2020-12-20

回归分析大作业

——关于国家税收的影响因素的回归分析报告

1807403021

邵云卿

18 统计

目录

一、问题背景

税收是国家（政府）最主要的收入形式和来源，对于国家经济生活和社会文明的重要作用，因此探究与税收相关的因素具有重要的作用。

本文选取中国国家统计局 1990 年-2019 年的统计数据，探究财政税收与国民总收入、国内生产总值、贸易进出口总额、总人口数、全社会固定资产投资、公路运输客运量及就业人数的关系。其中，自变量为国民总收入、国内生产总值、贸易进出口总额、总人口数、全社会固定资产投资、公路运输客运量及就业人数，均为连续型变量，因变量为财政税收。

二、变量说明

（1）符号说明：

变量	含义
Y	税收（亿元）
X1	国民总收入（亿元）
X2	国内生产总值（亿元）
X3	货物进出口总额（亿元）
X4	总人口（万人）
X5	全社会固定资产投资（亿元）
X6	公路运输客运量(万人)
X7	就业人数（万人）

（2）名词解释：

全社会固定资产投资：以货币形式表现的建造和购置固定资产活动的工作量。

问题求解

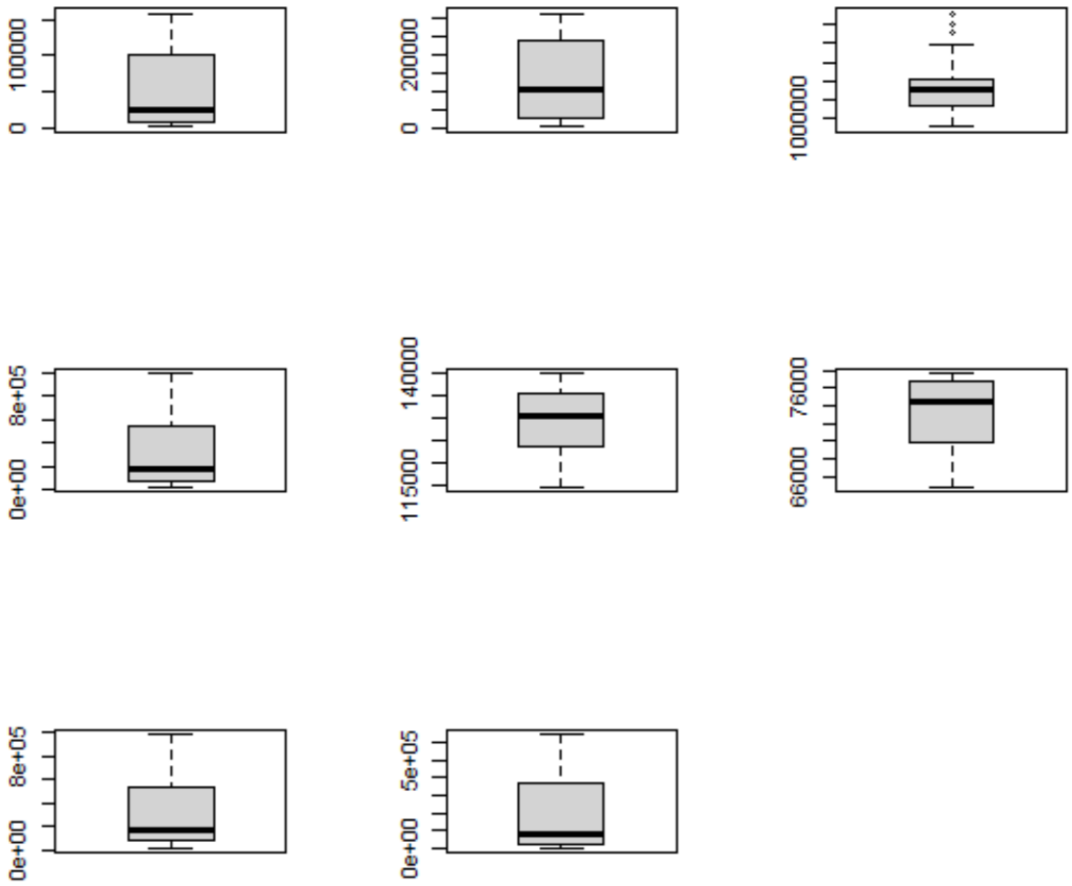
(1)数据的描述性分析

	Variables							
.function	Y	X1	X2	X3	X4	X5	X6	X7
All	All	All	All	All	All	All	All	All
max	158,000.460	988,528.900	990,865.110	315,627.320	140,005.000	645,675.000	3,804,035.000	77,640.000
mean	52,653.219	311,155.366	312,344.942	127,715.742	129,245.600	202,885.620	1,832,240.433	73,107.700
median	26,472.110	173,707.170	174,579.530	106,230.430	130,372.000	79,625.500	1,763,944.500	74,455.500
min	2,821.860	18,923.330	18,872.870	5,560.120	114,333.000	4,517.000	772,682.000	64,749.000
p25	8,491.230	80,056.535	81,085.158	26,879.070	123,909.750	25,807.375	1,339,249.750	70,024.250
p75	97,895.307	523,844.955	525,920.007	241,640.332	135,236.750	358,892.300	2,030,203.000	76,633.000
sd	53,294.657	297,780.718	298,592.416	107,245.018	7,531.881	227,652.151	780,669.372	4,157.763
se	9,730.229	54,367.072	54,515.267	19,580.172	1,375.127	41,563.406	142,530.075	759.100
var	2,840,320,494.306	88,673,355,809.325	89,157,430,706.871	11,501,493,800.008	56,729,238.455	51,825,501,781.339	609,444,668,871.219	17,286,989.390

R 代码：

```
> a<-read.csv('data.csv',head=TRUE)
> Y<-a[,1]
> X1<-a[,2]
> X2<-a[,3]
> X3<-a[,4]
> X4<-a[,5]
> X5<-a[,6]
> X6<-a[,7]
> X7<-a[,8]
> boxplot(Y)
> par(mfrow=c(2,3))
> boxplot(X1)
> boxplot(X2)
> boxplot(X3)
> boxplot(X4)
> boxplot(X5)
> boxplot(X6)
> boxplot(X7)
```

箱线图：



（2）建立多元线性回归模型

a. 采用普通最小二乘法（OLSE）建立Y关于 X1、X2、X3、X4、X5、X6、X7 的回归方程， 并对得到的回归方程进行假设检验。

R 代码：

```
> fit<-lm(Y~.,data=a)
> summary(fit)
```

结果：

```
Call:
lm(formula = Y ~ ., data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-2413.7  -599.6   138.3   520.5  2341.5
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.260e+04  1.562e+04   5.288 2.63e-05 ***
X1          -3.493e-01  1.780e-01  -1.962  0.06257 .
X2           4.706e-01  1.770e-01   2.658  0.01436 *
X3           6.184e-02  1.664e-02   3.717  0.00120 **
X4          -2.927e+00  8.188e-01  -3.575  0.00169 **
X5           7.055e-02  6.325e-03  11.155 1.59e-10 ***
X6           1.838e-03  6.055e-04   3.036  0.00606 **
X7           3.891e+00  1.267e+00   3.070  0.00560 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1260 on 22 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9994
F-statistic:  7410 on 7 and 22 DF,  p-value: < 2.2e-16
```

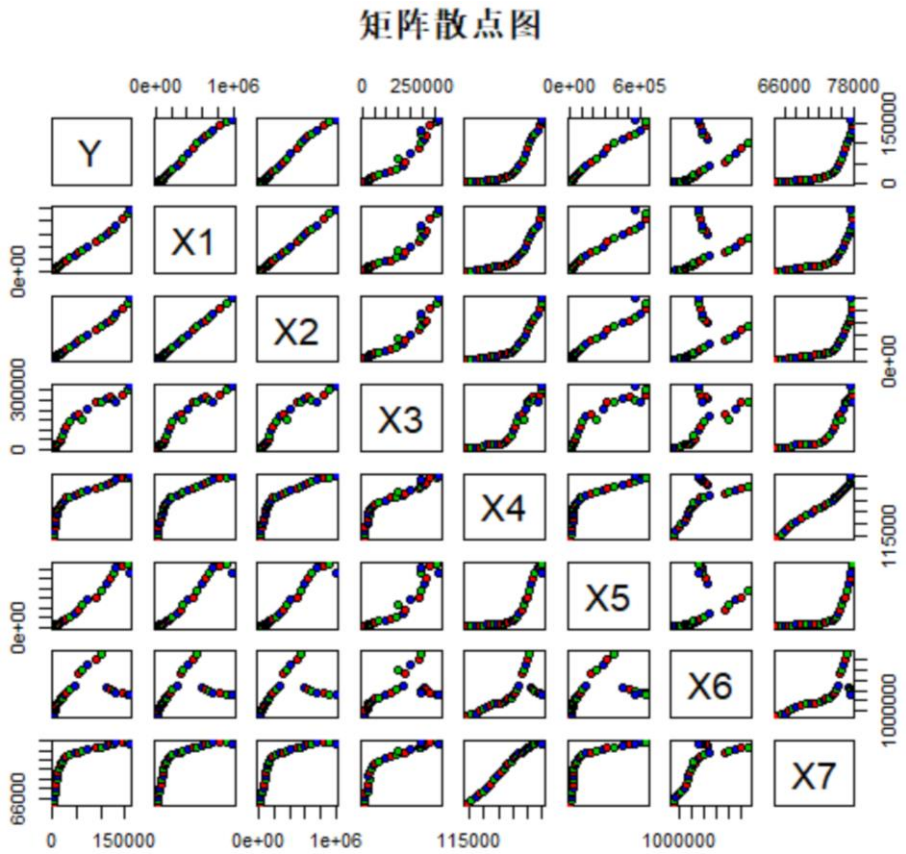
计算所得的回归方程为：$Y = 82600 - 0.3493X1 + 0.4706X2 + 0.06181X3 - 2.927X4 + 0.07055X5 + 0.001838X6 + 3.891X7$

b. 画矩阵散点图

R 代码：

```
> library(graphics)
> pairs(a[1:8],main="矩阵散点图",pch = 21,bg = c('red','green3','blue'))
```

结果：



c. 对参数进行区间估计

R 代码：

```
> confint(fit,level = 0.95)
```

结果：

```
                    2.5 %        97.5 %
(Intercept)  5.020422e+04  1.149984e+05
X1          -7.184782e-01  1.997382e-02
X2           1.034531e-01  8.377333e-01
X3           2.734183e-02  9.634199e-02
X4          -4.625264e+00 -1.228932e+00
X5           5.743594e-02  8.367003e-02
X6           5.826140e-04  3.093868e-03
X7           1.262512e+00  6.519712e+00
```

d. 方差分析

R 代码：

```
> anova(fit)
```

结果：

```
Analysis of Variance Table

Response: Y
          Df     Sum Sq    Mean Sq    F value    Pr(>F)
X1         1 8.1792e+10 8.1792e+10 51527.3162 < 2.2e-16 ***
X2         1 3.6176e+06 3.6176e+06     2.2790  0.145367
X3         1 1.6441e+08 1.6441e+08   103.5751 8.761e-10 ***
X4         1 6.5912e+07 6.5912e+07    41.5228 1.751e-06 ***
X5         1 2.7933e+08 2.7933e+08   175.9693 5.665e-12 ***
X6         1 1.3961e+07 1.3961e+07     8.7953  0.007139 **
X7         1 1.4960e+07 1.4960e+07     9.4246  0.005605 **
Residuals 22 3.4922e+07 1.5874e+06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

从 a. 输出结果中 t 统计量的 p 值可以看出，X1 的 p 值为 0.06257>0.05，回归系数不显著

从 d. 输出的方差分析表中的 F 统计量的 p 值可以看出，X2 的 p 值为 0.145367，不显著

(3)异常值点和强影响点的诊断

a.强影响点的诊断

R 代码：

```
> influence.measures(fit)
```

结果：

```
        dffit cov.r   cook.d   hat inf
1   -1.64736 1.093 3.15e-01 0.499   *
2   -0.22499 1.910 6.59e-03 0.278
3    0.09529 1.712 1.19e-03 0.167
4    0.27042 1.416 9.37e-03 0.138
5    0.19432 1.693 4.91e-03 0.193
6    0.18547 1.983 4.49e-03 0.291
7   -0.08789 1.751 1.01e-03 0.182
8   -0.01129 1.661 1.67e-05 0.127
9   -0.00752 1.639 7.41e-06 0.115
10   0.16307 1.522 3.45e-03 0.117
11   0.12840 1.577 2.15e-03 0.121
12   0.44955 1.414 2.56e-02 0.218
13   0.63708 0.934 4.90e-02 0.187
14   0.19882 1.531 5.12e-03 0.137
15  -0.28373 1.398 1.03e-02 0.139
16  -0.65690 0.813 5.14e-02 0.170
17  -0.82502 0.714 7.92e-02 0.206
18   0.01409 1.959 2.60e-05 0.259
19  -0.20574 2.065 5.52e-03 0.321
20  -0.33172 1.700 1.42e-02 0.248
21  -0.77498 0.711 7.01e-02 0.189
22   0.26999 2.534 9.51e-03 0.447   *
23   1.32424 0.886 2.03e-01 0.389   *
24   1.79369 0.701 3.56e-01 0.457   *
25   1.99284 0.235 3.91e-01 0.365   *
26   0.20553 1.806 5.50e-03 0.237
27  -1.71953 0.231 2.94e-01 0.307
28  -0.30852 1.982 1.23e-02 0.322
29  -0.43036 1.874 2.38e-02 0.329
30   1.78240 7.476 4.05e-01 0.844   *
```

可以看出第 1、22、23、24、25、30 个数据是强影响点。

b.异常值点诊断：

- 方法一：outlierTest()函数

R 代码：

```
> install.packages("car")
> library(car)
> outlierTest(fit)
```

结果：

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
25 2.626996           0.015755      0.47264
```

由此方法的输出结果可知，无异常值点。

- 方法二：（F 检验）

R 代码：

```
> r<-rstudent(fit)
> calcu_F<-function(p,r)
> {
>   n = length(r)
>   ans = (n-p-2)*(r**2)/(n-p-1-r**2)
>   return(ans)
> }
> ff<-calcu_F(9,r)
> df_val<-qf(1-0.05,1,12)
> ff[ff>df_val]
```

结果：

```
       25        27
10.010087  9.492269
```

由此方法的结果可知，第 25、27 个数据是异常点。

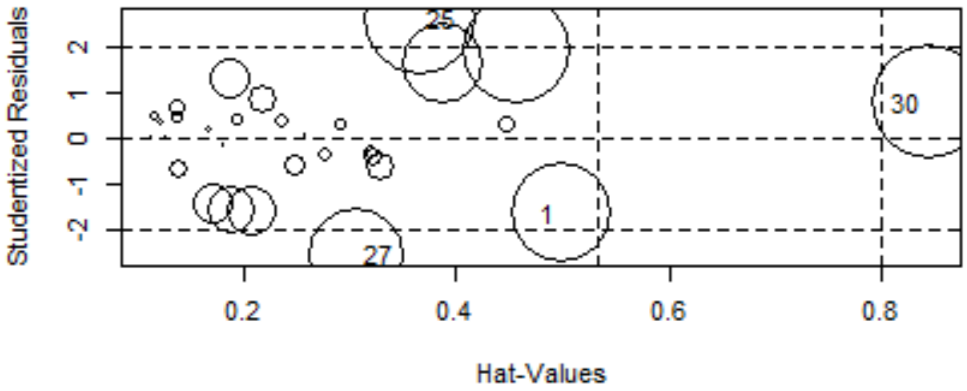
- 方法三：（使用 influencePlot()函数）

R 代码：

```
> influencePlot(fit)
```

结果：

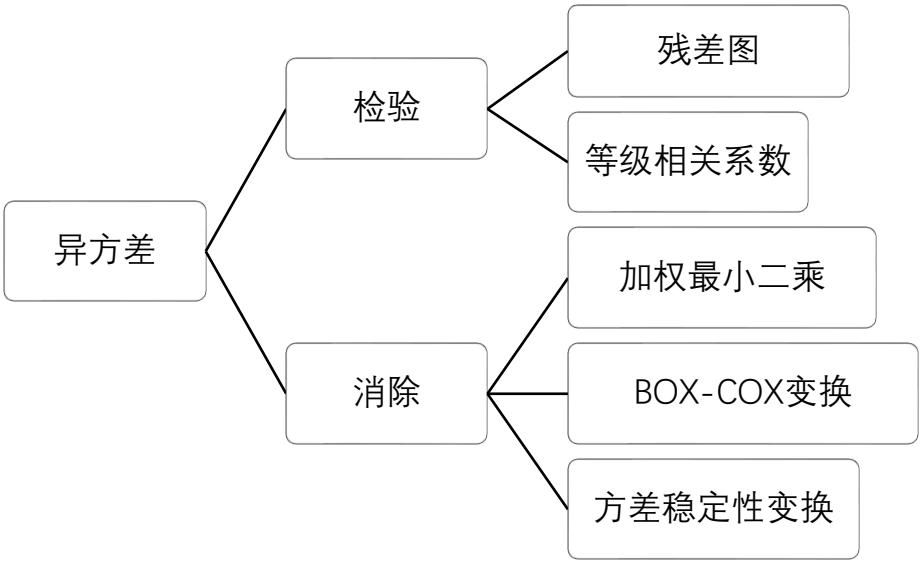
```
          StudRes        Hat     CookD
1  -1.6498344 0.4992486 0.3145993
25  2.6269959 0.3652703 0.3914314
27 -2.5812882 0.3073630 0.2939351
30  0.7654809 0.8442792 0.4047339
```



由此输出结果可知，第 1、25、27、30 个数据是异常点。

综上可看出不同方法对异常点的判断不同。

（4）异方差的检验和处理

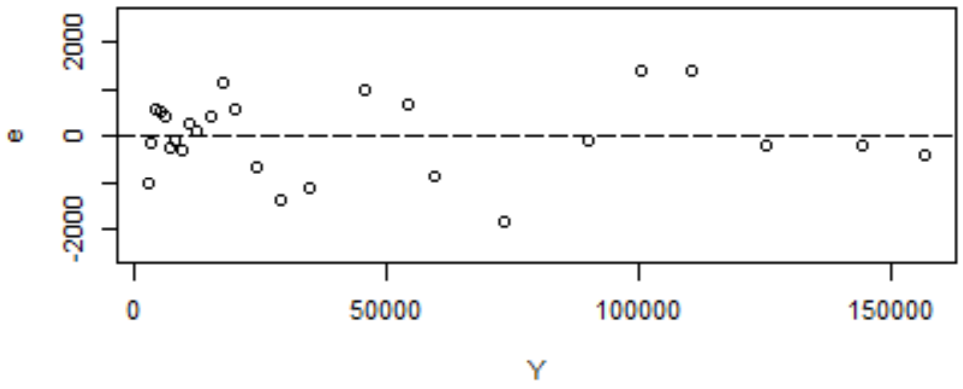


a. 首先用提出异常值的数据画出关于 y 的残差散点图

R 代码：

```
> e<-resid(fit)
> Y<-a[,1]
> plot(Y,e)
> attach(a)
> plot(Y,e,ylim=c(-2500,2500))
> abline(h=c(0),lty=5)
>detach(a)
```

结果：



通过该图，我们不难发现，残差图上的点的散布呈现出一定的趋势，可以初步判断所得的回归方程存在异方差性。

b. 为了进一步确认其异方差性，分别对各变量进行 Spearman 检验，

R 代码：

```
> abse<-abs(e)
> cor.test(a$X1,abse,alternative='two.side',method='spearman',conf.level=0.95)
> cor.test(a$X2,abse,alternative='two.side',method='spearman',conf.level=0.95)
> cor.test(a$X3,abse,alternative='two.side',method='spearman',conf.level=0.95)
> cor.test(a$X4,abse,alternative='two.side',method='spearman',conf.level=0.95)
> cor.test(a$X5,abse,alternative='two.side',method='spearman',conf.level=0.95)
> cor.test(a$X6,abse,alternative='two.side',method='spearman',conf.level=0.95)
> cor.test(a$X7,abse,alternative='two.side',method='spearman',conf.level=0.95)
```

结果：

自变量	等级相关系数	P 值
X1	0. 2102564	0. 3011
X2	0. 2102564	0. 3011
X3	0. 2177778	0. 2838
X4	0. 2102564	0. 3011
X5	0. 2102564	0. 3011
X6	0. 3873504	0. 05148
X7	0. 2068376	0. 3092

从 Spearman 检验的结果可以看出，回归方程确实存在异方差性。

c. 消除回归方程的异方差性

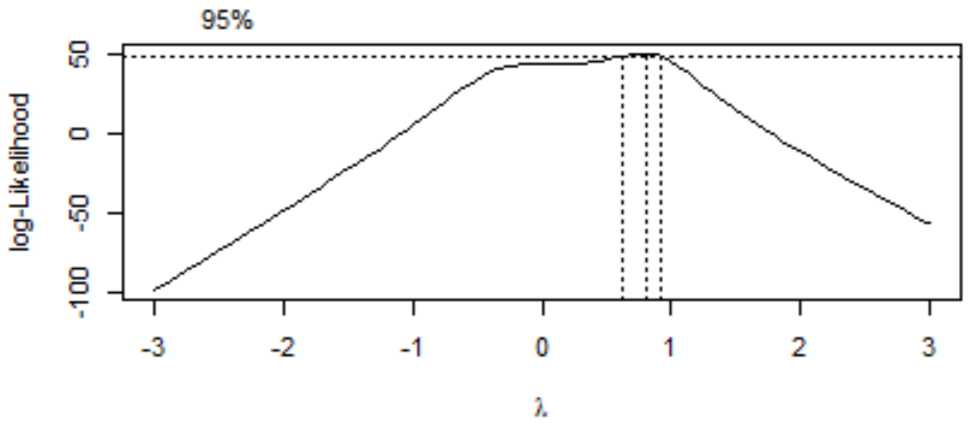
- 方法一：采用 BOX-COX 变换

取λ的值从[-3,3]以步长为 0.1，分别计算对应λ喜爱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 代码：

```
> install.packages("MASS")
> library (MASS)
> lambde_mle<-boxcox(Y~.,data=a,lambda=seq(-3,3,0.1))
> l=which(lambde_mle$Y==max(lambde_mle$Y))
> lambde_mle$x[l]
```

结果：



此时无使对数似然估计值取值最大的λ。

- 方法二：加权最小二乘

由 b. 中自变量的等级相关系数可知，X6 的等级相关系数最大，故选用 X6 构造权函数。

R 代码：

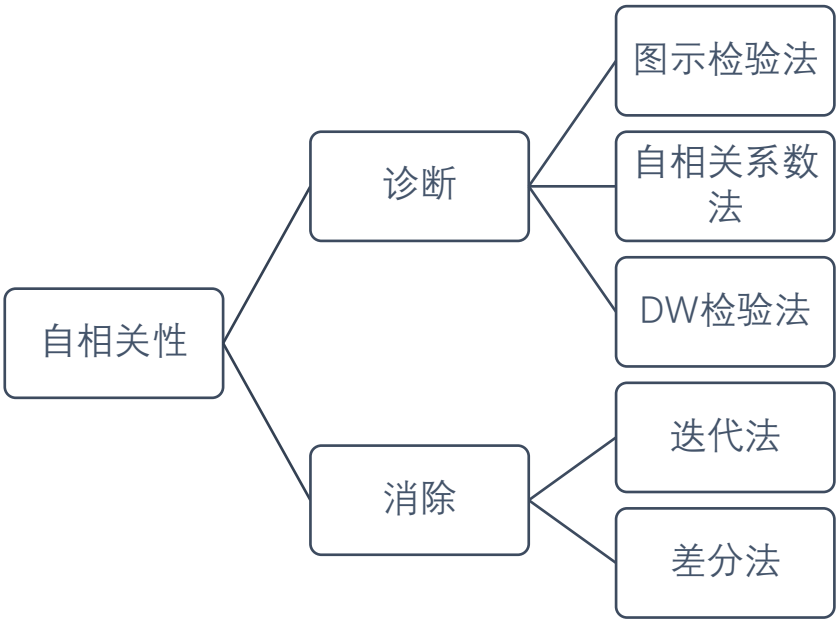
```
> s<-seq(1,5,0.5)
> result1<-vector(length=9,mode="list")
> result2<-vector(length=9,mode="list")
> for(j in 1:9)
> {w<-a$X6^(-s[j])
> lm1<-lm(Y~.,weights=w,data=a)
> result1[[j]]<-logLik(lm1)
> result2[[j]]<-summary(lm1)}
```

结果：

```
Warning messages:
1: In summary.lm(lm1) : essentially perfect fit: summary may be unreliable
2: In summary.lm(lm1) : essentially perfect fit: summary may be unreliable
```

上述输出结果说明可能存在的问题数据格式符合但不足以拟合导致结果不够理想，故无法消除异方差。

（5）自相关性的检验和处理



a. 对得到的回归方程做 DW 检验，进行自相关性分析，

R 代码：

```
> install.packages("lmtest")
> library(lmtest)
> dwtest(fit)
```

结果：

```
        Durbin-Watson test

data:  fit
DW = 1.2207, p-value = 0.0001271
alternative hypothesis: true autocorrelation is greater than 0
```

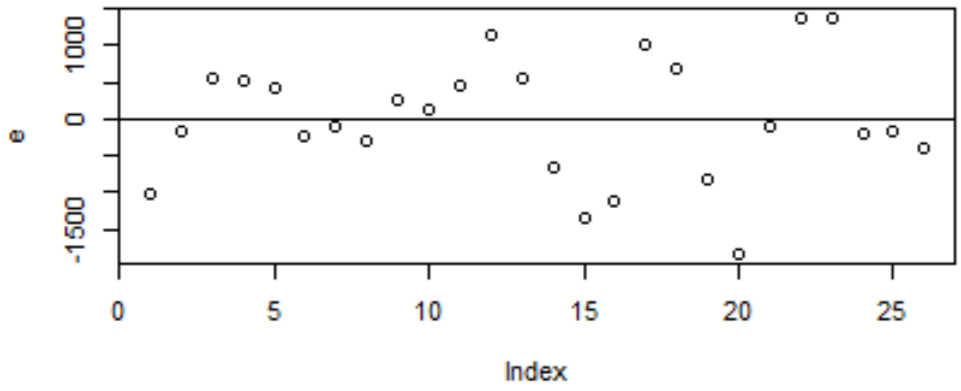
由运行结果可知 DW=1.2207，查表可得$d_U = 2.817$，$d_L = 0.816$，p 值为 0.0001271<<0.05，故存在自相关性。

b. 绘制时间序列残差图

R 代码：

```
> plot(e)
> abline(h = 0)
```

结果：



c. 差分法消除自相关性

R 代码：

```
> diffY <- diff(Y)
> diffX <- diff(X1-X7)
> diff.reg <- lm(diffY~diffX-1)
> summary(diff.reg)
> durbinWatsonTest(diff.reg)
```

结果：

```
Call:
lm(formula = diffY ~ diffX - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-4383.0  -432.6   338.6   950.3  4570.2

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
diffX 0.164422   0.007251   22.68   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1794 on 24 degrees of freedom
Multiple R-squared:  0.9554,    Adjusted R-squared:  0.9535
F-statistic: 514.2 on 1 and 24 DF,  p-value: < 2.2e-16
```

可以看出消除后$R^2 \to 1$，故方程显著。

（6）变量选择与逐步回归

R 代码：

```
> summary(step(fit,dirction='both'))
```

结果：

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, data = a)

Residuals:
     Min       1Q   Median       3Q      Max
-1832.93  -362.78   -92.45   539.75  1375.39

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.462e+04  1.728e+04   5.477 3.35e-05 ***
X1          -5.200e-01  1.524e-01  -3.412 0.003108 **
X2           6.429e-01  1.506e-01   4.269 0.000462 ***
X3           4.645e-02  1.402e-02   3.314 0.003863 **
X4          -3.621e+00  7.908e-01  -4.578 0.000233 ***
X5           7.806e-02  1.145e-02   6.819 2.20e-06 ***
X6           2.255e-03  4.823e-04   4.676 0.000188 ***
X7           4.939e+00  1.181e+00   4.182 0.000560 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 967 on 18 degrees of freedom
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9996
F-statistic:  8673 on 7 and 18 DF,  p-value: < 2.2e-16
```

```
Start:  AIC=363.9
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7

       Df Sum of Sq      RSS    AIC
<none>              16833130 363.90
- X3    1  10268394 27101523 374.28
- X1    1  10886739 27719868 374.87
- X7    1  16357169 33190298 379.55
- X2    1  17044763 33877892 380.08
- X4    1  19603428 36436558 381.98
- X6    1  20447432 37280561 382.57
- X5    1  43485825 60318955 395.08
```

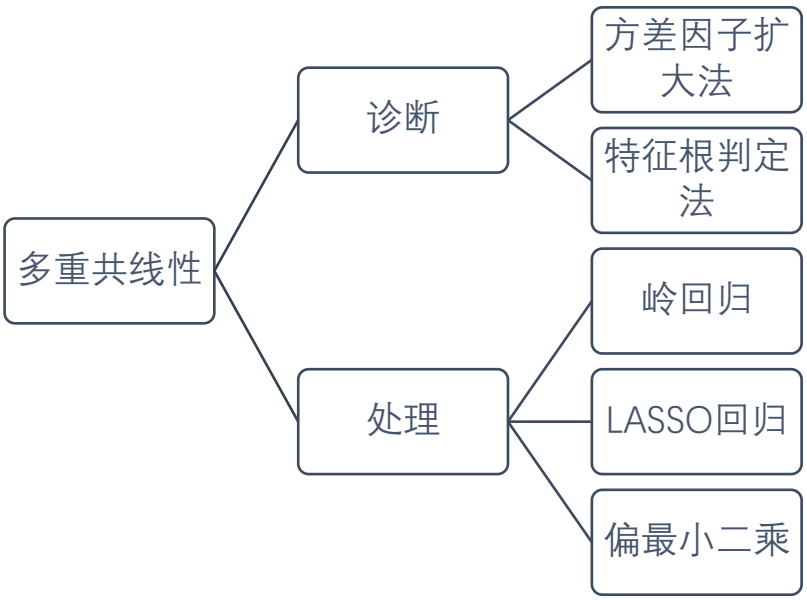
计算所得的回归方程为：

$$Y = 94620 - 0.52X1 + 0.6429X2 + 0.04645X3 - 3.621X4$$
$$+ 0.07806X5 + 0.002255X6 + 4.939X7$$

其相关系数为$R^2 = 0.996 \to 1$，回归效果良好，回归方程显著。

通过该表结果，我们可以发现各变量回归均显著。

（7）多重共线性的诊断和处理



a. 采用方差扩大因子法、回归方程进行多重共线性诊断，

R 代码：

```
> install.packages("car")
> library(car)
> vif(fit)
```

结果：

```
          X1           X2           X3           X4           X5           X6
41685.733843 40979.689961    51.989188   777.576378   146.814381     4.096942
          X7
  558.074470
```

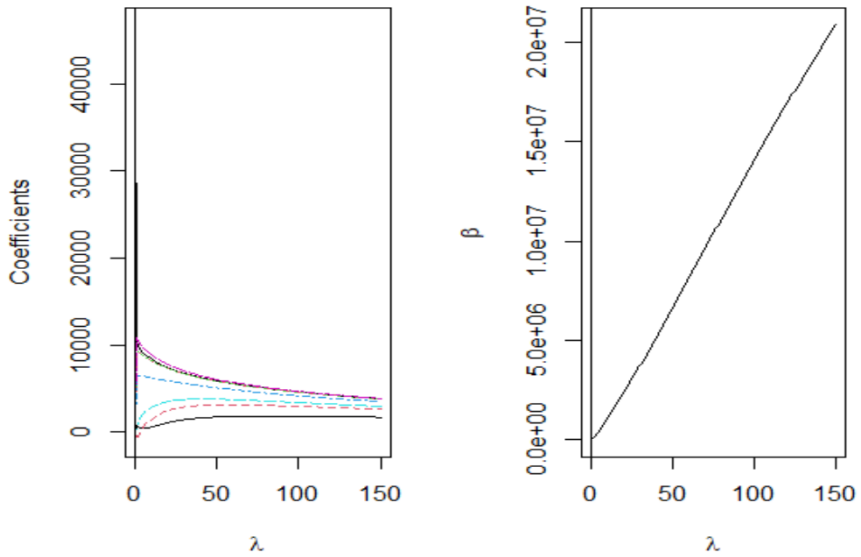
由输出结果可知，除 X6 外，各自变量的方差扩大因子 VIF 均大于 10，这说明自变量之间存在多重共线性。

b. 岭回归

R 代码：

```
> library(MASS)
> ridge.dia<-lm.ridge(y~.,lambda = seq(0,150,length.out = 151),data = a[,1:8],model = TRUE)
> names(ridge.dia)
> ridge.dia$lambda[which.min(ridge.dia$GCV)]
> ridge.dia$coef[,which.min(ridge.dia$GCV)]
> par(mfrow=c(1,2))
> matplot(ridge.dia$lambda,t(ridge.dia$coef),xlab = expression(lambda),ylab = "Coefficients",type = "l",lty = 1:20)
> abline(v=ridge.dia$lambda[which.min(ridge.dia$GCV)])
> plot(ridge.dia$lambda,ridge.dia$GCV,type = "l",xlab = expression(lambda),ylab = expression(beta))
> abline(v=ridge.dia$lambda[which.min(ridge.dia$GCV)])
> library(ridge)
> b=linearRidge(Y~X3+x4+x5+x7,data=a)
> summary(b)
```

结果：



```
Call:
linearRidge(formula = Y ~ X3 + X4 + X5 + X7, data = a)


Coefficients:
             Estimate Scaled estimate Std. Error (scaled) t value (scaled)
(Intercept) -1.849e+04              NA                  NA               NA
X3           1.517e-01       7.544e+04           7.415e+03           10.173
X4           6.019e-01       2.053e+04           4.805e+03            4.272
X5           1.547e-01       1.583e+05           5.459e+03           29.002
X7          -7.919e-01      -1.532e+04           4.359e+03            3.514
            Pr(>|t|)
(Intercept)       NA
X3           < 2e-16 ***
X4          1.94e-05 ***
X5           < 2e-16 ***
X7          0.000441 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge parameter: 0.01662145, chosen automatically, computed using 1 PCs

Degrees of freedom: model 2.724 , variance 2.387 , residual 3.061
```

当λ=0 时 GCM 最小,故最终得到的回归方程为:

$$Y = -18490 + 0.1517X3 + 0.6019X4 + 0.1547X5 - 0.7919X7$$

三、 附录：

数据来源：中国国家统计局《2020 年统计年鉴》

Y	X1	X2	X3	X4	X5	X6	X7
2821.86	18923.33	18872.87	5560.12	114333.00	4517.00	772682	64749
2990.17	22050.34	22005.63	7225.75	115823.00	5594.50	806048	65491
3296.91	27208.21	27194.53	9119.62	117171.00	8080.10	860855	66152
4255.30	35599.25	35673.23	11271.02	118517.00	13072.00	996634	66808
5126.88	48548.16	48637.45	20381.90	119850.00	17042.00	1092882	67455
6038.04	60356.64	61339.89	23499.94	121121.00	20019.30	1172596	68065
6909.82	70779.59	71813.63	24133.86	122389.00	22974.00	1245357	68950
8234.04	78802.86	79715.04	26967.24	123626.00	24941.10	1326094	69820
9262.80	83817.56	85195.51	26849.68	124761.00	28406.20	1378717	70637
10682.58	89366.48	90564.38	29896.23	125786.00	29854.70	1394413	71394
12581.51	99066.07	100280.14	39273.25	126743.00	32917.70	1478573	72085
15301.38	109276.15	110863.12	42183.62	127627.00	37213.50	1534122	72797
17636.45	120480.41	121717.42	51378.15	128453.00	43499.90	1608150	73280
20017.31	136576.25	137422.03	70483.45	129227.00	55566.60	1587497	73736
24165.68	161415.43	161840.16	95539.09	129988.00	70477.40	1767453	74264
28778.54	185998.91	187318.90	116921.77	130756.00	88773.60	1847018	74647
34804.35	219028.45	219438.47	140974.74	131448.00	109998.20	2024158	74978
45621.97	270704.02	270092.32	166924.07	132129.00	137323.90	2227761	75321
54223.79	321229.52	319244.61	179921.47	132802.00	172828.40	2867892	75564
59521.59	347934.88	348517.74	150648.06	133450.00	224598.80	2976898	75828
73210.79	410354.11	412119.26	201722.34	134091.00	278121.90	3269508	76105
89738.39	483392.79	487940.18	236401.95	134735.00	311485.10	3526319	76420
100614.28	537329.01	538579.95	244160.21	135404.00	374694.70	3804035	76704
110530.70	588141.21	592963.23	258168.89	136072.00	446294.10	2122992	76977
119175.31	644380.15	643563.10	264241.77	136782.00	512020.70	2032218	77253
124922.20	686255.74	688858.22	245502.93	137462.00	561999.80	1943271	77451
130360.73	743408.25	746395.06	243386.46	138271.00	606465.70	1900194	77603
144369.87	831381.20	832035.95	278099.24	139008.00	641238.40	1848620	77640
156402.86	914327.10	919281.13	305008.13	139538.00	645675.00	1793820	77586
158000.46	988528.90	990865.11	315627.32	140005.00	560874.30	1760436	77471