SCSI2143: PROBABILITY & STATISTICAL DATA ANALYSIS

CHAPTER 7

PART 2:
LINEAR REGRESSION MODEL

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

Introduction to Regression Analysis

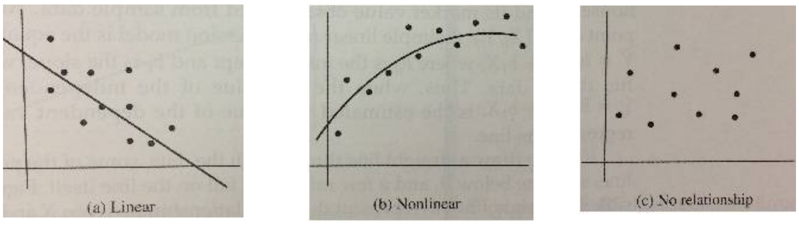
- A regression model that involves a single independent variable is called **simple regression**.
  - Example: imagine that your company wants to understand how past advertising expenditures have related to sales in order to make future decisions about advertising. The dependent variable in this instance is sales and the independent variable is advertising expenditures.

Introduction to Regression Analysis

- Usually, more than one independent variable influences the dependent variable.
- A regression model that involves two or more independent variables is called **multiple regression**.
  - Example: Sales are influenced by advertising as well as other factors, such as the number of sales representatives and the commission percentage paid to sales representatives

Introduction to Regression Analysis

- Regression models can be either linear or nonlinear.
- A linear model assumes the relationships between variables are straight-line relationships, while a nonlinear model assumes the relationships between variables are represented by curved lines.



Introduction to Regression Analysis

- The most basic type of regression is that of simple linear regression.
- A simple linear regression uses only one independent variable, and it describes the relationship between the independent variable and dependent variable as a straight line.
- This chapter will focus on the basic case of a **simple linear regression**.

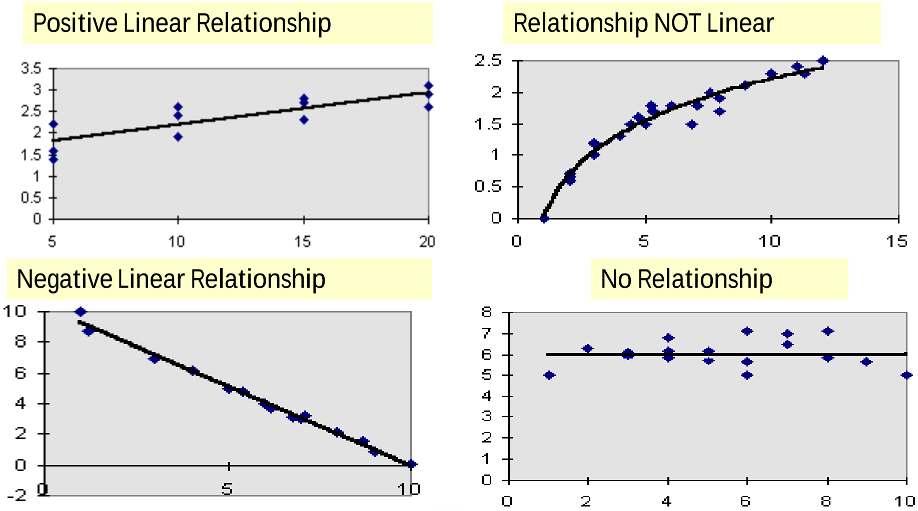
UTM UNIVERSITI TEKNOLOGI MALAYSIA

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*.

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Types of Regression Models



Population Linear Regression

The population regression model:

$$\underbrace{y}_{\text{Dependent Variable}} = \underbrace{\overbrace{\beta_0}^{\text{Population } y \text{ intercept}} + \overbrace{\beta_1}^{\text{Population Slope Coefficient}} \overbrace{x}^{\text{Independent Variable}}}_{\text{Linear component}} + \underbrace{\overbrace{\varepsilon}^{\text{Random Error term, or residual}}}_{\text{Random Error component}}$$

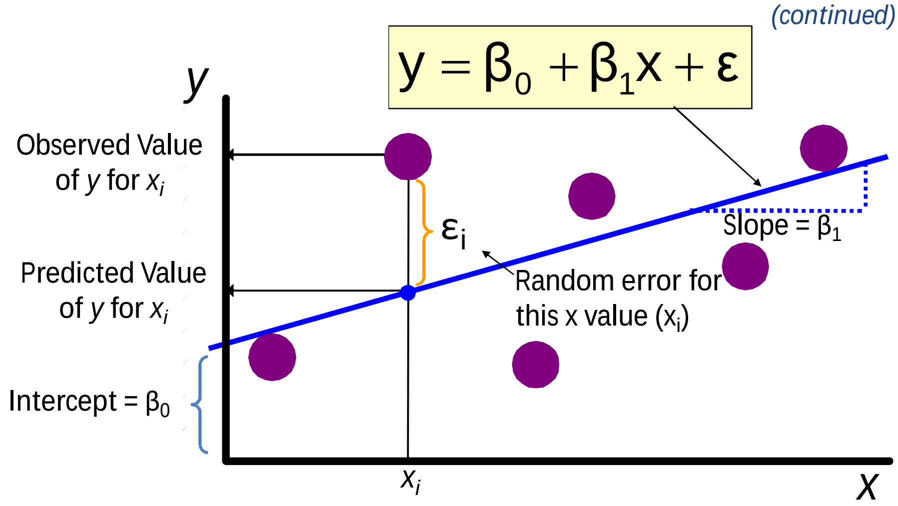
Linear Regression Assumptions

- 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

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Population Linear Regression

(continued)



UTM UNIVERSITI TEKNOLOGI MALAYSIA

Estimated Regression Model

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

Estimated (or predicted) y value

Estimate of the regression intercept

Estimate of the regression slope

Independent variable

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

The individual random error terms e_i have a mean of zero.

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

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

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

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

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*

Finding the Least Squares Equation

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

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

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

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

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

Example

y	x	xy	x^2
245	1400	343000	1960000
312	1600	499200	2560000
279	1700	474300	2890000
308	1875	577500	3515625
199	1100	218900	1210000
219	1550	339450	2402500
405	2350	951750	5522500
324	2450	793800	6002500
319	1425	454575	2030625
255	1700	433500	2890000
Σy=2865	Σx= 17150	Σxy= 5085975	Σx^2=30983750

Example

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

$$b_1 = \frac{5085975 - \frac{(17150)(2865)}{10}}{30983750 - \frac{(17150)^2}{10}}$$

$$= \frac{172500}{1571500} = 0.109767737$$

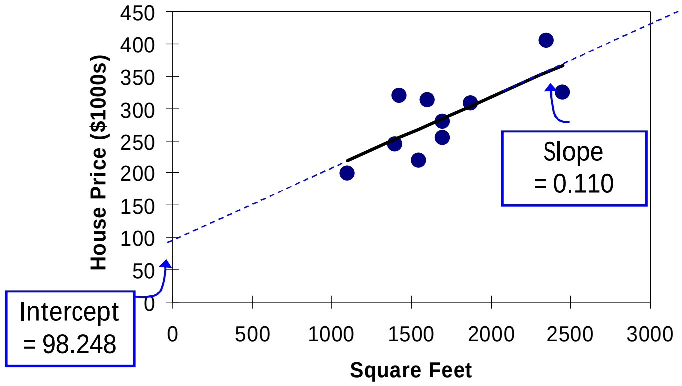
$$\boxed{b_0 = \bar{y} - b_1\bar{x}}$$

$$b_0 = 286.5 - 0.109767737(1715)$$

$$= 98.24832962$$

Graphical Presentation

- House price model: scatter plot and regression line



$$\hat{y} = 98.248 + 0.110\,x$$

Interpretation of the Intersection Coefficient, b_0

$$\hat{y} = \boxed{98.248} + 0.110x$$

- 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 = 98.248 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

Interpretation of the Slope Coefficient, b_1

$$\hat{y} = 98.248 + \boxed{0.110x}$$

- 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 = 0.110 tells us that the average value of a house increases by 0.110 ($1000) = $110, on average, for each additional one square foot of size

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

Explained and Unexplained Variation

- Total variation is made up of two parts:

$$SST = SSE + SSR$$

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

where:

$\bar{y}$ = Average value of the dependent variable
y = Observed values of the dependent variable
$\hat{y}$ = Estimated value of y for the given x value

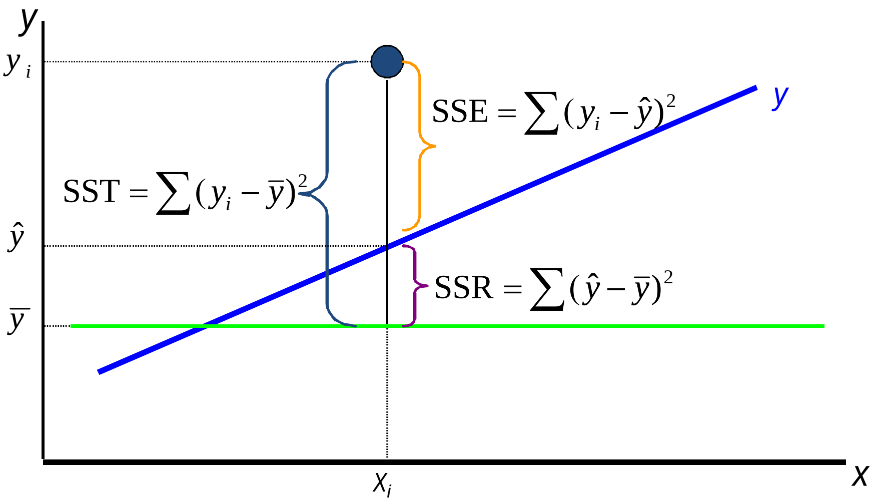
Explained and Unexplained Variation

(continued)

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean y
- SSE = error sum of squares
  - Variation attributable to factors other than the relationship between *x* and *y*
- SSR = regression sum of squares
  - Explained variation attributable to the relationship between *x* and *y*

Explained and Unexplained Variation

(continued)



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

$$\boxed{R^2 = \frac{SSR}{SST}} \quad \text{where} \quad \boxed{0 \le R^2 \le 1}$$

Coefficient of Determination, R^2

Coefficient of determination

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

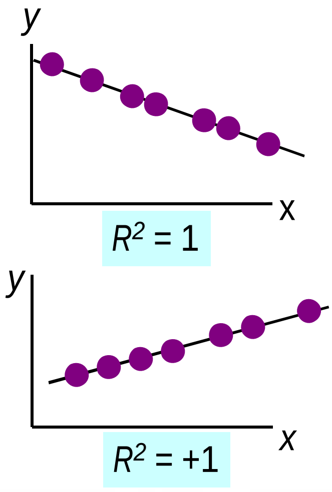
Note: In the single independent variable case, the coefficient of determination is

$$\boxed{R^2 = r^2}$$

where:

R^2 = Coefficient of determination
r = Simple correlation coefficient

Examples of Approximate R^2 Values



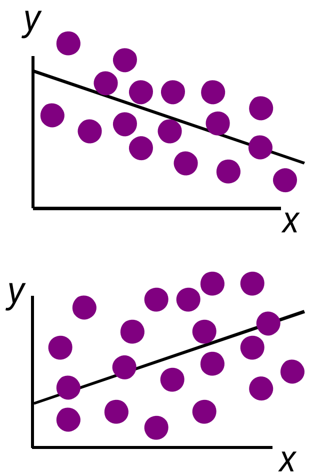
$R^2 = 1$

Perfect linear relationship between x and y:

100% of the variation in y is explained by variation in x

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Examples of Approximate R^2 Values



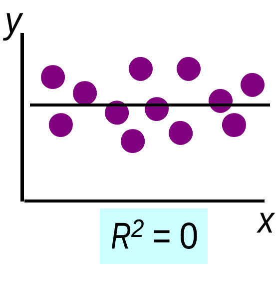
$0 < R^2 < 1$

Weaker linear relationship between x and y:

Some but not all of the variation in y is explained by variation in x

UTM UNIVERSITI TEKNOLOGI MALAYSIA

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)

Example

House Price in $1000s (y)	Square Feet (x)	$\hat{y}$	$(\hat{y}-\bar{y})^2$	$(y_i-\bar{y})^2$
245	1400	252.25	1173.06	1722.25
312	1600	274.25	150.06	650.25
279	1700	285.25	18.06	56.25
308	1875	304.50	324	462.25
199	1100	219.25	4522.56	7656.25
219	1550	268.75	315.05	4556.25
405	2350	356.75	4935.06	14042.25
324	2450	367.75	6601.56	1406.25
319	1425	255.00	992.25	1056.25
255	1700	285.25	1.56	992.25

$$\hat{y} = 98.248 + 0.110x$$

$$\bar{y} = \frac{\sum y}{n} = \frac{2865}{10} = 286.5$$

$$SSR = \sum(\hat{y}-\bar{y})^2 = 19033.22$$

$$SST = \sum(y_i-\bar{y})^2 = 13667.23$$

$$R^2 = \frac{SSR}{SST} = \frac{19033.22}{31700.5} = 0.60$$

60% 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

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$$

Where

SSE = Sum of squares error

n = Sample size

k = number of independent variables in the model

UTM UNIVERSITI TEKNOLOGI MALAYSIA

The Standard Deviation of the Regression Slope

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

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

where:

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

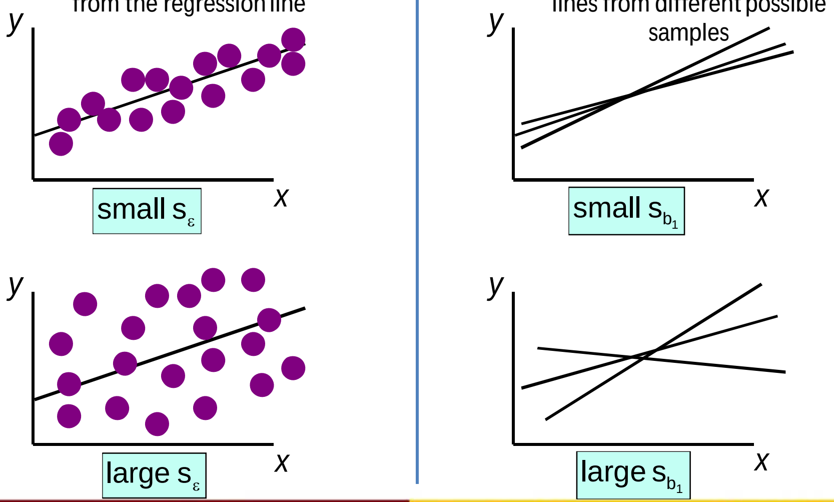
$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ = Sample standard error of the estimate

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Comparing Standard Errors

Variation of observed y values from the regression line

Variation in the slope of regression lines from different possible samples



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

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:

$$\hat{y} = 98.248 + 0.110x$$

The slope of this model is 0.110

Does square footage of the house affect its sales price?

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Inferences about the Slope: *t* Test Example

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

b_1

$$\hat{y} = 98.248 + 0.110x$$

$$s_\varepsilon = \sqrt{\frac{13667.23}{10-1-1}} = 41.33$$

$$s_{b_1} = \frac{41.33}{\sqrt{30983750 - \frac{294122500}{10}}} = 0.03$$

House Price in $1000s (y)	Square Feet (x)	$\hat{y}$	$(y_i - \hat{y})^2$
245	1400	252.25	52.56
312	1600	274.25	1425.06
279	1700	285.25	39.06
308	1875	304.50	12.25
199	1100	219.25	410.06
219	1550	268.75	2475.06
405	2350	356.75	2328.06
324	2450	367.75	1914.06
319	1425	255.00	4096
255	1700	285.25	915.06

$$SSE = \sum (y_i - \hat{y})^2 = 13667.23$$

www.utm.my innovative • entrepreneurial • global 39

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Inferences about the Slope: t Test Example

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0.110 - 0}{0.03} = 3.67$$

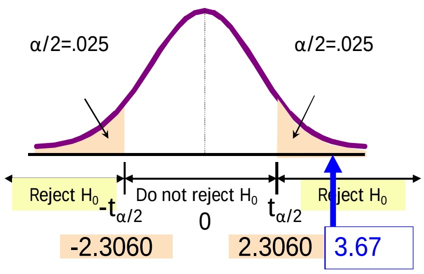
Test Statistic: t = 3.67

d.f. = 10-2 = 8

a =.05

a/2=.025

$t_{\alpha/2}$ = 2.3060 (refer to table)



Decision: Reject H_0

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

www.utm.my innovative • entrepreneurial • global 40

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Exercise 1

Representative data on x = carbonation depth (in millimeters) and y = strength (in mega pascals) for a sample of concrete core specimens taken from a particular building were read from a plot in the article "The Carbonation of Concrete Structures in the Tropical Environment of Singapore" (Magazine of Concrete Research [1996]: 293-300);

Depth, x	8	20	20	30	35	40	50	55	65
Strength, y	22.8	17.1	21.1	16.1	13.4	12.4	11.4	9.7	6.8

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Exercise 1

- Construct a scatterplot. Does the relationship between carbonation depth and strength appear to be linear?
- Find the equation of the least-square line.
- What would you predict for strength when carbonation depth is 25 mm?
- Explain why it would not be reasonable to use the least-square line to predict strength when carbonation depth is 100 mm.

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Exercise 2

The following data on sale, size, and land-to-building ratio for 10 large industrial properties appeared in the paper "Using Multiple Regression Analysis in Real Estate Appraisal" (Appraisal Journal [2002]: 424-430):

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Exercise 2

Property	Sale Price (millions of dollars)	Size (thousands of sq. ft.)	Land-to-Building Ratio
1	10.6	2166	2.0
2	2.6	751	3.5
3	30.5	2422	3.6
4	1.8	224	4.7
5	20.0	3917	1.7
6	8.0	2866	2.3
7	10.0	1698	3.1
8	6.7	1046	4.8
9	5.8	1108	7.6
10	4.5	405	17.2

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Exercise 2

a) Calculate and interpret the value of the correlation coefficient between sale price and size.
b) Calculate and interpret the value of the correlation co-efficient between sale price and land-to-building ratio.
c) If you wanted to predict sale price and you could use either size or land-to-building ratio as the basis for making predictions, which would you use? Explain.
d) Based on your choice in Part (c), find the equation of the least-square regression line you would use for predicting *y* = sale price.