《多元统计分析》第三点五次上机作业
多元线性回归分析

3180103000 许乐乐

- 实验目的与要求：

  通过本试验项目，能够理解并掌握如下内容：

  1. 能够利用统计软件对数据进行分析，变量筛选(可通过R中的step函数处理)和建模；
  2. 能够利用经典的大数据定律和中心极限定律探讨模型参数估计量的渐近性质。

习题一

利用多元线性回归模型探讨我们民航客运量与其他协变量之间的关系，数据见ex2.3.xls。其中，y（万人）为民航客运量，x1（亿元）为国民收入，x2（亿元）为销售额，x3（万人）为铁路客运量，x4（万公里）为民航航线里程。x5（万人）为来华旅游入境人数。

- 读取数据

```
d<-read_csv("ex2.3.csv")
```

- 建立多元线性回归模型

```
d.lm<-lm(y~x1+x2+x3+x4+x5,d)
summary(d.lm)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-252.31  -48.18  -12.79   52.81  193.02 

Coefficients:
               Estimate   Std. Error t value
(Intercept) -401.2240221  143.1418440  -2.803
x1             0.0144429    0.0240234   0.601
x2            -0.0208658    0.0865674  -0.241
x3             0.0000581    0.0014640   0.040
x4            30.4396638    8.2977116   3.668
x5             0.2004236    0.1114572   1.798
            Pr(>|t|)
(Intercept)  0.01870 *
x1           0.56109
x2           0.81440
x3           0.96912
x4           0.00433 **
x5           0.10235
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.6 on 10 degrees of freedom
Multiple R-squared:  0.9879,    Adjusted R-squared:  0.9818 
```

```
F-statistic: 162.8 on 5 and 10 DF,  p-value: 0.000000003043
```

故模型为：

$$y = -401.22 + 0.01444x_1 - 0.0209x_2 + 0.0001x_3 + 30.4397x_4 + 0.2004x_5$$

由t检验知，民航航线里程对民航客运量影响最大，来华旅游入境人数对民航客运量影响较大，国民收入、销售额、铁路客运量对民航客运量影响不大。

- 多重共线性

```
d.cor<-cor(d[,3:7]) # 计算相关矩阵
library(corrplot)
corrplot(d.cor)
```

```
          x1        x2        x3        x4        x5
x1 1.0000000 0.9722107 0.2310265 0.9599016 0.8775899
x2 0.9722107 1.0000000 0.2844786 0.9778043 0.9422928
x3 0.2310265 0.2844786 1.0000000 0.2316223 0.3583359
x4 0.9599016 0.9778043 0.2316223 1.0000000 0.8817976
x5 0.8775899 0.9422928 0.3583359 0.8817976 1.0000000
```

由相关矩阵图可知，x1与x2存在很强的正相关性，x3与x4存在很强的正相关性，因此回归模型存在多重共线性。

- 变量选择与逐步回归

用`step()`进行逐步回归。该函数根据AIC信息准则对变量进行选择，取使AIC最小的回归模型。

```
d.step<-step(d.lm)
```

```
Start:  AIC=160.15
y ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq    RSS    AIC
- x3    1        26 168042 158.15
- x2    1       976 168992 158.24
- x1    1      6073 174088 158.72
<none>              168015 160.15
- x5    1     54329 222344 162.63
- x4    1    226106 394122 171.79

Step:  AIC=158.15
y ~ x1 + x2 + x4 + x5

       Df Sum of Sq    RSS    AIC
- x2    1       957 168999 156.24
- x1    1      6050 174092 156.72
<none>              168042 158.15
- x5    1     55493 223535 160.72
- x4    1    228011 396052 169.87

Step:  AIC=156.24
y ~ x1 + x4 + x5

       Df Sum of Sq    RSS    AIC
- x1    1      5903 174902 154.79
<none>              168999 156.24
```

```
- x5    1    155567 324566 164.68
- x4    1    501855 670854 176.30

Step:  AIC=154.79
y ~ x4 + x5

       Df Sum of Sq     RSS    AIC
<none>               174902 154.79
- x5    1    180176  355078 164.12
- x4    1   1843682 2018584 191.93
```

```
summary(d.step)
```

```
Call:
lm(formula = y ~ x4 + x5, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-249.75  -42.72  -13.87   54.34  205.41

Coefficients:
             Estimate Std. Error t value
(Intercept) -410.07401   57.16828  -7.173
x4            31.47130    2.68842  11.706
x5             0.18680    0.05105   3.660
               Pr(>|t|)
(Intercept) 0.0000072278 ***
x4          0.0000000281 ***
x5               0.00288 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 116 on 13 degrees of freedom
Multiple R-squared:  0.9874,    Adjusted R-squared:  0.9854
F-statistic:   508 on 2 and 13 DF,  p-value: 0.0000000000004572
```

故模型为：

$$y = -410.07401 + 31.47130x_4 + 0.18680x_5$$

Multiple R-squared：0.9874，说明方程的拟合程度很高，我们认为民航客运量的变化可以由民航航线里程和来华旅游入境人数的变化来解释。

习题二

表ex2.5给出了我国1985年—2012年财政收入y （亿元），第一产业增加值x1（亿元），工业增加值x2（亿元），建筑业增加值x3（亿元），年末总人口x4（万人），社会消费品零售总额x5（亿元）和受灾面积x6（万公顷）的数据，试探讨我国财政收入与其他因素是否存在显著关系，并建立相应的线性回归模型。

- 完全仿照习题一
- 读取数据

```
d<-read_csv("ex2.5.csv")
```

- 建立多元线性回归模型

```
d.lm<-lm(y~x1+x2+x3+x4+x5+x6,d)
summary(d.lm)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-3078.2  -713.3  -118.6   674.8  2852.2 

Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept) 37159.50528 15846.84594   2.345 0.028937
x1             -0.77924     0.33507  -2.326 0.030138
x2              0.23081     0.05888   3.920 0.000786
x3              0.54249     0.89395   0.607 0.550460
x4             -0.30585     0.16361  -1.869 0.075580
x5              0.45995     0.15270   3.012 0.006636
x6             -0.57568     0.62738  -0.918 0.369255

(Intercept) *
x1          *
x2          ***
x3
x4          .
x5          **
x6
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1428 on 21 degrees of freedom
Multiple R-squared:  0.9985,    Adjusted R-squared:  0.9981 
F-statistic:  2338 on 6 and 21 DF,  p-value: < 0.00000000000000022
```

故模型为：

$$y = 37159.50528 - 0.77924x_1 + 0.23081x_2 + 0.54249x_3 - 0.30585x_4 + 0.45995x_5 - 0.57568x_6$$

由t检验知，

  - 工业增长值、社会消费品零售总额、第一产业增加值对财政收入影响很大
  - 年末总人口对财政收入影响较小
  - 建筑业增加值、受灾面积对财政收入影响不大

- 多重共线性

```
d.cor<-cor(d[,3:8]) # 计算相关矩阵
library(corrplot)
corrplot(d.cor)
```

	x1	x2	x3	x4	x5	x6
x1	1.0000000	0.9927753	0.9857972	0.8707332	0.9905320	-0.6400317
x2	0.9927753	1.0000000	0.9930891	0.8311175	0.9956199	-0.6616221
x3	0.9857972	0.9930891	1.0000000	0.7861198	0.9986965	-0.6804451
x4	0.8707332	0.8311175	0.7861198	1.0000000	0.8083851	-0.3509287
x5	0.9905320	0.9956199	0.9986965	0.8083851	1.0000000	-0.6760876
x6	-0.6400317	-0.6616221	-0.6804451	-0.3509287	-0.6760876	1.0000000

由相关矩阵图可知，

- 第一产业增加值与工业增加值、建筑业增加值、社会消费品零售总额存在很强正相关性
- 工业增加值与建筑业增加值、社会消费品零售总额存在很强正相关性
- 建筑业增加值与社会消费品零售总额存在很强正相关性
- 年末总人口与第一产业增加值、工业增加值、建筑业增加值、社会消费品零售总额存在较强正相关性
- 受灾面积与第一产业增加值、工业增加值、建筑业增加值、社会消费品零售总额存在较强负相关性

因此，模型存在多重共线性。

- 变量选择与逐步回归

```
d.step<-step(d.lm)
```

```
Start:  AIC=412.73
y ~ x1 + x2 + x3 + x4 + x5 + x6

       Df Sum of Sq      RSS    AIC
- x3    1    750912 43571799 411.22
- x6    1   1716840 44537727 411.83
<none>              42820887 412.73
- x4    1   7125754 49946641 415.04
- x1    1  11028364 53849251 417.15
- x5    1  18499558 61320445 420.78
- x2    1  31333958 74154845 426.10

Step:  AIC=411.22
y ~ x1 + x2 + x4 + x5 + x6

       Df Sum of Sq       RSS    AIC
- x6    1   1114248  44686047 409.92
<none>               43571799 411.22
- x1    1  10574111  54145911 415.30
- x4    1  21835301  65407100 420.59
- x2    1  32092087  75663886 424.67
- x5    1 112640001 156211801 444.97
```

```
Step:  AIC=409.92
y ~ x1 + x2 + x4 + x5

       Df Sum of Sq       RSS    AIC
<none>               44686047 409.92
- x1    1   9603822  54289869 413.37
- x4    1  31433342  76119389 422.84
- x2    1  32250373  76936420 423.14
- x5    1 111626871 156312918 442.98
```

```
summary(d.step)
```

```
Call:
lm(formula = y ~ x1 + x2 + x4 + x5, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-2972.2  -736.2  -232.6  1003.9  2551.2

Coefficients:
               Estimate  Std. Error t value
(Intercept) 45562.51553 10877.10693   4.189
x1             -0.64822     0.29156  -2.223
x2              0.23366     0.05735   4.074
x4             -0.41151     0.10231  -4.022
x5              0.53765     0.07093   7.580
               Pr(>|t|)
(Intercept)    0.000352 ***
x1             0.036302 *
x2             0.000468 ***
x4             0.000532 ***
x5          0.000000107 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1394 on 23 degrees of freedom
Multiple R-squared:  0.9984,    Adjusted R-squared:  0.9982
F-statistic:  3681 on 4 and 23 DF,  p-value: < 0.00000000000000022
```

故模型为：

$$y = 45562.51553 - 0.64822x_1 + 0.23366x_2 - 0.41151x_4 + 0.53765x_5$$

R-squared: 0.9984，说明方程的拟合程度很高，我们认为财政收入的变化可以由第一产业增加值、工业增加值、年末总人口、社会消费品零售总额的变化来解释。