迴歸分析作業一

（複回歸）

指導教授：沈葆聖 老師

學生：王奎貿

```
/* data analysis for multiple regression */
/* y = death rate per 1000 residents */
/* x1 = doctor availability per 100,000 residents */
/* x2 = hospital availability per 100,000 residents */
/* x3 = annual per capital income in thousands of dollars */
/* x4 = population density people per square mile */
/* Reference: Life In America's Small Cities, by G.S. Thomas */
data one;
infile 'C:/Users/PC01/Desktop/回歸作業/test.csv' dlm=',';     /*  讀資料    */
input y x1 x2 x3 x4;    /*  放資料的變數  */
proc reg;
model y = x1 x2 x3 x4; /*  做複回歸  */
*restrict intercept=0, hogs-sheep=0;
/*  輸出資料replot    找此資料在模型中的預測值(p)、學生化殘差、去除學生化殘差
*/
output out=resplot p=py student=ry rstudent=sry
/*  判斷影響點的標準*/
cookd=cook covratio=covinf dffits=dffit h=h press=press
/* stdi =  每個預測值的標準差，stdp =  平均預測測值的標準差，STDR =  殘差的標
準差*/
stdi=stdi stdp=stdp stdr=stdr;
title 'diagnostic for multiple regression';
proc print;var cook covinf dffit h press;
proc print;var stdi stdp stdr;
proc gplot;plot ry*py='*'/vref=0; /*y軸為學生化殘差，x軸為預測值，vref:水平
線0的位置  */
title 'plot for studentized residual';
proc gplot;plot sry*py='*'/vref=0; /*y軸為去除學生化殘差，x軸為預測值  */
title 'plot for deletd residual';


/*model selection based on adjusted Rsquare    aic bic;    compute VIF*/
proc reg data=one outest=est0; /*  用資料集one來做迴歸分析，輸出一個est0資
料集放基本統計摘要*/
model y = x1 x2 x3 x4/tol vif collin selection=adjrsq sse aic bic cp press ; /*
針對不同的標準做模型選擇*/
output out=out p=p r=r;    /*輸出資料檔命名為out，裡面放預測值(p)以及殘差(r)
*/
title 'selection of model';
proc print;
```

```
/* model selection based on forward*/
proc reg data=one;      /*用資料集one來做迴歸分析*/
model y = x1 x2 x3 x4/ slstay=0.15 slentry=0.15 /*slstay, slentry  為  forward
selection  選模的門檻*/
selection=forward ss2 sse aic; /* ss2 = Type 2 error */
output out=out1 p=p r=r; /*  輸出資料檔命名為out1，裡面放預測值(p)以及殘差
(r) */
title 'forward selection';
proc print;

/* model selection based on backward*/
proc reg data=one;      /*用資料集one來做迴歸分析*/
model y = x1 x2 x3 x4/ slstay=0.15 slentry=0.15    /* slstay, slentry  為
backward selection  選模的門檻*/
selection=backward ss2 sse aic;
output out=out2 p=p r=r;    /*輸出資料檔命名為out2，裡面放預測值(p)以及殘差
(r) */
title 'backward selection';
proc print;
run;
```

變異數的分析					
來源	DF	平方和	均方	F 值	Pr > F
模型	4	20.65433	5.16358	2.01	0.1075
誤差	48	123.07398	2.56404		
已校正的總計	52	143.72830			

根 MSE	1.60126	R 平方	0.1437
應變平均值	9.30566	調整 R 平方	0.0723
變異係數	17.20740		

參數估計值					
變數	DF	參數估計值	標準誤差	t 值	Pr > |t|
Intercept	1	12.26626	2.02015	6.07	<.0001
x1	1	0.00739	0.00693	1.07	0.2917
x2	1	0.00058372	0.00072191	0.81	0.4228
x3	1	-0.33023	0.23455	-1.41	0.1656
x4	1	-0.00946	0.00489	-1.94	0.0587

圖 1: 回歸分析的 ANOVA

從 ANOVA 表可以看到在顯著水準 α =0.05 下，拒絕 $H_0: \beta_1 = \ldots = \beta_4 = 0$之 pvalue = 0.1075 > 0.05，因此不拒絕虛無假設，代表全部的參數估計值皆為 0。

而且在$H_0: \beta_i = 0$, I = 1 ~ 4 下，每個 pvalue 皆大於 0.05，可以確定每個參數估計值皆為 0

但由於解釋變數的名詞意義上，確實是能夠影響死亡率，因此在之後的模型選擇，仍然是將所有變數放入模型，去作模型選擇。

Obs	cook	covinf	dffit	h	press
1	0.00378	1.11917	-0.13660	0.04663	-1.01957
2	0.00144	1.17049	0.08413	0.06162	0.54778
3	0.05756	0.98388	-0.54383	0.11289	-2.55671
4	0.10381	1.54649	-0.72017	0.35069	-1.94811
5	0.04928	1.10978	0.49892	0.14333	2.09953
6	0.00033	1.19469	-0.04027	0.07210	-0.24262
7	0.00070	1.19879	0.05838	0.07723	0.33982
8	0.00338	1.23947	-0.12877	0.11564	-0.61197
9	0.02103	1.00706	0.32616	0.06359	2.05896
10	0.03779	0.97808	0.43954	0.08506	2.38659

圖 2: 此為觀察值之影響點指標 (前 10 筆)

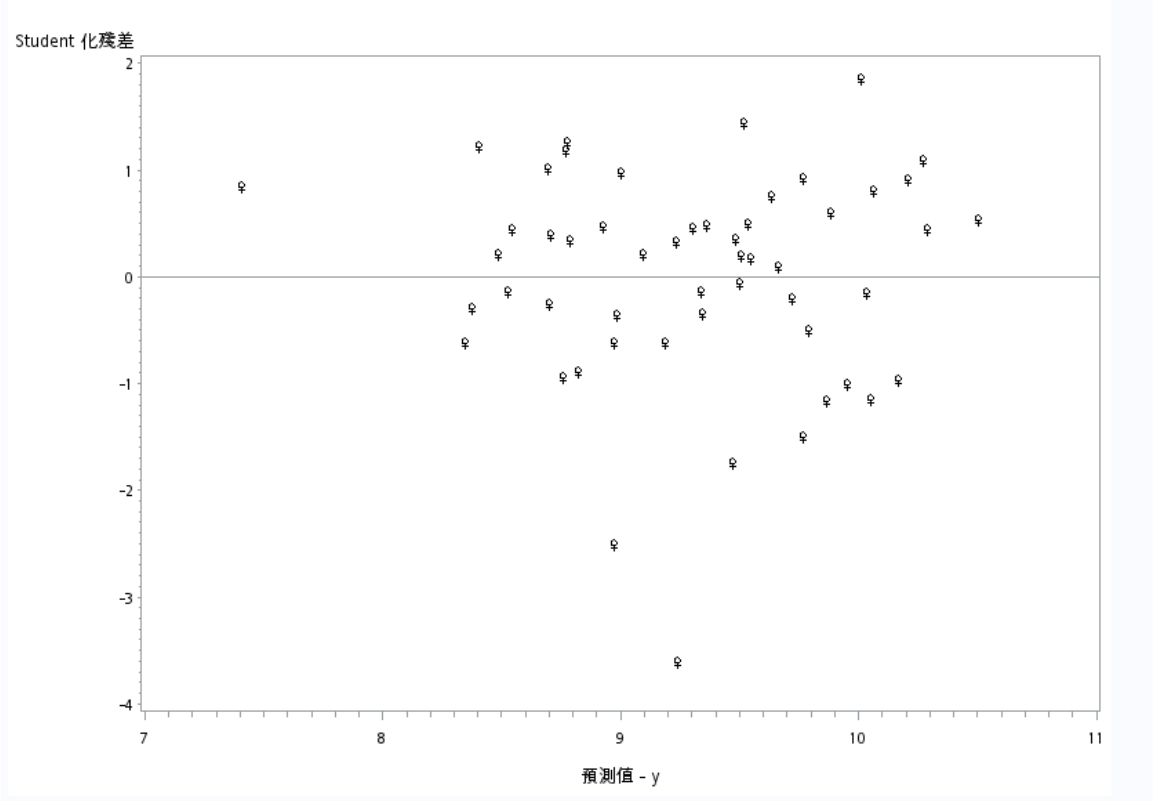

圖 3: 殘差圖

除了兩個離群點，大部分都落在-2~2 之間，滿足模型的變異數同質性。

模型中的變數目	調整的 R 平方	R 平方	C(p)	AIC	BIC	SSE	模型中的變數
3	0.0789	0.1320	3.6538	53.3692	56.0679	124.75030	x1 x3 x4
4	0.0723	0.1437	5.0000	54.6522	57.6721	123.07398	x1 x2 x3 x4
3	0.0698	0.1234	4.1365	53.8924	56.5091	125.98792	x2 x3 x4
2	0.0686	0.1044	3.2038	53.0313	55.3593	128.72455	x2 x4
2	0.0598	0.0959	3.6772	53.5288	55.7996	129.93855	x3 x4
1	0.0590	0.0771	2.7354	52.6241	54.7207	132.65179	x4
3	0.0538	0.1083	4.9822	54.7969	57.2726	128.15652	x1 x2 x4
2	0.0528	0.0892	4.0541	53.9214	56.1472	130.90470	x1 x4
2	0.0371	0.0742	4.8983	54.7907	56.9172	133.06946	x1 x3
3	0.0203	0.0768	6.7497	56.6387	58.8308	132.68834	x1 x2 x3
1	0.0106	0.0296	5.3972	55.2831	57.1790	139.47662	x3
2	0.0046	0.0429	6.6511	56.5512	58.4781	137.56376	x2 x3
1	-.0059	0.0134	6.3042	56.1595	57.9893	141.80212	x1
1	-.0071	0.0122	6.3698	56.2223	58.0475	141.97048	x2
2	-.0194	0.0198	7.9459	57.8150	59.6000	140.88365	x1 x2

圖 4: 不同標準下，模型選擇的結果

如果以調整後 R square 當作標準，最佳模型為: Y = X1 + X3 + X4。

參數估計值						
變數	DF	參數估計值	標準誤差	t 值	Pr > |t|	類型 II SS
Intercept	1	10.38800	0.56935	18.25	<.0001	865.86033
x4	1	-0.00978	0.00474	-2.06	0.0442	11.07651

圖 5: 向前加入法的模型選擇結果

向前加入法的最佳模型為: Y = X4

- 向前加入法的結果和向後去除法的結果相同