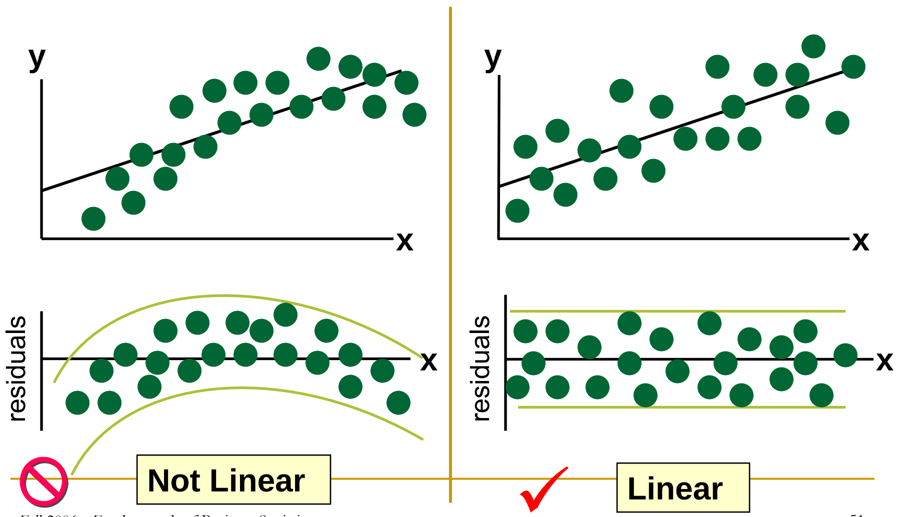
Fall 2006 – Fundamentals of Business Statistics 49

Residual Analysis

- Purposes
  - Examine for linearity assumption
  - Examine for constant variance for all levels of x
  - Evaluate normal distribution assumption
- Graphical Analysis of Residuals
  - Can plot residuals vs. x
  - Can create histogram of residuals to check for normality

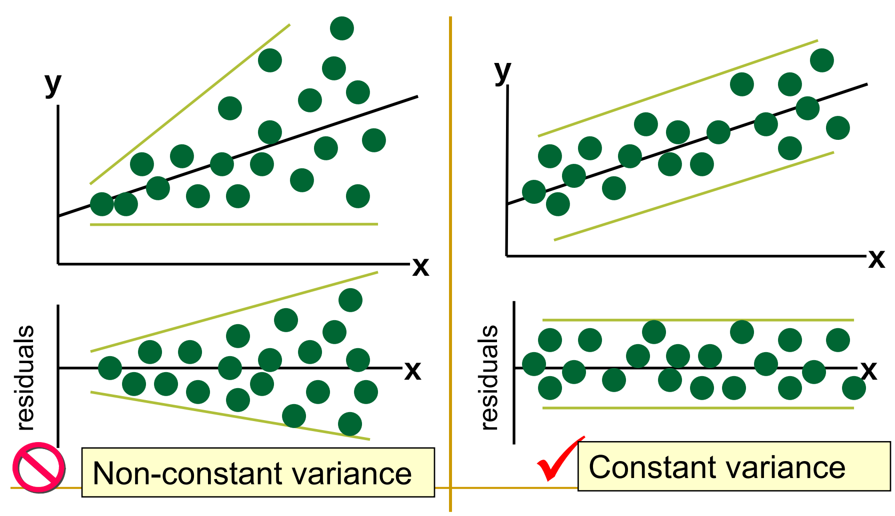
Fall 2006 – Fundamentals of Business Statistics 50

Residual Analysis for Linearity



Fall 2006 – Fundamentals of Business Statistics 51

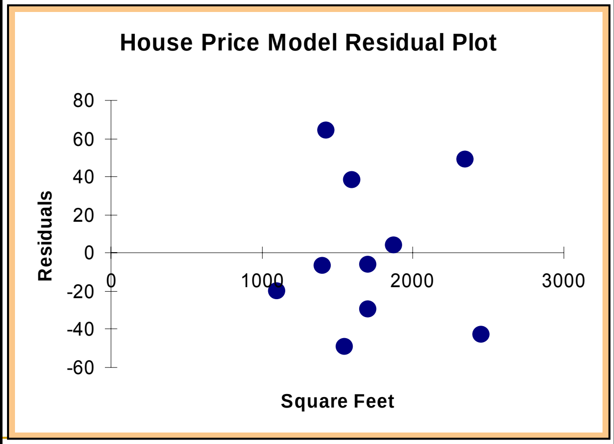
Residual Analysis for Constant Variance



Fall 2006 – Fundamentals of Business Statistics 52

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



Fall 2006 – Fundamentals of Business Statistics 53