第 4 小组　第 2 次作业

承子杰、张越、樊昊方、张涵、蔡若瑶、韩璐瑶、李昕燃
School of Mathematics Science , Soochow University

更新：*November 11, 2020*

摘　要

某医院外科为了预测做过某种肝手术患者的术后生存时间 Y，经过长期追踪研究，发现与三项指标:X_1：预后指数；X_2：酶化验值；X_3：肝功化验值存在线性关系。

经过长期追踪调查，现有 54 名患者的各项指标与术后生存时间的数据，我们可以建立多元线性回归模型，对所得模型的显著性分别进行 t 检验、方差检验和相关系数检验。

进一步，我们通过 *Spearman* 检验对回归方程做异方差性分析，并对可能存在的异方差性，我们对因变量做 $BOX-COX$ 变换。

我们通过 DW 检验对回归方程做自相关性分析，并对可能存在的自相关性，我们进行一阶方差变换。

最后，我们得到多元回归模型 $y=\exp(0.019712X_1+0.016773X_2+0.17605X_3+2.071719)$，可以通过该方程进行初步术后生存时间的预测。

关键词： *Spearman* 检验、$BOX-COX$ 变换、DW 检验

目录

1 问题背景

某医院外科为了预测做过某种肝手术患者的术后生存时间，随机选取了 54 位需要做此种手术的患者，手术前对每位患者化验和评估了如下三个指标：

X_1 : 预后指数；X_2 : 酶化验值；X_3 : 肝功化验值

术后随访得到各患者的生存时间 Y。各变量的观测值见附录 1(数据摘自 *Neter* 等，1990)

经过初步研究后，现已发现生存时间 Y 与各项指标之间存在一定的线性关系。

2 变量申明

2.1 符号说明

表 1: 变量申明

变量	含义
y	生存时间
X_1	预后指数
X_2	酶化验值
X_3	肝功化验值
n	患者编号

2.2 名词解释

1. 预后指数：在医学上，“预后指数”是指根据经验预测的疾病发展情况评估指数。

2. 酶化验值：肝功能检查中一些相关酶评估指数。

3. 肝功化验值：肝功能化验是了解肝脏健康状况的主要检查，主要目的是通过检查肝功能的具体指标得到定位、定性、定量的反应，有助于各种肝病的诊断，了解肝病的程度、转归及预后。

3 问题求解

3.1 建立多元线性回归模型

首先，我们采用普通最小二乘建立 y 关于 X_1、X_2、X_3 的回归方程，并对得到的回归方程进行假设检验。*R* 语言代码如下：

```
rm(list=ls());
data<-read.table("C:/data/data1.txt",head=TRUE);
head(data);
```

```r
y=data$y;
x1=data$x1;
x2=data$x2;
x3=data$x3
fit<-lm(y~x1+x2+x3+1);
summary(fit);
```

经过计算后得到的线性回归方程为：

$$Y = 4.1067X_1 + 3.1840X_2 + 45.5343X_3 - 429.1560$$

对各参数做 t 检验，检验结果可列为下表：

表 2: t 检验表

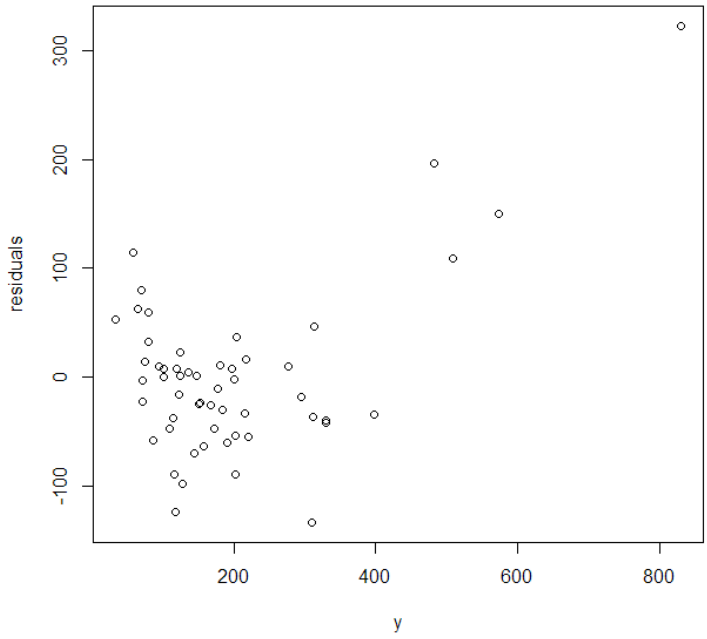
变量	t 值	P 值
x_1	6.056	1.79e-07
x_2	5.597	9.19e-07
x_3	3.422	0.00125

通过该表结果，我们可以发现各变量回归均显著。
下对所得的回归方程做方差分析，所得的方差分析表如下:

表 3: ANOVA

方差来源	自由度	平方和 (SS)	均方 (MS)	F 值	P 值
回归 (R)	3	931546	310515	82.85	<0.0001
误差 (E)	50	187385	3747.70927		
总和 (T)	53	1118932			

根据方差分析表，我们可以发现回归方程显著。
再计算回归方程的相关系数，$r^2 = 0.7238$，较为接近 1，回归效果良好。
我们进一步对其异方差性进行分析，首先先画出关于 y 的残差散点图如下：



通过该图，我们不难发现，残差散点具有明显的趋势，初步可以判断，回归方程存在异方差性。

3.2 异方差性检验与处理

为了进一步确认其异方差性，我们分别对各变量进行 *Spearman* 检验，检验结果如下：

表 4: *t* 检验表

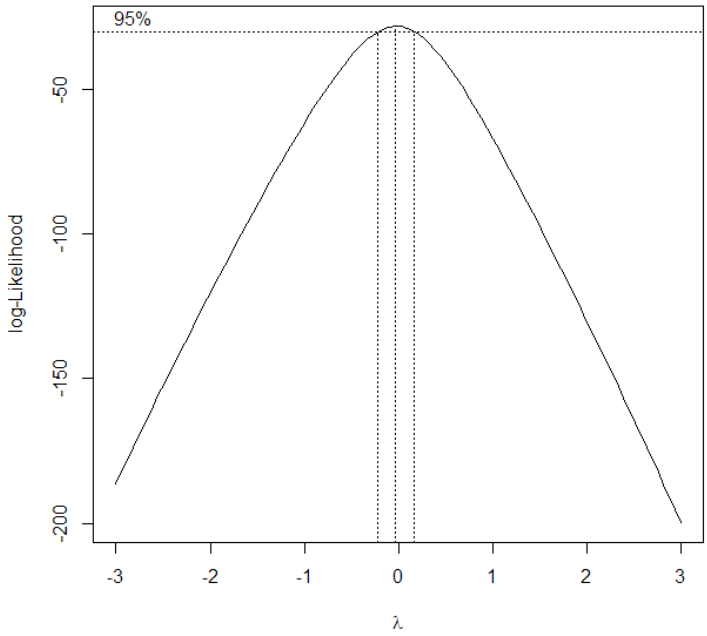
变量	*P* 值	等级相关系数
x_1	0.113	0.218141
x_2	0.1516	0.1978035
x_3	0.008586	0.3542509

从 *Spearman* 检验，我们可以发现回归方程确实存在异方差性。

为了消除回归方程的异方差性，下面我们采用 $BOX-COX$ 变换。我们取 λ 的值从 $[-3.3]$ 以步长为 0.1，分别计算对应 λ 下的对数极大似然估计，并取出对数似然估计值取值最大的 λ，对 y 做变换后重新进行线性回归，并作各项检验。*R* 语言代码如下：

```
library(MASS)
lambda_mle<-boxcox(y~x1+x2+x3+1,data=data,lambda=seq(-3,3,0.1))
```

我们得到的 MLE 与对应 λ 的图像如下：



为了便于计算，我们在此处取 $\lambda = 0$。在 $\lambda = 0$ 下，我们得到的回归方程为：

$$\ln y = 0.019712X_1 + 0.016773X_2 + 0.17605 + 2.071719$$

对各参数做 t 检验，检验结果可列为下表：

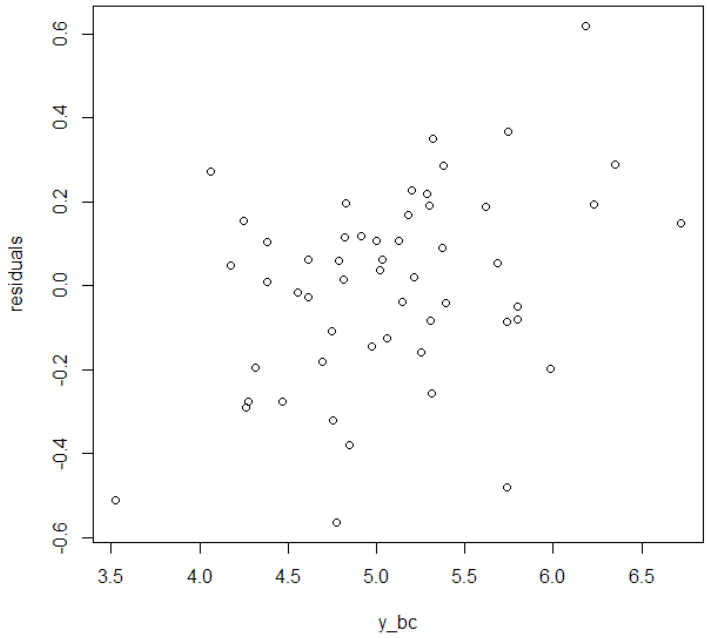
表 5: t 检验表

变量	t 值	P 值
x_1	9.559	7.18e-13
x_2	9.696	4.50e-13
x_3	4.351	6.69e-05

通过该表结果，我们可以发现各变量回归均显著。

再计算回归方程的相关系数，$r^2 = 0.8643$，较为接近 1，回归效果良好。

我们进一步对其异方差性进行分析，首先先画出关于 y 的残差散点图如下：



通过该图，我们不难发现，异方差性基本已消除。

最后我们在做 *Spearman* 检验，检验结果如下：

表 6: *t* 检验表

变量	*P* 值	等级相关系数
x_1	0.2661	0.3017127
x_2	0.8128	-0.03298631
x_3	0.731	-0.0478841

从 *Spearman* 检验，我们可以发现回归方程确实消除了异方差性。

3.3 自相关性检验与处理

对得到的回归方程做 *DW* 检验，进行自相关性进行分析，*R* 语言代码如下：

```
library(lmtest);
dwtest(fit2)
```

得到的结果为：*DW* 值为 2.0693，*P* 值为 0.6017，故回归方程不存在自相关性。

3.4 总结

经过初步回归，我们得到线性回归方程：

$$Y = 4.1067X_1 + 3.1840X_2 + 45.5343X_3 - 429.1560$$

进行显著性检验，发现回归方程显著，进一步，经过 $Spearman$ 检验，发现存在异方差性。我们采取 $BOX - COX$ 变换，λ 以 0.1 步长从 $[-3, 3]$ 取值，分别计算各 λ 下的对数似然值，取最大的对数似然值下的 λ 值为 $BOX - COX$ 变换中的 λ 值，对因变量进行变换后，重新进行线性回归，得回归方程：

$$y = \exp(0.019712X_1 + 0.016773X_2 + 0.17605X_3 + 2.071719)$$

重新进行显著性检验、$Spearman$ 检验和 DW 检验，发现方程显著，且不存在异方差性和自相关性。从而基本可以确立，回归模型建立完成，可使用该回归方程进行初步预测。

4 参考文献

参考文献

[1] 梅长林, 王宁. 近代回归分析方法 [M]. 科学出版社, 2012.

[2] 何晓群, 刘文卿. 应用回归分析 [M]. 中国人民大学出版社, 2001.

A 附录

文章所用的具体数据列为下表：

n	x1	x2	x3	y
1	62	81	2.59	200
2	59	66	1.70	101
3	57	83	2.16	204
4	73	41	2.01	101
5	65	115	4.30	509
6	38	72	1.42	80
7	46	63	1.91	80
8	68	81	2.57	127
9	67	93	2.50	202
10	76	94	2.40	203
11	84	83	4.13	329
12	51	43	1.86	65
13	96	114	3.95	830
14	83	88	3.95	330
15	62	67	3.40	168
16	74	68	2.40	217
17	85	28	2.98	87
18	51	41	1.55	34
19	68	74	3.56	215
20	57	87	3.02	172
21	52	76	2.85	109
22	83	53	1.12	136
23	26	68	2.10	70
24	67	86	3.40	220
25	59	100	2.95	276
26	61	73	3.50	144
27	52	86	2.45	181
28	76	90	5.59	574
29	54	56	2.71	72
30	76	59	2.58	178

31	64	65	0.74	71
32	45	23	2.52	58
33	59	73	3.50	116
34	72	93	3.30	295
35	58	70	2.64	115
36	51	99	2.60	184
37	74	86	2.05	118
38	8	119	2.85	120
39	61	76	2.45	151
40	52	88	1.81	148
41	49	72	1.84	95
42	28	99	1.30	75
43	86	88	1.81	483
44	56	77	2.85	153
45	77	93	1.48	191
46	40	84	3.00	123
47	73	106	3.05	311
48	86	101	4.10	398
49	67	77	2.86	158
50	82	103	4.55	310
51	77	46	1.95	124
52	85	40	1.21	125
53	59	85	2.33	198
54	78	72	3.20	313