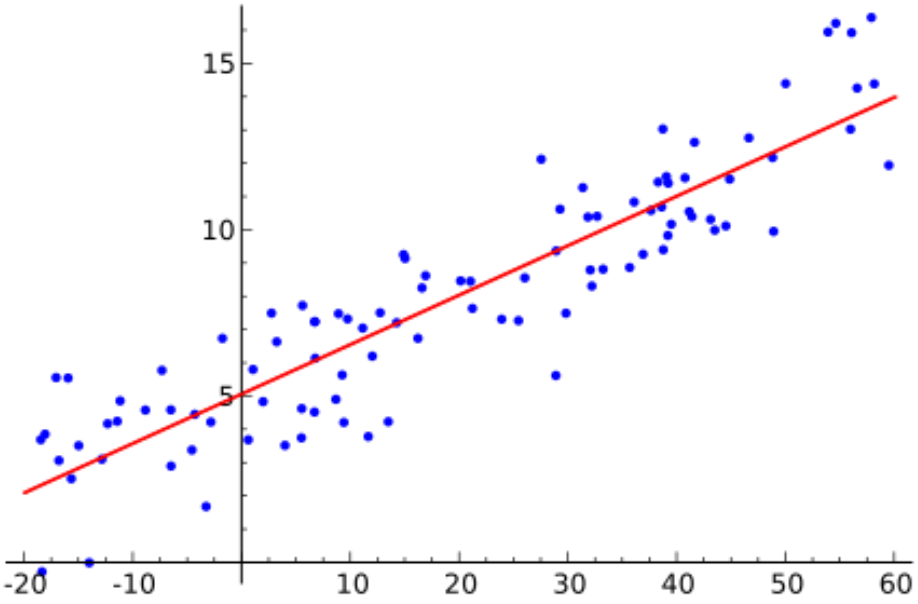
Regression Analysis

~ Basic concept ~

迴歸分析基本觀念

迴歸分析（Regression Analysis）起源

迴歸的最早形式是最小平方法，由1805年的勒壤得(Legendre)，和1809年的高斯(Gauss)出版。勒壤得和高斯都將該方法應用於從天文觀測中確定關於太陽的物體的軌道（主要是彗星，但後來是新發現的小行星）的問題。



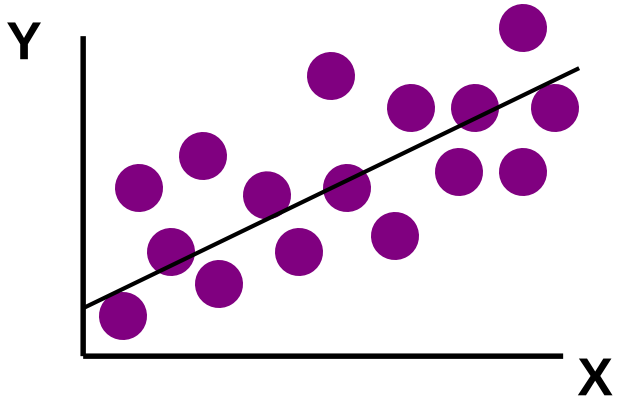
「迴歸」一詞最早由法蘭西斯·高爾頓（Francis Galton）所使用。他曾對親子間的身高做研究，發現父母的身高雖然會遺傳給子女，但子女的身高卻有逐漸「迴歸到中等（即人的平均值）」的現象，他把這個「極端」往「平均」移動的現象稱為「 regression to the mean 」。不過當時的迴歸和現在的迴歸在意義上已不盡相同。

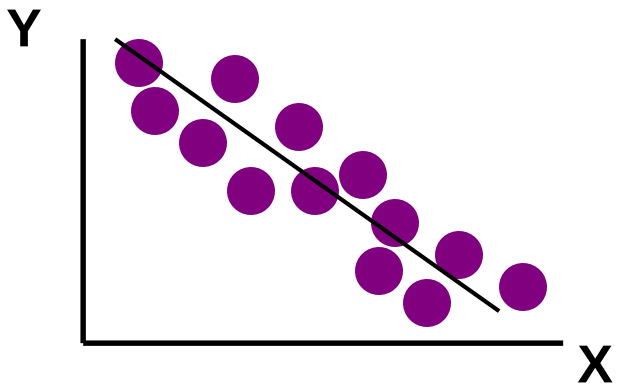
資料來源：維基百科

Types of Relationships

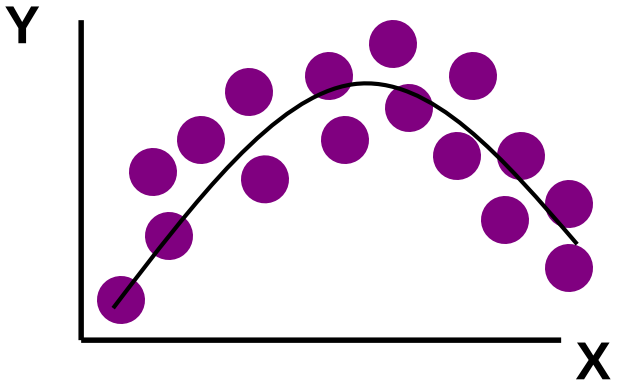
DCOVA

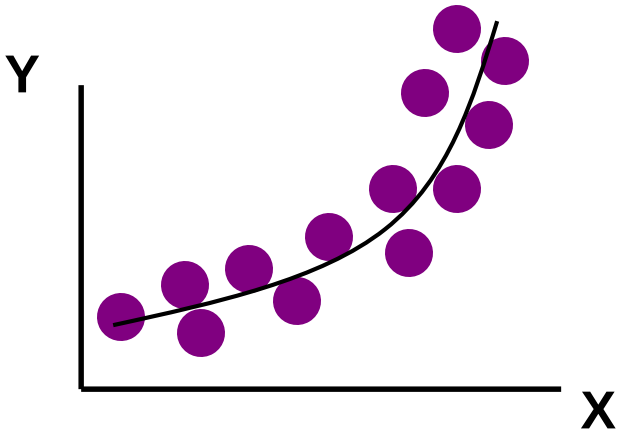
Linear relationships





Curvilinear relationships



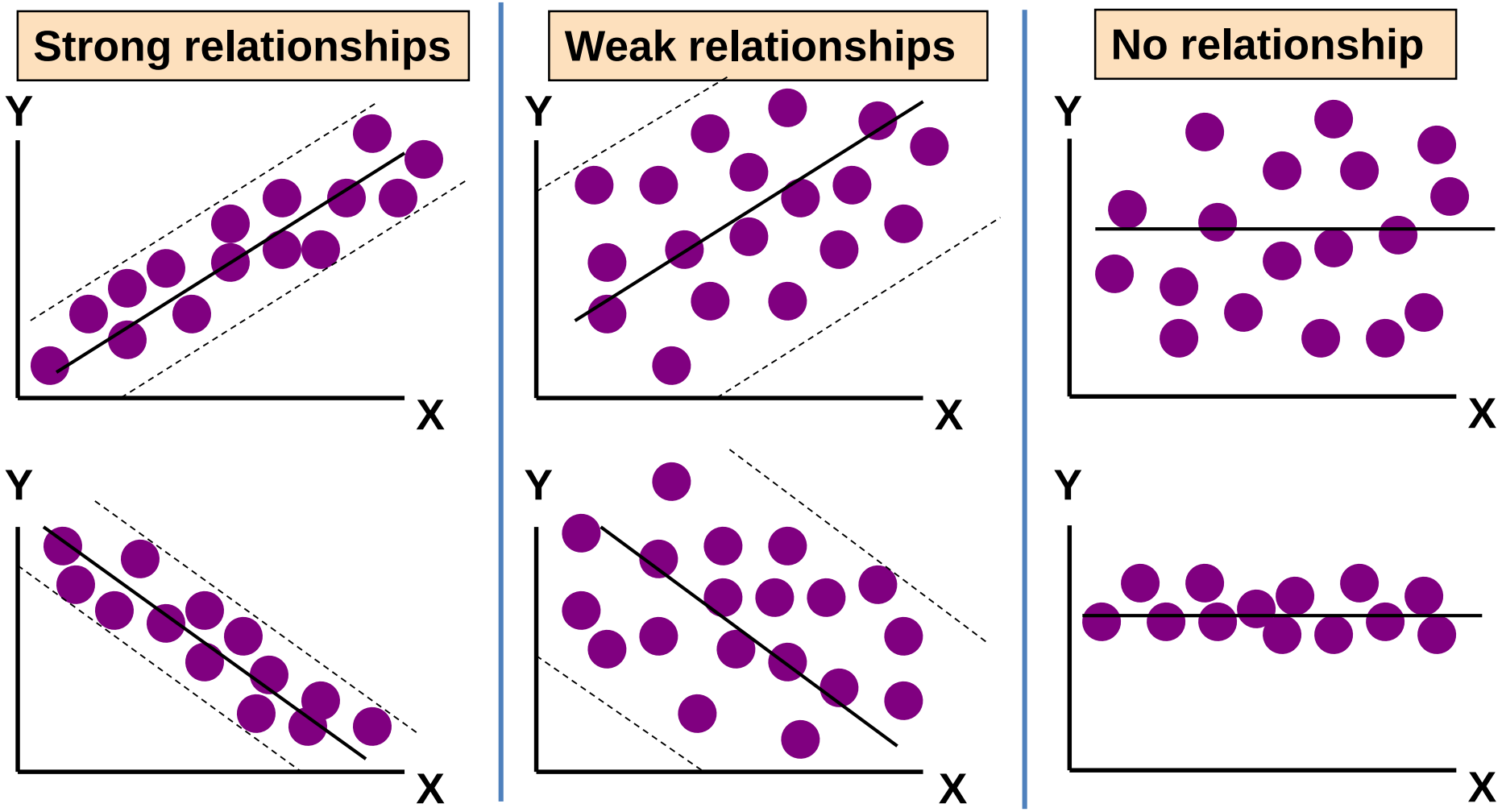


Copyright ©2014 Statistics for Managers Using Microsoft Excel 7th Edition, Global Edition

Types of Relationships

DCOVA

(continued)



Copyright ©2014 Statistics for Managers Using Microsoft Excel 7th Edition, Global Edition

基本觀念

利用數學方法，探索變數間的關係；一旦關係確認，也可以用來進行變數的預測與資料的差補（有空缺的資料）

自變數與依變數

自變數（independent variable，常以「X」表示總集合）

又稱為解釋變數（explanatory variable），通常是指用來解釋會造成影響的變數。以「因果關係」的概念來說，就是指「因」的角色；這種因果的關係性，通常是靠邏輯思考來設立應有〝關連性〞，統計算是協助驗證的一種工具，所以因果關係的確立，並不是由統計方法反推獲得的結果，而是先由研究者建立〝假設〞，經由統計結果獲得論證。

在簡單線性迴歸分析中，自變數必須是連續變項。

依變數（dependent variable ，常以「Y」表示總集合）

又稱為應變數（response variable），是指會隨著自變數變動而改變的數。以「因果關係」的概念來說，就是指「果」的角色。有時研究主題，何者是自變數，何者是依變數，必須依照研究相關資訊去界定，例如商統成績與上課出席間的關係，是成績影響出席？還是出席影響成績？有時會需要從設定的研究主題來決定兩者的角色定位。

依變數也必須是連續變項。

迴歸分析的類型

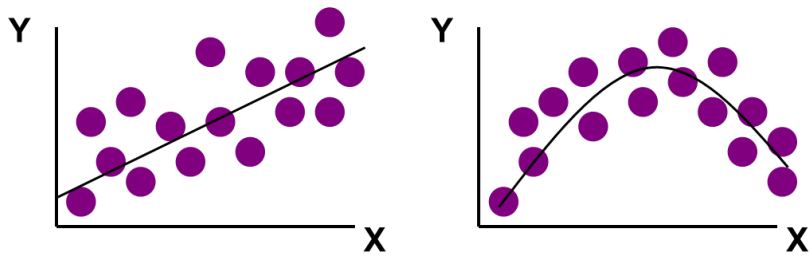
依照變數的多寡

		依變數（Y）	
		一個	多個
自變數(X)	一個	簡單迴歸分析 simple regression analysis $y=f(x)$	
	多個	多元迴歸分析 multiple regression analysis $y=f(x_1, x_2, x_3, ..., x_n)$	多變量迴歸分析 multivariate regression analysis $(y_1, y_2, .., y_m)=f(x_1, x_2, ..., x_n)$

依照變數間的關係型態

線性迴歸（linear regression）

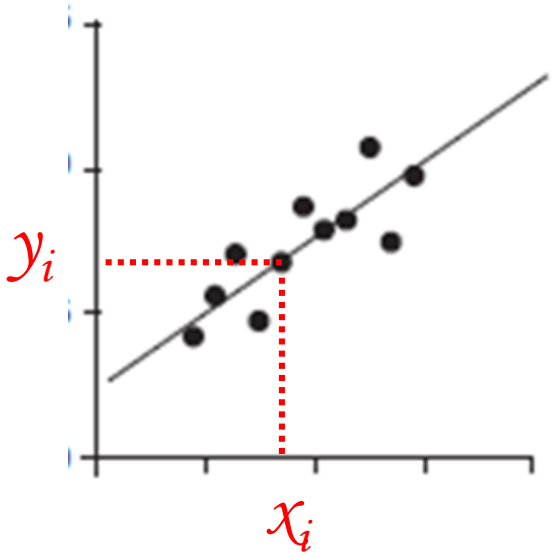
非線性迴歸（nonlinear regression）



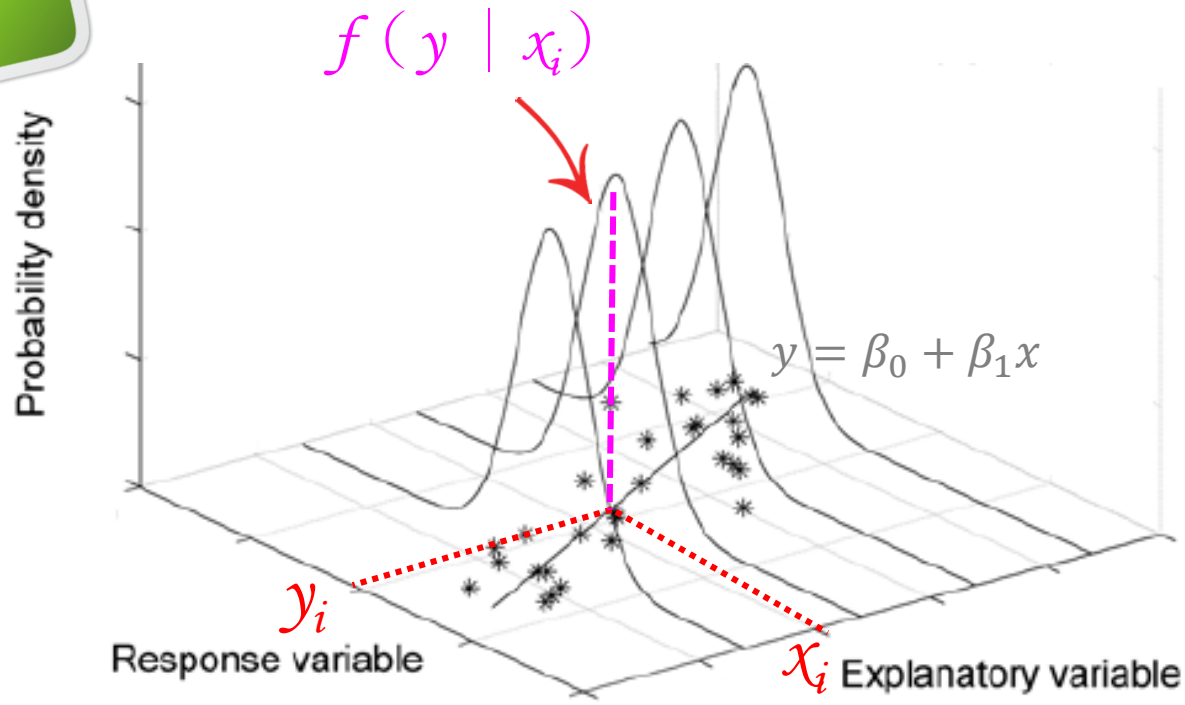
簡單迴歸分析
（線性迴歸）

simple regression analysis

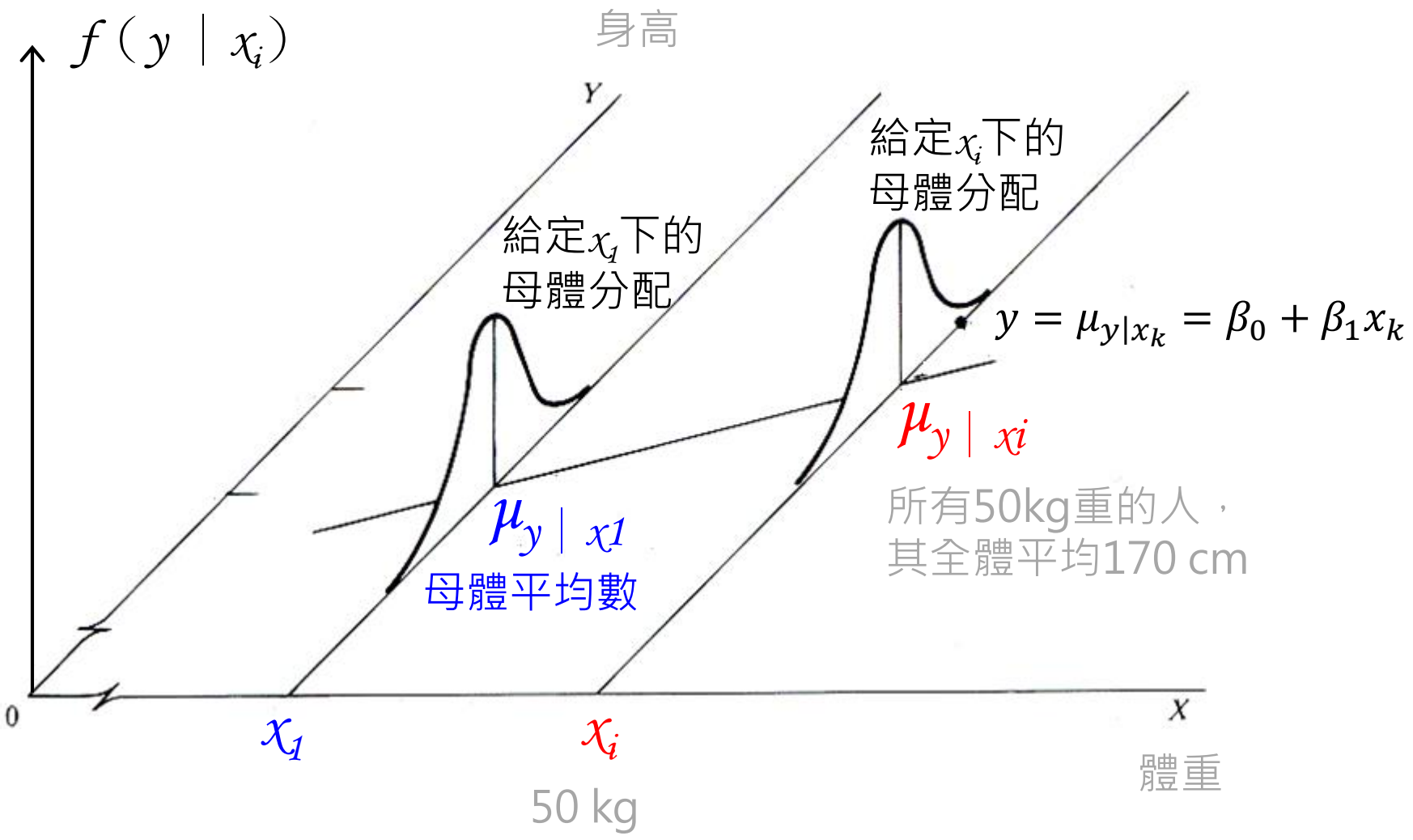
線性迴歸基本觀念示意圖



$y = \beta_0 + \beta_1 x$ 母體迴歸線（數學上的定義）



母體迴歸線與母體分配的關係



滿足母體線性迴歸基本假設

1.常態性：所有對應的y值都符合常態分配

2.齊一性（homocedasticity）：所有對應y值的變異數都相等，亦即~

Var(y)=$\sigma^2_{y|x_1} = \sigma^2_{y|x_2} = \cdots\cdots = \sigma^2_{y|x_n} = \sigma$

3.母體迴歸線即是所有小母體平均數通過的直線，函式表示~

$\mu_{y|x} = \alpha_0 + \alpha_1 x$ ，其中α_0、α_1為母體迴歸係數

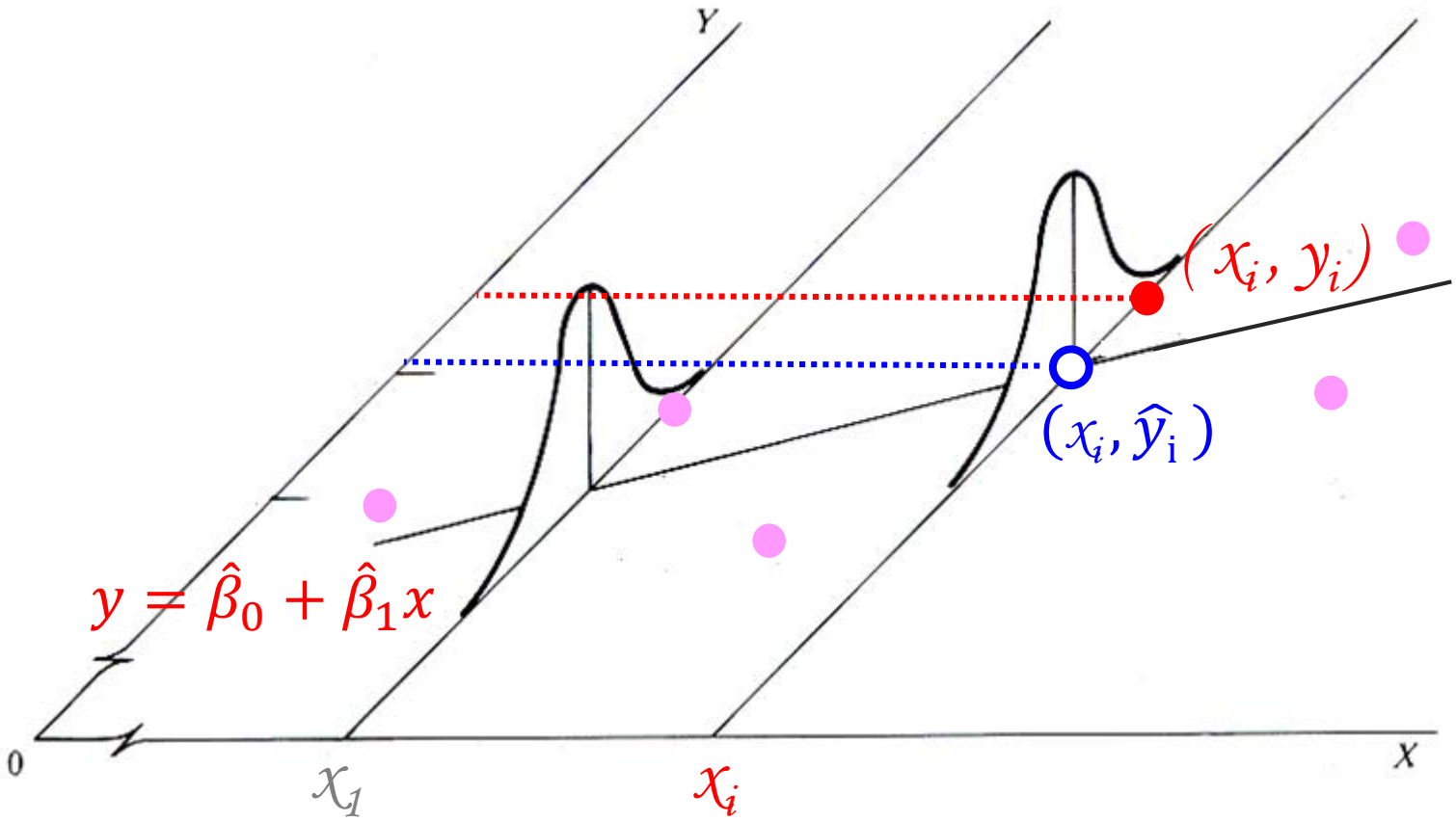
4.不同X下的Y彼此獨立

5.樣本資料與迴歸線上的差距（殘差），遵循常態分配，彼此互為獨立，且期望值為0，亦即~

ε_i ~ N（0, σ^2）且 Cov（ ε_i , ε_j ）=0

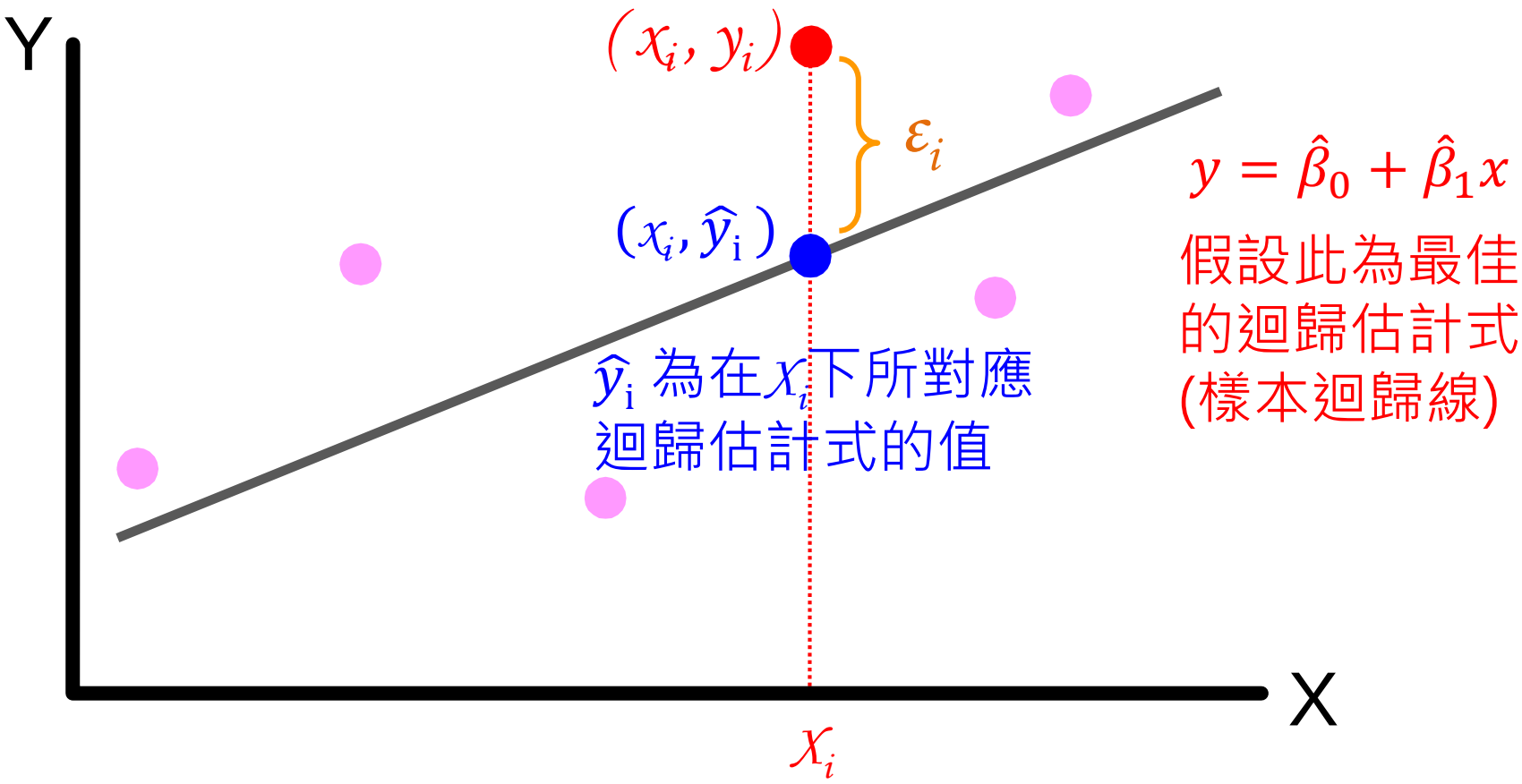
〔真實〕樣本迴歸線(方程式)

實際上，我們不知道母體的分配狀況，但可以利用所抽取的樣本觀測值(x_i, y_i)來推論。假設用樣本所推論出來的截距$\hat{\beta}_0$來表示β_0；斜率$\hat{\beta}_1$來表示β_1，其所形成的線性方程式，即為此樣本迴歸方程式。



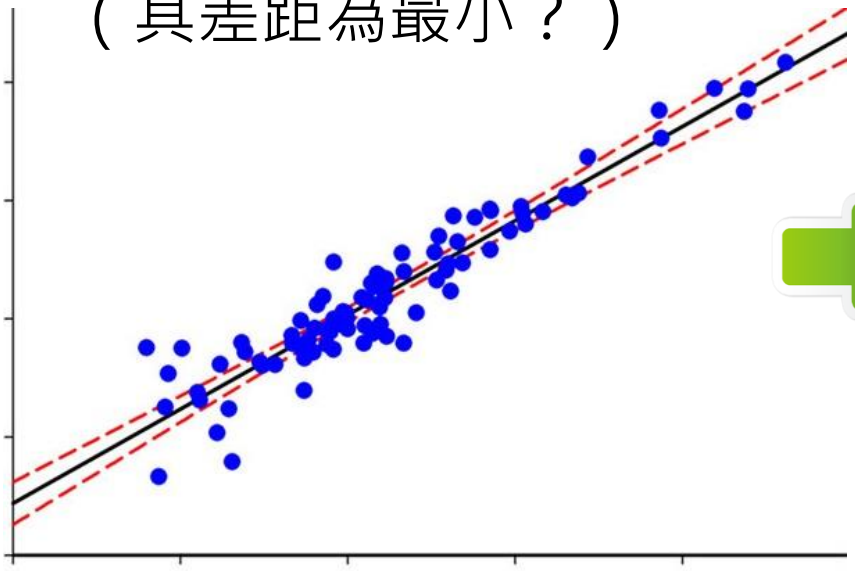
殘差：樣本值與迴歸線上的差距

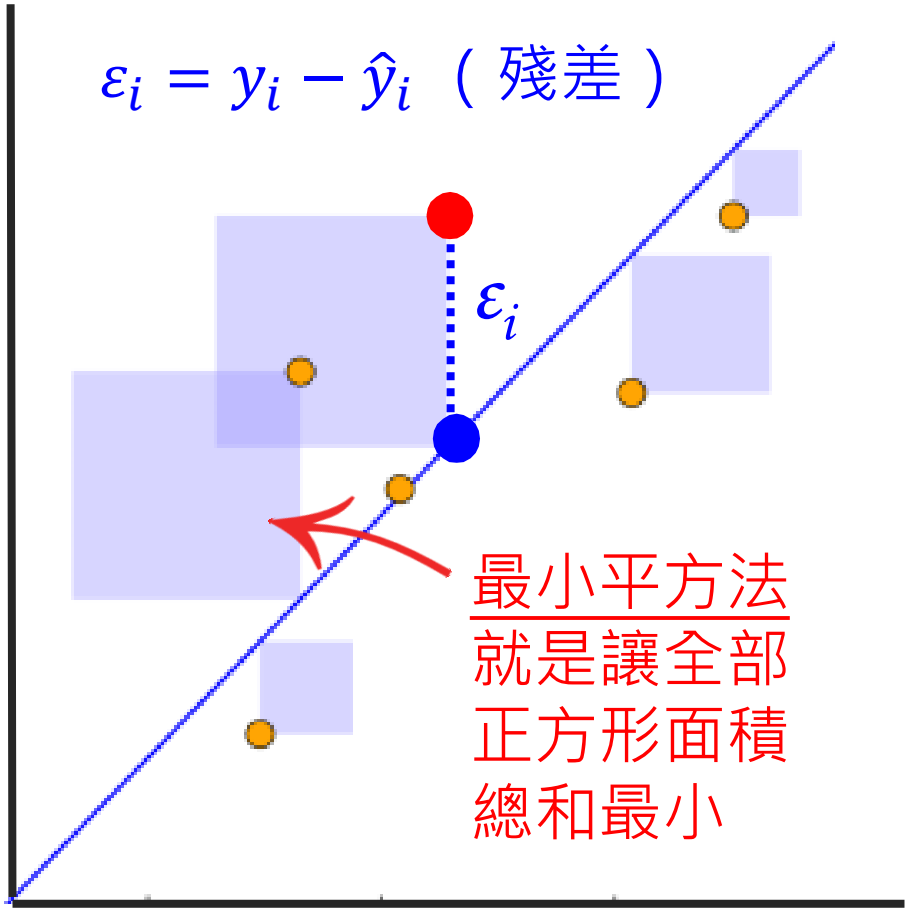
$\varepsilon_i = y_i - \hat{y}_i$（殘差）



推導樣本迴歸線～最小平方法

哪一條線比較具有代表性？
（其差距為最小？）





$$\min \sum \varepsilon_i^2 = \min \sum (y_i - \hat{y}_i)^2 = \min \sum (y_i - \beta_0 - \beta_1 x_i)^2 = \min (\text{SSE})$$

用最小平方法求得迴歸係數

求滿足 $\min \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \min\,(\text{SSE})$ 的 $\hat{\beta}_0$ ， $\hat{\beta}_1$ 的值

可利用微積分中一階偏導數的方法求得（令其值為0），所以：

$$\begin{cases} \dfrac{\partial \text{SSE}}{\partial \hat{\beta}_0} = 0 \\ \dfrac{\partial \text{SSE}}{\partial \hat{\beta}_1} = 0 \end{cases} \rightarrow \begin{cases} \sum 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-1) = 0 \\ \sum 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-x_i) = 0 \end{cases} \rightarrow \begin{cases} \sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i \quad \ldots(1) \\ \sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 \end{cases}$$

$$\rightarrow \begin{cases} \hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \dfrac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \dfrac{S_{xy}}{S_x^2} \\ \hat{\beta}_0 = \dfrac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} \end{cases}$$

其中 S_{xy} 為X與Y的共變異數
S_x^2 為樣本資料X所求得的變異數

實務上， $\hat{\beta}_1$比較好求與好記，所以會先計算出，從式(1)中可推得： $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ ，所以計算出 $\bar{x}$ 、 $\bar{y}$ 後，即可求得$\hat{\beta}_0$。

最小平方法求得迴歸係數的特質

1. $\hat{\beta}_0$，$\hat{\beta}_1$ 為 β_0，β_1 的不偏估計量，亦即E（ $\hat{\beta}_0$ ）= β_0 ；
   E（ $\hat{\beta}_1$ ）= β_1 。

2. $\hat{\beta}_0$，$\hat{\beta}_1$ 的變異數與母體變異數 σ^2 有下列關係：

   Var（ $\hat{\beta}_0$ ）= $\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\sigma^2$ ；　Var（ $\hat{\beta}_1$ ）= $\frac{\sigma^2}{\sum(x_i-\bar{x})^2}$ 。

3. $\hat{\beta}_0$，$\hat{\beta}_1$ 成常態分配：

   $\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\sigma^2\right)$ ；　$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right)$ 。

4. $\hat{\beta}_0$，$\hat{\beta}_1$ 均為最佳線性不偏估計量。

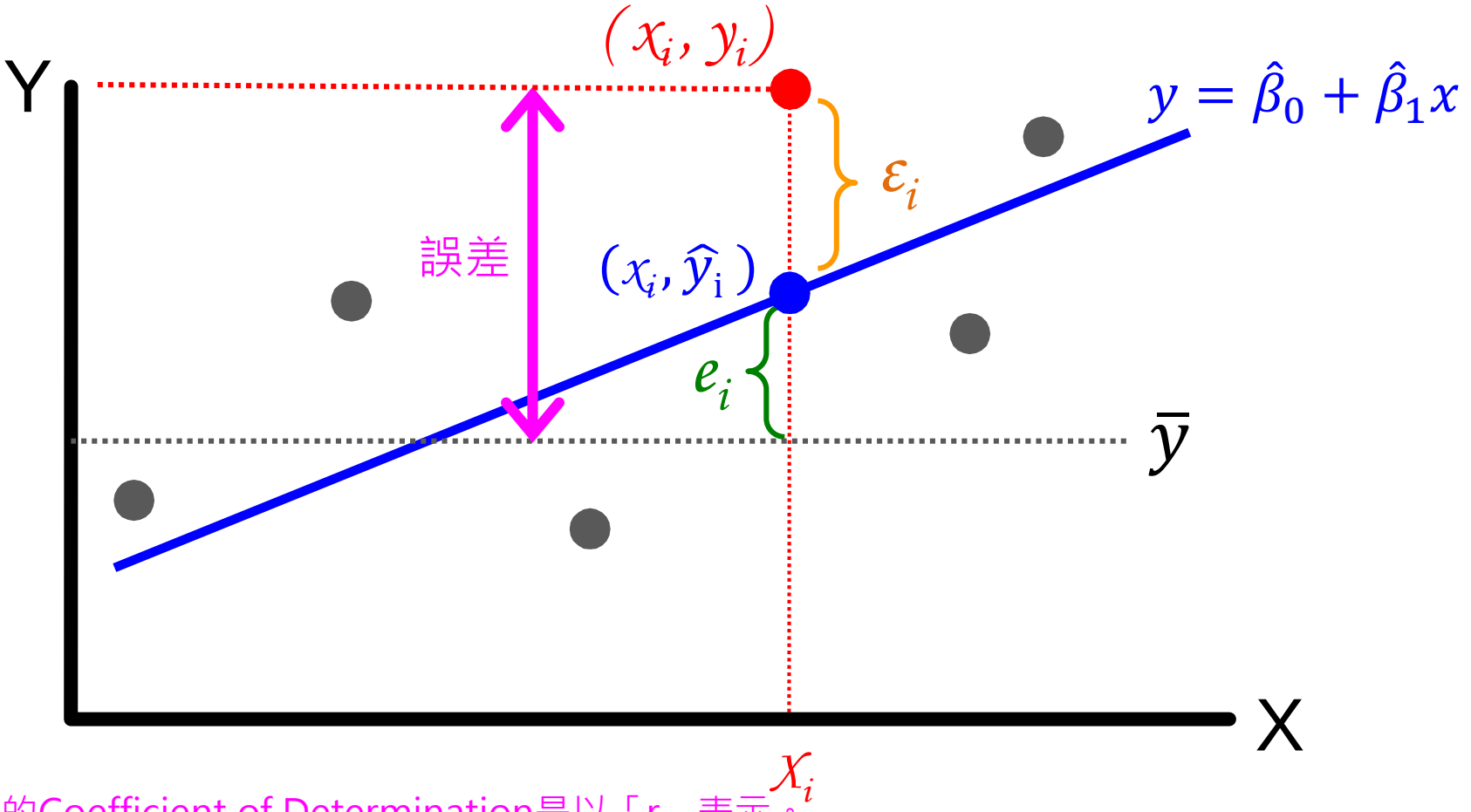
迴歸模型之
配適度檢定1

衡量迴歸方程式的解釋能力(判定係數)
也客觀檢定其適合度(F檢定)

判定係數(Coefficient of Determination)
~ 衡量迴歸方程式的解釋能力

$$y_i - \bar{y} \;=\; (\hat{y}_i - \bar{y}) \;+\; (y_i - \hat{y}_i)$$

誤差　＝可解釋誤差＋不可解釋誤差（隨機誤差）



※課本的Coefficient of Determination是以「r」表示。

判定係數的推導

將所有「誤差」平方後加總：

$$\sum (y_i - \bar{y})^2 = \sum \left((\hat{y}_i - \bar{y}) \ + \ (y_i - \hat{y}_i) \right)^2$$

➔ $$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

總變異	可解釋變異	隨機變異
SST	SSR	SSE

$$SST = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = (n-1)S_y^2 = n\,\hat{\sigma}_y^2$$

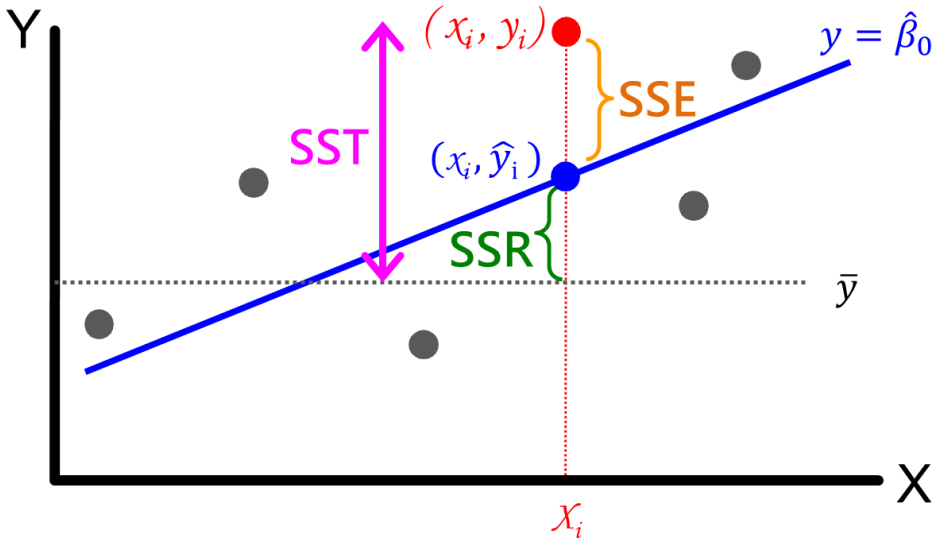
$$SSR = \sum (\hat{y}_i - \bar{y})^2 = \sum \left(\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_i - \bar{y} \right)^2 = \sum \left(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \hat{x}_i - \bar{y} \right)^2$$

$$= \hat{\beta}_1^2 \sum (\hat{x}_i - \bar{x})^2 = \hat{\beta}_1^2 \sum \hat{x}_i^2 - n\bar{x}^2$$

$$= \hat{\beta}_1^2 (n-1) \frac{1}{n-1} \sum (\hat{x}_i - \bar{x})^2 = \hat{\beta}_1^2 (n-1) S_x^2 = \hat{\beta}_1^2 \, n \, \hat{\sigma}_x^2$$

$$SSE = SST - SSR$$

判定係數的推導



適合度：希望觀測的值都落在迴歸線上。所以SSE越小越好；SSR越大越好。

判定係數 $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

➔ $R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \hat{x}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\hat{\beta}_1^2 \sum(\hat{x}_i - \bar{x})^2}{\sum(y_i - \bar{y})^2}$

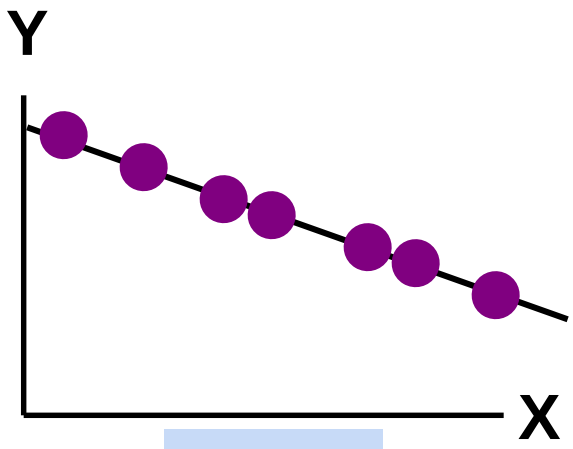
$= \frac{\hat{\beta}_1^2 (\sum x_i^2 - n\bar{x}^2)}{\sum y_i^2 - n\bar{y}^2} = \frac{\hat{\beta}_1^2 \frac{1}{n-1}\sum(\hat{x}_i - \bar{x})^2}{\frac{1}{n-1}\sum(y_i - \bar{y})^2} = \left(\hat{\beta}_1 \times \frac{S_x}{S_y}\right)^2 = \hat{\beta}_1^2 \frac{S_x^2}{S_y^2} = \frac{S_{\hat{y}}^2}{S_y^2}$

R^2越大，表示迴歸模型的解釋能力越強，配適度越大

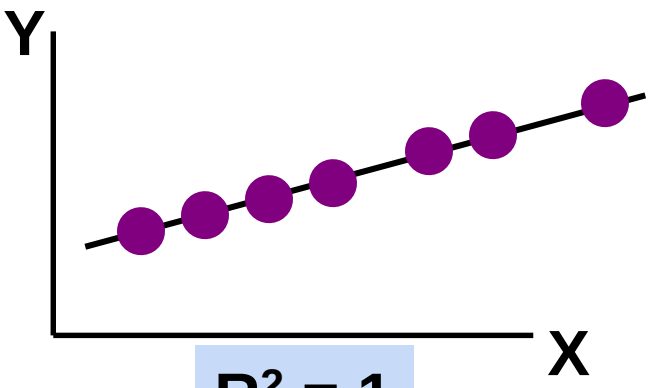
※課本的Coefficient of Determination是以「r」表示。

Examples of Approximate R^2 Values

DCOVA



$R^2 = 1$



$R^2 = 1$

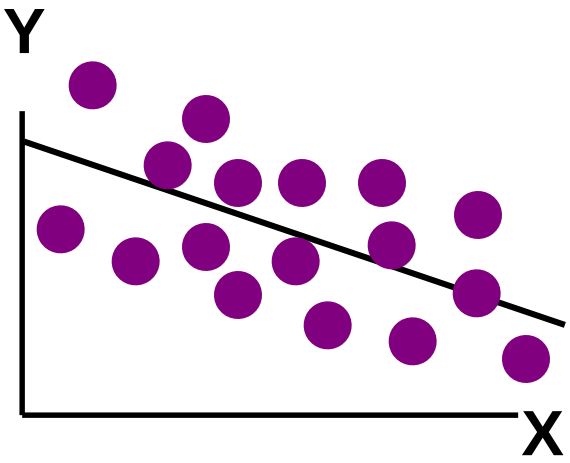
Perfect linear relationship between X and Y:

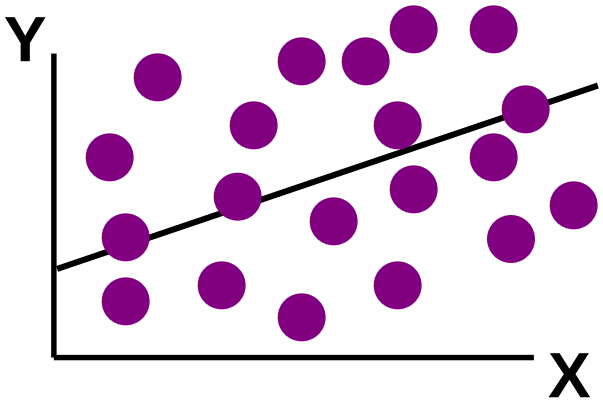
100% of the variation in Y is explained by variation in X

※課本的Coefficient of Determination是以「r」表示。

Examples of Approximate R^2 Values

DCOVA





$0 < R^2 < 1$

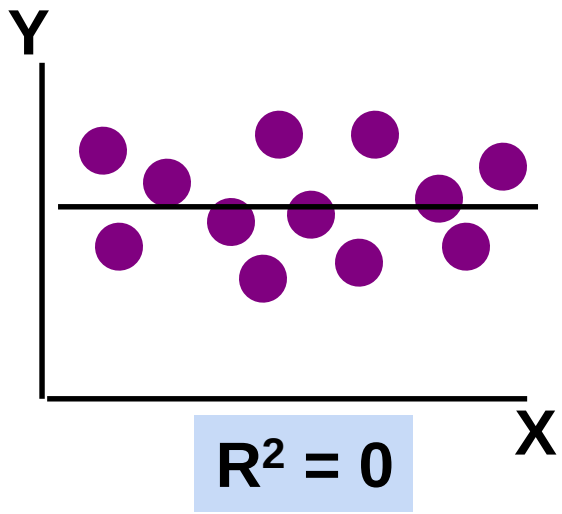
Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X

※課本的Coefficient of Determination是以「r」表示。

Examples of Approximate R^2 Values

DCOVA



$R^2 = 0$

$R^2 = 0$

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)

※課本的Coefficient of Determination是以「r」表示。

檢定迴歸模型的適合度

迴歸模型的適合度會受到樣本數與自變數數目所影響，因此僅憑判定係數大小來衡量迴歸模型的適合度還不夠，較客觀的方式，還可以透過假設檢定方式先來決定是否有顯著性的證明。

有關迴歸模型的檢定，可以利用F檢定來做：

$$\begin{cases} H_0 : \beta_1 = 0 \text{ (迴歸方程式不具有解釋力)} \\ H_1 : \beta_1 \neq 0 \text{ (迴歸方程式具有解釋力)} \end{cases}$$ (或x不可解釋y)

利用變異數分析的原理，將迴歸的總變異差解成：

SST = SSR(可解釋變異) + SSE(隨機變異)

迴歸模型的 F 檢定

變異來源	平方和	自由度	平均平方和	F值
迴歸	SSR	1	MSR=SSR/1	
隨機	SSE	n-2	MSE=SSE/(n-2)	$F^*=\frac{MSR}{MSE}$
總和	SST	n-1		

決策法則： $F^* > F_{\alpha, 1, n-2}$ ，則拒絕H_0

MSR越大（SSR越大），越容易拒絕虛無假設H_0，也就是迴歸方程式具有解釋力。

迴歸之判定係數R^2與F檢定

MSR越大（SSR越大），越容易拒絕虛無假設H_0，表示R^2也會越大，而且從兩邊的式子關係，兩者具有某種關連性：

$$R^2 = \frac{SSR}{SST}$$

$$F^* = \frac{MSR}{MSE} = \frac{(SSR/1)}{(SSE/n-2)}$$

$$F^* = \frac{MSR}{MSE} = \frac{(SSR/1)}{(SSE/n-2)} = \frac{(n-2)\frac{SSE}{SST}}{\frac{SSE}{SST}} = \frac{(n-2)R^2}{\frac{SST-SSR}{SST}} = \frac{(n-2)R^2}{1-R^2}$$

$$= \frac{R^2/1}{(1-R^2)/(n-2)}$$

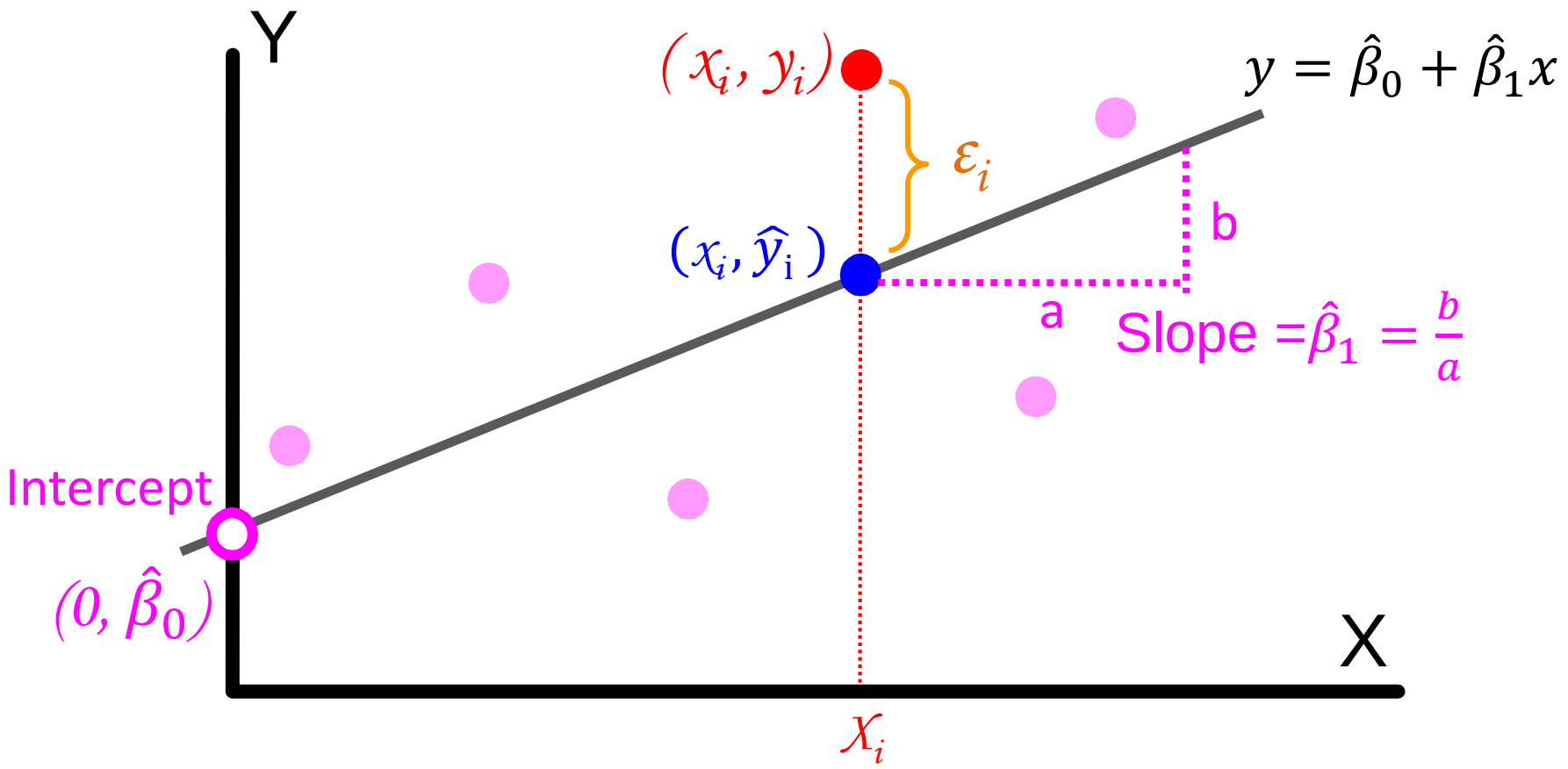
雖然R^2越大也越容易拒絕虛無假設H_0，但從式子中可知，若是樣本數多寡或是抽樣的偏差，也會造成R^2很大，可是F檢定卻不具解釋力（接受H_0）；或是R^2很小，F檢定卻具有解釋力。

迴歸模型之配適度檢定2

檢定斜率與截距的適合度

斜率項與截距項的檢定

除了用誤差（ε_i）最小平方法的作為檢定迴歸適合度外，也可以利用迴歸方程式的斜率（$\hat{\beta}_1$）與截距（$\hat{\beta}_0$）來作為判斷是否迴歸方程式具有解釋力。



斜率項 β_1 的檢定

雙尾檢定

$$\begin{cases} H_0 : \beta_1 = 0 \text{ (迴歸方程式不具有解釋力)} \\ H_1 : \beta_1 \neq 0 \text{ (迴歸方程式具有解釋力)} \end{cases}$$

已知 $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$ ，因母體變異數σ^2未知，依照

假設檢定方法，可用 t分配做檢定：

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}}}$$

決策法則： $|t^*| > t_{\alpha/2,\, n-2}$ 時，則拒絕虛無假設 H_0；
表示迴歸方程式具適配度，或自變數X可以解釋變數Y

斜率項 β_1 的信賴區間

也因為 $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$ ，且母體變異數σ^2未知，可用t分配做區間估計（信賴水準1-α）：

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}}}$$

以t分配做區間估計，所以1-α的信賴區間為：

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}}$$

斜率項 β_1 的檢定

右尾檢定

$$\begin{cases} H_0 : \beta_1 \leq 0 \text{ (自變數對依變數不具正向影響力)} \\ H_1 : \beta_1 > 0 \text{ (自變數對依變數具正向影響力)} \end{cases}$$

已知 $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$ ，母體變異數σ^2未知，依照假設檢定方法，可用 t分配做檢定：

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}}}$$

決策法則： $t^* > t_{\alpha, n-2}$ 時，則拒絕虛無假設 H_0；表示迴歸方程式具適配度，或自變數X可以解釋變數Y

斜率項 β_1 的檢定

左尾檢定

$$\begin{cases} H_0 : \beta_1 \geq 0 \text{ (自變數對依變數不具負向影響力)} \\ H_1 : \beta_1 < 0 \text{ (自變數對依變數具負向影響力)} \end{cases}$$

已知 $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$ ，母體變異數σ^2未知，依照假設檢定方法，可用 t分配做檢定：

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}}}$$

決策法則： $t^* < -t_{\alpha, n-2}$ 時，則拒絕虛無假設 H_0；表示自變數X對依變數Y有顯著的負面影響。

截距項 β_0的檢定

雙尾檢定

$$\begin{cases} \mathrm{H}_0 : \beta_0 = 0 \text{ (迴歸方程式沒有通過原點)} \\ \mathrm{H}_1 : \beta_0 \neq 0 \text{ (迴歸方程式沒有通過原點)} \end{cases}$$

已知 $\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\sigma^2\right)$ ，母體變異數σ^2未知，依照假設檢定方法，可用 t分配做檢定：

$$\mathrm{t}^* = \frac{\hat{\beta}_0 - \beta_0}{\mathrm{S}_{\hat{\beta}_0}} = \frac{\hat{\beta}_0}{\sqrt{\frac{\sum x_i^2}{n} \times \frac{MSE}{\sum(x_i-\bar{x})^2}}}$$

決策法則： $|\mathrm{t}^*| > \mathrm{t}_{\alpha/2,\, n-2}$ 時，則拒絕虛無假設 H_0；表示迴歸方程式沒有通過原點。

斜率項 β_0的信賴區間

也因為 $\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\,\sigma^2\right)$ ，母體變異數σ^2未知，可用t分配做區間估計（信賴水準1-α）：

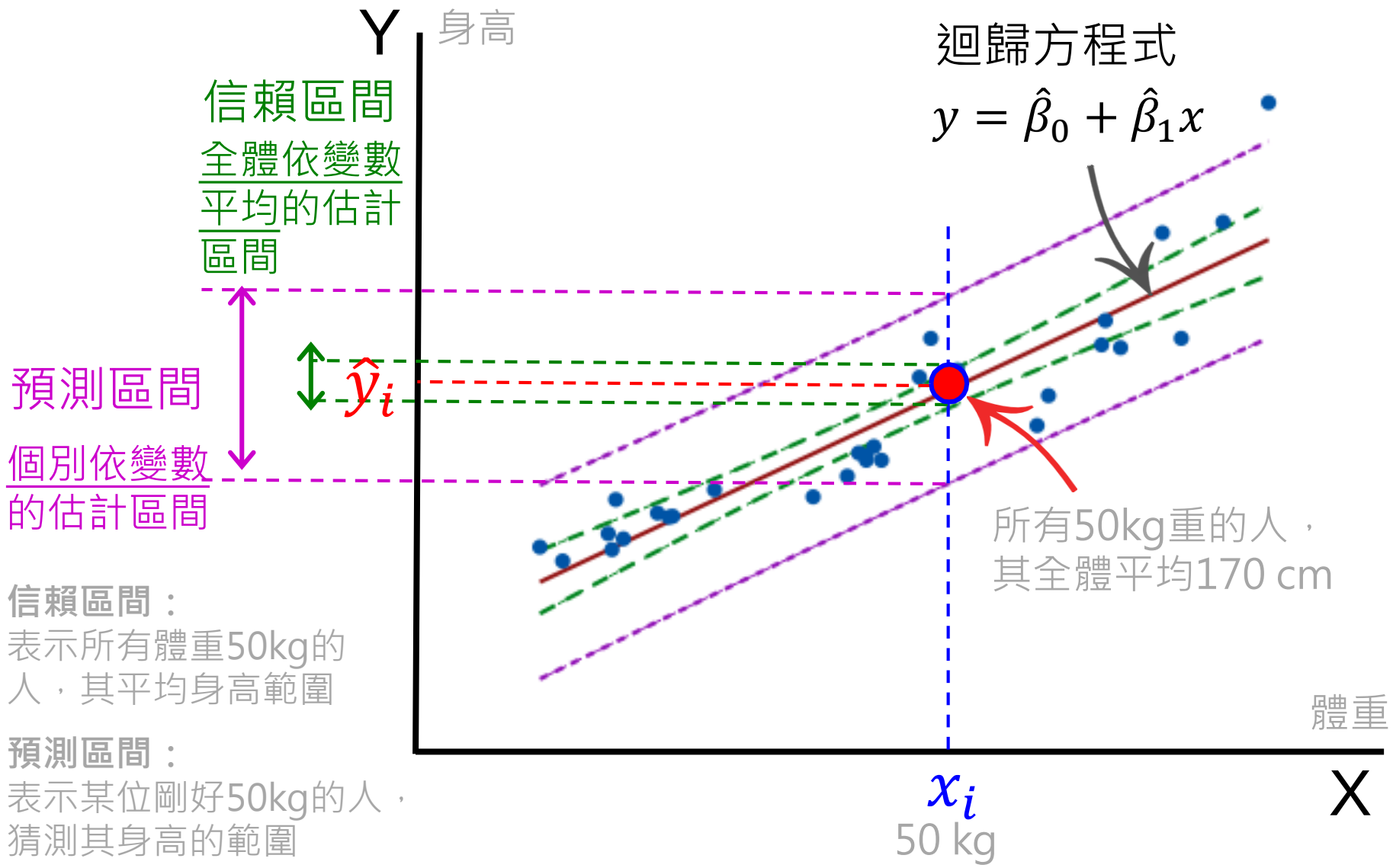
$$t^* = \frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}} = \frac{\hat{\beta}_0}{\sqrt{\frac{\sum x_i^2}{n} \times \frac{MSE}{\sum(x_i-\bar{x})^2}}}$$

以t分配做區間估計，所以1-α的信賴區間為：

$$\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\sum x_i^2}{n} \times \frac{MSE}{\sum(x_i-\bar{x})^2}} \leq \beta_0 \leq \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\sum x_i^2}{n} \times \frac{MSE}{\sum(x_i-\bar{x})^2}}$$

迴歸模型之信賴區間

迴歸方程式的信賴區間與預測區間

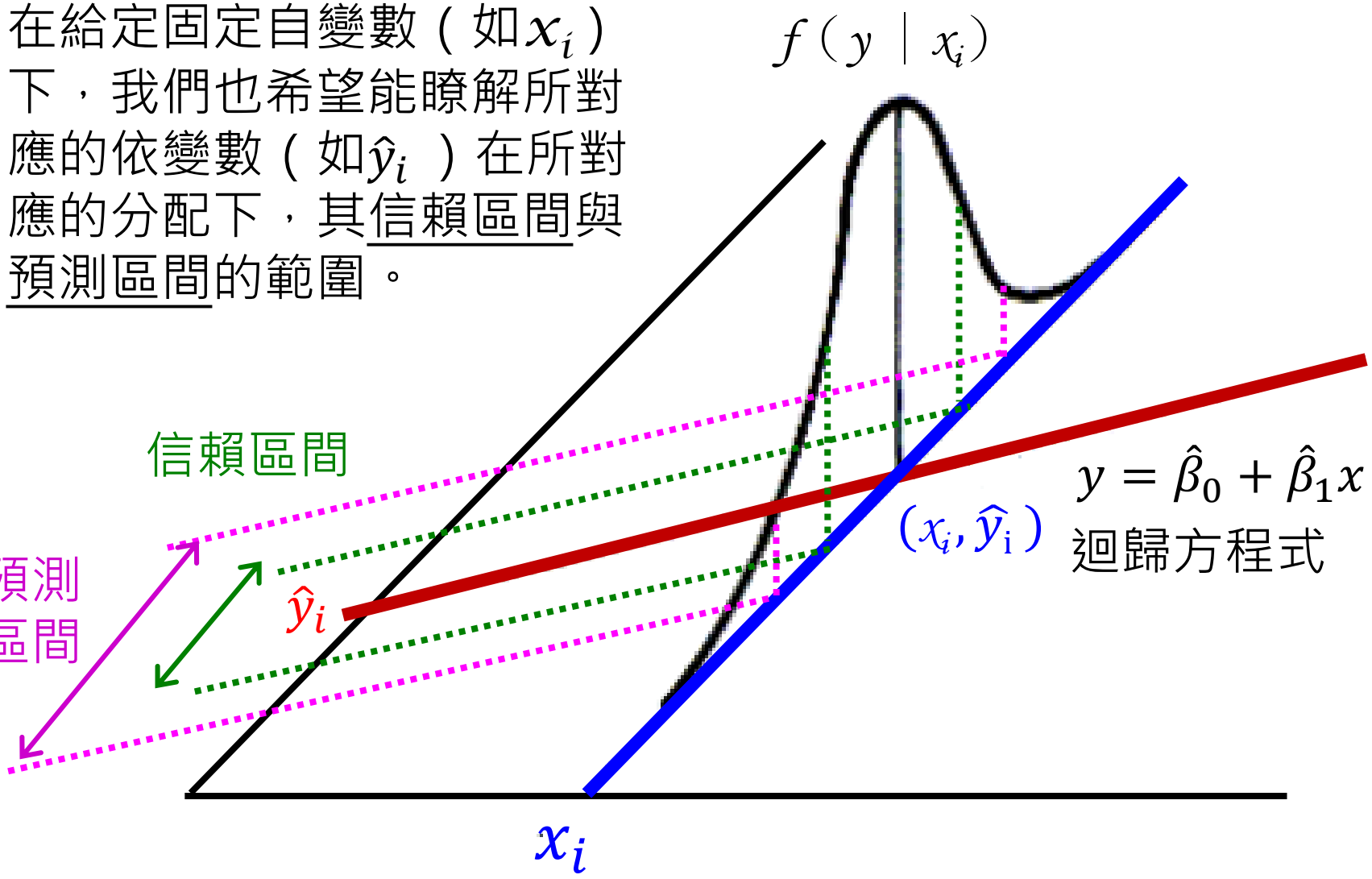


信賴區間：
表示所有體重50kg的人，其平均身高範圍

預測區間：
表示某位剛好50kg的人，猜測其身高的範圍

信賴區間與預測區間示意圖

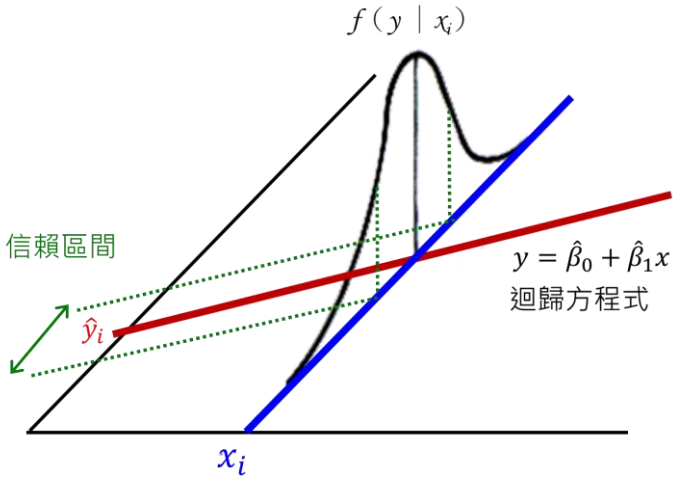
在給定固定自變數（如x_i）下，我們也希望能瞭解所對應的依變數（如$\hat{y}_i$）在所對應的分配下，其信賴區間與預測區間的範圍。



全體依變數平均數的信賴區間

所謂「迴歸方程式」的信賴區間，是指在給定自變數（x）的條件下，母體迴歸線依變數y期望值的信賴區間。

以E（y|x_i）表示在給定自變數x_i下，迴歸線的期望值。因母體變異數未知，根據~



$$\hat{y}_i \sim N\left(E(y|x_i), \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}\right]\sigma^2\right)$$

信賴區間 = 樣本統計量 $\pm t_{\frac{\alpha}{2},df}$ ×標準誤　　所以：

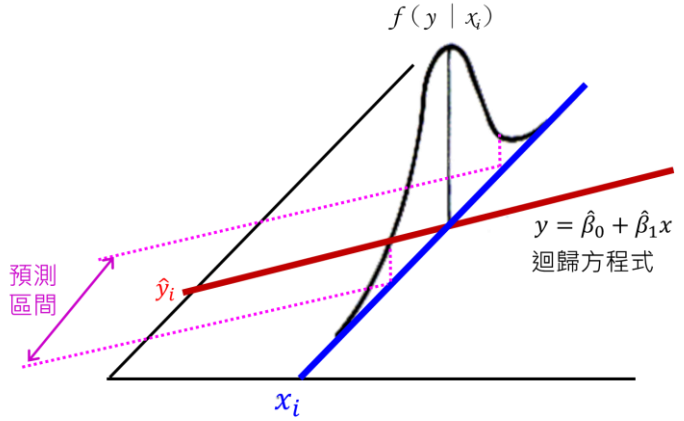
$$E(y|x_i) = \hat{y}_i \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}\right]}$$

(以MSE取代σ^2)

其中　$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$　，　x_i為所給定的自變數值

個別依變數的信賴區間（預測區間）

所謂個別依變數的信賴區間，是指在給定自變數（x）的條件下，預測依變數y的信賴區間。



以y_i表示在給定自變數x_i下所對應的依變數，因母體變異數未知，依信賴區間概念，可得~

$$(y_i - \hat{y}_i) \sim N(\mathrm{E}(y_i - \hat{y}_i), \mathrm{Var}(y_i - \hat{y}_i))$$

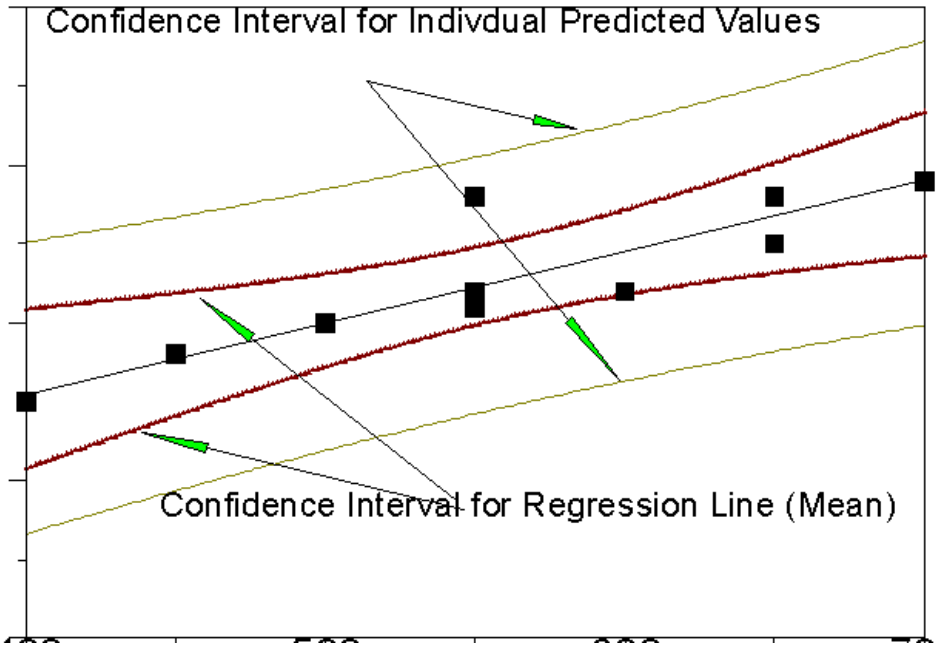
$$(y_i - \hat{y}_i) \sim N\left(0, \left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}\right] \sigma^2\right)$$

$$y_i = \hat{y}_i \pm t_{\frac{\alpha}{2}, n-2} \sqrt{MSE\left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}\right]}$$

其中 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ ， x_i為所給定的自變數值

影響信賴區間長度的因素

$$E(y|x_i) = \hat{y}_i \pm t_{\frac{\alpha}{2}, n-2} \sqrt{MSE\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}\right]}$$



從信賴區間公式可知：

1.顯著水準越大，信賴區間越短

2.MSE越大，信賴區間越長

3. x_i 離 $\bar{x}$ 越遠，信賴區間越長

4.樣本數 n 越大，信賴區間越短

5. $E(y|x_i)$的信賴區間比 y_i 的小，當x_i 越接近 $\bar{x}$ 則信賴區間越短，表示越接近平均數越準確，所以使用迴歸方程式時，不宜做大範圍的預測。

圖片來源：http://faculty.cas.usf.edu/mbrannick/regression/Prediction.html

相關分析(Correlation Analysis)

~ 迴歸模型之配適度檢定3

- 衡量兩隨機變數相關程度與變化的方向趨勢

相關分析(Correlation Analysis)

- 相關分析只能判斷兩變數間是否相關：正相關、負相關或無相關；也就是可以衡量兩變數具有線性關係強度的參考指標但無法判定是否具有線性關係的能力。
- 除了瞭解兩變數間的關係外，通常也用於進行迴歸分析前的初步判定。若是關係薄弱，即便找出迴歸方程式，其解釋與預測能力也不具代表性。
- 進行兩變數間的相關分析，是利用兩變數平均數間期望值與標準差來作為比較，所以母體相關係數可定義為：

$$\rho_{xy} = \frac{E\left[(x-\mu_x)(y-\mu_y)\right]}{\sigma_x\sigma_y} = \frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum(x_i-\mu_x)^2}\cdot\sqrt{\sum(y_i-\mu_y)^2}}$$

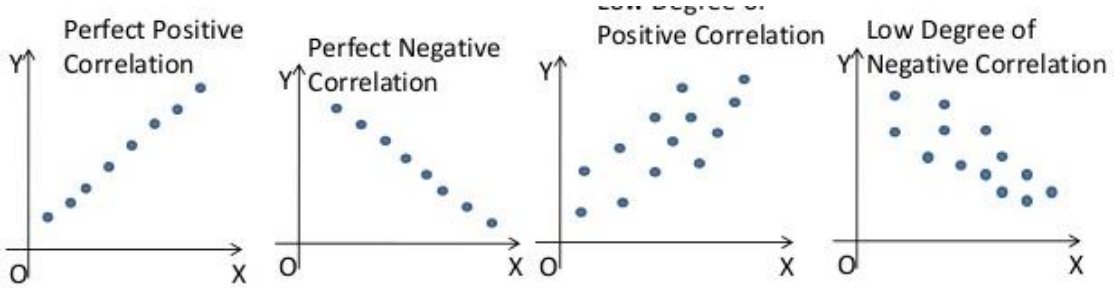
$$= \frac{Cov(x,y)}{\sigma_x\sigma_y} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \qquad \text{其中，} -1 \le \rho_{xy} \le 1$$

母體相關係數(correlation coefficient)性質

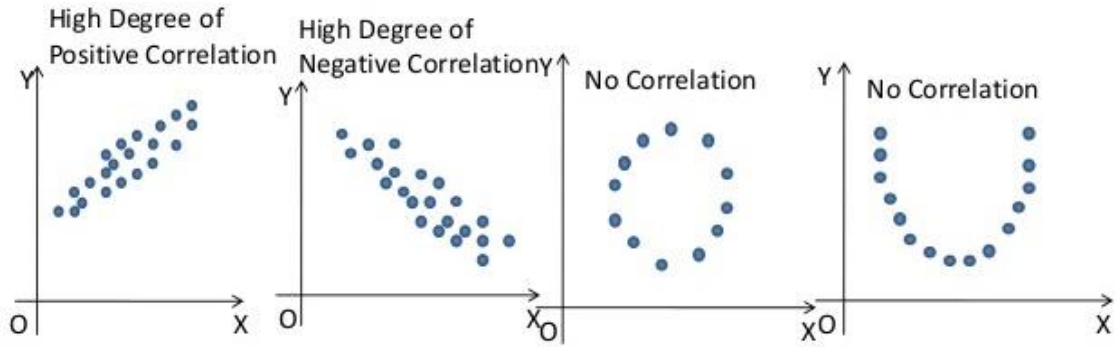
1.若x、y獨立，則$\rho_{xy} = 0$。

2. $\rho_{xy} = 0$ 時，表示x、y不具線性關係，但不一定獨立（例如x、y可能是曲線關係）。

3.若$\rho_{xy} = 1$，稱為「完全正相關」。



4.若$\rho_{xy} = -1$，稱為「完全負相關」。



樣本相關係數的估計

利用樣本資料來估算母體的相關係數：樣本相關係數

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{S_x S_y}$$

1. r_{xy} 為 ρ_{xy} 的一致估計量。

2. $\rho_{xy} = 0$ 表x、y不具線性相關。

3.判定係數等於相關係數平方，即 $R^2 = r_{xy}^2$ 。

4.即 $r_{xy} = \pm\sqrt{R^2}$（正負值與迴歸係數$\hat{\beta}_1$同號）。

相關係數與判定係數間的關係

$\because \hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$ 前面對斜率的推導而得　　$r_{xy} = \frac{S_{xy}}{S_x S_y}$ 定義

$$\Rightarrow \hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y} \times \frac{S_y}{S_x} = r_{xy} \cdot \frac{S_y}{S_x}$$

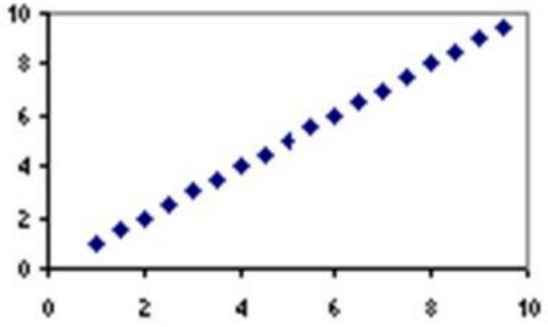
$$\Rightarrow r_{xy} = \hat{\beta}_1 \cdot \frac{S_x}{S_y}$$

另外已知判定係數 $R^2 = \hat{\beta}_1^2 \frac{S_x^2}{S_y^2}$

$$\Rightarrow R^2 = \left(\hat{\beta}_1 \times \frac{S_x}{S_y}\right)^2 = r_{xy}^2$$

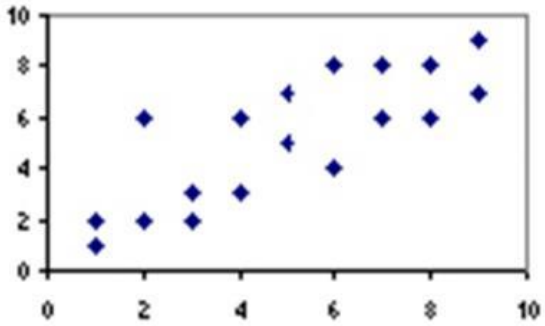
$\Rightarrow r_{xy} = \pm\sqrt{R^2}$　其正負號與斜率項$\hat{\beta}_1$相同

相關係數與對應的圖示關係



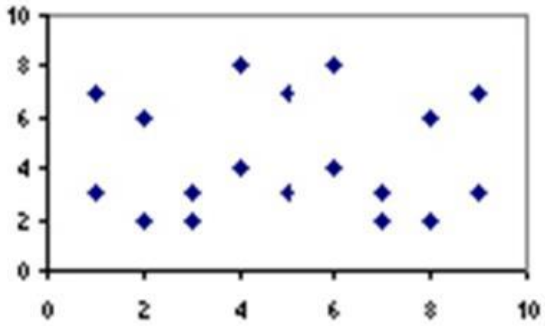
Maximum positive correlation

(r = 1.0)



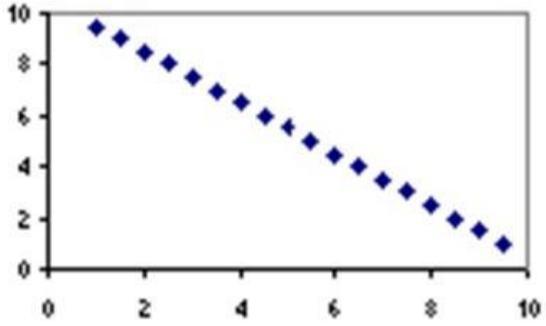
Strong positive correlation

(r = 0.80)



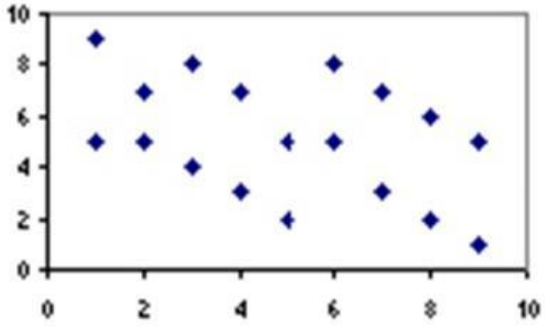
Zero correlation

(r = 0)



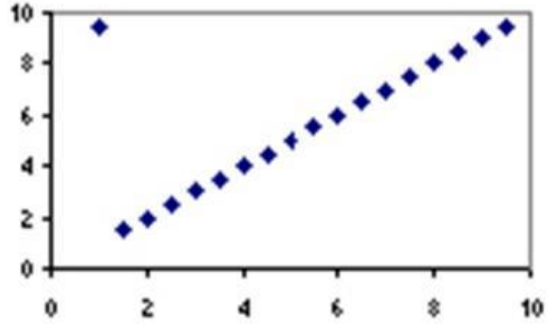
Maximum negative correlation

(r = -1.0)



Moderate negative correlation

(r = -0.43)



Strong correlation & outlier

(r = 0.71)

ρ_{xy}的統計推論

同樣地，我們利用樣本統計量 r_{xy} 來估算母體相關係數 ρ_{xy} ，也可以用來檢定兩變數母體間所抽取的樣本是否有顯著代表性。

假設檢定 (雙尾檢定)

$$\begin{cases} H_0 : \rho_{xy} = 0 \\ H_1 : \rho_{xy} \neq 0 \end{cases}$$

檢定統計量

$$t^* = \frac{r_{xy}}{\sqrt{\frac{1 - r_{xy}^2}{n-2}}}$$

可利用斜率檢定推導而得

$$t^* = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}}}$$

決策法則

當 $|t^*| > t_{\frac{\alpha}{2},n-2}$ 時，拒絕虛無假設H_0。

相同概念，也可以進行「右尾檢定」或「左尾檢定」。

The End