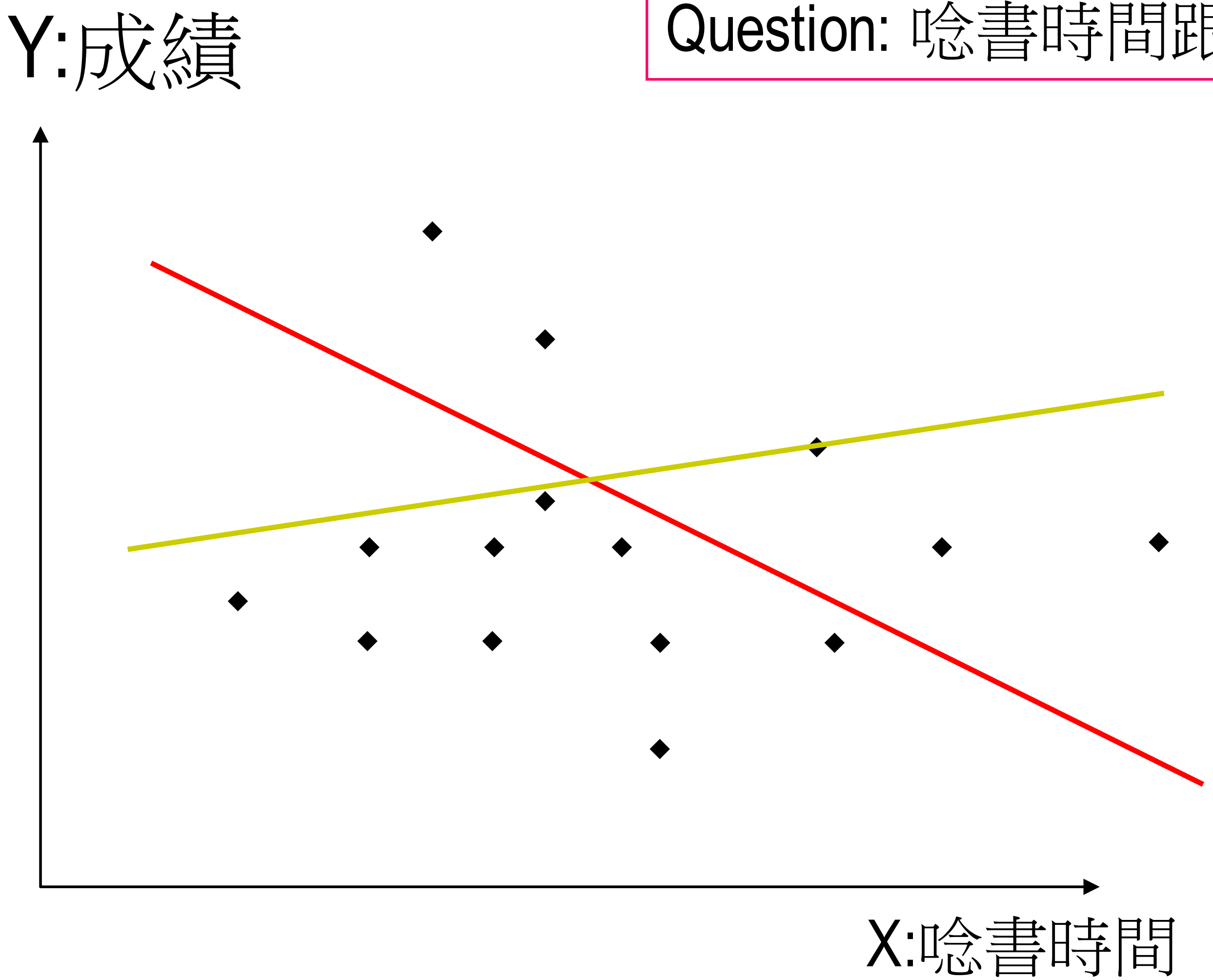
Ch 16 實習(1)

Agenda

- 如何求回歸式子？（單回歸）
- 如何檢測回歸的效力？
  - Standard error of estimate （標準誤 S_ε ）
  - Coefficient of determination (R^2)
  - T-test of coefficient of correlation (假設檢定ρ)
  - T-test of the slope (假設檢定b1 ）
- 例子：手算
- 例子：電腦報表

什麼是回歸？

Question: 唸書時間跟成績有關嗎？



什麼是回歸？

- 在給定過去歷史資料下，用來表示X和Y關聯的方程式（或是用X預測Y）

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

- 成績＝b0+b1唸書時數＋b2題目難度
- 何謂單回歸？何謂複回歸？

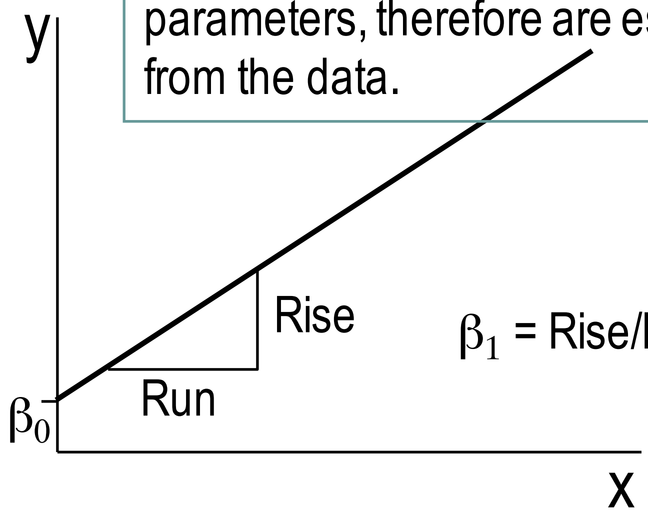
The Model

- The first order linear model (simple linear regression model)

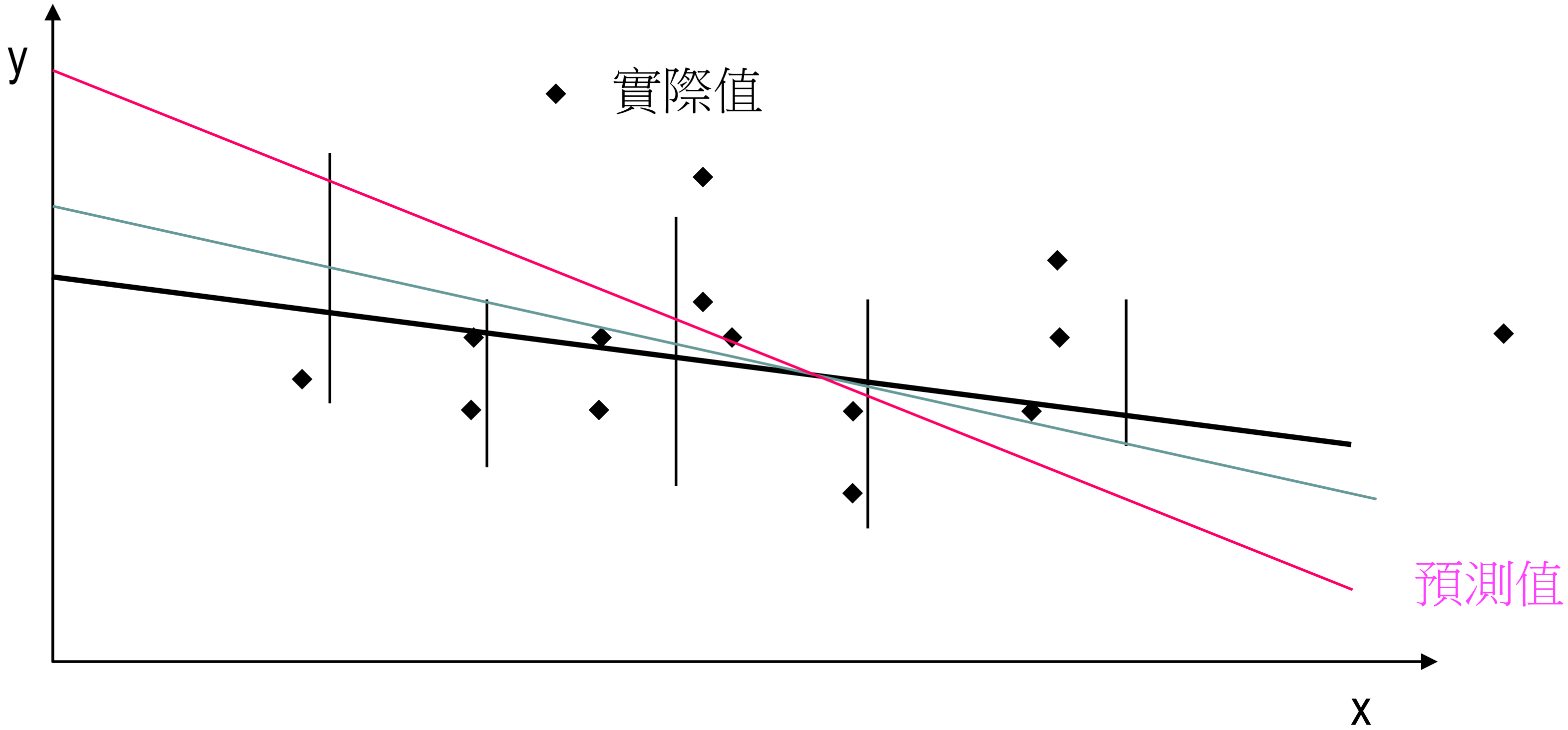
$$y = \beta_0 + \beta_1 x + \varepsilon$$

y = dependent variable (相依變數）
x = independent variable （獨立變數）
β_0 = y-intercept （截句項）
β_1 = slope of the line（斜率,解釋力）
ε = error variable（殘差）

β_0 and β_1 are unknown population parameters, therefore are estimated from the data.



The Least Squares (Regression) Line



A good line is one that minimizes the sum of squared differences between the points and the line.

OLS認為一條好的回歸式：目的在於找出一條回歸式子(預測值)，可以跟實際值之間的差異（誤差）平方加總最小

1. 回歸如何求得？

$$\hat{y} = b_0 + b_1 X + \varepsilon$$

$$b_1 = S_{xy}/S_X^2 \qquad b_0 = \bar{y}\text{-}b_1\bar{X}$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}]$$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1}{n-1}[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}]$$

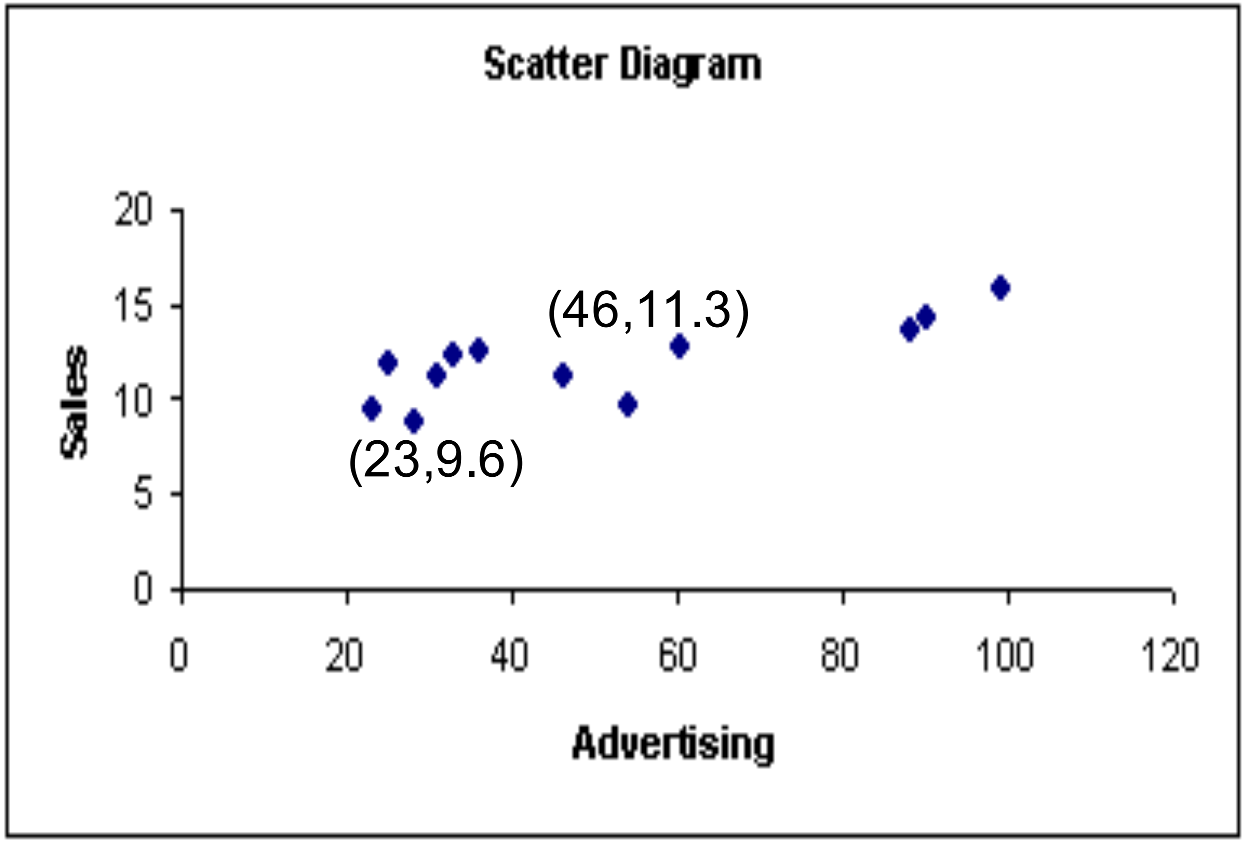
Example 1

- Attempting to analyze the relationship between advertising and sales, the owner of a furniture store recorded the monthly advertising budget ($thousands) and the sales ($millions) for a sample of 12 month. The data are listed here.

Advertising	23	46	60	54	28	33
Sales	9.6	11.3	12.8	9.8	8.9	12.5
Advertising	25	31	36	88	90	99
Sales	12.0	11.4	12.6	13.7	14.4	15.9

- a. Draw a scatter diagram. Does it appear that advertising and sales are linearly related?
- b. Calculate the least squares line and interpret the coefficients.

Solution 1 (1)



It seems that advertising and sales are linear related

Solution 1 (2)

b	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	23	9.6	529	92.16	220.8
	46	11.3	2,116	127.69	519.8
	60	12.8	3,600	163.84	768.0
	54	9.8	2,916	96.04	529.2
	28	8.9	784	79.21	249.2
	33	12.5	1,089	156.25	412.5
	25	12.0	625	144.00	300.0
	31	11.4	961	129.96	353.4
	36	12.6	1,296	158.76	453.6
	88	13.7	7,744	187.69	1205.6
	90	14.4	8,100	207.36	1296.0
	99	15.9	9,801	252.81	1,574.1
Total	613	144.9	39,561	1,795.77	7,882.2

$\sum_{i=1}^{n} x_i = 613$ $\quad \sum_{i=1}^{n} y_i = 144.9$ $\quad \sum_{i=1}^{n} x_i^2 = 39,561$ $\quad \sum_{i=1}^{n} x_i y_i = 7,882.2$

Solution 1 (3)

$$\sum_{i=1}^{n} x_i = 613 \rightarrow \sum_{i=1}^{n} y_i = 144.9 \rightarrow \sum_{i=1}^{n} x_i^2 = 39{,}561 \rightarrow \sum_{i=1}^{n} x_i y_i = 7{,}882.2$$

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{12-1}\left[7{,}882.2 - \frac{(613)(144.9)}{12}\right] = 43.66$$

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{12-1}\left[39{,}561 - \frac{(613)^2}{12}\right] = 749.7$$

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{43.66}{749.7} = .0582$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{613}{12} = 51.08$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{144.9}{12} = 12.08$$

$$b_0 = \bar{y} - b_1\bar{x} = 12.08 - (.0582)(51.08) = 9.107$$

$$b_1 = \frac{s_{xy}}{s_x^2}$$

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

The sample regression line is

$$\hat{y} = 9.107 + .0582x$$

每增加一單位的X（廣告費用），會增加0.0582單位的Y（銷售）

The slope tells us that for each additional thousand dollars of advertising sales increase on average by .0582 million. The y-intercept has no practical meaning.

2. Assessing the Model

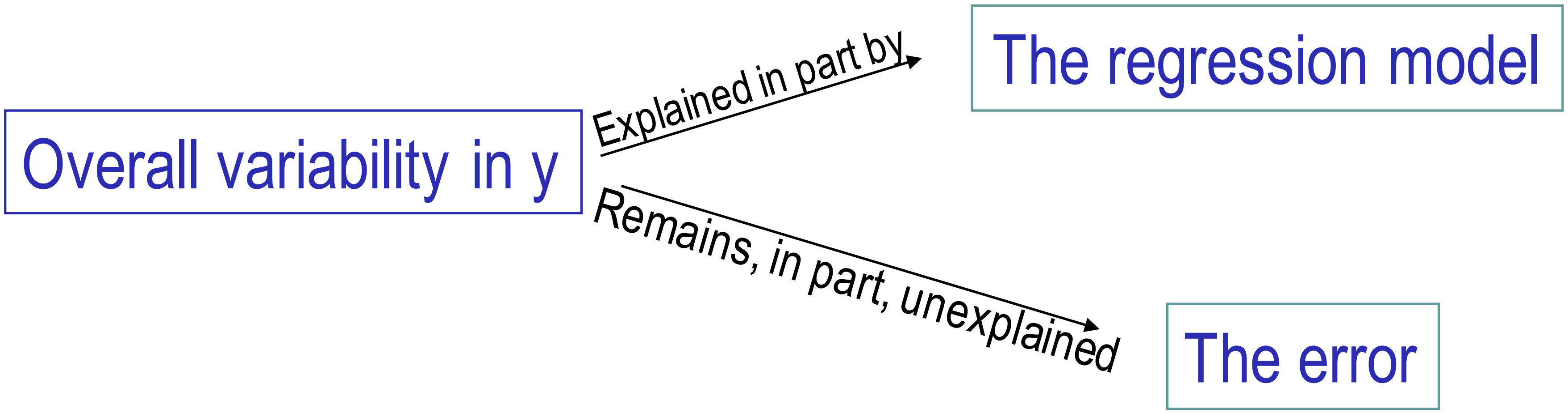
- 如何知道回歸式（模型）好不好？
  - 用讀書時間(X)來解釋成績(Y)，到底有沒有達到統計上顯著的關係？

- Standard error of estimate （標準誤 S_ε ）
- Coefficient of determination (判定系數 R^2)
- T-test of coefficient of correlation (假設檢定ρ)
- T-test of the slope (假設檢定b1）

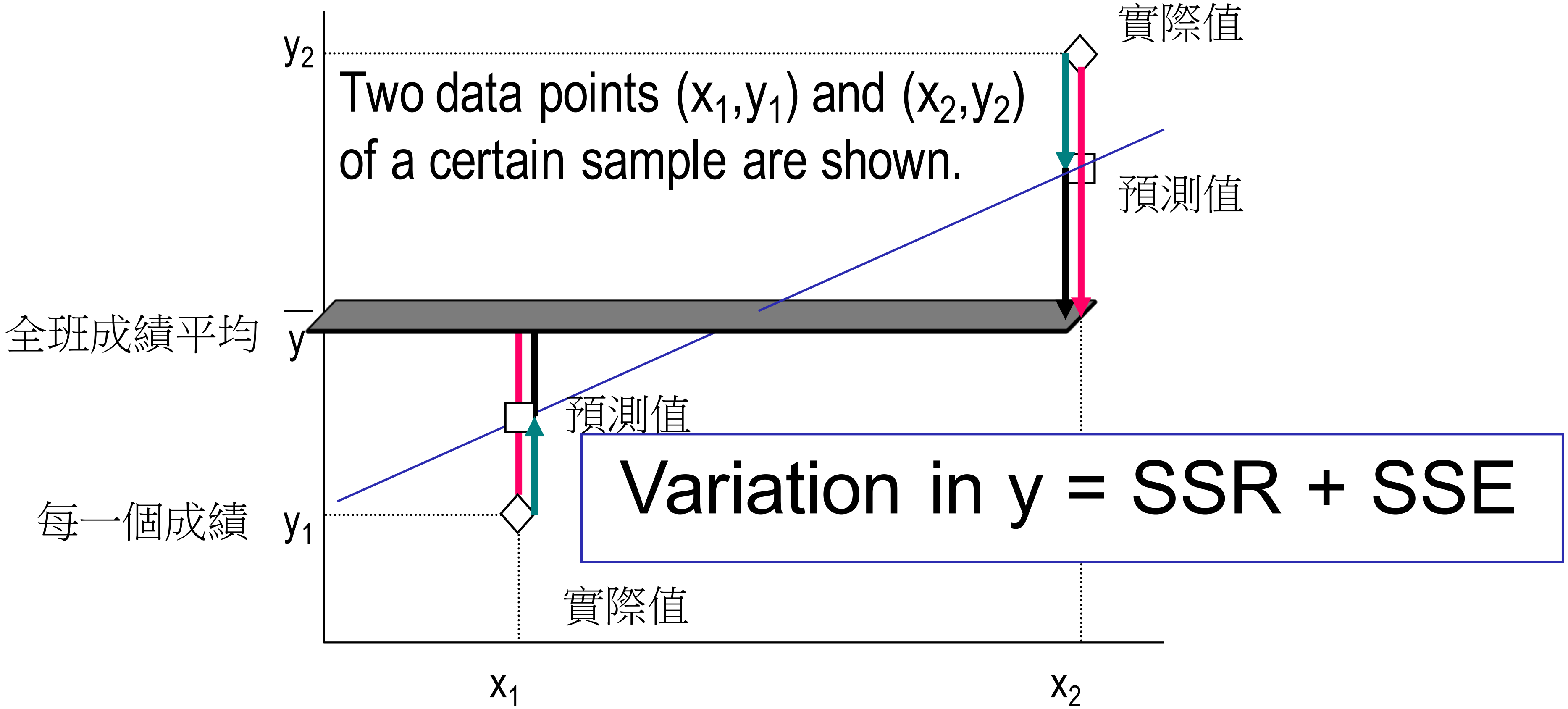
Variability in reg

Variation in y = SSR + SSE

- To understand the significance of this coefficient note:



Variability in reg



Total variation in y = Variation explained by the regression line + Unexplained variation (error)

$$(y_1-\bar{y})^2+(y_2-\bar{y})^2 = (\hat{y}_1-\bar{y})^2+(\hat{y}_2-\bar{y})^2 + (y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2$$

(1) Standard Error of Estimate （標準誤 S_ε）

- The mean error is equal to zero.
- If σ_ε is small the errors tend to be close to zero (close to the mean error). Then, the model fits the data well.
- Therefore, we can, use σ_ε as a measure of the suitability of using a linear model.
- An estimator of σ_ε is given by s_ε

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

實際值-預測值

$$SSE = (n-1)\left(s_Y^2 - \frac{s_{xy}^2}{s_x^2} \right)$$

– A shortcut formula

Standard Error of Estimate

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

(2) Coefficient of determination (R^2)

- To measure the strength of the linear relationship we use the coefficient of determination.
- 整條回歸式子的解釋力

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} \quad or \quad R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

無法解釋的變異

全部的變異

$$= \frac{SSR}{\sum (y_i - \bar{y})^2}$$

在單回歸中，$r^2 = R^2$

$$r = \frac{S_{xy}}{S_x S_y}$$

$$= \frac{SSR}{SST}$$

(2) Coefficient of determination (R^2)

- R^2 measures the proportion of the variation in y that is explained by the variation in x.

$$R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2} = \frac{\sum (y_i - \bar{y})^2 - SSE}{\sum (y_i - \bar{y})^2} = \frac{SSR}{\sum (y_i - \bar{y})^2}$$

- R^2 takes on any value between zero and one.
  - R^2 = 1: Perfect match between the line and the data points.
  - R^2 = 0: There are no linear relationship between x and y.
- 若欲利用判定係數來比較不同模型的配適能力，這些模型必須有相同的依變數(y)。

- Standard error of estimate (標準誤)

~~越小越好

- Coefficient of determination （判定係數 ,R^2)

~~越大越好

但何謂大小？沒有統計上的標準

(3) T-test of coefficient of correlation (假設檢定ρ)

Cor (X, Y)

Step 1:

H_0: ρ= 0

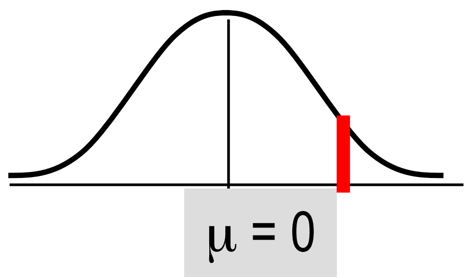
H_1: $\rho \neq 0$ (or < 0,or > 0)

Step2: Critical point: $t_{a/2,\ df=n-2}$

Step3: The test statistic is

$$t = r\sqrt{\frac{n-2}{1-r^2}} \qquad r = \frac{S_{xy}}{S_x S_y}$$

Step4: 結論



(4) T-test of the slope (假設檢定b1）

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

Step 1:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$ (or < 0,or > 0)

Step2: Critical point: $t_{a/2,\ df=n-2}$

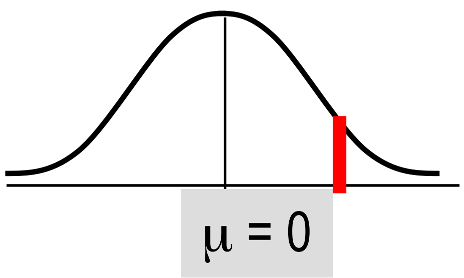
重要：

Standard error=√SSE/n-2

=√MSE

Step3: The test statistic is

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \qquad s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$



Step4: 結論

Example 2

Advertising	23	46	60	54	28	33
Sales	9.6	11.3	12.8	9.8	8.9	12.5
Advertising	25	31	36	88	90	99
Sales	12.0	11.4	12.6	13.7	14.4	15.9

(1) Calculate the least square line (求回歸線)

(2) Determine the standard error of estimate and describe what this statistic tells you about the regression line. (標準誤 s_ε)

(3) Determine the coefficient of determination and discuss what its value tells you about the two variables (判定係數, R^2)

(4) Calculate the Pearson correlation coefficient. What sign does it have? Why? (求相關係數r)

(5) Conduct a test of the population coefficient of correlation to determine at the 5% significance level whether a linear relationship exists between sale and adverting. (假設檢定ρ)

(6) Conduct a test of the population slope to determine at the 5% significance level whether a linear relationship exists between sale and adverting. (假設檢定b1)

1. Calculate the least square line (求回歸線)

b	x_i	y_i	x_i^2	y_i^2	x_iy_i
	23	9.6	529	92.16	220.8
	46	11.3	2,116	127.69	519.8
	60	12.8	3,600	163.84	768.0
	54	9.8	2,916	96.04	529.2
	28	8.9	784	79.21	249.2
	33	12.5	1,089	156.25	412.5
	25	12.0	625	144.00	300.0
	31	11.4	961	129.96	353.4
	36	12.6	1,296	158.76	453.6
	88	13.7	7,744	187.69	1205.6
	90	14.4	8,100	207.36	1296.0
	99	15.9	9,801	252.81	1,574.1
Total	613	144.9	39,561	1,795.77	7,882.2

$$\sum_{i=1}^{n} x_i = 613 \qquad \sum_{i=1}^{n} y_i = 144.9 \qquad \sum_{i=1}^{n} x_i^2 = 39{,}561 \qquad \sum_{i=1}^{n} x_i y_i = 7{,}882.2$$

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{12-1}\left[7{,}882.2 - \frac{(613)(144.9)}{12}\right] = 43.66$$

1. Calculate the least square line (求回歸線)

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{12-1}\left[39{,}561 - \frac{(613)^2}{12}\right] = 749.7$$

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{43.66}{749.7} = .0582$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{613}{12} = 51.08$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{144.9}{12} = 12.08$$

$$b_0 = \bar{y} - b_1\bar{x} = 12.08 - (.0582)(51.08) = 9.107$$

The sample regression line is

$$\hat{y} = 9.107 + .0582x$$

The slope tells us that for each additional thousand dollars of advertising sales increase on average by .0582 million. The y-intercept has no practical meaning.

Y=9.107+0.582 X

(2) Standard error of estimate (標準誤 S_ε)

- Determine the standard error of estimate and describe what this statistic tells you about the regression line. (標準誤 S_ε)

$$17.22\ \text{a}\ \ s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{12-1}\left[1{,}795.77 - \frac{(144.9)^2}{12}\right] = 4.191$$

$$SSE = (n-1)\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right) = (12-1)\left(4.191 - \frac{(43.66)^2}{749.7}\right) = 18.13$$

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{18.13}{12-2}} = 1.347\ (\text{Excel: } s_\varepsilon = 1.347)$$

標準誤很小，表示此模型可以有效解釋銷售的變異

(3) Coefficient of determination (判定係數,R ^{2})

- Determine the coefficient of determination and discuss what its value tells you about the two variables (判定係數,R ^{2})

$$R^2 = \frac{S_{Xy}^2}{S_X^2 S_y^2} = \frac{(43.66)^2}{(749.7)(4.191)} = 0.6076$$

表示此模型（廣告花費），可以解釋60.76%銷售的變異

which means that 60.067% of the variation in the sale is explained by the variation in the advertising .

(4) coefficient of correlation (求r)

Calculate the Pearson correlation coefficient. What sign does it have? Why? (求相關係數r)

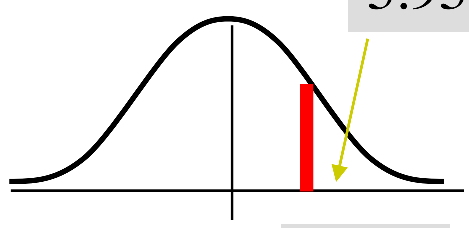
$$R^2 = \frac{S_{Xy}^2}{S_X^2 S_y^2} = \frac{(43.66)^2}{(749.7)(4.191)} = 0.6076$$

$$r = \frac{S_{xy}}{S_x S_y} = \sqrt{R^2} = 0.7794$$

只有在單回歸時，才可以這樣算喔！

(5) T-test of coefficient of correlation (假設檢定ρ)

Conduct a test of the population coefficient of correlation to determine at the 5% significance level whether a linear relationship exists between sale and adverting. (假設檢定ρ)



Step 1:

H_0: $\rho = 0$

H_1: $\rho \neq 0$ (or < 0,or > 0)

Step2:

Critical point: $t_{a/2,\ df=n-2} = t_{0.025,10} = 2.228$

Step3: The test statistic is

$$r = 0.7794, n = 12$$

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.7794 * \sqrt{\frac{12-2}{1-(0.7794)^2}} = 3.93$$

Step4:

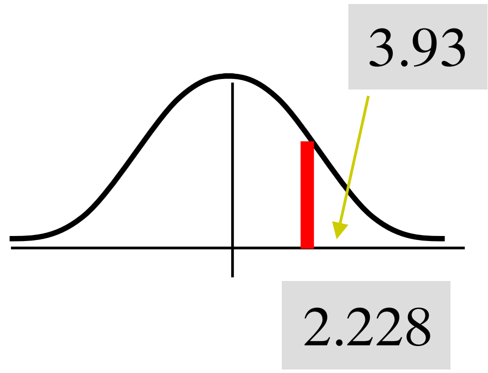
Reject H0. A positive linear relationship exists between sale and advertising , according to this data.

(6)請檢定 b1是否顯著

- Conduct a test of the population slope to determine at the 5% significance level whether a linear relationship exists between sale and adverting . (假設檢定b1)

b $H_0: \beta_1 = 0$

→ $H_1: \beta_1 \neq 0$



Rejection region: $t > t_{\alpha/2,n-2} = t_{.025,10} = 2.228$ or $t < -t_{\alpha/2,n-2} = -t_{.025,10} = -2.228$

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}} = \frac{1.347}{\sqrt{(12-1)(749.7)}} = .0148$$

Y=9.107+0.582 X

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{.0582 - 0}{.0148} = 3.93$$ (Excel: t = 3.93, p−value = .0028. There is enough evidence to infer a linear relationship between advertising and sales.

- 報表怎麼看？
- 該如何用報表數字，計算上述的問題

(1) Calculate the least square line (求回歸線)

(2) Determine the standard error of estimate and describe what this statistic tells you about the regression line. (標準誤 S_ε)

(3) Determine the coefficient of determination and discuss what its value tells you about the two variables (判定係數, R^2)

(4) Calculate the Pearson correlation coefficient. What sign does it have? Why? (求相關係數r)

(5) Conduct a test of the population coefficient of correlation to determine at the 5% significance level whether a linear relationship exists between sale and adverting. (假設檢定ρ)

(6) Conduct a test of the population slope to determine at the 5% significance level whether a linear relationship exists between sale and adverting. (假設檢定b1)

ANOVA table in the Linear Regression Model

Source	d.f.	Sums of Squares	Mean Squares	F Statistics
Regression	k	SSR	MSR= SSR/k	F=MSR/MSE
Error	n-k-1	SSE	MSE= SSE/(n-k-1)	
Total	n-1	Variation in y		

ANOVA table in the Simple Linear Regression Model

Source	d.f.	Sums of Squares	Mean Squares	F Statistics
Regression	1	SSR	MSR= SSR/1	F=MSR/MSE
Error	n-2	SSE	MSE= SSE/(n-2)	
Total	n-1	Variation in y		

- Calculate the least square line (求回歸線)
- Determine the standard error of estimate and describe what this statistic tells you about the regression line. （標準誤 S_ε）

Linear Regression

Regression Statistics							
R	0.77882						
R Square	0.60656						
Adjusted R Square	0.56722						
S	1.3468						
Total number of observations	12						
Sales($millions) = 9.1004 + 0.0582 * Advertising($thousands)							
ANOVA							
	d.f.	*SS*	*MS*	*F*	*p-level*		
Regression	1.	27.96391	27.96391	15.41681	0.00284		
Residual	10.	18.13859	1.81386				
Total	11.	46.1025					
	Coefficients	*Standard Error*	*LCL*	*UCL*	*t Stat*	*p-level*	*H0 (5%) rejected?*
Intercept	9.10037	0.85153	7.20305	10.9977	10.68711	0.	Yes
Advertising($thousands)	0.05823	0.01483	0.02519	0.09128	3.92642	0.00284	Yes
T (5%)	2.22814						
LCL - Lower value of a reliable interval (LCL)							
UCL - Upper value of a reliable interval (UCL)							

$S_\varepsilon = \sqrt{1.81366}$

Y=9.107+0.582 X

- Determine the coefficient of determination and discuss what its value tells you about the two variables (判定係數, R^2)
- Calculate the Pearson correlation coefficient. What sign does it have? Why? (求相關係數r)
- Conduct a test of the population coefficient of correlation to determine at the 5% significance level whether a linear relationship exists between sale and adverting. (假設檢定ρ)~請手算

Linear Regression							
Regression Statistics							
R	0.77882						
R Square	0.60656						
Adjusted R Square	0.56722						
S	1.3468						
Total number of observations	12						
Sales($millions) = 9.1004 + 0.0582 * Advertising($thousands)							
ANOVA							
	d.f.	*SS*	*MS*	*F*	*p-level*		
Regression	1.	27.96391	27.96391	15.41681	0.00284		
Residual	10.	18.13859	1.81386				
Total	11.	46.1025					
	Coefficients	*Standard Error*	*LCL*	*UCL*	*t Stat*	*p-level*	*H0 (5%) rejected?*
Intercept	9.10037	0.85153	7.20305	10.9977	10.68711	0.	Yes
Advertising($thousands)	0.05823	0.01483	0.02519	0.09128	3.92642	0.00284	Yes
T (5%)	2.22814						
LCL - Lower value of a reliable interval (LCL)							
UCL - Upper value of a reliable interval (UCL)							

- Conduct a test of the population slope to determine at the 5% significance level whether a linear relationship exists between sale and adverting. (假設檢定b1)

ANOVA					
	d.f.	*SS*	*MS*	*F*	*p-level*
Regression	1.	27.96391	27.96391	15.41681	0.00284
Residual	10.	18.13859	1.81386		
Total	11.	46.1025			

	Coefficients	*Standard Error*	*LCL*	*UCL*	*t Stat*	*p-level*	*H0 (5%) rejected?*
Intercept	9.10037	0.85153	7.20305	10.9977	10.68711	0.	Yes
Advertising($thousands)	b1 0.05823	S_b 0.01483	0.02519	0.09128	3.92642	0.00284	Yes
T (5%)	2.22814	T-critical			t		
LCL - Lower value of a reliable interval (LCL)							
UCL - Upper value of a reliable interval (UCL)							