实验概述：

【实验目的及要求】

实验目的：通过实际案例分析，使学生掌握如何使用 SAS 的 ANOVA，REG,GLM 和 logistic 等过程进行编程，完成方差分析，多元回归分析及逻辑回归分析。

一、下表记录了某种钻头的寿命 y 与钻头轴向振动频率 F 及振幅 A 的观测数据。

F	A	*y*
359	5.24	161
359	1.76	129
141	5.24	166
141	1.76	135
375	3.5	187
125	3.5	170
250	5.5	174
250	1.5	146
250	3.5	203
250	3.5	185
250	3.5	230

请通过编写 SAS 程序，要求分别 REG 和 GLM 过程，

（1）建立 y 关于 F 和 A 的多元线性回归方程；

（2）建立 y 关于 F 和 A 的二次多项式回归方程；

（3）指出每一个回归方程对应的复相关系数 R、标准差 σ 的估计值，不显著的项。

（4）在标准差 σ 的估计值为最小的准则下，哪个方程最好？

二、为了解大学校园附近的餐馆的月营业收入（千元）和学生人数（千人）的关系，$n = 10$ 个大学的记录数据如下

X: 学生人数	2	6	8	8	12	16	20	20	22	26
Y: 月营业收入	58	105	88	118	117	137	157	169	149	202

编写 SAS 代码并进行如下分析：（检验水平$\alpha = 0.05$）

（1） 月营业收入和学生人数之间的相关系数为多少？

（2） 将月营业收入作为因变量，学生人数作为自变量进行一元线性回归，问模型整体是否显著？模型整体的解释能力如何度量？

（3） 检验回归方程的截距项和斜率是否显著？求出回归方程。

（4） 假设某大学有 15（千人）个学生，那么该大学校园附近餐馆的月营业收入的预测值为多少？置信度 95%的预测区间是多少？

三、为了让探讨冠心病发生的有关危险因素，对 26 例冠心病病人和 28 例对照着进行病例对照研究（数据见 experiment3.txt），各因素的说明如下所示：

x1：年龄（岁）

x2: 高血压史（无=0）

x3: 高血压家族史（无=0）

x4: 吸烟（无=0）

x5: 高血脂史（无=0）

x6: 动物脂肪摄入（低=0，高=1）

x7: 体重指数 BMI（<24=1, 24-25=2, >25=3）

x8：A 型性格（否=0，是=1）

y： 冠心病（对照=0， 病例=1）

本例研究目的是找出与冠心病有关的影响因素及其影响作用的大小。x1-x8 是可能与冠心病有关的影响因素，对这些因素进行筛选，挑出与冠心病有关影响的因素，再分析这些因素对冠心病的影响成的大小。要求：

（1）使用逐步筛选法筛选自变量

（2）控制进入模型和留在模型中的显著性水平均为 0.1

【实验原理】

一&二、可先使用 SAS 的数据步建立数据集，然后使用 SAS 的 REG 和 GLM 过程建立相应的回归方程。但在使用 REG 和 GLM 建立多项式回归方程时，要注意其实现方法的区别。

三、可先使用 SAS 的数据步建立数据集，然后使用 SAS 的 Logisitic 过程建立相应的逻辑回归方程，分析影响因素的作业。

【实验环境】（使用的软硬件）

硬件： IBM PC 或其兼容机

软件： Microsoft Windows, Microsoft Word 2003 或更高版本, SAS 8.x 或更高版本.

实验内容：

【实验方案设计】

实验 1

（1）建立 y 关于 F 和 A 的多元线性回归方程；

```
data raw;
input frequency amplitude year @@;
datalines;
359 5.24 161 359 1.76 129 141 5.24 166 141 1.76 135
```

```
375 3.5 187 125 3.5 170 250 5.5 174 250 1.5 146
250 3.5 203 250 3.5 185 250 3.5 230
;
run;
data mod_raw;
set raw;
freq_sq=frequency*frequency;
amp_sq=amplitude*amplitude;
proc reg data=raw;
model year=frequency amplitude;
run;
proc glm data=raw;
model year=frequency amplitude;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: year

Number of Observations Read	11
Number of Observations Used	11

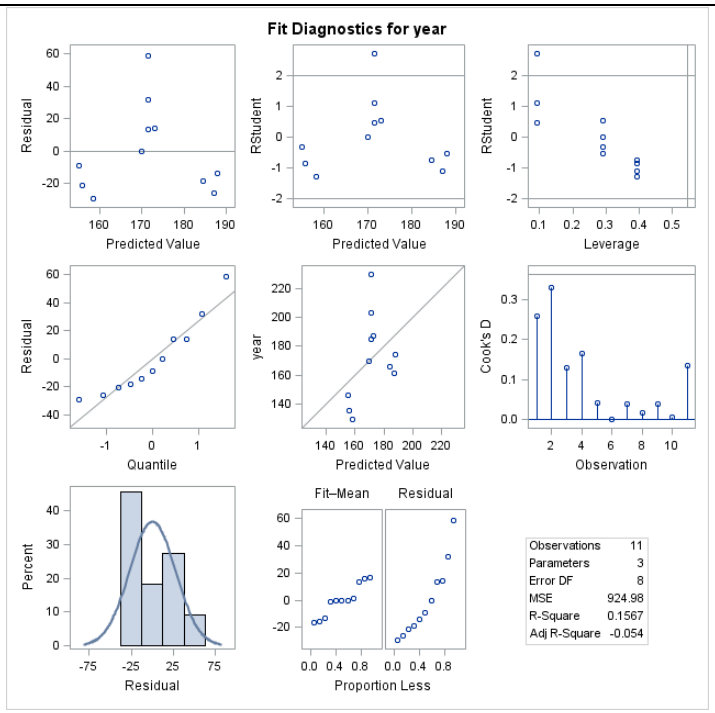
Analysis of Variance

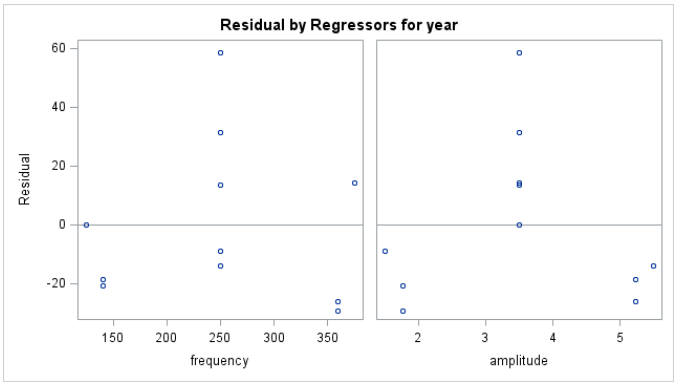
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1374.85537	687.42769	0.74	0.5058
Error	8	7399.87190	924.98399		
Corrected Total	10	8774.72727			

Root MSE	30.41355	R-Square	0.1567
Dependent Mean	171.45455	Adj R-Sq	-0.0541
Coeff Var	17.73855		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	139.69137	37.16749	3.76	0.0056
frequency	1	0.01176	0.10836	0.11	0.9163
amplitude	1	8.23554	6.78198	1.21	0.2592


Fit Diagnostics for year
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
year
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 11
Parameters 3
Error DF 8
MSE 924.98
R-Square 0.1567
Adj R-Square -0.054


Residual by Regressors for year
Residual
frequency
amplitude

The GLM Procedure

Dependent Variable: year

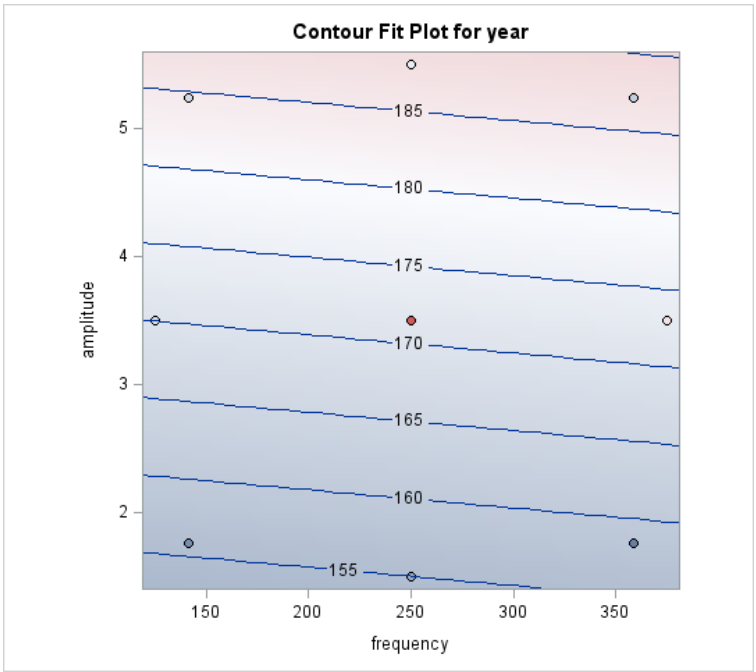
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1374.855372	687.427686	0.74	0.5058
Error	8	7399.871901	924.983988		
Corrected Total	10	8774.727273			

R-Square	Coeff Var	Root MSE	year Mean
0.156684	17.73855	30.41355	171.4545

Source	DF	Type I SS	Mean Square	F Value	Pr > F
frequency	1	10.885267	10.885267	0.01	0.9163
amplitude	1	1363.970105	1363.970105	1.47	0.2592

Source	DF	Type III SS	Mean Square	F Value	Pr > F
frequency	1	10.885267	10.885267	0.01	0.9163
amplitude	1	1363.970105	1363.970105	1.47	0.2592

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	139.6913692	37.16749329	3.76	0.0056
frequency	0.0117551	0.10836166	0.11	0.9163
amplitude	8.2355398	6.78198388	1.21	0.2592



$$year = 139.6913692 + 0.0117551 frequency + 8.2355398 amplitude$$

（2）建立 y 关于 F 和 A 的二次多项式回归方程；

```
data mod_raw;
set raw;
freq_sq=frequency*frequency;
amp_sq=amplitude*amplitude;
proc reg data=mod_raw;
title "Model2";
```

```
model year=freq_sq amp_sq frequency amplitude;
run;
proc glm data=mod_raw;
title "Model2";
model year=freq_sq amp_sq frequency amplitude;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: year

Number of Observations Read	11
Number of Observations Used	11

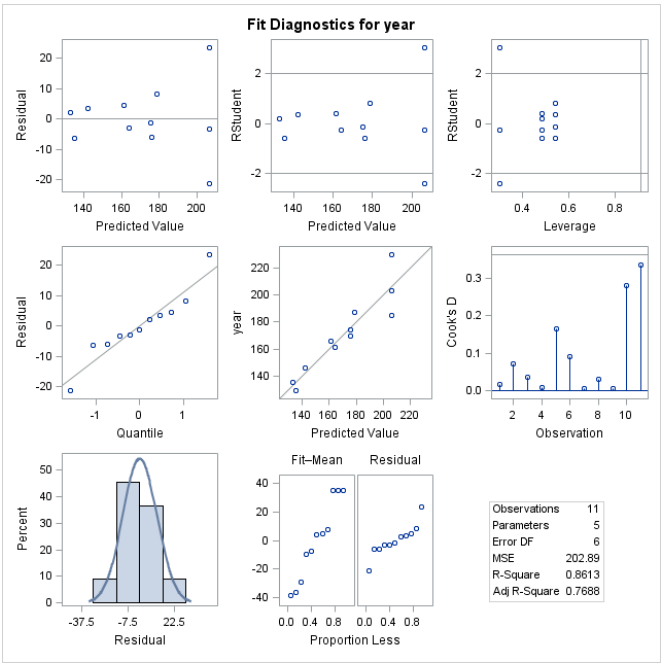
Analysis of Variance

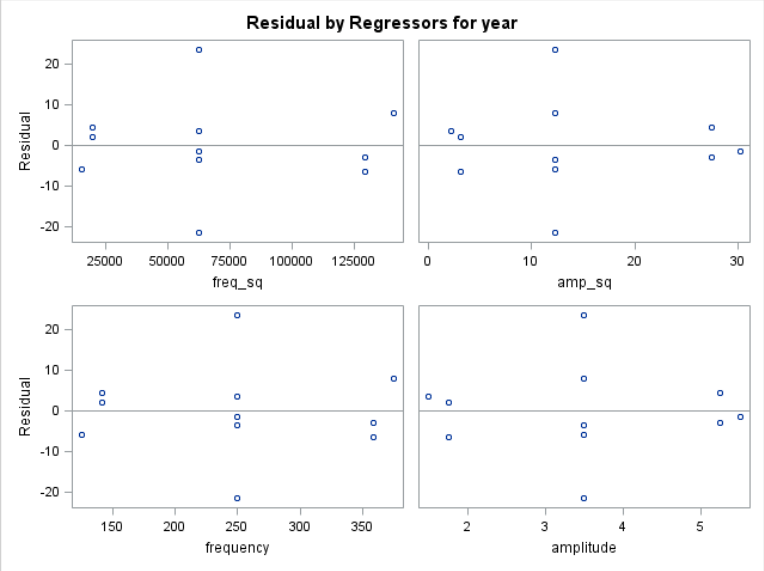
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7557.37906	1889.34477	9.31	0.0096
Error	6	1217.34821	202.89137		
Corrected Total	10	8774.72727			

Root MSE	14.24399	R-Square	0.8613
Dependent Mean	171.45455	Adj R-Sq	0.7688
Coeff Var	8.30774		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-86.90879	47.61277	-1.83	0.1177
freq_sq	1	-0.00186	0.00064427	-2.88	0.0279
amp_sq	1	-11.87971	2.52198	-4.71	0.0033
frequency	1	0.94051	0.32611	2.88	0.0279
amplitude	1	91.39351	17.93733	5.10	0.0022





Model2

The GLM Procedure

Dependent Variable: year

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7557.379061	1889.344765	9.31	0.0096
Error	6	1217.348212	202.891369		
Corrected Total	10	8774.727273			

R-Square	Coeff Var	Root MSE	year Mean
0.861267	8.307738	14.24399	171.4545

Source	DF	Type I SS	Mean Square	F Value	Pr > F
freq_sq	1	9.739823	9.739823	0.05	0.8338
amp_sq	1	599.013712	599.013712	2.95	0.1366
frequency	1	1681.440455	1681.440455	8.29	0.0281
amplitude	1	5267.185070	5267.185070	25.96	0.0022

Source	DF	Type III SS	Mean Square	F Value	Pr > F
freq_sq	1	1686.503404	1686.503404	8.31	0.0279
amp_sq	1	4501.856448	4501.856448	22.19	0.0033
frequency	1	1687.579665	1687.579665	8.32	0.0279
amplitude	1	5267.185070	5267.185070	25.96	0.0022

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-86.90879193	47.61276728	-1.83	0.1177
freq_sq	-0.00185751	0.00064427	-2.88	0.0279
amp_sq	-11.87971008	2.52198098	-4.71	0.0033
frequency	0.94051173	0.32610981	2.88	0.0279
amplitude	91.39351039	17.93733237	5.10	0.0022

$$year = -86.90879193 - 0.00185751frequency^2 - 11.87971008amplitude^2 + 0.94051173frequency + 91.39351039amplitude$$

（3）指出每一个回归方程对应的复相关系数 R、标准差 σ 的估计值，不显著的项。

多元线性回归方程中：

$$R^2 = 0.156684$$

$$\hat{\sigma} = 30.41355$$

Frequency, amplitude 在 5%水平下均为不显著的。

二次回归方程中：

$$R^2 = 0.861267$$

$$\hat{\sigma} = 14.24399$$

各项在 5%水平下均为显著的

（4）在标准差 σ 的估计值为最小的准则下，哪个方程最好？

二次回归方程的MSE更小（且其R^2更接近 1），因此二次方程拟合效果更好。

实验 2

（1）月营业收入和学生人数之间的相关系数为多少？

```
data campus;
input num income @@;
datalines;
2 58 6 105 8 88 8 118 12 117 16 137
20 157 20 169 22 149 26 202
;
run;
proc corr data=campus PEARSON KENDALL SPEARMAN;
title "correlation";
var num income;
run;
```

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
num	10	14.00000	7.94425	14.00000	2.00000	26.00000
income	10	130.00000	41.80643	127.50000	58.00000	202.00000

Pearson Correlation Coefficients, N = 10
Prob > |r| under H0: Rho=0

	num	income
num	1.00000	0.95012 <.0001
income	0.95012 <.0001	1.00000

Spearman Correlation Coefficients, N = 10
Prob > |r| under H0: Rho=0

	num	income
num	1.00000	0.92075 0.0002
income	0.92075 0.0002	1.00000

Kendall Tau b Correlation Coefficients, N = 10
Prob > |tau| under H0: Tau=0

	num	income
num	1.00000	0.79566 0.0016
income	0.79566 0.0016	1.00000

Pearson 相关系数为 0.95012,Spearman 与 Kendall Tau 相关系数如上图所示。

（2）将月营业收入作为因变量，学生人数作为自变量进行一元线性回归，问模型整体是否显著？模型整体的解释能力如何度量？

```
proc glm data=campus;
title "linear regression";
model income=num;
run;
```

linear regression

The GLM Procedure

Dependent Variable: income

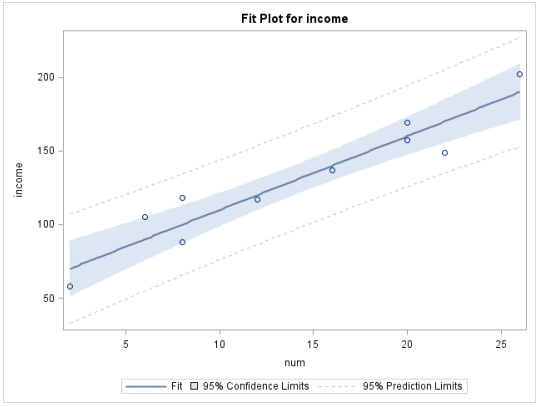
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	14200.00000	14200.00000	74.25	<.0001
Error	8	1530.00000	191.25000		
Corrected Total	9	15730.00000			

R-Square	Coeff Var	Root MSE	income Mean
0.902734	10.63794	13.82932	130.0000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
num	1	14200.00000	14200.00000	74.25	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
num	1	14200.00000	14200.00000	74.25	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	60.00000000	9.22603481	6.50	0.0002
num	5.00000000	0.58026524	8.62	<.0001



模型的F统计量的$p_{value} < 0.05$，因此模型通过检验，是显著的；

模型的解释性由复相关系数解释，$R^2 = 0.902734$，说明其中约90.2734%的信息是由模型解释的。

（3）检验回归方程的截距项和斜率是否显著？求出回归方程。

由各自$p_{value} < 0.05$可知，截距项和斜率项都是显著的，回归方程为：

$$income = 60 + 5num$$

（4）假设某大学有 15（千人）个学生，那么该大学校园附近餐馆的月营业收入的预测值为多少？置信度 95%的预测区间是多少？

```
data pred;
income=.;
num=15;
run;
```

```
data pred_data;
set pred campus;
run;
proc print data=pred_data;
run;
proc reg data=pred_data;
model income=num/p cli;
run;
```

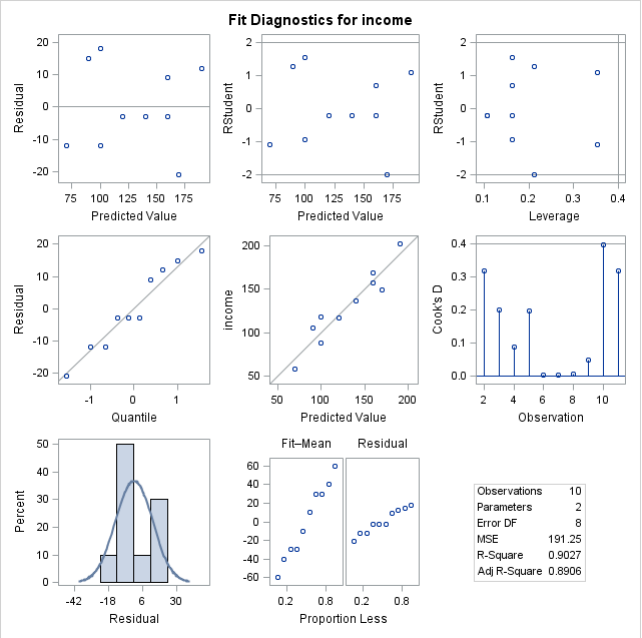
linear regression

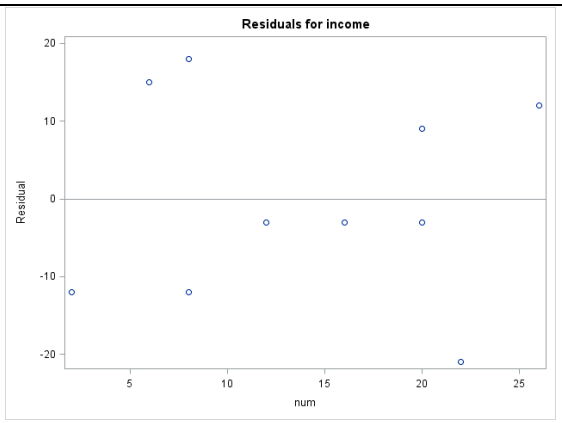
The REG Procedure
Model: MODEL1
Dependent Variable: income

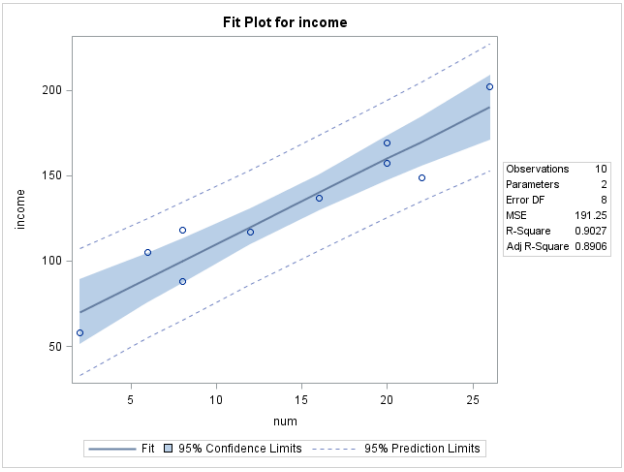
Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	.	135.0000	4.4115	101.5262	168.4738	.
2	58	70.0000	8.2226	32.8983	107.1017	-12.0000
3	105	90.0000	6.3776	54.8817	125.1183	15.0000
4	88	100.0000	5.5899	65.6029	134.3971	-12.0000
5	118	100.0000	5.5899	65.6029	134.3971	18.0000
6	117	120.0000	4.5246	86.4461	153.5539	-3.0000
7	137	140.0000	4.5246	106.4461	173.5539	-3.0000
8	157	160.0000	5.5899	125.6029	194.3971	-3.0000
9	169	160.0000	5.5899	125.6029	194.3971	9.0000
10	149	170.0000	6.3776	134.8817	205.1183	-21.0000
11	202	190.0000	8.2226	152.8983	227.1017	12.0000

Sum of Residuals	0
Sum of Squared Residuals	1530.00000
Predicted Residual SS (PRESS)	2583.29951







预测值为135；95%的置信区间为(101.5262,168.4738)

实验 3

本例研究目的是找出与冠心病有关的影响因素及其影响作用的大小。x1-x8 是可能与冠心病有关的影响因素，对这些因素进行筛选，挑出与冠心病有关影响的因素，再分析这些因素对冠心病的影响成的大小。要求：

（1）使用逐步筛选法筛选自变量

（2）控制进入模型和留在模型中的显著性水平均为 0.1

```
data raw;
infile "C:\Users\31949\iCloudDrive\By's Cloud\SUFE-lectures
\Assignment\hw5\experiment5.txt";
input id x1 x2 x3 x4 x5 x6 x7 x8 y;
run;
proc logistic data=raw descending;
model y=x1 x2 x3 x4 x5 x6 x7 x8 /
    selection=stepwise
    sle=0.1 sls=0.1;
run;
proc logistic data=raw descending;
model y=x1 x5 x6 x8 /
    sle=0.1 sls=0.1;
run;
```

model

The LOGISTIC Procedure

Model Information	
Data Set	WORK.RAW
Response Variable	y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	54
Number of Observations Used	54

Response Profile		
Ordered Value	y	Total Frequency
1	1	26
2	0	28

Probability modeled is y=1.

Stepwise Selection Procedure

Step 0. Intercept entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

-2 Log L	=	74.786

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
25.4181	8	0.0013

Step 1. Effect x6 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	76.786	67.467
SC	78.775	71.445
-2 Log L	74.786	63.467

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.3186	1	0.0008
Score	10.1174	1	0.0015
Wald	6.6570	1	0.0099

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
18.0210	7	0.0119

Note: No effects for the model in Step 1 are removed.

Step 2. Effect x5 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	76.786	61.480
SC	78.775	67.447
-2 Log L	74.786	55.480

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	19.3055	2	<.0001
Score	16.4702	2	0.0003
Wald	12.2010	2	0.0022

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
12.6157	6	0.0496

Note: No effects for the model in Step 2 are removed.

Step 3. Effect x8 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	76.786	58.402
SC	78.775	66.358
-2 Log L	74.786	50.402

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	24.3835	3	<.0001
Score	20.3833	3	0.0001
Wald	13.8847	3	0.0031

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
7.9650	5	0.1582

Note: No effects for the model in Step 3 are removed.

Step 4. Effect x1 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	76.786	56.224
SC	78.775	66.169
-2 Log L	74.786	46.224

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.5613	4	<.0001
Score	23.1563	4	0.0001
Wald	14.2827	4	0.0064

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
3.9490	4	0.4129

Note: No effects for the model in Step 4 are removed.

Note: No (additional) effects met the 0.1 significance level for entry into the model.

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	x6		1	1	10.1174		0.0015
2	x5		1	2	7.8749		0.0050
3	x8		1	3	4.9956		0.0254
4	x1		1	4	4.1370		0.0420

model

The LOGISTIC Procedure

Model Information	
Data Set	WORK.RAW
Response Variable	y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	54
Number of Observations Used	54

Response Profile		
Ordered Value	y	Total Frequency
1	1	26
2	0	28

Probability modeled is y=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	76.786	56.224
SC	78.775	66.169
-2 Log L	74.786	46.224

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.5613	4	<.0001
Score	23.1563	4	0.0001
Wald	14.2827	4	0.0064

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.7050	1.5433	9.2950	0.0023
x1	1	0.9239	0.4766	3.7583	0.0525
x5	1	1.4959	0.7439	4.0440	0.0443
x6	1	3.1355	1.2489	6.3031	0.0121
x8	1	1.9471	0.8466	5.2893	0.0215

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
x1	2.519	0.990	6.411
x5	4.464	1.039	19.181
x6	23.000	1.989	265.945
x8	7.008	1.333	36.834

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	87.1	Somers' D	0.766
Percent Discordant	10.4	Gamma	0.786
Percent Tied	2.5	Tau-a	0.390
Pairs	728	c	0.883

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
x1	2.519	0.990	6.411
x5	4.464	1.039	19.181
x6	23.000	1.989	265.945
x8	7.008	1.333	36.834

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	87.1	Somers' D	0.766
Percent Discordant	10.4	Gamma	0.786
Percent Tied	2.5	Tau-a	0.390
Pairs	728	c	0.883

经过筛选，筛选后的自变量为 x1：年龄（岁），x5：高血脂史（无=0），x6：动物脂肪摄入（低=0，高=1），x8：A 型性格（否=0，是=1）。其影响为：年龄增长 1 单位，患病概率增加 2.519 倍；有高血脂史的比无高血脂史的患病概率高 4.464 倍；动物脂肪摄入高的比动物脂肪摄入低的高 23 倍；拥有 A 性格比不具有的高 7.008 倍。

【小结】

本次实验主要内容为 GLM，REG，CORR，LOGISTIC 四种统计 PROC 的使用。在实验过程中对于其具体的参数确定，实验结果的解读有了更深的了解。希望能在最后这段时间加强对于 SAS 的编程理解，更好的完成编程内容。

指导教师评语及成绩：

成绩： **指导教师签名：**

批阅日期：

附件：

实验报告说明

1. **实验项目名称：**要用最简练的语言反映实验的内容。

2. **实验类型：**一般需说明是验证型实验还是设计型实验、综合型实验或其他实验。

3. **实验目的与要求：**目的要明确，要抓住重点。

4. **实验原理：**简要说明本实验项目所涉及的理论知识。

5. **实验环境：**实验用的软硬件环境（配置）。

6. **实验方案设计（思路、步骤和方法等）：**这是实验报告极其重要的内容，概括整个实验过程。

对于**验证型实验**，要写明依据何种原理、何仲操作方法进行实验，并写明需要经过哪几个步骤。

对于**设计型和综合型实验**，在上述内容基础上还应该画出流程图、设计思路和设计方法，再配以相应的文字说明。

7．实验过程（实验中涉及的记录、数据、分析）：写明具体上述实验方案的具体实施，包括实验过程中的记录、数据和相应的分析。

8．结论（结果）：即根据实验过程中所见到的现象和测得的数据，做出结论。

9．小结：对本次实验的心得体会、思考和建议。

10．指导教师评语及成绩：指导教师依据学生的实际报告内容，用简练语言给出本次实验报告的评价和价值。