Lecture 09. Regression Analysis

迴歸分析

Introduction

○ 前面所談的，大部份是 **"分佈"**（**distribution**）。比如，一組資料，來自某一個分佈 $F(\theta)$，θ 是 parameter (可以是一個 "向量"！) 我們可以根據 data，來對 $F(\theta)$ 做一個描述（description），比如 mean 是多少，variance 是多少，…等等。

○ 以 Normal distribution 為例：$F(\theta) \equiv N(\mu, \sigma^2)$，參數 $\theta = (\mu, \sigma^2)$，是 2-dimensional。你關心的是參數向量 $\theta = ?$，即

$$mean\ (\mu) = ?, \qquad \operatorname{var}iance\ (\sigma^2) = ?$$

或者，你給出一個 confidence interval；

或者，你檢定 $\theta = \theta_0$（某個 θ_0 值）是否成立。

這些主要都在想要了解 $F(\bullet) = ?$

○ 再者，在 ANOVA 中，不管是 one-way 或 two-way，背後還是關心幾組的 distribution 是不是一樣。（要看 distribution 是否一樣，最簡單的看法是：先看 mean 一不一樣！）

但是在 ANOVA 中，還可以引出一個問題：即不同的組（不同的 treatment），甚至不同的 block，若有不同的 mean，即是代表 treatment（或甚至 block）與 outcome variable 之間有一個 "關係" 存在。不同的 treatment，引出不同的 outcome 平均值；不同的 block，引出不同的平均 outcome,……。

○ 所以，以下便談談 **"關係"**（**relationship**）。

Regression

○ 兩個變數，比如 X 和 Y 之間若有“不同的 X 值對應出不同的 Y 值”，這種情形（可能）存在，則我們常用一個方程式去描述這種對應關係：$Y = f(x)$。

○ 統計上，你看到的多是具有 random error 存在的情形，這個 random error 怎麼來的，有多大？這種問題常無法回答，無論如何，上面的關係式常是以如下的形式呈現：

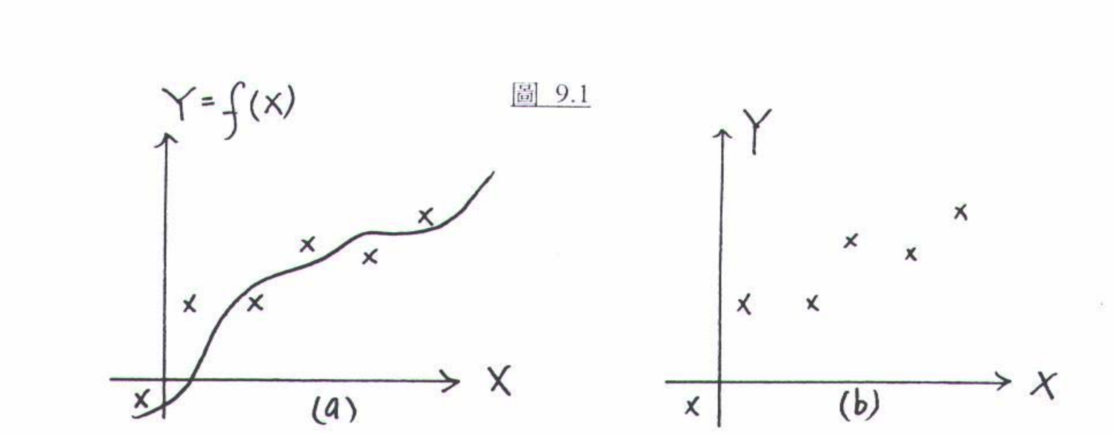
$$Y = f(x) + \varepsilon$$

○ ε 即是一個 random error。

○ 我們的目標，便是根據 data 來猜測 $f(\bullet) = ?$

○ 這時 data 是什麼呢？ Data 就是（Y_1, X_1），（Y_2, X_2），…，（Y_n, X_n）。根據這 n 個 data(sample size = n)，你能正確的描述出 $f(\bullet)$ 嗎？很困難，因為 $f(\bullet)$ 是一個函數，你要正確地描述它，你必須把所有可能的 X 值所對應出來的 $Y = f(x)$ 都描述正確，但這幾乎是不可能的！

○ 例如，$f(\bullet)$ 如下圖



實線部份是 $f(\bullet)$，而 data 只是 6 個點（”x”的部份），你只看到如上圖(b)的情況。根據圖(b)，你怎麼可能把 $f(\bullet)$ 從頭到尾都猜對？

[continued]

(i) 由上圖(b) ，你可以有兩個猜測法（比如說！）：

(a) 第一，你可以猜 Y 與 X 的關係是：

$$y = \sqrt{a + bx}$$ （拋物線）

然後根據 data 去 estimate a 和 b (parameters)

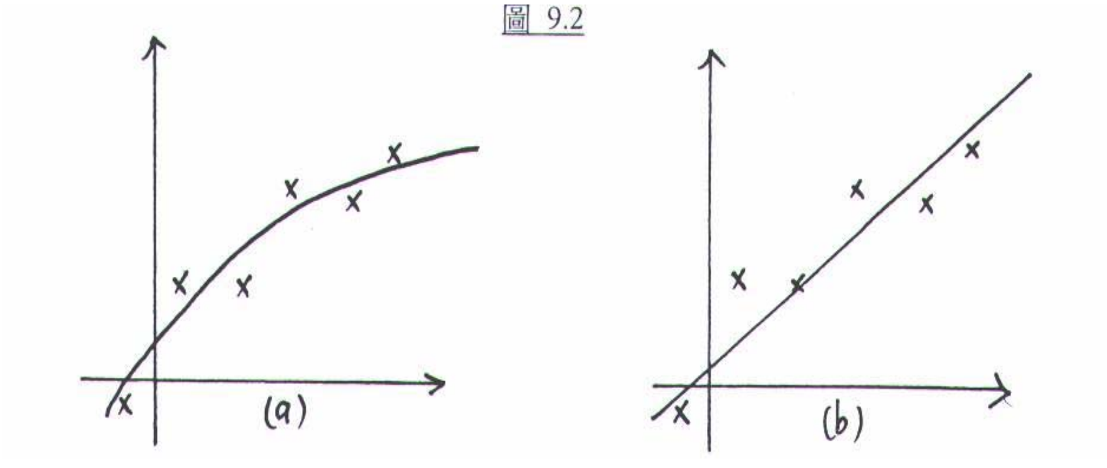
(b) 第二，你可以猜 Y 與 X 的關係是：

$$y = a + bx$$ （直線）

同樣根據 data 去 estimate a & b。

(ii) 根據 (ii)(a) & (ii)(b)的 estimates，你可能可以得到對 Y 和 X 之關係之兩種估計，示下圖 (a) & (b)

圖 9.2



[continued]

(iii) 在(ii)中，你等於把問題簡化了。即，你雖然不知道真正的答案 $f(\bullet)$，但你企圖把問題模式化（即給予一個模式（model））；或者說，你把 $f(\bullet)$ 參數化（parameterize），以粗略地求得 Y 與 X 之關係。

(iv) 求得這個參數化後之 $Y = f(x)$ 關係有一個好處，即你可以簡單地對不同的 X 值預測其 Y 值（但須注意其“有效範圍”及“可能誤差”！）

(v) (ii)(b)之模式 $y = a + bx$，即稱為一個線性模式（linear model 或 linear regression model）。

[continued]

(vi) 你的 data 可能更複雜：(Y_1, X_1, Z_1)，(Y_2, X_2, Z_2)，⋯，(Y_n, X_n, Z_n)，

而你建立如下之 model：

$$Y = a + bX + cZ$$

或者，再複雜一點：

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \ ,$$

而你的 data 乃是

$$\left.\begin{array}{llllll} y_1, & x_{11}, & x_{21}, & x_{31}, \cdots, & x_{p1} \\ y_2, & x_{12}, & x_{22}, & x_{32}, \cdots, & x_{p2} \\ \vdots & & & & \\ \vdots & & & & \\ y_n, & x_{1n}, & x_{2n}, & x_{3n}, \cdots, & x_{pn} \end{array}\right\} \text{共 n 個 data}$$

$Y = a + bX$ 的模式，叫一個 simple linear regression model，

$Y = \beta_0 + \beta_1 X_1 + \cdots\cdots + \beta_p X_p, \quad p \geq 2$，叫 **multiple linear regression model**。

[continued]

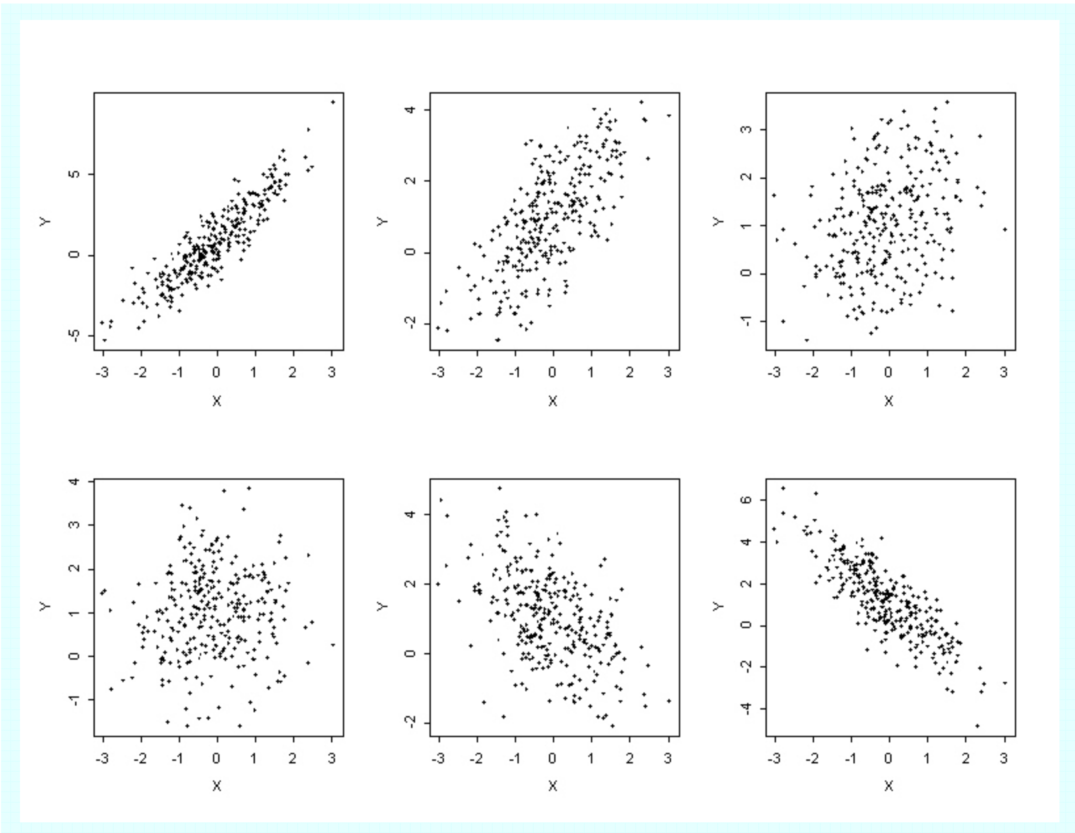
- **Correlation**

(i) $Y = a + bX$ ，是對 Y-X 關係的一種描述，其關心的是 Y 與 X 之相對變化（即 b 之大小）。

(ii)你也可以關心 Y-X 關係之“強度”。這種強度，一般用 correlation（相關性）或 correlation coefficient（相關係數）表示，其值一般介於 –1 和 +1 之間。

(iii) Several correlation type: (see below for bivariate-normal case).

A simulated illustration for correlation coefficient=**0.9, 0.7, 0.3, 0,-0.5,-0.8**; based on **300** pairs of observations.

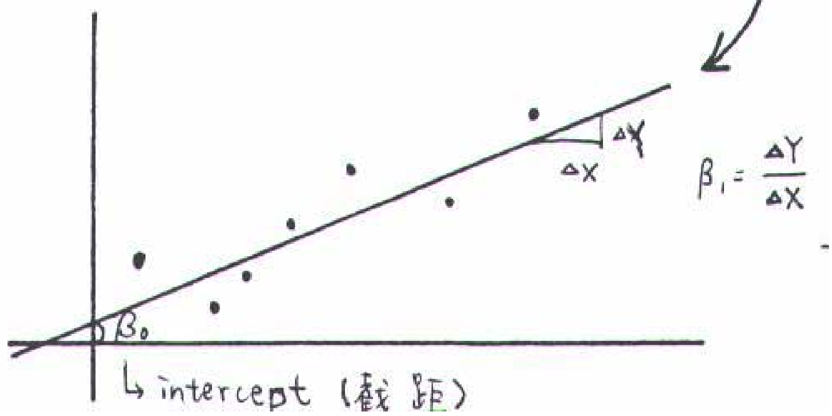


Regression Model

・Simple Linear Regression (Model)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$EY = \beta_0 + \beta_1 X \text{（或 } E(Y \mid X) = \beta_0 + \beta_1 X \text{）}$$



即<u>假設</u>真正的 model 為 $EY = \beta_0 + \beta_1 X$，或 $\mu_{Y|X} = \beta_0 + \beta_1 X$ 意即給定 X 時，<u>Y 的“期望值”</u>為 $\beta_0 + \beta_1 X$，而你看到的是 $\beta_0 + \beta_1 X + \varepsilon$，$\varepsilon$ 為 random error 且 $E(\varepsilon) = 0$ [**Q1**]。

X 稱為 regressor，或 independent variable。(常是可被控制的)

Y 稱為 response，或 dependent variable，或 outcome variable，……

（關於 **Q1**）

$E(\varepsilon)=0$ 是必須的；若 $E(\varepsilon)=a\neq 0$，(for some "a")，則 a 與 β_0 不可分辨，可以寫 $\beta_0^*=\beta_0+a$，則仍變成 $E(\varepsilon^*)=0$：

$$EY=\beta_0^*+\beta_1 X+\varepsilon^*$$

• Other assumptions

(i) X 的測量沒有誤差（without measurement error），若 X 的測量有誤差，則上面的 model 便叫做一個有 measurement error 的 regression model，或叫 measurement error model，或叫 error-in-variable (EIV) model，這時問題較複雜。

(ii)x 與 ε 沒有"關係"，或者說 $\mathrm{cov}(x,\ \varepsilon)=0$，即 x 與 ε 之 ***covariance = 0***，covariance 的定義為：

$$\begin{aligned}\mathrm{cov}(x,\varepsilon)&=E\{(x-Ex)(\varepsilon-E\varepsilon)\}\\&=E(x\varepsilon)-Ex\cdot E\varepsilon\end{aligned}$$

故若 $E\varepsilon=0$，則 $\mathrm{cov}(x,\varepsilon)=E(x\varepsilon)$

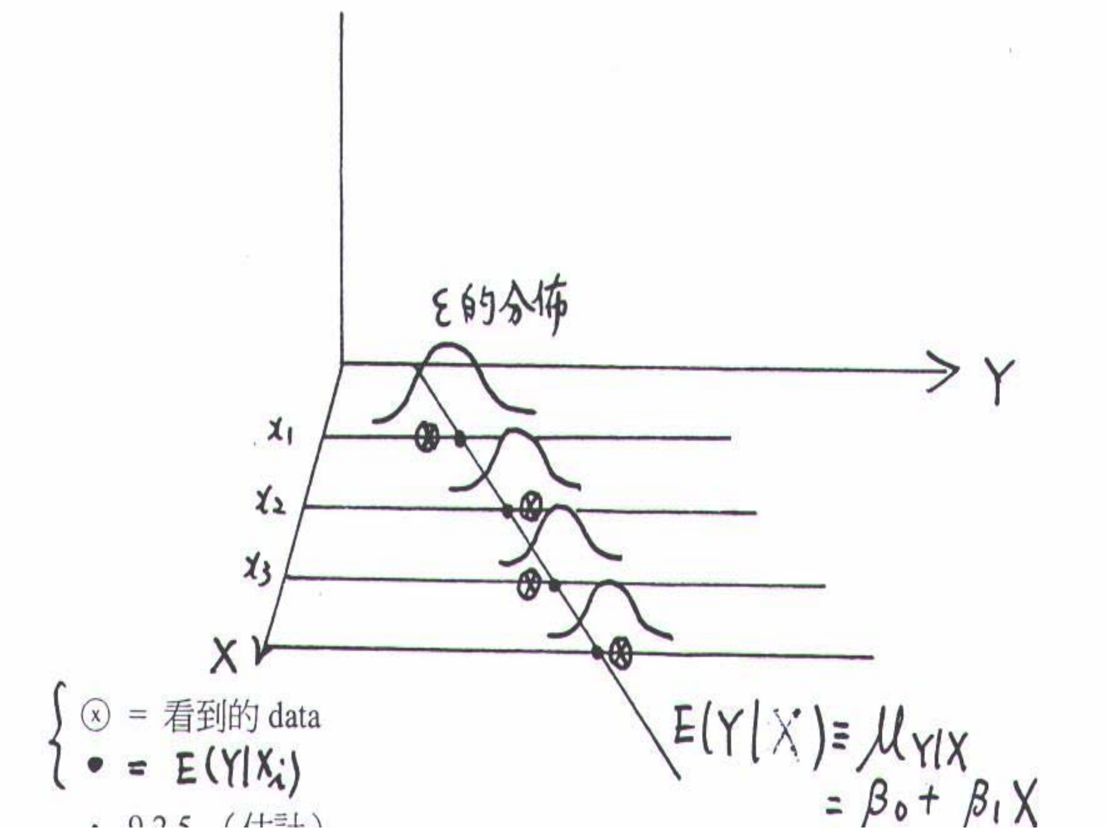
(iii) Y_1，Y_2，…，Y_n 彼此互相獨立，或

ε_1，ε_2，…，ε_n 彼此互相獨立。

(iv) ε_1，ε_2，…，ε_n 之 variance 均等於 σ^2（equal variance）

• (iv)' ε_1，ε_2，…，$\varepsilon_n \sim N(0,\ \sigma^2)$

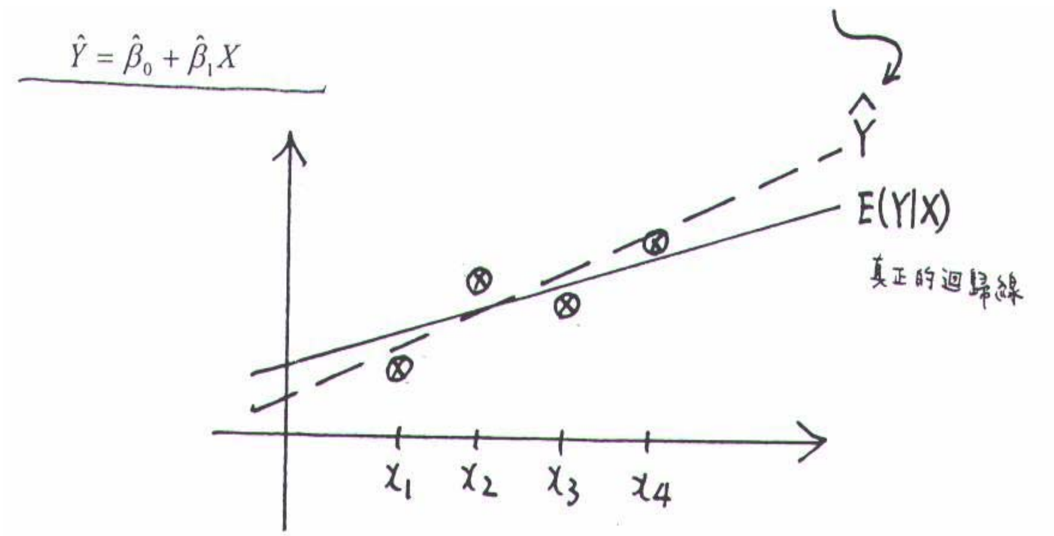
（圖形）



Estimation

根據 上圖 之 data（⊗），你可能得到一條如下之估計線：

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$



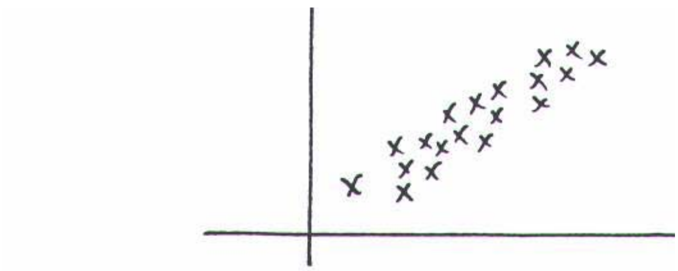
Estimation

・Model and Data

$$\begin{cases} Model & EY = \beta_0 + \beta_1 X \\ Data & (Y_1, X_1),\ (Y_2, X_2),\ \cdots,\ (Y_n, X_n) \end{cases}$$

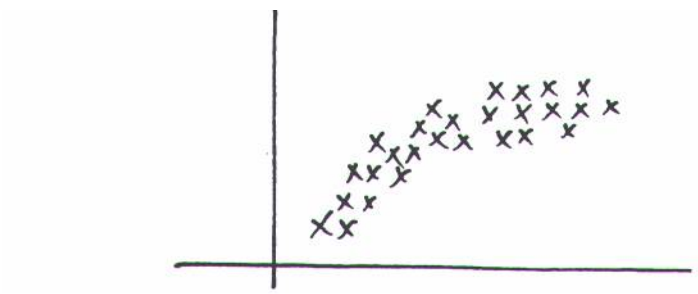
說明：Model 不是預先設下的，而是根據 data 的“長相”來決定的。

如果畫出 Y-X 的 scatter plot （散布圖），而得如下之“形式”。（pattern）：



你自然會覺得去 fit 一條直線（ $EY = \beta_0 + \beta_1 X$ ）是合理的。

但是，如果 Y-X（data）的 pattern 是這樣：



則你就該用別的 Model 了！

- **[Continued]**

Least Squares (LS) Method

前面所說的 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ 這條線是怎樣得到的呢？ 方法有很多種，其中一種最有名的叫最小平方法（least square method）：

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1,2,\cdots,n$$

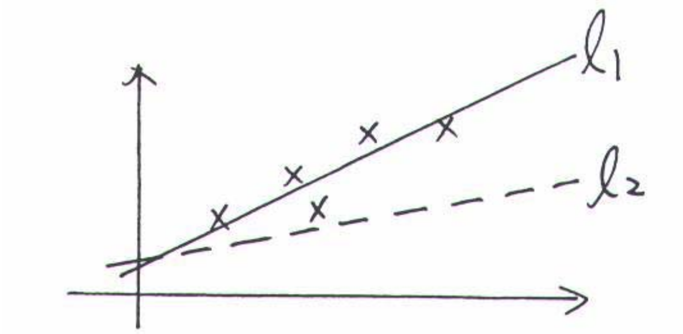
$$\rightarrow \quad \varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

$$\rightarrow \quad \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]^2 \equiv L$$

L 是一個“總誤差”（total error）的描述量。

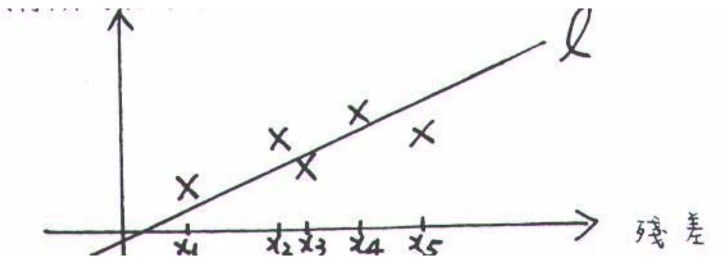
想法：

如果 data 長得如下圖所示：



你比較相信這些 data (“x”) 是從 l_1 這條線生成的（generate），還是從 l_2？（ANS: 當然是 l_1）

但 l_2 是很容易被排除的（相較於 l_1），因為它跟 data 差太多，很容易判別。如果有另一條 l_2’ 如下，則你要選 l_1 還是 l_2’ 呢？

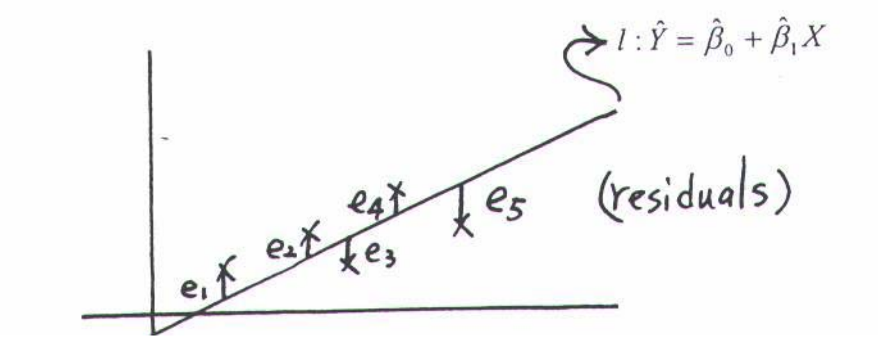


這時很難選了！

必須有一個“判據”（criterion）才行！

對於你所可能考慮的任一條線 l 而言，data 或觀測值（observations）與 l 的“縱垂距離”（可以有正負號），叫 **residual**，以 e 表示。

[continued]



把這些 residual 的平方加總起來得 residual sum of squares：

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left[Y_i - \hat{Y}_i\right]^2$$
$$= \sum_{i=1}^{n}\left[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right]^2$$

於是 least squares method 的想法便是考慮這樣的 criterion；你所得到的 $\hat{\beta}_0$, $\hat{\beta}_1$ 乃是所有可能的 l 線中，使得 residual sum of squares 為最小的（minimum）！

意思是說，找到 $\beta_0=\hat{\beta}_0$ 值而 $\beta_1=\hat{\beta}_1$ 值，使得

$$L=\sum_{i=1}^{n}\left[Y_i-\beta_0-\beta_1 X_i\right]^2$$

為 minimum。

[continued]

(i) 解法：

$$\text{解}\quad \begin{cases} \partial L / \partial \beta_0 = 0 \\ \partial L / \partial \beta_1 = 0 \end{cases}$$

（這組方程式叫 normal equations）

(ii)Normal Equations：(★)

$$\partial L / \partial \beta_0 = 0 \quad \Rightarrow \quad \sum 2[Y_i - \beta_0 - \beta_1 X_i] \cdot (-1) = 0$$

$$\Rightarrow \quad \sum [Y_i - \beta_0 - \beta_1 X_i] = 0$$

$$\partial L / \partial \beta_1 = 0 \quad \Rightarrow \quad \sum 2[Y_i - \beta_0 - \beta_1 X_i] \cdot (-X_i) = 0$$

$$\Rightarrow \quad \sum X_i [Y_i - \beta_0 - \beta_1 X_i] = 0$$

(iii) Solutions：

$$\begin{cases} \sum Y_i - n\beta_0 - \beta_1 \sum X_i = 0 & \cdots\cdots(1) \\ \sum X_i Y_i - \beta_0 \sum X_i - \beta_1 \sum X_i^2 = 0 & \cdots\cdots(2) \end{cases}$$

(1) 式 $\times \frac{1}{n} \sum X_i$ 得

$$\frac{1}{n} \sum X_i \cdot \sum Y_i - \beta_0 \sum X_i - \beta_1 \cdot \frac{1}{n} (\sum X_i)^2 = 0 \cdots\cdots(1)'$$

(1)’ – (2) 得

$$\beta_1 (\sum X_i^2 - \frac{1}{n} (\sum X_i)^2) + \frac{1}{n} \sum X_i \cdot \sum Y_i - \sum X_i Y_i = 0 \cdots\cdots(3)$$

解 (3) 式得

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - \frac{1}{n} \sum X_i \cdot \sum Y_i}{\sum X_i^2 - \frac{1}{n} (\sum X_i)^2} \cdots\cdots\cdots(4)$$

[continued]

或分子分母同除以 n：

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum X_i Y_i - (\frac{1}{n}\sum X_i)(\frac{1}{n}\sum Y_i)}{\frac{1}{n}\sum X_i^2 - (\frac{1}{n}\sum X_i)^2}$$

$$\equiv \frac{XY - \overline{X}\cdot\overline{Y}}{X^2 - (\overline{X})^2} \quad \cdots\cdots\cdots\cdots\cdots\cdots(5)$$

或分子分母同乘以 n：

$$\hat{\beta}_1 = \frac{n\sum X_i Y_i - \sum X_i \cdot \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \quad \cdots\cdots\cdots(5)'$$

將 (5) 或 (5)’ 代入 (1) 中，可得

$$\hat{\beta}_0 = \frac{1}{n}\sum Y_i - \hat{\beta}_1 \cdot \frac{1}{n}\sum X_i \qquad \text{（直接由 (1) 解得）}$$

$$= \overline{Y} - \hat{\beta}\cdot\overline{X} \qquad \text{（用這個式子即可！）}$$

$$= \overline{Y} - \overline{X}\cdot(\frac{XY - \overline{X}\cdot\overline{Y}}{X^2 - (\overline{X})^2}) \qquad \text{（將 (5) 代入）}$$

Model Evaluation （模式之評估）

General:

模式之評估分兩部份：

(i) 是迴歸係數之評估

(ii) 是整個模式的適合性之評估

迴歸係數的評估，指的是迴歸係數的“顯著性”評估，意即迴歸係數

（β_1）是否顯著地不等於 0（$\neq$ 0，significantly）！

整個模式的適合性評估又包括兩部份：

(a) 模式的“配適程度”（fitness, fits）

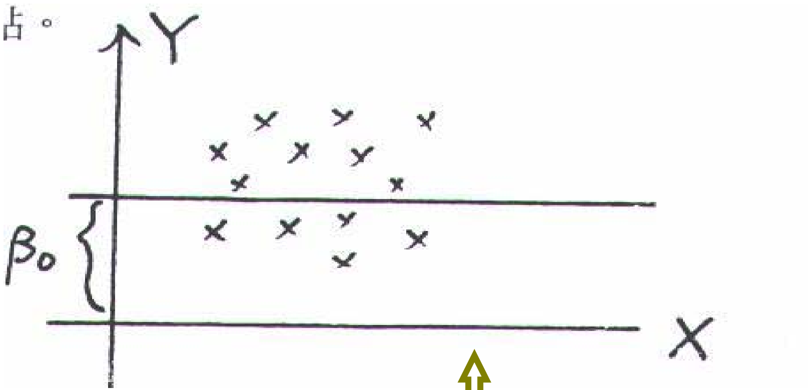
(b) 模式的“診斷”（diagnostics）

（H_0：$\beta_1=0$　v.s.　H_a：$\beta_1\neq0$）[單一參數]

Idea：

如果 β_1 顯著地不為 0，則 X 對 Y 才有解釋力（意即 X 的變化與 Y 的變化是有關係的）！　這時 β_0 不是重點。

如果 $\beta_1 = 0$，則你看到的 data 便是



如上圖，沒有什麼 pattern！

$$
\begin{aligned}
Y_i &= \beta_0 + 0 \cdot X_i + \varepsilon_i \\
&= \beta_0 + \varepsilon_i
\end{aligned}
$$

Test statistic

Under H_0,

$$\frac{\hat{\beta}_1 - 0}{\sqrt{\operatorname{var}\hat{\beta}_1}} \equiv \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}} \sim N(0,\ 1)$$

$\sigma_{\hat{\beta}_1}$ 裡有 $\sigma^2[\varepsilon \sim N(0,\ \sigma^2)]$，但 σ 為 unknown，要用一個 "代替品" $(\hat{\sigma})$，將這個估計子 $\hat{\sigma}$ 放入 $\sqrt{\operatorname{var}\hat{\beta}_1}$ $(\equiv \sigma_{\hat{\beta}_1})$ 裡時，記為 $S_{\hat{\beta}_1}$，結果是

$$\frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t_\nu, \qquad \nu = n - 2$$

sample size　　參數個數（參數：β_0, β_1）

在這裡 $\sigma_{\hat{\beta}_1} \equiv \sqrt{\operatorname{var}\ \hat{\beta}_1}$ 叫做 $\hat{\beta}_1$ 的 standard error（標準誤）！

Note：（$\hat{\beta}_1$ 的期望值與 confidence interval）

(a) $E\hat{\beta}_1 = \beta_1$ （true value）; that means the so obtained **least square estimator** is **unbiased** !!

(b) $\hat{\beta}_1$ 的 $100(1-\alpha)\%$ 省略，但若此 **CI** 包含 0，則表示不能 reject $H_0 : \beta_1 = 0$ 。

Extension (★)

對 Multiple regression $EY=\beta_0+\beta_1X_1+\cdots\cdots+\beta_pX_p$,

我們要做的是兩件事：

(i)

$$\left.\begin{array}{l} H_0:\beta_1=0 \\ H_0:\beta_2=0 \\ \quad\vdots \\ H_0:\beta_p=0 \end{array}\right\} \text{分別一個一個看 v.s.} \left\{\begin{array}{l} H_a:\beta_1\neq 0 \\ H_a:\beta_2\neq 0 \\ \quad\vdots \\ H_a:\beta_p\neq 0 \end{array}\right.$$

(ii) $H_0:\beta_1=\beta_2=\beta_3=\cdots=\beta_p=0$ （一起看！） v.s.

H_a： 至少有一個 $\beta_j\neq 0$

[➔ 至少有一個 X_j 可以解釋 Y]

對於(i)，我們得到的是個別的（分別對 $\beta_1=0$, $\beta_2=0$, $\cdots$; $\beta_p=0$, 之 test）檢定統計量 $\boldsymbol{t_1, t_2, \cdots, t_p}$。

對於 (ii) 我們會得到一個 F 統計量 $\boldsymbol{F_{a,b}}$：

$$\begin{cases} a=p \\ b=n-p-1 \end{cases}$$

•

• R^2 ；R-square

R^2 值又稱判定係數（coefficient of determination），是模式配適程度的一個指標，其定義及導出如下：

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

$$\Rightarrow\sum(Y_i-\bar{Y})^2=\sum(Y_i-\hat{Y}_i)^2+\sum(\hat{Y}_i-\bar{Y})^2+2\sum(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})$$

可以證明第三項 $=0$ ，則：

Total sum of squares：（TSS）

$$\sum(Y_i-\bar{Y})^2=\sum(Y_i-\hat{Y}_i)^2+\sum(\hat{Y}_i-\bar{Y})^2=RSS+MSS$$

$\bar{Y}$ 是一個不變的東西（∵ data 已知），$\hat{Y}_i$ 是你估的，如果你定的模式不離譜，$\hat{Y}_i$ 跟真正的 $EY_i=\beta_0+\beta_1X_i$ 應該很接近，則 $\sum(\hat{Y}_i-\bar{Y})^2$ 即表示來自於 model 的 sum of squares。剩下的 sum of squares 來自 random error ε_i，故 $(Y_i-\hat{Y}_i)^2$ 是模式沒辦法解釋的部份。R^2 的定義為

$$\begin{aligned}R^2=\frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}&=1-\frac{\sum(Y_i-\hat{Y}_i)^2}{\sum(Y_i-\bar{Y})^2}\\&=1-\frac{RSS}{TSS}\\&=\frac{MSS}{TSS}\end{aligned}$$

Residual and Residual plot: predicted Y vs. residual (★)

$Y_i - \hat{Y}_i$ 稱為 residual。

Residual 裡涵有許多在 model fitting 背後未知的訊息（information），

比如，若模式正確，則 $Y_i - \hat{Y}_i \ (\equiv e_i)$ 對 $\hat{Y}_i$ 的 plot 應該呈現隨機

分佈的情形（沒有 pattern）。(**)

如果有 pattern，則表示模式不正確，應重新考慮。

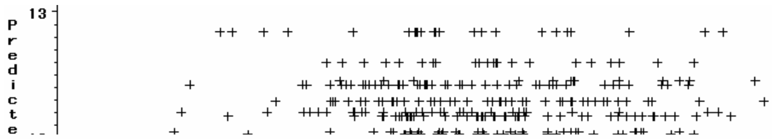
這個 plot，就叫 residual plot！[其他還有很多 diagnostic 的方法]

[: This can be easily justified through 'matrix algebra' and calculations, and is left as a homework.]**

Example:

Model: `model ln_x51=ind1trsf ind2trsf ind_age x16 ind_op ind_ISS ind_GCSM ind_RTS/include=2 selection=b;`

R-square=0.35.6



	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	203.62886430	40.72577286	60.08	0.0001
Error	543	368.08013585	0.67786397		
Total	548	571.70900015			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	10.61327698	0.08701823	10083.72635057	14875.7	0.0001
IND1TRSF	0.13118413	0.08299391	1.69360722	2.50	0.1145
IND2TRSF	0.55302609	0.09147250	24.77722443	36.55	0.0001
IND_OP	1.03361976	0.07889542	116.34844882	171.64	0.0001
IND_ISS	0.24761664	0.07766036	6.89133643	10.17	0.0015
IND_RTS	0.39342036	0.07527863	18.51453307	27.31	0.0001

The estimation of σ^2 (★)

設 model $\begin{cases} Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \\ EY = \beta_0 + \beta_1 X, \end{cases}$

Error term $\varepsilon \sim N(0, \sigma^2)$

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

Residual $e_i \equiv Y_i - \hat{Y}_i$

$$= Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

便形成用來"類比" ε_i 的量。考慮以下對 σ^2 的一個估計方式：

$$\hat{\sigma}^2 = \frac{\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n-2}$$

這個估計式（estimator）是對 σ^2 的一個 **unbiased estimator**。亦即

$$E\hat{\sigma}^2 = \sigma^2$$

Standard error (★)

標準誤（standard error）是對估計式（estimator）之精確度（precision）的一個描述 [in fact, the standard error is the “standard deviation” of the **sampling distribution** of the estimator]，一般即定為估計式之變異數之平方根，以前文之 $\hat{\beta}_1$ 為例，標準誤等於 $\sqrt{Var(\hat{\beta}_1)}$ 。但這個 standard error 裡含有 σ^2 ，是 unknown 的，所以一般便對這個 σ^2 給出一估計，將 $\hat{\sigma}^2$ 代入 $\sqrt{Var(\hat{\beta}_1)}$ 中（或即代入 $Var(\hat{\beta}_1)$ 中）。此時不寫 $Var(\hat{\beta}_1)$ ，而寫成 $V\hat{a}r(\hat{\beta}_1)$ ，代表 $Var(\hat{\beta}_1)$ 中仍有未知的成分是用其估計式代入的。

矩陣運算 (via matrix algebra) (★)

前述之估計方法若改用矩陣（matrix）的運算方式來表達，一切將變得更簡單：

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 X_1 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\Leftrightarrow \quad Y = X\beta + \varepsilon, \qquad X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

最小平方估計式 $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X'X)^{-1}X'Y$，而

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

Prediction

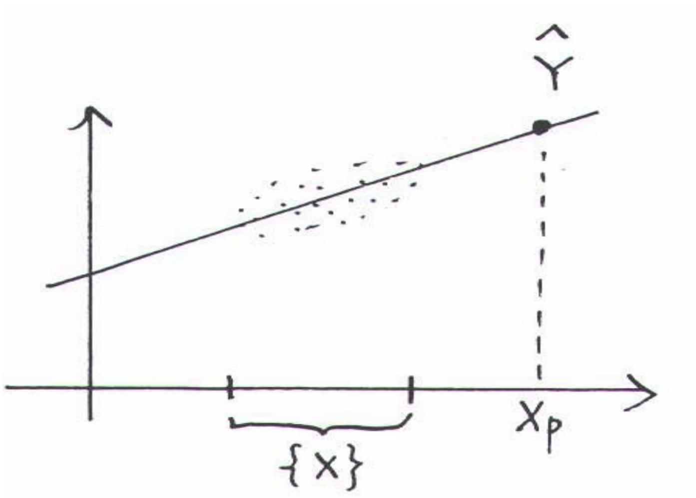
當你估計出了 $\hat{\beta}_0$ 及 $\hat{\beta}_1$ 後，你可以針對任一“合理範圍內”之 X 值來預測 Y 值（Q2）。但是這又分成兩部份：

因對一給定之 $X = x_p$ ，Y 可以有一個變動的範圍，其 variance 是 σ^2 （σ^2 可以用 $\hat{\sigma}^2$ 代替）。

我們可以預測兩樣東西，**一是 <u>Y 的 mean</u>，二是 <u>Y 的個別值</u>**，對於這兩種情況，均說為“預測 $\hat{Y}_p = \hat{\beta}_0 + \hat{\beta}_1 X_p$ ”。

On Q2

所謂“合理的範圍”指的是 X_p **不能離**



<u>“用來得到 $\hat{Y}$ 的那些 **x** 值(以 $\{X\}$ 表示)”</u> **太遠**。

(continued)

更精確地說，應該是“給定 $X = x_p$ ，利用 $\hat{\beta}_0$ 及 $\hat{\beta}_1$，來預測 Y 之 mean、或預測 Y 之單一觀察值 ”！

所謂預測，就是給出一個 100 (1 - α) % CI。我們可以使用矩陣運算符號很簡單地導出預測式。我們先給出其公式：

預測 mean Y

$$\hat{y} \pm t_{n-2,\ 1-\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Derivation: see the next page!

預測單一的 Y

(variance=預測 mean Y 時之 variance+σ^2)

$$\hat{y} \pm t_{n-2,\ 1-\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

顯然，這裡多出了一個 $\hat{\sigma}^2$ ！這可以理解。其餘的部分請參看本章之 Appendix 1.

Appendix 1 of Regression (★)

It can be shown by **standard argument** that, for a sample size n (not necessarily to be large),

$$\frac{\hat{y} - \mathrm{E}(\hat{y})}{\{\mathrm{Var}(\hat{y})\}^{1/2}} \sim N(0,1),$$

provided the design matrix is 'regular' (spread out very uniformly!). Because $\hat{y} = X\hat{\beta}$, and $\mathrm{E}(\hat{y}) = X\,\mathrm{E}(\hat{\beta})$; $\hat{\beta} = (X'X)^{-1}X'Y$, $\mathrm{E}(\hat{\beta}) = (X'X)^{-1}X'\mathrm{E}(Y) = (X'X)^{-1}X'X\beta = \beta$. So, $\mathrm{E}(\hat{y}) = X\beta = \mathrm{E}y$. The previous expression is equivalent to

$$\frac{\hat{y} - \mathrm{E}(y)}{\{\mathrm{Var}(X\hat{\beta})\}^{1/2}} \sim N(0,1).$$

It should be noted that, for calculating the variance at the denominator, X must be replaced by $X_p = (1, x_p)'$ because the prediction is implemented **on the value of** x_p. Moreover, $\mathrm{Var}(X\hat{\beta}) = X\mathrm{Var}(\hat{\beta})X'$. So, for the value of $x = x_p$, we need to calculate $(1, x_p)\mathrm{Var}(\hat{\beta})(1, x_p)' = \sigma^2(1, x_p)(X'X)^{-1}(1, x_p)'$, due to $\mathrm{Var}(\hat{\beta}) = (X'X)^{-1}\sigma^2$.

From a previous context,

$$(X'X)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}.$$

So we have

$$\begin{aligned}(1, x_p)(X'X)^{-1}(1, x_p)' &= \frac{\sum x_i^2 - 2x_p\sum x_i + nx_p^2}{n\sum x_i^2 - (\sum x_i)^2} \\ &= \ldots = \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}.\end{aligned}$$

I left the part "..." as an exercise for the reader. Becasue σ should be substituted by its consistent estimator $\hat{\sigma} = \hat{\sigma}_{\beta_0,\beta_1}$, the sampling distribution is no longer Gaussian, but t-distribution on $n-2$ degrees of freedom.

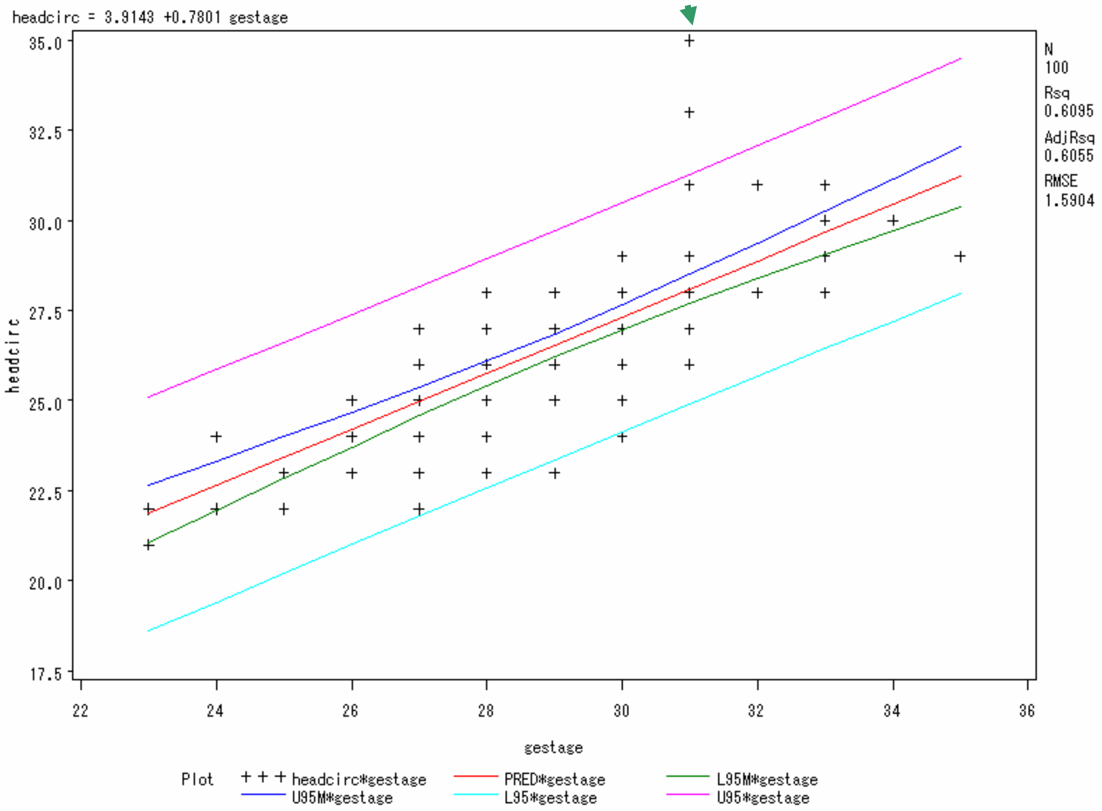
Example of SAS Code

```
data lbw;
infile 'd:\class\biostat_under_2004\SAS_lab\lab08_infant.dat';
input headcirc length gestage birthwt momage toxemia;

proc print;
run;

  proc reg graphics;
   model headcirc=gestage/cli;
   plot headcirc*gestage/conf95 pred95;
  run;

  proc reg graphics;
   model length=gestage/cli;
   plot length*gestage/conf95 pred95;
  run;
```



Correlation: the bivariate normal distribution (★)

X 與 Y 兩個變數的相關強度可以用 correlation coefficient（相關係數）來表示，先假設 X，Y 是 jointly normal（聯合常態分佈）。此時，固定任一 X 值來看，Y 均是常態分佈；相反的，固定任一 Y 值，X 也均是常態分佈，且假設其變異數分別為 σ_Y^2 、σ_X^2 。

Formulae: **Bivariate Normal Distribution**

$$f(X_1,X_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(X_1-\mu_1)^2}{\sigma_1^2}+\frac{(X_2-\mu_2)^2}{\sigma_2^2}-2\rho\frac{(X_1-\mu_1)(X_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\}$$

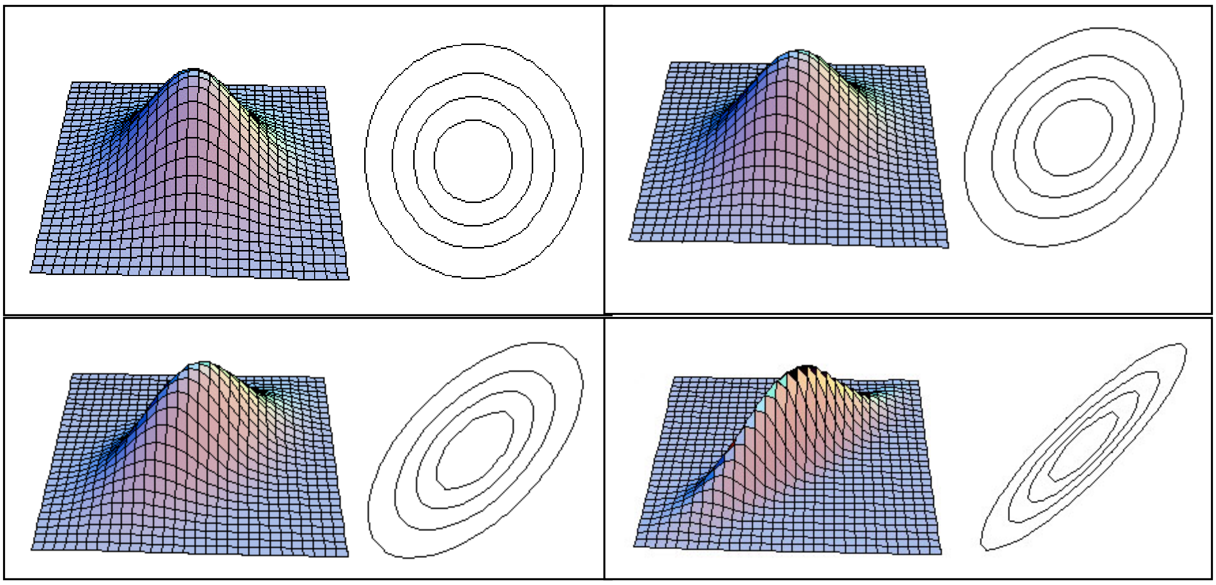
Formulae: **Multivariate Normal Distribution**

- There is a vector $\mu=[\mu_1,\ldots,\mu_N]$ and a symmetric, positive semi-definite covariance matrix Σ ($N\times N$ matrix) such that X has density

$$f_X(x_1,\ldots,x_N)=\frac{1}{(2\pi)^{N/2}\,|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right)$$

where $|\Sigma|$ is the determinant of Σ. Note how the equation above reduces to that of the univariate normal distribution if Σ is a scalar (i.e., a real number). The vector μ in these conditions is the expected value of X and the matrix $\Sigma = AA^T$ is the covariance matrix of the components X_i.

Figures: **Bivariate Normal Distribution, ρ=0, 0.3, 0.6, 0.9.**



Pearson's correlation coefficient

定義

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{VarX} \cdot \sqrt{VarY}}$$

Note: If the data is not ranked data instead of continuously distributed, the Spearman's rank correlation is used. Definition of the rank correlation has the same mathematical form, but with original X and Y replaced by ranks.

與迴歸係數之關係 (★)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad \varepsilon \sim N(0, \sigma^2)$$

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\varepsilon}, \qquad \bar{\varepsilon} \approx E\varepsilon = 0$$

$$\Rightarrow \quad Y_i - \bar{Y} \approx \beta_1 (X_i - \bar{X})$$

$$\Rightarrow \quad \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n} \approx \beta_1 \frac{\sum (X_i - \bar{X})^2}{n}$$

$$\Rightarrow \quad \hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})/n}{\sum (X_i - \bar{X})^2 / n}$$

$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})/n}{\sqrt{\sum (X_i - \bar{X})^2 / n}\sqrt{\sum (Y_i - \bar{Y})^2 / n}} \cdot \frac{\sqrt{\sum (Y_i - \bar{Y})^2 / n}}{\sqrt{\sum (X_i - \bar{X})^2 / n}}$$

$$= \hat{\rho}_{XY} \cdot \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}$$

[Note that $R^2 = (\hat{\rho}_{XY})^2$ in simple regression !!]

Multiple Regression （多變數迴歸）(★)

Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i \ , \quad \varepsilon_i \sim N(0,\ \sigma^2), \quad i = 1, 2, \cdots, n$$

$X_1,\ X_2, \cdots;\ X_p$ 乃是用來解釋 Y 的可能結果（outcome）的一組解釋變數，一位研究者（investigator）應當儘可能地收集可以解釋 outcome Y 的解釋變數 Xs。

遺漏了重要的 X，會使得估計結果有所偏差（biased）！

Estimation：Least Square (LS) Method

令 $L = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \cdots - \beta_p X_{pi})^2$

解 $\partial L / \partial \beta_j = 0, \quad j = 0,1,2,\cdots,p$ ，得 $\hat{\beta}_0, \ \hat{\beta}_1, \cdots, \ \hat{\beta}_p,$

Testing

(i) 個別參數之檢定：

$$\begin{cases} H_0 : \beta_j = 0 & (or \ \beta_j = \beta_{j0}) \ \ v.s. \\ H_a : \beta_j \neq 0 & (or \ \beta_j \neq \beta_{j0}) \end{cases}$$

$$\frac{\hat{\beta}_j - 0}{S_{\hat{\beta}_j}} \sim t_{n-(p+1)} \qquad (\because \beta_{j0} = 0)$$

(ii) 模式顯著性檢定（F-test）：

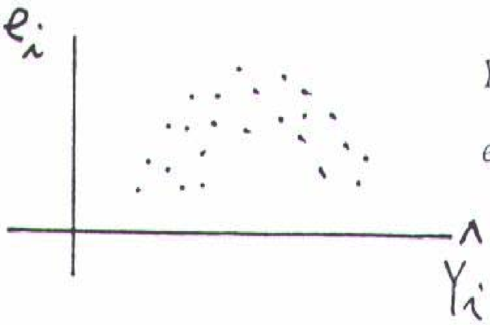
$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 & v.s. \\ H_a : \beta_j \neq 0, \quad for \ some \ j \end{cases}$$

建構出一個統計量 $F_n(Y,X) \sim F$ 分佈 $\equiv F_{a,b}$

$$d.f. \begin{cases} a = p \\ b = n - p - 1 \end{cases}$$

Diagnostics

一個初步的診斷（diagnostic）方式是做 $\hat{Y}_i$ 對 e_i 的散佈圖（scatter plot）：



$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_p X_{pi}$$

$$e_i = Y_i - \hat{Y}_i$$

如果前面關於迴歸的假設（Equal variance, normality, …）為正確的，則 $\hat{Y}_i$ 應與 e_i 為 uncorrelated，則其 scatter plot 應沒有什麼 pattern (即非常 random)。反之若呈現某種 pattern（如上圖，是弓形），表示有重要的 information 遺漏了，或變數搜集不足，或變數函數形式不對…。

R^2；multiple correlation coefficient

$$R^2 \equiv \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

迴歸係數之解釋

β_1 表當 $X_2, X_3, \cdots; X_p$ 固定時，X_1 變動一個單位，Y 的 mean 將變動 β_1 個單位。

（$\beta_2, \beta_3, \cdots, \beta_p$ 之意義仿之！）

Confounding，干擾 (★)

當 true model 是 $EY=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p$，而你卻遺漏了一個重要變數（比如 X_p）時，則你將會 fit 如下之 model：

$$EY=b_0+b_1X_1+b_2X_2+\cdots+b_{p-1}X_{p-1}$$

此時，你得到的估計式（LSE），$\hat{b}_j$ 將會是 biased 的，即 $E\hat{b}_j\neq\beta_j$。我們說此 bias 的大小為

$$\beta_j-E\hat{b}_j$$

而 Mean Square Error (MSE) = bias**²** + variance

$$=(\beta_j-E\hat{b}_j)^2+Var(\hat{b}_j) \qquad (*)$$

所以，若 X_p 是一個重要的解釋變數而被你遺漏時，其他變數的參數估計會有 bias 產生。(**Q**：如何評估？)

若 bias 很嚴重，我們稱 $\boldsymbol{X_p}$ 是（$\boldsymbol{X_1\cdots\cdots X_{p-1}}$）與 Y 之關係的一個干擾因子。故 $\boldsymbol{X_p}$ 一定要納入迴歸模式中，以避免干擾(i.e.：避免 bias)

(*)

$$
\begin{aligned}
MSE &\equiv E\left\{(\hat{b}_j - \beta_j)^2\right\} \\
&= E\left\{(\hat{b}_j - E\hat{b}_j + E\hat{b}_j - \beta_j)^2\right\} \\
&= E\left\{(\hat{b}_j - E\hat{b}_j)^2 + 2(\hat{b}_j - E\hat{b}_j)(E\hat{b}_j - \beta_j) + (E\hat{b}_j - \beta_j)^2\right\}
\end{aligned}
$$

$$\downarrow \qquad\qquad\qquad \downarrow \qquad\qquad\qquad \downarrow$$

$$(1) \qquad\qquad\qquad (2) \qquad\qquad\qquad (3)$$

$$= E\{(1)+(2)+(3)\}$$

$$= Var(\hat{b}_j) + E(\ (2)\ ) + E(\ (3)\ )$$ ，但 $E(\ (3)\ ) = (3)$

又，

$$E(\ (2)\ ) = 2(E(\hat{b}_j) - \beta_j)\cdot E(\hat{b}_j - E\hat{b}_j) = 0 \qquad (\because E(\hat{b}_j - E\hat{b}_j) = 0)$$

$$= Var(\hat{b}_j) + (bias)^2$$

分部相關，**partial correlation** (★)

如果真正的 model 是

$$EY=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p$$

比如 p=2，而你卻 fit $EY=b_0+b_1X_1$，根據你所相信的 model（只有 X_1 對 Y 有解釋力），計算出 Y 與 X_1 之 correlation coefficient 為 $\hat{\rho}_{Y,\,X_1}$。這個 $\hat{\rho}_{Y,\,X_1}$ 事實上是 biased 的，即它並非反映真正的 Y 與 X_1 之關係。如前，你並未“控制干擾”（control the confounding）。要達到控制干擾：

(i) 做迴歸，即必須做

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_pX_{pi}+\varepsilon_i$$

(ii) 求 correlation，即須求 $\rho_{Y,X_1|X_2,\cdots,X_p}$ 。

此時 β_1 即 $X_2,\cdots,\ X_p$ 固定時，X_1 改變一個單位，Y 之 mean 改變幾個單位，而 $\rho_{Y,X_1|X_2,\cdots,X_p}$ 即 $X_2,\cdots,\ X_p$ 固定時，X_1 與 Y 之 correlation，叫做 **partial correlation (coefficient)**。

用 regression 做兩組 data 之 mean 之比較(★)

設 $\left.\begin{matrix} X_{11},\cdots\cdots,X_{1n_1} \sim N(\mu_1,\sigma^2) \\ X_{21},\cdots\cdots,X_{2n_1} \sim N(\mu_2,\sigma^2) \end{matrix}\right\}$ $X_{1\bullet}$ 與 $X_{2\bullet}$ 獨立。

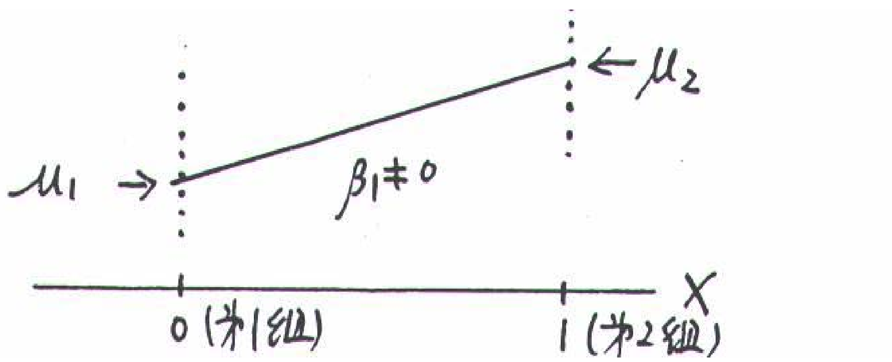
要看 μ_1 是否等於 μ_2 $(H_0:\mu_1=\mu_2 \quad v.s. \quad H_a:\mu_1\neq\mu_2)$ 可以這樣做：

$$\left\{\begin{array}{l} \left.\begin{array}{lcl} Y_1 & = & X_{11} \\ Y_2 & = & X_{12} \\ \vdots & & \\ Y_{n_1} & = & X_{1n_1} \end{array}\right\} Group \quad 1 \\ \left.\begin{array}{lcl} Y_{n_1+1} & = & X_{21} \\ Y_{n_1+2} & = & X_{22} \\ \vdots & & \\ Y_{n_1+n_2} & = & X_{2n_2} \end{array}\right\} Group \quad 2 \end{array}\right. = \beta_0+\beta_1 \left\{\begin{array}{l} \left.\begin{array}{lcl} X_1 & = & 0 \\ X_2 & = & 0 \\ \vdots & & \\ X_{n_1} & = & 0 \end{array}\right\} Group \quad 1 \\ \left.\begin{array}{lcl} X_{n_1+1} & = & 1 \\ X_{n_1+2} & = & 1 \\ \vdots & & \\ X_{n_1+n_2} & = & 1 \end{array}\right\} Group \quad 2 \end{array}\right.$$

對上面 $n\,(=n_1+n_2)$ 個 data $(Y_1,X_1),\cdots;(Y_{n_1},X_{n_1}),(Y_{n_1+1},X_{n_1+1}),\cdots(Y_n,X_n)$

做 simple linear regression，看 $H_0:\beta_1=0 \quad v.s. \quad H_a:\beta_1\neq 0$

以圖形表示，如果 $\beta_1\neq 0$ 表示下兩組人之 mean 有差異



如果 $\beta_1=0$，則表示上面的兩組人之 mean 沒有差異。

用 regression 作 one-way ANOVA（➔ two-way 呢？）(★)

假設有 3 組人：

$$\begin{cases} X_{11},\cdots\cdots,X_{1n_1} \sim N(\mu_1,\sigma^2) \\ X_{21},\cdots\cdots,X_{2n_2} \sim N(\mu_2,\sigma^2) \\ X_{31},\cdots\cdots,X_{3n_3} \sim N(\mu_3,\sigma^2) \end{cases}$$

令

$$\begin{cases} \left.\begin{matrix} Y_1 = X_{11} \\ Y_2 = X_{12} \\ \vdots \\ Y_{n_1} = X_{1n_1} \end{matrix}\right\} \\ \left.\begin{matrix} Y_{n_1+1} = X_{21} \\ Y_{n_1+2} = X_{22} \\ \vdots \\ Y_{n_1+n_2} = X_{2n_2} \end{matrix}\right\} \\ \left.\begin{matrix} Y_{n_1+n_2+1} = X_{31} \\ \vdots \\ Y_{n_1+n_2+n_3} = X_{3n_3} \end{matrix}\right\} \end{cases} = \beta_0 + \beta_1 \begin{cases} \begin{cases} X_{11} = 0 \\ X_{12} = 0 \\ \vdots \\ X_{1n_1} = 0 \end{cases} \\ \begin{cases} X_{1,n_1+1} = 1 \\ X_{1,n_1+2} = 1 \\ \vdots \\ X_{1,n_1+n_2} = 1 \end{cases} \\ \begin{cases} X_{1,n_1+n_2+1} = 0 \\ \vdots \\ X_{1,n_1+n_2+n_3} = 0 \end{cases} \end{cases} + \beta_2 \begin{cases} \begin{cases} X_{21} = 0 \\ X_{22} = 0 \\ \vdots \\ X_{2n_1} = 0 \end{cases} \\ \begin{cases} X_{2,n_1+1} = 0 \\ X_{2,n_1+2} = 0 \\ \vdots \\ X_{2,n_1+n_2} = 0 \end{cases} \\ \begin{cases} X_{2,n_1+n_2+1} = 1 \\ \vdots \\ X_{2,n_1+n_2+n_3} = 1 \end{cases} \end{cases}$$

則 $$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \equiv \quad H_0' : \beta_1 = \beta_2 = 0$$

Dummy Variable Technique

前兩節中對解釋變數的設定是一種“技巧”，這種變數叫 dummy variable（啞變數）。

若定第 1 組人為 baseline（基準組），β_1 代表的數便是第 2 組人與第 1 組人之關係；β_3 代表的則是第 3 組與第 1 組之關係。

SAS-example for coding dummy variables:

```
data transf;
infile 'd:\Dr_ChenRayJ\transfer\work1230.csv' dlm="," missover;
input id $ name $ sex age ind_age lag $ ind_lag $ ind2_lag $
      x9 x10 $ x11-x21 GCSM ind_GCSM ind_RTS x25-x36 ISS X38 surv x40-x51 transfer;
          if sex="." then delete;
          xx=x9-transfer; /* to check whether or not x9=x52 */
      ind_trsf=(transfer=1);  /* indicator variable for direct/indirect transfer */
          ind1trsf=(transfer=2); /* dummy variable 1 for transfer */
          ind2trsf=(transfer=3); /* dummy variable 2 for transfer */
          ln_x51=log(x51);
      RTS=0.93678*((x13>3)+(x13>5)+(x13>8)+(x13>12))+
           0.7326*((x16>0)+(x16>49)+(x16>75)+(x16>89))+
           0.2908*((x19>0)+(x19>5)+(x19>9)+(x19>29));
          ind_ISS=(ISS>=25);
          ind_age=(age>=40);
          ind_op=(x41>=1);
        data temp; set transf;/* keep id name sex age iss ind_iss surv;*/
        proc sort; by surv ind_ISS   iss;
proc print; run;
```

x51	transfer	xx	ind_trsf	ind1trsf	ind2trsf	ln_x51	RTS	ind_ISS	ind_op
27364	1	0	1	0	0	10.2170	7.54992	0	0
11462	1	0	1	0	0	9.3468	7.54992	0	0
114158	1	0	1	0	0	11.6453	7.54992	0	1
44125	3	0	0	0	1	10.6948	7.54992	0	0
12034	1	0	1	0	0	9.3955	7.54992	0	0
151345	1	0	1	0	0	11.9273	7.54992	0	1
485189	1	0	1	0	0	13.0923	4.73958	0	1
15630	1	0	1	0	0	9.6569	7.54992	0	0
61153	1	0	1	0	0	11.0211	7.54992	0	1
68333	1	0	1	0	0	11.1321	5.67636	0	0
110923	2	0	0	1	0	11.6166	7.54992	0	1
204101	2	0	0	1	0	12.2264	4.73958	0	1
349940	2	0	0	1	0	13.6529	4.80396	0	1

Interaction（交互作用）(★)

Model（2-variable）

$EY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 \cdot X_2$

當 $\beta_{12} \neq 0$，叫 X_1 與 X_2 之間有 interaction。

Slope dummy

假設 X_2 只有兩種可能： $X_2 = \begin{cases} 0 \\ 1 \end{cases}$

則當 $X_2 = 0$ 時，

$$E(Y \mid X_2 = 0) = \beta_0 + \beta_1 X_1 + 0 + 0$$

當 $X_2 = 1$，

$$E(Y \mid X_2 = 1) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_{12} X_1$$

∴固定，(i) 在 $X_2 = 0$ 時，X_1 改變一個單位，EY 改變 β_1。

(ii) 在 $X_2 = 1$ 時，X_1 改變一個單位，EY 改變 $\beta_1 + \beta_{12}$。

Interaction 之意義

X_1 在不同之 X_2 值時，其對 Y 的“解釋力”不同！

A two-variable example

給定 Table 11.2.2 & Table 11.2.3（課本），設 fitted model 為

$$\hat{y}_i = 6.21 + 1.03X_{1i} + 41.3X_{2i} + 22.7X_{3i} - 0.703X_{1i} \cdot X_{2i} - 0.510X_{1i} \cdot X_{3i}$$

且有一個 computer output 為 figure 11.2.5（課本），問：

(i) 如何描述 $\mathbf{X_1}$ 的 effect（$\mathbf{X_1}$ = age）?

(ii) 如何描述 $\mathbf{X_1}$ treatment 的 effect？

(iii) Age 與 treatment 是否有 interaction？