第五小组应用回归分析实验报告

组员：陈霄　王一龙　曹米纳　顾辰菲　杨思哲　陈佳宁　叶子文　张晔

摘要

本报告利用多元线性回归模型对申请大学成功率的数据集进行分析，通过模型拟合可以根据申请人的各项水平，对其申请大学的成功率进行预测。首先初步对全体数据进行多元线性的最小二乘回归，剔除不显著变量得到回归方程为 $\hat{y} = -1.298 + 0.0018x_1 + 0.0030x_2 + 0.0228x_5 + 0.1210x_6 + 0.0246x_7$, 接着检验模型是否满足建立时对其所做的基本假设 (G-M 条件)，即残差项是否存在异方差与自相关。消除异方差与自相关后，再次检验得到的新拟合方程的显著性及变量显著性，得到最终拟合结果为:

$$\frac{\widehat{y}_{t+1}^{2.27}-1}{2.27} = 0.6215\frac{y_t^{2.27}-1}{2.27} + (-0.255, 0.001, 0.002, 0.009, 0.052, 0.014)\begin{pmatrix} 1 \\ x_{1,t+1} - 0.6215x_{1t} \\ x_{2,t+1} - 0.6215x_{2t} \\ x_{5,t+1} - 0.6215x_{5t} \\ x_{6,t+1} - 0.6215x_{6t} \\ x_{7,t+1} - 0.6215x_{7t} \end{pmatrix}$$

关键字: 多元线性回归，最小二乘，G-M 条件，Box-Cox 变换，DW 检验

目录

1 建立线性回归模型

对申请大学成功率的数据集进行分析。原数据集共包含 9 个数据列，其中第一列为学生编号，可剔除考虑; 2-8 列分别代表七个解释变量，其含义依次为：GRE 成绩、TOEFL 成绩、申请大学排名、个人陈述、推荐信、平均绩点以及研究经历次数; 第 9 列为因变量，意义为申请大学的成功概率。

首先读取文件中的全体数据，将因变量对所有自变量进行多元线性回归建模，并对回归结果进行初步检验，即检验回归方程的显著性、解释变量的显著性。

```
a<-read.table("D:/R/data/Admission_Predict.csv",sep=",",header=T)
y<-a[,9]
x1<-a[,2]
x2<-a[,3]
x3<-a[,4]
x4<-a[,5]
x5<-a[,6]
x6<-a[,7]
x7<-a[,8]
z<-lm(y~x1+x2+x3+x4+x5+x6+x7)
summary(z)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)

Residuals:
      Min        1Q    Median        3Q       Max
-0.26259 -0.02103  0.01005  0.03628  0.15928

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.2594325  0.1247307 -10.097  < 2e-16 ***
x1           0.0017374  0.0005979   2.906  0.00387 **
x2           0.0029196  0.0010895   2.680  0.00768 **
x3           0.0057167  0.0047704   1.198  0.23150
x4          -0.0033052  0.0055616  -0.594  0.55267
x5           0.0223531  0.0055415   4.034  6.6e-05 ***
x6           0.1189395  0.0122194   9.734  < 2e-16 ***
x7           0.0245251  0.0079598   3.081  0.00221 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06378 on 392 degrees of freedom
Multiple R-squared:  0.8035,    Adjusted R-squared:  0.8
F-statistic: 228.9 on 7 and 392 DF,  p-value: < 2.2e-16
```

回归方程通过方程的显著性检验。由对参数的显著性检验可以剔除不显著的变量 x_3, x_4(分别代表申请大学排名以及个人陈述)。剔除不显著的变量后，对剩余变量重新进行多元线性回归的最小二乘拟合。

2 建立显著变量的线性回归模型

将上述未通过 t 检验的变量剔除，重新建立多元线性最小二乘模型如下。

```
z1<-lm(y~x1+x2+x5+x6+x7)
summary(z1)

Call:
lm(formula = y ~ x1 + x2 + x5 + x6 + x7)

Residuals:
      Min        1Q    Median        3Q       Max
-0.263542 -0.023297  0.009879  0.038078  0.159897

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.2984636  0.1172905 -11.070  < 2e-16 ***
x1           0.0017820  0.0005955   2.992  0.00294 **
x2           0.0030320  0.0010651   2.847  0.00465 **
x5           0.0227762  0.0048039   4.741 2.97e-06 ***
```

```
x6          0.1210042  0.0117349  10.312  < 2e-16 ***
x7          0.0245769  0.0079203   3.103  0.00205 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06374 on 394 degrees of freedom
Multiple R-squared:  0.8027,    Adjusted R-squared:  0.8002
F-statistic: 320.6 on 5 and 394 DF,  p-value: < 2.2e-16
```

此时所有变量通过显著性检验，回归方程通过显著性检验，可以初步建立回归模型：

$$\hat{y} = -1.298 + 0.0018x_1 + 0.0030x_2 + 0.0228x_5 + 0.1210x_6 + 0.0246x_7$$

接下来检验模型是否违背以上模型的基本假设。

3 模型检验

在回归模型的基本假设中，要求随机误差项 $\varepsilon_1, \cdots, \varepsilon_n$ 满足 $G-M$ 条件即：

$$\begin{cases} E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2, & i = 1, 2, \cdots, n \\ Cov(\varepsilon_i, \varepsilon_j) = 0, & i \neq j \end{cases};$$

但在建立实际问题的回归模型时，经常存在于此假设相违背的情况，分别为异方差性，即：

$$Var(\varepsilon_i) \neq Var(\varepsilon_j), \quad \exists i \neq j$$

与自相关性：

$$Cov(\varepsilon_i, \varepsilon_j) \neq 0, \quad \exists i \neq j$$

虽然在上述过程中我们已经确定了模型的显著性以及所有变量的显著性，但仍需对以上两种违背基本假设情况的存在性进行检验。

3.1 异方差检验

首先检验异方差的存在性，可以离用残差图检验法绘制残差图直接观察分析是否存在异方差。

```
e<-resid(z1)
plot(e)
```

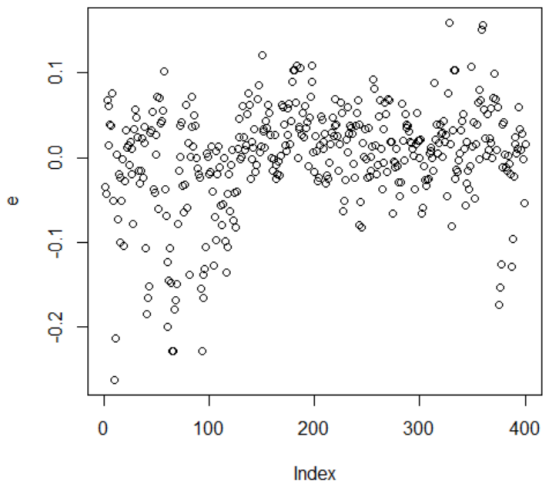

图 1: 离差项散点图

观察残差图可以初步确认异方差的存在性。为了进一步确认异方差的存在性，再利用 R 语言的 *bptest* 函数检验。

```
bptest(z1)

        studentized Breusch-Pagan test

data:  z1
BP = 22.428, df = 5, p-value = 0.0004341
```

由检验结果，模型存在异方差，需要后续通过 Box-Cox 变换消除。

3.2 自相关性检验

利用 DW 检验模型是否存在自相关性。

```
library(lmtest)
载入需要的程辑包：zoo

载入程辑包：‘zoo’

The following objects are masked from ‘package:base’:

    as.Date, as.Date.numeric

Warning messages:
1: 程辑包‘lmtest’是用R版本3.6.3 来建造的
2: 程辑包‘zoo’是用R版本3.6.3 来建造的
library(zoo)
dw<-dwtest(z1)
dw

        Durbin-Watson test

data:  z1
DW = 0.74991, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

显然检验的 DW 值小于 $n = 400$ 对应的 d_L, 故原数据存在自相关，需要通过后续处理消除。

4 模型修正

由以上过程，需要对模型进行修正来消除异方差性和自相关性。

4.1 Box-Cox 变换

首先利用 Box-Cox 变换消除异方差性，求出使得对数似然函数值达到最大的 λ 值，再对新变量 $y^{(\lambda)} = \dfrac{y^{\lambda} - 1}{\lambda}$ 对自变量做回归，得到新的回归方程。

```
re<-boxcox(y~x1+x2+x5+x6+x7,lambda=seq(-3,3,0.1))
lambda<-re$x[which.max(re$y)]
lambda
[1] 2.272727
new_y<-y^(lambda-1)/lambda
new_z<-lm(new_y~x1+x2+x5+x6+x7)
summary(new_z)

Call:
lm(formula = new_y ~ x1 + x2 + x5 + x6 + x7)

Residuals:
      Min        1Q    Median        3Q       Max
-0.129101 -0.011955  0.004405  0.018661  0.079480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -0.7391922  0.0577983 -12.789  < 2e-16 ***
x1           0.0009156  0.0002935   3.120  0.00194 **
x2           0.0015821  0.0005249   3.014  0.00274 **
x5           0.0114251  0.0023673   4.826 1.99e-06 ***
x6           0.0612394  0.0057827  10.590  < 2e-16 ***
x7           0.0128628  0.0039029   3.296  0.00107 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03141 on 394 degrees of freedom
Multiple R-squared:  0.8132,    Adjusted R-squared:  0.8108
F-statistic:   343 on 5 and 394 DF,  p-value: < 2.2e-16
```

4.2 检验自相关性

对经过 Box-Cox 变换得到的新回归方程进行 DW 检验判断是否仍存在自相关性。

```
new_dw<-dwtest(new_z)
new_dw

        Durbin-Watson test

data:  new_z
DW = 0.75709, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

显然新模型对应的 DW 值仍然在 $(0, d_L)$ 的区间内，因此可以判断模型对应的残差项仍然存在一阶自相关，需要进一步消除来完善模型。

4.3 迭代法消除自相关

```
new_y.dif<-vector(length=length(y)-1)
x1.dif<-vector(length=length(x1)-1)
x2.dif<-vector(length=length(x2)-1)
x5.dif<-vector(length=length(x5)-1)
x6.dif<-vector(length=length(x6)-1)
x7.dif<-vector(length=length(x7)-1)
rho<-1-0.75709/2
for(i in 1:399){new_y.dif[i]=new_y[i+1]-rho*new_y[i]
x1.dif[i]=x1[i+1]-rho*x1[i]
x2.dif[i]=x2[i+1]-rho*x2[i]
x5.dif[i]=x5[i+1]-rho*x5[i]
x6.dif[i]=x6[i+1]-rho*x6[i]
x7.dif[i]=x7[i+1]-rho*x7[i]}
new_z.dif<-lm(new_y.dif~x1.dif+x2.dif+x5.dif+x6.dif+x7.dif)
summary(new_z.dif)

Call:
lm(formula = new_y.dif ~ x1.dif + x2.dif + x5.dif + x6.dif +
    x7.dif)

Residuals:
      Min        1Q    Median        3Q       Max
-0.109482 -0.011197  0.002566  0.014574  0.073381

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2546826  0.0158577 -16.060  < 2e-16 ***
x1.dif       0.0008939  0.0002146   4.165 3.84e-05 ***
x2.dif       0.0018686  0.0003893   4.800 2.26e-06 ***
x5.dif       0.0089477  0.0017455   5.126 4.65e-07 ***
x6.dif       0.0516554  0.0045322  11.397  < 2e-16 ***
x7.dif       0.0143360  0.0026551   5.400 1.16e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02418 on 393 degrees of freedom
Multiple R-squared:  0.8002,    Adjusted R-squared:  0.7977
```

```
F-statistic: 314.8 on 5 and 393 DF,  p-value: < 2.2e-16
```

经过迭代法处理得到的回归方程通过方程的显著性检验，所有变量通过显著性 t 检验。接下来对此模型进行异方差检验与自相关检验。

```
new_dw.dif<-dwtest(new_z.dif)
new_dw.dif

        Durbin-Watson test

data:  new_z.dif
DW = 1.8901, p-value = 0.1496
alternative hypothesis: true autocorrelation is greater than 0

plot(new_e.dif)
library(car)
bptest(new_z.dif)

        studentized Breusch-Pagan test

data:  new_z.dif
BP = 9.126, df = 5, p-value = 0.1041
```

模型已消除异方差和自相关，通过检验。

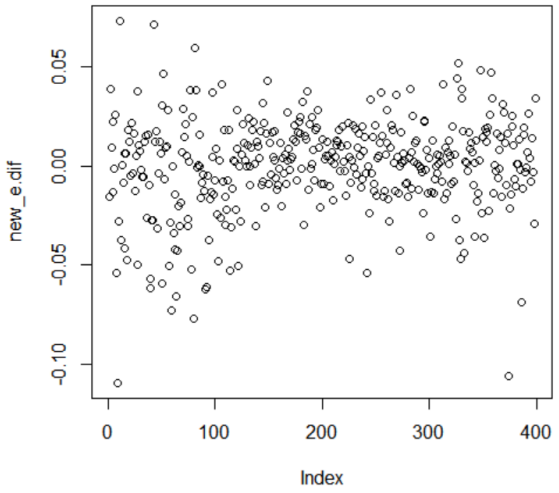

图 2: 离差项散点图

5 最终拟合结果

经过对全体数据进行回归 $\longrightarrow$ 变量显著性的 t 检验剔除不显著变量 $\longrightarrow$ 残差异方差、自相关存在性检验及消除 $\longrightarrow$ 处理后新模型显著性检验、变量显著性检验一系列过程，可以得到最终模型为：

$$\widehat{y_t'} = -0.2546826 + 0.0008939x_{1t}' + 0.0018686x2t' + 0.0089477x_{5t}' + 0.0516554x_{6t}' + 0.0143360x_{7t}'.$$

其中，

$$\begin{cases} \widehat{y_t'} = \widehat{y}_{t+1}^{(\lambda)} - \rho y_t^{(\lambda)}, \quad \rho = 1 - \frac{0.75709}{2}, \lambda = 2.272727 \\ y^(\lambda) = \dfrac{y^\lambda - 1}{\lambda} \\ x_{it} = x_{i,t+1} - \rho x_{it}, \quad i = 1, 2, 5, 6, 7 \end{cases}$$

附录

[1] 数据来源：Admission_Predict.csv