Discussion on the Application of R Software in Statistics Teaching

Fengyan Yu

ABSTRACT

Inspiration and practicality are important to improve the teaching effect of higher education, especially for the statistical teaching which has strong theoretical content, abstract content and complicated calculation. R software is free and open, with a powerful statistical analysis and mapping functions. This paper discusses the application of R software in statistical teaching for improving the teaching effect of statistics, and gives the concrete application examples, and provides some reference for the current statistical teaching.

1. Introduction

Statistics is a science of data collection, collation, analysis, interpretation and inference. In 2011, the Ministry of Education has officially classified statistics as a discipline (with mathematics, economics, etc.), then the importance of statistics is evident. In the domestic higher education teaching, most students think that statistics is a very difficult course, in which cumbersome and complicated formulas often make the students discouraged. Making full use of statistical software is an important way to improve the effectiveness of statistical teaching. Because students will be interested in the data processing process of some practical problems.

In recent years, R software with free use, open source code and other characteristics quickly comes into the analysis of statistical researchers and applications. R software can be used in place of SAS, SPSS, EVIEWS and other software in teaching. This paper provides some reference for the current statistical teaching by some examples of application.

Fengyan Yu, Shanghai Normal University Tianhua College, Shanghai China
2. R Software

R software is free and open with statistical analysis capabilities and powerful mapping capabilities, developed by Robert Gentleman and Ross Ihaka of Auckland University and other volunteers. Users can understand the latest information on R software and instructions, and get the latest version of the R software and R-based application statistics package in R software website (http://www.r-project.org). Compared with the expensive commercial copyright professional software, such as SAS, SPSS, S-Plus, Matlab and other software, R software not only has almost all of these similar software features, but sometimes even has more features in scientific research and business practice.

Basic programming language in R software refer to S language, including input and output programs from Excel files, TXT files, memory, database import and export data, data processing programs, mathematical procedures, character operation procedures, date and time procedures, drawing related, Statistical model programs, and so on. R software master has now provided more than seven thousand packages, covering basic statistics, sociology, economic management, finance, ecology, biological information and other aspects. Various packages are editable and recompiled, and the user can tailor a unique package to suit their needs. At the same time for the convenience of use, R software package will be packaged, including 34 categories of packages, covering Bayesian, ChemPhys, ClinicalTrials, Cluster and so on, user can quickly install and update.\[1\]

3. Application Analysis of R Software

The main theoretical contents of statistical teaching include regression analysis, discriminant analysis, cluster analysis, principal component analysis, factor analysis, correlation analysis and so on. In the following we will apply R software with specific cases to explore these theories.\[2\]

3.1 Regression Analysis

Regression analysis is a statistical method to determine the quantitative relationship between two or more variables.

Here's an example about how to use R software for regression analysis. According to Keynes' absolute income hypothesis, we estimate the marginal propensity to consume of Chinese residents from 2000 to 2015 by establishing a simple consumption function. The data used in the analysis were based on the data on consumer spending and gross domestic product (GDP) of Chinese residents in the China Statistical Yearbook from 2000 to 2015, and the consumer price index (1978 = 100) was used to eliminate the price factor. The dependent variable was consumption Expenditure C, the independent variable is the gross domestic product (Y).
The analysis result is shown below. We can conclude that the independent variables \( x_1 \) have a positive correlation with the response variable \( y \), and the regression coefficient after the t-test is less than 0.5, and the regression is significant. So the regression analysis is successful.

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

Residuals:
     Min      1Q  Median       3Q      Max
-2590.9  -1779.2   -591.9   1803.9  3361.7

Coefficients:
                     Estimate       Std. Error   t value  Pr(>|t|)
(Intercept)  2.619e+03     1.387e+03     1.888    0.07999 .
income$x1    2.542e+00      5.392e-02     47.150  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2235 on 14 degrees of freedom
Multiple R-squared:  0.9937,    Adjusted R-squared:  0.9933
F-statistic: 2223 on 1 and 14 DF,  p-value: < 2.2e-16
```

3.2 Discriminant analysis

Discriminant analysis is a method for classification and prediction. The following is the examples of the operation of the enterprise by Fisher discriminant analysis. We apply the four economic indicators used to determine whether the enterprise is in bankruptcy or normal operation, namely the total debt ratio (variable 1), yield index (variable 2), short-term payment capacity (Variable 3), production efficiency index (variable 4). While the category is divided into two categories of the bankruptcy status and the normal operation.

Data are shown in Table 2. R software in the Fisher analysis function is lda(). We use Fisher function to analysis the new data. The new data is shown as Category 1.
Table 2. Discriminant analysis of classification reference data.

<table>
<thead>
<tr>
<th>Variable1</th>
<th>Variable2</th>
<th>Variable3</th>
<th>Variable4</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.45</td>
<td>-0.41</td>
<td>1.09</td>
<td>0.45</td>
<td>1</td>
</tr>
<tr>
<td>-0.56</td>
<td>-0.31</td>
<td>1.51</td>
<td>0.16</td>
<td>1</td>
</tr>
<tr>
<td>0.06</td>
<td>0.02</td>
<td>1.01</td>
<td>0.4</td>
<td>1</td>
</tr>
<tr>
<td>-0.07</td>
<td>-0.09</td>
<td>1.45</td>
<td>0.26</td>
<td>1</td>
</tr>
<tr>
<td>-0.06</td>
<td>-0.06</td>
<td>1.37</td>
<td>0.4</td>
<td>1</td>
</tr>
<tr>
<td>-0.13</td>
<td>-0.14</td>
<td>1.42</td>
<td>0.44</td>
<td>1</td>
</tr>
<tr>
<td>0.38</td>
<td>0.11</td>
<td>3.27</td>
<td>0.55</td>
<td>2</td>
</tr>
<tr>
<td>0.19</td>
<td>0.05</td>
<td>2.25</td>
<td>0.33</td>
<td>2</td>
</tr>
<tr>
<td>0.32</td>
<td>0.07</td>
<td>4.24</td>
<td>0.63</td>
<td>2</td>
</tr>
<tr>
<td>0.04</td>
<td>0.01</td>
<td>1.5</td>
<td>0.71</td>
<td>2</td>
</tr>
<tr>
<td>0.07</td>
<td>-0.01</td>
<td>1.37</td>
<td>0.34</td>
<td>2</td>
</tr>
<tr>
<td>0.15</td>
<td>0.06</td>
<td>2.23</td>
<td>0.56</td>
<td>2</td>
</tr>
<tr>
<td>0.16</td>
<td>0.05</td>
<td>2.31</td>
<td>0.2</td>
<td>2</td>
</tr>
</tbody>
</table>

3.3 Cluster analysis

Cluster analysis is also called group analysis, which is a multivariate statistical method for studying the classification of samples (samples or indicators).

R software for clustering analysis is the function hclust(), using four clustering algorithms (the shortest example, the longest distance, the middle distance, ward method) to compare and analyze the application characteristics. The following is the use of the per capita consumption expenditure data (yuan/person) of rural residents in various regions in China, including food and tobacco, clothing, living, daily necessities and services, transportation and communication, education, culture and entertainment, health care, other supplies and services Cluster analysis, the data from the China Statistical Yearbook in 2016.

Using four clustering algorithms (the shortest example, the longest distance, the
middle distance, ward method) to compare the application characteristics. Figure 1 shows the best results in the longest distance analysis.

3.4 Principal components analysis

Principal components analysis was first proposed by Hotelling in 1933. This method was later widely used in various fields. In the analysis of real problems, often involve a lot of indicators, these indicators if you have to consider, it will be very complicated. The principal component analysis method integrates multiple indicators into a few representative indicators with little loss of information, and usually refers to these representative indicators as principal components.

The main component analysis function in the software is princomp(). The following uses this function to perform the principal component analysis on the pollution data from the China Statistical Yearbook in 2016, which contains 5 components.

The results are shown in Figure 2. It can be seen that the contribution of component 1 is 82.6%, which is the component that needs to be analyzed.
3.5 Factor analysis

Factor analysis is an important component of multivariate statistical analysis. Using the R software factanal() function, we select five factors that affect the larger factor from the 15 factors of the reference data. The decomposition results are shown below. The results show that the proportion of the five factors reached 77%, to achieve a better factor analysis and factor selection.

```
Call:
factanal(x = ~., factors = 5, data = X)

Uniquenesses:
X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12
 0.439 0.597 0.509 0.197 0.118 0.005 0.292 0.140 0.365 0.223 0.098 0.119
X13 X14 X15
 0.084 0.005 0.267

Loadings:
Factor1 Factor2 Factor3 Factor4 Factor5
X1  0.127  0.722  0.102 -0.117
X2  0.451  0.134  0.270  0.206  0.258
X3  0.129  0.686
X4  0.222  0.246  0.827
X5  0.917  0.147
X6  0.851  0.125  0.279  0.420
X7  0.228 -0.220  0.777
X8  0.850  0.266  0.111
X9  0.773
X10 0.754  0.393  0.199  0.114
X11 0.909  0.187  0.112  0.165
X12 0.783  0.295  0.354  0.148 -0.181
X13 0.717  0.362  0.446  0.267
X14 0.418  0.399  0.563 -0.585
X15 0.351  0.764  0.148

Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings  5.490  2.507  2.188  1.028  0.331
Proportion Var  0.366  0.167  0.146  0.069  0.022
Cumulative Var  0.366  0.533  0.679  0.748  0.770

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 60.97 on 40 degrees of freedom.
The p-value is 0.0179
```
3.6 Related analysis

Correlation analysis is a statistical method to study the correlation between the two sets of variables, which can effectively reveal the linear dependencies between the two sets of variables.

R software for correlation analysis of the function of cor(), cancor() and so on. Data are statistics on R&D activities and patents of industrial enterprises above designed size by region from the China Statistical Yearbook in 2016. We uses the cor() function to calculate the autocorrelation and interrelated values of all parameters X1-X4. The autocorrelation coefficient of each variable is 1. The correlation coefficient reveals the correlation between the variables. From the results below, we can conclude that the interrelated coefficients of all parameters are also very high, means they are all somehow related.

```r
> ca<-cor(X);
> ca;
          X1     X2     X3     X4
X1 1.000000 0.9690314 0.9654416 0.8428251
X2 0.9690314 1.0000000 0.9156423 0.8442731
X3 0.9654416 0.9156423 1.0000000 0.7127088
X4 0.8428251 0.8442731 0.7127088 1.0000000
```

4. Concluding remarks

Teaching practice proves that R software is a powerful tool in statistical teaching. The introduction of R software into the process of statistical teaching can deepen students 'understanding of statistical methods, better grasp the statistical theory and methods, but also enable students to get rid of cumbersome calculations, stimulate students' interest in learning, increase the motivation of classroom extracurricular learning, The application of professional knowledge to solve practical problems, so as to cultivate more for the community to adapt to the needs of the new economic situation of high quality, business ability of complex, application-based statistical professionals.

In this paper, the application of R software in statistical teaching is introduced. From these examples, R software can effectively solve the traditional "teacher-led" teaching and learning, can become a powerful tool for teaching. This article only briefly describes the R software several test design package, but R software is not limited to a certain type of problem, with different functions of the expansion package, R software can be applied to a wide range of its simple and easy to operate Other similar software is difficult to match.

REFERENCES
1. http://www.r-project.org
2. http://cran.r-project.org/bin/windows/base/