Chapter 10

Correlation and Regression

1

Can temp. predict crime?

ORIGINAL RESEARCH

Temperature and Violent Crime in Dallas, Texas: Relationships and Implications of Climate Change

Janet L. Gamble, PhD* Jeremy J. Hess, MD, MPH†

* United States Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment, Washington, DC
† Emory University School of Medicine, Department of Emergency Medicine and Department of Environmental Health, Emory Rollins School of Public Health, Atlanta, Georgia

Supervising Section Editor: Abigail Hankin, MD
Submission history: Submitted January 12, 2012; Revision received March 13, 2012; Accepted March 14, 2012
Reprints available through open access at http://www. http://escholarship.org/uc/ucem_westjem
DOI: 10.5811/westjem.2012.3.11748

Month	Average temperature	Total offenses
January	36	83
February	35	82
March	42	81
April	52	102
May	60.5	122
June	71.5	117
July	77	126
August	77.5	115
September	73	84
October	63	123
November	53	82
December	45	102

2

Do Dust Storms Affect Respiratory Health?

Southeast Washington state has a long history of seasonal dust storms. Several researchers decided to see what effect, if any, these storms had on the respiratory health of the people living in the area. They undertook (among other things) to see if there was a relationship between the amount of dust and sand particles in the air when the storms occur and the number of hospital emergency room visits for respiratory disorders at three community hospitals in southeast Washington. Using methods of correlation and regression, which are explained in this chapter, they were able to determine the effect of these dust storms on local residents. See Statistics Today—Revisited at the end of the chapter.

Source: B. Hefflin, B. Jalaludin, N. Cobb, C. Johnson, L. Jecha, and R. Etzel, "Surveillance for Dust Storms and Respiratory Diseases in Washington State, 1991," *Archives of Environmental Health* 49, no. 3 (May–June 1994), pp. 170–74. Reprinted with permission of the Helen Dwight Reid Education Foundation. Published by Heldref Publications, 1319 18th St. N.W., Washington, D.C. 20036-1802. Copyright 1994.

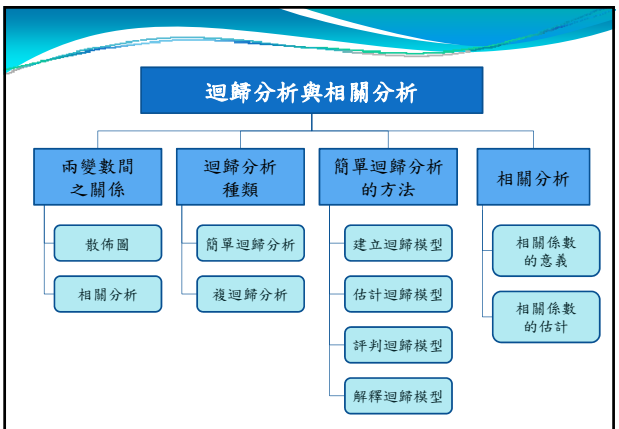
3

Chapter 10 Overview

Introduction

- 10-1 Scatter Plots and Correlation
- 10-2 Regression
- 10-3 Coefficient of Determination and Standard Error of the Estimate
- 10-4 Multiple Regression (Optional)

4



5

Chapter 10 Objectives

1. Draw a scatter plot for a set of ordered pairs.
2. Compute the correlation coefficient.
3. Test the hypothesis H_0: $\rho = 0$.
4. Compute the equation of the regression line.
5. Compute the coefficient of determination.
6. Compute the standard error of the estimate.
7. Find a prediction interval.
8. Be familiar with the concept of multiple regression.

6

Introduction

- In addition to hypothesis testing and confidence intervals, inferential statistics involves determining whether a relationship between two or more numerical or quantitative variables exists.
- **Correlation** is a statistical method used to determine whether a linear relationship between variables exists.
- **Regression** is a statistical method used to describe the nature of the relationship between variables—that is, positive or negative, linear or nonlinear.

7

Introduction

- The purpose of this chapter is to answer these questions statistically:
  1. Are two or more variables related?
  2. If so, what is the strength of the relationship?
  3. What type of relationship exists?
  4. What kind of predictions can be made from the relationship?

8

Introduction

1. Are two or more variables related?

2. If so, what is the strength of the relationship?

To answer these two questions, statisticians use the **correlation coefficient**, a numerical measure to determine whether two or more variables are related and to determine the strength of the relationship between or among the variables.

9

Introduction

3. What type of relationship exists?

There are two types of relationships: simple and multiple.

In a simple relationship, there are two variables: an **independent variable** (predictor variable) and a **dependent variable** (response variable).

In a multiple relationship, there are two or more independent variables that are used to predict one dependent variable.

10

Introduction

4. What kind of predictions can be made from the relationship?

Predictions are made in all areas and daily. Examples include weather forecasting, stock market analyses, sales predictions, crop predictions, gasoline price predictions, and sports predictions. Some predictions are more accurate than others, due to the strength of the relationship. That is, the stronger the relationship is between variables, the more accurate the prediction is.

11

Section 10-1

Scatter Plots and Correlation

12

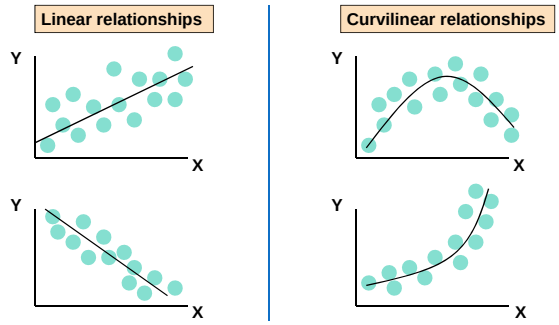
Example: concept of the relationship

- Whether there is a relationship between number of hours of study and test scores on an exam.
- Collect data:
  - two numerical or quantitative variables
  - ➔ To test whether a relationship exists between the variables

Student	Hours of study x	Grade y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

13

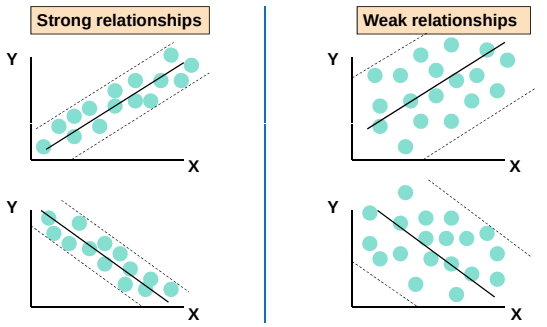
Types of Relationships



Chap 13-14

Chap 13-14

Types of Relationships (continued)



15

10.1 Scatter Plots and Correlation

- A **scatter plot** is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x and the dependent variable y.

- Procedure of drawing a scatter plot

Step 1: Draw and label the x and y axes.

Step 2: Plot each point on the graph.

Step 3: Determine the type of relationship (if any) that exits for the variables.

16

Example 10-1: Car Rental Companies

Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

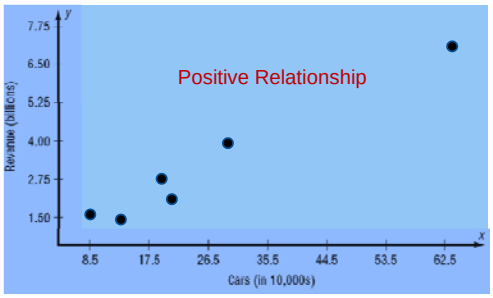
Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Step 1: Draw and label the x and y axes.

Step 2: Plot each point on the graph.

17

Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5



18

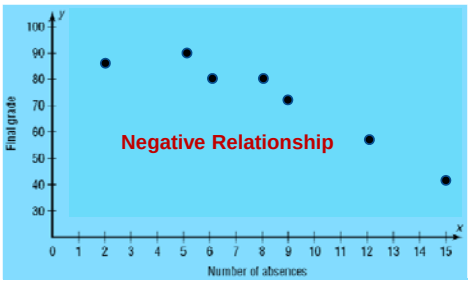
Example 10-2: Absences/Final Grades

Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.

Student	Number of absences x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

Step 1: Draw and label the *x* and *y* axes.
Step 2: Plot each point on the graph.

Example 10-2: Absences/Final Grades



Example 10-3: Age and Wealth

A researcher wishes to see if there is a relationship between the ages of the wealthiest people in the world and their net worth. The data shows a random sample of 10 persons selected from the *Forbes* list of the 400 richest people for a recent year.

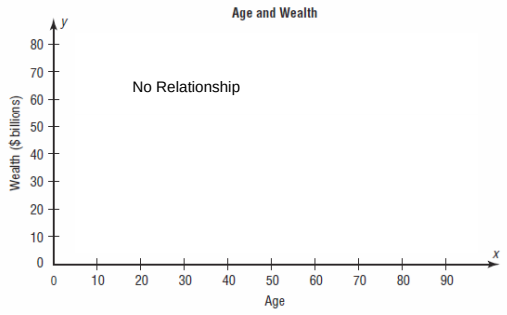
Step 1: Draw and label the *x* and *y* axes.

Step 2: Plot each point on the graph.

Person	Age x	Net worth y (in billions of dollars)
A	60	11
B	72	69
C	56	11.9
D	55	30
E	83	12.2
F	67	36
G	38	18.7
H	62	10.2
I	62	23.3
J	46	10.6

Source: *Forbes* magazine.

Example 10-3: Age and Wealth

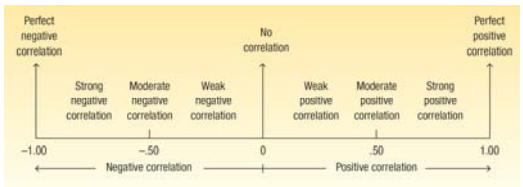


Correlation

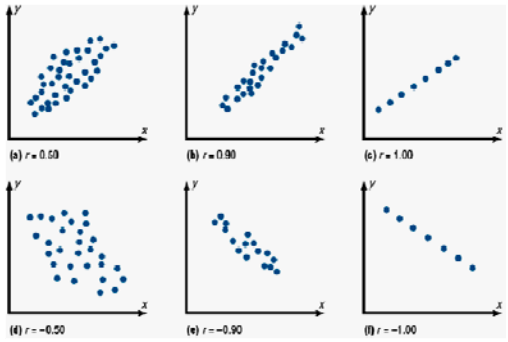
- The **correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two variables.
- There are several types of correlation coefficients. The one explained in this section is called the **Pearson product moment correlation coefficient (PPMC)**.
- The symbol for the sample correlation coefficient is *r*. The symbol for the population correlation coefficient is ρ.

Correlation

- The range of the correlation coefficient is from –1 to +1.
- If there is a **strong positive linear relationship** between the variables, the value of *r* will be close to +1.
- If there is a **strong negative linear relationship** between the variables, the value of *r* will be close to –1.



Correlation



簡單迴歸分析的方法

概念與思考 **同一數據，不同視覺。**

繪製散佈圖（或觀察散佈圖）可以看出兩個變數之間的關係。

但是散佈圖會因縱軸或橫軸的數值間距的不同，以致於看起來（目測法）兩變數間的關係會有極大的差異。

簡單迴歸分析的方法

圖14.8 數值軸間距較小的散佈圖

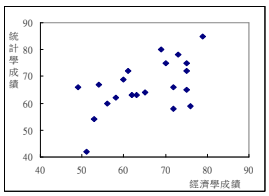
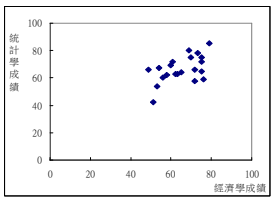


圖14.9 數值軸間距較大的散佈圖



Correlation Coefficient

The formula for the correlation coefficient is

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

where n is the number of data pairs.

Rounding Rule: Round to three decimal places.

Example 10-4: Car Rental Companies

Compute the correlation coefficient for the data in Example 10–1.

Company	Cars x (in 10,000s)	Income y (in billions)	xy	x^2	y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	2.75	72.25	2.25
	$\Sigma x =$ 153.8	$\Sigma y =$ 18.7	$\Sigma xy =$ 682.77	$\Sigma x^2 =$ 5859.26	$\Sigma y^2 =$ 80.67

Example 10-4: Car Rental Companies

Compute the correlation coefficient for the data in Example 10–1.

$\Sigma x = 153.8$, $\Sigma y = 18.7$, $\Sigma xy = 682.77$, $\Sigma x^2 = 5859.26$, $\Sigma y^2 = 80.67$, $n = 6$

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

Example 10-5: Absences / Final Grades

Compute the correlation coefficient for the data in Example 10–2.

Student	Number of absences, x	Final Grade y (pct.)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x =$ 57	$\Sigma y =$ 511	$\Sigma xy =$ 3745	$\Sigma x^2 =$ 579	$\Sigma y^2 =$ 38,993

Example 10-5: Absences / Final Grades

Compute the correlation coefficient for the data in Example 10–2.

$\Sigma x = 57$, $\Sigma y = 511$, $\Sigma xy = 3745$, $\Sigma x^2 = 579$, $\Sigma y^2 = 38{,}993$, $n = 7$

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

Example 10-6: Age and Wealth

Compute the value of the correlation coefficient for the data given in Example 10–3 for the age and wealth of the richest persons in the United States.

Person	Age x	Net wealth y	xy	x^2	y^2
A	60	11	660	3,600	121
B	77	69	4,966	5,184	4,761
C	56	11.9	666.4	3,136	141.61
D	55	30	1,650	3,025	900
E	83	12.2	1,012.6	6,889	148.84
F	67	36	2,412	4,489	1,296
G	38	18.7	710.6	1,444	349.69
H	62	10.2	632.4	3,844	104.04
I	62	23.3	1,444.6	3,844	542.89
J	46	10.6	487.6	2,116	112.36
	$\Sigma x = 601$	$\Sigma y = 232.9$	$\Sigma xy = 14{,}644.2$	$\Sigma x^2 = 37{,}571$	$\Sigma y^2 = 8{,}477.43$

Example 10-6: Age and Wealth

$\sum x = 601$, $\sum y = 232.9$, $\sum xy = 14{,}644.2$,

$\sum x^2 = 37571$, $\sum y^2 = 8477.43$

Exercise: Alumni Contributions

- The director of an alumni association for a small college wants to determine whether there is any type of relationship between the amount of an alumnus's contribution (in dollars) and the years the alumnus has been out of school.

Years x	1	5	3	10	7	6
Contribution y	500	100	300	50	75	80

a. Draw the scatter plot for the variables.

b. Compute the value of the correlation coefficient.

The Assumptions

The assumptions for testing the significance of the Linear Correlation Coefficient

1. The data are **quantitative** and are obtained from a simple **random sample**.
2. The **scatter plot** shows that the data are approximately **linear related**.
3. There are **no outlier**s in the data.
4. The variables x and y must come from **normally distributed populations**.

Test the significance of the correlation coefficient

- Three common used methods:
  - The ***t-test*** method
  - The ***p-value*** method
  - Critical method of using the **PPMC**, refer to Table I

37

The Significance of the Correlation Coefficient

- **Hypothesis-testing procedure**
  - Step 1: State the hypotheses.

H_0: $\rho = 0$	This null hypothesis means that there is no correlation between the x and y variables in the population.
H_1: $\rho \neq 0$	This alternative hypothesis means that there is a significant correlation between the variables in the population.

  - Step 2: Find the critical values.
    - **T-test, df=n-2**
  - Step 3: Compute the test value.
    - **Test-value,** $t = r\sqrt{\frac{n-2}{1-r^2}}$
  - Step 4: Make the decision.
  - Step 5: Summarize the results.

38

Example 10-7: Car Rental Companies

Test the significance of the correlation coefficient found in Example 10–4. Use $\alpha = 0.05$ and $r = 0.982$.

Step 1: State the hypotheses.

Step 2: Find the critical value.

39

Example 10-7: Car Rental Companies

Step 3: Compute the test value.

Step 4: Make the decision.

Step 5: Summarize the results.

40

Test the significance of the correlation coefficient

- Three common used methods:
  - The *t-test* method
  - The *p-value* method
  - Critical method of using the **PPMC**, refer to Table I
    - This table shows the values of the correlation coefficient that are significant for a specific α level and a specific number of degrees of freedom.

41

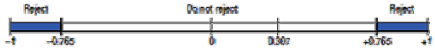
Example 10-8: Car Rental Companies

Using Table I, test the significance at α = 0.01 of the correlation coefficient r = 0.307, obtained in Example 10–6.

Step 1: State the hypotheses. **Step 2: Find the critical value.**

For a significant relationship, r must be greater than ( ) or less than ( ). Since $r = 0.307$, do not reject the null.

Hence, there is not enough evidence to say that there is a significant linear relationship between the variables.



42

Possible Relationships Between Variables

When the null hypothesis has been rejected for a specific α value, any of the following five possibilities can exist.

1. There is a *direct cause-and-effect* relationship between the variables. That is, *x* causes *y*.
2. There is a *reverse cause-and-effect* relationship between the variables. That is, *y* causes *x*.
3. The relationship between the variables may be *caused by a third variable*.
4. There may be a *complexity of interrelationships* among many variables.
5. The relationship may be *coincidental*.

Exercise: Alumni Contributions

- The director of an alumni association for a small college wants to determine whether there is any type of relationship between the amount of an alumnus's contribution (in dollars) and the years the alumnus has been out of school.

Years x	1	5	3	10	7	6
Contribution y	500	100	300	50	75	80

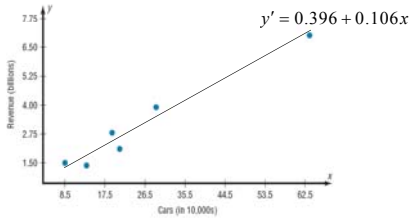
a. Draw the scatter plot for the variables.
b. Compute the value of the correlation coefficient.
c. State the hypotheses.
d. Test the significance of the correlation coefficient at $\alpha = 0.05$, using t-test and Table I.
e. Give a brief explanation *of the type of relationship.*

Section 10-2

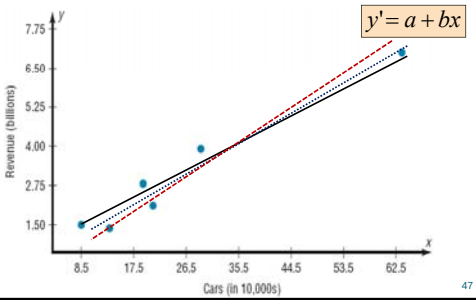
Regression

Regression

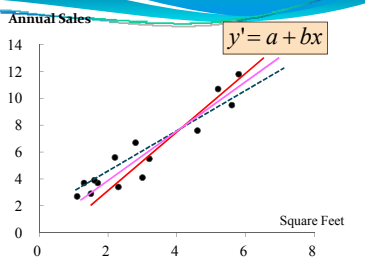
- If the value of the correlation coefficient is significant, the next step is to determine the equation of the **regression line** which is the data's line of best fit.



Company	Cars (in 10,000s)	Income (in billions)
A	63.0	7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

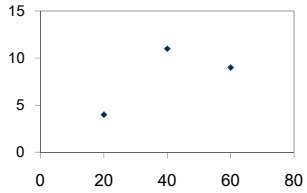


	Square Feet	Annual Sales
1	1.7	3.7
2	1.6	3.9
3	2.8	6.7
4	5.6	9.5
5	1.3	3.4
6	2.2	5.6
7	1.3	3.7
8	1.1	2.7
9	3.2	5.5
10	1.5	2.9
11	5.2	10.7
12	4.6	7.6
13	5.8	11.8
14	3.0	4.1



Least-Squares Method

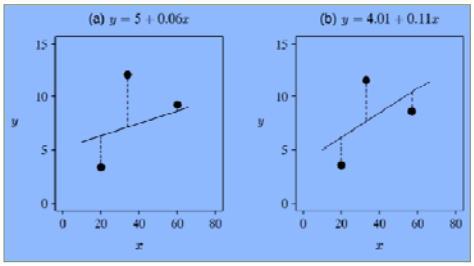
x	y
20	4
40	11
60	9



Two straight line:
(a) $y = 5 + 0.06x$
(b) $y = 4.01 + 0.11x$

Which Line is the best fit?

Least-Squares Method



Least-Squares Method

a. $y = 5 + 0.06x$

Sum square of error:

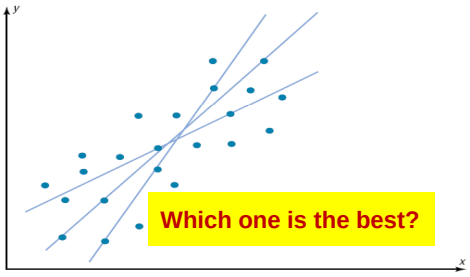
x	y
20	4
40	11
60	9

b. $y = 4.01 + 0.11x$

Sum square of error:

➔ The straight line with equation of _____ is the ***best fit*** to the data.

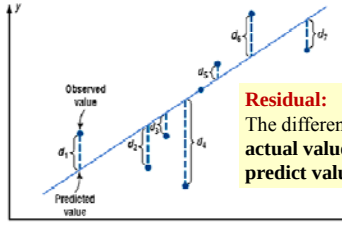
Regression



Which one is the best?

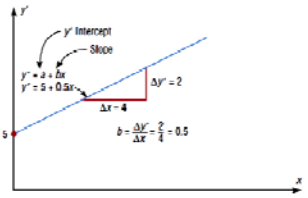
Regression

- **Best fit** means that the sum of the squares of the vertical distance from each point to the line is at a minimum.



Residual:
The difference between the **actual value *y*** and the **predict value *y'*.**

- $E(y_i) = a + b\,x_i$; $i=1,2,.$
- $Var(y_i) = \sigma^2$ 或 $Sd(y_i)=\sigma$
- $y_i \sim N(a + b\,x_i, \sigma)$



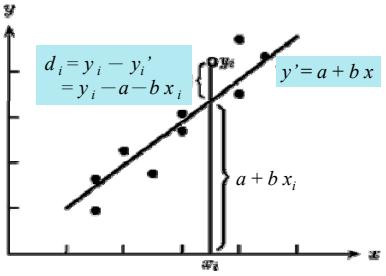
Constant term or y-intercept — Slope

$$y_i = a + b\,x_i$$

Dependent variable — Independent variable

Estimation of regression coefficients *a*, *b*

Least Square Estimation, LSE

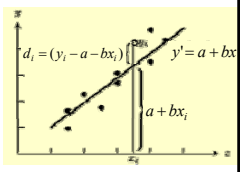


觀察值與直線 $y' = a + b x$ 之離差

55

Least Square Estimation, LSE

- Data : $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Sum of square error, distance



Min. $D = SSE = \sum_{i=1}^{n} (y_i - a - bx_i)^2$

$$\begin{cases} \frac{\partial D}{\partial a} = \sum_{i=1}^{n} -2(y_i - a - bx_i) = 0 \\ \frac{\partial D}{\partial b} = \sum_{i=1}^{n} -2x_i(y_i - a - bx_i) = 0 \end{cases}$$

56

Regression Line $y' = a + bx$

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

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

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

57

$$\sum \hat{e}_i = \sum (y_i - \hat{y}_i)$$

殘差

$$= \sum (y_i - a - b x_i) = 0$$

$$\sum \hat{e}_i^2 = \sum (y_i - \hat{y}_i)^2 = \text{最小}$$

殘差平方和 or 誤差平方和, sum of square due to error, SSE

58

Example 10-9: Car Rental Companies

Find the equation of the regression line for the data in Example 10–4, and graph the line on the scatter plot.

Company	Cars *x* (in 10,000s)	Income *y* (in billions)
A	63.0	7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

59

Example 10-9: Car Rental Companies

Find the equation of the regression line for the data in Example 10–4, and graph the line on the scatter plot. ($n = 6$)

$\Sigma x = 153.8$, $\Sigma y = 18.7$, $\Sigma xy = 682.77$, $\Sigma x^2 = 5859.26$, $\Sigma y^2 = 80.67$

Solution:

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

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

$y' = a + bx \quad \rightarrow \quad y' =$ _____

60

Example 10-9: Car Rental Companies

Find two points to sketch the graph of the regression line.

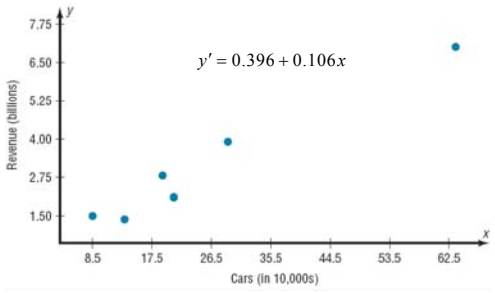
Use any x values between 10 and 60. For example, let x equal 15 and 40. Substitute in the equation and find the corresponding y value.

➔ **Plot (15,_____) and (40, _____), and sketch the resulting line.**

61

Example 10-9: Car Rental Companies

Find the equation of the regression line for the data in Example 10–4, and graph the line on the scatter plot.



62

Example 10-10: Absences and Final Grades

Find the equation of the regression line for the data in Example 10–5, and graph the line on the scatter plot. ($n = 6$)

$\Sigma x = 153.8$, $\Sigma y = 18.7$, $\Sigma xy = 682.77$, $\Sigma x^2 = 5859.26$, $\Sigma y^2 = 80.67$,

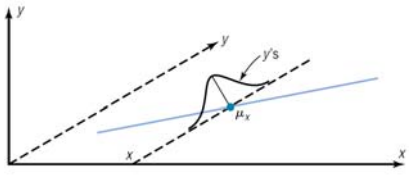
63

Source: Reprinted with special permission of King Features Syndicate.

64

Assumptions for Valid Predictions

1. The sample is a random sample.
2. For any specific value of the independent variable x, the value of the dependent variable y must be normally distributed about the regression line.

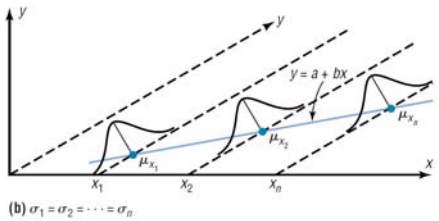


(a) Dependent variable y normally distributed

65

Assumptions for Valid Predictions

3. The standard deviation of each of the dependent variables must be the same for each value of the independent variable.



(b) $\sigma_1 = \sigma_2 = \cdots = \sigma_n$

66

Example 10-11: Car Rental Companies

Use the equation of the regression line to predict the income of a car rental agency that has 200,000 automobiles.

Linear equation: $y' = 0.396 + 0.106x$

$x = 20$ corresponds to 200,000 automobiles.

$$y' = 0.396 + 0.106x$$
$$= 0.396 + 0.106(20)$$
$$= 2.516$$

Hence, when a rental agency has 200,000 automobiles, its revenue will be approximately \$2.516 billion.

Regression

- The magnitude of the change in one variable when the other variable changes exactly 1 unit is called a **marginal change.** The value of slope *b* of the regression line equation represents the marginal change.
- For valid predictions, the value of the correlation coefficient must be significant.
- When *r* is significantly different from 0, the best predictor of *y* is the mean of the data values of *y*.

Extrapolations (Future Predictions)

- **Extrapolation**, or making predictions beyond the bounds of the data, must be interpreted cautiously.
- Remember that when predictions are made, they are based on present conditions or on the premise that present trends will continue. This assumption may or may not prove true in the future.

Finding the Correlation Coefficient and the Regression Line Equation

Step 1 Make a table, as shown in step 2.

Step 2 Find the values of xy, x^2, and y^2. Place them in the appropriate columns and sum each column.

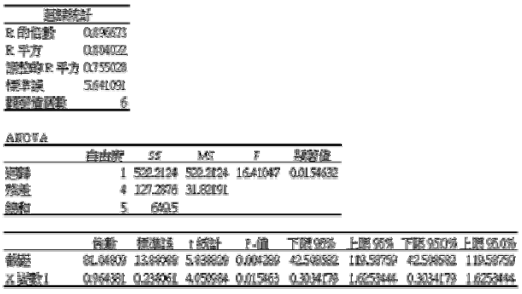
x	y	xy	x^2	y^2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$\Sigma x =$	$\Sigma y =$	$\Sigma xy =$	$\Sigma x^2 =$	$\Sigma y^2 =$

Step 3 Substitute in the formula to find the value of r.

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

Step 4 When r is significant, substitute in the formulas to find the values of a and b for the regression line equation $y' = a + bx$.

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \bar{y} - b\bar{x} \qquad b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$



Exercise: Alumni Contributions

◆The director of an alumni association for a small college wants to determine whether there is any type of relationship between the amount of an alumnus's contribution (in dollars) and the years the alumnus has been out of school.

Years x	1	5	3	10	7	6
Contribution y	500	100	300	50	75	80

a. Find the equation of the regression line.

b. Find the y' value when $x = 4$ years.

Section 10-3

Coefficient of Determination and Standard Error of the Estimate

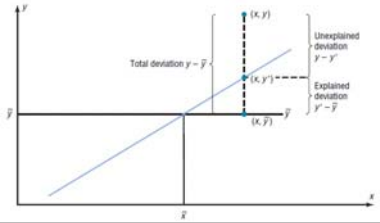
x	1	2	3	4	5
y	10	8	12	16	20

The regression line:
$y' = 0.396 + 0.106x$,
$r = 0.919$

➔ The predicted value, when $x = 1$

$y' = 4.8 + 2.8x = 4.8 + (2.8)(1) = 7.6$

➔ The observed value, when $x = 1$, $y = 10$



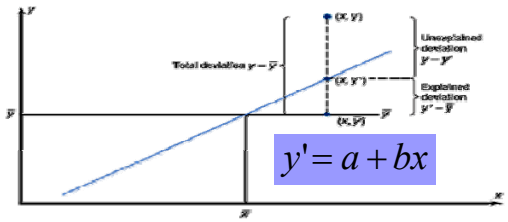
Total variation

- The **total variation** $\sum (y - \bar{y})^2$ is the sum of the squares of the vertical distances each point is from the mean.
- The total variation can be divided into two parts:
  - ***Explained variation***
  - ***Unexplained variation***

Variation

- **Explained variation**:
  - The variation obtained from the relationship (i.e., from the predicted y' values)
  - Found by: $\sum (y' - \bar{y})^2$
- **Unexplained variation**
  - This variation cannot be attributed to the relationships.
  - Variation due to chance
  - Found by: $\sum (y' - y)^2$

Variation



$y' = a + bx$

Total variation = explained + unexplained variation

$$\sum (y - \bar{y})^2 = \sum (y' - \bar{y})^2 + \sum (y - y')^2$$

residual

◆The procedure for finding the three types of variation

x	1	2	3	4	5
y	10	8	12	16	20

x	y	y'
1	10	7.6
2	8	10.4
3	12	13.2
4	16	16.0
5	20	18.8

◆The procedure for finding the three types of variation

x	1	2	3	4	5
y	10	8	12	16	20

Step 2 Find the mean of the y values.

Step 3 Find the total variation $\Sigma(y - \bar{y})^2$.

◆The procedure for finding the three types of variation

x	y	y'
1	10	7.6
2	8	10.4
3	12	13.2
4	16	16.0
5	20	18.8

Step 4 Find the explained variation $\Sigma(y' - \bar{y})^2$.

Step 5 Find the unexplained variation $\Sigma(y - y')^2$.

Total variation = Explained variation + Unexplained variation

ANOVA table

Source	Sum of squares (SS)	d.f.	Mean square (MS)	F-value (test value)
Regression	$\sum(y'-\bar{y})^2 = SSR$ (Explained variation)	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	$\sum(y_i - y')^2 = SSE$ (Unexplained variation)	$n-k-1$	$MSE = \frac{SSE}{n-k-1}$	
Total	$\sum(y_i - \bar{y})^2 = SST$ (Total variation)	$n-1$		

Simple linear regression: $k = 1$

Critical value: $F_{(k,\, n-k-1)}$

ANOVA

Source	Sum of squares (SS)	d.f.	Mean square (MS)	F-value (test value)
Regression				
Error				
Total				

Residual Plots

Residual: prediction errors, $(y\text{-}y')$

Coefficient of Determination

- The **coefficient of determination** is the ratio of the explained variation to the total variation.
- The symbol for the coefficient of determination is r^2.
- Another way to arrive at the value for r^2 is to square the correlation coefficient.

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

85

Coefficient of Nondetermiation

- The **coefficient of nondetermination** is a measure of the unexplained variation.
- The formula for the coefficient of nondetermination is $1.00 - r^2$.

86

Standard Error of the Estimate

- The **standard error of estimate**, denoted by s_{est} is the standard deviation of the observed y values about the predicted y' values. The formula for the standard error of estimate is:

$$s_{est} = \sqrt{\frac{\sum (y - y')^2}{n - 2}}$$

87

Example 10-12: Copy Machine Costs

A researcher collects the following data and determines that there is a significant relationship between the age of a copy machine and its monthly maintenance cost. The regression equation is $y' = 55.57 + 8.13x$. Find the standard error of the estimate.

Machine	Age x (years)	Monthly cost y
A	1	$ 62
B	2	78
C	3	70
D	4	90
E	4	93
F	6	103

88

Example 10-12: Copy Machine Costs

Machine	Age x (years)	Monthly cost, y	y'	$y - y'$	$(y - y')^2$
A	1	62	63.70	-1.70	2.89
B	2	78	71.83	6.17	38.0689
C	3	70	79.96	-9.96	99.2016
D	4	90	88.09	1.91	3.6481
E	4	93	88.09	4.91	24.1081
F	6	103	104.35	-1.35	1.8225
					169.7392

$y' = 55.57 + 8.13x$

$y' = 55.57 + 8.13(1) = 63.70$

$y' = 55.57 + 8.13(2) = 71.83$

$y' = 55.57 + 8.13(3) = 79.96$

$y' = 55.57 + 8.13(4) = 88.09$

$y' = 55.57 + 8.13(6) = 104.35$

$$s_{est} = \sqrt{\frac{\sum (y - y')^2}{n - 2}}$$

$$s_{est} = \sqrt{\frac{169.7392}{4}} = 6.51$$

89

Example 10-13: Copy Machine Costs

$$s_{est} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n - 2}}$$

$$= \sqrt{\frac{\text{unexplained deviation}}{n - 2}}$$

$$= \sqrt{\frac{\sum (y - y')^2}{n - 2}} = \sqrt{MSE}$$

90

Example 10-13: Copy Machine Costs

Machine	Age x (years)	Monthly cost, y	xy	y^2
A	1	62	62	3,844
B	2	78	156	6,084
C	3	70	210	4,900
D	4	90	360	8,100
E	4	93	372	8,649
F	6	103	618	10,609
		496	1778	42,186

$$s_{est} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}}$$

$$s_{est} = \sqrt{\frac{42{,}186 - 55.57(496) - 8.13(1778)}{4}} = 6.48$$

例題

在研發新的治療頭疼症狀的藥劑研究中，研究者對不同的藥劑服用量 X（毫克）能解除頭疼症狀的藥效持續時間 Y（小時）常感到興趣。下表即是對八名接受實驗者給予不同的藥劑服用量，並記載服用藥劑所能持續的時間所獲得之資料。

- 試求出樣本迴歸直線及誤差項變異數σ^2的估計值。
- 試求判定係數，並建立變異數分析表及相關檢定。

服用量	持續時間
4	20
6	40
3	11
5	30
2	9
2	12
3	15
3	21

解：

服用量	持續時間
4	20
6	40
3	11
5	30
2	9
2	12
3	15
3	21

i	x_i	y_i	x_i^2	y_i^2	x_iy_i
1	4	20	16	400	80
2	6	40	36	1,600	240
3	3	11	9	121	33
4	5	30	25	900	150
5	2	9	4	81	18
6	2	12	4	144	24
7	3	15	9	225	45
8	3	21	9	441	63
合計	28	158	112	3,912	653

$$b = \frac{8(653) - (28)(158)}{8(112) - (28)^2} = 7.1429$$

$$a = \frac{158}{8} - (7.1429)\left(\frac{28}{8}\right) = -5.2502$$

因此，樣本迴歸直線為

$$y' = -5.2502 + 7.1429x$$

$$\sum_{i=1}^{8}(x_i - \bar{x})^2 = 112 - 8\left(\frac{28}{8}\right)^2 = 14$$

$$\sum_{i=1}^{8}(y_i - \bar{y})^2 = 3912 - 8\left(\frac{158}{8}\right)^2 = 791.5$$

➔ $SSE = 791.5 - (7.1429)^2(14) = 77.2057$

故σ^2的估計值為

$$\hat{\sigma}^2 = \frac{77.2057}{8-2} = 12.8676$$

$$SST = 3{,}912 - 8\left(\frac{158}{8}\right)^2 = 791.5$$

$$SSE = 77.2057$$

$$SSR = SST - SSE = 791.5 - 77.2057 = 714.2943$$

因此，判定係數 $R^2 = \frac{714.2943}{791.5} = 0.9025$

樣本迴歸直線已解釋了總變異的90.25%。

變異來源	平方和 (SS)	自由度 ($d.f.$)	均方 (MS)	F值
迴歸	714.2943	1	714.2943	55.5111
誤差	77.2057	6	12.8676	
總和	791.5	7		

在顯著水準 $\alpha = 0.05$，$F_{0.05}(1, 6) = 5.99$，故拒絕虛無假設 H_0，即藥劑服用量對解除頭疼症狀所能持續的時間有影響。

Regression analysis of MINITAB

```
Regression Analysis: y versus x

The regression equation is
(1) y = - 5.25 + 7.14 x

Predictor      Coef  SE Coef      T      P
Constant     -5.250    3.587  -1.46  0.194
x       (2)7.1429   0.9588   7.45  0.000

(3)S = 3.58735   R-Sq = 90.2%   R-Sq(adj) = 88.6%

(4)Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  714.29  714.29  55.50  0.000
Residual Error    6   77.21   12.87
Total             7  791.50
```

Formula for the Prediction Interval about a Value y'

$$y' - t_{\alpha/2} s_{est} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum (x - \bar{X})^2}} < y <$$

$$y' + t_{\alpha/2} s_{est} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum (x - \bar{X})^2}}$$

Example 10-14: Copy Machine Costs

For the data in Example 10–12, find the 95% prediction interval for the monthly maintenance cost of a machine that is 3 years old.

Step 1: Find $\sum x, \sum x^2$, and $\bar{X}$.

$$\sum x = 20 \quad \sum x^2 = 82 \quad \bar{X} = \frac{20}{6} = 3.3$$

Step 2: Find y' *for* $x = 3$.

$$y' = 55.57 + 8.13(3) = 79.96$$

Step 3: Find s_{est}.

$s_{est} = 6.48$ (as shown in Example 10-13)

Example 10-14: Copy Machine Costs

Step 4: Substitute in the formula and solve.

$$y' - t_{\alpha/2} s_{est} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum (x - \bar{X})^2}} < y < y' + t_{\alpha/2} s_{est} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum (x - \bar{X})^2}}$$

$$79.96 \pm 2.776 \times 6.48 \times \sqrt{1 + \frac{1}{6} + \frac{(3 - 3.3)^2}{82 - (20^2/6)}}$$

$$79.96 - 19.43 < y < 79.96 + 19.43$$

$$60.53 < y < 99.39$$

Hence, you can be 95% confident that the interval $60.53 < y < 99.39$ contains the actual value of y.

Exercise:

相關的意義

(1)依變數個數區分
- 簡單相關：僅探討二個變數。
- 複相關：探討三個或以上的變數。

(2)依線性或非線性區分
- 線性相關：兩變數間之相關可用一直線方程式做適當表示者。
- 非線性相關：無法用直線來描述者，亦稱為曲線相關。

(3)依相關程度區分
- 完全相關：兩變數間的相關可用一直線或曲線完全表示者。
- 零相關：兩變數間不具任何關係者。
- 非完全相關：介於完全相關與零相關之間。

(4)依相關方向區分
- 正相關：兩變數之間的相關為兩變數同時增加或同時減少。
- 負相關：變數間此增故減或此減故增。

例題

X	6	4	10	2	8
Y	7	1	9	3	5

請檢定是否存在迴歸直線 $y = a + bx$，且 $b \neq 0$。
請建立迴歸之變異數分析表進行檢定，並說明其涵義。

解： (a) ① 假設：$H_0: b = 0; H_1: b \neq 0$

② 顯著水準：$\alpha = 0.05$

③ 放棄域：$F > F_{(0.95,1,3)} = 10.128$

④ 計算：

$\sum X = 30, \sum Y = 25, \sum X^2 = 220, \sum Y^2 = 165,$

$\sum XY = 182, n = 5, \overline{X} = 6, \overline{Y} = 5$

$\text{SST} = \sum (Y - \overline{Y})^2 = \sum Y^2 - n\overline{Y}^2 = 165 - (5)(5)^2 = 40$

$\text{MSE} = \hat{\sigma}^2 = 4.8$（詳見 11-5 之例題 11-10 的計算）

$\therefore \quad \text{SSE} = (5-2)(4.8) = 14.4$

$\text{SSR} = \text{SST} - \text{SSE} = 40 - 14.4 = 25.6$

或 $\text{SSR} = b^2 \sum (X - \overline{X})^2$

$= b^2 [\sum X^2 - n\overline{X}^2]$

$= (0.8)^2 (40)$

$= 25.6$

⑤ 迴歸分析之變異數分析表：

變異來源	SS	f	MS	F
迴　歸	25.6	1	25.6	$F = \frac{25.6}{4.8} = 5.33$
誤　差	14.4	3	4.8	
總　和	40	4		

⑥ 判斷：$\because 5.33 < 10.128$，落在接受域，差異不顯著，接受 H_0，即 β 可能爲 0，表示母體迴歸直線可能與橫軸平行，自變數 X 對依變數 Y 解釋能力低。

例題

X	6	4	10	2	8
Y	7	1	9	3	5

(a) 判定係數並解釋其涵義。

(b) 非判定係數並解釋其涵義。

(c) 相關係數。

解： (a) $r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{25.6}{40} = 0.64 = 64\%$，即依變數 Y 的變異可由自變數 X 的變異解釋 64%。

(b) $1 - r^2 = 1 - 64\% = 36\%$，即依變數 Y 的變異有 36% 無法由自變數 X 所解釋。

(c) $r = \pm\sqrt{r^2} = \pm\sqrt{0.64} = 0.8$ ($\because b = 0.8 > 0, \therefore r$ 取爲正數)。

另外我們亦可直接由相關係數 r 的公式計算

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{(5)(182) - (30)(25)}{\sqrt{5(220) - (30)^2}\sqrt{5(165) - (25)^2}}$$

$$= 0.8$$