Ch 13 線性迴歸與相關分析

Linear Regression
And
Correlation Analysis

1

相關分析

Correlation Analysis

2

Correlation Analysis

- Correlation Analysis is the study of the relationship between two variables.
  - Scatter Diagram（散佈圖）
  - Coefficient of Correlation（相關係數）
  - Coefficient of Determination（判定係數）

3

Scatter Diagram

- A Scatter Diagram is a chart that portrays the relationship between the two variables. It is the usual first step in correlations analysis

4

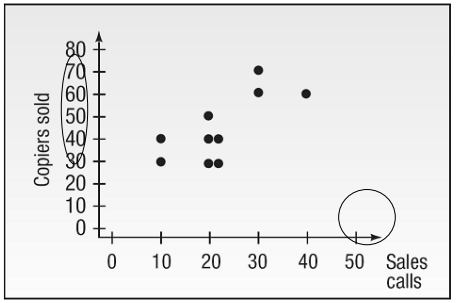
Scatter Diagram Example

The sales manager of Copier Sales of America, which has a large sales force throughout the United States and Canada, wants to determine whether there is a relationship between the number of sales calls made in a month and the number of copiers sold that month. The manager selects a random sample of 10 representatives and determines the number of sales calls each representative made last month and the number of copiers sold.

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

5

Scatter Diagram



6

(母體) 相關係數 (Correlation Coefficient) ρ (1)

- X 與 Y 兩變數的相關方向與相關程度，可以由相關係數 ρ_{XY} 來衡量。

$$\rho_{XY} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

若 X 與 Y 爲 N 個成對資料，則:

$$\rho_{XY} \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu_X)^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i-\mu_Y)^2}},$$

其中 σ_{XY} 稱爲 X 與 Y 的共變異數(Covariance)。

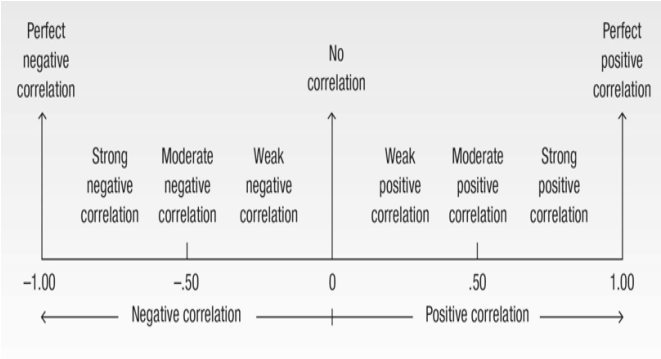
7

(母體) 相關係數 (Correlation Coefficient) ρ (2)

- 相關係數 ρ_{XY} 僅能衡量 X 與 Y 的線性相關程度。
  - $-1 \leqq \rho_{XY} \leqq 1$
  - $\rho_{XY} = +1$ 完全正(線性)相關
  - $\rho_{XY} = -1$ 完全負(線性)相關
  - $\rho_{XY} = 0$ 無(線性)相關
- $\rho_{XY} = 0$表示X 與 Y 的無線性相關，並不表示X 與 Y 無關。若 X 與 Y 的無關， 則 $\rho_{XY} = 0$。

8

Correlation Coefficient - Interpretation



(樣本) 相關係數 r_{XY} (1)

- 母體相關係數 ρ_{XY} 無法得知，因此必須利用樣本估計。

若有 n 組樣本： $(X_1, Y_1)\,(X_2, Y_2), ..., (X_n, Y_n)$ ：

$$r_{XY} = \frac{S_{XY}}{S_X S_Y},$$

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \quad , \quad S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad ,$$

$S_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$ ， 其中 S_{XY} 稱爲 X 與 Y 的樣本共變異數。

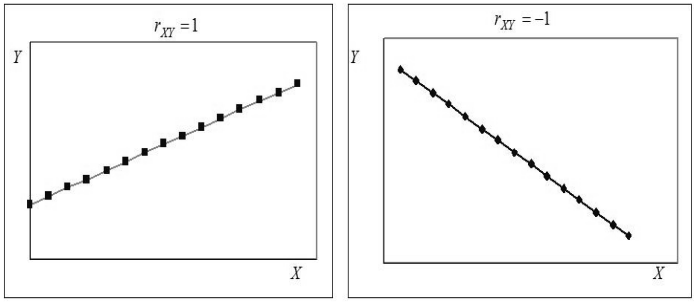
10

(樣本) 相關係數 r_{XY} (2)

- 樣本相關係數 r_{XY} 能估計 X 與 Y 的（母體）相關係數。
  - $-1 \leqq r_{XY} \leqq 1$
  - $r_{XY} = +1$ 完全正(線性)相關
  - $r_{XY} = -1$ 完全負(線性)相關
  - $r_{XY} = 0$ 無(線性)相關

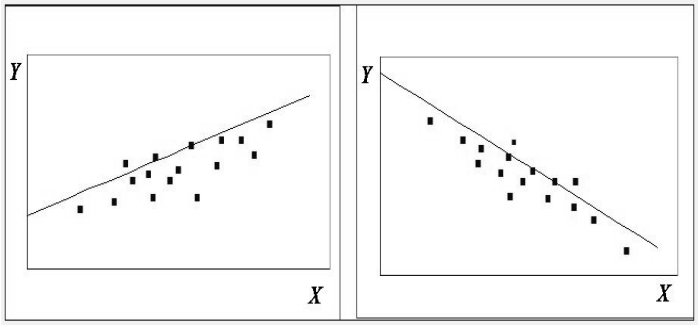
11

r_{XY} 在各種可能的散佈圖



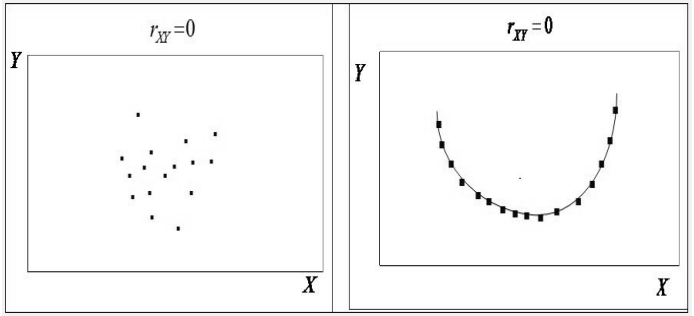
12

r_{XY} 在各種可能的散佈圖



13

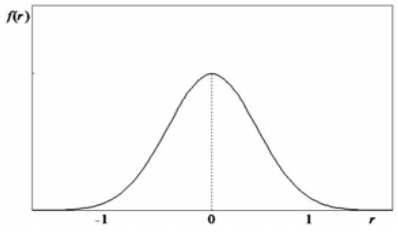
r_{XY} 在各種可能的散佈圖



14

r_{XY} 的抽樣分配

- r_{XY} 的抽樣分配可分兩種情形討論：
  - 當 $\rho_{XY}=0$， r_{XY} 為對稱分配 (0, 1/n-2) 。
  - 當 $\rho_{XY}\neq 0$， r_{XY} 為偏態分配，$E[r_{XY}] \neq \rho_{XY}$ 。



15

檢定 H_0：$\rho_{XY}=0$

$$H_0 : \rho_{XY} = 0$$
$$H_1 : \rho_{XY} \neq 0$$

t 檢定統計量：$\dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}} \sim t_{n-2}$

16

EX13.1
廣告支出與銷售額間的相關係數？

- 大發汽車公司的廣告支出與銷售額資料如下表，求廣告支出與銷售額之間的相關係數。

	A	B	C
1	分公司名稱	廣告支出 X	年銷售收入 Y
2	大通	300	9,500
3	大德	400	10,300
4	大信	500	11,000
5	大道	500	12,000
6	大方	800	12,400
7	大立	1,000	13,400
8	大興	1,000	14,500
9	大展	1,300	15,300

17

EX13.1
廣告支出與銷售額間的相關係數？

表15.3 最小平方法的計算

	X_i	Y_i	$(X_i-\bar{X})$	$(Y_i-\bar{Y})$	$(X_i-\bar{X})^2$	$(Y_i-\bar{Y})^2$	$(X_i-\bar{X})\times(Y_i-\bar{Y})$
	300	9,500	-425	-2,800	180,625	7,840,000	1,190,000
	400	10,300	-325	-2,000	105,625	4,000,000	650,000
	500	11,000	-225	-1,300	50,625	1,690,000	292,500
	500	12,000	-225	-300	50,625	90,000	67,500
	800	12,400	75	100	5,625	10,000	7,500
	1,000	13,400	275	1,100	75,625	1,210,000	302,500
	1,000	14,500	275	2,200	75,625	4,840,000	605,000
	1,300	15,300	575	3,000	330,625	9,000,000	1,725,000
總 合	5,800	98,400	0	0	875,000	28,680,000	4,840,000
平均數	725	12,300	0	0	109,375	3,585,000	605,000

EX13.1
廣告支出與銷售額間的相關係數？

$$r_{XY}=\frac{4,840,000}{\sqrt{875,000}\sqrt{28,680,000}}=0.966$$

- 將所得的樣本相關係數 r_{XY} 作為母體相關係數 ρ_{XY} 的估計值，表示大發汽車公司的汽車銷售額與廣告支出有很大的相關性。

19

EX13.2
檢定廣告支出與銷售額間的關係是否為0

在大發公司的汽車銷售額與廣告支出關係中，若要檢定廣告支出與銷售額間的相關係數是否爲零，此時檢定假設爲：

H_0: $\rho_{XY}=0$

H_1: $\rho_{XY}\neq 0$

計算統計量如下： $t=\dfrac{0.966}{\sqrt{\dfrac{1-(0.966)^2}{6}}}=9.15$

在顯著水準 α=0.05，檢定統計量 t = 9.15 大於臨界值 $t_{6,0.025}=2.447$，故拒絕 H_0, 亦即廣告支出與銷售額有關係。

20

檢定 $H_0: \rho_{XY} = \rho_0 \neq 0$

$H_0 : \rho_{XY} = \rho_0 \neq 0$

$H_1 : \rho_{XY} \neq \rho_0 \neq 0$

檢定統計量：

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \sim N\left(\frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right),\ \frac{1}{n-3}\right)$$

21

EX13.3
看電視的時間(X)與教育年數(Y)的相關

爲調查看電視的時間 (X) 與教育年數 (Y) 的相關，隨機抽取 50 個人，得 $r_{XY} = -0.5$。請檢定下面的假設 (α=0.05)：

H_0: $\rho_{XY} = -0.6$

H_1: $\rho_{XY} \neq -0.6$

22

EX13.3
看電視的時間(X)與教育年數(Y)的相關

因 $Z_r = \frac{1}{2}\ln\frac{1-0.5}{1+0.5} = -0.549$，$Z_\rho = \frac{1}{2}\ln\frac{1-0.6}{1+0.6} = -0.693$

故可得 $\dfrac{Z_r - Z_\rho}{\sqrt{\frac{1}{n-3}}} = \dfrac{-0.549-(-0.693)}{\sqrt{\frac{1}{47}}} = \dfrac{0.144}{0.146}$

$= 0.986 < Z_{0.025} = 1.96$

因此不拒絕 H_0, 即看電視時間與教育年數的相關係數爲 $\rho_{XY} = -0.6$。

23

EX13.3
看電視的時間(X)與教育年數(Y)的相關

若欲求 ρ 之信賴區間，則先求：

$$Z_r - Z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}} \leq Z_\rho \leq Z_r + Z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}}$$

再將 Z_r, Z_ρ代入上式，再求 ρ 之信賴區間。

Z_ρ 的信賴區間爲：

$$0.188 \leq \frac{1+\rho}{1-\rho} \leq 0.591$$

$$\Rightarrow\ -0.684 \leq \rho \leq -0.257$$

24

線性迴歸模型

Linear Regression Model

25

變異數分析 vs. 迴歸模型

- 變異數分析是探討因子（獨立變數）對相依變數是否有影響的統計方法，但有兩點限制：
  - 僅能探討有無影響，無法估計影響程度。
  - 僅能處理離散自變數，無法處理連續自變數。
- 迴歸模型能解決這兩個問題：
  - 肥料用量與產量的關係。
  - 商品價格與需求量的關係。
  - 廣告支出與銷售額的關係。

26

Regression Analysis - Uses

Some examples.

- Is there a relationship between the amount Healthtex spends per month on advertising and its sales in the month?
- Can we base an estimate of the cost to heat a home in January on the number of square feet in the home?
- Is there a relationship between the miles per gallon achieved by large pickup trucks and the size of the engine?
- Is there a relationship between the number of hours that students studied for an exam and the score earned?

27

獨立變數與相依變數

- 獨立變數或稱自變數或解釋變數（ Independent Variable ）

  ---- The Independent Variable provides the basis for estimation. It is the predictor variable.

- 相依變數或稱被解釋變數 (Dependent Variable ）

  ---- The Dependent Variable is the variable being predicted or estimated.

28

迴歸模型

- 迴歸模型是用來分析一個或一個以上的自變數與依變數的數量關係，以瞭解自變數為某一數量時，依變數反應的數量。
- 迴歸模型的主要功能：
  - 瞭解自變數與依變數的關係式及影響方向與程度。
  - 利用自變數與相依變數的關係式，對依變數做預測。

簡單迴歸模型與複迴歸模型

- 依自變數的多寡，迴歸模型可分為：
  - 簡單迴歸模型：只有一個自變數
    - 消售額與廣告支出
    - 家庭支出與所得
  - 複迴歸模型：兩個或兩個以上自變數
    - 銷售額與廣告支出、季節、地區等
    - 庭支出與所得、人口、財富等

簡單線性迴歸模型 (1)

Recall the example involving Copier Sales of America. The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 10 sales representatives. Use the least squares method to determine a linear equation to express the relationship between the two variables.

What is the expected number of copiers sold by a representative who made 20 calls?

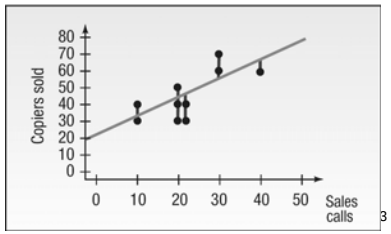
Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

簡單線性迴歸模型 (2)

- 業務經理想以一條直線方程式，描述『撥電話數』與『銷售額』之間的關係 。
  - 自變數（X）：撥電話數
  - 依變數（Y）：銷售額
- Goal: $Y=\alpha+\beta X$, α=?, β=?

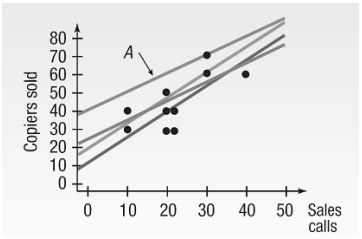
簡單線性迴歸模型 (3)

- 任意決定一條直線，皆會產生誤差。
  - 假設：Y = 20+1.5 X，則 X=10 ➔ Y =35。
    - Welch: (X, Y) = (10, 30) Error = -5
    - Ramirez: (X, Y) = (10, 40) Error = 5
- 以 $Y=\alpha+\beta X+\varepsilon$ 描述迴歸模型，其中 ε 表示誤差。



簡單線性迴歸模型 (4)

- 任意決定一條 $Y=\alpha+\beta X$，皆會產生誤差，因此最適當的估計式，必須使誤差最小。



簡單線性迴歸模型 (5)

- 簡單線性迴歸模型：

$$Y_i=\alpha+\beta X_i+\varepsilon_i$$，i=1,2,…,n，

其中 α 稱為截距 (Intercept)， β 稱為迴歸係數 (Regression Coefficient)， ε_i 為隨機誤差。

- ε 的來源：
  - 人類行為或自然現象的隨機性
  - 測量誤差
  - 其他因素

簡單線性迴歸模型 (6)

- 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, i.e. $E[Y|X]=\alpha+\beta X$.
  - 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.

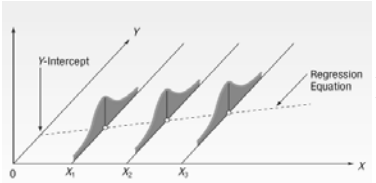
簡單線性迴歸模型 (7)

以上假設可以簡寫成：

$Y_i = \alpha + \beta X_i + \varepsilon_i$，其中 $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

(i.i.d. = identical and independent distribution)

因此，$Y_i \sim N(\alpha + \beta X_i,\ \sigma^2)$ 且 $E[Y_i] = \alpha + \beta X_i$ 。



$E[Y] = \alpha + \beta X$

最小平方法估計 α 與 β (1)

我們分別以 $\hat{\alpha}$ 與 $\hat{\beta}$ 表示 α 與 β 的估計值。

最小平方法是使樣本觀察值與估計值的差異平方和最小，以求取 $\hat{\alpha}$ 與 $\hat{\beta}$ 的方法。

$\underset{\alpha,\ \beta}{\text{Min}}(\text{SSE}) \Rightarrow \hat{\alpha}, \hat{\beta}$，其中

SSE (Sum squares of error) $= \sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2$。

最小平方法估計 α 與 β (2)

$\hat{\alpha}$ 與 $\hat{\beta}$ 可由下列聯立方程式解得：

$$\begin{cases} \dfrac{\partial SSE}{\partial \hat{\alpha}} = 0 \\ \dfrac{\partial SSE}{\partial \hat{\beta}} = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)(-1) = 0 \\ \sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)(-2X_i) = 0 \end{cases}$$

整理可得標準方程式 (Normal Equation) 如下：

$$\begin{cases} \sum Y = n\alpha + \beta \sum X \quad \text{----------} \quad (1) \\ \sum XY = \alpha \sum X + \beta \sum X^2 \text{----------} (2) \end{cases}$$

最小平方法估計 α 與 β (3)

(1)×$(\sum X)$ - (2)×n : 解得 $\hat{\beta} = \dfrac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$ 。

分子分母同除 n，可得 $\hat{\beta} = \dfrac{\sum (X-\bar{X})(Y-\bar{Y})}{\sum (X-\bar{X})^2}$ 。

分子分母同除 n-1，可得 $\hat{\beta} = \dfrac{S_{XY}}{S_X^2}$

最小平方法估計 α 與 β (4)

(1)÷n，可得：$\bar{Y} = \hat{\alpha} + \hat{\beta}\,\bar{X}$，

$\Rightarrow\ \hat{\alpha} = \bar{Y} - \hat{\beta}\,\bar{X}$

因此，可得迴歸估估計式 :

$$\hat{Y} = \hat{\alpha} + \hat{\beta}\,X$$

簡單線性迴歸模型 (8)

Step 1 – Find the slope (b) of the line

$$b = r\left(\frac{S_Y}{S_X}\right) = .759\left(\frac{14.337}{9.189}\right) = 1.1842$$

Step 2 – Find the y-intercept (a)

$$a = \bar{Y} - b\bar{X} = 45 - 1.1842(22) = 18.9476$$

The regression equation is :

$$\hat{Y} = a + bX$$

$$\hat{Y} = 18.9476 + 1.1842\ X$$

$$\hat{Y} = 18.9476 + 1.1842\ (20)$$

$$\hat{Y} = 42.6316$$

估計 σ^2

由於 σ^2 是母體殘差項 (Residual) ε 的變異數，因此表示爲：

$$\sigma^2 = \frac{\sum_{i=1}^{N}(\varepsilon_i - E[\varepsilon_i])^2}{N} = \frac{\sum_{i=1}^{N}(\varepsilon_i)^2}{N}。$$

因此，估計 σ^2 可以使用樣本殘差項 $e_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$ 來估計。

σ^2 估計式： $S^2_{Y|X} = \dfrac{\sum_{i=1}^{n}(e_i)^2}{n-2}$

$S^2_{Y|X}$ (or $S^2_{Y\cdot X}$)

$S^2_{Y|X}$ 是估計的標準誤 (Standard Error) ，可用來描述估計的準確性。

由標準方程式 (Normal Equation) 可以將 $S^2_{Y|X}$ 改寫如下，以方便運算。

$$S^2_{Y|X} = \frac{\sum_{i=1}^{n}Y_i^2 - \hat{\alpha}\sum_{i=1}^{n}Y_i - \hat{\beta}\sum_{i=1}^{n}X_iY_i}{n-2}$$

Standard Error of the Estimate - Example

Recall the example involving Copier Sales of America. The sales manager determined the least squares regression equation is given below.

Determine the standard error of estimate as a measure of how well the values fit the regression line.

$$\hat{Y} = 18.9476 + 1.1842X$$

Sales Representative	Actual Sales, (Y)	Estimated Sales, ($\hat{Y}$)	Deviation, ($Y - \hat{Y}$)	Deviation Squared, $(Y - \hat{Y})^2$
Tom Keller	30	42.6316	−12.6316	159.557
Jeff Hall	60	66.3156	−6.3156	39.887
Brian Virost	40	42.6316	−2.6316	6.925
Greg Fish	60	54.4736	5.5264	30.541
Susan Welch	30	30.7896	−0.7896	0.623
Carlos Ramirez	40	30.7896	9.2104	84.831
Rich Niles	40	42.6316	−2.6316	6.925
Mike Kiel	50	42.6316	7.3684	54.293
Mark Reynolds	30	42.6316	−12.6316	159.557
Soni Jones	70	54.4736	15.5264	241.069
			0.0000	784.211

$$s_{y.x} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{784.211}{10-2}} = 9.901$$

Confidence Interval and Prediction Interval

- A confidence interval reports the **mean value** of Y for a given X.
- A prediction interval reports the value of Y for a particular value of X.

CONFIDENCE INTERVAL FOR THE MEAN OF Y, GIVEN X

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \quad [13\text{–}7]$$

PREDICTION INTERVAL FOR Y, GIVEN X

$$\hat{Y} \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \quad [13\text{–}8]$$

Confidence Interval Estimate - Example

We return to the Copier Sales of America illustration. Determine a 95 percent confidence interval for **all** sales representatives who make 25 calls.

CONFIDENCE INTERVAL FOR THE MEAN OF Y, GIVEN X

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \quad [13\text{–}7]$$

where

$\hat{Y}$ is the predicted value for any selected X value.
X is any selected value of X.
$\bar{X}$ is the mean of the Xs, found by $\Sigma X/n$.
n is the number of observations.
$s_{y \cdot x}$ is the standard error of estimate.
t is the value of t from Appendix B.2 with $n - 2$ degrees of freedom.

Confidence Interval Estimate - Example

CONFIDENCE INTERVAL FOR THE MEAN OF Y, GIVEN X

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \quad [13\text{–}7]$$

Step 1 – Compute the point estimate of Y

In other words, determine the number of copiers we expect a sales representative to sell if he or she makes 25 calls.

The regression equation is :

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(25)$$

$$\hat{Y} = 48.5526$$

Confidence Interval Estimate - Example

CONFIDENCE INTERVAL FOR THE MEAN OF *Y*, GIVEN *X*

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad [13\text{–}7]$$

Step 2 – Find the value of t

- To find the *t* value, we need to first know the number of degrees of freedom. In this case the degrees of freedom is $n - 2 = 10 - 2 = 8$.
- We set the confidence level at 95 percent.
- The value of *t* is 2.306.

Confidence Interval Estimate - Example

CONFIDENCE INTERVAL FOR THE MEAN OF *Y*, GIVEN *X*

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad [13\text{–}7]$$

Step 3 – Compute $(x - \bar{x})^2$ and $\sum(x - \bar{x})^2$

Sales Representative	Sales Calls, (*X*)	Copier Sales, (*Y*)	$(X - \bar{X})$	$(X - \bar{X})^2$
Tom Keller	20	30	−2	4
Jeff Hall	40	60	18	324
Brian Virost	20	40	−2	4
Greg Fish	30	60	8	64
Susan Welch	10	30	−12	144
Carlos Ramirez	10	40	−12	144
Rich Niles	20	40	−2	4
Mike Kiel	20	50	−2	4
Mark Reynolds	20	30	−2	4
Soni Jones	30	70	8	64
			0	760

Confidence Interval Estimate - Example

CONFIDENCE INTERVAL FOR THE MEAN OF *Y*, GIVEN *X*

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad [13\text{–}7]$$

Step 4 – Use the formula above by substituting the numbers computed in previous slides

$$\text{Confidence Interval} = \hat{Y} \pm ts_{y \cdot x}\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}}$$

$$= 48.5526 \pm 2.306(9.901)\sqrt{\frac{1}{10} + \frac{(25 - 22)^2}{760}}$$

$$= 48.5526 \pm 7.6356$$

Thus, the 95 percent confidence interval for the average sales of all sales representatives who make 25 calls is from 40.9170 up to 56.1882 copiers.

Prediction Interval Estimate - Example

We return to the Copier Sales of America illustration. Determine a 95 percent prediction interval for Sheila Baker, a West Coast sales representative who made 25 calls.

Prediction Interval Estimate - Example

PREDICTION INTERVAL FOR Y, GIVEN X

$$\hat{Y} \pm ts_{y \cdot x}\sqrt{1+\frac{1}{n}+\frac{(X-\bar{X})^2}{\Sigma(X-\bar{X})^2}} \qquad [13\text{–}8]$$

Step 1 – Compute the point estimate of Y

In other words, determine the number of copiers we expect a sales representative to sell if he or she makes 25 calls.

The regression equation is :

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(25)$$

$$\hat{Y} = 48.5526$$

Prediction Interval Estimate - Example

PREDICTION INTERVAL FOR Y, GIVEN X

$$\hat{Y} \pm ts_{y \cdot x}\sqrt{1+\frac{1}{n}+\frac{(X-\bar{X})^2}{\Sigma(X-\bar{X})^2}} \qquad [13\text{–}8]$$

Step 2 – Using the information computed earlier in the confidence interval estimation example, use the formula above.

$$\text{Prediction Interval} = \hat{Y} \pm ts_{y \cdot x}\sqrt{1+\frac{1}{n}+\frac{(X-\bar{X})^2}{\Sigma(X-\bar{X})^2}}$$

$$= 48.5526 \pm 2.306(9.901)\sqrt{1+\frac{1}{10}+\frac{(25-22)^2}{760}}$$

$$= 48.5526 \pm 24.0746$$

If Sheila Baker makes 25 sales calls, the number of copiers she will sell will be between about 24 and 73 copiers.

Ex. 13.4

在大發汽車的例子，迴歸係數可使用最小平方法估計如下：

$$\hat{\beta} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{4,840,000}{875,000} = 5.53$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 12,300 - 5.53 \times 725 = 8,290.75$$

因此，迴歸式爲：

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X = 8,290.75 + 5.53 \times X$$

Ex. 13. 5

- 大發公司的廣告支出與銷售額的例子中，總經理想要預測若廣告支出為1400萬元時，汽車的平均銷售額為多少？求 95%的信賴區間為何？

Ex. 13. 5

由已經估得的迴歸估計式: $\hat{Y}=8290.75+5.53X$

將 $X_0=1400$ 帶入可得

$\hat{Y}_0 = 8290.75 + 5.53 \times 1400 = 16032.75$

即當廣告支出爲 1400 萬元時，根據估計的迴歸模型預測銷售額爲 16032.75 萬元。

57

Ex. 13. 5

求得 $\hat{Y}_0$ 樣本變異數

$$S^2_{\hat{Y}_0} = S^2_{Y|X}\left[\frac{1}{n} + \frac{\sum X_0 - \bar{X}}{\sum_{i=1}^{n} X_i^2}\right] = S^2_{Y|X}\left[\frac{1}{8} + \frac{(1400 - 725)^2}{875000}\right]$$

$= 320286 \times 0.6457 = 206809$

58

Ex. 13. 5

E(Y|X=1400) 的95% 信賴區間爲

$\hat{Y}_0 \pm t_{6,0.025}S_{\hat{Y}_0} = 16032.75 \pm t_{6,0.025}S_{\hat{Y}_0}$

$=16032.75 \pm 2.45 \times\sqrt{206809} = 16032.75 \pm 2.45\times455$

$=16032.75 \pm 1114.75$

因此，在 95% 信賴區間水準下，廣告支出爲 1400 萬元的年平均銷售額的信賴區間爲 14918 到 17147.5 萬元

59

Ex. 13. 6

- 在大發公司的廣告支出與銷售額關係的例子中，戴經理想知道如果明年的廣告支出為1400萬元時，明年的銷售額為何？

60

Ex. 13. 6

先求變異數：

$$S_{e_0}^2 = S_{Y|X}^2 \left[1 + \frac{1}{n} + \frac{\sum X_0 - \bar{X}}{\sum_{i=1}^{n} X_i^2} \right]$$

$$= S_{Y|X}^2 \left[1 + \frac{1}{8} + \frac{(1400-725)^2}{875000} \right]$$

$$= 320286 \times 1.6457 = 527095$$

Ex. 13. 6

得信賴區間爲：

$\hat{Y}_0 \pm t_{6,0.025} S_{e_0} = 16032.75 \pm t_{6,0.025} S_{e_0}$

$= 16032.75 \pm 2.45 \times \sqrt{527029} = 16032.75 \pm 2.45 \times 726$

$= 16032.75 \pm 1778.7$

因此，若明年的廣告支出爲 1400 萬元時，在 95% 信賴區間水準下，

明年汽車銷售額的信賴區間爲 14254.25 到 17811.45 萬元。

依變數的總差異

$\left(Y_i - \bar{Y}\right)$ = 依變數總差異

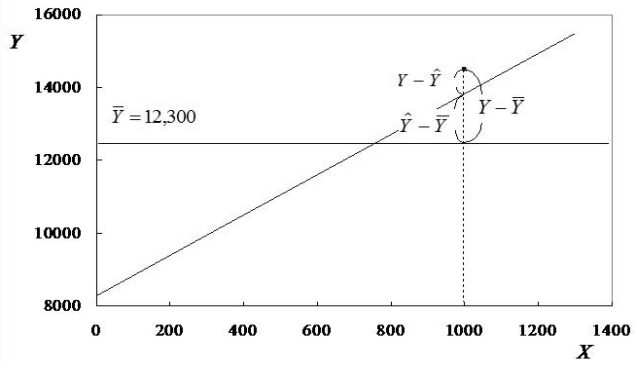
$\left(\hat{Y}_i - \bar{Y}\right)$ = 依變數可解釋差異

$\left(Y_i - \hat{Y}_i\right)$ = 依變數不可解釋差異

$\left(Y_i - \bar{Y}\right) = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$

總差異 = 可解釋差異 + 不可解釋差異

依變數的總差異



依變數的總變異

$SST = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$ = 依變數總變異 (total sum of squares)

$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ = 依變數可解釋變異 (sum of squares due to regression)

$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ = 依變數不可解釋變異 (sum of squares due to error)

SST = SSR + SSE

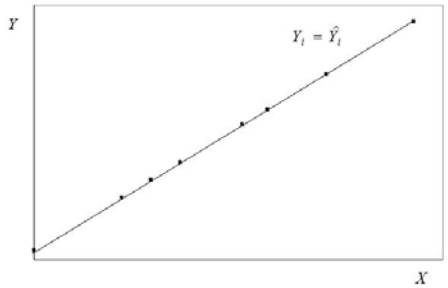
總變異 = 可解釋變異 + 不可解釋變異

R^2, 判定係數 (Coefficient of Determination)

- 可解釋變異 (SSR) 佔總變異 (SST) 的比例稱為判定係數 (R^2)。
- 判定係數可以用來衡量迴歸方程式的配適度，並衡量迴歸方程式的解釋能力。
- R^2 = SSR/SST = 1- (SSE/SST)
- $0 \leqq R^2 \leqq 1$
  - R^2 =1: 自變數完全解釋依變數。
  - R^2 =0: 自變數完全無法解釋依變數。

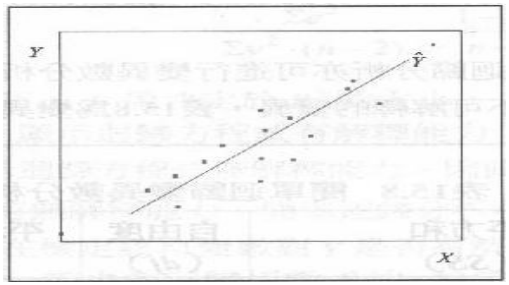
R^2, 判定係數 (Coefficient of Determination)

R^2 =1



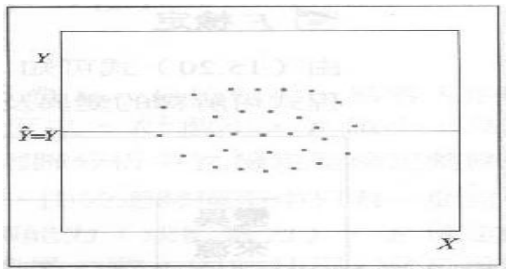
R^2, 判定係數 (Coefficient of Determination)

R^2 = 0.8



R^2, 判定係數 (Coefficient of Determination)

$R^2 = 0$



ANOVA Table

$$SST = \sum\left(Y - \bar{Y}\right)^2$$

$$SSR = \sum\left(\hat{Y} - \bar{Y}\right)^2$$

$$SSE = \sum\left(Y - \hat{Y}\right)^2 = \sum\left(Y - \hat{\alpha} - \hat{\beta}X\right)^2$$

ANOVA Table

Source	df	SS	MS
Regression	1	SSR	SSR/1
Error	n-2	SSE	SSE/(n-2)
Total	n-1	SST	SST/n-1

迴歸解釋能力的檢定

H_0 : 迴歸無解釋能力 ($\beta = 0$)

H_1 : 迴歸有解釋能力 ($\beta \neq 0$)

F檢定統計量：$F_0 = \dfrac{SSR/1}{SSE/n-2} = \dfrac{MSR}{MSE} \sim F_{1,n-2}$

決策法則：

(1) $F_0 > F_{1,n-2,\alpha} \Rightarrow$ 拒絕H_0

(2) $F_0 \leq F_{1,n-2,\alpha} \Rightarrow$ 接受H_0

Ex. 13.6

- 大發公司的廣告支出與銷售額的變異數分析結果如下表，請問迴歸模型可接受嗎?

變異來源	平方和(SS)	自由度df	平均平方和MS	F
回歸	26758287	1	26758287	83.54
隨機	1921713	6	320286	
總和	28680000	7		

Ex. 13.6

計算公式與結果如下

$SSR=\sum(\hat{Y}-\bar{Y})^2 = 26758287$

$SST=\sum(Y-\bar{Y})^2 = 28680000$

$SSE=\sum(Y-\hat{Y})^2 = SST - SSR$

$=28680000-26758287=1921713$

73

Ex. 13.6

而 MSR = SSR / 1 = 26758287 / 1 = 26758287

MSE = SSE / (n-2) = 1921713 / 6 = 320286

故 $F = \frac{MSR}{MSE} = \frac{26758287}{320286} = 83.54$

F 檢定統計量大於臨界值 $F_{1,6,0.05} = 5.99$ ，因此拒絕 H_0。

此即表示迴歸模型是可接受的，自變數與依變數有顯著關係，

迴歸方程式有解釋能力。

74

Exercise

- 1, 3, 5, 7, 9, 11, 13, 15, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 43, 45, 47, 51, 55, 57, 59,

75