Multivariate Regression Analysis Using Statistics with R

Li Xiumin

Hebei University of Science and Technology, P.R.China, 050018

li_xiumin@163.com

Key words：regression analysis model;statistics with R; linear

Abstract：Multiple regression analysis is a useful model in econometrics. It can be applied in many fields. Statistics software plays an important role in processing data. This paper gives a method to use R, constructs regression model, and explains the result.

Introduction of statistics of R

R is a free software environment for statistical computing and graphics, established by Ross Ihaka and Robert Gentleman from the University of Auckland, New Zealand [1]. They chose to write a reduced version of S for teaching purposes. R owns many powerful statistical packages, which were offered by different experts. Besides some elementary packages such as regression analysis, ANOVA, there are some new research results in R, for example, extreme value statistical package, dependence structures function package and so on. The installation files can be downloaded from the Internet Comprehensive R Archive Network (CRAN): http://cran.r-project.org. It is easy to install R for pressing the key “enter”. The statistical analysis is finished by writing some procedures. This paper gives some examples of regression analysis, introduces the method of analyzing problem using R, and explains the result [2].

Multivariate linear regression

Multivariate linear regression generates an equation to describe the statistical relationship between the response variable and several predictors. The basic model for multiple regression analysis is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots \hat{\beta}_k x_k$$

where $x_1, \ldots, x_k$ are explanatory variables (also called predictors), $\hat{\beta}_0$ is a constant, the $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are regression coefficients, which can be estimated using the method of least squares [3].

Example1. Multi-factor analysis of financial income

In a certain period of time, financial income y is affected by many factors, such as revenue income x_1, GDP x_2, which are considered as independent variable and analyzed the extend of impact on financial income. The data are from the statistical yearbooks of China between 1978 and 1995 [4].

Data input

There are several ways to read data into R. While the number of data is less, we can input the data form the key directly; but while the number is large, we need to read data from a text file. For instance, if the data save format is .txt, we can read the data below:

```
>income =read.table("d:/shuju.txt",col.names=c("y","x1","x2")),
```

The result is a data frame, which is put into the variable income and looks as follows:

```
>incom
        y          x1        x2
1    1132.62    3624.1    519.28
            .........
18   6242.20   57277.3   6038.04
```

The left of the output shows the length of the data.

Linear regression analysis

For linear regression analysis, the function lm is used:

```
>lm.income=lm(income$y~income$x1+income$x2)
>summary(lm.income)
```

The argument to lm is a model formula, in which the tilde symbol (~) should be read as "described by". The result is:

Call: lm(formula = income$y ~ income$x1 + income$x2)

Residuals:

Min	1Q	Median	3Q	Max
-140.03	-75.76	-15.12	51.81	456.77

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.036e+02	6.844e+01	11.742	5.82e-09 ***
income$x1	6.082e-02	9.781e-03	6.218	1.65e-05 ***
income$x2	3.233e-01	9.134e-02	3.540	0.00297 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 144.7 on 15 degrees of freedom

Multiple R-Squared: 0.9915, Adjusted R-squared: 0.9904

F-statistic: 873.4 on 2 and 15 DF, p-value: 2.992e-16

From above, the best-fitted straight line is seen to be $y = 803.6 + 0.06x_1 + 0.32x_2$. T statistics are 6.218 and 3.540 respectively. The p-value is less than 0.05, that is to say, two regression coefficients are both unequal to zero at the 5 percent level of significance. Fatherly, the residual variation is 144.72, F statistics is 873.4, p-value is 2.992e-16, and this shows that regression equation is significant. Multiple R-Squared is 0.9915, and adjusted R-squared is 0.9904, this shows the fitting is better.

Fitted plot

We have financial income as x axis, regression fitted value as y axis, line the points on the plane, see figure 1. Using the function:

plot(income$y,fitted(lm.1),type="l"),

We get the plot as follow:

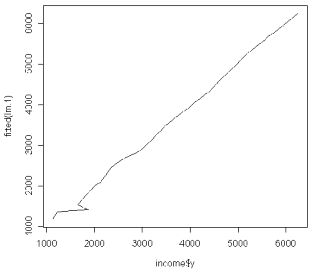

Figure 1. Regression value versus financial income

Confidence and prediction intervals

The narrow intervals, confidence intervals, reflect the uncertainty about the line itself, like the precision with which a mean is known. The wide intervals, prediction intervals, include the uncertainty about future observations. These intervals should capture the majority of the observed points and will not collapse to a line as the number of observations increases. The command is as follows:

> predict(lm.1,int="c")

	fit	lwr	upr
1	1191.902	1072.927	1310.876
	………		
18	6239.037	6014.831	6463.244

> predict(lm.1,int="p")

	fit	lwr	upr
1	1191.902	861.2898	1522.513
	………		
18	6239.037	5857.7003	6620.374

Residual analysis

Residual analysis can test whether the random errors in regression model are independent and identically distributed and judge the outlier. The command is plot(lm.1). The figure 2 shows the residual plot. Four different plots are in the set: the left top shows residuals versus fitted values, we can see all plots randomly scatted between y-axis -1 and 1 except the sixth plot, this is to say the random error have the same variation; the left below is of the square root of the absolute value of the standardized residuals; the right top is a Q-Q normal distribution plot of standardized residuals, this plot shows random error are from normal distribution since the Q-Q plot is a straight line; the right below is of "Cook's distance" which is a measure of the influence of each observation on the regression coefficients, this shows the sixth plot is a outlier too, this plot play an important role in regression equation. We need to consider this outlier's background in practice.

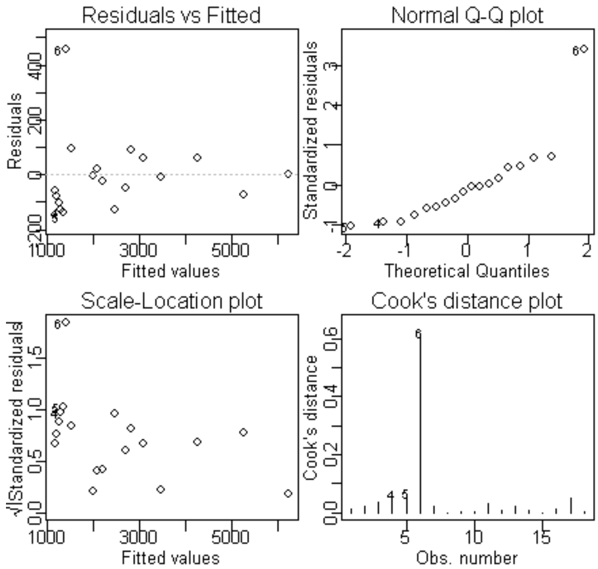

Figure 2. Residual analysis Plot

Construction the best regression equation

In practice, the response variable is often influenced by several predictors. But if all the predictors are taken into account, we would find the large amount calculation and the dissatisfactory result.

Example 2. A certain kind of cement release the heat y (cal/g) when it is solidified is related to four chemical compositions. The data are shown in table 1 [5] .

Table 1. Heat versus four chemical compositions

compositions	1	2	3	4	5	6	7	8	9	10	11	12	13
x_1	7	1	11	11	7	11	3	1	2	21	1	11	10
x_2	26	29	56	31	52	55	71	31	54	47	40	66	68
x_3	6	15	8	8	6	9	17	22	18	4	23	9	8
x_4	60	52	20	47	33	22	6	44	22	26	34	12	12
y	78.5	74.3	104.3	87.6	95.9	109.2	102.7	72.5	93.1	115.9	83.8	113.3	109.4

Find the linear regression equation, the command is as follows:

```
>concrete = read.table ("d:/example2.txt",col.names = c ("x1", "x2","x3","x4","y"))
>lm.1 = lm (concrete$y~concrete$x1+concrete$x2+concrete$x3+concrete$x4)
>summary (lm.1)
  Call:
lm(formula = concrete$y ~ concrete$x1 + concrete$x2 + concrete$x3 + concrete$x4)
          Residuals:
            Min        1Q       Median       3Q       Max
          -3.1750    -1.6709   0.2508       1.3783   3.9254
          Coefficients:
                       Estimate    Std. Error   t value   Pr(>|t|)
          (Intercept)   62.4054     70.0710     0.891     0.3991
          concrete$x1    1.5511      0.7448     2.083     0.0708 .
```

concrete$x2	0.5102	0.7238	0.705	0.5009
concrete$x3	0.1019	0.7547	0.135	0.8959
concrete$x4	-0.1441	0.7091	-0.203	0.8441

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-Squared: 0.9824, Adjusted R-squared: 0.9736
F-statistic: 111.5 on 4 and 8 DF, p-value: 4.756e-07

From the result, the four regression coefficients are equal to zero at the 5 percent level of significance. However, F statistics is 111.5, p-value is 4.756e-07, and this shows that regression equation is significant. The reason for this contradictory result is there is collinearity between the four predictors. Below we adopt the "backward regression method" to construct the best regression equation. This method construct the regression equation covered all the predictors, then test the regression coefficients, and keep back the significant factors, remove other factors. Firstly, remove the variable whose t statistics is the least, namely x_4, the command as follows:

>lm.2 = lm(concrete$y~concrete$x1+concrete$x2+concrete$x3)

>summary(lm.2), Similarly remove x_3:

>lm.3=lm(concrete$y~concrete$x1+concrete$x2)

>summary(lm.3)

Call: lm(formula = concrete$y ~ concrete$x1 + concrete$x2)

Residuals:

Min	1Q	Median	3Q	Max
-2.893	-1.574	-1.302	1.362	4.048

Coefficients:

	Estimate	Std.	Error	t value	Pr(>|t|)
(Intercept)	52.57735	2.28617	23.00	5.46e-10 ***	
concrete$x1	1.46831	0.12130	12.11	2.69e-07 ***	
concrete$x2	0.66225	0.04585	14.44	5.03e-08 ***	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-Squared: 0.9787, Adjusted R-squared: 0.9744
F-statistic: 229.5 on 2 and 10 DF, p-value: 4.407e-09

Final we get the best regression equation is $\hat{y} = 52.5773 + 1.4683x_1 + 0.6623x_2$. From above analysis, we can conclude R plays a very important role in regression analysis, for its command is easy, the result is precise, it is a useful statistical analysis software.

Acknowledgment

This work is supportted by the teaching and research project of Polytechnic College of HeBei University of Science and Technologe(2011Z08). It is also supported by the Soft Science Foundation of Hebei Province(12457203D−53**).**

References

[1] P. Dalgaard. Introductory statistics with R. Springer, New York, 2002.
[2] Xue Yi, Zhang Liping. Statistical Model and Software R. Beijing: Tsinghua university press, 2006.
[3] Yang Hu, Zhong Bo, Liu Qiongsun. Applied Mathematical Statistics. Beijing: Tsinghua university press. 2010.
[4] Yi Danhui. Data analysis and application for EViews. Beijing: China Statistics Press, 2002.
[5] Zhang Xiaodong. A guide to using EViews. Tianjin: Nankai University Press, 2004.