# 報告題名：教師薪資的探討

作者：施珮玉、陳弘明、徐戎萱、李育瑄、錢佩君、吳憶瀅、郭根連

系級：統計系三年乙班

學號： D9765780、M9807073、D9765763、D9765818、D9765759、
D9725092、D9659685

開課老師：陳婉淑 教授

課程名稱：迴歸分析

開課系所：統計學系

開課學年：99 學年度 第一學期

# 中文摘要

本次的分析為調查教師薪資的多寡是否受到其他因素影響，我們選用 Journal of Royal Statistical Society 中的 Salary 資料作為分析依據，試圖探討**任教年資**、**性別**、**學位**、**任教學校**、**研究所學位**以及**任教期間間斷是否超過兩年**等變數是否影響薪資的高低。

在研究過程中我們採用迴歸分析尋找各變數與教師薪資的相關性，並以**向前選取**、**向後消去**、**逐步迴歸**、Adj R-square 及 Cp **選取**等法則輔以作業，再以**殘差分析**檢視是否符合殘差常態假設，數據及圖表資料則以 SAS 統計軟體分析、採 MHT 格式輸出呈現。

最後研究結果發現，教師**任教年資**、**學位**、**任教學校**與**研究所學位**等四個變數對於薪資多寡有顯著的影響。希望這回的研究結果能夠應用在各級學校中對於教師薪資的給付作為參考依據。

**關鍵字：**迴歸分析、選模、殘差分析

目次

# 一、資料蒐集

## 1-1. 資料來源

原始資料來自 Journal of Royal Statistical Society, A-137 ,1974, 245-258，作者為 Turnbull 與 Williams，資料中含有 90 筆資料，七個變數，包含變數 Y 是教師一年的薪資，六個變數 X 分別是教師任教年資、性別、學位、任教學校、研究所學位與任教期間間斷是否超過兩年，將這些變數做迴歸分析並觀察其之間的關連性，推測出影響教師薪資的重要變數有哪些。

## 1-2. 變數解釋

Y：salary 一年的薪資(以英鎊計)

$X_1$：service 教師任教的年資(以月為單位)

$X_2$：sex 性別(1=男，0=女)

$X_3$：degree 學位(次序變項，介於 1-36 之間，愈高學位其值愈高)

$X_4$：school 任教學校(1=公立，0=私立)

$X_5$：grad 研究所學位(1=有，0=無)

$X_6$：break 教師任教期間間斷教職超過兩年(1=有，0=無)

# 二、資料分析

## 2-1.原始資料分析

### ◎散佈圖

圖 2.1.1 變數 Y 與$X_1$散佈圖(相關係數 0.86764)

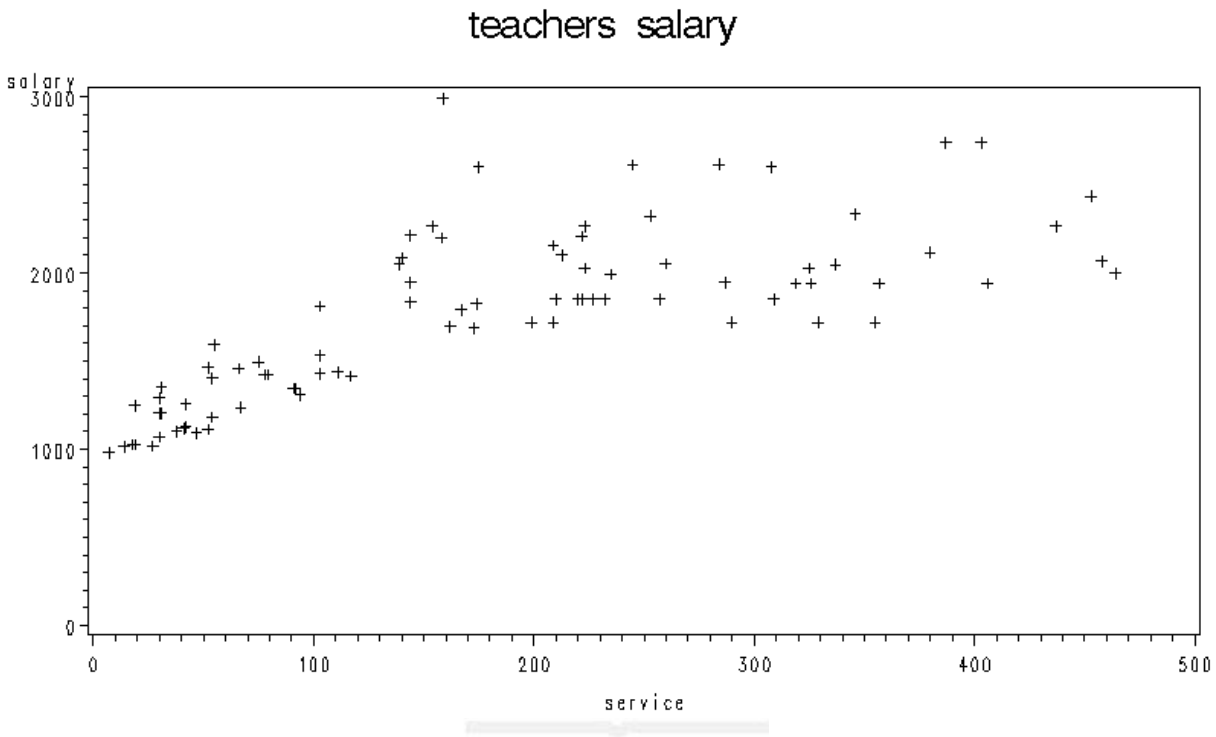
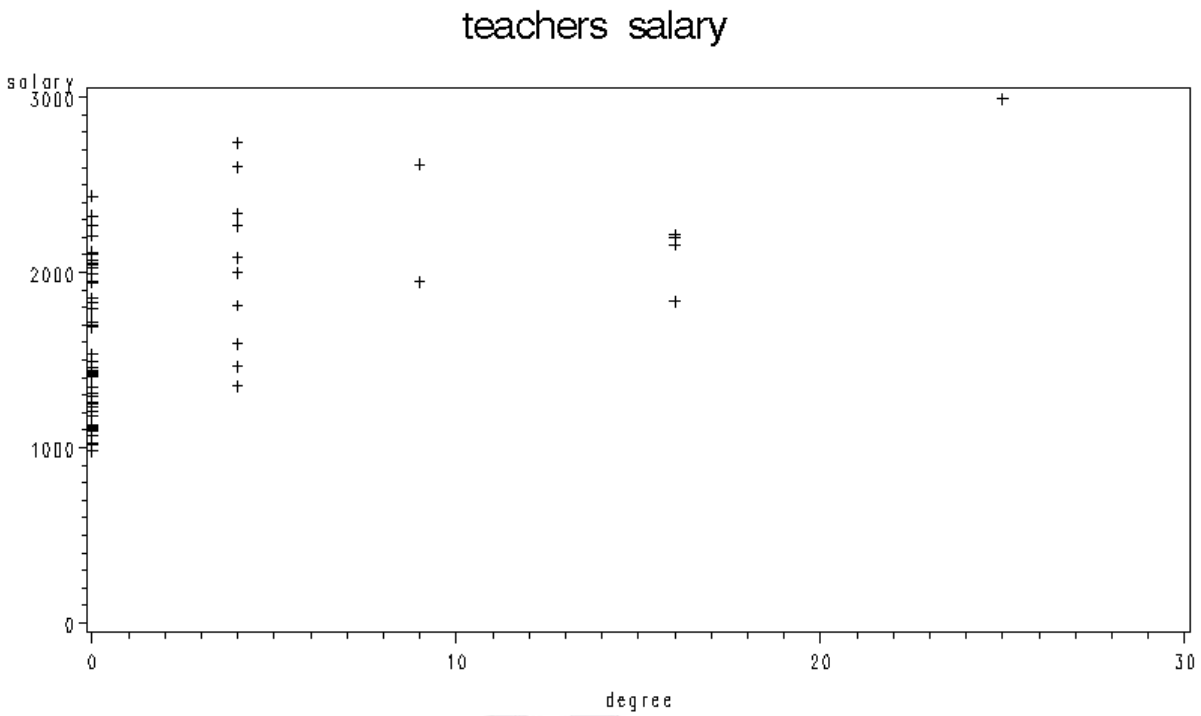



圖 2.1.1 中，由於範圍包含太廣，且變異有漸增之趨勢，無法看出教師一年的薪資與教師任教的年資有很明顯的線性關係。

圖 2.1.2 變數 Y 與$X_3$散佈圖




由圖 2.1.2 中，看出教師一年的薪資與學位(次序變項)高低沒有線性關係。$X_2$、$X_4$、$X_5$、$X_6$為虛擬變數，散佈圖省略。

表 2.1.1

| | | | |
|---|---|---|---|
| Root MSE | 221.74474 | R-Square | 0.8067 |
| Dependent Mean | 1727.84667 | Adj R-Sq | 0.7927 |
| Coeff Var | 12.83359 | | |

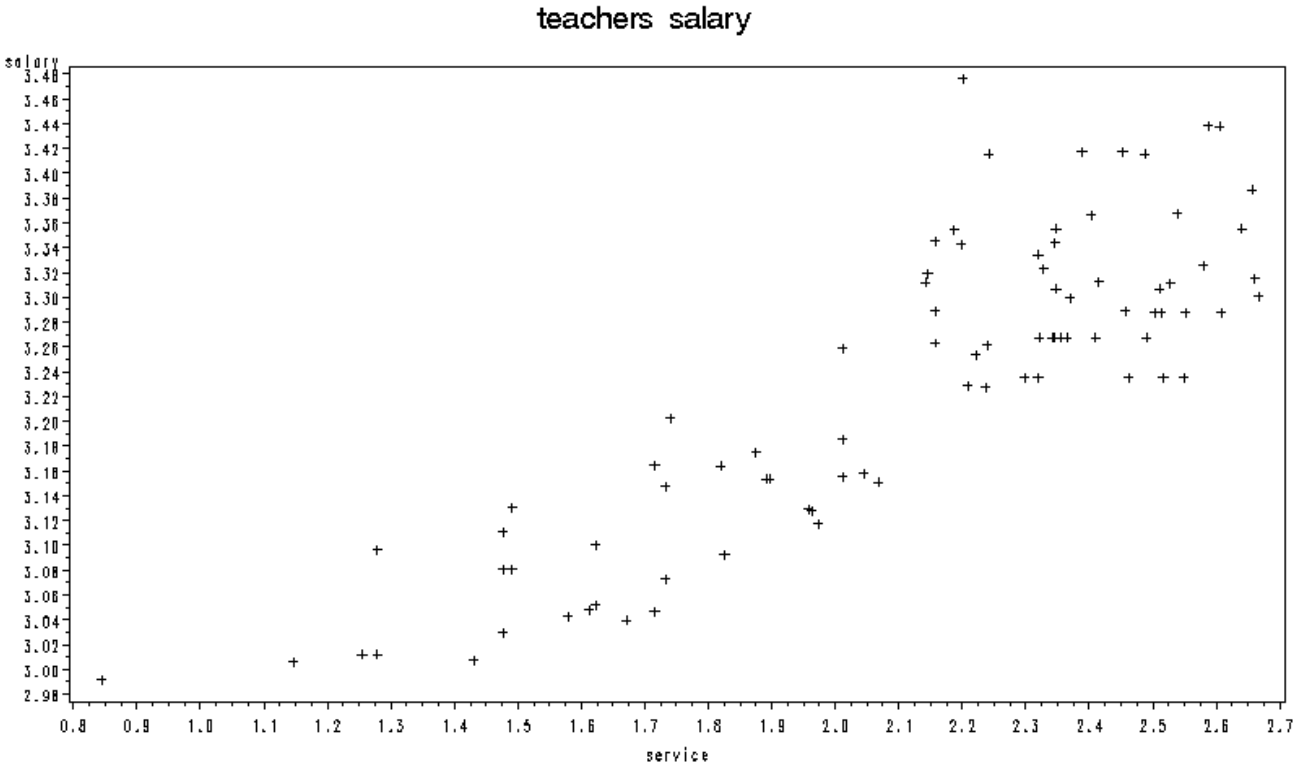
經由表 2.1.1 可以看到 R-Square 為 0.8067，代表迴歸模型對這筆資料有 80.67%的解釋能力,所以我們將對資料做轉換看是否能使我們的解釋能力上升。

## 2-2. 轉換後資料分析

由於原始資料的範圍包含域廣，所呈現出的散佈圖變異有漸增之趨勢，不容易看出線性相關，所以我們決定將薪資(Y)與年資($X_1$)作 log 轉換，以便觀察。

$$Y \to \log_{10} Y, X_1 \to \log_{10} X_1$$

圖 2.2.1 變數 Y 與$X_1$散佈圖




由圖 2.2.1 可以看出變數 Y 教師一年的薪資與變數$X_1$教師任教的年資有線性相關，呈現正相關。(相關係數 0.86764)

學位($X_3$)為次序變項，$X_2$、$X_4$、$X_5$、$X_6$為虛擬變數，散佈圖省略。

## 2-3. 設立模型

### ◎是否用 log 轉換後模型

$$\log_{10} Y_i = \beta_0 + \beta_1 \log_{10} X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon_i$$

表 2.3.1 轉換後的 ANOVA 表

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 6 | 1.23546 | 0.20591 | 109.53 | <.0001 |
| Error | 83 | 0.15603 | 0.00188 | | |
| Corrected Total | 89 | 1.39150 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.04336 | R-Square | 0.8879 |
| Dependent Mean | 3.22004 | Adj R-Sq | 0.8798 |
| Coeff Var | 1.34651 | | |

由表 2.3.1 中， 檢定是否存在迴歸模型

$$\begin{cases} H_0: \beta_j = 0 & , j = 1,2 \ldots, 6 \\ H_1: \beta_j \text{至少有一個不等於} 0 & , j = 1,2, \ldots, 6 \end{cases}$$

檢定統計量:$F^*$近似於 109.53

檢定規則:Reject $H_0$ if $F^* > F_{0.05}(6, 83)$近似於 2.210

所以$F^*$=109.53 > 2.210

Reject $H_0$ at $\alpha$=0.05 level 可知迴歸模型存在

另外 R-Square=0.8879 及 Adj R-Sq=0.8798，對變數 Y 能解釋的變異有 88.79% ，解釋能力相對未轉換的模型(R-Square=0.8067 及 Adj R-Sq=0.7927)改善很多。另外，殘差分析結果也有明顯的改善，因此我們**將模型做 log 轉換**。

表 2.3.2 轉換後的 Parameter Estimates 表

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
| Intercept | Intercept | 1 | 2.67736 | 0.02389 | 112.08 | <.0001 | 0 |
| X1 | service | 1 | 0.24578 | 0.01141 | 21.53 | <.0001 | 1.11005 |
| X2 | sex | 1 | -0.00028515 | 0.01049 | -0.03 | 0.9784 | 1.20746 |
| X3 | degree | 1 | 0.00580 | 0.00138 | 4.22 | <.0001 | 1.79716 |
| X4 | school | 1 | 0.03298 | 0.01010 | 3.26 | 0.0016 | 1.21949 |
| X5 | grad | 1 | 0.04432 | 0.01471 | 3.01 | 0.0034 | 1.79084 |
| X6 | break | 1 | -0.01611 | 0.01147 | -1.40 | 0.1639 | 1.16384 |

可初步觀察可能要的重要變數：教師任教的年資$X_1$、學位$X_3$、任教學校$X_4$與研究所學位$X_5$，因其四個變數有明顯的顯著效果，而性別($X_2$)與間斷教學期間是否超過兩年($X_6$)無顯著效果。

# 三、選擇重要變數

## 使用的重要變數選取法：

- 向後消去法（Backward Elimination）
- 向前選取法(Forward Selection)
- 逐步迴歸法(Stepwise Regression)
- Adjusted R-Square 選取法
- CP 選取法

## 3-1 向後消去法（Backward Elimination）

首先把所有變數都放進模型裡面，一次評估一個變數,把不顯著的移除,直到沒有不顯著的,且被剔除的變數就不再考慮進模型中。

表 3.1 Backward Elimination: Summary

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | 2.68240 | 0.02332 | 24.88879 | 13235.9 | <.0001 |
| X1 | 0.24165 | 0.01101 | 0.90513 | 481.35 | <.0001 |
| X3 | 0.00564 | 0.00136 | 0.03227 | 17.16 | <.0001 |
| X4 | 0.03355 | 0.00946 | 0.02364 | 12.57 | 0.0006 |
| X5 | 0.04222 | 0.01463 | 0.01566 | 8.33 | 0.0050 |

由表可知變數$X_1$、$X_3$、$X_4$、$X_5$在0.1的顯著水準下，都被選入模式中。

表 3.2

| Summary of Backward Elimination | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable Removed | Label | Number Vars In | Partial R-Square | Model R-Square | C(p) | F Value | Pr > F |
| 1 | X2 | sex | 5 | 0.0000 | 0.8879 | 5.0007 | 0.00 | 0.9784 |
| 2 | X6 | break | 4 | 0.0027 | 0.8851 | 5.0209 | 2.04 | 0.1565 |

由表 3.2 發現向後消去法消去的變數有$X_2$(sex)和$X_6$(break)。

## 3-2 向前選取法(Forward Selection):

自變數的選取以是否達到統計顯著水準的變數，依解釋力的大小，依次選取進入迴歸方程式中，以**逐步增加**的方式，完成選取的動作。

表 3.3

| Summary of Forward Selection | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable Entered | Label | Number Vars In | Partial R-Square | Model R-Square | C(p) | F Value | Pr > F |
| 1 | X1 | service | 1 | 0.7528 | 0.7528 | 96.9776 | 267.98 | <.0001 |
| 2 | X3 | degree | 2 | 0.0986 | 0.8514 | 25.9832 | 57.74 | <.0001 |
| 3 | X4 | school | 3 | 0.0225 | 0.8739 | 11.3492 | 15.32 | 0.0002 |
| 4 | X5 | grad | 4 | 0.0113 | 0.8851 | 5.0209 | 8.33 | 0.0050 |

由向前選取法，在 0.1 的顯著水準下，$X_1$、$X_3$、$X_4$、$X_5$都被選入模式中。

## 3-3 逐步迴歸法(Stepwise Regression):

先選取最顯著的變數，選進模型後再進行檢驗是否有要刪除的變數，重複直到所有的變數都達到顯著水準為止，以得到迴歸最佳模式。

表 3.4

| Summary of Stepwise Selection | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Step | Variable Entered | Variable Removed | Label | Number Vars In | Partial R-Square | Model R-Square | C(p) | F Value | Pr > F |
| 1 | X1 | | service | 1 | 0.7528 | 0.7528 | 96.9776 | 267.98 | <.0001 |
| 2 | X3 | | degree | 2 | 0.0986 | 0.8514 | 25.9832 | 57.74 | <.0001 |
| 3 | X4 | | school | 3 | 0.0225 | 0.8739 | 11.3492 | 15.32 | 0.0002 |
| 4 | X5 | | grad | 4 | 0.0113 | 0.8851 | 5.0209 | 8.33 | 0.0050 |

$X_1$、$X_3$、$X_4$、$X_5$這四個變數皆為顯著，因此皆保留下來，沒有變數應被剔除,所以我們最後選擇模型的解釋變數是$X_1$(service)、$X_3$(degree) 、$X_4$(school)、$X_5$(grad)。

## 3-4 Adjusted R-Square Selection Method

表 3.5

| Number in Model | Adjusted R-Square | R-Square | AIC | Variables in Model |
|---|---|---|---|---|
| 5 | 0.8812 | 0.8879 | -560.1730 | X1 X3 X4 X5 X6 |
| 6 | 0.8798 | 0.8879 | -558.1738 | X1 X2 X3 X4 X5 X6 |
| 4 | 0.8797 | 0.8851 | -560.0088 | X1 X3 X4 X5 |
| 5 | 0.8784 | 0.8852 | -558.0600 | X1 X2 X3 X4 X5 |
| 4 | 0.8697 | 0.8756 | -552.8206 | X1 X3 X4 X6 |
| 3 | 0.8695 | 0.8739 | -553.5983 | X1 X3 X4 |

從表 3.10 發現 Adjusted R-Square Selection Method 選取變數$X_1$，$X_3$, $X_4$, $X_5$, $X_6$的組合， AIC 值最小，但是四個解釋變數的 AIC 與 Adjusted R-Square 都很接近，因此選擇變數較少的組合：

$X_1$(service)，$X_3$(degree), $X_4$(school)， $X_5$(grad)

## 3-5 CP 選取法

根據 CP 的準則:

1. CP 值要小

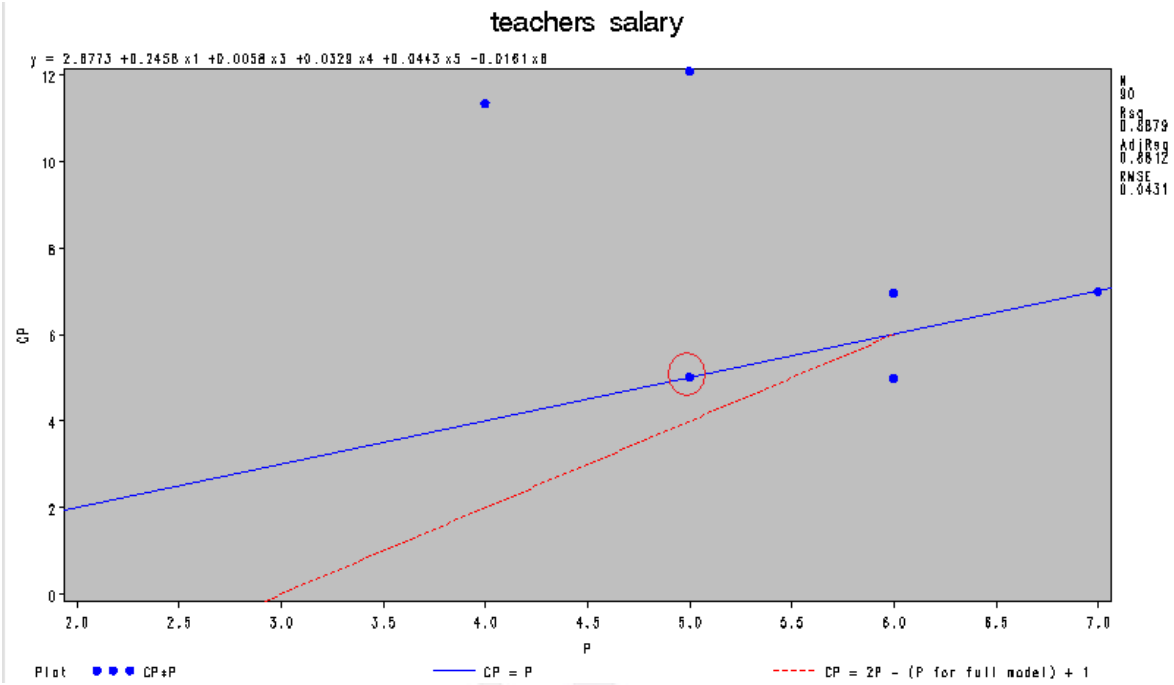
2. CP 值要接近 P(參數個數)

表 3.6

| Number in Model | Adjusted R-Square | R-Square | C(p) | Variables in Model |
|---|---|---|---|---|
| 5 | 0.8812 | 0.8879 | 5.0007 | X1 X3 X4 X5 X6 |
| 6 | 0.8798 | 0.8879 | 7.0000 | X1 X2 X3 X4 X5 X6 |
| 4 | 0.8797 | 0.8851 | 5.0209 | X1 X3 X4 X5 |
| 5 | 0.8784 | 0.8852 | 6.9725 | X1 X2 X3 X4 X5 |
| 4 | 0.8697 | 0.8756 | 12.0899 | X1 X3 X4 X6 |
| 3 | 0.8695 | 0.8739 | 11.3492 | X1 X3 X4 |

CP 值 5.0007 最小，但不滿足準則 2：CP 值要接近參數個數 6 而 CP 值 5.0209 小且又接近參數個數 5,故選取重要變數為 $X_1$(service)，$X_3$(degree), $X_4$(school) , $X_5$(grad)

圖 3.1




根據圖 3.1 可以得知，實線下方的點為 CP 最小,但是根據準則 2，CP 值要接近 P,即越接近實線：Cp = p，故選擇圈起來的點。

## ◎最終模式:

表 3.7

| 變數選取方法 | $\alpha$ | 最佳組合 |
|---|---|---|
| 向前選取法 | 0.1 | X1、X3、X4、X5 |
| 向後選取法 | 0.1 | X1、X3、X4、X5 |
| 逐步選取法 | stay：0.15；entry：0.1 | X1、X3、X4、X5 |
| Adj R-Square 選取法 | | X1、X3、X4、X5 |
| CP 選取法 | | X1、X3、X4、X5 |

綜合以上五種變數選取法，接著檢定是否剔除$X_2$(sex)、$X_6$(break)，並且留$X_1$(service)、$X_3$(degree)、$X_4$(school)、$X_5$(grad)四個變數在模型中。

## 3-6 Partial F test

表 3.8 ANOVA(partial test)

| Test 1 Results for Dependent Variable y | | | | |
|---|---|---|---|---|
| Source | DF | Mean Square | F Value | Pr > F |
| Numerator | 2 | 0.00190 | 1.01 | 0.3685 |
| Denominator | 83 | 0.00188 | | |

檢定是否考慮去除 $X_2$、$X_6$

$$\begin{cases} H_0 : \beta_2 = \beta_6 = 0 \\ H_1 : \beta_2、\beta_6 不全為 0 \end{cases}$$

檢定統計量:$F^*$近似於 1.01

檢定規則:Reject $H_0$ if

$$F^* = \frac{MSR(X_2、X_6|X_1、X_3、X_4、X_5)}{MSE(Full)} = 1.01 < F_{0.05}(2、83)$$近似於 3.107

所以$F^*$=1.01 < 3.107　　Do not reject $H_0$ at $\alpha$=0.05 level

不拒絕 $\beta_2 = \beta_6 = 0$ 的假設　　表示$\beta_2$、$\beta_6$可以都視為 0，

即可以考慮剔除 $X_2$(sex)、$X_6$(break) 這兩個變數，與選取法的結果一致!

◎$X_2$(sex)、$X_6$(break)對模型的解釋能力

表 3.9 選取後的 R-Square

| | | | |
|---|---|---|---|
| Root MSE | 0.04336 | R-Square | 0.8851 |
| Dependent Mean | 3.22004 | Adj R-Sq | 0.8797 |
| Coeff Var | 1.34668 | | |

由表可知，R-Square =0.8851 代表所選取的解釋變數有 88.51%的解釋能力，與選取前的 R-Square =0.8879 沒有甚麼差異，表示$X_2$(sex)及$X_6$(break)對模型解釋的能力很低，可以考慮去除。

由以上選取法及 Partial F test 及 R-Square 的結果，我們自模

型中剔除$X_2$(sex)、$X_5$(break)

## 3-7 選取後的模型

表 3.10

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | 2.68240 | 0.02332 | 115.05 | <.0001 |
| X1 | service | 1 | 0.24165 | 0.01101 | 21.94 | <.0001 |
| X3 | degree | 1 | 0.00564 | 0.00136 | 4.14 | <.0001 |
| X4 | school | 1 | 0.03355 | 0.00946 | 3.55 | 0.0006 |
| X5 | Grad | 1 | 0.04222 | 0.01463 | 2.89 | 0.0050 |

由以上五種重要變數選取法、Partial F test 及 R-Square 的結果，我們自模型中剔除$X_2$(sex)、$X_5$(break)，並且留$X_1$(service)、$X_3$(degree)、$X_4$(school)、$X_5$(grad)四個變數在模型中。

由表 3.10 可知，最終配適的迴歸線為:

$$\log_{10}\hat{Y}_i = 2.6824 + 0.24165 * \log_{10}X_1 + 0.00564 * X_3 + 0.03355 * X_4 + 0.04222 * X_5$$

# 四、殘差分析

## I. $E(\varepsilon_i)=0$

表 4-1：檢定 $E(\varepsilon_i)$是否=0

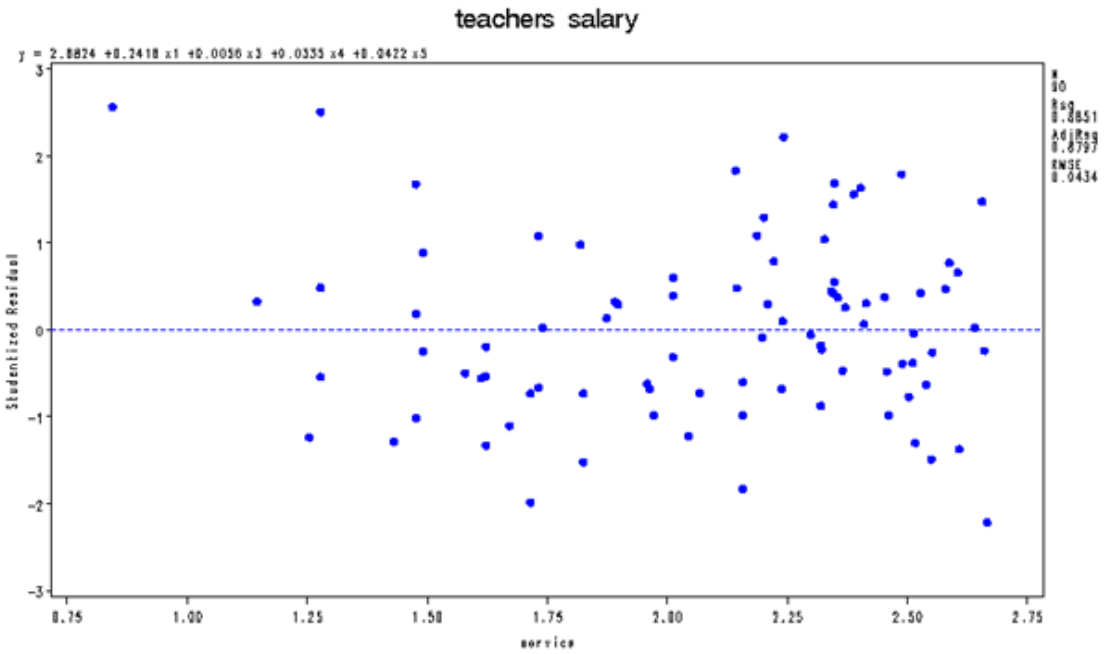
| Tests for Location: Mu0=0 | | | | |
|---|---|---|---|---|
| **Test** | **Statistic** | | **p Value** | |
| **Student's t** | **t** | 0.003924 | **Pr > \|t\|** | 0.9969 |
| **Sign** | **M** | -2 | **Pr >= \|M\|** | 0.7520 |
| **Signed Rank** | **S** | -106.5 | **Pr >= \|S\|** | 0.6707 |

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

由表4-1可知Student's t 、Sign及Signed Rank的 p-value >0.05，不拒絕$H_0$，符合基本假設 $E(\varepsilon_i)=0$。

## II. $Var(\varepsilon_i) = \sigma^2$(常數)

圖4-1 ： X1殘差圖




由圖4-1 可觀察$\varepsilon_i$變異是固定常數

## III. $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$

表4-2：自我相關檢定

| | |
|---|---|
| Durbin-Watson D | 1.625 |
| Pr < DW | 0.0250 |
| Pr > DW | 0.9750 |
| Number of Observations | 90 |
| 1st Order Autocorrelation | 0.128 |

◎殘差正自我相關檢定：

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{cases}$$ 由表：Pr < DW＝0.025，大於0.01，不拒絕虛無假設

◎殘差負自我相關檢定：

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho < 0 \end{cases}$$ 由表：Pr > DW＝0.975，大於0.01，不拒絕虛無假設

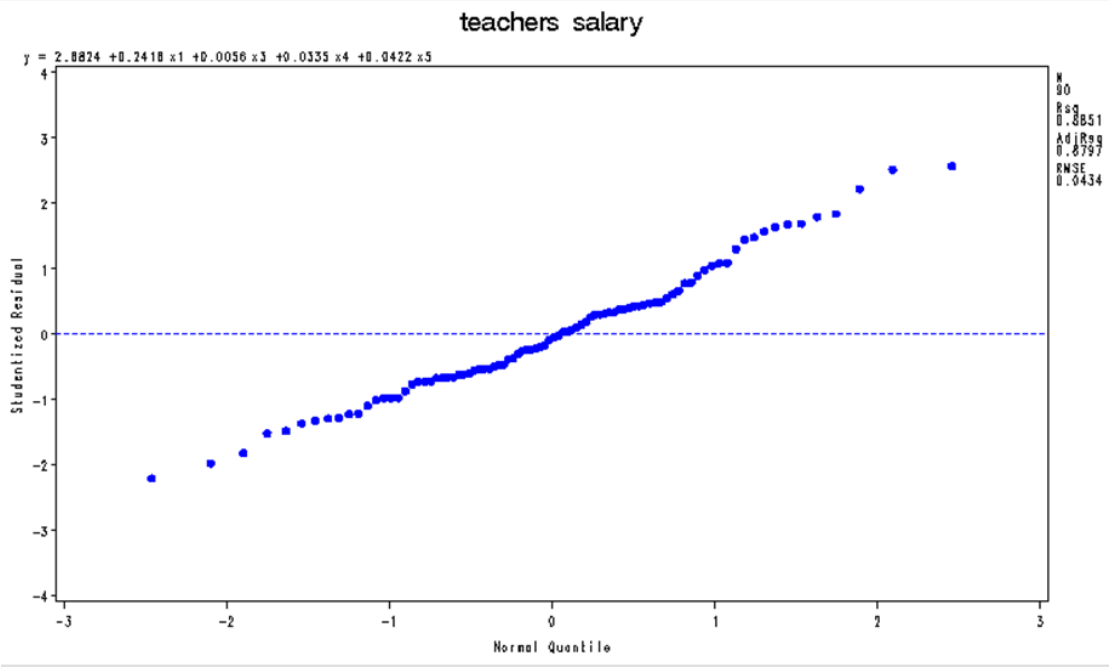
檢定結果：殘差之間無自我相關。

## IV. $\varepsilon_i$為常態分配

表 4-3 ：檢定$\varepsilon_i$是否為常態分配

| Tests for Normality | | | | |
|---|---|---|---|---|
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.983 | **Pr < W** | 0.2896 |
| **Kolmogorov-Smirnov** | **D** | 0.070085 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.078263 | **Pr > W-Sq** | 0.2224 |
| **Anderson-Darling** | **A-Sq** | 0.492776 | **Pr > A-Sq** | 0.2203 |

圖 4-2 ：殘差是否常態




$$\begin{cases} H_0：\varepsilon_i \text{服從常態分配} \\ H_1：\varepsilon_i \text{不服從常態分配} \end{cases}$$

由表4-2可看出 S-W 及 K-S 的p-value大於0.05，所以不拒絕$H_0$，圖4-2也有常態的趨勢，故$\varepsilon_i$服從常態分配。

# 五、檢查影響點&異常點

表： 影響點&異常點的檢測

| Obs | Student Residual | Cook's D | RStudent | Hat Diag H | Cov Ratio | DFFITS | DFBETAS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Intercept | x1 | x3 | x4 | x5 |
| 1 | 2.564 | 0.168 | 2.6538 | 0.1135 | 0.8002 | 0.9495 | 0.9344 | -0.8505 | 0.1147 | -0.1815 | -0.1331 |
| 2 | 0.328 | 0.002 | 0.326 | 0.0899 | 1.1584 | 0.1025 | 0.0817 | -0.0832 | 0.008 | 0.0504 | -0.0263 |
| 3 | -1.232 | 0.05 | -1.2353 | 0.1406 | 1.1283 | -0.4997 | -0.229 | 0.2505 | 0.2594 | -0.1192 | -0.35 |
| 4 | 2.508 | 0.079 | 2.5911 | 0.0591 | 0.7677 | 0.6496 | 0.603 | -0.5119 | 0.064 | -0.1964 | -0.102 |
| 5 | -0.54 | 0.01 | -0.538 | 0.1472 | 1.2229 | -0.2235 | -0.1078 | 0.0955 | 0.1105 | 0.0753 | -0.1743 |
| 7 | -1.286 | 0.021 | -1.2915 | 0.0599 | 1.0229 | -0.326 | -0.2259 | 0.2277 | -0.0163 | -0.1884 | 0.0946 |
| 9 | 1.679 | 0.025 | 1.6973 | 0.0424 | 0.936 | 0.357 | 0.3076 | -0.245 | 0.0282 | -0.1348 | -0.0591 |
| 20 | -1.98 | 0.033 | -2.0157 | 0.0404 | 0.873 | -0.4136 | -0.2061 | 0.203 | -0.0026 | -0.2787 | 0.134 |
| 25 | 0.0293 | 0 | 0.0291 | 0.0717 | 1.1428 | 0.0081 | 0.0021 | -0.0025 | -0.0027 | 0.0023 | 0.0058 |
| 40 | -0.601 | 0.011 | -0.5989 | 0.1285 | 1.1918 | -0.23 | -0.0129 | 0.0225 | -0.1748 | -0.0366 | 0.0234 |
| 41 | -1.826 | 0.119 | -1.8522 | 0.1512 | 1.0233 | -0.7819 | -0.0783 | 0.034 | -0.5595 | 0.3239 | 0.0009 |
| 46 | -0.087 | 0 | -0.0865 | 0.151 | 1.2491 | -0.0365 | -0.0027 | 0.0006 | -0.026 | 0.0152 | 0 |
| 47 | 1.296 | 0.189 | 1.301 | 0.3606 | 1.5017 | 0.977 | 0.0782 | -0.1091 | 0.9001 | 0.0792 | -0.3489 |
| 52 | 2.218 | 0.07 | 2.2711 | 0.0665 | 0.8437 | 0.6063 | -0.1222 | 0.103 | -0.2523 | 0.1535 | 0.479 |
| 55 | -0.874 | 0.055 | -0.873 | 0.2635 | 1.3769 | -0.5222 | -0.0216 | 0.022 | -0.4882 | -0.1113 | 0.3769 |
| 61 | 1.686 | 0.02 | 1.705 | 0.0345 | 0.927 | 0.3224 | -0.0939 | 0.1075 | -0.0402 | 0.2119 | -0.0901 |
| 66 | 1.564 | 0.044 | 1.5776 | 0.0826 | 0.9994 | 0.4734 | -0.0779 | 0.1238 | 0.0819 | -0.2637 | 0.2363 |
| 73 | 1.792 | 0.033 | 1.8162 | 0.0488 | 0.92 | 0.4112 | -0.1326 | 0.1461 | 0.1804 | 0.2157 | -0.2393 |
| 88 | 1.475 | 0.022 | 1.4856 | 0.0475 | 0.9784 | 0.3316 | -0.1679 | 0.2408 | -0.0446 | -0.1561 | -0.0144 |
| 90 | -2.213 | 0.122 | -2.2659 | 0.1105 | 0.8866 | -0.7987 | 0.3277 | -0.4071 | 0.2832 | 0.3887 | -0.5973 |

## (一) Cook's Distance

$$\log_{10}\hat{Y}_i = 2.6824 + 0.24165 * \log_{10}X_1 + 0.00564 * X_2 + 0.03355 * X_4 + 0.04222 * X_5$$

準則：$Di > F_{0.5}(p, n-p) \sim 1$，表示可能有影響點。

結論；沒有影響點。

## (二) DFFITS

$$\log_{10}\hat{Y}_i = 2.6824 + 0.24165 * \log_{10}X_1 + 0.00564 * X_2 + 0.03355 * X_4 + 0.04222 * X_5$$

準則：樣本數為50 >30 為大樣本， |DFFITS|＞$2\sqrt{\frac{p}{n}}$，表示有影響點。

|DFFITS|＞$2\sqrt{\frac{5}{90}}$＝0.4714

**第1點**為0.9495，**第3點**為-0.4997，第4點為0.6496，第41點為-0.7819，第47點為0.977，第52點為0.6063，第55點為-0.5222，第66點為0.4734，第90點為-0.7987 皆大於0.4714，表示這些點可能為影響點。

## （三）DFBETAS

準則：|DFBETAS|＞$\frac{2}{\sqrt{n}}$，表示可能有影響點。

|DFBETAS|＞$\frac{2}{\sqrt{90}}$＝0.2108

Intercept：**第1點**為0.9344，**第3點**為-0.229，第4點為0.603，第7點為-0.2259，第9點為0.3076，第20點為-0.2061。

$X_1$：**第1點**為-0.8505，**第3點**為0.2505，第4點為-0.5119，第7點為0.2277，第9點為0.245，第20點為0.203，第88點為0.2408 。

X3：**第3點**為0.2594，第41點為-0.5595。

X4：第20點為-0.2787，第41點為0.3239，第61點為0.2119，第66點為-0.2637，第73點為0.2157

X5：**第3點**為-0.35，第52點為0.479，第55點為0.3769，第66點為0.2363

所以**第1點**、**第3點**、第4點、第7點、第9點、第20點、第41點、第52點、第55點、第66點、第88點，皆可能為影響點。

(四) Hat value

$$\log_{10}\hat{Y}_i = 2.6824 + 0.24165 * \log_{10}X_1 + 0.00564 * X_2 + 0.03355 * X_4 + 0.04222 * X_5$$

準則：$h_{ii} > 2\frac{p}{n}$，表示可能有影響點。

$h_{ii} > 2x\frac{5}{90} = 0.1111$

**第1點**為0.1135，**第3點**為0.1406，第5點為0.1472，第40點為0.1285，第41點為0.1512，第46點為0.151，表示這些點可能為影響點。

(五) COVRATIO

準則：$COVRATIO_i > 1+3*\frac{p}{n}$ 或 $COVRATIO_i < 1-3*\frac{p}{n}$

，表示可能有影響點。

$COVRATIO_i > 1.1667$ 或 $COVRATIO_i < 0.8333$

**第1點**為0.8002，第4點為0.7677，第5點為1.2229，第40點為1.1918，第46點為1.2491，第47點為1.5017，第55點為1.3769表示這些點可能為影響點。

(六)student residual

準則：$|r_i| > 3$，表示可能有異常點。

無異常點。

(七)Rstudent

準則：$|t_i|>3$，表示有異常點。

無異常點。

結論：

由以上異常點跟影響點檢測第1點、第3點影響最為嚴重，將這兩筆資料刪除後，發現R-square上升1.41%，且殘差分析及選模結果與未刪除前都相近，所以建議刪除此兩筆資料。

以下刪除異常點後重新做迴歸分析，因方法步驟都跟前述相同，故不再多加詳述，只列出刪除異常點後的分析結果。

# 六、刪除異常點後迴歸分析

## ◎重要變數選取法與未刪除異常點之結果相同

| 變數選取方法 | $\alpha$ | 最佳組合 |
|---|---|---|
| 向前選取法 | 0.1 | X1、X3、X4、X5 |
| 向後選取法 | 0.1 | X1、X3、X4、X5 |
| 逐步選取法 | stay：0.15；entry：0.1 | X1、X3、X4、X5 |
| Adj R-Square 選取法 | | X1、X3、X4、X5 |
| CP 選取法 | | X1、X3、X4、X5 |

◎殘差分析經檢定後亦符合假設：

$$E(\varepsilon_i) = 0$$

$$Var(\varepsilon_i) = \sigma^2$$

$$Cov(\varepsilon_i, \varepsilon_j) = 0$$

$$\varepsilon_i \rightarrow Normal$$

◎最終配適迴歸線

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
| Intercept | Intercept | 1 | **2.66656** | 0.02430 | 109.75 | <.0001 | 0 |
| x1 | service | 1 | **0.24806** | 0.01141 | 21.75 | <.0001 | 1.02401 |
| x3 | degree | 1 | **0.00518** | 0.00134 | 3.86 | 0.0002 | 1.82961 |
| x4 | school | 1 | **0.03615** | 0.00919 | 3.93 | 0.0002 | 1.05897 |
| x5 | grad | 1 | **0.04854** | 0.01468 | 3.31 | 0.0014 | 1.83379 |

最終配適的迴歸線為:

$$\log_{10}\hat{Y}_i = 2.66656 + 0.24806 * \log_{10}X_1 + 0.00518 * X_3 + 0.03615 * X_4 + 0.04854 * X_5$$

◎多重共線性檢查

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
| Intercept | Intercept | 1 | 2.66656 | 0.02430 | 109.75 | <.0001 | 0 |
| x1 | service | 1 | 0.24806 | 0.01141 | 21.75 | <.0001 | **1.02401** |
| x3 | degree | 1 | 0.00518 | 0.00134 | 3.86 | 0.0002 | **1.82961** |
| x4 | school | 1 | 0.03615 | 0.00919 | 3.93 | 0.0002 | **1.05897** |
| x5 | grad | 1 | 0.04854 | 0.01468 | 3.31 | 0.0014 | **1.83379** |

VIF 準則：Variance Inflation Factor＞10，表示變數間有高度。

上表VIF 值皆小於10，表示變數間無多重共線性的問題。

# 七、結論

由上面的分析我們可以發現重要變數為教師任教的年資（$X_1$）、學位($X_3$)、任教學校($X_4$)、研究所學位($X_5$)。

最佳模式為：

$$\log_{10}\hat{Y}_1 = 2.66656 + 0.24806 * \log_{10}X_1 + 0.00518 * X_3 + 0.03615 * X_4 + 0.04854 * X_5$$

還原後

$$\hat{Y}_1 = 10^{2.66656} * X_1^{0.24806} * 10^{0.00518*X_3} * 10^{0.03615*X_4} * 10^{0.04854*X_5}$$

由迴歸模型可知：

教師任教的年資（$X_1$)增加一倍將使一年的薪資(Y)增加 $2^{0.24806} - 1 \approx 18.76\%$，學校對於資歷深越深的教師，給予較高的薪資；

學位($X_3$)增加1單位將使一年的薪資(Y)增加$10^{0.00518} - 1 \approx 1.20\%$，符合學位越高薪資越高的直覺；

任教學校($X_4$)為公立比私立的多$10^{0.03615} - 1 \approx 8.68\%$的薪資，公立學校給予平均較高的薪資；

有研究所學位($X_5$)比沒有研究所學位的多$10^{0.04854} - 1 \approx 11.18\%$的薪資，可見有無研究所學位對薪資的影響幅度蠻大的。

# 八、參考文獻

1.

彭照英.唐麗

SAS 1-2-3 , 5th

儒林圖書公司

2.

Kutner , M. H.  Nachtsheim , C. J.  Neter J.

APPLIED  LINEAR  REGRESSION  MODELS

Mc Graw Hill

3.

Journal of Royal Statistical Society,

P.Turnbull、G.Williams

# 附錄

**附錄一、SAS 程式(取 log 前)**

**附錄二、SAS 程式(取log 後)**

## 附錄一

```
dm "output;clear;log;clear;program;recall;graph;cler;";
options ps=55;
title 'teachers salary ';
data salary;
input y X1 X2 X3 X4 X5 X6;
label y='salary'  X1='service'  X2='sex' X3='degree' X4='school' X5='grad' X6='break' ;
cards;
980.2 7    0    0    0    0    0
1015  14   1    0    1    0    0
1028  18   1    0    1    1    0
1250  19   0    0    0    0    0
1028  19   0    0    0    1    0
1028  19   0    0    0    0    0
1018  27   0    0    1    0    0
1072  30   0    0    1    0    0
1290  30   1    0    0    0    0
1204  30   0    0    1    0    0
1352  31   1    4    1    1    0
1204  31   0    0    0    0    0
1104  38   0    0    0    0    0
1118  41   0    0    0    0    0
1127  42   1    0    1    0    0
1259  42   1    0    1    0    0
1127  42   0    0    0    0    0
1127  42   1    0    0    0    0
1095  47   0    0    0    0    1
1113  52   0    0    1    0    1
1462  52   1    4    1    1    0
1182  54   0    0    0    0    0
1404  54   0    0    0    0    0
1182  54   0    0    0    0    0
1594  55   0    4    1    1    0
1459  66   0    0    0    0    0
```

| | | | | | | |
|---|---|---|---|---|---|---|
| 1237 | 67 | 0 | 0 | 0 | 0 | 0 |
| 1237 | 67 | 1 | 0 | 1 | 0 | 0 |
| 1496 | 75 | 1 | 0 | 1 | 0 | 0 |
| 1424 | 78 | 0 | 0 | 0 | 0 | 0 |
| 1424 | 79 | 1 | 0 | 0 | 0 | 0 |
| 1347 | 91 | 0 | 0 | 0 | 0 | 0 |
| 1343 | 92 | 0 | 0 | 0 | 0 | 1 |
| 1310 | 94 | 1 | 0 | 0 | 0 | 0 |
| 1814 | 103 | 1 | 4 | 0 | 1 | 0 |
| 1534 | 103 | 0 | 0 | 0 | 0 | 0 |
| 1430 | 103 | 0 | 0 | 0 | 0 | 0 |
| 1439 | 111 | 1 | 0 | 1 | 0 | 0 |
| 1946 | 144 | 0 | 9 | 1 | 1 | 0 |
| 2216 | 144 | 1 | 16 | 1 | 1 | 0 |
| 1834 | 144 | 0 | 16 | 0 | 1 | 1 |
| 1416 | 117 | 1 | 0 | 0 | 0 | 1 |
| 2052 | 139 | 0 | 0 | 1 | 0 | 0 |
| 2087 | 140 | 1 | 4 | 1 | 1 | 1 |
| 2264 | 154 | 0 | 4 | 1 | 1 | 1 |
| 2201 | 158 | 0 | 16 | 0 | 1 | 1 |
| 2992 | 159 | 1 | 25 | 1 | 1 | 1 |
| 1695 | 162 | 1 | 0 | 0 | 0 | 0 |
| 1792 | 167 | 1 | 0 | 0 | 0 | 0 |
| 1690 | 173 | 0 | 0 | 1 | 0 | 1 |
| 1827 | 174 | 0 | 0 | 1 | 0 | 1 |
| 2604 | 175 | 0 | 4 | 1 | 1 | 0 |
| 1720 | 199 | 0 | 0 | 0 | 0 | 0 |
| 1720 | 209 | 0 | 0 | 0 | 0 | 0 |
| 2159 | 209 | 1 | 16 | 1 | 0 | 0 |
| 1852 | 210 | 1 | 0 | 1 | 0 | 0 |
| 2104 | 213 | 1 | 0 | 1 | 0 | 0 |
| 1852 | 220 | 0 | 0 | 0 | 0 | 1 |
| 1852 | 222 | 0 | 0 | 0 | 0 | 0 |
| 2210 | 222 | 0 | 0 | 1 | 0 | 0 |
| 2266 | 223 | 1 | 0 | 1 | 0 | 0 |
| 2027 | 223 | 1 | 0 | 1 | 0 | 0 |
| 1852 | 227 | 0 | 0 | 0 | 0 | 0 |
| 1852 | 232 | 0 | 0 | 1 | 0 | 1 |
| 1995 | 235 | 1 | 0 | 1 | 0 | 1 |

```
2616  245  0  9  0  1  0
2324  253  1  0  1  0  0
1852  257  0  0  0  0  1
2054  260  0  0  1  0  0
2617  284  1  9  1  1  0
1948  287  1  0  1  0  0
1720  290  0  0  0  0  1
2604  308  0  4  1  0  0
1852  309  0  0  0  0  1
1942  319  1  0  1  0  0
2027  325  0  0  1  0  0
1942  326  0  0  0  0  0
1720  329  0  0  0  0  0
2048  337  1  0  0  0  0
2334  346  0  4  1  1  1
1720  355  0  0  0  0  1
1942  357  0  0  0  0  0
2117  380  0  0  0  0  1
2742  387  1  4  1  1  1
2740  403  0  4  1  1  1
1942  406  0  0  1  0  0
2266  437  0  0  1  0  0
2436  453  0  0  0  0  0
2067  458  0  0  0  0  0
2000  464  0  4  0  1  0
;
ods html;
ods graphics on;
proc gplot;
plot y*(X1 X3);
proc reg outest=est;
FULL:  MODEL  y=X1 - X6/vif;
run;

ods graphics off;
ods html close;
run;
quit;
```

## 附錄二(取log之後)

```
dm "output;clear;log;clear;program;recall;graph;cler;";
options ps=55;
title 'teachers salary ';
data salary;
input y X1 X2 X3 X4 X5 X6;
y=log10(y);
X1=log10(X1);

label y='salary'  X1='service'  X2='sex' X3='degree' X4='school' X5='grad' X6='break';
cards;
```

與上方的原始資料相同(略)

```
ods html;
ods graphics on;

proc gplot;
plot y*(x1 x3);

proc corr ;
var y x1 - x6 ;
run;

proc reg outest=est;
FULL:    MODEL y=x1 x2 x3 x4 x5 x6/vif;
M1:      MODEL y=x1 x2 x3 x4 x5 x6/SELECTION=FORWARD SLE=0.1;
M2:      MODEL y=x1 x2 x3 x4 x5 x6/SELECTION=BACKWARD SLS=0.1;
M3:      MODEL y=x1 x2 x3 x4 x5 x6/SELECTION=STEPWISE SLS=0.15 SLE=0.1;
M4:      MODEL y=x1 x2 x3 x4 x5 x6/SELECTION=adjrsq aic;
M5:      MODEL y=x1 x2 x3 x4 x5 x6/SELECTION=cp;

plot cp.*np.
          / chocking=red  cmallows=blue
            vaxis= 0 to 12 by 2 cframe=ligr;
symboll v=dot  c=blue;
run;
```

```
proc reg data=salary;
M1:      MODEL y=x1 x3 x4 x5 /vif DWPROB;
symboll v=dot  c=blue;
run;


proc reg data=salary;
Improve: model y=x1 x3 x4 x5/p r influence clm cli dw;
output out=all student=student rstudent=rstudent;
plot y* (x1);
plot y* (x1) / pred;
plot (student. rstudent. )*(x1 predicted.);
plot student. * nqq./vaxis=-4 to 4 by 1 haxis=-3 to 3 by 1.0;
symboll v=dot  c=blue;
run;

proc univariate normal plot data=all;
var student;
histogram student / cfill=ltgray  kernel(color=red)  name='HIST';
run;

ods graphics off;
ods html close;
run;
quit;
```

**附錄三、會議記錄**

第一次
組員討論時間：12/06　中午12:00~13:00
地點：人言203
組員：施珮玉、錢佩君、李育瑄、徐戎萱、郭根連 、吳憶瀅 、陳弘明
討論內容：解釋變數、散佈圖、full model、相關係數

第二次
組員討論時間：12/15　晚上6:00~8:30
地點：福星路麥當勞
組員：施珮玉、錢佩君、李育瑄、徐戎萱、郭根連 、吳憶瀅 、陳弘明
討論內容：矩陣散佈圖、檢查是否多重共線性、選擇重要變數、殘差檢查

第三次
組員討論時間：12/19　晚上5:30~8:00
地點：商學十樓
組員：施珮玉、錢佩君、李育瑄、徐戎萱、郭根連 、吳憶瀅 、陳弘明
討論內容：如何報告與工作分配

第四次
組員討論時間：12/23　早上11:00~12:00
地點：商學十樓
組員：施珮玉、錢佩君、李育瑄、徐戎萱、郭根連 、吳憶瀅 、陳弘明
討論內容：如何報告

第五次
組員討論時間：1/4　晚上6:30~10:00
地點：商學十樓
組員：施珮玉、錢佩君、李育瑄、徐戎萱、郭根連 、吳憶瀅 、陳弘明
討論內容：報告演練