

R 語言問卷分析(4)

複選題及卡方檢定

平均數差異檢定

變異數分析、事後檢定

迴歸分析

徑路分析

吳漢銘

國立政治大學 統計學系



<https://hmwu.idv.tw>

■ 複選題

- 題項編號: a1m1, a1m2, a1m3, a1m4
 - a1: 第1題。m1: 選項1。
- 選項有被勾選者，編碼為1; 未勾選者，編碼為0。

■ 排序題

- 題項編號: a2m1, a2m2, a2m3, a2m4, a2m5
- 數值介於1至5。

子女學習意見調查問卷

【基本資料】

1. 您是學童的： 父親 母親

2. 您的年齡： 35歲以下 36歲~44歲 45歲以上

【題項】

一、您未來選擇孩子就讀的國中時，會考量哪些因素？（可複選）

1. 學校辦學的口碑

2. 校長的領導風格

3. 學校升學率高低

4. 住家交通的因素

二、對於子女國小高年級的學習科目，您重視的重要性次序為何？
（1 最重視、2 次重視、……）

國語 數學 英文 自然 社會

三、對於子女國小的學習，您最重視項目是哪一項？

考試成績 生活常規 同儕關係 品德行為

四、您對於目前子女就讀學校的整體滿意度如何？

非常滿意 滿意 不滿意 非常不滿意

五、您對於目前子女就讀班級的整體滿意度如何？

非常滿意 滿意 不滿意 非常不滿意

問題(1): 求全體樣本在複選題， 各選項勾選的次數及百分比為何？

- 了解全體受試者選填的情形。
- 不同背景變項的樣本選填情形。(即複選題的次數分配表)

```

> study <- read.csv("data/子女學習問卷_1.csv")
> head(study)
  編號 關係 年齡 a1m1 a1m2 a1m3 a1m4 a2m1 a2m2 a2m3 a2m4 a2m5 a3 a4 a5
1     1    1    1     1    0    1    0    1    1    2    3    4    5    1    2    2
2     3    1    1     1    1    1    0    0    3    4    1    2    5    4    2    1
3     5    1    1     1    1    0    1    1    2    1    3    4    5    3    1    1
4     7    1    1     1    1    0    1    0    1    2    3    4    5    1    2    1
5    29    1    1     1    1    1    1    0    2    1    3    5    4    3    4    2
6    30    1    1     1    1    0    1    0    2    3    1    4    5    3    4    2
> dim(study)
[1] 120  15
> # 關係: 1: 父親, 2: 母親
> # 年齡: 1: 35歲以下, 2: 36-44歲, 3: 45歲以上
>
>
> a1_variable_names <- c("學校辦學的口碑", "校長的領導風格",
+ "學校升學率高低", "住家交通的因素")
> a1_count <- colSums(study[, 4:7])
> a1_percentage <- a1_count/sum(a1_count)
> a1_obs_percentage <- a1_count/nrow(study)

```

複選題各選項勾選的次數及百分比

```
> a1_freq <- data.frame(就讀國中考量因素 = a1_variable_names,  
+                       個數 = a1_count,  
+                       百分比 = paste0(round(a1_percentage*100, 1), "%"),  
+                       觀察值百分比 = paste0(round(a1_obs_percentage*100, 1), "%"))  
> a1_freq  
  就讀國中考量因素  個數  百分比  觀察值百分比  
a1m1  學校辦學的口碑  81  29.3%  67.5%  
a1m2  校長的領導風格  66  23.9%  55%  
a1m3  學校升學率高低  78  28.3%  65%  
a1m4  住家交通的因素  51  18.5%  42.5%
```

```
> a1_freq_all <- rbind(a1_freq,  
+                      c("總數", sum(a1_count),  
+                        paste0(round(sum(a1_percentage)*100, 1), "%"),  
+                        paste0(round(sum(a1_obs_percentage)*100, 1), "%")))  
> a1_freq_all  
  就讀國中考量因素  個數  百分比  觀察值百分比  
a1m1  學校辦學的口碑  81  29.3%  67.5%  
a1m2  校長的領導風格  66  23.9%  55%  
a1m3  學校升學率高低  78  28.3%  65%  
a1m4  住家交通的因素  51  18.5%  42.5%  
5      總數  276  100%  230%
```



問題(2-1): 不同親子關係的樣本在 複選題各選項勾選的次數及百分比為何?

```

> n <- nrow(study)
> study$關係父母 <- c("父親", "母親")[study$關係]
> table(study$關係父母)
父親 母親
  63   57
>
> # 個數
> a1_cross_parents_freq <- aggregate(study[, 4:7], list(study$關係父母), sum)
> colnames(a1_cross_parents_freq) <- c("關係", a1_variable_names)
> a1_cross_parents_freq
關係 學校辦學的口 校長的領導風格 學校升學率高低 住家交通的因素
1 父親           51           33           36           33
2 母親           30           33           42           18
>
> # 總數的百分比
> a1_cross_parents_percentage <- aggregate(study[, 4:7], list(study$關係父母),
+                                         function(x) sum(x)/n)
> colnames(a1_cross_parents_percentage) <- c("關係", a1_variable_names)
> a1_cross_parents_percentage
關係 學校辦學的口 校長的領導風格 學校升學率高低 住家交通的因素
1 父親           0.425           0.275           0.30           0.275
2 母親           0.25            0.275           0.35           0.150

```

複選題分析交叉表

關係*%a1 交叉表列

		就讀國中考量因素 ^a				總數	
		學校辦學的口 碑	校長的領導風 格	學校升學率高 低	住家交通的因 素		
關係	父親	個數	51	33	36	33	63
		關係中的 %	81.0%	52.4%	57.1%	52.4%	
		%a1 中的 %	63.0%	50.0%	46.2%	64.7%	
		總數的 %	42.5%	27.5%	30.0%	27.5%	
母親	母親	個數	30	33	42	18	57
		關係中的 %	52.6%	57.9%	73.7%	31.6%	
		%a1 中的 %	37.0%	50.0%	53.8%	35.3%	
		總數的 %	25.0%	27.5%	35.0%	15.0%	
總數		個數	81	66	78	51	120
		總數的 %	67.5%	55.0%	65.0%	42.5%	

百分比及總數是根據應答者而來的。

a. 二分法群組表列於值 1。



問題(2-1): 不同親子關係的樣本在 複選題各選項勾選的次數及百分比為何?

```

> # 關係中的百分比
> tmp <- a1_cross_parents_freq[, 2:5]/table(study$關係父母)
> a1_cross_parents_percentage_within <- data.frame(關係 = c("父親", "母親"),
+                                               round(tmp, 3))
> a1_cross_parents_percentage_within
關係 學校辦學的口碑 校長的領導風格 學校升學率高低 住家交通的因素
1 父親          0.810          0.524          0.571          0.524
2 母親          0.526          0.579          0.737          0.316

> # a1中的百分比
> tmp <- t(apply(a1_cross_parents_freq[, 2:5], 1, function(x) x/a1_count))
> a1_cross_parents_percentage_a1 <- data.frame(關係 = c("父親", "母親"),
+                                               round(tmp, 3))
> a1_cross_parents_percentage_a1
關係 學校辦學的口碑 校長的領導風格 學校升學率高低 住家交通的因素
1 父親          0.63          0.5          0.462          0.647
2 母親          0.37          0.5          0.538          0.353
    
```

關係*%a1 交叉表列

			就讀國中考量因素 ^a				總數
			學校辦學的口 碑	校長的領導風 格	學校升學率高 低	住家交通的因 素	
關係	父親	個數	51	33	36	33	63
		關係中的 %	81.0%	52.4%	57.1%	52.4%	
	%a1 中的 %	63.0%	50.0%	46.2%	64.7%		
	總數的 %	42.5%	27.5%	30.0%	27.5%		
母親	母親	個數	30	33	42	18	57
		關係中的 %	52.6%	57.9%	73.7%	31.6%	
	%a1 中的 %	37.0%	50.0%	53.8%	35.3%		
	總數的 %	25.0%	27.5%	35.0%	15.0%		
總數		個數	81	66	78	51	120
		總數的 %	67.5%	55.0%	65.0%	42.5%	100.0%

百分比及總數是根據應答者而來的。

a. 二分法群組表列於值 1。



問題(2-2): 不同年齡父母的樣本在 複選題各選項勾選的次數及百分比為何?

7/55

```
> age <- c("35歲以下", "36-44歲", "45歲以上")
> study$父母年齡 <- age[study$年齡]
> table(study$父母年齡)
35歲以下 36-44歲 45歲以上
    45      36      39
```

```
> table(study$父母年齡, study$a1m1)
           0  1
35歲以下 15 30
36-44歲  21 15
45歲以上  3 36
```

```
> # 個數
> a1_cross_age_freq <- sapply(study[, 4:7], function(x) table(study$父母年齡, x)[, 2])
> colnames(a1_cross_age_freq) <- a1_variable_names
> a1_cross_age_freq
      學校辦學的口碑  校長的領導風格  學校升學率高低  住家交通的因素
35歲以下           30           27           30           18
36-44歲            15           24           18           15
45歲以上           36           15           30           18
>
> # 總數百分比
> round(a1_cross_age_freq/n, 3)
      學校辦學的口碑  校長的領導風格  學校升學率高低  住家交通的因素
35歲以下           0.250           0.225           0.25           0.150
36-44歲            0.125           0.200           0.15           0.125
45歲以上           0.300           0.125           0.25           0.150
> # 年齡中的百分比
> # wrong: round(prop.table(a1_cross_age_freq, 1), 3)
> round(a1_cross_age_freq/c(table(study$父母年齡)), 3)
      學校辦學的口碑  校長的領導風格  學校升學率高低  住家交通的因素
35歲以下           0.667           0.600           0.667           0.400
36-44歲            0.417           0.667           0.500           0.417
45歲以上           0.923           0.385           0.769           0.462
> # a1中的百分比
> round(prop.table(a1_cross_age_freq, 2), 3)
      學校辦學的口碑  校長的領導風格  學校升學率高低  住家交通的因素
35歲以下           0.370           0.409           0.385           0.353
36-44歲            0.185           0.364           0.231           0.294
45歲以上           0.444           0.227           0.385           0.353
```



問題(3): 在不同親子關係及不同年齡父母的樣本中， 8/55 複選題各選項勾選的次數及百分比為何？

課堂練習!

年齡*關係 交叉表列

關係	年齡			就讀國中考量因素 ^a				總數
				學校辦學的口 碑	校長的領導風 格	學校升學率高 低	住家交通的因 素	
父親	35歲以下	個數	21	15	15	9	24	
		年齡 中的 %	87.5%	62.5%	62.5%	37.5%		
		\$a1 中的 %	41.2%	45.5%	41.7%	27.3%		
		總數的 %	33.3%	23.8%	23.8%	14.3%	38.1%	
	36-44歲	個數	9	12	9	9	18	
		年齡 中的 %	50.0%	66.7%	50.0%	50.0%		
		\$a1 中的 %	17.6%	36.4%	25.0%	27.3%		
		總數的 %	14.3%	19.0%	14.3%	14.3%	28.6%	
	45歲以上	個數	21	6	12	15	21	
		年齡 中的 %	100.0%	28.6%	57.1%	71.4%		
		\$a1 中的 %	41.2%	18.2%	33.3%	45.5%		
		總數的 %	33.3%	9.5%	19.0%	23.8%	33.3%	
總數		個數	51	33	36	33	63	
		總數的 %	81.0%	52.4%	57.1%	52.4%	100.0%	
母親	35歲以下	個數	9	12	15	9	21	
		年齡 中的 %	42.9%	57.1%	71.4%	42.9%		
		\$a1 中的 %	30.0%	36.4%	35.7%	50.0%		
		總數的 %	15.8%	21.1%	26.3%	15.8%	36.8%	
	36-44歲	個數	6	12	9	6	18	
		年齡 中的 %	33.3%	66.7%	50.0%	33.3%		
		\$a1 中的 %	20.0%	36.4%	21.4%	33.3%		
		總數的 %	10.5%	21.1%	15.8%	10.5%	31.6%	
	45歲以上	個數	15	9	18	3	18	
		年齡 中的 %	83.3%	50.0%	100.0%	16.7%		
		\$a1 中的 %	50.0%	27.3%	42.9%	16.7%		
		總數的 %	26.3%	15.8%	31.6%	5.3%	31.6%	
總數		個數	30	33	42	18	57	
		總數的 %	52.6%	57.9%	73.7%	31.6%	100.0%	

百分比及總數是根據應答者而來的。

a. 二分法群組表列於值 1。

描述性統計量

```

> subjects <- c("國語","數學", "英語", "自然", "社會")
> a2 <- paste0("a2m", 1:5)
> a2_summary <- summary(study[, a2])
> colnames(a2_summary) <- subjects
> a2_summary
  國語          數學          英語          自然          社會
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:4.000
Median :2.000   Median :2.000   Median :3.000   Median :4.000   Median :5.000
Mean   :2.383   Mean   :2.167   Mean   :2.783   Mean   :3.333   Mean   :4.333
3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:5.000
Max.   :5.000   Max.   :4.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
>
> library(psych)
> a2_describe <- describe(study[, a2])
> rownames(a2_describe) <- subjects
> a2_describe
  vars   n mean  sd median trimmed  mad min max range  skew kurtosis  se
國語   1 120 2.38 1.32     2   2.23 1.48   1  5   4  0.73   -0.63 0.12
數學   2 120 2.17 1.00     2   2.08 1.48   1  4   3  0.47   -0.85 0.09
英語   3 120 2.78 1.16     3   2.76 1.48   1  5   4 -0.12   -0.61 0.11
自然   4 120 3.33 1.19     4   3.42 1.48   1  5   4 -0.57   -0.74 0.11
社會   5 120 4.33 1.24     5   4.65 0.00   1  5   4 -1.85    2.05 0.11
> describeBy(study[, a2], study$關係父母)
> describeBy(study[, a2], study$父母年齡)

```

問題(5): 求不同年齡父母對排序題的等級平均數?

- 想了解不同年齡父母對於學習科目重要性看法。

子女學習問卷_1.sav [DataSet1] - SPSS Statistics Data Editor

檔案(E) 編輯(E) 檢視(V) 資料(D) 轉換(I) 分析(A) 統計圖(G) 公用程式(U) 增益集(O) 視窗(W) 說明(H)

顯示: 15 個變數 (共有 15 個)

	編號	關係	a1m2	a1m3	a1m4	a2m1	a2m2	a2m3	a2m4	a2m5	a3	a4	a5	var	v
1	001	父親	1	0	1	1	2	3	4	5	考試...	不滿...	不滿...		
2	003	父親	1	0	0	3	4	1	2	5	品德...	不滿...	非常...		
3	005	父親	0	1	1	2									
4	007	父親	0	1	0	1									
5	029	父親	1	1	0	2									
6	030	父親	0	1	0	2									
7	031	父親	1	1	1	2									
8	034	父親	1	0	0	1									
9	041	父親	1	0	1	1									
10	043	父親	1	0	0	3									
11	045	父親	0	1	1	2									
12	047	父親	0	1	0	1									
13	069	父親	1	1	0	2									
14	070	父親	0	1	0	2									
15	071	父親	1	1	1	2									
16	074	父親	1	0	0	1									
17	081	父親	35歲以下	0	1	0	1								

分割檔案

分析所有觀察值，勿建立群組(A)
 比較群組(C)
 依群組組織輸出(O)
 依此群組(G)：
 年齡
 依分組變數排序檔案(S)
 檔案已排序(E)

目前狀態：「依組別分析」已關閉。

描述性統計量

SPSS Statistics Data Editor window showing the 'Analyze' menu path: Analyze > Descriptive Statistics > Descriptive Statistics (D)...

編號	關係	年齡
1	父親	35歲以下
2	父親	35歲以下
3	父親	35歲以下
4	父親	35歲以下
5	父親	35歲以下
6	父親	35歲以下
7	父親	35歲以下
8	父親	35歲以下
9	父親	35歲以下
10	父親	35歲以下
11	父親	35歲以下
12	父親	35歲以下
13	父親	35歲以下
14	父親	35歲以下
15	父親	35歲以下
16	父親	35歲以下
17	父親	35歲以下

描述性統計量 (Descriptive Statistics) dialog box. The '變數(V):' (Variables) list contains: 關係, 年齡, 學校辦學的口碑 [a1m1], 校長的領導風格 [a1m2], 學校升學率高低 [a1m3], 住家交通的因素 [a1m4], 重視項目 [a3]. The '顯示: 15 個變數 (共有 15 個)' (Display: 15 variables (out of 15)) is shown at the top right.

年齡 = 35歲以下

敘述統計^a

	個數	範圍	最小值	最大值	平均數	標準差
國語	45	4	1	5	2.00	.977
數學	45	3	1	4	1.91	1.041
英語	45	3	1	4	2.69	.900
自然	45	3	2	5	3.73	.780
社會	45	4	1	5	4.67	1.022
有效的 N (完全排除)	45					

a. 年齡 = 35歲以下

年齡 = 45歲以上

敘述統計^a

	個數	範圍	最小值	最大值	平均數	標準差
國語	39	3	1	4	2.15	1.159
數學	39	3	1	4	2.03	.778
英語	39	3	1	4	2.49	1.073
自然	39	4	1	5	3.56	1.209
社會	39	1	4	5	4.77	.427
有效的 N (完全排除)	39					

a. 年齡 = 45歲以上

年齡 = 36-44歲

敘述統計^a

	個數	範圍	最小值	最大值	平均數	標準差
國語	36	4	1	5	3.11	1.582
數學	36	3	1	4	2.64	1.018
英語	36	4	1	5	3.22	1.416
自然	36	4	1	5	2.58	1.273
社會	36	4	1	5	3.44	1.594
有效的 N (完全排除)	36					

a. 年齡 = 36-44歲

分割檔案

- 分析所有觀察值，勿建立群組(Δ)
- 比較群組(C)
- 依群組組織輸出(O)

依此群組(S):

- 依分組變數排序檔案(S)
- 檔案已排序(E)

目前狀態：安排輸出依據：年齡

確定 貼上之後(E) 重設(R) 取消 輔助說明



問題(6): 全體樣本對「學習重視項目」「學校整體滿意度」「班級滿意度」各選項勾選的次數及百分比為何?

```
> item <- c("考試成績", "生活常規", "同儕關係", "品德行為")
> satisfactory <- c("非常不滿意", "不滿意", "滿意", "非常滿意")
>
> study$重視項目 <- item[study$a3]
> study$學校滿意 <- factor(satisfactory[study$a4], levels = satisfactory, ordered = T)
> study$班級滿意 <- factor(satisfactory[study$a5], levels = satisfactory, ordered = T)
>
> table(study$重視項目)
生活常規  同儕關係  考試成績  品德行為
      35      29      17      39
> round(prop.table(table(study$重視項目)), 3)
生活常規  同儕關係  考試成績  品德行為
  0.292    0.242    0.142    0.325
> table(study$學校滿意)
非常不滿意  不滿意  滿意  非常滿意
      28      35      26      31
> round(prop.table(table(study$學校滿意)), 3)
非常不滿意  不滿意  滿意  非常滿意
  0.233    0.292    0.217    0.258
> table(study$班級滿意)
非常不滿意  不滿意  滿意  非常滿意
      32      30      34      24
> round(prop.table(table(study$班級滿意)), 3)
非常不滿意  不滿意  滿意  非常滿意
  0.267    0.250    0.283    0.200
```

問題(7): 全體樣本對「學習重視項目」、「學校整體滿意度」、「班級滿意度」各選項勾選的次數是否有顯著不同? 14/55

- 卡方適合度檢定 (goodness of fit test)
 - 檢定某一變項的「實際觀察次數分配」與「期望理論次數分配」是否符合?
 - H_0 : 實際觀察次數與期望理論次數之間無顯著差異。

```
> table(study$重視項目)
生活常規  同儕關係  考試成績  品德行為
      35      29      17      39
> table(study$學校滿意)
非常不滿意  不滿意  滿意  非常滿意
      28      35      26      31
> table(study$班級滿意)
非常不滿意  不滿意  滿意  非常滿意
      32      30      34      24
```

```
> chisq.test(table(study$重視項目))

      Chi-squared test for given probabilities

data:  table(study$重視項目)
X-squared = 9.2, df = 3, p-value = 0.02675

> chisq.test(table(study$學校滿意))

      Chi-squared test for given probabilities

data:  table(study$學校滿意)
X-squared = 1.5333, df = 3, p-value = 0.6746

> chisq.test(table(study$班級滿意))

      Chi-squared test for given probabilities

data:  table(study$班級滿意)
X-squared = 1.8667, df = 3, p-value = 0.6005
```

- 卡方檢定: 百分比同質性檢定
 - 自變項: 不同「年齡」為3個類別變項。
 - 反應變項(依變項): 「學校整體滿意度」為4個選項的次數百分比。屬類別反應變項。
 - 探討不同自變項在依變項反應的差需採用「百分比同質性檢定」。

```

> tb <- table(study$學校滿意, study$父母年齡)
> tb

           35歲以下  36-44歲  45歲以上
非常不滿意         14         10          4
不滿意             13         11         11
滿意               6          7         13
非常滿意          12          8         11

> chisq.test(tb)

        Pearson's Chi-squared test

data:  tb
X-squared = 8.6201, df = 6, p-value = 0.1961
    
```

p-value = 0.1961，接受虛無假設。表示不同年齡的學童家長，在學校滿意度4個反應變項沒有一個選項選擇的百分比間有顯著差異。

課堂練習:

- 不同「年齡」父母對「班級滿意度」4個選項反應百分比是否有顯著不同?
- 不同「親子關係」對「學校整體滿意度」4個選項反應百分比是否有顯著不同?
- 不同「親子關係」對「班級滿意度」4個選項反應百分比是否有顯著不同?

- 積差相關 (Product-moment correlation): 即皮爾森相關係數。

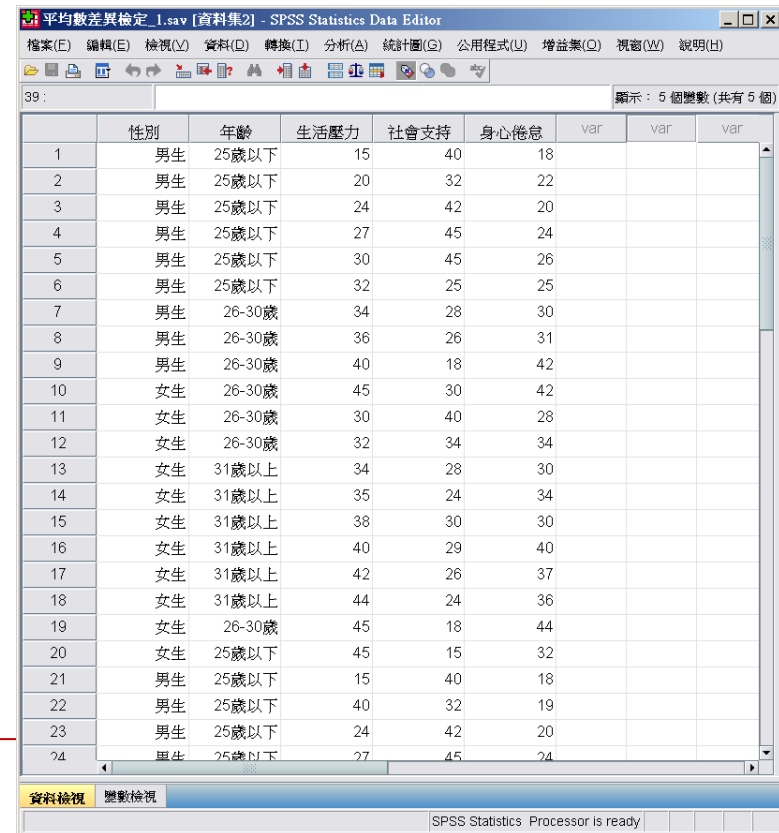
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- 「積差相關係數」的平方為「決定係數」 coefficient of determination。
- 簡單線性迴歸中: 依變數的總變量中，可以被自變項解釋的變異量百分比。

相關係數絕對值	關聯程度	決定係數
$r < .40$	低度相關	$< .16$
$.40 \leq r \leq .70$	中度相關	$.16 \leq r^2 \leq .49$
$r > .70$	高度相關	$> .49$

- 運動員的「生活壓力」、「社會支持」及「身心倦怠」關係。
- 「生活壓力量表」、「社會支持量表」、「身心倦怠量表」
- 各量表得分愈高，表示「生活壓力愈大」、「社會支持愈高」、「身心倦怠愈大」
- 問題：
 - 運動員的「生活壓力」、「社會支持」與「身心倦怠」是否有顯著的相關？

```
> example_data <- read.csv("data/平均數差異檢定_1.csv")
> head(example_data)
  性別 年齡 生活壓力 社會支持 身心倦怠
1    1    1    15      40      18
2    1    1    20      32      22
3    1    1    24      42      20
4    1    1    27      45      24
5    1    1    30      45      26
6    1    1    32      25      25
> dim(example_data)
[1] 40  5
```



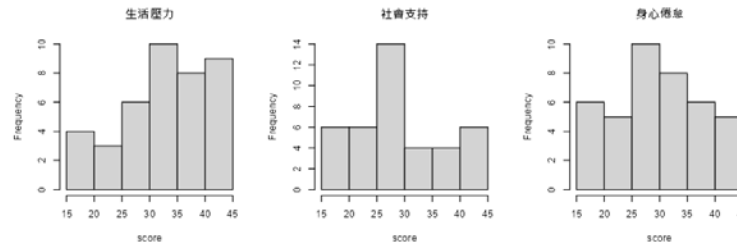
平均數差異檢定_1.sav [資料集2] - SPSS Statistics Data Editor

	性別	年齡	生活壓力	社會支持	身心倦怠	var	var	var
1	男生	25歲以下	15	40	18			
2	男生	25歲以下	20	32	22			
3	男生	25歲以下	24	42	20			
4	男生	25歲以下	27	45	24			
5	男生	25歲以下	30	45	26			
6	男生	25歲以下	32	25	25			
7	男生	26-30歲	34	28	30			
8	男生	26-30歲	36	26	31			
9	男生	26-30歲	40	18	42			
10	女生	26-30歲	45	30	42			
11	女生	26-30歲	30	40	28			
12	女生	26-30歲	32	34	34			
13	女生	31歲以上	34	28	30			
14	女生	31歲以上	35	24	34			
15	女生	31歲以上	38	30	30			
16	女生	31歲以上	40	29	40			
17	女生	31歲以上	42	26	37			
18	女生	31歲以上	44	24	36			
19	女生	26-30歲	45	18	44			
20	女生	25歲以下	45	15	32			
21	男生	25歲以下	15	40	18			
22	男生	25歲以下	40	32	19			
23	男生	25歲以下	24	42	20			
24	男生	25歲以下	27	45	24			

```

> example_data$性別2 <- as.factor(c("男生", "女生")[example_data$性別])
> example_data$年齡2 <- as.factor(c("25歲以下", "26-30歲", "31歲以上")[example_data$年齡])
>
> item <- c("生活壓力", "社會支持", "身心倦怠")
> par(mfrow =c (1, 2))
> sapply(6:7, function(x) barplot(table(example_data[, x]),
+                                 main = colnames(example_data)[x]))
>
> par(mfrow =c (1, 3))
> sapply(3:5, function(x) hist(example_data[, x],
+                               main = colnames(example_data)[x], xlab = "score"))
>
> cor(example_data[, item])
      生活壓力  社會支持  身心倦怠
生活壓力  1.0000000 -0.6703826  0.8124818
社會支持 -0.6703826  1.0000000 -0.6185917
身心倦怠  0.8124818 -0.6185917  1.0000000
> my_cor_test_data <- stack(example_data[, item])
> colnames(my_cor_test_data) <- c("總分", "分量表")
> pairwise.cor.test(my_cor_test_data$總分, my_cor_test_data$分量表)
Pairwise comparisons using Pearson's product-moment correlation

```



data: my_cor_test_data\$總分 and my_cor_test_data\$分量表

```

      生活壓力  社會支持
社會支持 4.4e-06 -
身心倦怠 5.8e-10 2.1e-05

P value adjustment method: holm

```

- 運動員感受的「社會支持」程度愈高(低)，則其知覺的「生活壓力」愈小(大)。(r=-0.670)
- 判定係數=0.449 (-0.670²)，表示「社會支持」變項可以解釋「生活壓力」變項總變異的44.9%。或表示「生活壓力」變項可以解釋「社會支持」變項總變異的44.9%。
- 若「社會支持」有三層面: 老師支持，家人支持，同儕支持。若「身心倦怠」有三層面: 生理症狀，心理情緒，行為表現。則需同時求出各層面間的相關。

■ 平均數的比較

- 獨立樣本t檢定(two-sample t-test, unpaired): 二個母體平均數是否有差異?
- 相依樣本t檢定(two-sample t-test, paired): 一個母體的兩次平均數是否有差異?
- 單因子變異數分析(ANOVA): 三個(以上)母體平均數是否有差異?
- 單因子多變量變異數分析(MANOVA): 三個(以上)母體多變量平均數是否有差異?

■ 事後比較 (post-hoc test) (multiple comparison)

- Tukey最實在顯著差法: HSD法, honestly significant difference
- Newman-Keul' s method: N-K法
- Scheffe' s method: S法
- 最小顯著差異法: Least significant difference: LSD法。

平均數檢定 in R

Hypothesis Testing	One Sample	Two Samples		> two Groups
	-	Paired data	Unpaired data	Complex data
Parametric (variance equal)	t-test <code>t.test(x, mu = 0)</code>	t-test <code>t.test(x-y, var.equal = TRUE)</code> <code>t.test(x, y, paired = TRUE, var.equal = TRUE)</code>	t-test <code>t.test(x, y, var.equal = TRUE)</code>	One-Way Analysis of Variance (ANOVA) <code>aov(x~g, data)</code> <code>oneway.test(x~g, data, var.equal = TRUE)</code>
Parametric (variance not equal)		Welch t-test <code>t.test(x-y)</code> <code>t.test(x, y, paired = TRUE)</code>	Welch t-test <code>t.test(x, y)</code>	Welch ANOVA <code>oneway.test(x~g, data)</code>
Non-Parametric (無母數檢定)	Wilcoxon Signed-Rank Test <code>wilcox.test(x, mu = 0)</code>	Wilcoxon Signed-Rank Test <code>wilcox.test(x-y)</code> <code>wilcox.test(x, y, paired = TRUE)</code>	Wilcoxon Rank-Sum Test (Mann-Whitney U Test) <code>wilcox.test(x, y)</code>	Kruskal-Wallis Test <code>kruskal.test(x, g)</code>

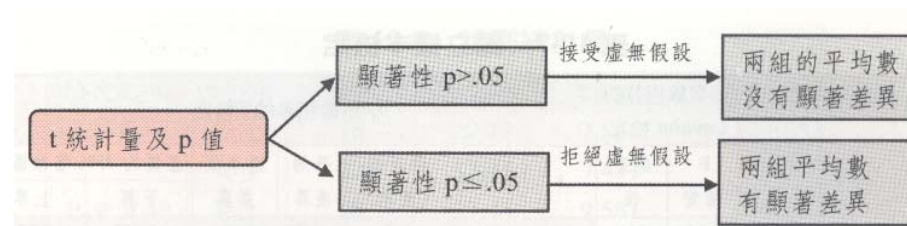
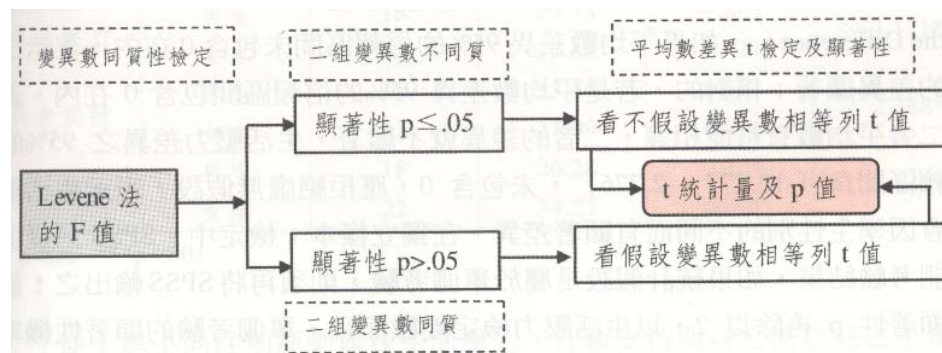
`pairwise.t.test {stats}`: Calculate pairwise comparisons between group levels with corrections for multiple testing

`TukeyHSD {stats}`: Compute Tukey Honest Significant Differences



獨立樣本t檢定 (t-test)

- 問題:
 - 不同性別(男、女)的運動員在「生活壓力」是否有顯著的差異?
 - 不同性別(男、女)的運動員在「社會支持」是否有顯著的差異?
 - 不同性別(男、女)的運動員在「身心倦怠」是否有顯著的差異?
 - 高生活壓力組學生，低生活壓力組學生的學業成就是否有顯著不同?
 - 男女生的工作壓力是否有顯著不同?
- 自變項(分組變數): 二分類別變項: 性別(男、女)、分組(高分、低分)
- 依變項(檢定變數): 連續變項: 生活壓力、社會支持、分數。



獨立樣本t檢定 (t-test)

```

> library(car) # Companion to Applied Regression
> alpha <- 0.05
> attach(example_data)
> my_t_test <- function(x){
+   md <- eval(parse(text = paste0(item[x], " ~ 性別2")))
+   is.equal.variance <- leveneTest(md)$"Pr(>F)"[1] > alpha
+   t.test(md, var.equal = is.equal.variance)
+ }
> lapply(1:3, my_t_test)
[[1]]
      Two Sample t-test
data: 生活壓力 by 性別2
t = 3.1712, df = 38, p-value = 0.002999
alternative hypothesis: true difference in means between group 女生 and group 男生 is not equal to 0
95 percent confidence interval:
 2.776191 12.577344
sample estimates:
mean in group 女生 mean in group 男生
   37.45455         29.77778

[[2]]
      Welch Two Sample t-test
data: 社會支持 by 性別2
t = -2.368, df = 29.79, p-value = 0.02457
alternative hypothesis: true difference in means between group
95 percent confidence interval:
-11.8346388 -0.8724319
sample estimates:
mean in group 女生 mean in group 男生
   27.09091         33.44444

[[3]]
      Two Sample t-test
data: 身心倦怠 by 性別2
t = 3.848, df = 38, p-value = 0.0004421
alternative hypothesis: true difference in means between group
95 percent confidence interval:
 3.788891 12.201008
sample estimates:
mean in group 女生 mean in group 男生
   34.27273         26.27778

```

Statistical tests for comparing variances

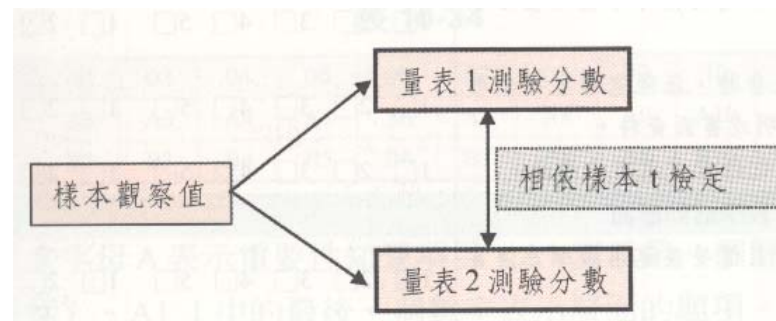
- **F-test**: Compare the variances of two samples. The data must be normally distributed.
- **Bartlett's test**: Compare the variances of k samples, where k can be more than two samples. The data must be normally distributed. The Levene test is an alternative to the Bartlett test that is less sensitive to departures from normality.
- **Levene's test**: Compare the variances of k samples, where k can be more than two samples. It's an alternative to the Bartlett's test that is less sensitive to departures from normality.
- **Fligner-Killeen test**: a non-parametric test which is very robust against departures from normality.

■ 就性別在「生活壓力」的差異比較而言，「變異數相等的 Levene檢定」之 f 值未達顯著差異 ($f=0.01, p=0.976>0.05$)，表示二組樣本變異數同質。

■ 在假設「變異數相等」之下的 t 檢定已達 0.05 顯著水準 ($t=-3.171, p=0.003<0.05$)。平均數的差異值為 -7.667，表示男女運動員的「生活壓力」感受有顯著差異存在。其中女生的生活壓力感受顯著高於男生。

相依樣本t檢定

- **適用時機:** 同一組受試者接受前後兩次測驗時，二次測驗量值之平均數的差異比較。
- **範例:** 一份有效教學指標問卷中
 - **三個層面:** 教學經營(a01~a05), 教學活動(a06~a10), 輔導追蹤(a11~a15)
 - **目的:** 探討教師對13題題項指標重要性的看法，了解教師在班級中實踐的程度。
 - **問題:** 問卷樣本，教師對三大指標層面的重要性知覺與實踐程度間是否有差異？





成對樣本問卷範例： 教學檢核問卷

	重 要 性		實 踐 程 度		
	非常不重要	←→非常重要	很少做到	←→常常做到	
一、教學經營層面					
01.教師有完整班級經營計畫與實施教學過程資料。	1□	2□	3□	4□	5□
02.教學情境布置能善用資源，重視整潔、美綠化效果及資源回收等。	1□	2□	3□	4□	5□
03.能於教育活動中適切指導學生之生活教育。	1□				
04.召開班級家長會時，任課老師能提出教學實施相關說明或書面資料。	1□				
05.教師能採多樣化的方式與家長溝通。	1□				

二、教學活動層面					
06.教師能依教學目標妥善運用教學方法實施教學。	1□	2□	3□	4□	5□
07.學習領域教學能適切結合學校本位課程或融入重要議題與時事隨機教學。	1□	2□	3□	4□	5□
08.教師於教學過程中能充分提供多樣化教學素材，讓學生親自操作或體驗學習。	1□				
09.能善用教學資源協助教學。	1□				
10.學生作業批改認真詳實並有助學習。	1□				

三、輔導追蹤層面					
11.針對未達學習目標、行為偏差學童能分析原因，進行適性化教學與輔導措施。	1□	2□	3□	4□	5□
12.隨時留意學生身心健康及學習情形，如發現異常，能通知家長並採取相關輔導措施或尋求支援。	1□	2□	3□	4□	5□
13.積極落實輔導工作與輔導資料的建立，並能妥善維護管理及有效地運用。	1□	2□	3□	4□	5□

教學經營_重要性
教學活動_實踐程度

教學活動_重要性
輔導追蹤_實踐程度

輔導追蹤_重要性
教學經營_實踐程度

題號	01	02	03	04	05	06	07	08	09	10	11	12	13
重要性	A1_1	A1_2	A1_3	A1_4	A1_5	A2_1	A2_2	A2_3	A2_4	A2_5	A3_1	A3_2	A3_3
實踐程度	B1_1	B1_2	B1_3	B1_4	B1_5	B2_1	B2_2	B2_3	B2_4	B2_5	B3_1	B3_2	B3_3

題號	01	02	03	04	05	06	07	08	09	10	11	12	13
重要性	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13
實踐程度	B1	B2	B3	B4	B5	B6	B7	B8	B8	B10	B11	B12	B13



想了解樣本對每層面重要性及實踐性看法是否有顯著差異： 25/55 採用相依樣本t檢定

```
> teaching_survey <- read.csv("data/教學檢核問卷.csv")
> head(teaching_survey)
  經營_重要性 活動_重要性 輔導_重要性 經營_實踐性 活動_實踐性 輔導_實踐性
1           20           24           15           18           20           14
...
6           20           24           14           20           18           13
> dim(teaching_survey)
[1] 12 6
> summary(teaching_survey)
  經營_重要性 活動_重要性 輔導_重要性 經營_實踐性 活動_實踐性 輔導_實踐性
Min.   :18.00  Min.   :20.00  Min.   :10.00  Min.   :18.00  Min.   :15.00  Min.   :10.00
1st Qu.:20.00  1st Qu.:23.00  1st Qu.:12.00  1st Qu.:19.00  1st Qu.:16.75  1st Qu.:11.75
Median :24.00  Median :24.00  Median :13.50  Median :20.00  Median :19.00  Median :12.00
Mean   :22.58  Mean   :23.42  Mean   :13.08  Mean   :20.33  Mean   :19.08  Mean   :12.42
3rd Qu.:24.25  3rd Qu.:24.00  3rd Qu.:14.00  3rd Qu.:21.00  3rd Qu.:21.00  3rd Qu.:14.00
Max.   :25.00  Max.   :25.00  Max.   :15.00  Max.   :24.00  Max.   :24.00  Max.   :15.00
> mapply(cor, teaching_survey[, 1:3], teaching_survey[, 4:6])
  經營_重要性 活動_重要性 輔導_重要性
0.5318767    0.6697263    0.7556197
>
> mapply(t.test, teaching_survey[, 1:3], teaching_survey[, 4:6], paired = TRUE)

  經營_重要性 活動_重要性
statistic    3.446738      6.5
parameter    11          11
p.value      0.00545938    4.427706e-05
...
  輔導_重要性
statistic    2.15211
parameter    11
p.value      0.05444671
...
```

- 想了解樣本對兩兩層面的重要性看法是否有顯著差異=> 採用相依樣本t檢定。
- 由於三個層面包含的題項數不同，可先求出各層面單題的平均得分，再進行比較。

```

> teaching_survey$經營_重要性_單題平均 <- teaching_survey$經營_重要性/5
> teaching_survey$活動_重要性_單題平均 <- teaching_survey$活動_重要性/5
> teaching_survey$輔導_重要性_單題平均 <- teaching_survey$輔導_重要性/3
> t.test(teaching_survey$經營_重要性_單題平均, teaching_survey$活動_重要性_單題平均,
+        paired = TRUE)
      Paired t-test

data:  teaching_survey$經營_重要性_單題平均 and teaching_survey$活動_重要性_單題平均
t = -0.78789, df = 11, p-value = 0.4474
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.6322542  0.2989209
sample estimates:
mean difference
 -0.1666667

> t.test(teaching_survey$經營_重要性_單題平均, teaching_survey$輔導_重要性_單題平均,
+        paired = TRUE)
      Paired t-test

...
t = 0.69295, df = 11, p-value = 0.5027
...

> t.test(teaching_survey$活動_重要性_單題平均, teaching_survey$輔導_重要性_單題平均,
+        paired = TRUE)
      Paired t-test

...
t = 2.5901, df = 11, p-value = 0.02513
...

```



單因子變異數分析(One-way ANOVA)

27/55

- Using Analysis of Variance, which can be considered to be a **generalization of the t -test**, when
 - compare more than two groups (e.g., *drug 1*, *drug 2*, and *placebo*), or
 - compare groups created by more than one independent variable while controlling for the separate influence of each of them (e.g., *Gender*, *type of Drug*, and *size of Dose*).
- For two group comparisons, ANOVA will give results **identical to a t -test**.
- **One-way** ANOVA compares groups using **one parameter**.
- We can test the following:
 - Are all the means from **more than two populations** equal?
 - Are all the means from **more than two treatments** on one population equal? (This is equivalent to asking whether the treatments have any overall effect.)

One-way ANOVA

Assumptions

- The subjects are sampled **randomly**.
- The groups are **independent**.
- The population variances are **homogenous**.
- The population distribution is **normal** in shape.

As with t tests, violation of homogeneity is particularly a problem when we have quite **different sample sizes**.

- **Homogeneity of variance test**
 - Bartlett's test (1937)
 - Levene's test (Levene 1960)
 - O'Brien (1979)

ANOVA Table

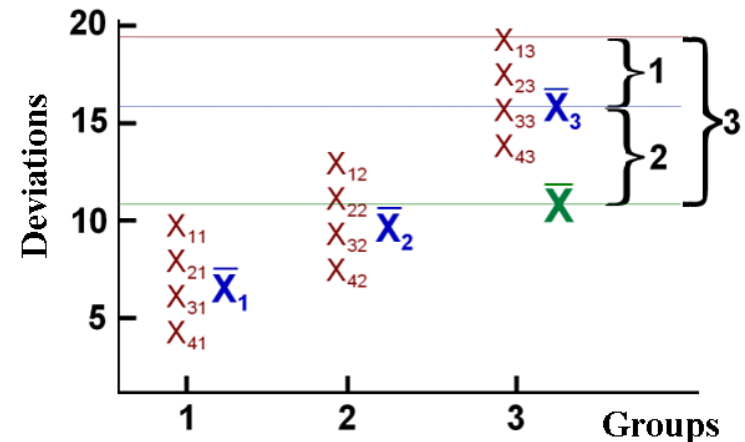
Groups

1	2	...	j	...	k
X_{11}	X_{12}	...	X_{1j}	...	X_{1k}
X_{21}	X_{22}	...	X_{2j}	...	X_{2k}
			...		
X_{i1}	X_{i2}	...	X_{ij}	...	X_{ik}
\vdots			\vdots		$X_{n_k k}$
$X_{n_{11}}$	$X_{n_{22}}$...	$X_{n_{ij}}$...	

$$T_j = \sum_{i=1}^{n_j} X_{ij} \quad \bar{X}_j = \frac{T_j}{n_j}$$

$$T = \sum_{j=1}^k T_j \quad \bar{X} = \frac{T}{N}$$

$$S^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} \frac{(X_{ij} - \bar{X})^2}{N - 1}$$



$$(X_{ij} - \bar{X}) = (X_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X})$$

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$X_{ij} = \mu_j + \epsilon_{ij} \quad \begin{matrix} i = 1, \dots, n_j \\ j = 1, \dots, k \end{matrix}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} [(X_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X})]^2$$

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{X}_j - \bar{X})^2$$

ANOVA Table

Source	SS	df	MS	F	p
Between	SS_B	$p - 1$	MS_B	MS_B / MS_W	< 0.05
Within	SS_W	$N - p$	MS_W		
Total	SS_T	$N - 1$			

$$SS_{Total} = SS_{Within} + SS_{Between}$$

$$F = \frac{MS_{Between}}{MS_{Within}}$$

Reject H_0 , if $F_{obs} > F_{\{\alpha, k-1, N-k\}}$

Post Hoc Tests

Applicable when comparing more than 2 groups.

- One-way ANOVA model : $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 (\dots = \mu_n)$
- If H_0 is rejected for a **gene**, there is still no information about where differences are observed.

How does one determine which specific differences are significant?

Test Name	How it works
Tukey	All means for each condition are ranked in order of magnitude; group with lowest mean gets a ranking of 1. The pairwise differences between means, starting with the largest mean compared to the smallest mean, are tabulated between each group pair and divided by the standard error. This value, q , is compared to a Studentized range critical value. If q is larger than the critical value, then the expression between that group pair is considered to be statistically different.
Student-Newman-Keuls (SNK) test:	This test is similar to the Tukey test, except with regard to how the critical value is determined. All q 's in Tukey's test are compared to the same critical value determined for that experiment; whereas all q 's determined from SNK test are compared to a different critical value. This makes the SNK test slightly less conservative than the Tukey test.



Student-Newman-Keuls (SNK) Test

assuming
equal sample sizes and
homogeneity of variance

Group	A	B	C	D
Mean	2	3	7	8

alpha = 0.01
n = 5
df = 16

$$\sqrt{\frac{MSE}{n}} = \sqrt{\frac{.5}{5}} = 0.316$$

$$q = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{MSE}{n}}}$$

SNK.test {agricolae}
snk.test {GAD}

“r” is the number of means spanned by a given comparison.
r, df, alpha → studentized range statistic q

1. r = 4, q_{.01} = 5.19

A vs D: $q = \frac{8 - 2}{0.316} = 18.99, p < 0.01$

2. r = 3, q_{.01} = 4.79

a. A vs C: $q = \frac{7 - 2}{0.316} = 15.82, p < 0.01$

b. B vs D: $q = \frac{8 - 3}{0.316} = 15.82, p < 0.01$

3. r = 2, q_{.01} = 4.13

a. A vs B: $q = \frac{3 - 2}{.316} = 3.16, p > 0.01$

b. B vs C: $q = \frac{7 - 3}{.316} = 12.66, p < 0.01$

c. C vs D: $q = \frac{8 - 7}{.316} = 3.16, p > 0.01$

Honestly Significant Difference (HSD)

$$HSD = q \sqrt{\frac{MS_{within}}{n}} \quad \frac{M_1 - M_2}{\sqrt{MS_w \left(\frac{1}{n}\right)}}$$

Tukey's HSD Post-hoc test is applied in exactly the same way that the Student-Newman-Keuls is, with the exception that r is set at k for all comparisons.
 (k vs 1, k vs 2,..., k vs k-1) (k-1 vs 1, k-1 vs 2,..., k-1 vs k-2) ...(...2 vs 1)

$r = k$, df , α → studentized range statistic q

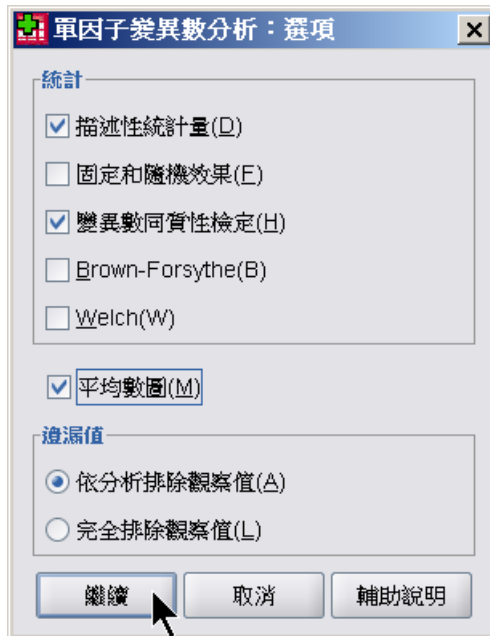
- All alpha's in Tukey's test are compared to the same critical value.
- All alpha's in SKN test are compared to a different critical value.
- This test is more conservative (less powerful) than the Student-Newman-Keuls.

單因子變異數分析(ANOVA) 範例

33/55

■ 問題:

- 不同年齡的運動員在「生活壓力」(「社會支持」、「身心倦怠」)是否有顯著的差異?
- 若F值達到顯著，表示至少有二個組別平均數的差達到顯著水準。
- 進行「事後比較」，找出是哪幾個配對組的平均數有顯著差異。



one-way anova

```

> example_data <- read.csv("data/平均數差異檢定_1.csv")
> head(example_data)
  性別  年齡  生活壓力  社會支持  身心倦怠
1    1    1    15      40      18
...
6    1    1    32      25      25
> dim(example_data)
[1] 40  5
>
> example_data$性別2 <- as.factor(c("男生", "女生")[example_data$性別])
> example_data$年齡2 <- as.factor(c("25歲以下", "26-30歲", "31歲以上")[example_data$年齡])
>
> item <- c("生活壓力", "社會支持", "身心倦怠")
> describeBy(example_data$生活壓力, example_data$年齡2)

```

```

Descriptive statistics by group
group: 25歲以下
  vars  n mean   sd median trimmed  mad min max range skew kurtosis  se
X1    1 14  29 9.54  28.5  28.83 6.67 15 45  30 0.25  -1.01 2.55
-----
group: 26-30歲
  vars  n mean   sd median trimmed  mad min max range skew kurtosis  se
X1    1 14 37.43 5.81  36  37.42 5.93 30 45  15 0.2  -1.67 1.55
-----
group: 31歲以上
  vars  n mean   sd median trimmed  mad min max range skew kurtosis  se
X1    1 12 35.83 7.44  36.5  36.6 5.19 20 44  24 -0.84  -0.46 2.15

```

one-way anova

```
> library(car) # Companion to Applied Regression
> my_model <- example_data$生活壓力 ~ example_data$年齡2
> leveneTest(md)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group 1    8e-04 0.9773
      38
> summary(aov(my_model))
              Df Sum Sq Mean Sq F value Pr(>F)
example_data$年齡2  2  554.9   277.45   4.597 0.0165 *
Residuals          37 2233.1    60.35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> oneway.test(my_model)

      One-way analysis of means (not assuming equal variances)

data:  example_data$生活壓力 and example_data$年齡2
F = 3.9072, num df = 2.000, denom df = 23.287, p-value = 0.03442

> oneway.test(my_model, var.equal = T)

      One-way analysis of means

data:  example_data$生活壓力 and example_data$年齡2
F = 4.5971, num df = 2, denom df = 37, p-value = 0.01648
```

Perform SNK Test, Tukey HSD Test in R

```

> library(agricolae)
> my_aov <- aov(生活壓力 ~ 年齡2, data = example_data)
> result <- SNK.test(my_aov, "年齡2")
> result
$statistics
  MSerror Df Mean      CV
 60.35393 37   34 22.84935

$parameters
 test name.t ntr alpha
  SNK  年齡2   3  0.05

...

$means
      生活壓力      std  r      se Min Max  Q25  Q50  Q75
25歲以下 29.00000 9.543423 14 2.076293  15  45 24.0 28.5 32.00
26-30歲  37.42857 5.813966 14 2.076293  30  45 32.5 36.0 43.75
31歲以上 35.83333 7.444746 12 2.242653  20  44 34.0 36.5 40.50

...

$groups
      生活壓力  groups
26-30歲  37.42857    a
31歲以上 35.83333    a
25歲以下 29.00000    b

attr(,"class")
[1] "group"

```

```

> TukeyHSD(aov(my_model))
Tukey multiple comparisons of means
 95% family-wise confidence level

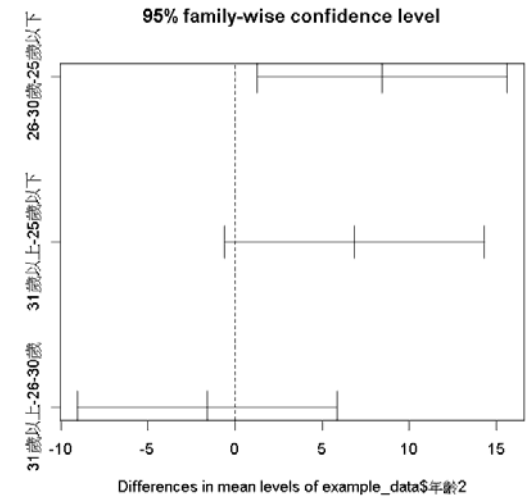
Fit: aov(formula = my_model)

$`example_data$年齡2`
      diff      lwr      upr      p adj
26-30歲-25歲以下  8.428571  1.2595839 15.597559  0.0180742
31歲以上-25歲以下  6.833333 -0.6283855 14.295052  0.0782887
31歲以上-26-30歲 -1.595238 -9.0569569  5.866481  0.8611283
>
> plot(TukeyHSD(aov(my_model)))

```

注意事項:

- 變異數分析中達顯著，但採用「Scheffe」事後檢定法，則無出現成對組的平均數差異達到顯著的結果。
- 原因: Scheffe法是最嚴格，統計定力最低的一種多重比較法。
- 通常發生在F值的p-value在0.05附近。可改用「HSD法」，以便和整體檢定F值的顯著性相呼應。



- 學生性別、數學焦慮、數學態度、數學投入動機是否可有效預測學生的數學成就？其預測力如何？
- 工作壓力五個層面: 工作負荷，人際關係，專業能力，學生行為，角色衝突，是否可以有效預測學校效能？
- 國中學生的智力，畢業成績，三年級模擬考平均成績，期望動機等四個變項是否可以有效預測其基本學力測驗成績？
- 警察人員婚姻態度三個面向; 親子關係有三個面向。此六個變項是否可以有效預測警察人員的幸福感？

- **簡單線性迴歸分析 (simple linear regression):**
 - 探討一個自變項(independent variable)對一個依變項(dependent variable)的影響。
 - 自變項(X): 又稱預測變項 (predictor) , 解釋變項 (explanatory variable) 。
 - 依變項(Y): 效標變項 (criterion) , 反應變項 (response) 。
- **迴歸分析目的:**
 - 找出一個自變數的線性組合(迴歸方程式) , 來描述一組自變項與依變項的關係。
 - 描述 , 解釋 , 預測 。
- **複迴歸(多重迴歸)分析(multiple linear regression):**
 - 又稱多重迴歸分析、多元線性迴歸分析。
 - 探討多個自變項(X's)對一個依變項(Y)的影響。

變數的型態

- 多重迴歸，一般而言依變數與自變數皆需為「連續變數」。
- 依變數(Y)是二分類別變數:
 - 區別分析(discriminant analysis)
 - 二元logistic迴歸分析
- 依變數(Y)是多分類別變數:
 - 區別分析(discriminant analysis)
 - 多項logistic迴歸分析
- 自變項(X's)若為類別變項，則需轉化為 dummy variables (啞變數，虛擬變數)

一般迴歸分析所包含的內容

- 虛擬變數轉換
- 參數估計: beta's
- 模型檢定
 - F檢定: 探討迴歸模式中的所有斜率係數是否全部為0 (變異數分析)。
 - t檢定: 探討個別迴歸係數是否顯著異於0: 共p個t檢定。
- 解釋
 - 迴歸係數
 - 判定係數 R^2
- 自變數選擇
- 殘差分析
- 處理共線性問題

虛擬變數 (dummy variable)

- 類別變數在投入迴歸分析時，必須轉換為虛擬變數。
- 以「0」、「1」的方式表示，虛擬變項個數等於「水準」個數減一。
- 虛擬變項之迴歸係數解釋和非虛擬變項之迴歸係數不同。
- 非虛擬變項之迴歸係數解釋是和參照組相比較，採用相關係數的解釋原理 ($\text{cor}(X, y)$)。
- 例如: 樣本的學習動機(X)愈強，其學業成績(y)愈佳。

原變項	虛擬變項
學生性別	sexd
男性 1	0
女性 2	1

原變項	虛擬變項	
家庭狀況	homd1	homd2
單親家庭組 1	1	0
他人照顧組 2	0	1
雙親家庭組 3	0	0

原變項	虛擬變項		
地理位置	locd1	locd2	locd3
北部 1	1	0	0
中部 2	0	1	0
南部 3	0	0	1
東部 4	0	0	0

職務 (原始變項)	職務虛擬_1	職務虛擬_3	職務虛擬_4
1 主任	1	0	0
2 組長 (參照組)	0	0	0
3 科任	0	1	0
4 級任	0	0	1

職務 (原始變項)	職務虛擬_1	職務虛擬_2	職務虛擬_4
1 主任	1	0	0
2 組長	0	1	0
3 科任 (參照組)	0	0	0
4 級任	0	0	1

職務 (原始變項)	職務虛擬_1	職務虛擬_3	職務虛擬_3
1 主任	1	0	0
2 組長	0	1	0
3 科任	0	0	1
4 級任 (參照組)	0	0	0

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

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

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (\text{total variability in the observations})$$

$$SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (\text{residual sum of squares})$$

$$SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (\text{regression sum of squares})$$

$$S_{yy} = SS_R + SS_E$$

Source of variation	Sum of squares	Degree of freedom	Mean square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R / MS_E
Residuals	$SS_E = S_{yy} - \hat{\beta}_1 S_{xy}$	$n - 1$	MS_E	
Total	S_{yy}	$n - 2$		



判定係數 (Determination Coefficient)^{43/55}

■ R^2

- 總變異中可被解釋之百分比例
- 模式配適度(Goodness of Fit)之指標。

$$R^2 = \frac{SSR}{SST} \quad 0 \leq R^2 \leq 1$$

■ Adjusted R2

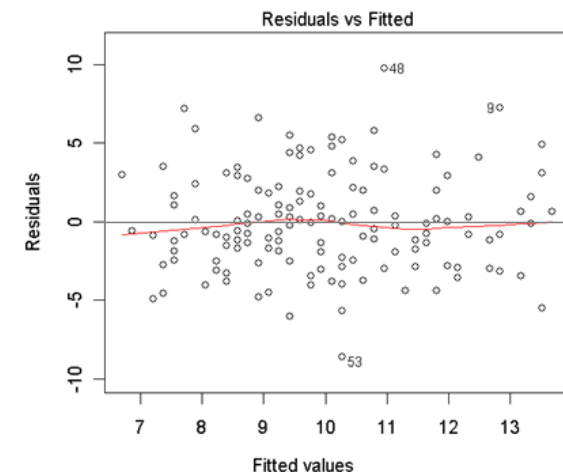
- 調整後的判定係數：
- 在迴歸分析中，如果自變項的個數很多，有時候就要用調整後的判定係數代替原先的判定係數，因為增加新的自變項後，均會使 R^2 變大。

- 迴歸分析乃以殘差值(e_i , Residual)為誤差項(ϵ_i)之估計，等於樣本觀察值與預測值之差，即：

$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

- 在探討誤差項是否符合常態性、變異數恆常性、獨立性等三項假定。
- Durbin-Watson檢定為觀察樣本獨立性檢定，檢定模型中是否存在自我相關：
 - DW範圍(0-2)，表示殘差項間呈現正相關。
 - DW範圍(2-4)，表示殘差項間呈現負相關。
 - DW愈接近2，表示殘差項間無自我相關。

- 殘差圖 (residual plot): e_i against \hat{y}_i



- **共線性(collinearity):**
 - 變項間有共線性問題，表示一個自變數是其它自變項的線性組合。
 - 自變項間的相關太高，則變項迴歸係數的估計值不夠穩定，其計算值也會有很大誤差。

- **判別方法:**
 - **容忍度(tolerance) = $1 - R_i^2$; (允差)**
 - R_i^2 是此自變項與其它自變項間的多元相關係數的平方，
 - 模式中其它自變項對這個變項的有效解釋能力。
 - 值介於0至1間。
 - 容忍度接近於0，表示此變項與其它自變項間有共線性問題。

 - **變異數膨脹因素(variance inflation factor ; VIF)**
 - 為容忍度的倒數，VIF的值愈大，表示自變項的容忍度愈小，愈有共線性問題。
 - >10: 變項間愈有重的問題。

The Variance Inflation Factors

$$X_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon. \quad VIF_j = \frac{1}{1 - R_j^2}$$

- A VIF for a single explanatory variable is obtained using the R-squared value of the regression of **that variable** X_j against all other explanatory variables.
- A VIF measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.

VIF	Status of predictors
VIF = 1	Not correlated
$1 < VIF < 5$	Moderately correlated
$VIF > 5$ to 10	Highly correlated

■ 以邏輯基礎選擇:

- 理論實證、邏輯推理、專家共識。

■ 以統計量基礎選擇:

- 利用每一解釋變數對應之偏F統計量值之大小決定刪去或留在模式中
- 其方法有
 1. 輸入法 (enter)(強迫進入變數法)
 2. 後退淘汰法(Backward Elimination Procedure)
 3. 後退選擇法(Backward Selection Procedure)
 4. 前進選擇法(Forward Selection Procedure)
 5. 逐步迴歸法(Stepwise Regression Procedure)

迴歸分析範例(1)

- 想了解「企業組織環境，企業組織學習，企業知識管理」對「企業組織效能」的影響。
- 「企業組織現況調查」問卷四個量表：
 - 企業組織環境 (二個層面): 福利措施，同儕關係
 - 企業組織學習 (二個層面): 適應學習，創新學習
 - 企業知識管理 (三個層面): 知識獲取，知識流通，知識創新
 - 企業組織效能 (一個層面)。
- 分層隨機抽樣1300員工，回收有效問卷1200份。
- **研究問題:** 七個層面變項對企業組織效能是否有顯著的解釋力，其聯合解釋變異量多少？



基本統計量

```
> organization_data <- read.csv("data/組織效能_1.csv")
> head(organization_data)
  福利措施 同儕關係 適應學習 創新學習 知識獲取 知識流通 知識創新 財務控管
1         7         8         7         5         9         8         6         5
...
6         8         9        14        14        16        13        10        6
  顧客認同 內部運作 學習成長 組織效能
1         5         5         5        20
...
6        10         9        10        35
> dim(organization_data)
[1] 1200  12
> apply(organization_data, 2, mean)
福利措施 同儕關係 適應學習 創新學習 知識獲取 知識流通 知識創新 財務控管
13.87167 14.46750 25.04500 17.38750 28.41667 20.92333 13.43333 17.53083
顧客認同 內部運作 學習成長 組織效能
17.67083 17.37417 17.48167 70.05750
> apply(organization_data, 2, sd)
福利措施 同儕關係 適應學習 創新學習 知識獲取 知識流通 知識創新
2.924150 2.725522 4.785610 3.572765 5.504844 3.655625 3.071658
財務控管 顧客認同 內部運作 學習成長 組織效能
3.615007 3.388025 3.320395 3.465979 12.384210
> cor(organization_data)
      福利措施 同儕關係 適應學習 創新學習 知識獲取 知識流通
福利措施 1.0000000 0.7614193 0.7009475 0.7174237 0.6427968 0.6583684
同儕關係 0.7614193 1.0000000 0.6935755 0.6641817 0.6474550 0.6587031
...
組織效能 0.6298855 0.8667116 0.9179857 0.9276875 0.8830393 1.0000000
```

anova

```
> org_lm <- lm(組織效能 ~ 福利措施 + 同儕關係 + 適應學習 + 創新學習 +
+ 知識獲取 + 知識流通 + 知識創新, data = organization_data)
> org_lm
> summary(org_lm)
```

Call:

```
lm(formula = 組織效能 ~ 福利措施 + 同儕關係 + 適應學習 +
  創新學習 + 知識獲取 + 知識流通 + 知識創新,
  data = organization_data)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-27.2142	-4.5164	0.0287	4.6076	30.1969

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	14.61246	1.45305	10.056	< 2e-16 ***
福利措施	0.33095	0.14307	2.313	0.0209 *
同儕關係	0.95540	0.14420	6.625	5.23e-11 ***
適應學習	0.54651	0.08648	6.320	3.69e-10 ***
創新學習	0.05366	0.11677	0.460	0.6459
知識獲取	0.40335	0.07578	5.323	1.22e-07 ***
知識流通	0.14072	0.12296	1.144	0.2527
知識創新	0.59594	0.12826	4.646	3.76e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.048 on 1192 degrees of freedom

Multiple R-squared: 0.5802, Adjusted R-squared: 0.5777

F-statistic: 235.3 on 7 and 1192 DF, p-value: < 2.2e-16

```
> # aov(org_lm)
> anova(org_lm)
Analysis of Variance Table
```

Response: 組織效能

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
福利措施	1	75143	75143	1160.2794	< 2.2e-16 ***
同儕關係	1	12619	12619	194.8408	< 2.2e-16 ***
適應學習	1	11198	11198	172.9093	< 2.2e-16 ***
創新學習	1	986	986	15.2294	0.0001005 ***
知識獲取	1	4809	4809	74.2574	< 2.2e-16 ***
知識流通	1	538	538	8.3039	0.0040268 **
知識創新	1	1398	1398	21.5878	3.756e-06 ***
Residuals	1192	77198	65		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- 七個自變項與「組織效能」的多元相關係數: 0.762。
- 決定係數為0.580，表示七個自變項可解釋「組織效能」58%的變異量。
- 七個自變項的標準化迴歸係數均為正數，表示七個自變項對「組織效能」的影響均為正向。
- 對「組織效能」有顯著影響為「福利措施」、「同儕關係」、「適應學習」、「知識獲取」、「知識創新」。
- 「適應學習」及「同儕關係」beta係數絕對值較大，表示這兩個預測變數對「企業組織效能」有較高的解釋變異量。

```
> library(car)
> vif(org_lm)
福利措施 同儕關係 適應學習 創新學習 知識獲取 知識流通 知識創新
3.240216 2.859886 3.170670 3.222117 3.221654 3.740568 2.873672
```

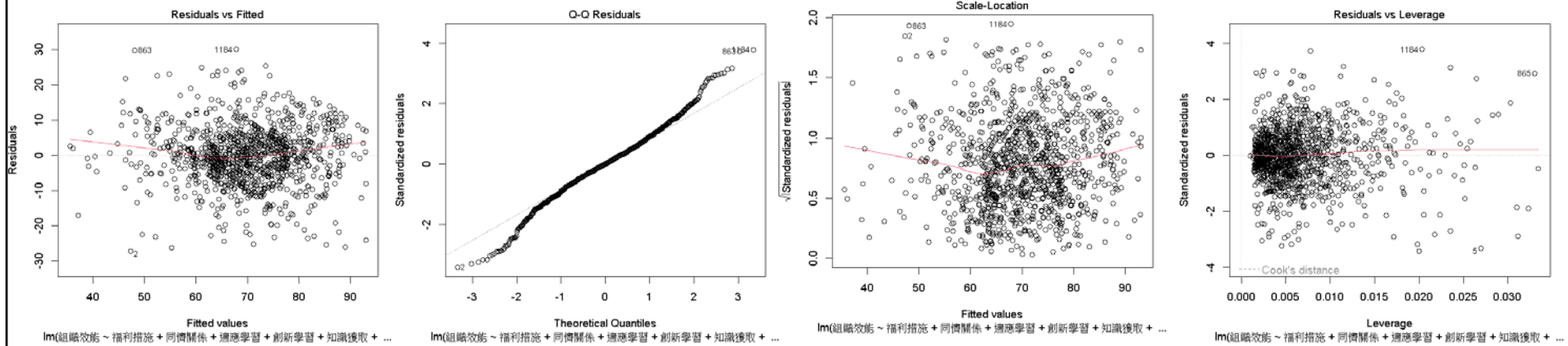
VIF	Status of predictors
VIF = 1	Not correlated
1 < VIF < 5	Moderately correlated
VIF > 5 to 10	Highly correlated

The leverage score for the i th independent observation

$$Le_i = \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

$$h_{ii} = [\mathbf{H}]_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$$

```
> plot(org_lm)
```



迴歸分析 (含虛擬變項) 範例

- 探討「學校規模」，「教育人員職務」與「工作壓力」的關係。
- 進行「單因子變異數分析」：不同學校規模的國中教師，其工作壓力有顯著差異。

```
> school_data <- read.csv("data/虛擬迴歸.csv")
> head(school_data)
  規模 職務 工作壓力 規模_虛擬1 規模_虛擬2 職務_虛擬1 職務_虛擬2 職務_虛擬4
1    1    1         30          1          1          0          1          0          0
2    1    1         29          1          1          0          1          0          0
3    1    1         28          1          1          0          1          0          0
4    1    3         27          1          1          0          0          0          0
5    1    3         27          1          1          0          0          0          0
6    1    1         26          1          1          0          1          0          0

> dim(school_data)
[1] 48  8

> str(school_data)
'data.frame':
  48 obs. of  8 variables:
 $ 規模      : int  1 1 1 1 1 1 3 3 3 3 ...
 $ 職務      : int  1 1 1 3 3 1 1 1 1 1 ...
 $ 工作壓力  : int  30 29 28 27 27 26 28 27 25 26 ..
 $ 規模_虛擬1: int  1 1 1 1 1 1 0 0 0 0 ...
 $ 規模_虛擬2: int  0 0 0 0 0 0 0 0 0 0 ...
 $ 職務_虛擬1: int  1 1 1 0 0 1 1 1 1 1 ...
 $ 職務_虛擬2: int  0 0 0 0 0 0 0 0 0 0 ...
 $ 職務_虛擬4: int  0 0 0 0 0 0 0 0 0 0 ...
```

原始變項		規模變項之虛擬變數		職務變項之虛擬變數		
規模	職務	規模-虛擬1	規模-虛擬2	職務-虛擬1	職務-虛擬2	職務-虛擬4
1 大型	1 主任	1	0	1	0	0
2 中型	2 組長	0	1	0	1	0
3 小型#	3 科任#	0	0	0	0	0
1 大型	4 級任	1	0	0	0	1
2 中型	1 主任	0	1	1	0	0
3 小型#	2 組長	0	0	0	1	0
1 大型	3 科任#	1	0	0	0	0

#為該變項的參照組

【備註】：上述虛擬變項的變項標記如下：「規模_虛擬1」為大型&小型對比、「規模_虛擬2」為中型&小型對比、「職務_虛擬1」為主任&科任對比、「職務_虛擬2」為組長&科任對比、「職務_虛擬4」為級任&科任對比。

迴歸分析

```
> size <- c("大型", "中型", "小型")
> position <- c("主任", "組長", "科任", "級任")
> school_data$學校規模 <- factor(size[school_data$規模], levels = rev(size), ordered = TRUE)
> levels(school_data$學校規模)
[1] "小型" "中型" "大型"
> school_data$教育人員職務 <- factor(position[school_data$職務])
> levels(school_data$教育人員職務)
[1] "主任" "科任" "級任" "組長"
> school_lm <- lm(工作壓力 ~ 學校規模 + 教育人員職務, data = school_data)
> summary(school_lm)
```

Call:

```
lm(formula = 工作壓力 ~ 學校規模 + 教育人員職務,
    data = school_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.7564	-1.5074	-0.1519	2.0646	7.6534

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	25.1601	1.0518	23.921	< 2e-16
學校規模.L	2.3872	0.8834	2.702	0.00990 **
學校規模.Q	4.1687	0.9554	4.363	8.16e-05 ***
教育人員職務科任	-9.2033	1.4614	-6.297	1.48e-07 ***
教育人員職務級任	-2.4202	1.4743	-1.642	0.10815
教育人員職務組長	-4.8210	1.5018	-3.210	0.00254 **

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.35 on 42 degrees of freedom

Multiple R-squared: 0.7313, Adjusted R-squared: 0.6993

F-statistic: 22.86 on 5 and 42 DF, p-value: 5.254e-11

```
> anova(school_lm)
```

Analysis of Variance Table

Response: 工作壓力

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
學校規模	2	808.69	404.34	36.02	7.763e-10 ***
教育人員職務	3	474.16	158.05	14.08	1.711e-06 ***
Residuals	42	471.47	11.23		

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

迴歸分析: 指定baseline

```
> school_data$學校規模 <- relevel(factor(size[school_data$規模]), "小型")
> levels(school_data$學校規模)
[1] "小型" "大型" "中型"
> school_data$教育人員職務 <- relevel(factor(position[school_data$職務]), "科任")
> levels(school_data$教育人員職務)
[1] "科任" "主任" "級任" "組長"
> school_lm <- lm(工作壓力 ~ 學校規模 + 教育人員職務, data = school_data)
> summary(school_lm)
```

```
Call:
lm(formula = 工作壓力 ~ 學校規模 + 教育人員職務,
    data = school_data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.7564  -1.5074  -0.1519   2.0646   7.6534
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      15.971      1.205  13.258 < 2e-16 ***
學校規模大型       3.376      1.249   2.702  0.00990 **
學校規模中型     -3.418      1.260  -2.713  0.00962 **
教育人員職務主任   9.203      1.461   6.297 1.48e-07 ***
教育人員職務級任   6.783      1.507   4.501 5.29e-05 ***
教育人員職務組長   4.382      1.292   3.392  0.00152 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.35 on 42 degrees of freedom
Multiple R-squared:  0.7313,    Adjusted R-squared:  0.6993
F-statistic: 22.86 on 5 and 42 DF,  p-value: 5.254e-11
```

```
> anova(school_lm)
Analysis of Variance Table
```

```
Response: 工作壓力
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
學校規模	2	808.69	404.34	36.02	7.763e-10 ***
教育人員職務	3	474.16	158.05	14.08	1.711e-06 ***
Residuals	42	471.47	11.23		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

迴歸分析: 若未指定正確的類別

```
> school_data$規模_文字 <- size[school_data$規模]
> school_data$職務_文字 <- position[school_data$職務]
> test_lm <- lm(工作壓力 ~ 規模_文字 + 職務_文字, data = school_data)
> summary(test_lm)
```

Call:

```
lm(formula = 工作壓力 ~ 規模_文字 + 職務_文字, data = school_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.7564	-1.5074	-0.1519	2.0646	7.6534

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	28.550	1.263	22.603	< 2e-16 ***
規模_文字小型	-3.376	1.249	-2.702	0.00990 **
規模_文字中型	-6.794	1.390	-4.887	1.53e-05 ***
職務_文字科任	-9.203	1.461	-6.297	1.48e-07 ***
職務_文字級任	-2.420	1.474	-1.642	0.10815
職務_文字組長	-4.821	1.502	-3.210	0.00254 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.35 on 42 degrees of freedom

Multiple R-squared: 0.7313, Adjusted R-squared: 0.6993

F-statistic: 22.86 on 5 and 42 DF, p-value: 5.254e-11

Wrong

```
test_lm <- lm(工作壓力 ~ 規模 + 職務, data = school_data)
summary(test_lm)
anova(test_lm)
```

```
> anova(test_lm)
```

Analysis of Variance Table

Response: 工作壓力

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
規模_文字	2	808.69	404.34	36.02	7.763e-10 ***
職務_文字	3	474.16	158.05	14.08	1.711e-06 ***
Residuals	42	471.47	11.23		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

若X是類別型自變數，在R中為「character」class，是否轉成factor再進行迴歸分析，無差別。跟指定的baseline category有關。