回归分析期中作业

姓名：蔡佺佑
学号：201928017515019

题目：进行相应的线性回归分析，确立各因素影响，预测右眼视力状况

目的：掌握线性模型的基本方法，熟练运用R的lm，并能够对结果进行相应的分析

具体实施：

1. 数据预处理
   将原始数据"2019年期中考试数据说明.xlsx"另存为"2019年期中考试数据说明.csv"；在R中，使用read.csv()函数来加载用于分析的数据，如下：

```
trainning_data<-read.csv(file = '2019年回归分析期中考试数据.csv',header = 
T,encoding = 'UTF-8',stringsAsFactors = TRUE)
dim(trainning_data)
```

   读入原始数据集并查看数据集，数据集结构为2094个样本数据，每个样本有72个变量，变量所代表的内容记录在"2019年期中考试数据说明.xlsx"中。

   读入数据后，对原始数据的缺失值进行处理，将其中缺失值比例超过50%的变量提取并删除：

```
#缺失值比例计算
missing_data<-as.matrix(apply(trainning_data,2, function(x)
{sum(is.na(x))/length(x)*100}))
#找出缺失值大于50%的特征，删除该特征
miss_feature<-names(which(missing_data[,1]>=50))
clean_data<-trainning_data[,-match(miss_feature,names(trainning_data))]
```

   处理后共计删除4个变量：CHGLASS1; DRWBASE; NRWBASE; JNGLASS，剩余68个变量；对于缺失值未超过50%的变量，考虑到若存在较大偏离的异常值，平均数不能反应大部分数据特征，故而采取用中位数代替的方法。

```
clean_data=apply(clean_data,2,function(x){
  x[is.na(x)]=mean(x,na.rm = T)
 return(x)
})
clean_data<-as.data.frame(clean_data)
#这里也找到简单的Hmisc包里面的impute实现同样的功能
for(i in seq(1,68)){clean_data[,i]<-impute(clean_data[,i],median)}
```

2. 模型设计
   假设我们要预测的右眼视力与其他因素存在一定的线性关系，以读入的数据中心右眼视力为因变量，以其他变量为自变量拟合模型，即表示为：$Y = X\beta + e$的形式，为了能够解释最小二乘法的系数，这里假设数据满足以下统计假设：

   1. 数据存在线性关系；
   2. 符合高斯马尔科夫假设。

      用lm()函数拟合线性回归模型：

```
lm_1<-
lm(RA~GENDER+DRWITHGL+NRWITHGL+RAXISLEG+RANTECHA+RWTW+RNCONPRE+HEIGHT+WEIGHT
+TOUW+SBLOPRE+DBLOPRE+PULSE+CHAOD1+ETEST+TRT+AMB11+AMB21+AMB31+AMB41+DRUGIF+
PREGQ+SGAR+LSMK+COSTM+BULBP+TUTOR1+TUTOR2+NTON+MSMK+HSMK+MH+FH+FW+BED+IFTAI+
BULB+TUTOR+CELLP+NOONS+EYE+EYE1+SITE+STR+PLACE5+SHIFT+IFDOU+ALLERGY+DRNBASE+
NRNBASE+DOMEYE+RPR+JTR+YTR+RCORCU+DAI1+DAI2+MB+ACI+BCI+CCI+DCI+total+st1+st2
+INCM+PREG,data= clean_data)
```

查看初步建立的模型

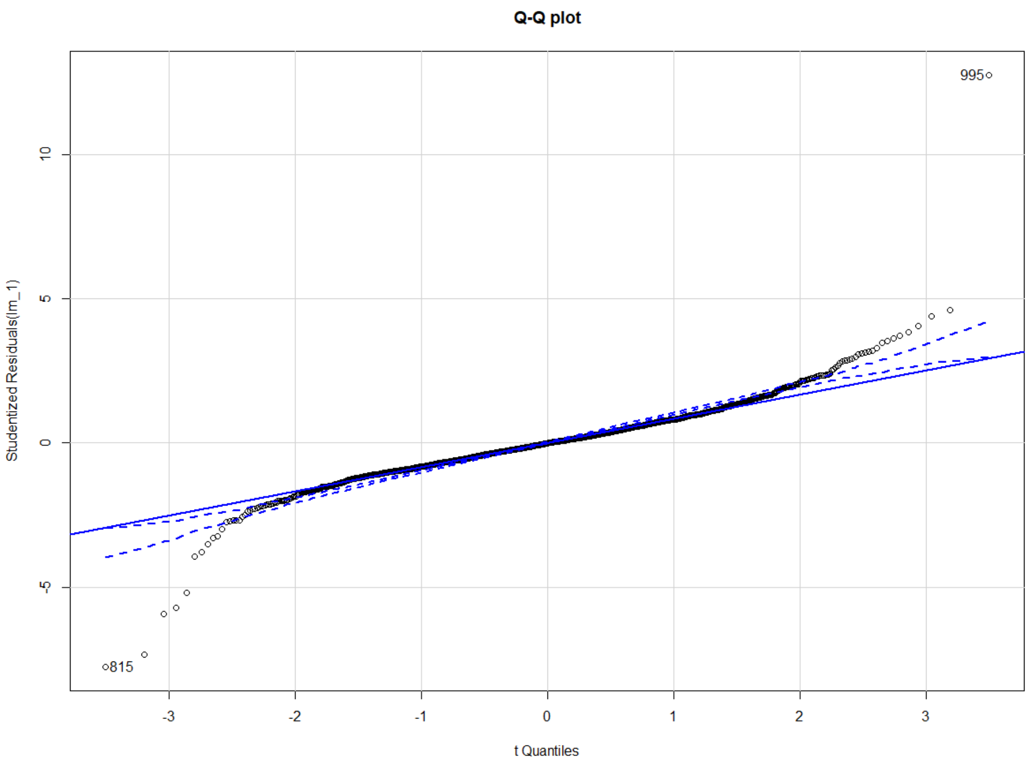
```
> summary(lm_1)

Call:
lm(formula = RA ~ GENDER + DRWITHGL + NRWITHGL + RAXISLEG +
    RANTECHA + RWTW + RNCONPRE + HEIGHT + WEIGHT + TOUW + SBLOPRE +
    DBLOPRE + PULSE + CHAOD1 + ETEST + TRT + AMB11 + AMB21 +
    AMB31 + AMB41 + DRUGIF + PREGQ + SGAR + LSMK + COSTM + BULBP +
    TUTOR1 + TUTOR2 + NTON + MSMK + HSMK + MH + FH + FW + BED +
    IFTAI + BULB + TUTOR + CELLP + NOONS + EYE + EYE1 + SITE +
    STR + PLACE5 + SHIFT + IFDOU + ALLERGY + DRNBASE + NRNBASE +
    DOMEYE + RPR + JTR + YTR + RCORCU + DAI1 + DAI2 + MB + ACI +
    BCI + CCI + DCI + total + st1 + st2 + INCM + PREG, data = clean_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6229 -0.3479 -0.0067  0.3241  7.2541

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.294e+01  1.555e+00  27.620  < 2e-16 ***
GENDER       2.039e-01  3.173e-02   6.428 1.61e-10 ***
DRWITHGL    -2.015e-01  1.991e-01  -1.012  0.31180
NRWITHGL    -1.112e-01  2.000e-01  -0.556  0.57824
RAXISLEG    -1.050e+00  3.421e-02 -30.683  < 2e-16 ***
RANTECHA     4.871e-01  7.541e-02   6.460 1.31e-10 ***
RWTW         5.836e-02  3.920e-02   1.489  0.13667
RNCONPRE    -1.302e-02  4.162e-03  -3.129  0.00178 **
HEIGHT       1.746e-03  1.945e-03   0.898  0.36953
WEIGHT      -2.774e-05  1.527e-04  -0.182  0.85583
TOUW         3.052e-04  6.926e-03   0.044  0.96485
SBLOPRE     -3.461e-04  1.645e-03  -0.210  0.83342
DBLOPRE      1.775e-03  1.876e-03   0.946  0.34415
PULSE        8.557e-05  1.239e-03   0.069  0.94497
CHAOD1      -1.601e+00  2.572e-01  -6.223 5.92e-10 ***
ETEST       -1.334e-02  1.954e-02  -0.682  0.49505
TRT         -5.988e-02  4.250e-02  -1.409  0.15901
AMB11        5.233e-02  4.240e-02   1.234  0.21727
AMB21       -1.158e-02  5.976e-02  -0.194  0.84642
AMB31        3.040e-03  5.889e-02   0.052  0.95883
AMB41        1.223e-01  7.148e-02   1.711  0.08717 .
DRUGIF       3.794e-03  2.801e-02   0.135  0.89226
PREGQ       -6.065e-02  6.912e-02  -0.877  0.38038
SGAR         1.464e-01  1.573e-01   0.931  0.35186
LSMK        -1.212e-02  2.445e-02  -0.496  0.62030
COSTM        2.178e-02  2.155e-02   1.011  0.31220
BULBP       -2.024e-02  2.951e-02  -0.686  0.49300
```

```
TUTOR1        1.474e-02  2.689e-02   0.548  0.58369    
TUTOR2        6.928e-03  2.748e-02   0.252  0.80098    
NTON          1.992e-04  2.462e-02   0.008  0.99355    
MSMK          8.456e-02  1.952e-01   0.433  0.66490    
HSMK          1.327e-02  3.030e-02   0.438  0.66134    
MH            4.050e-04  1.987e-03   0.204  0.83850    
FH           -3.426e-03  1.710e-03  -2.004  0.04525 *  
FW            2.038e-03  1.280e-03   1.592  0.11160    
BED          -1.207e-02  1.773e-02  -0.680  0.49635    
IFTAI        -2.246e-02  3.330e-02  -0.674  0.50018    
BULB         -6.761e-03  3.284e-02  -0.206  0.83691    
TUTOR         2.992e-02  2.975e-02   1.006  0.31471    
CELLP         5.133e-02  2.917e-02   1.760  0.07856 .  
NOONS         2.889e-03  1.832e-02   0.158  0.87471    
EYE          -4.364e-02  3.943e-02  -1.107  0.26845    
EYE1         -2.784e-02  1.979e-02  -1.407  0.15967    
SITE          8.375e-03  2.514e-02   0.333  0.73906    
STR           3.397e-02  2.157e-02   1.574  0.11555    
PLACE5       -1.779e-03  2.069e-02  -0.086  0.93149    
SHIFT         3.858e-02  3.005e-02   1.284  0.19930    
IFDOU         1.031e-02  3.164e-02   0.326  0.74448    
ALLERGY       4.233e-03  1.790e-02   0.237  0.81306    
DRNBASE      -7.070e-01  7.333e-02  -9.642  < 2e-16 ***
NRNBASE      -4.115e-01  7.022e-02  -5.860 5.38e-09 ***
DOMEYE       -4.416e-02  2.686e-02  -1.644  0.10033    
RPR           3.602e-01  1.379e-02  26.126  < 2e-16 ***
JTR           5.565e-03  2.242e-02   0.248  0.80396    
YTR           1.676e-01  1.766e-02   9.492  < 2e-16 ***
RCORCU       -4.464e-01  1.688e-02 -26.441  < 2e-16 ***
DAI1         -7.540e-02  3.965e-02  -1.901  0.05740 .  
DAI2         -5.863e-02  3.990e-02  -1.469  0.14187    
MB            5.702e-03  3.775e-03   1.510  0.13115    
ACI          -4.965e-04  1.066e-03  -0.466  0.64148    
BCI           2.323e-04  1.288e-03   0.180  0.85688    
CCI          -2.778e-04  6.782e-04  -0.410  0.68215    
DCI           1.792e-04  8.525e-04   0.210  0.83349    
total         1.625e-02  9.500e-03   1.710  0.08738 .  
st1           1.653e-02  1.634e-02   1.012  0.31179    
st2           1.403e-03  7.946e-03   0.177  0.85986    
INCM         -2.398e-02  2.172e-02  -1.104  0.26971    
PREG          1.455e-03  7.121e-04   2.043  0.04121 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6193 on 2026 degrees of freedom
Multiple R-squared:  0.9112,	Adjusted R-squared:  0.9082 
F-statistic: 310.1 on 67 and 2026 DF,  p-value: < 2.2e-16
```

可以看出与右眼视力影响显著的变量并不多，大部分变量的相关性较低，这说明，当控制其它预测变量不变时，那些不显著的变量与因变量存在不显著的线性相关关系。后面将会基于此对模型进一步优化。由Multiple R-squared: 0.9112可知，所以预测变量解释了91.12%的方差。由Residual standard error: 0.6193知估计标准误差为0.6193，说明用以上变量来估计时，平均的估计误差为0.6193。

- 正态性假设

```
qqPlot(lm_1,id.method='identify',simulate =
TRUE,labels=row.names(RA),main='Q-Q plot')
```



可以看到大部分点都在直线附近，并基本落在置信区间内，这表明正态性假设基本符合。

- 独立性假设

```
> durbinWatsonTest(lm_1)
 lag Autocorrelation D-W Statistic p-value
   1     -0.01473945      2.028863   0.582
 Alternative hypothesis: rho != 0
```

进行D-W检验：P=0.582>0.05不显著，说明因变量之间无自相关性，互相独立。

- 同方差性假设

```
> ncvTest(lm_1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 10.60838, Df = 1, p = 0.0011258
```

p值说明目前拟合的线性模型不符合同方差性假设。

3. 选择最佳模型
选择最终的预测变量有以下两种使用较多的方法：逐步回归法、全子集回归。最终获取一个回归方程时，实际上就是从众多可能的模型选择最佳的一个。该回归方程能最好的平衡预测精度和模型简洁度的关系。如果两个模型预测精度没有显著差别，则选择包含预测变量较少的模型。

- 1. 逐步回归法

```
step(lm_1, direction = "backward")
```

选择AIC最小的结果（即最后的结果）

```
Step:  AIC=-2005.75
RA ~ GENDER + DRWITHGL + RAXISLEG + RANTECHA + RNCONPRE + 
    CHAOD1 + AMB41 + FH + FW + CELLP + STR + DRNBASE + NRNBASE + 
    DOMEYE + RPR + YTR + RCORCU + DAI1 + DAI2 + MB + total + 
    PREG

           Df Sum of Sq     RSS     AIC
<none>                   786.04 -2005.8
- MB        1      0.79  786.83 -2005.6
- DOMEYE    1      0.85  786.89 -2005.5
- STR       1      1.00  787.04 -2005.1
- FW        1      1.01  787.05 -2005.1
- DAI2      1      1.03  787.07 -2005.0
- total     1      1.12  787.16 -2004.8
- FH        1      1.44  787.48 -2003.9
- CELLP     1      1.49  787.53 -2003.8
- PREG      1      1.78  787.82 -2003.0
- DAI1      1      2.41  788.45 -2001.3
- AMB41     1      4.00  790.04 -1997.1
- RNCONPRE  1      4.40  790.44 -1996.1
- NRNBASE   1     13.60  799.64 -1971.8
- CHAOD1    1     15.24  801.28 -1967.5
- GENDER    1     18.57  804.61 -1958.8
- DRWITHGL  1     20.01  806.05 -1955.1
- RANTECHA  1     21.08  807.12 -1952.3
- DRNBASE   1     37.86  823.90 -1909.2
- YTR       1     41.25  827.29 -1900.7
- RPR       1    269.66 1055.70 -1390.1
- RCORCU    1    301.71 1087.75 -1327.5
- RAXISLEG  1    373.74 1159.78 -1193.2

Call:
lm(formula = RA ~ GENDER + DRWITHGL + RAXISLEG + RANTECHA + 
    RNCONPRE + CHAOD1 + AMB41 + FH + FW + CELLP + STR + DRNBASE + 
    NRNBASE + DOMEYE + RPR + YTR + RCORCU + DAI1 + DAI2 + MB + 
    total + PREG, data = clean_data)

Coefficients:
(Intercept)       GENDER     DRWITHGL     RAXISLEG     RANTECHA     
RNCONPRE       CHAOD1  
  44.428492     0.209517    -0.316275    -1.051908     0.522612    
-0.013780    -1.609074  
      AMB41           FH           FW        CELLP          STR     
DRNBASE      NRNBASE  
   0.148423    -0.003167     0.002031     0.055765     0.031463    
-0.714418    -0.411098  
     DOMEYE          RPR          YTR       RCORCU         DAI1     
    DAI2           MB  
  -0.039570     0.360612     0.168415    -0.452840    -0.095614    
-0.061638     0.005345  
      total         PREG  
   0.015119     0.001519  
```

再次进行模型拟合：

```
lm_2<-lm(RA~ GENDER + DRWITHGL + RAXISLEG + RANTECHA + RNCONPRE +
               CHAOD1 + AMB41 + FH + FW + CELLP + STR + DRNBASE +
NRNBASE +
               DOMEYE + RPR + YTR + RCORCU + DAI1 + DAI2 + MB +
total +
               PREG,data= clean_data)
summary(lm_2)
```

得到如下结果:

```
Call:
lm(formula = RA ~ GENDER + DRWITHGL + RAXISLEG + RANTECHA +
    RNCONPRE + CHAOD1 + AMB41 + FH + FW + CELLP + STR + DRNBASE +
    NRNBASE + DOMEYE + RPR + YTR + RCORCU + DAI1 + DAI2 + MB +
    total + PREG, data = clean_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6771 -0.3519 -0.0053  0.3198  7.2892

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.4284917  1.3263544  33.497  < 2e-16 ***
GENDER       0.2095166  0.0299517   6.995 3.56e-12 ***
DRWITHGL    -0.3162748  0.0435621  -7.260 5.44e-13 ***
RAXISLEG    -1.0519076  0.0335216 -31.380  < 2e-16 ***
RANTECHA     0.5226119  0.0701190   7.453 1.33e-13 ***
RNCONPRE    -0.0137799  0.0040462  -3.406 0.000673 ***
CHAOD1      -1.6090737  0.2539243  -6.337 2.87e-10 ***
AMB41        0.1484225  0.0457171   3.247 0.001187 **
FH          -0.0031671  0.0016245  -1.950 0.051356 .
FW           0.0020310  0.0012432   1.634 0.102481
CELLP        0.0557649  0.0281358   1.982 0.047612 *
STR          0.0314633  0.0193369   1.627 0.103866
DRNBASE     -0.7144183  0.0715301  -9.988  < 2e-16 ***
NRNBASE     -0.4110981  0.0686800  -5.986 2.53e-09 ***
DOMEYE      -0.0395699  0.0264440  -1.496 0.134710
RPR          0.3606121  0.0135290  26.655  < 2e-16 ***
YTR          0.1684145  0.0161546  10.425  < 2e-16 ***
RCORCU      -0.4528402  0.0160613 -28.194  < 2e-16 ***
DAI1        -0.0956139  0.0379317  -2.521 0.011787 *
DAI2        -0.0616375  0.0374461  -1.646 0.099909 .
MB           0.0053449  0.0036994   1.445 0.148661
total        0.0151188  0.0088045   1.717 0.086097 .
PREG         0.0015192  0.0007012   2.167 0.030379 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6161 on 2071 degrees of freedom
Multiple R-squared:  0.9101,    Adjusted R-squared:  0.9092
F-statistic: 953.5 on 22 and 2071 DF,  p-value: < 2.2e-16
```

从结果中可以看出仍有显著性不高的变量，将这些变量删去后再进行拟合：

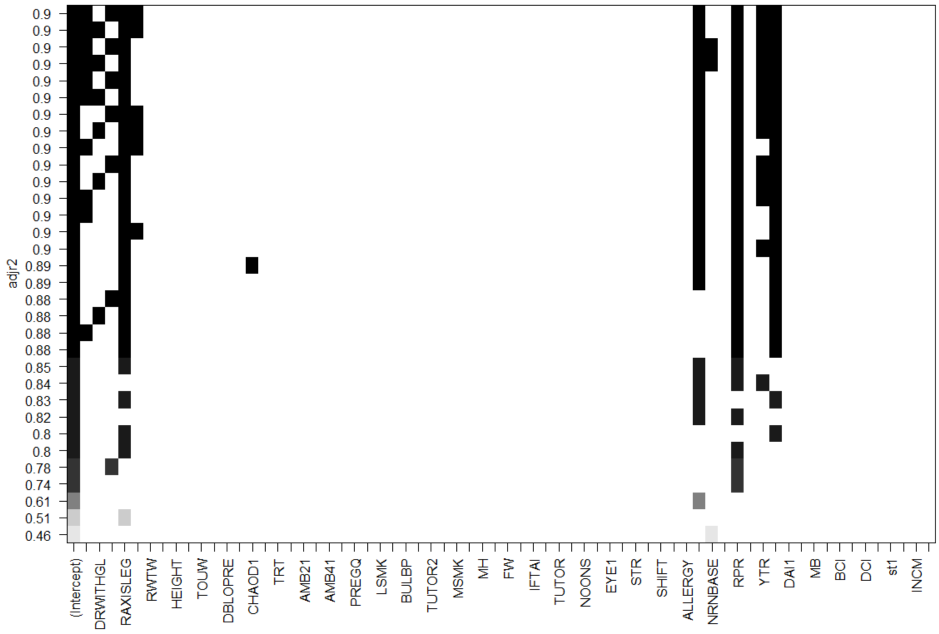
```
lm_3<-lm(RA ~ GENDER + DRWITHGL + RAXISLEG + RANTECHA + CHAOD1 +
               AMB41 + CELLP + DRNBASE + NRNBASE +
               RPR + YTR + RCORCU + DAI1 + PREG,data=clean_data)
summary(lm_3)
```

得到如下结果:

```
Call:
lm(formula = RA ~ GENDER + DRWITHGL + RAXISLEG + RANTECHA +
    CHAOD1 + AMB41 + CELLP + DRNBASE + NRNBASE + RPR + YTR +
    RCORCU + DAI1 + PREG, data = clean_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7369 -0.3524 -0.0081  0.3217  7.4629

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.1293125  1.2944345  34.092  < 2e-16 ***
GENDER       0.2181295  0.0298534   7.307 3.88e-13 ***
DRWITHGL    -0.3170487  0.0437382  -7.249 5.90e-13 ***
RAXISLEG    -1.0550180  0.0335841 -31.414  < 2e-16 ***
RANTECHA     0.5056479  0.0703930   7.183 9.44e-13 ***
CHAOD1      -1.6060539  0.2548034  -6.303 3.55e-10 ***
AMB41        0.1501098  0.0457401   3.282  0.00105 **
CELLP        0.0671354  0.0275722   2.435  0.01498 *
DRNBASE     -0.7157988  0.0717533  -9.976  < 2e-16 ***
NRNBASE     -0.4135288  0.0687067  -6.019 2.07e-09 ***
RPR          0.3616669  0.0135474  26.696  < 2e-16 ***
YTR          0.1709473  0.0162258  10.536  < 2e-16 ***
RCORCU      -0.4543807  0.0160955 -28.230  < 2e-16 ***
DAI1        -0.1051687  0.0375896  -2.798  0.00519 **
PREG         0.0016395  0.0007031   2.332  0.01980 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6193 on 2079 degrees of freedom
Multiple R-squared:  0.9089,    Adjusted R-squared:  0.9082
F-statistic:  1481 on 14 and 2079 DF,  p-value: < 2.2e-16
```

2. 全子集回归法

```
lm_sub<-
regsubsets(RA~GENDER+DRWITHGL+NRWITHGL+RAXISLEG+RANTECHA+RWTW+RNCON
PRE+HEIGHT+WEIGHT+TOUW+SBLOPRE+DBLOPRE+PULSE+CHAOD1+ETEST+TRT+AMB11
+AMB21+AMB31+AMB41+DRUGIF+PREGQ+SGAR+LSMK+COSTM+BULBP+TUTOR1+TUTOR2
+NTON+MSMK+HSMK+MH+FH+FW+BED+IFTAI+BULB+TUTOR+CELLP+NOONS+EYE+EYE1+
SITE+STR+PLACE5+SHIFT+IFDOU+ALLERGY+DRNBASE+NRNBASE+DOMEYE+RPR+JTR+
YTR+RCORCU+DAI1+DAI2+MB+ACI+BCI+CCI+DCI+total+st1+st2+INCM+PREG,dat
a=clean_data,really.big=T,nbest=4)
plot(lm_sub,scale="adjr2")
```



选择最大adjr2模型，进行拟合：

```
> lm_4<-lm(RA~GENDER+NRWITHGL+RAXISLEG+RANTECHA+DRNBASE+RPR+YTR+
RCORCU,data = clean_data)
> summary(lm_4)

Call:
lm(formula = RA ~ GENDER + NRWITHGL + RAXISLEG + RANTECHA + 
    DRNBASE + RPR + YTR + RCORCU, data = clean_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4670 -0.3633 -0.0146  0.3228  7.7264 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 44.79002    1.28247  34.925  < 2e-16 ***
GENDER       0.23209    0.03037   7.643 3.22e-14 ***
NRWITHGL    -0.35336    0.04438  -7.962 2.75e-15 ***
RAXISLEG    -1.09423    0.03391 -32.266  < 2e-16 ***
RANTECHA     0.52100    0.07141   7.296 4.20e-13 ***
DRNBASE     -0.76076    0.07092 -10.726  < 2e-16 ***
RPR          0.38916    0.01324  29.382  < 2e-16 ***
YTR          0.15916    0.01624   9.800  < 2e-16 ***
RCORCU      -0.46953    0.01631 -28.791  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6322 on 2085 degrees of freedom
Multiple R-squared:  0.9047,    Adjusted R-squared:  0.9044 
F-statistic:  2475 on 8 and 2085 DF,  p-value: < 2.2e-16
```

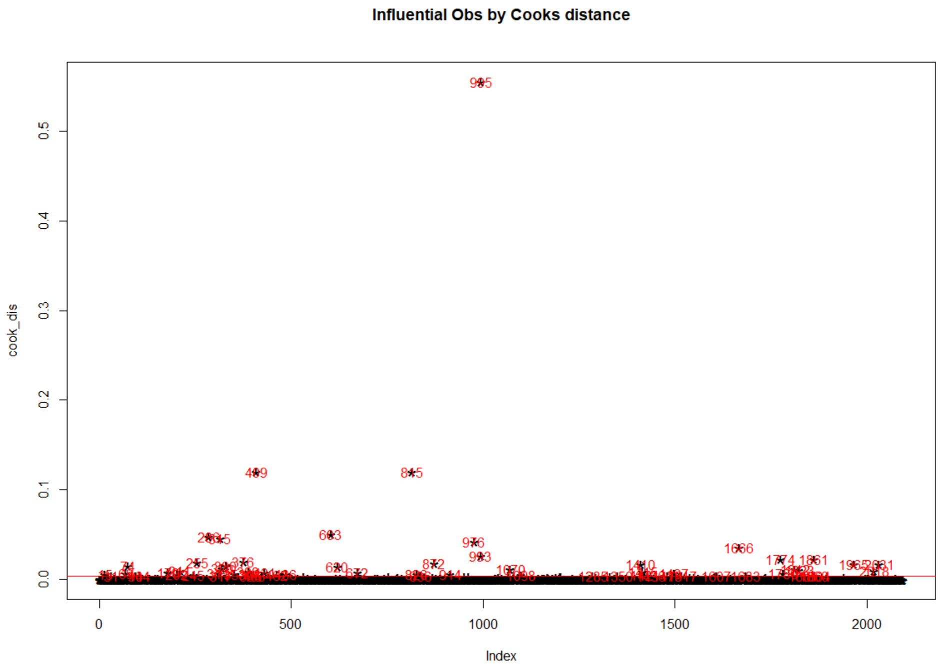
虽然采用该模型将使得调整R平方有所降低，但各预测变量的回归系数却都在统计学意义上变得显著。综上比较全子集回归法得到的模型变量更少，故优先考虑lm_4这个模型：RA~GENDER+NRWITHGL+RAXISLEG+RANTECHA+DRNBASE+RPR+YTR+RCORCU。

3. 异常值检验：

```
outlierTest(lm_4)
yichang<-c(995,409,815,872,286,315,603,2018,993)
c1_data<-clean_data[-yichang,]
cook_dis <- cooks.distance(lm_4)
plot(cook_dis, pch="*", cex=2, main="Influential Obs by Cooks
distance")
abline(h = 4*mean(cook_dis, na.rm=T), col="red")
text(x=1:length(cook_dis)+1, y=cook_dis,
labels=ifelse(cook_dis>4*mean(cook_dis,
na.rm=T),names(cook_dis),""), col="red")
influential<-which(cook_dis>4*mean(cook_dis, na.rm=T))
c2_data<-c1_data[-influential,]
```

此处借鉴了从网上找到的处理异常值的方法：

首先，先用car包的outlierTest函数找出部分离群点，删除后再找强影响点，强影响点是那种若删除则模型的系数会产生明显的变化的点。一种方法是计算Cook距离，一般来说， Cook's D值大于4/(n-k -1)，则表明它是强影响点，其中n 为样本量大小， k 是预测变量数目。图中红色虚线以上就返回了强影响点。

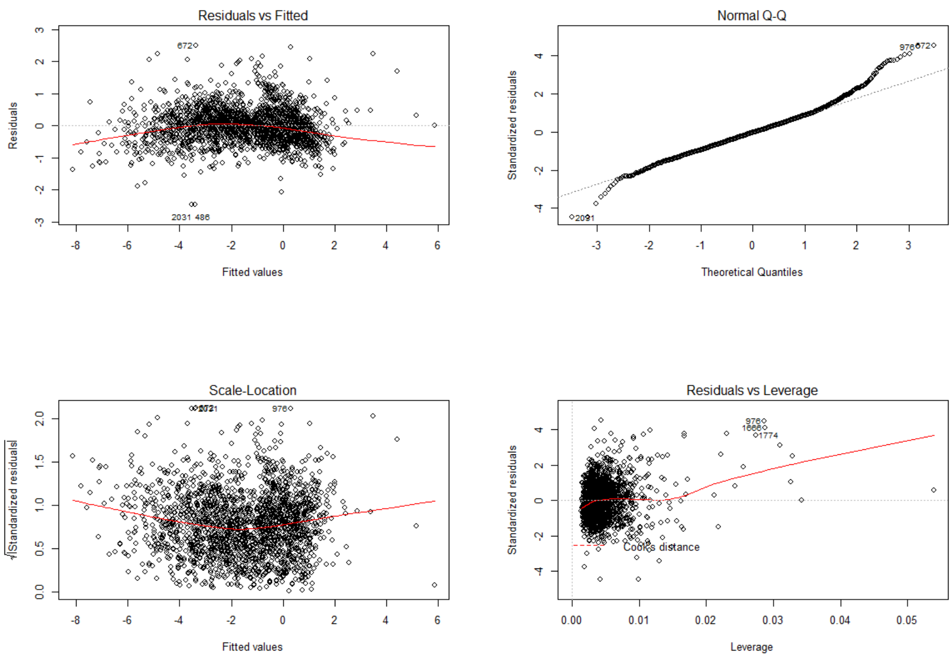


去除异常值后对数据重新拟合模型：

```
> lm_5<-lm(RA~GENDER+NRWITHGL+RAXISLEG+RANTECHA+DRNBASE+RPR+YTR+
RCORCU,data = c2_data)
> summary(lm_5)

Call:
lm(formula = RA ~ GENDER + NRWITHGL + RAXISLEG + RANTECHA +
    DRNBASE + RPR + YTR + RCORCU, data = c2_data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.46886 -0.35372 -0.01285  0.31400  2.50479 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.09678    1.18557  33.821  < 2e-16 ***
GENDER       0.19295    0.02710   7.119 1.50e-12 ***
NRWITHGL    -0.39598    0.03973  -9.966  < 2e-16 ***
RAXISLEG    -0.96422    0.03140 -30.706  < 2e-16 ***
RANTECHA     0.48627    0.06409   7.587 4.98e-14 ***
DRNBASE     -0.85477    0.06364 -13.432  < 2e-16 ***
RPR          0.40219    0.01251  32.139  < 2e-16 ***
YTR          0.19240    0.01456  13.213  < 2e-16 ***
RCORCU      -0.42594    0.01495 -28.495  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5533 on 2013 degrees of freedom
Multiple R-squared:  0.9221,	Adjusted R-squared:  0.9218 
F-statistic:  2977 on 8 and 2013 DF,  p-value: < 2.2e-16
```

4. 对选择的模型重新进行前面所做的假设检验



1.残差和拟合值之间数据点均匀分布在y=0两侧，呈现出随机的分布，红色线呈现出一条平稳的曲线并没有明显的形状特征。

2.残差Q-Q图，数据点按对角直线排列，趋于一条直线，并被对角直接穿过，但右边界的点都不在直线上，大体符合正态分布。

3.标准化残差平方根和拟合值，水平线周围的点呈随机分布，说明满足同方差性假设。

4.标准化残差和杠杆值(右下)，没有出现红色的等高线，则说明数据中没有特别影响回归结果的异常点。

4. 预测未知数据

通过之前建立的模型lm_5对测试数据集进行视力预测：

```
test_data<-read.csv(file = '测试集.CSV')
yucezhi = predict(lm_5, newdata = test_data)
test_data$RA_预测=yucezhi
for(i in seq(1,8)){
  if(test_data$RA_预测[i]>0.5){
    test_data$视力[i]<-"远视"
  }
  else if(test_data$RA_预测[i]<(-0.5)){
    test_data$视力[i]<-"近视"
  } else{
    test_data$视力[i]<-"正常"
  }
}
test_data[,c(72,73)]
write.csv(test_data,"预测结果.csv")
```

结果如下:

```
    RA_yuce shili
1 -3.320847  近视
2 -3.511353  近视
3 -0.129761  正常
4 -3.939994  近视
5 -3.140864  近视
6 -2.194960  近视
7 -3.813365  近视
8  4.641970  远视
```

本次环境:

```
> sessionInfo()
R version 3.6.1 (2019-07-05)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18362)

Matrix products: default

locale:
[1] LC_COLLATE=Chinese (Simplified)_China.936  LC_CTYPE=Chinese
(Simplified)_China.936
[3] LC_MONETARY=Chinese (Simplified)_China.936 LC_NUMERIC=C

[5] LC_TIME=Chinese (Simplified)_China.936

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] leaps_3.0         car_3.0-3         carData_3.0-2     MASS_7.3-51.4
[5] Hmisc_4.3-0       ggplot2_3.2.0     Formula_1.2-3     survival_2.44-1.1
[9] lattice_0.20-38

loaded via a namespace (and not attached):
```

```
 [1] tidyselect_0.2.5    xfun_0.8            purrr_0.3.2         splines_3.6.1

 [5] haven_2.1.1         colorspace_1.4-1    vctrs_0.2.0         htmltools_0.3.6

 [9] base64enc_0.1-3     rlang_0.4.0         pillar_1.4.2        foreign_0.8-71

[13] glue_1.3.1          withr_2.1.2         RColorBrewer_1.1-2  readxl_1.3.1

[17] stringr_1.4.0       cellranger_1.1.0    munsell_0.5.0       gtable_0.3.0

[21] zip_2.0.3           htmlwidgets_1.3     latticeExtra_0.6-28 knitr_1.23

[25] rio_0.5.16          forcats_0.4.0       curl_4.0
htmlTable_1.13.1
[29] Rcpp_1.0.2          acepack_1.4.1       backports_1.1.4     scales_1.0.0

[33] checkmate_1.9.4     abind_1.4-5         gridExtra_2.3       hms_0.5.0

[37] digest_0.6.20       openxlsx_4.1.0.1    stringi_1.4.3       dplyr_0.8.3

[41] grid_3.6.1          tools_3.6.1         magrittr_1.5        lazyeval_0.2.2

[45] tibble_2.1.3        cluster_2.1.0       zeallot_0.1.0       crayon_1.3.4

[49] pkgconfig_2.0.2     Matrix_1.2-17       data.table_1.12.2
assertthat_0.2.1
[53] rstudioapi_0.10     R6_2.4.0            rpart_4.1-15        nnet_7.3-12

[57] compiler_3.6.1
```

完整代码：

```
rm(list = ls())
Sys.setenv(R_MAX_NUM_DLLS=999)
options(stringsAsFactors = F)
library(leaps)
library(Hmisc)
library(MASS)
library(car)
#读取数据
trainning_data<-read.csv(file = '2019年回归分析期中考试数据.csv',header = T,encoding
= 'UTF-8',stringsAsFactors = TRUE)
#缺失值处理
missing_data<-as.matrix(apply(trainning_data,2, function(x)
{sum(is.na(x))/length(x)*100}))
#找出缺失值大于50%的特征，删除该特征
miss_feature<-names(which(missing_data[,1]>=50))
clean_data<-trainning_data[,-match(miss_feature,names(trainning_data))]
for(i in seq(1,68)){clean_data[,i]<-impute(clean_data[,i],median)}
lm_1<-
lm(RA~GENDER+DRWITHGL+NRWITHGL+RAXISLEG+RANTECHA+RWTW+RNCONPRE+HEIGHT+WEIGHT+TOU
W+SBLOPRE+DBLOPRE+PULSE+CHAOD1+ETEST+TRT+AMB11+AMB21+AMB31+AMB41+DRUGIF+PREGQ+SG
AR+LSMK+COSTM+BULBP+TUTOR1+TUTOR2+NTON+MSMK+HSMK+MH+FH+FW+BED+IFTAI+BULB+TUTOR+C
ELLP+NOONS+EYE+EYE1+SITE+STR+PLACE5+SHIFT+IFDOU+ALLERGY+DRNBASE+NRNBASE+DOMEYE+R
PR+JTR+YTR+RCORCU+DAI1+DAI2+MB+ACI+BCI+CCI+DCI+total+st1 +st2+INCM+PREG,data=
clean_data)
#假设检验
```

```
qqPlot(lm_1,id.method='identify',simulate = TRUE,labels=row.names(RA),main='Q-Q plot')
durbinWatsonTest(lm_1)
ncvTest(lm_1)
#par(mfrow=c(2,2))
#plot(lm_1)
#逐步向后回归
step(lm_1, direction = "backward")
#MASS包的stepAIC函数进行逐步向后回归
##stepAIC(lm_1, direction = "backward")
#选取AIC最小的模型结果
lm_2<-lm(RA~ GENDER + DRWITHGL + RAXISLEG + RANTECHA + RNCONPRE + CHAOD1 + AMB41
+ FH + FW + CELLP + STR + DRNBASE + NRNBASE +DOMEYE + RPR + YTR + RCORCU + DAI1
+ DAI2 + MB + total +PREG,data= clean_data)
summary(lm_2)
lm_3<-lm(RA ~ GENDER + DRWITHGL + RAXISLEG + RANTECHA + CHAOD1 +AMB41 + CELLP +
DRNBASE + NRNBASE + RPR + YTR + RCORCU + DAI1 + PREG,data=clean_data)
summary(lm_3)
#全子集回归法
lm_sub<-
regsubsets(RA~GENDER+DRWITHGL+NRWITHGL+RAXISLEG+RANTECHA+RWTW+RNCONPRE+HEIGHT+WEIGHT+TOUW+SBLOPRE+DBLOPRE+PULSE+CHAOD1+ETEST+TRT+AMB11+AMB21+AMB31+AMB41+DRUGIF+PREGQ+SGAR+LSMK+COSTM+BULBP+TUTOR1+TUTOR2+NTON+MSMK+HSMK+MH+FH+FW+BED+IFTAI+BULB+TUTOR+CELLP+NOONS+EYE+EYE1+SITE+STR+PLACE5+SHIFT+IFDOU+ALLERGY+DRNBASE+NRNBASE+DOMEYE+RPR+JTR+YTR+RCORCU+DAI1+
DAI2+MB+ACI+BCI+CCI+DCI+total+st1+st2+INCM+PREG,data=clean_data,really.big=T,nbest=4)
plot(lm_sub,scale="adjr2")
#选最大adjr2拟合
lm_4<-lm(RA~GENDER+NRWITHGL+RAXISLEG+RANTECHA+DRNBASE+RPR+YTR+ RCORCU,data =
clean_data)
summary(lm_4)
#异常值检验
outlierTest(lm_4)
yichang<-c(995,409,815,872,286,315,603,2018,993)
c1_data<-clean_data[-yichang,]
cook_dis <- cooks.distance(lm_4)
plot(cook_dis, pch="*", cex=2, main="Influential Obs by Cooks distance")
abline(h = 4*mean(cook_dis, na.rm=T), col="red")
text(x=1:length(cook_dis)+1, y=cook_dis, labels=ifelse(cook_dis>4*mean(cook_dis,
na.rm=T),names(cook_dis),""), col="red")
influential<-which(cook_dis>4*mean(cook_dis, na.rm=T))
c2_data<-c1_data[-influential,]
#重新拟合
lm_5<-lm(RA~GENDER+NRWITHGL+RAXISLEG+RANTECHA+DRNBASE+RPR+YTR+ RCORCU,data =
c2_data)
summary(lm_5)
#对所选模型进行假设检验
par(mfrow=c(2,2))
plot(lm_5)
#测试数据集预测
test_data<-read.csv(file = '测试集.CSV')
yucezhi = predict(lm_5, newdata = test_data)
test_data$RA_预测=yucezhi
for(i in seq(1,8)){
  if(test_data$RA_预测[i]>0.5){
    test_data$视力[i]<-"远视"
  }
```

```
  else if(test_data$RA_预测[i]<(-0.5)){
    test_data$视力[i]<-"近视"
  } else{
    test_data$视力[i]<-"正常"
  }
}
test_data[,c(72,73)]
write.csv(test_data,"预测结果.csv")

sessionInfo()
save.image(file = "mid_test.RData")
```