統計軟體期末報告 SAS

（SAS內建Heart資料集）

找導教授: 黃俞閔 老師

學生: 王奎貿

```
/*建立一個資料集heart，並將內建資料存入此資料集*/
data heart;set sashelp.heart;
/* 將缺失值移除 */
if DeathCause = " " then delete;
if AgeCHDdiag = " " then delete;
if Cholesterol = "." then delete;
if Chol_Status = " " then delete;
if Smoking =. then delete;
if Weight =. then delete;
if MRW =. then delete;
if Height =. then delete;
/* 只留想分析的變數 */
drop AgeAtStart AgeAtDeath AgeCHDdiag;
/*  先針對這幾個變數做排序*/
proc sort data= heart ; by DeathCause Chol_Status BP_Status
Weight_Status Smoking_Status;
/* 再針對性別做排序 */
proc sort data=heart;by sex;
proc print;run;

/*max min mean std p1 p10 p50*/
/* 針對身高、體重、MRW找最大值、最小值並取小數點第二位*/
proc means data = heart max min maxdec = 2;
var Height Weight MRW;
run;
/* 計算男女的平均身高體重、MRW、膽固醇並取小數第二位*/
proc means data = heart MEAN   maxdec = 2;
class sex;
var Height Weight MRW Cholesterol;
run;
/* 計算男女且不同死亡原因下的身高體重、MRW的平均值、標準差、
第一、第十、第五十百分位數、最大值、最小值並取小數第二位*/
proc means data = heart MEAN STD p1 p10 p50 max min maxdec = 2;
class DeathCause sex;
var Height Weight MRW;
run;

/*frequency*/
```

```
/* 從資料集heart取出感興趣的變數並建立一個新的表格heart1 */
PROC SQL;
create table heart1 as
SELECT DeathCause, sex, Height, Weight, Diastolic, Systolic, MRW,
Smoking, Cholesterol, Chol_Status, BP_Status, Weight_Status,
Smoking_Status
    FROM
    heart
;
RUN;

/* 在男女下，根據資料heart1中的變數計算次數分配表(每個表都列出) */
proc FREQ data = heart1 order = freq;
tables DeathCause BP_Status Weight_Status Smoking_Status;
by sex;
run;

/* 在男女下，根據資料heart1中的變數計算次數分配表 (整合成男女比較表)*/
proc FREQ data = heart1 order = freq;
tables sex*(DeathCause BP_Status Weight_Status Smoking_Status);
run;

/*histogram*/
/* 針對男女的變數做直方圖，並調整組中心點    */
proc univariate data = heart;
histogram Height
/ midpoints = 54.75 to 74.5 by 5;
by Sex;
histogram Weight
/ midpoints = 83 to 271 by 15;
by Sex;
histogram MRW
/ midpoints = 81 to 249 by 15;
by Sex;
run;

/*boxplot*/
```

```
/* 針對男女的變數做箱型圖    */
PROC SGPLOT   DATA = heart1;
    VBOX   Height
    / category = sex;
    run;
PROC SGPLOT   DATA = heart1;
    VBOX   Weight
    / category = sex;
    run;
PROC SGPLOT   DATA = heart1;
    VBOX   MRW
    / category = sex;
    run;

/*scatterplot*/
/*  針對男女的變數做散佈圖，並比較不同變數的分布情形    */
PROC sgscatter   DATA = HEART1;
    matrix height weight mrw
    / group = sex;
    title 'height vs. weight vs. mrw by sex';
RUN;

/*TEST*/
/* 在alpha = 0.05下，做雙尾T檢定，比較男女的身高體重、MRW是否
有差異 */
proc ttest data = heart1 sides = 2 alpha = 0.05 h0 = 0;
    title "Two sample t-test example";
    class sex;
    var height Weight mRW;
    run;

 /*回歸分析*/
 /* 模型: 收縮壓 = 舒張壓 + 身高 + 體重 + 膽固醇 + MRW */
proc reg data = HEART;
model Systolic=Diastolic Height weight Cholesterol MRW;
run;

/*model selection*/
/* 根據不同標準做模型選擇，例如: VIF, collinearity, adjusted R2,..etc.
```

```
*/
proc reg data=heart outest=est0;
model Systolic=Diastolic Height weight Cholesterol MRW/tol vif collin
selection=adjrsq sse aic bic cp press ;
output out=out p=p r=r;
title 'selection of model';
proc print;

/*SQL*/
/* 用原有的變數建立一個新的變數，並做成表格 */
PROC SQL ;
SELECT sex,    Smoking_Status,       /* 挑選變數 */
CASE       /*  建立變數值 */
    WHEN sex='Female' and Smoking_status NE 'Non-smoker' THEN
"smoker_F"
    WHEN sex='Male' and Smoking_status NE 'Non-smoker' THEN
"smoker_M"
    else   'NON'
end as if_smoke /* 新變數名   */
FROM heart
;
quit;
/*MACRO*/
/* 命名一個macro巨集資料叫做show_result，並建立變數名稱  */
%MACRO show_result(sex_ ,
Chol_Status_,BP_Status_,Weight_Status_,Smoking_Status_);
proc print DATA= HEART;   /* 從heart資料集挑選變數 */
where sex = "&sex_" and Chol_Status = "&Chol_Status_" and BP_Status =
"&BP_Status_" and Weight_Status = "&Weight_Status_" and
Smoking_Status = "&Smoking_Status_" ;
TITLE "Probable DeathCause due to Chol, BP, Weight";
run;
%MEND; /* 將有興趣的變數值全部呈列表格 */
%show_result(Female,High,Normal,Normal,Non-smoker);
%show_result(Female,Desirable,High,Normal,Non-smoker);
%show_result(Female,Desirable,Normal,Overweight,Non-smoker);
%show_result(Male,High,Normal,Normal,Non-smoker);
%show_result(Male,Desirable,High,Normal,Non-smoker);
%show_result(Male,Desirable,Normal,Overweight,Non-smoker);
```

程式結果

Probable DeathCause due to Chol, BP, Weight

Obs	Status	DeathCause	Sex	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol	Chol_Status	BP_Status	Weight_Status	Smoking_Status
1	Dead	Cancer	Female	60.50	189	104	178	173	0	224	Borderline	High	Overweight	Non-smoker
2	Dead	Cancer	Female	62.50	154	88	150	133	0	217	Borderline	High	Overweight	Non-smoker
3	Dead	Cancer	Female	60.25	133	100	210	122	0	227	Borderline	High	Overweight	Non-smoker
4	Dead	Cancer	Female	60.00	156	90	120	143	0	234	Borderline	High	Overweight	Non-smoker
5	Dead	Cancer	Female	54.75	105	110	200	115	0	210	Borderline	High	Overweight	Non-smoker
6	Dead	Cancer	Female	62.25	136	96	156	117	0	209	Borderline	High	Overweight	Non-smoker

圖 1: 原資料(前六筆)

Sex	N Obs	Variable	Label	Mean
Female	308	Height	Metropolitan Relative Weight	62.00
		Weight		151.68
		MRW		131.86
		Cholesterol		256.01
Male	556	Height	Metropolitan Relative Weight	67.28
		Weight		170.30
		MRW		121.98
		Cholesterol		239.45

圖 2: 男女下，不同變數的平均數

Cause of Death	Sex	N Obs	Variable	Label	Mean	Std Dev	1st Pctl	10th Pctl	50th Pctl	Maximum	Minimum
Cancer	Female	28	Height	Metropolitan Relative Weight	62.78	2.85	54.75	59.50	62.63	68.25	54.75
			Weight		152.96	25.29	105.00	124.00	148.50	208.00	105.00
			MRW		130.14	23.01	91.00	103.00	125.50	173.00	91.00
	Male	59	Height	Metropolitan Relative Weight	67.29	2.43	62.00	64.00	67.25	73.50	62.00
			Weight		167.95	22.43	120.00	138.00	169.00	240.00	120.00
			MRW		120.44	14.34	87.00	99.00	120.00	153.00	87.00
Cerebral Vascular Disease	Female	43	Height	Metropolitan Relative Weight	62.15	2.37	56.50	59.50	61.75	67.50	56.50
			Weight		155.88	26.93	115.00	121.00	150.00	211.00	115.00
			MRW		135.53	24.28	94.00	106.00	133.00	192.00	94.00
	Male	56	Height	Metropolitan Relative Weight	66.83	2.53	62.00	63.75	67.00	71.50	62.00
			Weight		168.86	23.13	117.00	137.00	171.00	229.00	117.00
			MRW		122.50	15.84	91.00	103.00	121.50	170.00	91.00
Coronary Heart Disease	Female	199	Height	Metropolitan Relative Weight	62.02	2.38	56.50	59.00	62.00	68.25	55.00
			Weight		151.72	31.82	91.00	117.00	147.00	271.00	83.00
			MRW		131.63	26.65	86.00	103.00	128.00	249.00	81.00
	Male	382	Height	Metropolitan Relative Weight	67.31	2.68	61.25	64.00	67.50	74.50	58.75
			Weight		170.45	26.20	119.00	137.00	170.50	256.00	106.00
			MRW		121.96	16.80	88.00	101.00	122.00	178.00	82.00

圖 3: 根據不同的男女及死因下，統計摘要表

Sex * DeathCause的表格						
Sex	DeathCause(Cause of Death)					
	Coronary Heart Disease	Cerebral Vascular Disease	Cancer	Other	Unknown	總計
Male	382	56	59	48	11	556
	44.21	6.48	6.83	5.56	1.27	64.35
	68.71	10.07	10.61	8.63	1.98	
	65.75	56.57	67.82	61.54	57.89	
Female	199	43	28	30	8	308
	23.03	4.98	3.24	3.47	0.93	35.65
	64.61	13.96	9.09	9.74	2.60	
	34.25	43.43	32.18	38.46	42.11	
總計	581	99	87	78	19	864
	67.25	11.46	10.07	9.03	2.20	100.00

圖 4 男女及有興趣的變數之雙變量次數表(因表格太多，所以僅放一張作代表)

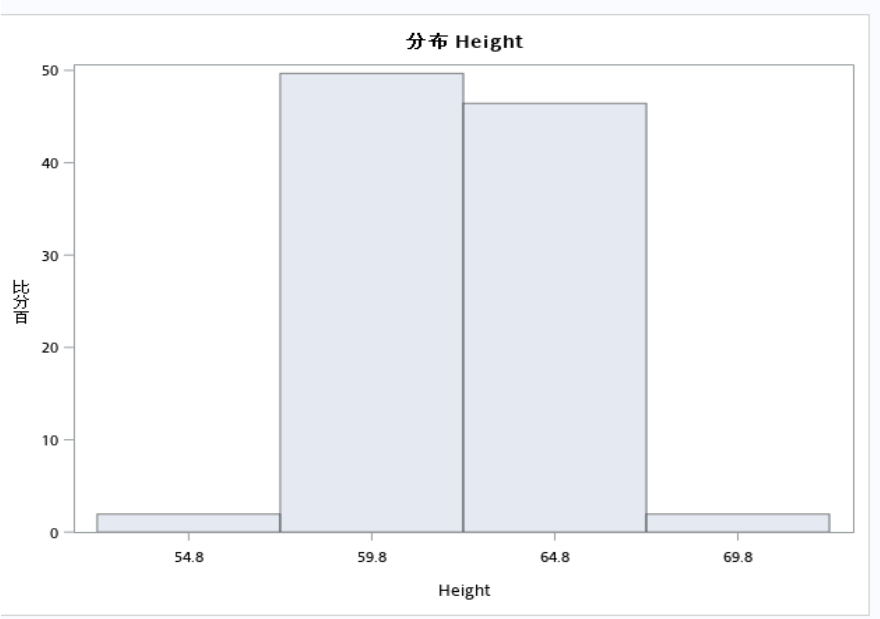

圖 5: 男女及有興趣變數的直方圖(因表格太多，所以僅放一張作代表)

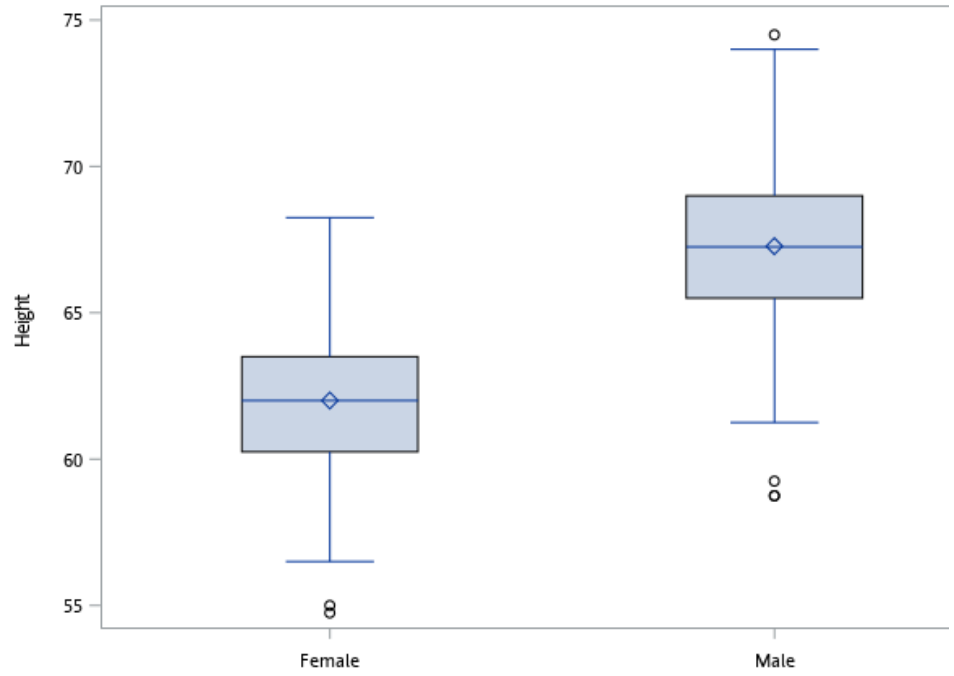

圖 6: 男女及有興趣變數的箱型圖(因表格太多，所以僅放一張作代表)

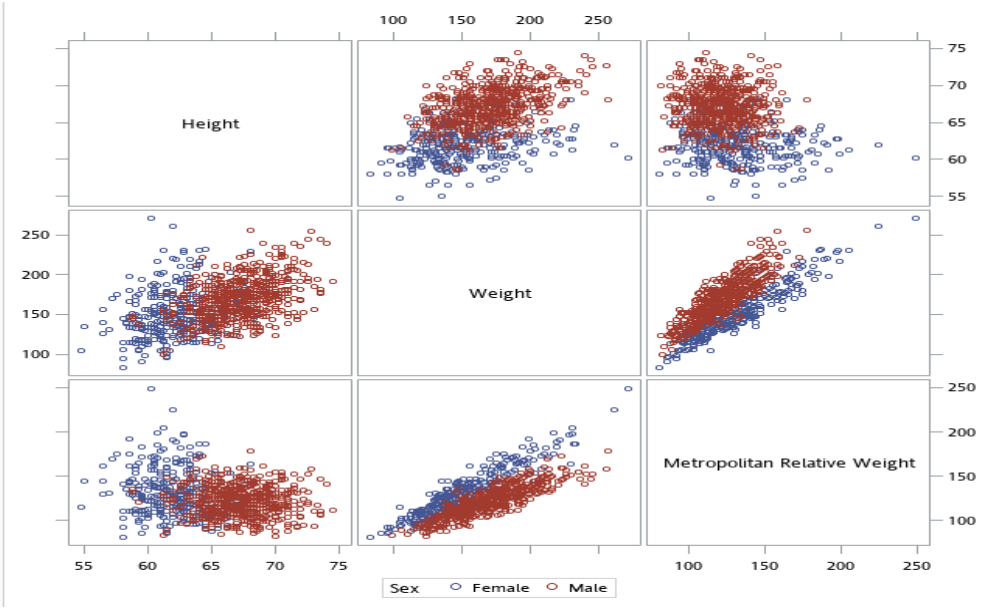

圖 7: 男女及身高體重、MRW 之交叉散佈圖

方法	變異數	DF	t 值	Pr > |t|
集區	均等	862	-29.19	<.0001
Satterthwaite	不均等	670.11	-29.75	<.0001

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	555	307	1.14	0.1917

圖 8: 男女下身高是否有差異之 T 檢定

可以從表上看到，在顯著水準 α=0.05 下，pvalue <0.05，因此兩者身高有差異!且在變異數同質性檢定中，兩者的變異數相等。

方法	變異數	DF	t 值	Pr > |t|
集區	均等	862	-9.4	<.0001
Satterthwaite	不均等	547.72	-9.01	<.0001

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	307	555	1.42	0.0004

圖 9: 男女下體重是否有差異之 T 檢定

可以從表上看到，在顯著水準 α=0.05 下，pvalue <0.05，因此兩者體重有差異!且在變異數同質性檢定中，兩者的變異數不相等。

方法	變異數	DF	t 值	Pr > |t|
集區	均等	862	6.75	<.0001
Satterthwaite	不均等	445.87	5.96	<.0001

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	307	555	2.10	<.0001

圖 10: 男女下 MRW 是否有差異之 T 檢定

可以從表上看到，在顯著水準 α=0.05 下，pvalue <0.05，因此兩者 MRW 有差異!且在變異數同質性檢定中，兩者的變異數不相等。

Two sample t-test example

REG 程序
模型: MODEL1
應變數: Systolic

讀取的觀測值數目	864
使用的觀測值數目	864

變異數的分析

來源	DF	平方和	均方	F 值	Pr > F
模型	5	467286	93457	331.82	<.0001
誤差	858	241801	281.81932		
已校正的總計	863	709087			

根 MSE	16.78748	R 平方	0.6590
應變平均值	150.17130	調整 R 平方	0.6570
變異係數	11.17888		

參數估計值

變數	標籤	DF	參數估計值	標準誤差	t 值	Pr > |t|
Intercept	Intercept	1	-125.34631	50.46746	-2.48	0.0132
Diastolic		1	1.54280	0.04237	36.41	<.0001
Height		1	2.01560	0.78145	2.58	0.0101
Weight		1	-0.55321	0.14333	-3.86	0.0001
Cholesterol		1	0.00351	0.01190	0.30	0.7680
MRW	Metropolitan Relative Weight	1	0.73361	0.17927	4.09	<.0001

圖 11: 迴歸分析的報表

ANOVA 表結果顯示在模型（顯著水準 0.05 下):

收縮壓 = 舒張壓 + 身高 + 體重 + 膽固醇 + MRW

至少有一個參數估計值不為 0（pvalue < 0.05)。

而在報表最下面的變數 cholesterol，在 H_0:$\beta_{cholestreol} = 0$下，p-value =0.76 > 0.05 不拒絕虛無假設，表示此變數的參數估計值為 0

模型中的數目	調整的 R 平方	R 平方	C(p)	AIC	BIC	SSE	模型中的變數
4	0.6574	0.6590	4.0871	4878.1208	4880.1896	241826	Diastolic Height Weight MRW
5	0.6570	0.6590	6.0000	4880.0330	4882.1169	241801	Diastolic Height Weight Cholesterol MRW
3	0.6551	0.6563	8.7070	4882.7607	4884.7541	243691	Diastolic Weight MRW
4	0.6548	0.6564	10.6528	4884.7065	4886.6990	243676	Diastolic Weight Cholesterol MRW
3	0.6518	0.6530	16.9803	4890.9880	4892.9054	246023	Diastolic Height MRW
4	0.6515	0.6531	18.8961	4892.9046	4894.8024	245999	Diastolic Height Cholesterol MRW
3	0.6511	0.6523	18.8409	4892.8274	4894.7280	246547	Diastolic Height Weight
2	0.6510	0.6518	18.0087	4891.9799	4893.8973	246876	Diastolic Height
4	0.6507	0.6523	20.7464	4894.7340	4896.6109	246520	Diastolic Height Weight Cholesterol

圖 12: 模型選擇

（因為變數眾多，所以只放一部份的模型選擇結果）

Sex	Smoking Status	if_smoke
Female	Non-smoker	NON
Female	Non-smoker	NON
Female	Non-smoker	NON
Female	Non-smoker	NON
Female	Non-smoker	NON
Female	Non-smoker	NON
Female	Non-smoker	NON
Female	Very Heavy (> 25)	smoker_F
Female	Very Heavy (> 25)	smoker_F
Female	Non-smoker	NON
Female	Heavy (16-25)	smoker_F
Female	Non-smoker	NON
Female	Moderate (6-15)	smoker_F
Female	Moderate (6-15)	smoker_F
Female	Moderate (6-15)	smoker_F
Female	Non-smoker	NON

圖 13: 用 proc SQL 建立新變數的結果

Obs	Status	DeathCause	Sex	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol	Chol_Status	BP_Status	Weight_Status	Smoking_Status
243	Dead	Coronary Heart Disease	Female	62.00	124	80	140	107	0	347	High	Normal	Normal	Non-smoker
244	Dead	Coronary Heart Disease	Female	62.75	126	80	130	109	0	242	High	Normal	Normal	Non-smoker
245	Dead	Coronary Heart Disease	Female	65.00	136	74	122	106	0	255	High	Normal	Normal	Non-smoker
246	Dead	Coronary Heart Disease	Female	63.50	115	80	112	96	0	250	High	Normal	Normal	Non-smoker
297	Dead	Other	Female	59.25	106	80	140	100	0	267	High	Normal	Normal	Non-smoker
306	Dead	Unknown	Female	59.00	113	78	130	107	0	242	High	Normal	Normal	Non-smoker

圖 14: 用 %MACRO 針對有興趣的變數做的表格（僅放一張做為代表）