



國立臺灣大學

National Taiwan University

使用R語言進行資料分析 Using R for Data Analysis

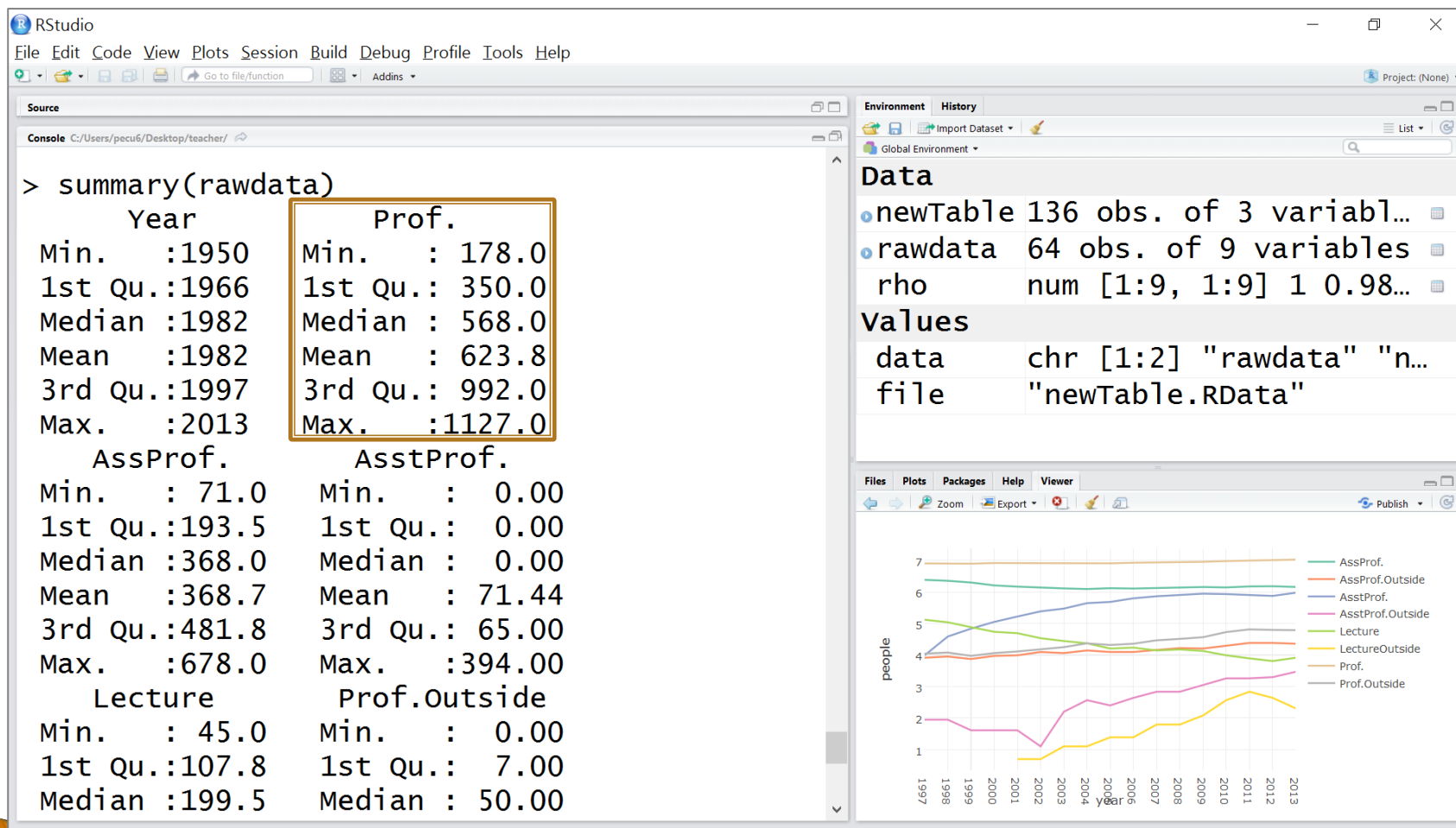
國立臺灣大學共同教育中心

助理教授 蔡芸琇


Chapter 05

» 多變量分析

統計敘述



設計問題

- ▶ 哪一年發生了1127 位正教授的高峰？（有意義？）
 - ▶ 哪一年增聘了最多教授？
 - ▶ 增聘最多教授的那一年，學校經費預算是否有增加？
 - ▶ 哪一年減少了最多教授？
 - ▶ ...
- 

討論變數間的相關性，但不包含因果關係

- ▶ 目的：主要衡量兩變數間線性關聯性的高低程度
- ▶ 方法：相關係數
- ▶ 公式說明：<http://wiki.mbalib.com/zh-tw/%E7%9B%B8%E5%85%B3%E7%B3%BB%E6%95%B0>
- ▶ R 語言：<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cor.html>

相關係數分析

- ▶ 變項間的相關程度高或低，得到的相關係數只能說明這兩個變項間是正相關、負相關，或者是無關。
- ▶ 相關程度之高低，在正負0.3之間（即0.3至-0.3之間）稱為低度相關；在正負0.3-0.6之間（即指介於0.3至0.6，-0.3至-0.6之間）稱為中度相關；而在正負0.6至0.9之間（即指在0.6至0.9，-0.6至-0.9之間）則稱為高度相關；若是為正負1，即表示完全相關；若是為0，即表示無關。

基本函數	意義
cor()	計算相關係數
cor.test()	相關係數分析

討論變數間的相關性，但不包含因果關係

- ▶ 相關係數值介於 -1 至 1 之間。
- ▶ 相關係數值 $= -1$ ：兩變數為完全負相關。
- ▶ $-1 < \text{相關係數值} < 0$ ：兩變數為負相關。
- ▶ 相關係數值 $= 0$ ：兩變數為無相關。
- ▶ $0 < \text{相關係數值} < 1$ ：兩變數為正相關。
- ▶ 相關係數值 $= 1$ ：兩變數為完全正相關。

數據閱讀

RStudio Source Editor

rho x

Filter

	Year	Prof.	AssProf.	AsstProf.	Lecture	Prof.Outside	AssProf.Outside	AsstProf.Outside	LectureOutside
Year	1.0000000	0.9857888	0.84621975	0.7485752	-0.31570368	0.9556784	0.9755560	0.7841485	0.6041368
Prof.	0.9857888	1.0000000	0.84640694	0.7411748	-0.39576535	0.9239644	0.9536310	0.7261133	0.5783464
AssProf.	0.8462198	0.8464069	1.00000000	0.3132312	0.05812416	0.7656979	0.8070489	0.6679175	0.2259859
AsstProf.	0.7485752	0.7411748	0.31323116	1.0000000	-0.75618717	0.7755827	0.7411235	0.6298275	0.8568931
Lecture	-0.3157037	-0.3957653	0.05812416	-0.7561872	1.00000000	-0.3406727	-0.3179593	-0.1832406	-0.6259633
Prof.Outside	0.9556784	0.9239644	0.76569787	0.7755827	-0.34067270	1.0000000	0.9838035	0.8960301	0.7329668
AssProf.Outside	0.9755560	0.9536310	0.80704889	0.7411235	-0.31795927	0.9838035	1.0000000	0.8141089	0.6370524
AsstProf.Outside	0.7841485	0.7261133	0.66791754	0.6298275	-0.18324064	0.8960301	0.8141089	1.0000000	0.7170201
LectureOutside	0.6041368	0.5783464	0.22598586	0.8568931	-0.62596333	0.7329668	0.6370524	0.7170201	1.0000000

Showing 1 to 9 of 9 entries

數據閱讀 (每年師資變化情況)

RStudio Source Editor

rhodiff x

Filter

	Prof.	AssProf.	AsstProf.	Lecture	Prof.Outside	AssProf.Outside	AsstProf.Outside	LectureOutside
Prof.	1.00000000	0.182781012	-0.16609075	-0.08968256	0.04257345	0.03919189	0.04559873	-0.030178343
AssProf.	0.18278101	1.000000000	-0.58656357	0.24894658	-0.03096503	-0.20260016	0.17895141	-0.008136201
AsstProf.	-0.16609075	-0.586563569	1.00000000	-0.35470678	0.02644027	0.02561974	0.06515885	-0.101601231
Lecture	-0.08968256	0.248946575	-0.35470678	1.00000000	0.12750541	0.02817573	0.25332165	-0.126145720
Prof.Outside	0.04257345	-0.030965034	0.02644027	0.12750541	1.00000000	0.76558914	0.66250191	0.516472720
AssProf.Outside	0.03919189	-0.202600155	0.02561974	0.02817573	0.76558914	1.00000000	0.09987466	0.314732145
AsstProf.Outside	0.04559873	0.178951408	0.06515885	0.25332165	0.66250191	0.09987466	1.00000000	0.088347644
LectureOutside	-0.03017834	-0.008136201	-0.10160123	-0.12614572	0.51647272	0.31473214	0.08834764	1.000000000

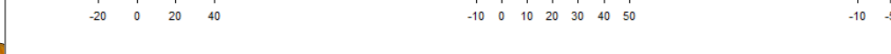
Showing 1 to 8 of 8 entries

數據閱讀

- ▶ 引起動機：<http://news.tvbs.com.tw/life/652591>
- ▶ 學生希望師資能再提升，但台灣真的能吸引好的師資嗎？根據了解以鄰近亞洲國家開出的薪資條件，日本、香港都是台灣的2到3倍，新加坡至少3倍起跳，而大陸通常是4到5倍，而這還不包括研究經費，而原本政府編列5年5百億元的預算，台大能分到31億元，但預算卻一直刪減，現在只剩下16億元，讓台灣大學副校長同時也是準教育部次長的陳良基坦言，台灣沒錢也沒條件跟人家搶人才。
- ▶ Q：臺大師資的吸引力如何？(少了和其他學校的比較)
- ▶ Prof. 與 Year 的相關係數最高，代表正教授是逐年應聘人數增加的趨勢
- ▶ Prof. 與 Lecture 變化的相關係數是負值，代表正教授增聘時，講師的聘任減少
- ▶ Prof. 與 Lecture Outside 變化的相關係數是負值，代表正教授增聘時，合聘講師減少

討論變數間的相關性，包含因果關係

- ▶ 目的：解釋資料過去的現象
 - 由自變數來預測依變數未來可能產生之數值。
 - 簡單線性迴歸分析是用一直線來解釋一個自變數 (因, x) 與一個依變數 (果, y) 的關係。
 - 例如，利率的變化影響股價的漲跌，股價即為依變數，而利率就是自變數。利率的變動是因，股價的波動為果。
- ▶ R 語言：<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html>
- ▶ 參考資料：<http://molecular-service-science.com/2012/09/12/statistics-regression/>



TIMSS & PIRLS International Study Center

- ▶ 國際數學與科學教育成就趨勢調查
- ▶ 收集台灣 2011 年八年級學生問卷資料
- ▶ 以數學能力為依變數 (果, y)，以
 - 性別、數學投入、數學興趣、教育資源與父母教育程度為自變數 (因, x)
- ▶ <http://myweb.ncku.edu.tw/~cpcheng/Rbook/03/data/TIMSS2011TW.txt>

TIMSS & PIRLS International Study Center

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins Project: (None)

ch03utf.R dta TIMSS2011TW.txt

	gender	math	math.interest	math.evaluation	math.input	math.hours	science	science.interest	science.evaluation	science.input	science.hours	parental.education	educational.resources
1	girl	729.3937	8.93041	9.34439	9.15641	45min - 3hours	682.7541	7.64598	9.15956	8.30491	<= 45min	high school	9.60097
2	girl	776.1965	13.46507	9.34439	12.42205	<= 45min	663.3682	7.64598	6.52660	7.91938	<= 45min	high school	8.91919
3	girl	718.1735	9.60333	10.35139	10.15325	<= 45min	667.1151	9.41832	11.09147	9.54897	<= 45min	elementary school	6.33067
4	girl	607.1847	13.46507	10.35139	8.70884	<= 45min	575.0923	9.41832	7.60129	7.91938	<= 45min	junior high school	10.25396
5	girl	658.1759	8.26761	8.20673	7.85736	45min - 3hours	649.7578	9.03893	8.23142	8.69954	<= 45min	university above	10.92551
6	girl	478.5763	6.36452	7.27410	7.85736	<= 45min	491.3467	12.94370	9.15956	10.57281	<= 45min	college	10.92551
7	girl	675.6004	10.80246	9.81093	9.15641	<= 45min	573.9230	8.32900	8.84836	8.69954	<= 45min	university above	11.64917
8	girl	601.0425	7.91186	7.88323	7.42878	45min - 3hours	531.5194	6.42609	8.54023	6.62923	<= 45min	junior high school	9.60097
9	girl	628.6167	9.96500	6.97345	10.15325	<= 45min	592.5403	12.94370	6.91883	12.10159	<= 45min	high school	10.92551
10	girl	639.9817	10.35777	8.20673	10.15325	<= 45min	582.9994	9.03893	8.23142	9.11017	45min - 3hours	high school	9.60097
11	girl	510.3174	9.60333	9.34439	9.63460	<= 45min	493.1658	7.64598	8.54023	7.91938	<= 45min	university above	11.64917
12	girl	499.4729	11.33851	7.88323	10.15325	<= 45min	468.6448	6.42609	6.05617	5.23495	45min - 3hours	high school	8.91919
13	girl	716.3315	13.46507	11.94250	14.34298	<= 45min	632.3760	9.41832	11.77900	11.22017	45min - 3hours	high school	10.92551
14	boy	599.8161	7.03246	5.96004	9.63460	<= 45min	558.7069	4.51274	6.91883	3.55609	<= 45min	high school	9.60097
15	boy	642.5979	13.46507	13.70742	11.44903	45min - 3hours	505.3455	6.89527	5.40984	7.91938	<= 45min	high school	8.91919
16	boy	661.8374	9.60333	8.20673	10.15325	<= 45min	597.0517	6.89527	7.60129	7.52384	<= 45min	university above	12.51743
17	boy	603.0308	9.26151	8.92971	9.15641	<= 45min	620.7838	9.03893	9.47921	8.30491	45min - 3hours	junior high school	9.60097
18	boy	752.2938	13.46507	8.92971	9.15641	45min - 3hours	666.1333	10.27988	9.47921	9.11017	45min - 3hours	junior high school	8.91919
19	boy	590.1165	9.96500	7.88323	8.27843	45min - 3hours	541.8461	7.98897	7.91985	7.10299	45min - 3hours	high school	8.91919
20	boy	486.3618	9.96500	8.92971	9.15641	<= 45min	478.6224	8.67804	7.60129	7.91938	<= 45min	junior high school	8.19314
21	boy	493.1744	13.46507	8.55351	11.44903	<= 45min	396.0681	5.78990	6.05617	5.23495	45min - 3hours	high school	8.19314
22	boy	555.6859	10.35777	8.13449	8.27843	45min - 3hours	532.0276	9.03893	8.54023	7.52384	<= 45min	high school	8.19314
23	boy	577.6692	8.93041	8.92971	9.15641	<= 45min	575.2649	9.41832	10.17921	9.11017	<= 45min	high school	9.60097
24	boy	485.9597	10.80246	7.57539	14.34298	<= 45min	558.4281	12.94370	9.81631	13.83345	<= 45min	high school	11.64917
25	boy	442.7694	8.93041	7.57539	7.42878	45min - 3hours	448.7002	10.82037	9.15956	8.62591	<= 45min	high school	8.91919
26	boy	478.7631	8.93041	6.97345	8.70884	45min - 3hours	475.8722	9.41832	8.84836	9.54897	<= 45min	high school	9.60097
27	boy	403.2912	9.96500	8.49785	10.15325	>= 3hours	400.6288	12.94370	8.84836	10.02885	<= 45min	high school	8.91919
28	boy	652.9431	9.96500	7.88323	10.74331	<= 45min	591.2669	7.64598	7.60129	7.52384	<= 45min	high school	8.91919
29	boy	475.2748	5.03748	5.46149	3.76515	<= 45min	501.4180	9.41832	6.91883	8.30491	<= 45min	high school	10.25396

Showing 1 to 29 of 4,467 entries

Console

Environment History

Global Environment

Data

- dta 4467 ob...
- dt... 4467 ob...
- fi... 4467 ob...

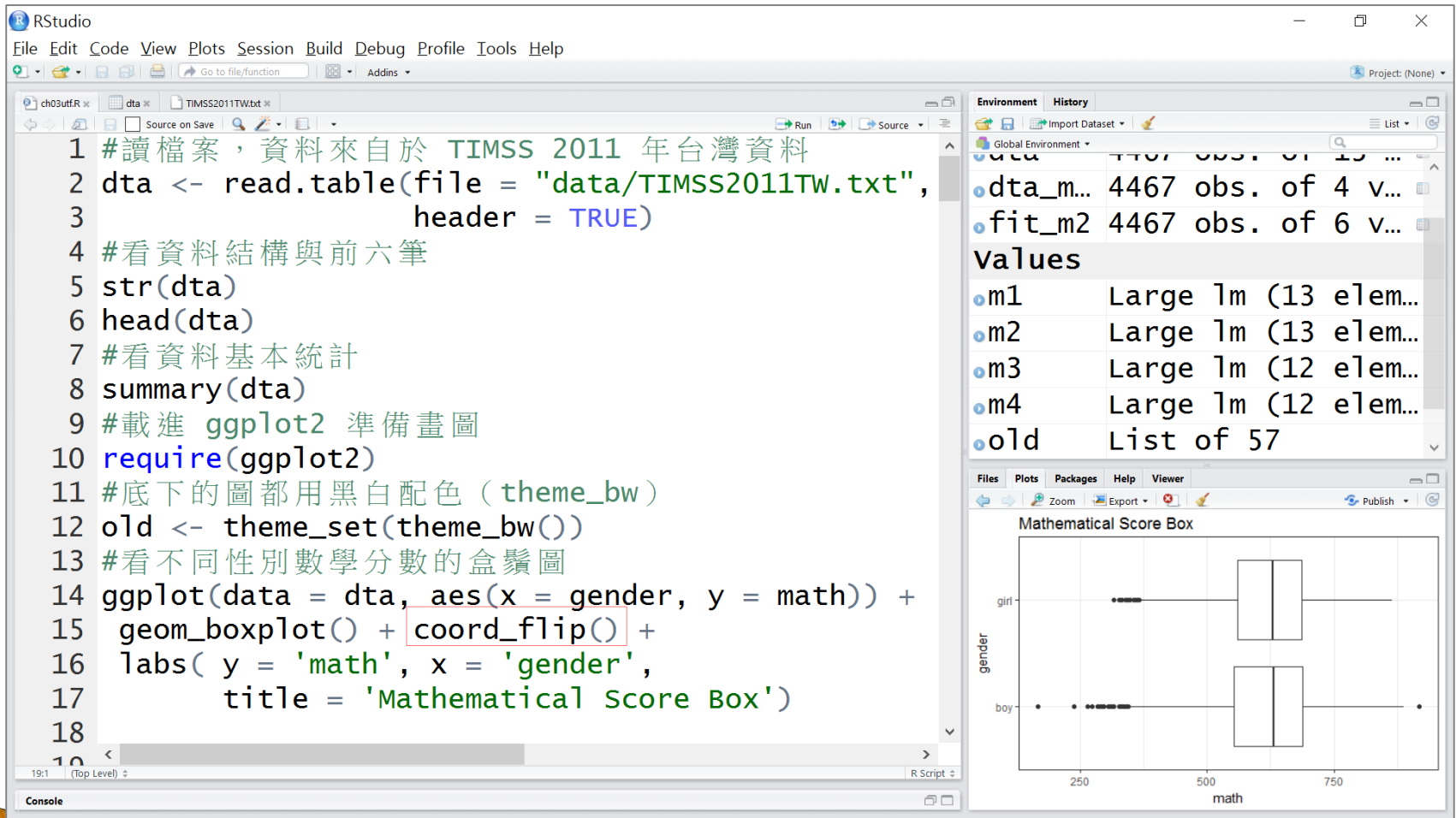
Values

- m1 Large 1m...
- m2 Large 1m...
- m3 Large 1m...
- m4 Large 1m...

Files Plots Packages Help Viewer

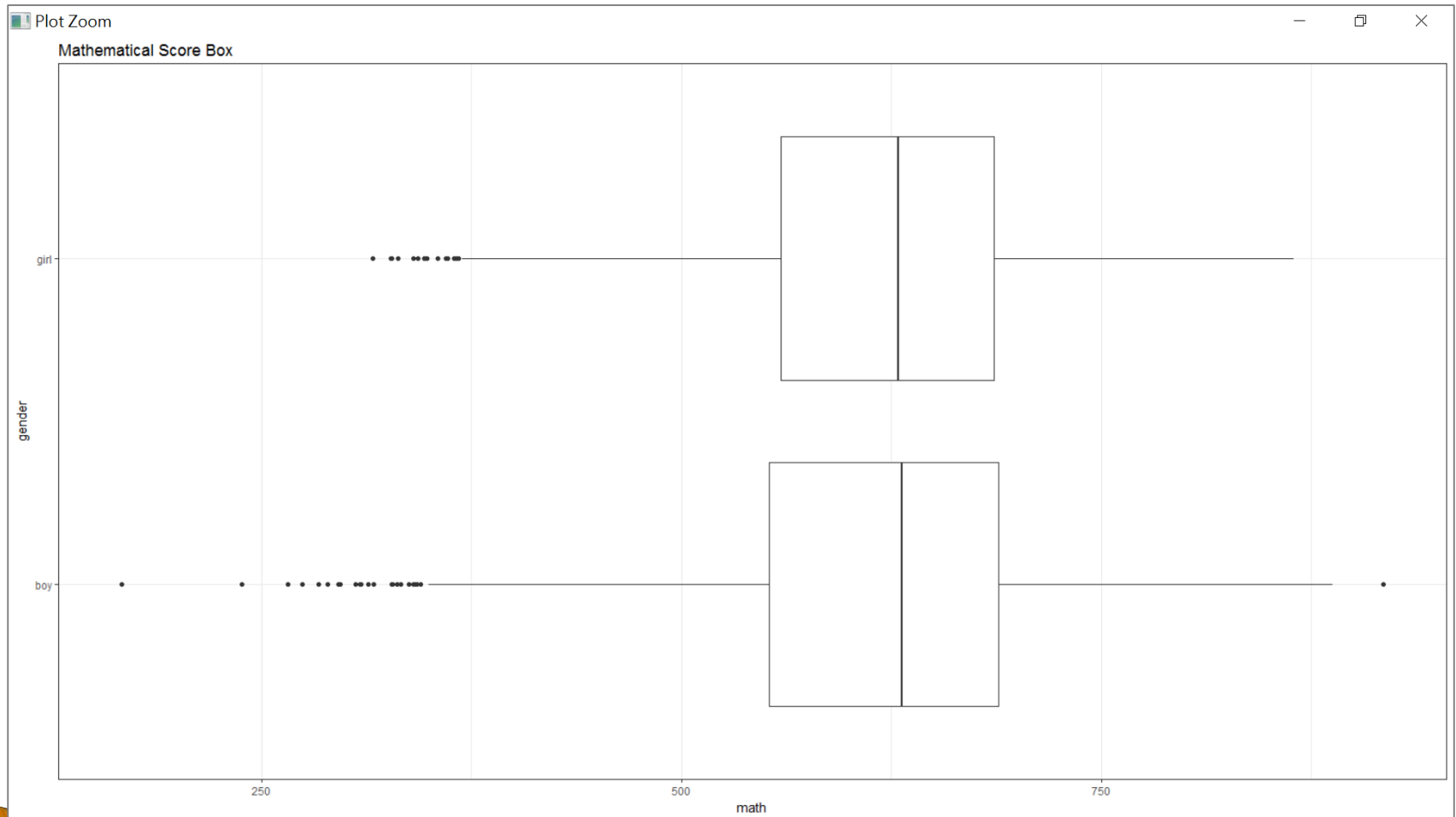
Zoom Export

TIMSS & PIRLS International Study Center



如果要改成水平方向，可以使用：`coord_flip()`。

TIMSS & PIRLS International Study Center



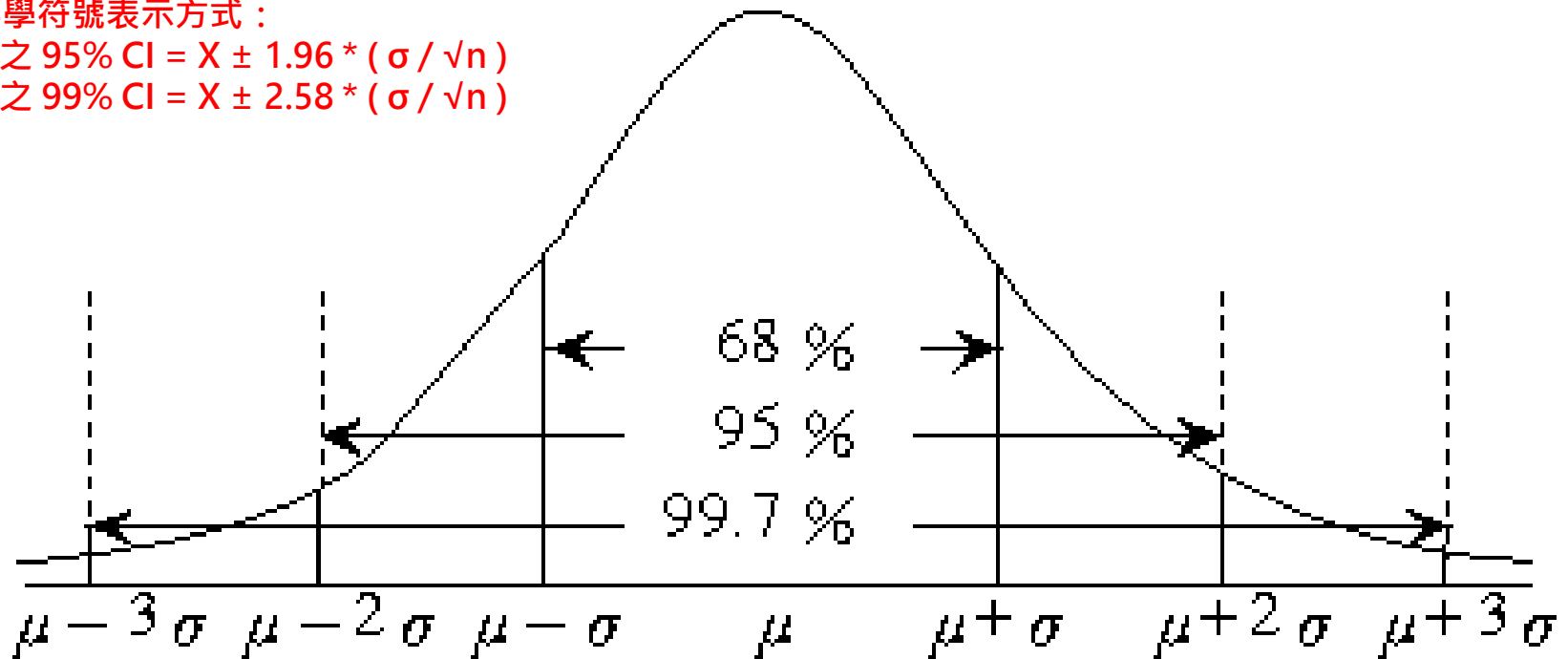
數學分數因性別差異是不顯著的

Confidence Interval, CI

科學符號表示方式：

μ 之 95% CI = $X \pm 1.96 * (\sigma / \sqrt{n})$

μ 之 99% CI = $X \pm 2.58 * (\sigma / \sqrt{n})$



TIMSS & PIRLS International Study Center

The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains R code for calculating confidence intervals for math scores by gender.
- Console:** Shows the execution of the code and the resulting confidence interval estimates for boys and girls.
- Environment:** Lists the objects created in the global environment, including 'dta', 'fit', and 'm1' through 'm4'.
- Plots:** A box plot titled 'Mathematical Score Box' showing the distribution of math scores for girls and boys.

```
#看信賴區間
with(dta,
      tapply(math, gender,
              function(x)
                c(mean(x) + c(-2, 2) * sd(x)/sqrt(length(x))))))
```

Console Output:

```
> with(dta,
+       tapply(math, gender,
+               function(x)
+                 c(mean(x) + c(-2, 2) * sd(x)/sqrt(length(x))))))
$boy
[1] 612.0913 621.0584

$girl
[1] 615.4899 623.5705

>
```

Environment:

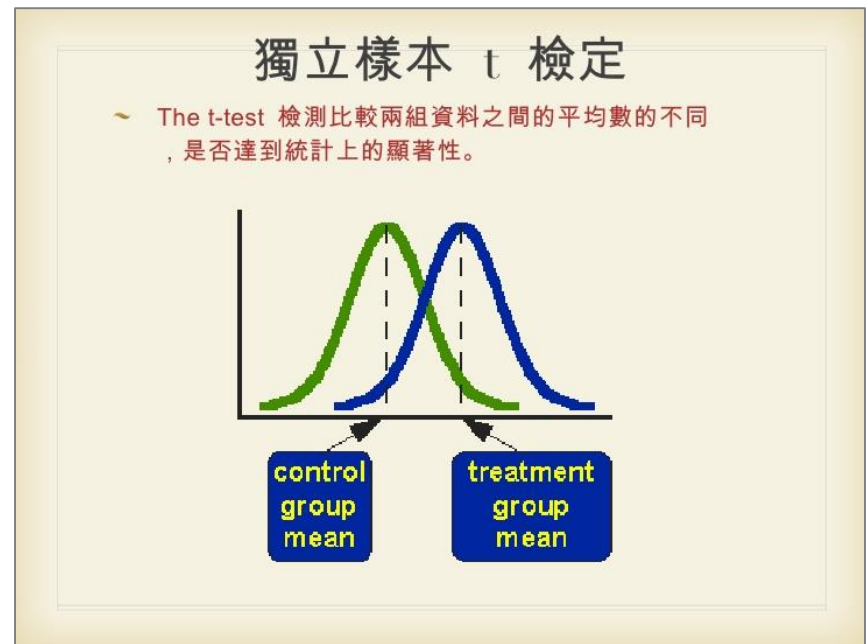
- dta... 4467 obs. o...
- fit... 4467 obs. o...
- m1 Large 1m (1...
- m2 Large 1m (1...
- m3 Large 1m (1...
- m4 Large 1m (1...
- old List of 57

Mathematical Score Box

Box plot showing the distribution of math scores for girls and boys. The y-axis is labeled 'gender' with categories 'girl' and 'boy'. The x-axis is labeled 'math' with values 250, 500, and 750. The plot shows that boys generally have higher math scores than girls, with a wider distribution.

T-Test

- ▶ T-Test 是對兩樣本平均數 (mean) 差別的顯著性進行檢驗。
- ▶ $H_0: \mu_1 = \mu_2$
- ▶ T-Test 須知道兩個總體的變異數 (Variances) 是否相等。
- ▶ T-Test 值的計算會因變異數是否相等而有所不同。



TIMSS & PIRLS International Study Center

The screenshot displays the RStudio interface. The main editor window contains R code for a t-test. The console shows the output of the `t.test` function. The Environment pane on the right lists objects in the global environment. The Plots pane at the bottom right shows a box plot titled 'Mathematical Score Box'.

```
30
31 #以t檢定比較不同性別的數學差異
32 #預設作法會做 welch 校正，處理兩樣本變異數不相同的問題
33 t.test(math ~ gender, data = dta)
34 #這是一般假設變異數同值下的 t 檢定
35 t.test(math ~ gender, data = dta, var.equal = TRUE)
36 |
37 <
```

```
> t.test(math ~ gender, data = dta)

Welch Two Sample t-test

data:  math by gender
t = -0.97932, df = 4414, p-value = 0.3275
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.871550  2.960942
sample estimates:
```

Environment: Global Environment
dta... 4467 obs. o...
fit... 4467 obs. o...
Values
m1 Large 1m (1...
m2 Large 1m (1...
m3 Large 1m (1...
m4 Large 1m (1...
old List of 57

Mathematical Score Box

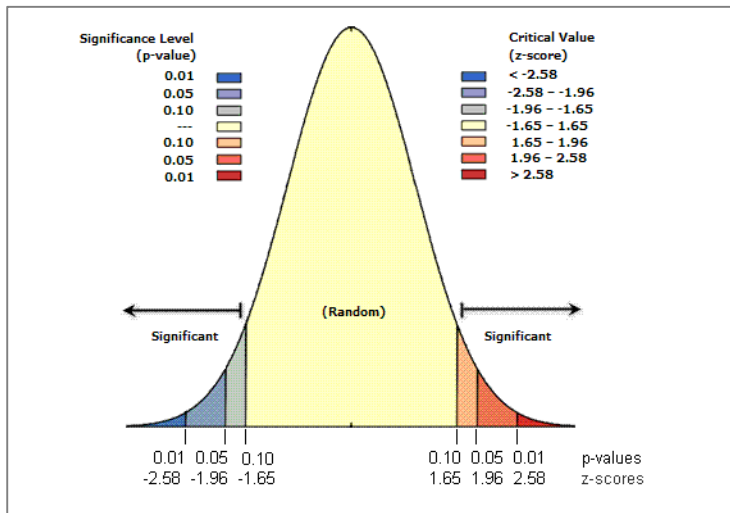
girl

boy

250 500 750

math

P-Value



- ▶ 因為 p-value = 0.3275，該值遠大於 $1 - 95\% = 0.05$ 。
- ▶ 因為 95% 信賴區間為 $(-8.87155, 2.96094)$ 。
- ▶ 這兩者都代表無法否認虛無假設 H_0 。【接受】

父母教育程度與數學成績的關係

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

ch03utf.R dta TIMSS2011TW.txt

```
29 #看不同父母教育背景者的數學成績差異
30 #先把父母教育各個水準順序定下來
31 dta$parental.education <- factor(dta$parental.education,
32                                 levels = c('elementary school',
33                                             'junior high school',
34                                             'high school',
35                                             'college',
36                                             'university above'))
37 #看不同父母教育程度下的數學分數平均數
38 tapply(dta$math, dta$parental.education, mean)
39 <
```

40:1 (Top Level)

Console C:/Users/pecu/Desktop/03/

```
+                                     college,
+                                     'university above'))
> tapply(dta$math, dta$parental.education, mean)
elementary school junior high school high school
          536.5940           558.7106           598.8742
      college university above
          645.2816           660.9434
```

Environment History

Global Environment

Data

- dta 4467 obs. o...
- dta... 4467 obs. o...
- fit... 4467 obs. o...

Values

- m1 Large 1m (1...
- m2 Large 1m (1...
- m3 Large 1m (1...
- m4 Large 1m (1...

Files Plots Packages Help Viewer

Mathematical Score Box

girl

boy

math

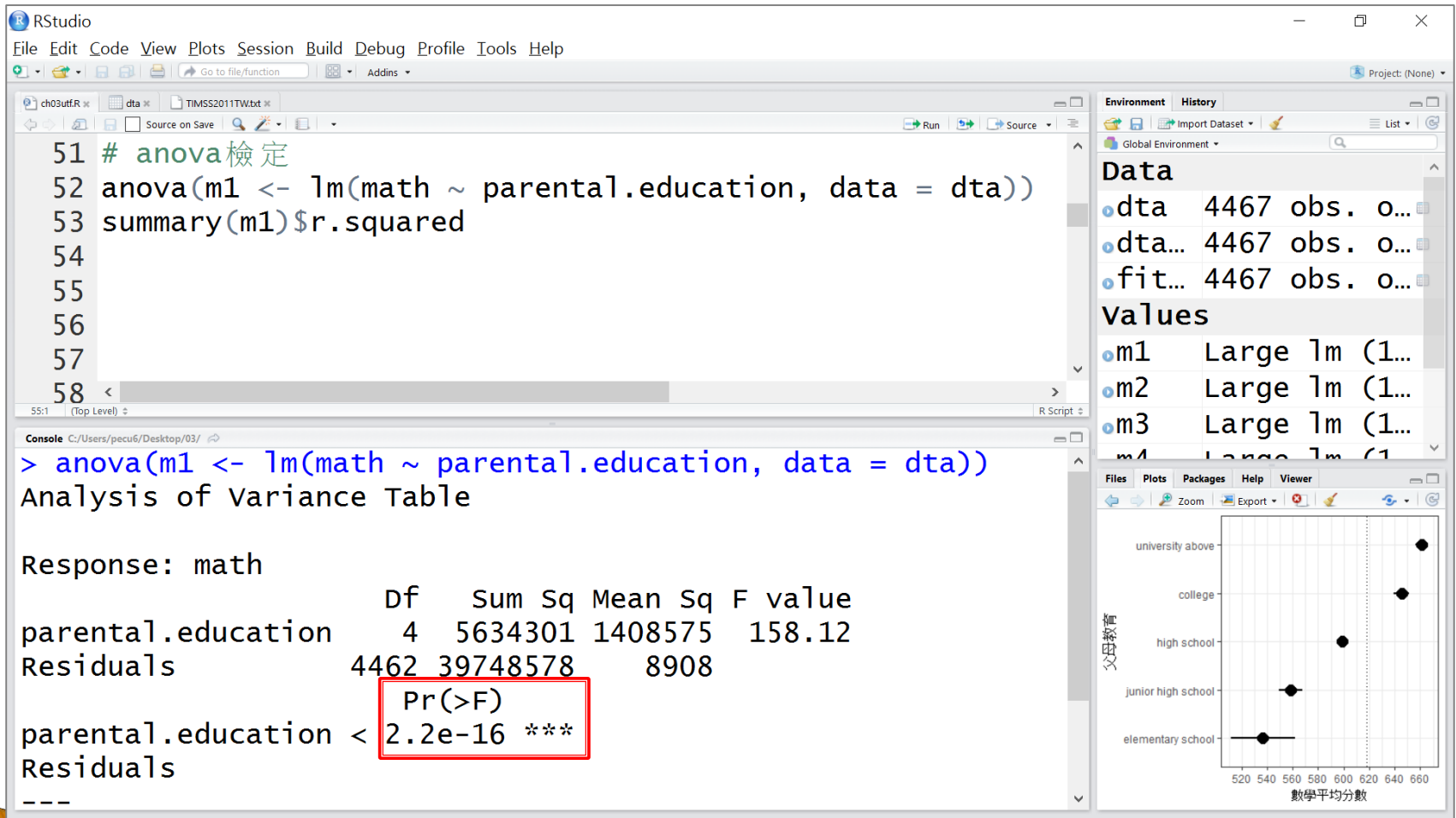
tapply()

- ▶ `tapply()` 允許根據某些變數的值，把原始資料分割為若干組。
- ▶ 對每一組資料應用特定的操作。
- ▶ `tapply(dta$math, dta$parental.education, mean)`
- ▶ `tapply(dta$math, dta$parental.education, summary)`
- ▶ 表示將 `dta$math` 的資料按照 `dta$parental.education` 的值進行分組，並將分組後資料進行 `mean or summary`。

ANOVA 分析

- ▶ 變異數分析 (**Analysis of variance** , 簡稱**ANOVA**) 為資料分析中常見的統計模型。
- ▶ 主要為探討連續型 (**Continuous**) 資料型態之依變數 (果, y) 與類別型資料型態之自變數 (因, x) 的關係。

ANOVA 分析

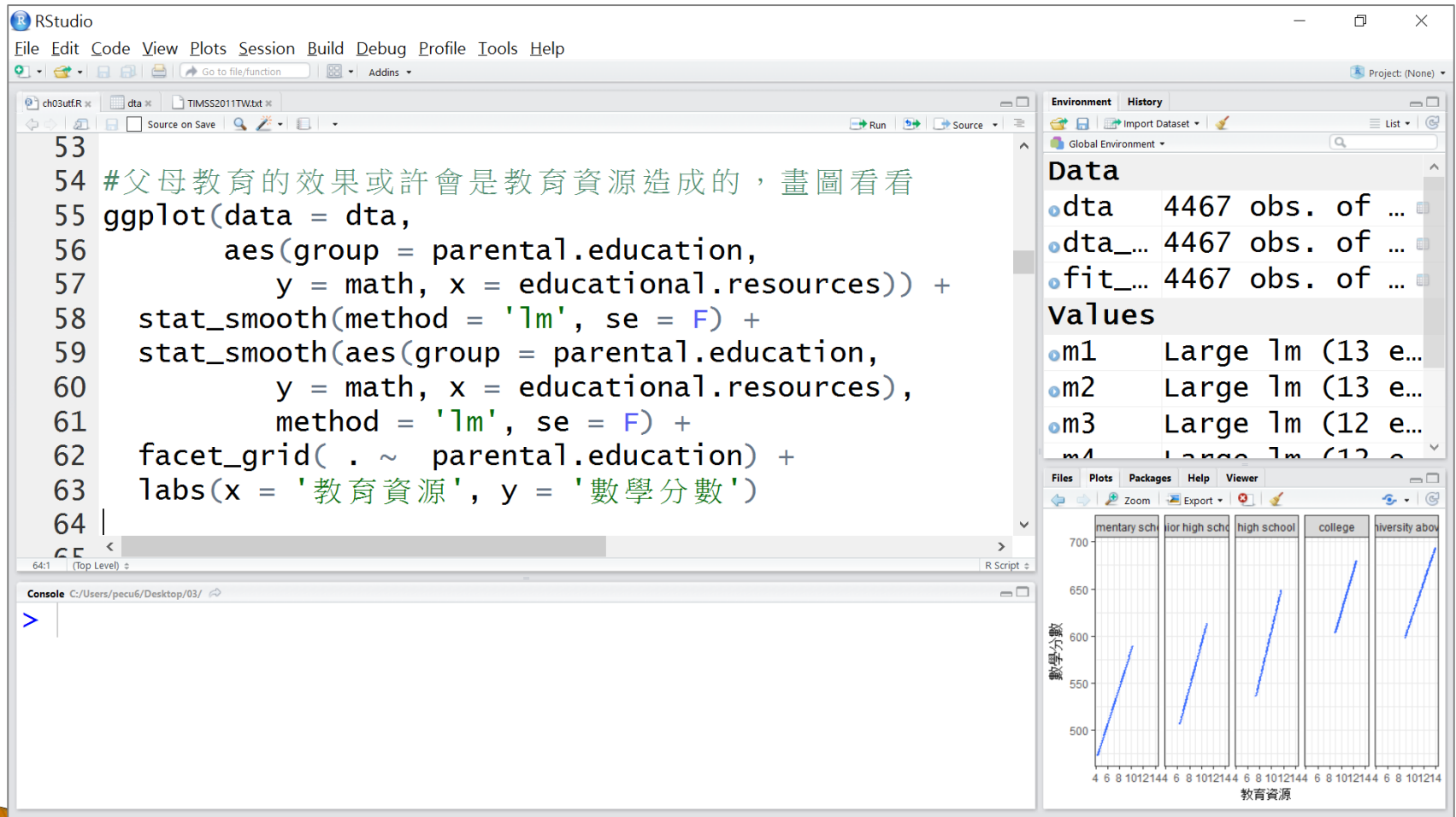


迴歸分析

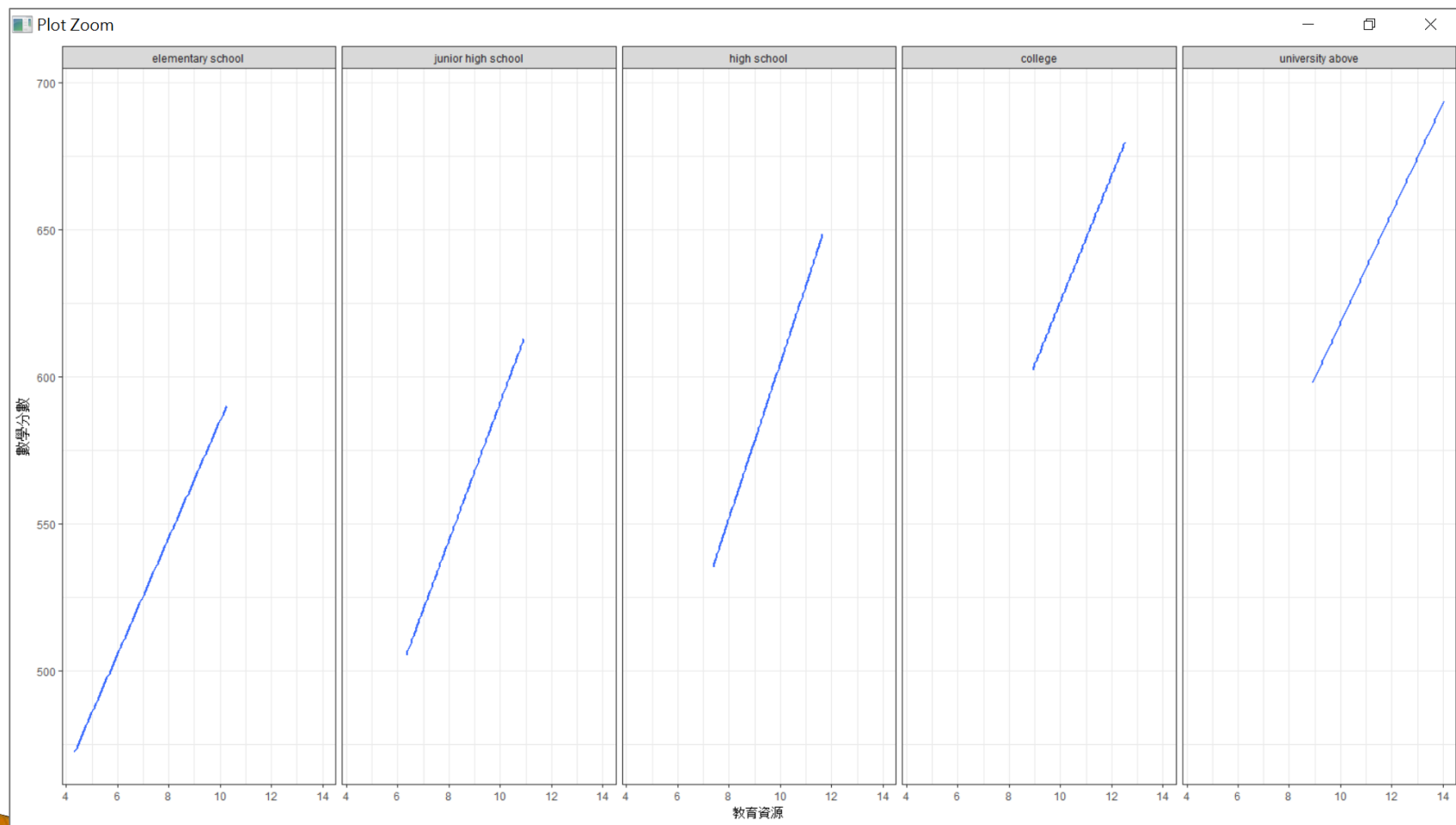
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \text{softmax} \left(\begin{bmatrix} W_{1,1}x_1 + W_{1,2}x_1 + W_{1,3}x_1 + b_1 \\ W_{2,1}x_2 + W_{2,2}x_2 + W_{2,3}x_2 + b_2 \\ W_{3,1}x_3 + W_{3,2}x_3 + W_{3,3}x_3 + b_3 \end{bmatrix} \right)$$

YX^{-1} W XX^{-1}

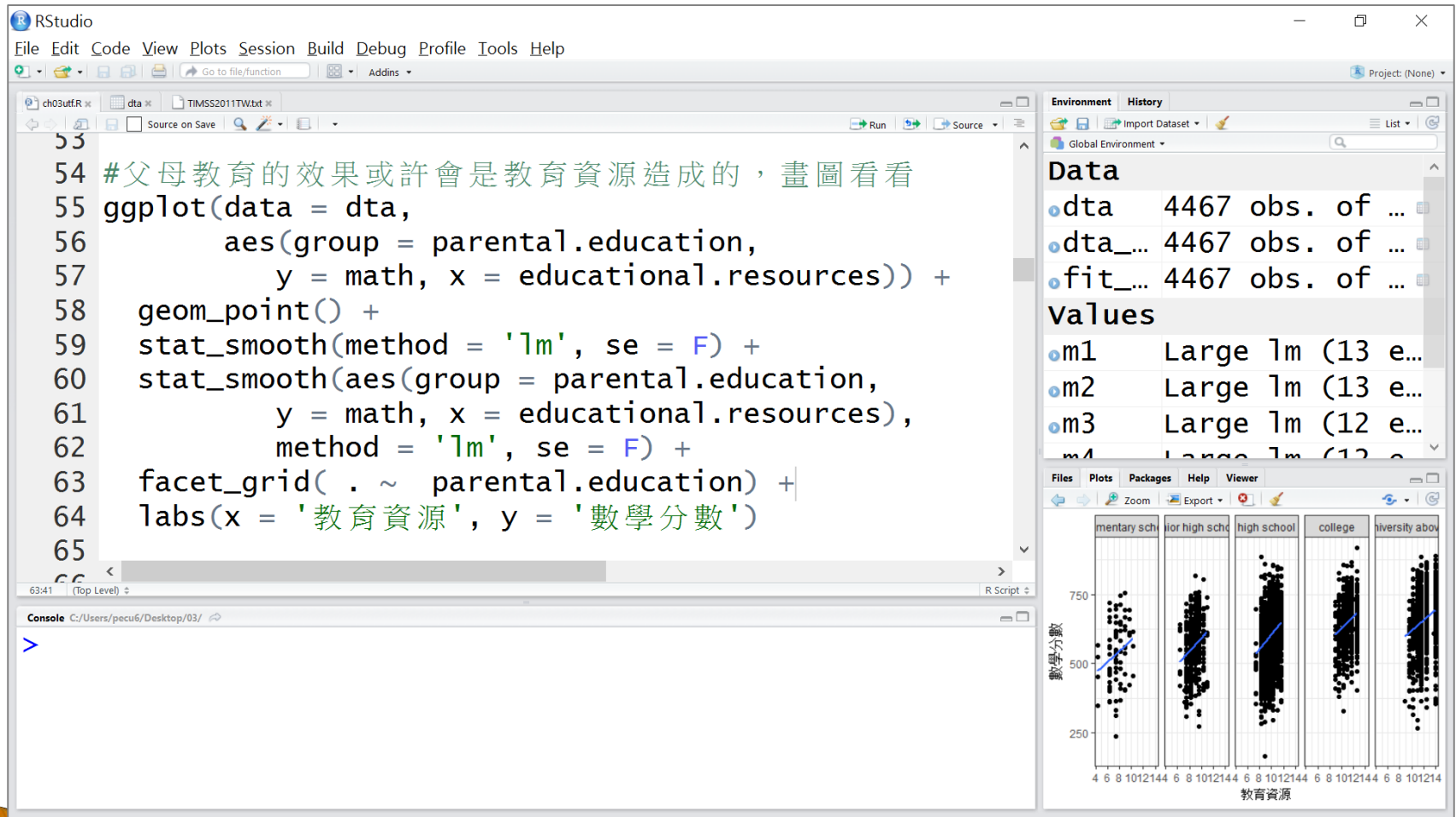
迴歸分析



父母教育程度與教育資源的影響



迴歸分析



父母教育程度與教育資源的影響

