MATLAB

Assignment Series

The Project

-The Research on Students' Admission in India-

By Kinsgley Cheng

Student ID: 202228000243001

Academy of Mathematics and System Science

2022 年 9 月 29 日

Abstract

本文主要通过建立多元线性回归模型研究影响印度本科学生升学录取机率的因素，研究的因素范围包括 GRE 成绩、TOFEL 成绩、大学评分、个人陈述、推荐信、本科 GPA 和科研经历这 7 个因素。

首先，我们对数据进行预处理，画出各变量的频数直方图、箱型图、计算各变量的基本统计量，包括最大值、最小值、中位数和众数、画出各变量之间的相关热力图。从而能够对所处理数据的分布状况、相关性和异常值的情况有一个初步的了解。

接下来，我们对回归方程进行全模型回归，了解数据的整体线性性质，以测试线性回归模型的可行性，发现方程整体显著性高，只存在少量变量不显著。这说明该问题大致可以采用线性模型进行解释。为了处理不显著的变量，我们依次采用逐步回归、LASSO 回归、Elastic Net 等方法分别进行变量选择，期望通过统计性质剔除不相关因素，以提高模型整体的拟合度与可解释性。通过比较各种方法的结果，我们发现在这里 LASSO 回归与 Elastci Net 对变量的分离性能并不太好，所以最终采用了逐步回归的结果，即删除变量—个人陈述。

然后，我们对变量选择后的子集重新进行回归，对得到的回归方程进行统计性检验。首先进行异方差诊断。在画出残差散点图后，无法判断其异方差性，接着分别采用 Spearman 检验和 White 检验分别对回归方程的变量和整体进行诊断，发现方程存在严重异方差。对此，我们采用 BOX-COX 变换进行处理，通过极大似然最小 MSE 的方法选取 λ，处理后发现重新检验，发现异方差已消除。

下一步，我们进行自相关检验。我们画出残差时序图，初步判断回归方程可能存在一定自相关性，再通过 DW 检验，进一步确定了模型自相关性的存在。我们再一次观测时序图，发现残差间存在近似线性相关性，故推断其存在一阶自相关，进而采用迭代法进行处理，处理后再次进行检验，发现方程自相关性被消除。

最后，我们对所得的回归方程，进行多重共线性的检验。主要是计算其方差膨胀因子 (VIF) 和条件数，发现不存在共线性，至此回归方程建立完成。

对于上述得到的回归方程，我们进行初步的回归诊断，分别画出其学生化残差散点图、Cook 距离散点图和各个杠杆值相较于平均杠杆值倍数的散点图，基本判断回归方程存在极少的异常值，但可能存在一定的高杠杆值点。

最后通过得到的回归方程，我们对模型进行一定的初步解释。可以发现，对于印度学生而言，大学的录取率和个人推荐信的好坏并没有很大的关系，而和本科 GPA 与是否具有科研经历具有密切关系。

Keywords: 多元线性回归、变量选择、BOX-COX 变换、升学录取

目录

1 问题背景

对于每一个学生而言，大学的选择与录取都对他们今后的人生发展有着至关重要的影响。而大学的录取率受多种因素影响，且其中的关系复发，预测每位学生的录取机率是一个极其复杂的问题。相关部门经过初步调查和对往年录取情况的一定的研究分析后，发现学校的录取与否主要取决于 GRE 成绩、TOFEL 成绩、大学评分、个人陈述、推荐信、本科 GPA 和科研经历这 7 个因素的相关性较高。

现在我们有 500 名印度学生的 7 项指标与他们的录取机率的数据[1]，我们希望借此建立线性回归模型，能够帮助将来的学生更加清晰准确地来预测与评估自己的录取机率，以便更有效及时的获取大学的录取。与此同时，也希望给一些在籍本科生提供准备方向的参考和侧重，能够在将来申请时到更好的结果。

2 基本假设

对于所处理的问题与建立的模型，我们有如下几条基本假设：

1. 假定数据的来源均真实可靠，录入与统计误差不会对模型产生严重的影响。
2. 假定 500 名学生相互独立，各个数据间不存在相关关系。
3. 假定各个学校对于每个学生的推荐信和个人陈述的评分尺度一致，不存在主观人为因素，各个学校间的差异可以忽略。同时，数据中的大学评分与录取几率的评估具有一定的权威性，即能够反应真实情况。

3 符号说明

变量符号	变量含义	取值范围	单位
n	学生序号	1、2、$\cdots$、500	/

[1]data is from Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

x_1	GRE 成绩	0∼340	分
x_2	TOEFL 成绩	0∼120	分
x_3	大学评分	1、2、3、4、5	分
x_4	个人陈述评分 (SOP)	1、2、3、4、5	分
x_5	推荐信评分 (LOR)	1、2、3、4、5	分
x_6	本科 GPA(CGPA)	0∼10	/
x_7	科研经历	0、1	/
y	录取机率	0∼1	/

4 模型建立与求解

4.1 数据预处理与基本分析

缺失值检验

首先我们需要对数据进行缺失值检验。通过检验，可以发现原数据集并无缺失值存在。

直方图—数据分布形态的分析

接下来，我们分别做出各变量的频率直方图，以此来了解数据的大致分布情况。

图 1: 前四个变量的直方图

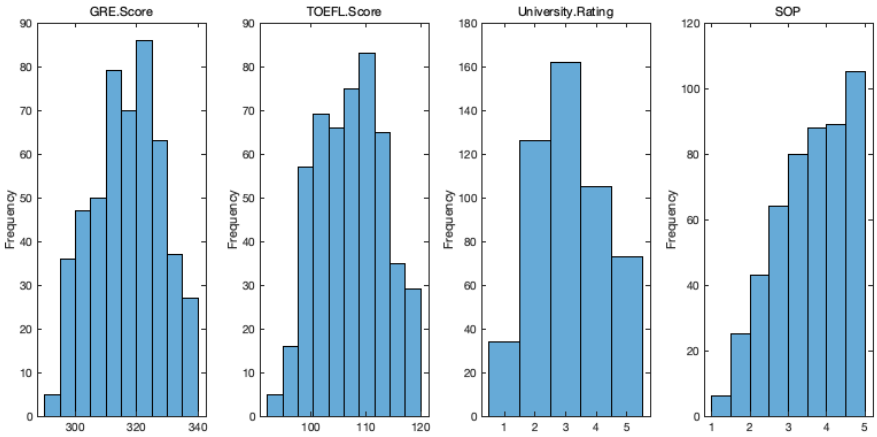
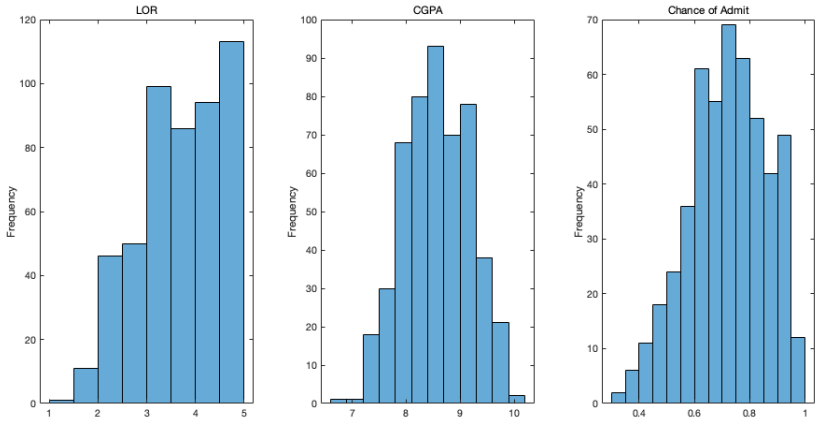


图 2: 后三个变量的直方图



基本统计量

我们可以计算出各变量的基本统计量，结果统计如下：

表 2: 各变量的基本统计量

变量	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
最大值	340.0	120.0	5.0	5.0	5.0	9.9	1.0	0.97
最小值	290.0	92.0	1.0	1.0	1.0	6.8	0.0	0.34
平均值	316.5	107.2	3.1	3.4	3.5	8.6	0.6	0.72
中位数	317	107.0	3.0	3.5	3.5	8.6	81.0	0.72

箱形图—异常值初步判断

下面我们画出箱型图，对一些异常值进行初步判断

图 3: 前 4 个变量的箱形图

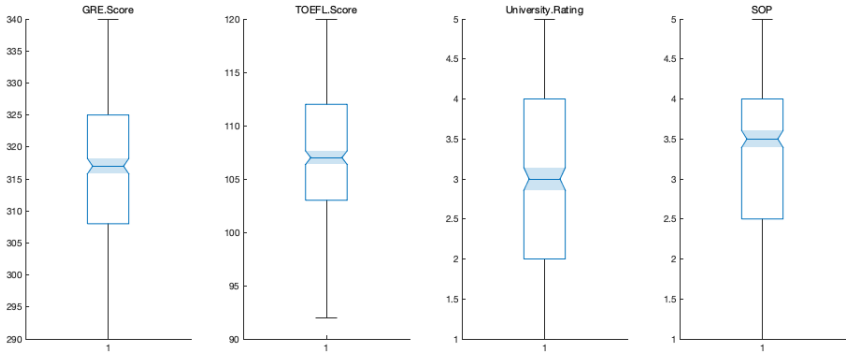
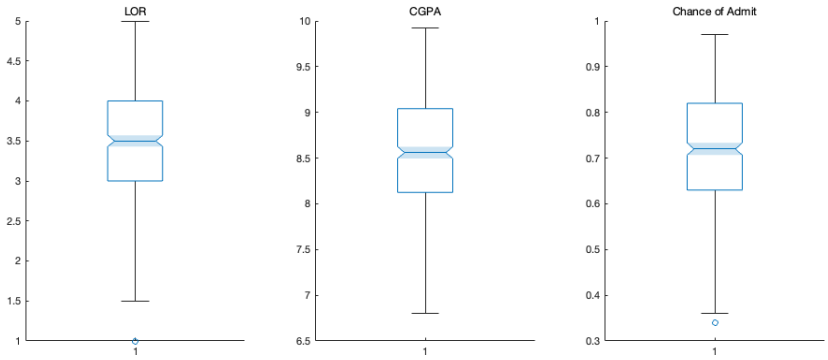


图 4: 后 3 个变量的箱形图

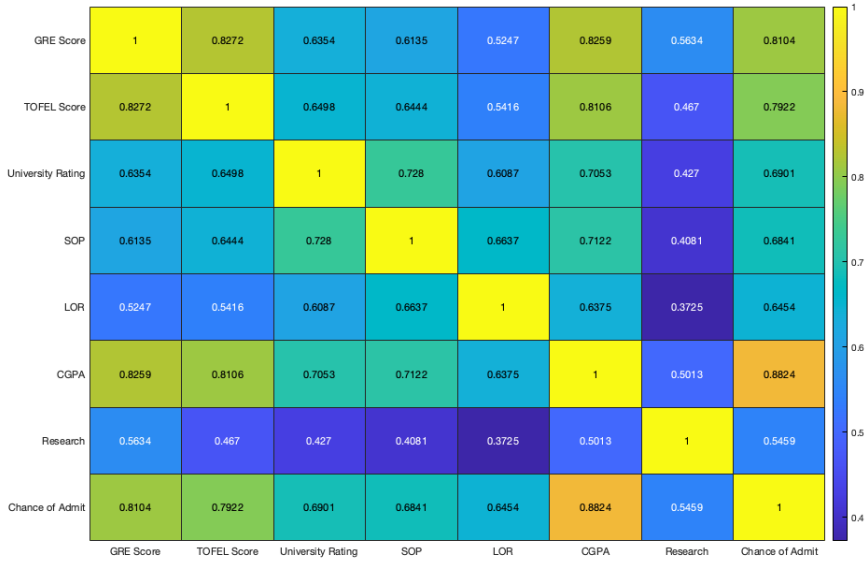


通过箱形图我们可以发现变量 LOR 和因变量录取几率存在异常值，考虑到异常值数量较少，我们将其删除（93、348、377 行数据）。

相关热力图—变量相关性的初步分析

下面我们画出各变量之间的热力图，对它们之间的线性相关性有一个比较直观地了解。

图 5: 相关系数热力图



从上图我们可以发现录取几率和自变量 GRE 分数、CGPA、TOEFL 成绩的相关性较高。同时 CGPA、TOFEL 分数、GRE 分数之间的相关性也都很高。

4.2 建立多元线性回归模型

建立全模型

首先，我们对所有预测变量进行回归，得到全模型的回归方程。再对回归方程做相关的显著性检验。

鉴于已有的 MATLAB 函数中对于线性回归的 regress 函数所能提供的输出过少，并没有相关的统计指标与统计检验，无法满足需求。因此，我单独编写了 myregression 函数，将相应的指标都进行了一定的计算，对于对应的检验同时输出了 p 值，以供判断。具体的代码如下：

```
function [beta,r2,adjr2,F,Ftest,t,ttest,residuals] = myregression(x,y)
n = size(x,1);
m = size(x,2);
b = ones(n,1);
```

```
x = [b,x];
beta = (x'*x)\(x'*y);
ttest = zeros(1,m+1);
sst=(y-mean(y))'*(y-mean(y));
sse=(y-x*beta)'*(y-x*beta);
ssr=sst-sse;
F=(ssr/m)/(sse/(n-m-1));
r2=ssr/sst;
adjr2=1-(n-1)/(n-m-1)*(1-r2);
Ftest=1-fcdf(F,m,n-m-1);
c=diag(inv(x'*x));
t=zeros(1,m+1);
sigma=sqrt(sse/(n-m-1));
for i=1:m+1
t(i)=beta(i)/(sqrt(c(i))*sigma);
ttest(i)=2*(1-tcdf(abs(t(i)),n-m-1));
end
residuals = y-x*beta;
```

参数估计、t 检验结果、R^2 值和调整的 R^2 值等各项统计指标汇总于下表

表 3: 全模型各项指标汇总

变量	Intercept	x_1	x_2	x_3	x_4	x_5	x_6	x_7
估计	-1.235102	0.001787	0.002581	0.005412	0.003872	0.016011	0.118506	0.023749
t 值	-12.010	3.618	3.005	1.447	0.858	3.927	12.423	3.659
P 值	.000***	.000***	.002**	.149	.391	.000***	.000***	.000***

R^2	调整 R^2
0.8223	0.8198

通过上表数据我们可以发现回归方程整体效果良好，但存在参数并不显著。下面我们进行变量选择，我们将依次采用逐步回归、LASSO 回归和 Elastic Net。

变量选择 1—逐步回归

首先我们通过逐步回归进行变量选择，并指定模型选择标准为 AIC 准则。

我们选择 MATLAB 函数 stepwiselm 来实现该项功能，并选择“Criterion”为“aic”，“Upper”为”linear”，忽略交叉项的考虑。

选择的结果列于下表：

表 4: 逐步回归子模型各项指标汇总

变量	Intercept	x_1	x_2	x_3	x_5	x_6	x_7
估计	-1.2462544	0.0017757	0.0026542	0.0065987	0.0170505	0.1199758	0.0238714
t 值	-12.220	3.597	3.107	1.9	4.381	12.788	3.680
P 值	.000***	.000**	.002**	.058.	.000***	.000***	.000***

R^2	调整 R^2
0.8221	0.8199

通过上表的各项指标，我们发现逐步回归建立的子模型删去了变量 SOP，整体的 R^2 和调整后的 R^2 基本没有太大改变，但相应的各预测变量的显著性得到了进一步的改进。

变量选择 2—LASSO 回归

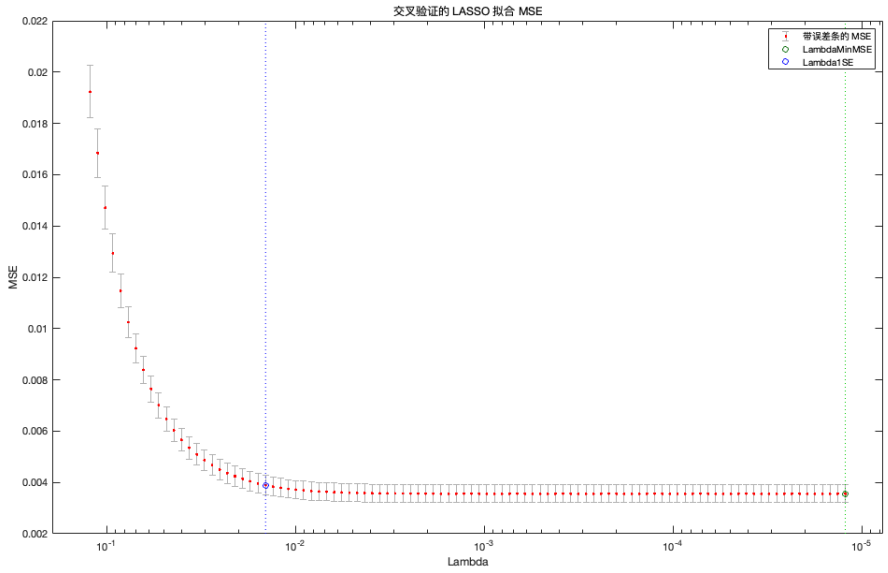
由于逐步回归的惩罚函数不连续，且每一次选择的跨度过大，所以下面我们尝试采用 LASSO 回归的稀疏性进行变量选择，并观察其选择结果。其中 λ 的选择采用交叉验证 (CV)。

采用 CV 交叉验证选取 λ 值为 0.0000125(由于 CV 分组存在随机性，故可能每次运行结果不一致), 此时各变量参数列于下表

表 5: LASSO 回归变量参数

变量	Intercept	x_1	x_2	x_3	x_4	x_5	x_6	x_7
估计	-1.1054	0.001682	0.001876	0.002508	0.001178	0.009416	0.121944	0.007376

下面我们给出 CV 验证下不同 λ 的交叉验证误差图。



综上所述，我们得到 LASSO 回归并没有删去任何变量，选择全模型。

变量选择 3—Elastic Net

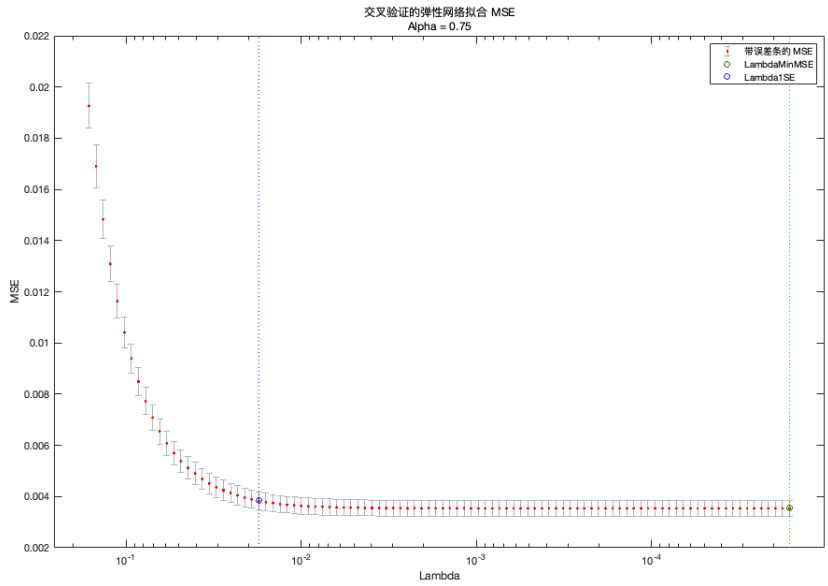
由于 LASSO 回归的惩罚函数存在有偏性，下面我将尝试采用 Elastic Net 进行变量选择，其中 λ 的选择采用交叉验证 (CV)。

采用 CV 交叉验证选取 λ 值为 0.0000163(由于 CV 分组存在随机性，故可能每次运行结果不一致), 此时各变量参数列于下表

表 6: Elastic Net 变量参数

变量	Intercept	x_1	x_2	x_3	x_4	x_5	x_6	x_7
估计	-1.115540	0.001717	0.001995	0.002913	0.001662	0.010131	0.119603	0.008938

下面我们给出 CV 验证下不同组别的交叉验证误差图。



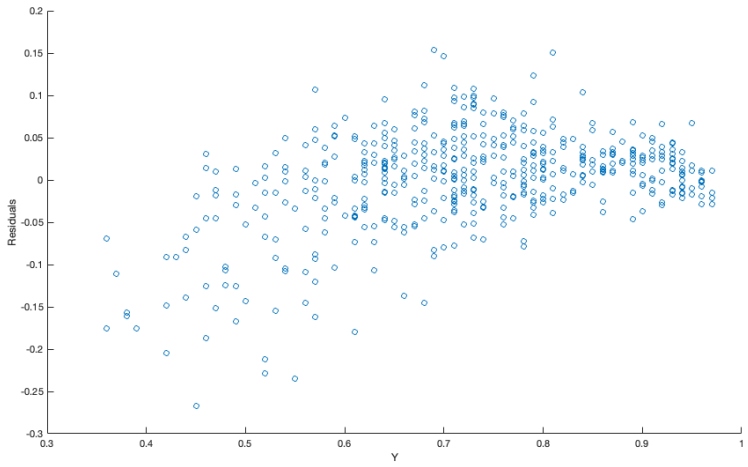
综上所述，我们得到 Elastic Net 并没有删去任何变量，选择全模型。

我们通过结果发现这里 LASSO 和 Elastic Net 对变量的分离效果不是很好，综合上述三种方法得到的选择结果，我们最终选择子模型 $\{x_1, x_2, x_3, x_5, x_6, x_7\}$。

4.3 异方差的检验与处理

首先，我们画出响应变量 Y 关于残差的散点图，通过图像进行初步判断。

图 6: 修正前残差散点图



通过上图，我并不能准确判断回归方程是否存在异方差，需要进行进一步检验

异方差检验 1—Spearman 检验

为了确定回归方程的异方差性，下面我们对各个预测变量做 Spearman 检验。

由于 MATLAB 函数中并没有相关检验的函数，因此我们编写了相应的检验函数 spearmantest，输入值为任意一个回归变量的观测样本向量和回归后的残差向量，输出值为 spearman 相关系数，对应检验的 t 值和检验 p 值。具体代码如下：

```
function [rho,tvalue,pvalue] = spearmantest(x,residual)
n = size(x,1);
[rho,p] = corr(x,abs(residual),"type","Spearman");
tvalue = (sqrt(n-2)*rho)/sqrt(1-rho^2);
pvalue = 2*(1-tcdf(abs(tvalue),n-2));
end
```

检验的结果列于下表中

表 7: Spearman 检验结果

变量	x_1	x_2	x_3	x_5	x_6	x_7
P 值	0	0	0	0	0	0.001521
rho	-0.316197	-0.2512531	-0.2112627	-0.2772145	-0.3240891	-0.1418634

由此可见，各变量的 P 值均很低，存在强烈的异方差。

异方差检验 2—white 检验

我们通过 white 检验，我们可以对回归方程整体的异方差性进行检验

由于 MATLAB 函数中并没有相关检验的函数，因此我们编写了相应的检验函数 whitetest，输入值为回归变量的观测样本矩阵和回归后的残差向量，输出值为 White 统计量和检验 p 值。具体代码如下：

```
function [W,pvalue] = whitetest(x,residual)
n=size(x,1);
m=size(x,2);
res2 = residual.^2;
xtest =x;
for i =1:m
xtest = [xtest,x(:,i).^2];
end
[betat,r2t,adjr2t,Ft,Ftestt,tt,ttestt,residualst] = myregression(xtest,res2);
W=n*r2t;
pvalue = 1-chi2cdf(W,size(xtest,2));
end
```

检验结果如下：

```
W=34.8798, p-value = 0.0004894
```

white 检验同样显示，回归方程存在异方差性。

异方差处理———BOX-COX 变换

BOX-COX 变换是在 1964 年由 BOX 和 COX 提出的一种通过对因变量处理来消除异方差的方法，主要进行的变换为：

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda} & ,\lambda \neq 0 \\ \ln y & ,\lambda = 0 \end{cases}$$

其中 λ 为待定系数。

此变换要求 y 的各分量都大于 0。否则可先将因变量 y 做一个平移变换，使得平移后各分量均大于 0，再进行 BOX-COX 变换。从而我们得到推广的 BOX-COX 变换：

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+a)^{\lambda}-1}{\lambda} & ,\lambda \neq 0 \\ \ln(y+a) & ,\lambda = 0 \end{cases}$$

很显然，在不同的 λ 下，因变量 y 所作的变换是不一样的，产生的效果也是不一样的。为了消除异方差，我们希望寻找到合适的 λ，使得便后有：

$$y^{(\lambda)}=\begin{pmatrix} y_1^{(\lambda)} \\ y_2^{(\lambda)} \\ \vdots \\ y_n^{(\lambda)} \end{pmatrix} \sim N(X\beta,\sigma^2 I)$$

事实上，由此我们可以发现，BOX-COX 变换不仅可以处理异方差性问题，还能处理自相关、误差非正态、回归函数非线性等问题。

寻找合适的 λ 的方法有很多，我们可以通过计算 λ 的对数极大似然估计得到。λ 的极大似然估计：

$$L_{\max}(\lambda)=(2\pi e\hat{\sigma}_\lambda^2)^{-\frac{n}{2}}|J|$$

式中，$\hat{\sigma}_\lambda^2=\frac{1}{n}SSE(\lambda,y^{(\lambda)}),|J|=\prod_{i=1}^{n}y^{(\lambda)}$

令 $z^{(\lambda)}=\frac{y^{(\lambda)}}{|J|}$，对 $L_{\max}$ 取对数并忽略与 λ 无关的常数项，可得：

$$\ln L_{\max}(\lambda)=-\frac{n}{2}\ln SSE(\lambda,z^{(\lambda)})$$

此时只需取使 $SSE(\lambda,z^{(\lambda)})$ 最小的 λ 即可。但在实际情况中，其解析解一般无法得到，我们可以通过大量 λ 值计算，得到近似最优解。

在这里我们从-5 到 5 以步长为 0.1 依次取 λ 的取值进行回归，通过最大对数似然值来选取最终的 λ。

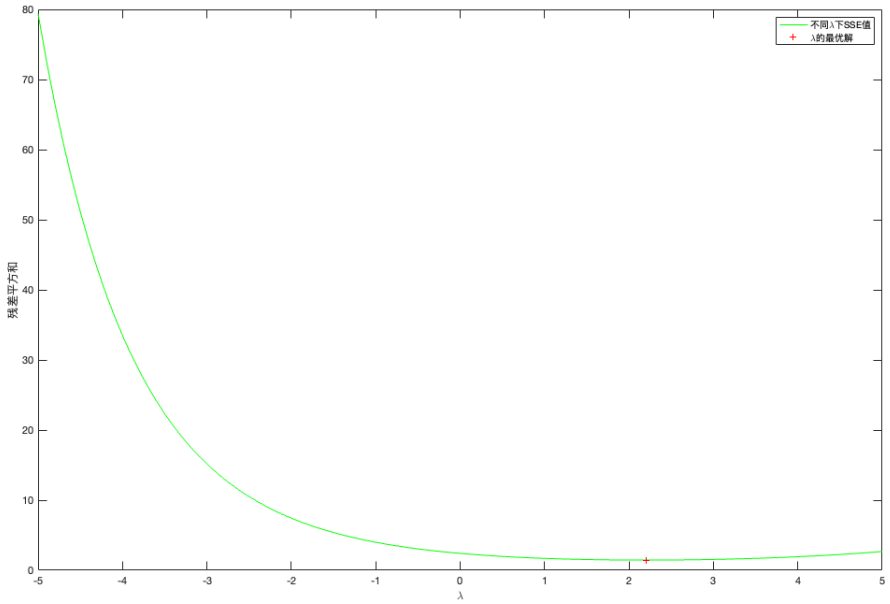
在这里我们编写了一个在给定 λ 下计算 SSE 值大小的函数，名称为 bocc，具体代码如下：

```
function SSE=bocc(X,Y,lambda)
H=X*inv(X'*X)*X';
n=length(Y);
switch lambda
case 0
z=log(Y)*prod(Y)^(1/n);
otherwise
```

```
z=(Y.^lambda-1)/lambda/(prod(Y)^((lambda-1)/n));
end
SSE=z'*(eye(n)-H)*z;
end
```

各 λ 下的极大似然图如下所示

图 7: BOX-COX 变换下极大似然值



λ 最终取值为 2.2, 经过 BOX-COX 变换后，重新进行线性回归。我们将参数估计、t 检验结果、R^2 值和调整的 R^2 值汇总于下表：

表 8: BOX-COX 变换后回归方程各项指标汇总

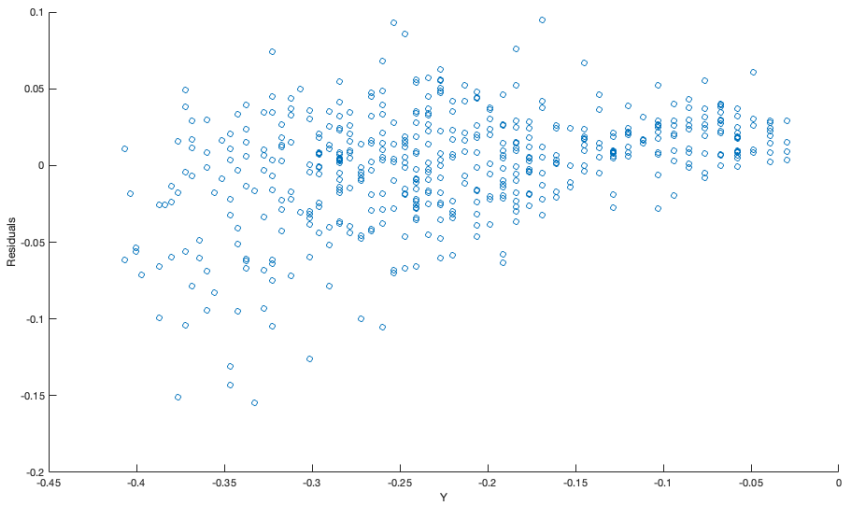
变量	Intercept	x_1	x_2	x_3	x_5	x_6	x_7
估计	-1.5480818	0.0011814	0.0021030	0.0069611	0.0098556	0.0771804	0.0167611
t 值	-24.758	3.903	4.015	3.268	4.130	13.417	4.214
P 值	.000***	.000***	.000***	.001**	.000***	.000***	.000***

R^2	调整 R^2
0.8501	0.8483

经过变换后可以发现，P 值，R^2 和调整的 R^2 均有明显的提高，回归方程的显著性有一定的提升, 并且所有系数均通过了 t 检验。

我们重新再画出残差的散点图：

图 8: BOX-COX 变换后的残差散点图



可以发现，残差有显著的减小。为了进一步地确定，我们再对各变量进行 Spearman 检验，检验结果列于下表

表 9: Spearman 检验结果

变量	x_1	x_2	x_3	x_5	x_6	x_7
秩 S	20367287	20338509	20547614	20806322	20308580	20533394
P 值	0.9193	0.8945	0.9246	0.707	0.8689	0.9369
rho	0.004555572	0.005962061	-0.004257869	-0.01690215	0.007424821	-0.003562844

再进行 white 检验，检验结果如下文本框

```
studentized Breusch-Pagan test
W = 20.6335, p-value = 0.0560
```

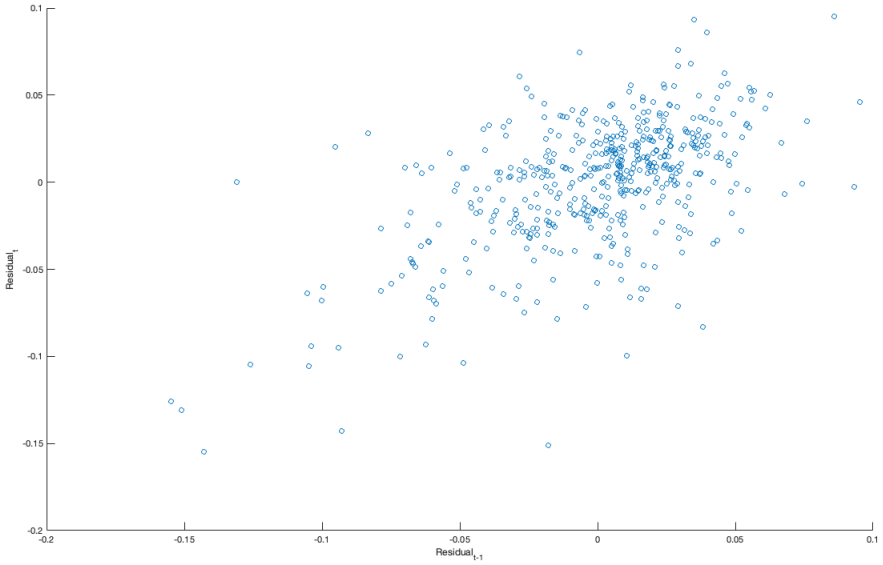
通过上面两表中的数据，可以肯定，异方差已消除。

4.4 自相关的检验与处理

自相关检验—DW 检验

首先，我们画出残差 e_t 与 e_{t-1} 的散点图，通过图形来初步判断自相关性。散点图如下：

图 9: 残差时序图



通过图形，我们可以初步判断回归方程存在一定的自相关性。为了进一步的确定，我们下面对回归模型进行 DW 检验。

由于 MATLAB 并没有相关的计算函数，因此我们编写了 dwtest 函数，输入值为残差向量，输出值为 DW 值和相关系数估计值 $\hat{\rho}$

检验结果列于下表：

表 10: DW 检验检验结果

D.W. 值	P 值	D_L	D_U
0.8723	0.000	1.70	1.83

由于 $0 \leq D.W. \leq D_L$，故回归方程的误差项之间存在正自相关性。

自相关的处理—迭代法

我们通过图 10可以初步判断，残差项存在一阶自回归形式，故我们采用迭代法进行处理。

对于一元回归方程而言，我们设一元线性回归模型的误差项存在一阶自相关性，即:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + \mu_t$$

其中 μ_t 满足 $Gauss - Markov$ 条件，即:

$$\begin{cases} E(\mu_t) = 0 & t = 1, 2\cdots, n \\ cov(\mu_t, \mu_s) = \begin{cases} \sigma^2 & t = s \\ 0 & t \neq s \end{cases} & t, s = 1, 2, \cdots, n \end{cases}$$

通过上述条件，我们可以作如下变形:

$$(y_t - \rho y_{t-1}) = (\beta_0 - \rho\beta_0) + \beta_1(x_t - \rho x_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1})$$

下面我们令:

$$y'_t = y_t - \rho y_{t-1}$$
$$x'_t = x_t - \rho x_{t-1}$$
$$\beta'_0 = \beta_0 - \rho\beta_0$$

从而获得新的回归方程如下:

$$y'_t = \beta'_0 + \beta_1 x'_t + \mu_t$$

很容易发现，获得的新的回归方程满足基本假设。在一般情况下，对于迭代法中的自相关系数 ρ，我们用 $\hat{\rho} = 1 - \frac{1}{2}\mathrm{DW}$ 来估计。

对于现在的多元情况，我们可以类似一元的情形进行处理。在这里，我们取 $\hat{\rho} = 1 - \frac{1}{2}\mathrm{DW} = 0.5638$, 迭代后重新进行线性回归，我们将参数估计、t 检验结果、R^2 值和调整的 R^2 值汇总于下表:

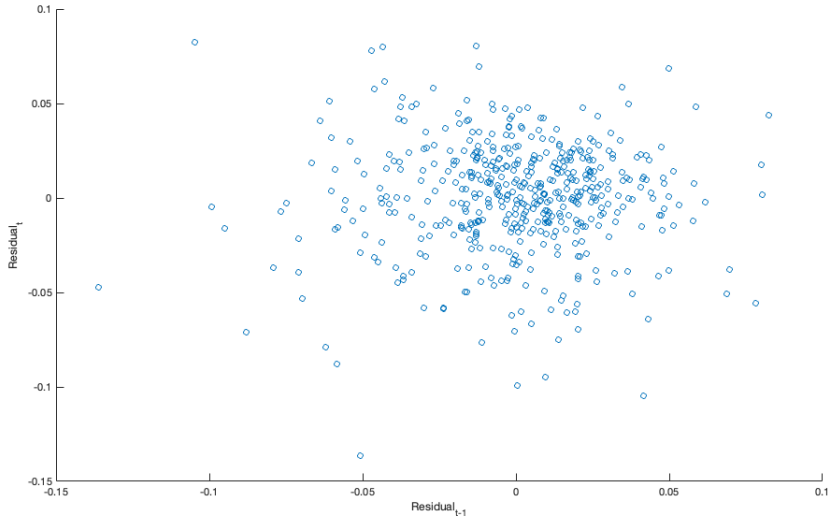
表 11: 迭代后回归方程各项指标汇总

变量	Intercept	x_1	x_2	x_3	x_5	x_6	x_7
估计	-0.655759	0.001134	0.002534	0.005213	0.009079	0.069185	0.018304
t 值	-29.684	4.744	6.021	2.828	4.847	14.023	6.176
P 值	.000***	.000***	.000***	.006**	.000***	.000***	.000***

R^2	调整 R^2
0.8351	0.8331

对于新得到的方程，我们重新画出残差时序图

图 10: 残差时序图



重新进行 DW 检验，检验结果列于下表：

表 12: D.W. 检验结果

D.W. 值	P 值	D_L	D_U
1.9717	0.3822	1.70	1.83

此时已通过 DW 检验，方程的共线性已消除。

4.5 共线性检验与处理

共线性检验———VIF 与条件数

最后我们需要对得到的回归模型做共线性检验。我们首先计算各自变量的方差膨胀因子 (VIF)。

由于 MATLAB 并没有提供相关的计算函数，因此我们编写了 vif 函数来计算。该函数的输入值为回归变量的样本观测矩阵，输出为每个回归变量的方差膨胀因子。具体的代码如下:

```
function [diagvalue] = vif(x)
x = corr(x,"type","Pearson");
diagvalue = diag(inv(x));
end
```

利用函数的计算结果如下表所示。

表 13: 各自变量的 VIF 值

变量	x_1	x_2	x_3	x_5	x_6	x_7
VIF	3.31510	2.94016	1.75433	1.32270	3.23408	1.26914

由上表结果，我们发现回归方程不存在共线性。下面，我们再计算其条件数 k，进一步确认结果。

由于条件数的计算 MATLAB 并没有提供相应的函数,因此我们编写了对应的计算函数 condvaluecal。其中输入值为回归变量的样本观测矩阵，输出为每个回归变量的条件数。具体的代码如下:

```
function [condvalue] = condvaluecal(x)
x = corr(x,type="Pearson");
eigvalue = eig(x);
eigmax = max(eigvalue);
condvalue = sqrt(eigmax./eigvalue);
end
```

计算得条件数如下表

表 14: 各自变量的条件数

变量	x_1	x_2	x_3	x_5	x_6	x_7
k	1	4.172	3.748	2.699	2.223	2.170

由上表结果，我

由此可以确定，方程不存在共线性。

5 模型分析与评价

5.1 模型分析

通过上述变换后我们最终得到了如下回归方程：

$$\begin{cases} z' = --0.65575 + 0.001134x_1' + 0.002534x_2' + 0.005213x_3' + 0.009079x_5' + 0.069185x_6' + 0.018304x_7' \\ z' = \dfrac{z^{\lambda} - 1}{\lambda} \\ \lambda = 2.2 \end{cases}$$

下面我们进行回归诊断，主要为残差分析与影响分析，用来检验回归方程是否存在大量异常值点。

异常值通常分为三类，具体定义如下：

- 离群点 (Outlier)：通常指残差非常大的点。模型预测的 y 值与真实的 y 值相差很大，是关于 y 异常的点。
- 高杠杆值点 (High-leverage point)：一般是指关于 x 异常的点，与响应变量 y 没有直接关系。
- 强影响点 (Influnce point)：关于模型异常的点，不一定关于 x 或关于 y 异常。

-离群点 (Outlier)

在残差分析中，我们一般认为残差超过 $\pm 3\hat{\sigma}$ 的点为异常值点。通常使用的残差主要为：

- 普通残差：$e_i = y_i - \hat{y}_i$

- 标准化残差：$ZRE_i = \dfrac{e_i}{\hat{\sigma}}$
- 学生化残差：$SRE_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$，其中 h_{ii} 为帽子矩阵 $H = X(X^TX)'X^T$ 对于的主对角线元素。

然而对于离群点而言，标准残差、学生化残差和普通残差都不在适用，因为离群点会把回归线拉向自身，从而使其本身的残差减小，其余观测的残差增大，回归的标准差 $\hat{\sigma}$ 也相应增大。此时我们采用删除残差来判断。

删除残差构造的主要思想：在计算第 i 个观测值的残差时，我们使用剩余的 $n-1$ 个观测进行线性回归拟合，此时得到的拟合值 $\hat{y}_{(i)}$ 与第 i 个观测值无关，由此定义其删除残差为

$$e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1-h_{ii}}$$

进一步，此时对应的删除学生残差为：

$$SRE_{(i)} = SRE_i \cdot (\frac{n-p-2}{n-p-1-SRE_i^2})^{\frac{1}{2}}$$

一般情况下，当 $|SRE_{(i)}| > 3$ 时，我们判定该点为离群点。

-高杠杆值点 (High-leverage point) 在多元线性回归中，$D(e_i) = (1-h_{ii})\sigma^2$，其中 h_{ii} 为帽子矩阵中主对角线的第 i 个元素，它是调节 e_i 方差大小的杠杆，故我们常称 h_{ii} 为第 i 个观测的杠杆值。同时，h_{ii} 也代表了自变量的第 i 次观测与自变量平均值之间距离的远近。具有较大杠杆值的观测点远离样本中心，能够把回归拉向自身。

一般情况下，我们用杠杆值的平均值作为评判标准，来区分高杠杆值点。由于 $tr(H) = \sum\limits_{i=1}^{n} h_{ii} = p+1$, 所以：

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{p+1}{n}$$

当一个观测点的杠杆值 h_{ii} 大于 2 倍或 3 倍的 $\bar{h}$ 时，我们可以认为该点为高杠杆值点。

-强影响点 (Influnce point) 对于强影响点，我们一般采用库克距离来判别，库克距离的计算公式为：

$$D_i = \frac{\sum\limits_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{(p+1)\hat{\sigma}^2} = \frac{e_i^2}{(p+1)\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$$

通过定义式，我们可以发现库克距离是反映了杠杆值 h_{ii} 和残差 e_i 的综合效应。对于强影响点，我们粗略的判断准则如下：

- 当 $D_i < 0.5$ 时，认为该点不是异常值点。
- 当 $D_i > 1$ 时，认为该点为异常值点。

下面我们画出学生化残差图、cook 距离图与杠杆值倍数图。

图 11: 学生化残差

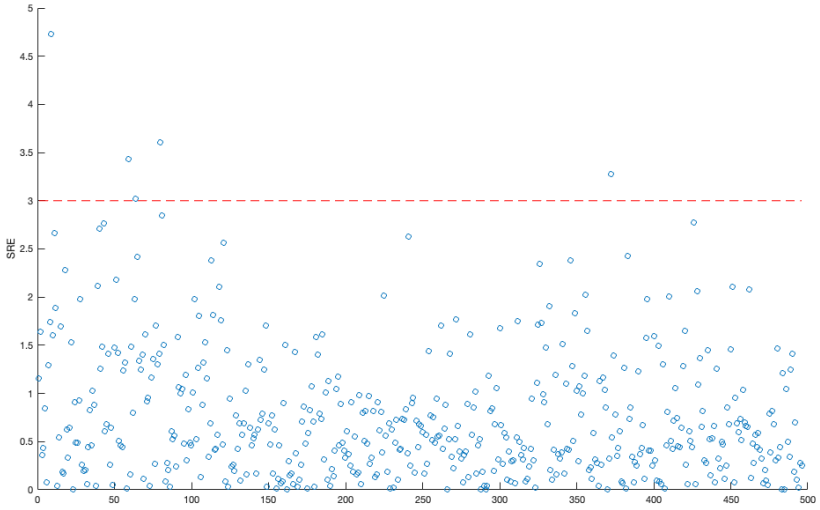


图 12: Cook-Distance

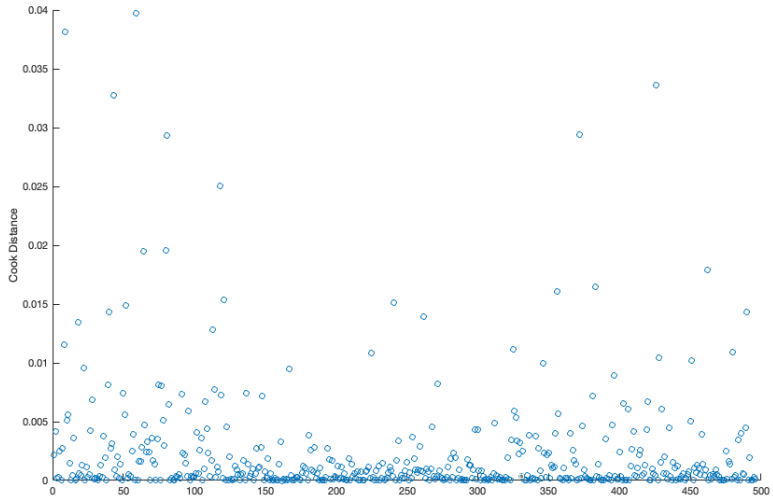
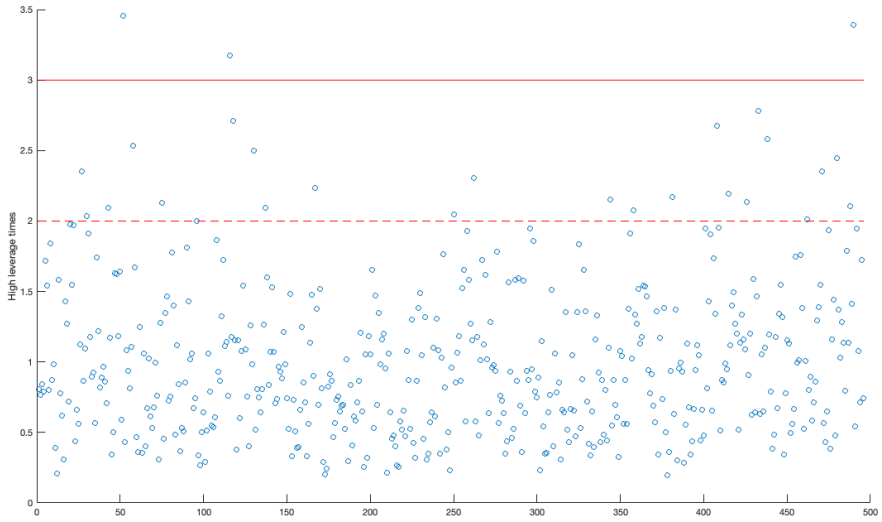


图 13: High-leverage



从上面三张图可以看出，所得回归方程的学生化残差基本落在 [-2,2] 之间，只有少量点落在-4 以外，Cook 距离均在 0.04 以下，没有强影响点存在，但存在部分的高杠杆值点。

5.2 模型解释

从所得回归方程我们可以发现，大学的录取率和个人推荐信的好坏并没有很大的关系，而和本科 GPA 与是否具有科研经历具有密切关系。从而，对于印度的本科生而言，需要更加注重 GPA 和科研方面。

5.3 模型优缺点

优点

1 采用线性回归模型，通俗易懂，便于理解。可以很好的了解各变量对因变量的影响程度。

2 回归拟合效果好，显著很高。

缺点

1 存在一些离群点和高杠杆值点未处理。

2 数据没有划分出训练集与测试集，对于所得方程的应用效果有待检验。

Reference

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

[2] Mazu, Michael J . Regression Analysis by Example[J]. Technometrics, 1977, 35(1):86-87.

[3] Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

A 原始数据

Serial No.	GRE	TOEFL	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4	4.5	8.87	1	0.76
3	316	104	3	3	3.5	8	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.8
5	314	103	2	2	3	8.21	0	0.65
6	330	115	5	4.5	3	9.34	1	0.9
7	321	109	3	3	4	8.2	1	0.75
8	308	101	2	3	4	7.9	0	0.68
9	302	102	1	2	1.5	8	0	0.5
10	323	108	3	3.5	3	8.6	0	0.45
11	325	106	3	3.5	4	8.4	1	0.52
12	327	111	4	4	4.5	9	1	0.84
13	328	112	4	4	4.5	9.1	1	0.78
14	307	109	3	4	3	8	1	0.62
15	311	104	3	3.5	2	8.2	1	0.61
16	314	105	3	3.5	2.5	8.3	0	0.54
17	317	107	3	4	3	8.7	0	0.66
18	319	106	3	4	3	8	1	0.65
19	318	110	3	4	3	8.8	0	0.63
20	303	102	3	3.5	3	8.5	0	0.62
21	312	107	3	3	2	7.9	1	0.64
22	325	114	4	3	2	8.4	0	0.7
23	328	116	5	5	5	9.5	1	0.94
24	334	119	5	5	4.5	9.7	1	0.95
25	336	119	5	4	3.5	9.8	1	0.97
26	340	120	5	4.5	4.5	9.6	1	0.94
27	322	109	5	4.5	3.5	8.8	0	0.76
28	298	98	2	1.5	2.5	7.5	1	0.44

29	295	93	1	2	2	7.2	0	0.46
30	310	99	2	1.5	2	7.3	0	0.54
31	300	97	2	3	3	8.1	1	0.65
32	327	103	3	4	4	8.3	1	0.74
33	338	118	4	3	4.5	9.4	1	0.91
34	340	114	5	4	4	9.6	1	0.9
35	331	112	5	4	5	9.8	1	0.94
36	320	110	5	5	5	9.2	1	0.88
37	299	106	2	4	4	8.4	0	0.64
38	300	105	1	1	2	7.8	0	0.58
39	304	105	1	3	1.5	7.5	0	0.52
40	307	108	2	4	3.5	7.7	0	0.48
41	308	110	3	3.5	3	8	1	0.46
42	316	105	2	2.5	2.5	8.2	1	0.49
43	313	107	2	2.5	2	8.5	1	0.53
44	332	117	4	4.5	4	9.1	0	0.87
45	326	113	5	4.5	4	9.4	1	0.91
46	322	110	5	5	4	9.1	1	0.88
47	329	114	5	4	5	9.3	1	0.86
48	339	119	5	4.5	4	9.7	0	0.89
49	321	110	3	3.5	5	8.85	1	0.82
50	327	111	4	3	4	8.4	1	0.78
51	313	98	3	2.5	4.5	8.3	1	0.76
52	312	100	2	1.5	3.5	7.9	1	0.56
53	334	116	4	4	3	8	1	0.78
54	324	112	4	4	2.5	8.1	1	0.72
55	322	110	3	3	3.5	8	0	0.7
56	320	103	3	3	3	7.7	0	0.64
57	316	102	3	2	3	7.4	0	0.64
58	298	99	2	4	2	7.6	0	0.46

59	300	99	1	3	2	6.8	1	0.36
60	311	104	2	2	2	8.3	0	0.42
61	309	100	2	3	3	8.1	0	0.48
62	307	101	3	4	3	8.2	0	0.47
63	304	105	2	3	3	8.2	1	0.54
64	315	107	2	4	3	8.5	1	0.56
65	325	111	3	3	3.5	8.7	0	0.52
66	325	112	4	3.5	3.5	8.92	0	0.55
67	327	114	3	3	3	9.02	0	0.61
68	316	107	2	3.5	3.5	8.64	1	0.57
69	318	109	3	3.5	4	9.22	1	0.68
70	328	115	4	4.5	4	9.16	1	0.78
71	332	118	5	5	5	9.64	1	0.94
72	336	112	5	5	5	9.76	1	0.96
73	321	111	5	5	5	9.45	1	0.93
74	314	108	4	4.5	4	9.04	1	0.84
75	314	106	3	3	5	8.9	0	0.74
76	329	114	2	2	4	8.56	1	0.72
77	327	112	3	3	3	8.72	1	0.74
78	301	99	2	3	2	8.22	0	0.64
79	296	95	2	3	2	7.54	1	0.44
80	294	93	1	1.5	2	7.36	0	0.46
81	312	105	3	2	3	8.02	1	0.5
82	340	120	4	5	5	9.5	1	0.96
83	320	110	5	5	4.5	9.22	1	0.92
84	322	115	5	4	4.5	9.36	1	0.92
85	340	115	5	4.5	4.5	9.45	1	0.94
86	319	103	4	4.5	3.5	8.66	0	0.76
87	315	106	3	4.5	3.5	8.42	0	0.72
88	317	107	2	3.5	3	8.28	0	0.66

89	314	108	3	4.5	3.5	8.14	0	0.64
90	316	109	4	4.5	3.5	8.76	1	0.74
91	318	106	2	4	4	7.92	1	0.64
92	299	97	3	5	3.5	7.66	0	0.38
93	298	98	2	4	3	8.03	0	0.34
94	301	97	2	3	3	7.88	1	0.44
95	303	99	3	2	2.5	7.66	0	0.36
96	304	100	4	1.5	2.5	7.84	0	0.42
97	306	100	2	3	3	8	0	0.48
98	331	120	3	4	4	8.96	1	0.86
99	332	119	4	5	4.5	9.24	1	0.9
100	323	113	3	4	4	8.88	1	0.79
101	322	107	3	3.5	3.5	8.46	1	0.71
102	312	105	2	2.5	3	8.12	0	0.64
103	314	106	2	4	3.5	8.25	0	0.62
104	317	104	2	4.5	4	8.47	0	0.57
105	326	112	3	3.5	3	9.05	1	0.74
106	316	110	3	4	4.5	8.78	1	0.69
107	329	111	4	4.5	4.5	9.18	1	0.87
108	338	117	4	3.5	4.5	9.46	1	0.91
109	331	116	5	5	5	9.38	1	0.93
110	304	103	5	5	4	8.64	0	0.68
111	305	108	5	3	3	8.48	0	0.61
112	321	109	4	4	4	8.68	1	0.69
113	301	107	3	3.5	3.5	8.34	1	0.62
114	320	110	2	4	3.5	8.56	0	0.72
115	311	105	3	3.5	3	8.45	1	0.59
116	310	106	4	4.5	4.5	9.04	1	0.66
117	299	102	3	4	3.5	8.62	0	0.56
118	290	104	4	2	2.5	7.46	0	0.45

119	296	99	2	3	3.5	7.28	0	0.47
120	327	104	5	3	3.5	8.84	1	0.71
121	335	117	5	5	5	9.56	1	0.94
122	334	119	5	4.5	4.5	9.48	1	0.94
123	310	106	4	1.5	2.5	8.36	0	0.57
124	308	108	3	3.5	3.5	8.22	0	0.61
125	301	106	4	2.5	3	8.47	0	0.57
126	300	100	3	2	3	8.66	1	0.64
127	323	113	3	4	3	9.32	1	0.85
128	319	112	3	2.5	2	8.71	1	0.78
129	326	112	3	3.5	3	9.1	1	0.84
130	333	118	5	5	5	9.35	1	0.92
131	339	114	5	4	4.5	9.76	1	0.96
132	303	105	5	5	4.5	8.65	0	0.77
133	309	105	5	3.5	3.5	8.56	0	0.71
134	323	112	5	4	4.5	8.78	0	0.79
135	333	113	5	4	4	9.28	1	0.89
136	314	109	4	3.5	4	8.77	1	0.82
137	312	103	3	5	4	8.45	0	0.76
138	316	100	2	1.5	3	8.16	1	0.71
139	326	116	2	4.5	3	9.08	1	0.8
140	318	109	1	3.5	3.5	9.12	0	0.78
141	329	110	2	4	3	9.15	1	0.84
142	332	118	2	4.5	3.5	9.36	1	0.9
143	331	115	5	4	3.5	9.44	1	0.92
144	340	120	4	4.5	4	9.92	1	0.97
145	325	112	2	3	3.5	8.96	1	0.8
146	320	113	2	2	2.5	8.64	1	0.81
147	315	105	3	2	2.5	8.48	0	0.75
148	326	114	3	3	3	9.11	1	0.83

149	339	116	4	4	3.5	9.8	1	0.96
150	311	106	2	3.5	3	8.26	1	0.79
151	334	114	4	4	4	9.43	1	0.93
152	332	116	5	5	5	9.28	1	0.94
153	321	112	5	5	5	9.06	1	0.86
154	324	105	3	3	4	8.75	0	0.79
155	326	108	3	3	3.5	8.89	0	0.8
156	312	109	3	3	3	8.69	0	0.77
157	315	105	3	2	2.5	8.34	0	0.7
158	309	104	2	2	2.5	8.26	0	0.65
159	306	106	2	2	2.5	8.14	0	0.61
160	297	100	1	1.5	2	7.9	0	0.52
161	315	103	1	1.5	2	7.86	0	0.57
162	298	99	1	1.5	3	7.46	0	0.53
163	318	109	3	3	3	8.5	0	0.67
164	317	105	3	3.5	3	8.56	0	0.68
165	329	111	4	4.5	4	9.01	1	0.81
166	322	110	5	4.5	4	8.97	0	0.78
167	302	102	3	3.5	5	8.33	0	0.65
168	313	102	3	2	3	8.27	0	0.64
169	293	97	2	2	4	7.8	1	0.64
170	311	99	2	2.5	3	7.98	0	0.65
171	312	101	2	2.5	3.5	8.04	1	0.68
172	334	117	5	4	4.5	9.07	1	0.89
173	322	110	4	4	5	9.13	1	0.86
174	323	113	4	4	4.5	9.23	1	0.89
175	321	111	4	4	4	8.97	1	0.87
176	320	111	4	4.5	3.5	8.87	1	0.85
177	329	119	4	4.5	4.5	9.16	1	0.9
178	319	110	3	3.5	3.5	9.04	0	0.82

179	309	108	3	2.5	3	8.12	0	0.72
180	307	102	3	3	3	8.27	0	0.73
181	300	104	3	3.5	3	8.16	0	0.71
182	305	107	2	2.5	2.5	8.42	0	0.71
183	299	100	2	3	3.5	7.88	0	0.68
184	314	110	3	4	4	8.8	0	0.75
185	316	106	2	2.5	4	8.32	0	0.72
186	327	113	4	4.5	4.5	9.11	1	0.89
187	317	107	3	3.5	3	8.68	1	0.84
188	335	118	5	4.5	3.5	9.44	1	0.93
189	331	115	5	4.5	3.5	9.36	1	0.93
190	324	112	5	5	5	9.08	1	0.88
191	324	111	5	4.5	4	9.16	1	0.9
192	323	110	5	4	5	8.98	1	0.87
193	322	114	5	4.5	4	8.94	1	0.86
194	336	118	5	4.5	5	9.53	1	0.94
195	316	109	3	3.5	3	8.76	0	0.77
196	307	107	2	3	3.5	8.52	1	0.78
197	306	105	2	3	2.5	8.26	0	0.73
198	310	106	2	3.5	2.5	8.33	0	0.73
199	311	104	3	4.5	4.5	8.43	0	0.7
200	313	107	3	4	4.5	8.69	0	0.72
201	317	103	3	2.5	3	8.54	1	0.73
202	315	110	2	3.5	3	8.46	1	0.72
203	340	120	5	4.5	4.5	9.91	1	0.97
204	334	120	5	4	5	9.87	1	0.97
205	298	105	3	3.5	4	8.54	0	0.69
206	295	99	2	2.5	3	7.65	0	0.57
207	315	99	2	3.5	3	7.89	0	0.63
208	310	102	3	3.5	4	8.02	1	0.66

209	305	106	2	3	3	8.16	0	0.64
210	301	104	3	3.5	4	8.12	1	0.68
211	325	108	4	4.5	4	9.06	1	0.79
212	328	110	4	5	4	9.14	1	0.82
213	338	120	4	5	5	9.66	1	0.95
214	333	119	5	5	4.5	9.78	1	0.96
215	331	117	4	4.5	5	9.42	1	0.94
216	330	116	5	5	4.5	9.36	1	0.93
217	322	112	4	4.5	4.5	9.26	1	0.91
218	321	109	4	4	4	9.13	1	0.85
219	324	110	4	3	3.5	8.97	1	0.84
220	312	104	3	3.5	3.5	8.42	0	0.74
221	313	103	3	4	4	8.75	0	0.76
222	316	110	3	3.5	4	8.56	0	0.75
223	324	113	4	4.5	4	8.79	0	0.76
224	308	109	2	3	4	8.45	0	0.71
225	305	105	2	3	2	8.23	0	0.67
226	296	99	2	2.5	2.5	8.03	0	0.61
227	306	110	2	3.5	4	8.45	0	0.63
228	312	110	2	3.5	3	8.53	0	0.64
229	318	112	3	4	3.5	8.67	0	0.71
230	324	111	4	3	3	9.01	1	0.82
231	313	104	3	4	4.5	8.65	0	0.73
232	319	106	3	3.5	2.5	8.33	1	0.74
233	312	107	2	2.5	3.5	8.27	0	0.69
234	304	100	2	2.5	3.5	8.07	0	0.64
235	330	113	5	5	4	9.31	1	0.91
236	326	111	5	4.5	4	9.23	1	0.88
237	325	112	4	4	4.5	9.17	1	0.85
238	329	114	5	4.5	5	9.19	1	0.86

239	310	104	3	2	3.5	8.37	0	0.7
240	299	100	1	1.5	2	7.89	0	0.59
241	296	101	1	2.5	3	7.68	0	0.6
242	317	103	2	2.5	2	8.15	0	0.65
243	324	115	3	3.5	3	8.76	1	0.7
244	325	114	3	3.5	3	9.04	1	0.76
245	314	107	2	2.5	4	8.56	0	0.63
246	328	110	4	4	2.5	9.02	1	0.81
247	316	105	3	3	3.5	8.73	0	0.72
248	311	104	2	2.5	3.5	8.48	0	0.71
249	324	110	3	3.5	4	8.87	1	0.8
250	321	111	3	3.5	4	8.83	1	0.77
251	320	104	3	3	2.5	8.57	1	0.74
252	316	99	2	2.5	3	9	0	0.7
253	318	100	2	2.5	3.5	8.54	1	0.71
254	335	115	4	4.5	4.5	9.68	1	0.93
255	321	114	4	4	5	9.12	0	0.85
256	307	110	4	4	4.5	8.37	0	0.79
257	309	99	3	4	4	8.56	0	0.76
258	324	100	3	4	5	8.64	1	0.78
259	326	102	4	5	5	8.76	1	0.77
260	331	119	4	5	4.5	9.34	1	0.9
261	327	108	5	5	3.5	9.13	1	0.87
262	312	104	3	3.5	4	8.09	0	0.71
263	308	103	2	2.5	4	8.36	1	0.7
264	324	111	3	2.5	1.5	8.79	1	0.7
265	325	110	2	3	2.5	8.76	1	0.75
266	313	102	3	2.5	2.5	8.68	0	0.71
267	312	105	2	2	2.5	8.45	0	0.72
268	314	107	3	3	3.5	8.17	1	0.73

269	327	113	4	4.5	5	9.14	0	0.83
270	308	108	4	4.5	5	8.34	0	0.77
271	306	105	2	2.5	3	8.22	1	0.72
272	299	96	2	1.5	2	7.86	0	0.54
273	294	95	1	1.5	1.5	7.64	0	0.49
274	312	99	1	1	1.5	8.01	1	0.52
275	315	100	1	2	2.5	7.95	0	0.58
276	322	110	3	3.5	3	8.96	1	0.78
277	329	113	5	5	4.5	9.45	1	0.89
278	320	101	2	2.5	3	8.62	0	0.7
279	308	103	2	3	3.5	8.49	0	0.66
280	304	102	2	3	4	8.73	0	0.67
281	311	102	3	4.5	4	8.64	1	0.68
282	317	110	3	4	4.5	9.11	1	0.8
283	312	106	3	4	3.5	8.79	1	0.81
284	321	111	3	2.5	3	8.9	1	0.8
285	340	112	4	5	4.5	9.66	1	0.94
286	331	116	5	4	4	9.26	1	0.93
287	336	118	5	4.5	4	9.19	1	0.92
288	324	114	5	5	4.5	9.08	1	0.89
289	314	104	4	5	5	9.02	0	0.82
290	313	109	3	4	3.5	9	0	0.79
291	307	105	2	2.5	3	7.65	0	0.58
292	300	102	2	1.5	2	7.87	0	0.56
293	302	99	2	1	2	7.97	0	0.56
294	312	98	1	3.5	3	8.18	1	0.64
295	316	101	2	2.5	2	8.32	1	0.61
296	317	100	2	3	2.5	8.57	0	0.68
297	310	107	3	3.5	3.5	8.67	0	0.76
298	320	120	3	4	4.5	9.11	0	0.86

299	330	114	3	4.5	4.5	9.24	1	0.9
300	305	112	3	3	3.5	8.65	0	0.71
301	309	106	2	2.5	2.5	8	0	0.62
302	319	108	2	2.5	3	8.76	0	0.66
303	322	105	2	3	3	8.45	1	0.65
304	323	107	3	3.5	3.5	8.55	1	0.73
305	313	106	2	2.5	2	8.43	0	0.62
306	321	109	3	3.5	3.5	8.8	1	0.74
307	323	110	3	4	3.5	9.1	1	0.79
308	325	112	4	4	4	9	1	0.8
309	312	108	3	3.5	3	8.53	0	0.69
310	308	110	4	3.5	3	8.6	0	0.7
311	320	104	3	3	3.5	8.74	1	0.76
312	328	108	4	4.5	4	9.18	1	0.84
313	311	107	4	4.5	4.5	9	1	0.78
314	301	100	3	3.5	3	8.04	0	0.67
315	305	105	2	3	4	8.13	0	0.66
316	308	104	2	2.5	3	8.07	0	0.65
317	298	101	2	1.5	2	7.86	0	0.54
318	300	99	1	1	2.5	8.01	0	0.58
319	324	111	3	2.5	2	8.8	1	0.79
320	327	113	4	3.5	3	8.69	1	0.8
321	317	106	3	4	3.5	8.5	1	0.75
322	323	104	3	4	4	8.44	1	0.73
323	314	107	2	2.5	4	8.27	0	0.72
324	305	102	2	2	2.5	8.18	0	0.62
325	315	104	3	3	2.5	8.33	0	0.67
326	326	116	3	3.5	4	9.14	1	0.81
327	299	100	3	2	2	8.02	0	0.63
328	295	101	2	2.5	2	7.86	0	0.69

329	324	112	4	4	3.5	8.77	1	0.8
330	297	96	2	2.5	1.5	7.89	0	0.43
331	327	113	3	3.5	3	8.66	1	0.8
332	311	105	2	3	2	8.12	1	0.73
333	308	106	3	3.5	2.5	8.21	1	0.75
334	319	108	3	3	3.5	8.54	1	0.71
335	312	107	4	4.5	4	8.65	1	0.73
336	325	111	4	4	4.5	9.11	1	0.83
337	319	110	3	3	2.5	8.79	0	0.72
338	332	118	5	5	5	9.47	1	0.94
339	323	108	5	4	4	8.74	1	0.81
340	324	107	5	3.5	4	8.66	1	0.81
341	312	107	3	3	3	8.46	1	0.75
342	326	110	3	3.5	3.5	8.76	1	0.79
343	308	106	3	3	3	8.24	0	0.58
344	305	103	2	2.5	3.5	8.13	0	0.59
345	295	96	2	1.5	2	7.34	0	0.47
346	316	98	1	1.5	2	7.43	0	0.49
347	304	97	2	1.5	2	7.64	0	0.47
348	299	94	1	1	1	7.34	0	0.42
349	302	99	1	2	2	7.25	0	0.57
350	313	101	3	2.5	3	8.04	0	0.62
351	318	107	3	3	3.5	8.27	1	0.74
352	325	110	4	3.5	4	8.67	1	0.73
353	303	100	2	3	3.5	8.06	1	0.64
354	300	102	3	3.5	2.5	8.17	0	0.63
355	297	98	2	2.5	3	7.67	0	0.59
356	317	106	2	2	3.5	8.12	0	0.73
357	327	109	3	3.5	4	8.77	1	0.79
358	301	104	2	3.5	3.5	7.89	1	0.68

359	314	105	2	2.5	2	7.64	0	0.7
360	321	107	2	2	1.5	8.44	0	0.81
361	322	110	3	4	5	8.64	1	0.85
362	334	116	4	4	3.5	9.54	1	0.93
363	338	115	5	4.5	5	9.23	1	0.91
364	306	103	2	2.5	3	8.36	0	0.69
365	313	102	3	3.5	4	8.9	1	0.77
366	330	114	4	4.5	3	9.17	1	0.86
367	320	104	3	3.5	4.5	8.34	1	0.74
368	311	98	1	1	2.5	7.46	0	0.57
369	298	92	1	2	2	7.88	0	0.51
370	301	98	1	2	3	8.03	1	0.67
371	310	103	2	2.5	2.5	8.24	0	0.72
372	324	110	3	3.5	3	9.22	1	0.89
373	336	119	4	4.5	4	9.62	1	0.95
374	321	109	3	3	3	8.54	1	0.79
375	315	105	2	2	2.5	7.65	0	0.39
376	304	101	2	2	2.5	7.66	0	0.38
377	297	96	2	2.5	2	7.43	0	0.34
378	290	100	1	1.5	2	7.56	0	0.47
379	303	98	1	2	2.5	7.65	0	0.56
380	311	99	1	2.5	3	8.43	1	0.71
381	322	104	3	3.5	4	8.84	1	0.78
382	319	105	3	3	3.5	8.67	1	0.73
383	324	110	4	4.5	4	9.15	1	0.82
384	300	100	3	3	3.5	8.26	0	0.62
385	340	113	4	5	5	9.74	1	0.96
386	335	117	5	5	5	9.82	1	0.96
387	302	101	2	2.5	3.5	7.96	0	0.46
388	307	105	2	2	3.5	8.1	0	0.53

389	296	97	2	1.5	2	7.8	0	0.49
390	320	108	3	3.5	4	8.44	1	0.76
391	314	102	2	2	2.5	8.24	0	0.64
392	318	106	3	2	3	8.65	0	0.71
393	326	112	4	4	3.5	9.12	1	0.84
394	317	104	2	3	3	8.76	0	0.77
395	329	111	4	4.5	4	9.23	1	0.89
396	324	110	3	3.5	3.5	9.04	1	0.82
397	325	107	3	3	3.5	9.11	1	0.84
398	330	116	4	5	4.5	9.45	1	0.91
399	312	103	3	3.5	4	8.78	0	0.67
400	333	117	4	5	4	9.66	1	0.95
401	304	100	2	3.5	3	8.22	0	0.63
402	315	105	2	3	3	8.34	0	0.66
403	324	109	3	3.5	3	8.94	1	0.78
404	330	116	4	4	3.5	9.23	1	0.91
405	311	101	3	2	2.5	7.64	1	0.62
406	302	99	3	2.5	3	7.45	0	0.52
407	322	103	4	3	2.5	8.02	1	0.61
408	298	100	3	2.5	4	7.95	1	0.58
409	297	101	3	2	4	7.67	1	0.57
410	300	98	1	2	2.5	8.02	0	0.61
411	301	96	1	3	4	7.56	0	0.54
412	313	94	2	2.5	1.5	8.13	0	0.56
413	314	102	4	2.5	2	7.88	1	0.59
414	317	101	3	3	2	7.94	1	0.49
415	321	110	4	3.5	4	8.35	1	0.72
416	327	106	4	4	4.5	8.75	1	0.76
417	315	104	3	4	2.5	8.1	0	0.65
418	316	103	3	3.5	2	7.68	0	0.52

419	309	111	2	2.5	4	8.03	0	0.6
420	308	102	2	2	3.5	7.98	1	0.58
421	299	100	3	2	3	7.42	0	0.42
422	321	112	3	3	4.5	8.95	1	0.77
423	322	112	4	3.5	2.5	9.02	1	0.73
424	334	119	5	4.5	5	9.54	1	0.94
425	325	114	5	4	5	9.46	1	0.91
426	323	111	5	4	5	9.86	1	0.92
427	312	106	3	3	5	8.57	0	0.71
428	310	101	3	3.5	5	8.65	1	0.71
429	316	103	2	2	4.5	8.74	0	0.69
430	340	115	5	5	4.5	9.06	1	0.95
431	311	104	3	4	3.5	8.13	1	0.74
432	320	112	2	3.5	3.5	8.78	1	0.73
433	324	112	4	4.5	4	9.22	1	0.86
434	316	111	4	4	5	8.54	0	0.71
435	306	103	3	3.5	3	8.21	0	0.64
436	309	105	2	2.5	4	7.68	0	0.55
437	310	110	1	1.5	4	7.23	1	0.58
438	317	106	1	1.5	3.5	7.65	1	0.61
439	318	110	1	2.5	3.5	8.54	1	0.67
440	312	105	2	1.5	3	8.46	0	0.66
441	305	104	2	2.5	1.5	7.79	0	0.53
442	332	112	1	1.5	3	8.66	1	0.79
443	331	116	4	4.5	4.5	9.44	1	0.92
444	321	114	5	4.5	4.5	9.16	1	0.87
445	324	113	5	4	5	9.25	1	0.92
446	328	116	5	4.5	5	9.08	1	0.91
447	327	118	4	5	5	9.67	1	0.93
448	320	108	3	3.5	5	8.97	1	0.84

449	312	109	2	2.5	4	9.02	0	0.8
450	315	101	3	3.5	4.5	9.13	0	0.79
451	320	112	4	3	4.5	8.86	1	0.82
452	324	113	4	4.5	4.5	9.25	1	0.89
453	328	116	4	5	3.5	9.6	1	0.93
454	319	103	3	2.5	4	8.76	1	0.73
455	310	105	2	3	3.5	8.01	0	0.71
456	305	102	2	1.5	2.5	7.64	0	0.59
457	299	100	2	2	2	7.88	0	0.51
458	295	99	1	2	1.5	7.57	0	0.37
459	312	100	1	3	3	8.53	1	0.69
460	329	113	4	4	3.5	9.36	1	0.89
461	319	105	4	4	4.5	8.66	1	0.77
462	301	102	3	2.5	2	8.13	1	0.68
463	307	105	4	3	3	7.94	0	0.62
464	304	107	3	3.5	3	7.86	0	0.57
465	298	97	2	2	3	7.21	0	0.45
466	305	96	4	3	4.5	8.26	0	0.54
467	314	99	4	3.5	4.5	8.73	1	0.71
468	318	101	5	3.5	5	8.78	1	0.78
469	323	110	4	4	5	8.88	1	0.81
470	326	114	4	4	3.5	9.16	1	0.86
471	320	110	5	4	4	9.27	1	0.87
472	311	103	3	2	4	8.09	0	0.64
473	327	116	4	4	4.5	9.48	1	0.9
474	316	102	2	4	3.5	8.15	0	0.67
475	308	105	4	3	2.5	7.95	1	0.67
476	300	101	3	3.5	2.5	7.88	0	0.59
477	304	104	3	2.5	2	8.12	0	0.62
478	309	105	4	3.5	2	8.18	0	0.65

479	318	103	3	4	4.5	8.49	1	0.71
480	325	110	4	4.5	4	8.96	1	0.79
481	321	102	3	3.5	4	9.01	1	0.8
482	323	107	4	3	2.5	8.48	1	0.78
483	328	113	4	4	2.5	8.77	1	0.83
484	304	103	5	5	3	7.92	0	0.71
485	317	106	3	3.5	3	7.89	1	0.73
486	311	101	2	2.5	3.5	8.34	1	0.7
487	319	102	3	2.5	2.5	8.37	0	0.68
488	327	115	4	3.5	4	9.14	0	0.79
489	322	112	3	3	4	8.62	1	0.76
490	302	110	3	4	4.5	8.5	0	0.65
491	307	105	2	2.5	4.5	8.12	1	0.67
492	297	99	4	3	3.5	7.81	0	0.54
493	298	101	4	2.5	4.5	7.69	1	0.53
494	300	95	2	3	1.5	8.22	1	0.62
495	301	99	3	2.5	2	8.45	1	0.68
496	332	108	5	4.5	4	9.02	1	0.87
497	337	117	5	5	5	9.87	1	0.96
498	330	120	5	4.5	5	9.56	1	0.93
499	312	103	4	4	5	8.43	0	0.73
500	327	113	4	4.5	4.5	9.04	0	0.84

B Rcode

```
format long;
clc,clearvars;
data = readtable("Data/Admission_Predict_2.csv",
... VariableNamingRule="preserve");
data = data(:,2:9);
```

```
name = {'GRE Score', 'TOFEL Score' ,'University Rating',
          ... 'SOP','LOR','CGPA','Research','Chance of Admit'};
temp = data{:,:};
% 判断数据是否有空值
sum(isnan(temp),'all');
% 绘制自变量的直方图
subplot(1,4,1);
histogram(Data =data.("GRE Score"),NumBins=10);
title('GRE.Score');
ylabel('Frequency');
subplot(1,4,2);
histogram(Data =data.("TOEFL Score"),NumBins=10);
title('TOEFL.Score');
ylabel('Frequency');
subplot(1,4,3);
histogram(Data =data.("University Rating"));
title('University.Rating');
ylabel('Frequency');
subplot(1,4,4);
histogram(data.("SOP"));
title('SOP');
ylabel('Frequency');
subplot(1,3,1)
histogram(Data =data.("LOR"));
title('LOR');
ylabel('Frequency');
subplot(1,3,2)
histogram(Data =data.("CGPA"));
title('CGPA');
ylabel('Frequency');
subplot(1,3,3)
histogram(Data =data.("Chance of Admit"));
```

```
38 title('Chance of Admit');
39 ylabel('Frequency');
40 % 各个变量的基本统计量
41 minvalue = min(temp);
42 maxvalue = max(temp);
43 averagevalue = mean(temp);
44 medianvalue = median(temp);
45 % 绘制各变量的箱形图
46 subplot(1,4,1);
47 boxchart(data.("GRE Score"),Notch="on");
48 title('GRE.Score');
49 subplot(1,4,2);
50 boxchart(data.("TOEFL Score"),Notch="on");
51 title('TOEFL.Score');
52 subplot(1,4,3);
53 boxchart(data.("University Rating"),Notch="on");
54 title('University.Rating');
55 subplot(1,4,4);
56 boxchart(data.("SOP"),Notch="on");
57 title('SOP');
58 subplot(1,3,1)
59 boxchart(data.("LOR"),Notch="on");
60 title('LOR');
61 subplot(1,3,2)
62 boxchart(data.("CGPA"),Notch="on");
63 title('CGPA');
64 subplot(1,3,3)
65 boxchart(data.("Chance of Admit"),Notch="on");
66 title('Chance of Admit');
67 % 删除异常值
68 data([93,348,377],:)=[];
69 close all;
```

```
% 绘制各个变量间的相关热力图
subplot(1,1,1);
corr_matrix = corr(temp);
heatmap(corr_matrix,"ColorbarVisible","on","ColorData",corr_matrix,
... "Colormap",parula,XDisplayLabels=name,YDisplayLabels=name);

% 建立线性回归模型
x=table2array(data(:,1:7));
y=table2array(data(:,8));
n=length(y);
b = ones(n,1);
x_zeros = [b,x];
[beta,r2,adjr2,F,Ftest,t,ttest,residuals] = myregression(x,y);
% 逐步回归
stepwiselm(x,y,Criterion="aic",Upper="linear");

% Lasso 回归
[B,FitInfo] = lasso(x,y,'CV',10,'Alpha',1);
lassoPlot(B,FitInfo, 'PlotType', 'CV');
legend('show') % 显示图例
% 筛选稀疏变量
idxLambda1SE = FitInfo.Index1SE;
coef = B(:,idxLambda1SE);%回归系数
coef0 = FitInfo.Intercept(idxLambda1SE);%常系数
lanmda=FitInfo.LambdaMinMSE;

% elastic net
[B,FitInfo] = lasso(x,y,'CV',10,'Alpha',0.75);
lassoPlot(B,FitInfo, 'PlotType', 'CV');
legend('show') % 显示图例
% 筛选稀疏变量
idxLambda1SE = FitInfo.Index1SE;
```

```
coef = B(:,idxLambda1SE);%回归系数
coef0 = FitInfo.Intercept(idxLambda1SE);%常系数
lanmda=FitInfo.LambdaMinMSE;
close all;
% 异方差检验
% 画Y与残差的散点图
x = x(:,[1,2,3,5,6,7]);
[beta,r2,adjr2,F,Ftest,t,ttest,residuals] = myregression(x,y);
scatter(y,residuals);
xlabel("Y");
ylabel("Residuals");
close all;
% spearman检验
[rho,tvalue,pvalue]=spearmantest(x(:,1),residuals);
% white检验
[W,pvalue] = whitetest(x,residuals);
% Box–Cox 变换
% 求对数似然下最大的lambda值
lambda = -5:0.1:5;
SSE_lam=arrayfun(@(t) (bocc(x_zeros,y,t)),lambda);
[value,index] = min(SSE_lam);
lambda_min=lambda(index);
SSE_min = SSE_lam(index);
plot(lambda,SSE_lam,'g',lambda_min,SSE_min,'r+');
xlabel('\lambda');ylabel('残差平方和');
legend('不同\lambda下SSE值','\lambda的最优解');
close all;
% 做 BOX-COX 变换
y = boxcox(lambda_min,y);
% 重新拟合
[beta,r2,adjr2,F,Ftest,t,ttest,residuals] = myregression(x,y);
% 画Y与残差的散点图
```

```
scatter(y,residuals);
xlabel("Y");
ylabel("Residuals");
close all;
% spearman检验
[rho,tvalue,pvalue]=spearmantest(x(:,1),residuals);
% white检验
[W,pvalue] = whitetest(x,residuals);

% 自相关检验
% 残差时序图
scatter(residuals(1:end-1),residuals(2:end));
xlabel("Residual_{t-1}");
ylabel("Residual_t");
close all;
%DW检验
[rho_hat,DW]=dwtest(residuals);
%wls变换
newx = x(2:end,:)-rho_hat.*x(1:end-1,:);
newy = y(2:end,:)-rho_hat.*y(1:end-1,:);
[beta,r2,adjr2,F,Ftest,t,ttest,residuals] = myregression(newx,newy);
scatter(residuals(1:end-1),residuals(2:end));
xlabel("Residual_{t-1}");
ylabel("Residual_t");
close all;
%DW检验
[rho_hat,DW]=dwtest(residuals);

% 共线性检验
diagvalue = vif(newx);
condvalue = condvaluecal(newx);

```

```
166 % 删除学生残差
167 [sre,delsre] =sre(newx,residuals);
168 scatter(1:size(delsre,2),abs(delsre));
169 hold on;
170 plot(1:size(delsre,2),3*ones(size(delsre,2),2),'r--');
171 ylabel("SRE")
172 close all;
173 % cook distance
174 D = cook(newx,residuals);
175 n= size(D,1);
176 scatter(1:n,D);
177 ylabel("Cook Distance")
178 close all;
179 % 杠杆值点
180 [h,h_mean,times] = High_leverage(newx,residuals);
181 scatter(1:n,times);
182 ylabel("High leverage times")
183 hold on;
184 plot(1:n,2*ones(n,1),'r--',1:n,3*ones(n,1),'r-');
185 % close all
186
187
188
189 ————functions————
190
191
192
193 function SSE=bocc(X,Y,lambda)
194 H=X*inv(X'*X)*X';
195 n=length(Y);
196 switch lambda
197 case 0
```

```
198 z=log(Y)*prod(Y)^(1/n);
199 otherwise
200 z=(Y.^lambda-1)/lambda/(prod(Y)^((lambda-1)/n));
201 end
202 SSE=z'*(eye(n)-H)*z;
203 end
204
205
206
207
208 function [condvalue] = condvaluecal(x)
209 x = corr(x,type="Pearson");
210 eigvalue = eig(x);
211 eigmax = max(eigvalue);
212 condvalue = sqrt(eigmax./eigvalue);
213 end
214
215
216
217
218 function [D] = cook(x,residual)
219 n = size(x,1);
220 m = size(x,2);
221 b = ones(size(x,1),1);
222 x = [b,x];
223 H = x*inv(x'*x)*x';
224 h = diag(H);
225 sse=residual'*residual;
226 sigma=sqrt(sse/(n-m-1));
227 D = zeros(n,1);
228 for i=1:n
229 D(i) = ((residual(i)^2)/((m+1)*sigma^2))*(h(i)/((1-h(i))^2));
```

```
230 end
231 end
232
233
234
235
236 function [rho_hat,DW] = dwtest(residual)
237 residual1 = residual(1:end-1,1);
238 residual2 = residual(2:end,1);
239 DW = ((residual2-residual1)'*(residual2-residual1))/(residual'*residual);
240 rho_hat =1-DW/2;
241 end
242
243
244
245
246 function [h,h_mean,times] = High_leverage(x,residual)
247 n = size(x,1);
248 m = size(x,2);
249 b = ones(size(x,1),1);
250 x = [b,x];
251 H = x*inv(x'*x)*x';
252 h = diag(H);
253 h_mean = (m+1)/n;
254 times = h./h_mean;
255 end
256
257
258
259
260 function [beta,r2,adjr2,F,Ftest,t,ttest,residuals] = myregression(x,y)
261 n = size(x,1);
```

```
m = size(x,2);
b = ones(n,1);
x = [b,x];
beta = (x'*x)\(x'*y);
ttest = zeros(1,m+1);
sst=(y-mean(y))'*(y-mean(y));
sse=(y-x*beta)'*(y-x*beta);
ssr=sst-sse;
F=(ssr/m)/(sse/(n-m-1));
r2=ssr/sst;
adjr2=1-(n-1)/(n-m-1)*(1-r2);
Ftest=1-fcdf(F,m,n-m-1);
c=diag(inv(x'*x));
t=zeros(1,m+1);
sigma=sqrt(sse/(n-m-1));
for i=1:m+1
t(i)=beta(i)/(sqrt(c(i))*sigma);
ttest(i)=2*(1-tcdf(abs(t(i)),n-m-1));
end
residuals = y-x*beta;



function [rho,tvalue,pvalue] = spearmantest(x,residual)
n = size(x,1);
[rho,p] = corr(x,abs(residual),"type","Spearman");
tvalue = (sqrt(n-2)*rho)/sqrt(1-rho^2);
pvalue = 2*(1-tcdf(abs(tvalue),n-2));
end


```

```
294
295
296 function [sre,delsre] = sre(x,residual)
297 n = size(x,1);
298 m = size(x,2);
299 b = ones(size(x,1),1);
300 x = [b,x];
301 H = x*inv(x'*x)*x';
302 h = diag(H);
303 sse=residual'*residual;
304 sigma=sqrt(sse/(n-m-1));
305 sre = zeros(size(x,1),1);
306 for i=1:size(x,1)
307 sre(i) = residual(i)/(sigma*sqrt(1-h(i)));
308 delsre(i) = sre(i)*sqrt((n-m-2)/(n-m-1-sre(i)^2));
309 end
310 end
311
312
313
314
315 function [diagvalue] = vif(x)
316 x = corr(x,"type","Pearson");
317 diagvalue = diag(inv(x));
318 end
319
320
321
322
323 function [W,pvalue] = whitetest(x,residual)
324 n=size(x,1);
325 m=size(x,2);
```

```
326 res2 = residual.^2;
327 xtest =x;
328 for i =1:m
329 xtest = [xtest,x(:,i).^2];
330 end
331 [betat,r2t,adjr2t,Ft,Ftestt,tt,ttestt,residualst] = myregression(xtest,res2);
332 W=n*r2t;
333 pvalue = 1-chi2cdf(W,size(xtest,2));
334 end
```