多元统计分析 作业 2 回归分析

地信一班 罗皓文 15303096

实验环境：

OS: Windows 7 Pro SP1 x64
CPU: Intel Xeon E3-1241 v3 @ 3.50GHz 3.50GHz
RAM: 8.00 Gb
Soft: SPSS Statistics 19

一、一元线性回归

为了检验美国电力行业是否存在规模经济，Nerlove (1963)收集了 1955 年 145 家美国电力企业的总成本(TC)、产量(Q)、工资率(PL)、燃料价格(PF)及资本租赁价格(PK)的数据资料所示。利用该数据(例 2.2.sav)，以产量为自变量，工资率为因变量，对工资率和产量做一元线性回归分析；

部分数据示例：

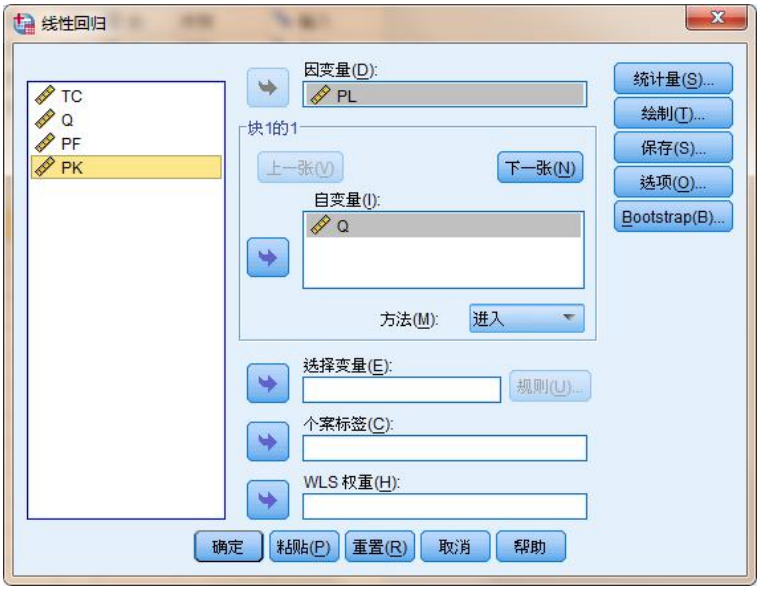
	TC	Q	PL	PF	PK
1	.082	2	2.1	17.9	183
2	.661	3	2.1	35.1	174
3	.990	4	2.1	35.1	171
4	.315	4	1.8	32.2	166
5	.197	5	2.1	28.6	233

数据变量视图：

	名称	类型	宽度	小数	标签	值	缺失	列	对齐	度量标准	角色
1	TC	数值(N)	11	3		无	无	11	右	未知	输入
2	Q	数值(N)	11	0		无	无	11	右	未知	输入
3	PL	数值(N)	11	1		无	无	11	右	未知	输入
4	PF	数值(N)	11	1		无	无	11	右	未知	输入
5	PK	数值(N)	11	0		无	无	11	右	未知	输入

（1）使用 SPSS 线性回归模块：分析(A) - 回归(R) - 线性(L)...

（2）选取自变量和因变量，使用进入回归，进行一元线性回归。设置参数如下：



（3）输出结果与分析：

表 1 Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.171[a]	.029	.023	.2341

a. Predictors: (Constant), Q

如表 1 模型汇总表中，R 和 R 方可以用来衡量模型对观测值的拟合程度。R 越接近 1 说明模型越好。调整 R 方比 R 方更准确一些，本例中的最终调整 R 方为 0.023，表示自变量一共可以解释因变量 2.3%的变化（variance）。本例 R=0.171，一般认为，拟合优度 R 达到 0.1 为小效应，0.3 为中等，0.5 为大效应。

表 2 ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.237	1	.237	4.331	.039[a]
	Residual	7.838	143	.055		
	Total	8.075	144			

a. Predictors: (Constant), Q

b. Dependent Variable: PL

如表 2，对模型进行方差分析，模型的 F 统计量为 4.331，显著性水平 p 值为 0.039，模型是显著的。从假设检验来说，F 值是对模型的显著性检验，表示的是模型中被解释变量与所有解释变量之间的线性关系在总体上是否显著做出推断。

表 3 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.943	.024		80.696	.000
	Q	1.385E-5	.000	.171	2.081	.039

a. Dependent Variable: PL

如表 3，Q 的系数很小（近似为 0），在本例中 Q 对 PL 显著不相关。

二、多元线性回归

利用给定数据(作业 multi.sav)，以 s80、expr80、tenure80 和 iq 为自变量，lw80 因变量做多元线性回归分析。

部分数据示例：

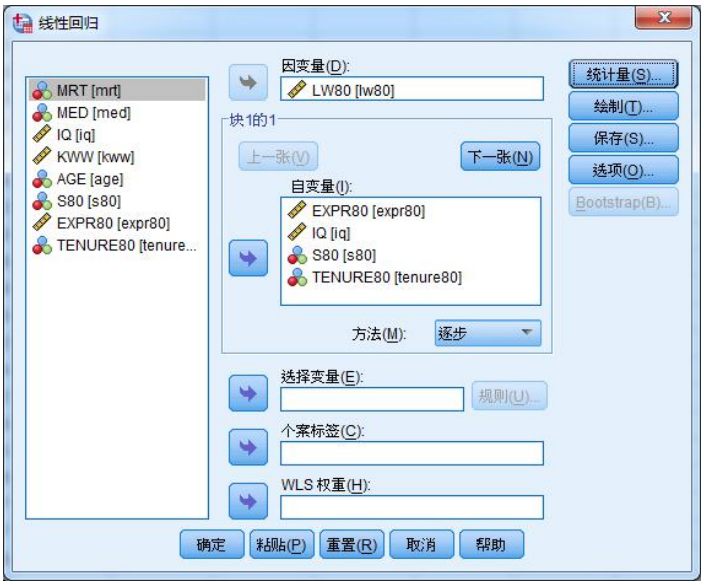
	mrt	med	iq	kww	age	s80	expr80	tenure80	lw80
1	0	8	93	35	19	12	10.64	2	6.64
2	0	14	119	41	23	18	11.37	16	6.69
3	0	14	108	46	20	14	11.03	9	6.72
4	0	12	96	32	18	12	13.09	7	6.48
5	1	6	74	27	26	11	14.40	5	6.33
6	0	8	91	24	16	10	13.43	0	6.40
7	1	8	114	50	30	18	7.55	14	6.99
8	1	14	111	37	23	15	12.62	1	7.05
9	1	12	95	44	22	12	15.87	16	6.91

数据变量视图：

	名称	类型	宽度	小数	标签	值	缺失	列	对齐	度量标准	角色
1	mrt	数值(N)	1	0	MRT	无	无	5	右	名义(N)	输入
2	med	数值(N)	2	0	MED	无	无	5	右	名义(N)	输入
3	iq	数值(N)	3	0	IQ	无	无	5	右	度量(S)	输入
4	kww	数值(N)	2	0	KWW	无	无	5	右	度量(S)	输入
5	age	数值(N)	2	0	AGE	无	无	5	右	名义(N)	输入
6	s80	数值(N)	2	0	S80	无	无	5	右	名义(N)	输入
7	expr80	数值(N)	5	2	EXPR80	无	无	8	右	度量(S)	输入
8	tenure80	数值(N)	2	0	TENURE80	无	无	6	右	名义(N)	输入
9	lw80	数值(N)	4	2	LW80	无	无	6	右	度量(S)	输入

（1）使用 SPSS 线性回归模块：分析(A) - 回归(R) - 线性(L)...

（2）选取自变量和因变量，使用逐步回归，设置参数如下：



（4）输出结果与分析：

表1 模型汇总[e]

模型	R	R 方	调整 R 方	标准估计的误差	更改统计量				
					R 方更改	F 更改	df1	df2	Sig. F 更改
1	.309[a]	.095	.094	.39012	.095	79.798	1	756	.000
2	.376[b]	.141	.139	.38042	.046	40.057	1	755	.000
3	.394[c]	.155	.152	.37755	.014	12.513	1	754	.000
4	.402[d]	.162	.157	.37632	.007	5.952	1	753	.015

a. 预测变量: (常量), S80。

b. 预测变量: (常量), S80, EXPR80。

c. 预测变量: (常量), S80, EXPR80, IQ。

d. 预测变量: (常量), S80, EXPR80, IQ, TENURE80。

e. 因变量: LW80

表2 输入／移去的变量[a]

模型	输入变量	移去变量	方法
1	S80	.	步进（准则: F-to-enter 概率<= .050，F-to-remove 概率>= .100）。
2	EXPR80	.	步进（准则: F-to-enter 概率<= .050，F-to-remove 概率>= .100）。
3	IQ	.	步进（准则: F-to-enter 概率<= .050，F-to-remove 概率>= .100）。
4	TENURE80	.	步进（准则: F-to-enter 概率<= .050，F-to-remove 概率>= .100）。

a. 因变量: LW80

如表 1 模型汇总表中，R 和 R 方可以用来衡量模型对观测值的拟合程度。R 越接近 1 说明模型越好。调整 R 方比 R 方更准确一些，本例中的最终调整 R 方为 0.157，表示自变量一共可以解释因变量 15.7%的变化（variance）。

另外，由于使用的是 StepWise Linear Regression (SWLR)，分析——回归——线性——“方法”选择“逐步”，变量的输入、移除情况如表 2 所示。由表 1 所示，模型 1、2、3、4 的 R 方逐渐增大，标准误差逐渐减小。一般认为，拟合优度 R 达到 0.1 为小效应，0.3 为中等，0.5 为大效应。经过比较模型 4 的拟合优度最好，因此采用模型 4。

Anova[e]

模型		平方和	df	均方	F	Sig.
4	回归	20.569	4	5.142	36.310	.000[d]
	残差	106.638	753	.142		
	总计	127.206	757			

d. 预测变量: (常量), S80, EXPR80, IQ, TENURE80。

e. 因变量: LW80

如表 3，对模型进行方差分析，模型的 F 统计量为 36.310，显著性水平 p 值为 0.000，模型是显著的。从假设检验来说，F 值是对模型的显著性检验，表示的是模型中被解释变量与所有解释变量之间的线性关系在总体上是否显著做出推断。

本例自由度 k=4，n=758，查表得到在显著性 $\alpha = 0.005$ 时，F(4,753)=3.747，F>>F(4,753)，故认为模型线性关系显著。

表4 系数[a]

模型	非标准化系数		标准系数	t	Sig.	相关性			共线性统计量	
	B	标准误差	试用版			零阶	偏	部分	容差	VIF
4 (常量)	5.269	.139		37.870	.000					
S80	.062	.008	.337	7.952	.000	.309	.278	.265	.620	1.613
EXPR80	.020	.004	.209	5.345	.000	.054	.191	.178	.727	1.376
IQ	.004	.001	.134	3.448	.001	.260	.125	.115	.734	1.361
TENURE80	.007	.003	.086	2.440	.015	.133	.089	.081	.903	1.107

回归模型为：

$$y = 0.062x_1 + 0.020x_2 + 0.004x_3 + 0.007x_4 + 5.269,$$

其中 y 为因变量 lw80，x_1, x_2, x_3, x_4 分别为自变量 s80，expr80，tenure80，iq。

三、二项分类 Logistic 回归

给定数据为 20 名前列腺癌患者的相关数据。试用二项分类 Logistic 回归方法分析患者前列腺细胞癌转移情况（有转移 y=1、无转移 y=0）与患者年龄、前列腺细胞癌血管内皮生长因子（其阳性表述由低到高共 3 个等级）、术前探针活检病理分级（从低到高共 4 级）、酸性磷酸酯酶、前列腺细胞癌分期之间的关系。

部分数据示例：

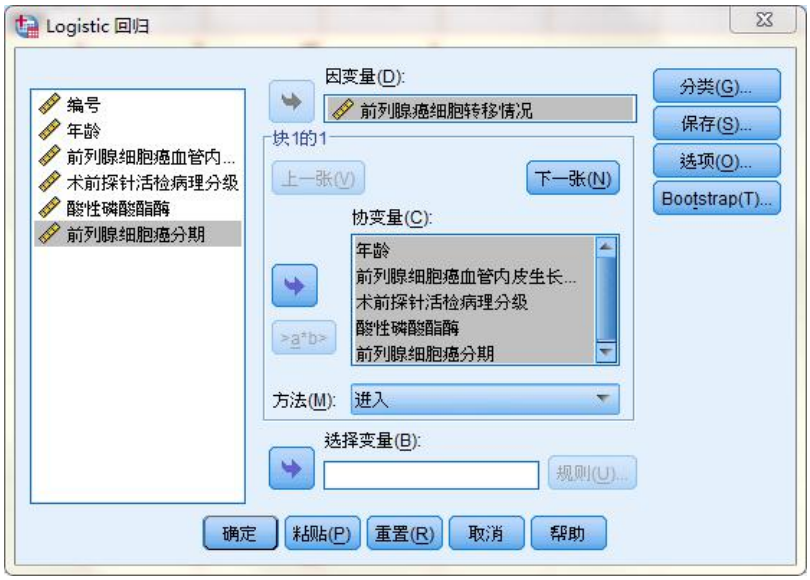
	编号	前列腺癌细胞转移情况	年龄	前列腺细胞癌血管内皮生长因子	术前探针活检病理分级	酸性磷酸酯酶	前列腺细胞癌分期
1	1	0	66	3	3	46	1
2	2	1	45	2	2	60	2
3	3	1	79	1	1	50	3
4	4	0	65	2	3	50	2
5	5	0	55	3	4	60	3
6	6	0	58	3	3	43	2
7	7	1	43	1	2	70	1

数据变量视图：

	名称	类型	宽度	小数	标签	值	缺失	列	对齐	度量标准	角色
1	编号	数值(N)	8	0		无	无	8	右	度量(S)	输入
2	前列腺癌细...	数值(N)	8	0		无	无	8	右	度量(S)	输入
3	年龄	数值(N)	8	0		无	无	8	右	度量(S)	输入
4	前列腺细胞...	数值(N)	8	0		无	无	8	右	度量(S)	输入
5	术前探针活...	数值(N)	8	0		无	无	8	右	度量(S)	输入
6	酸性磷酸酯酶	数值(N)	8	0		无	无	8	右	度量(S)	输入
7	前列腺细胞...	数值(N)	8	0		无	无	8	右	度量(S)	输入

（1）使用 SPSS 的 logistics 回归模块，分析(A) - 回归(R) - 二元 logistics...

（2）选取自变量和因变量，使用逐步回归，设置参数如下：



（3）输出结果与分析

如表 5 所示，总体分类精度为 80%，达到较好的分类效果。

表 5 Classification Table[a]

Observed			Predicted		
			前列腺癌细胞转移情况		
			0	1	Percentage Correct
Step 1	前列腺癌细胞转移情况	0	10	2	83.3
		1	2	6	75.0
	Overall Percentage				80.0

a. The cut value is .500

表 5 Classification Table[a]

Observed			Predicted		
			前列腺癌细胞转移情况		
			0	1	Percentage Correct
Step 1	前列腺癌细胞转移情况	0	10	2	83.3
		1	2	6	75.0
	Overall Percentage				80.0

表 6 Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	年龄	-.064	.058	1.220	1	.269	.938
	前列腺细胞癌血管内皮生长因子	-.431	.745	.335	1	.563	.650
	术前探针活检病理分级	-1.506	.892	2.854	1	.091	.222
	酸性磷酸酯酶	-.007	.069	.010	1	.921	.993
	前列腺细胞癌分期	2.057	1.072	3.681	1	.055	7.819
	Constant	3.224	6.396	.254	1	.614	25.140

a. Variable(s) entered on step 1: 年龄，前列腺细胞癌血管内皮生长因子，术前探针活检病理分级，酸性磷酸酯酶，前列腺细胞癌分期.

如表 6 所示，最终模型为：

$$y=\frac{1}{1+\exp(0.064x_1+0.431x_2+1.506x_3+0.007x_4-2.057x_5)-3.224)}$$

其中，因变量 y 为前列腺细胞癌转移情况，自变量 x_1,x_2,x_3,x_4,x_5 分别为年龄、前列腺细胞癌血管内皮生长因子、术前探针活检病理分级、酸性磷酸酯酶和前列腺细胞癌分期。