Advances in Social Science, Education and Humanities Research (ASSEHR), volume 184
2nd International Conference on Education Science and Economic Management (ICESEM 2018)

# Campus Network User Demand Forecasting Model Based on Multiple Linear Regression

Xuefeng li
Zhongnan University of Economics and Law
Wuhan, China

***Abstract*—By analyzing the relationship between campus network users and online processes, a campus network user demand forecasting model based on multiple linear regression is proposed. Firstly, influencing factors are analyzed through scatter plots and trend lines, and the main factors affecting the Internet access needs of campus users are selected, namely, service reachability, service response time, service interruption rate, service quality, and service ease of use. Secondly, the statistical software SPSS 22 is used to construct a multiple linear regression model based on least squares method, and the multi-collinearity problem between independent variables is eliminated by ridge regression analysis to obtain modified model. Finally, the model is used to meet the needs of campus users. The prediction is carried out, and the result shows that the accuracy of the predicted estimation value is high, indicating that the model has good fitting degree and has certain practicability and reference significance.**

***Keywords—Campus network; User demand; Least squares method; Ridge regression***

## I. INTRODUCTION

In the "Statistical Report on the Development of China's Internet Network", it is pointed out that the growth of demand for network users has become the driving force for the development of the Internet environment [1]. The Internet access needs of campus network users are the driving force for the development of campus network environment. According to the theory of supply and demand, in a certain period of time, when the environmental quality of the campus network can not meet the needs of users, it will inhibit the demand; on the contrary, when the supply exceeds the demand, it will inevitably cause waste of supply. Therefore, user demand is the basis for the development of campus network environment. User demand forecast analysis is based on this theory to provide an important theoretical basis for school information construction, to achieve a balance between the needs of campus network users and the campus network environment. Sex. At present, the empirical research on the factors affecting the demand of campus network users in China is relatively weak, and most of the research focuses on the analysis between the needs of campus network users and network service technologies.

This paper starts from the perspective of the close relationship between the online users of the campus network, and introduces the user behavior factors in human behavior, combined with the multiple linear regression method which has advantages in dealing with multi-factor influence problems and has the advantage of eliminating multi-factor multi-colinearity. The ridge regression method is used to construct a campus network user demand forecasting model based on multiple linear regression. The model analyzes the factors affecting user demand and realizes the forecasting function, and the empirical effect is good. Compared with the analysis method that affects the needs of campus network users from the perspective of network service technology, the model has deep data analysis, strong explanatory ability, high accuracy of prediction results, and has certain practical value and reference significance.

## II. CONSTRUCTING MULTIPLE LINEAR REGRESSION MODELS

In linear regression analysis, with two or more independent variables, called multiple linear regression, regression analysis can solve different problems more easily and practically. Regression analysis is one of the most commonly used statistical tools to describe the variation of a dependent variable and an independent variable. The goal of regression analysis is to find a suitable mathematical model and determine the best fit coefficient of the model from the given data. . Since the output variable spans a continuous range of values and the effect of the input on the output is known, using regression techniques is a viable solution for developing predictive models.

### *A. Sample screening*

The number of users' Internet connection usage in the past six years was selected as the test data set, including service reachability, business response time, service interruption rate, service quality, and service ease of use [11]. Due to the relationship between users' online processes, if we want to predict user needs, the factors that affect user needs may involve the subjective perception of users. From the subjective and objective aspects of this paper, the business accessibility, business response time, business interruption rate, service quality, and business usability are selected as the influencing factors from the test data set.

### *B. Selecting independent variables*

The linear relationship between the independent variable and the dependent variable is obtained by combining the scatter plot and the trend line, and the independent variable [2] is selected. X1, X2, X3, X4, and X5 are independent variables, and Y is a dependent variable, where X1 is service reachability, X 2 is service response time, X3 is service interruption rate, X4 is service quality, and X5 is service ease of use. . Establish user requirements and scatter plots of various influencing factors.

Through analysis, it can be seen that there is a positive linear correlation between service reachability, service response time, service interruption rate, service quality, service ease of use and campus network user demand. The positive linear correlation between service reachability and service quality and campus network user requirements is particularly significant. The linear correlation between business response time and service interruption rate and campus network user demand is second, and business ease of use. The linear correlation with the needs of campus network users is general.

By observing the detailed information of each trend line listed in Table 1, the determination of the influencing factors is comprehensively determined, and the least square method is combined with the trend line. When R2≤1, it is most reliable to describe the trend line with R2. . Therefore, these five factors affecting the needs of the campus network can be used as input variables.

TABLE I INFORMATION ON EACH TREND LINE

| Column | Row | R2 | Standard error | P value |
|---|---|---|---|---|
| Y | X1 | 0.9477 | 233.1171 | <0.0001 |
| Y | X2 | 0.7799 | 354.1240 | <0.0001 |
| Y | X3 | 0.7571 | 220.7612 | <0.0001 |
| Y | X4 | 0.8822 | 430.3400 | <0.0001 |
| Y | X5 | 0.6193 | 130.0000 | <0.0001 |

### *C. Constructing multiple linear regression models by least squares method*

Through the statistical software SPSS 22, a multivariate linear regression model is established for the five variables of business accessibility, business response time, business interruption rate, service quality and business usability, and the fitting results are calculated as follows:

TABLE II FITNESS TEST

| R | Before adjusting R2 | Adjusted R2 | Standard estimated error |
|---|---|---|---|
| 0.997 | 0.994 | 0.993 | 42.4685 |

The larger the R2 value, the higher the covariate ratio of the two variables reflected. When the R2 value of the trend line is close to 1, the trend line is the most reliable, and the fit of the model to the data is better. Analysis Table 2 shows that the adjusted fitness R2=99.3%, and P<0.0001, indicating that the multivariate linear regression model obtained by the least squares method has good fitting degree, and the independent variable has a significant influence on the dependent variable.

TABLE III LEAST SQUARES ESTIMATION RESULTS

| | Unnormalized coefficient | Standardization coefficient |
|---|---|---|
| | B | β |
| constant | 5838.37 | |
| X1 | 0.4308 | 0.259 |
| X2 | -8.8519 | -0.14 |
| X3 | 0.0738 | 0.357 |
| X4 | 1.1036 | 0.911 |
| X5 | -1.3331 | -0.03 |

TABLE IV T VALUE, SIGNIFICANCE, VIF

| | t value | Significant | VIF |
|---|---|---|---|
| constant | 2.6240 | 0.0250 | |
| X1 | 6.8300 | 0.0000 | 18.9900 |
| X2 | -9.7300 | 0.0000 | 17.9100 |
| X3 | 22.7800 | 0.0000 | 7.7740 |
| X4 | 5.2900 | 0.0000 | 9.4410 |
| X5 | -1.1910 | 0.7300 | 5.5770 |

According to the multiple linear regression results of the least squares method, the linear regression equation of Y to five independent variables can be obtained as follows:

Y=5838.37+0.4308X1-8.8519X2+0.0738X3+1.1036X4-1.3331X5

## III. MULTICOLLINEARITY ANALYSIS

Multicollinearity was proposed by Frisch in 1934 [3]. Complex collinearity is a strong correlation between features that affect the target vector at the same time, making the model estimation distorted or difficult to estimate accurately. Complex collinearity is a problem in regression analysis. For example, the least square method is used to establish an excessively complex unstable model, and the significance check of various variables is distorted, resulting in a large VIF value of the variance expansion factor, resulting in distortion of the predicted value. According to the analysis table 5, the values of the various conditional index values are weak, and the variance of the parameter estimates are greater than 2, indicating that the collinearity between the variables is strongly correlated, and the collinearity between the independent variables needs to be solved problem.

TABLE V COLLINEAR DIAGNOSIS

| Eigenvalues | Conditional indicator | constant | X1 | X2 | X3 | X4 | X5 |
|---|---|---|---|---|---|---|---|
| 7.69 | 1.00 | 0.00 | 0.00 | 0.00 | 0.60 | 0.00 | 0.00 |
| 0.02 | 15.77 | 0.00 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 |
| 0.02 | 25.00 | 0.00 | 0.30 | 0.00 | 0.50 | 0.02 | 0.00 |
| 0.00 | 35.08 | 0.00 | 0.20 | 0.40 | 0.86 | 0.07 | 0.00 |
| 0.00 | 59.30 | 0.00 | 0.80 | 0.20 | 0.28 | 0.09 | 0.10 |

## IV. MODELING USING RIDGE REGRESSION

In 1962, Hoerl first proposed the ridge regression analysis. After several years, he systematically studied the ridge regression analysis with Kennard [4]. Ridge regression analysis is essentially a kind of collinearity problem ($|XTX| \approx 0$) for solving the independent variables in linear regression analysis. By adding a unit matrix kI(k>0), the degree of alienation is reduced. The improved least squares estimation method, that is, the estimation model of normalized ridge regression, where K is called the ridge parameter.

When the ridge regression analysis was performed using SPSS 22, the calculation results are shown in Table 6. Observed that the change of K value leads to the change of the coefficient value of each corresponding variable, so the choice of K value should be based on the combination of quantitative data analysis and theoretical reasoning.

TABLE VI CHANGES IN REGRESSION COEFFICIENTS OF DIFFERENT RIDGES

| K value | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| 0.010000 | 5.8604 | 7.0244 | 5.8972 | 9.3540 | 9.8590 |
| 0.020000 | 4.7172 | 5.5528 | 4.7834 | 6.2177 | 6.5976 |
| 0.030000 | 3.9382 | 4.5230 | 4.0254 | 4.5146 | 4.8127 |
| 0.040000 | 3.3710 | 3.7725 | 3.4700 | 3.4822 | 3.7219 |
| 0.050000 | 2.3240 | 2.4246 | 2.4253 | 1.9938 | 2.1278 |
| 0.060000 | 2.0974 | 2.1460 | 2.1942 | 1.7362 | 1.8477 |
| 0.070000 | 2.9390 | 3.2073 | 3.0428 | 2.8055 | 3.0014 |
| 0.080000 | 2.5988 | 2.7702 | 2.7028 | 2.3355 | 2.4970 |
| 0.090000 | 1.9075 | 1.9178 | 1.9988 | 1.5361 | 1.6290 |
| 0.100000 | 1.7461 | 1.7282 | 1.8313 | 1.3767 | 1.4540 |
| 0.200000 | 0.8998 | 0.8160 | 0.9334 | 0.6765 | 0.6803 |
| 0.300000 | 0.5739 | 0.5076 | 0.5827 | 0.4473 | 0.4323 |
| 0.400000 | 0.4072 | 0.3588 | 0.4053 | 0.3309 | 0.3113 |
| 0.500000 | 0.3086 | 0.2729 | 0.3021 | 0.2599 | 0.2402 |
| 0.600000 | 0.2446 | 0.2718 | 0.2364 | 0.2122 | 0.1938 |
| 0.700000 | 0.2004 | 0.1797 | 0.1917 | 0.1780 | 0.1614 |
| 0.800000 | 0.1684 | 0.1521 | 0.1599 | 0.1525 | 0.1376 |
| 0.900000 | 0.1443 | 0.1312 | 0.1362 | 0.1328 | 0.1195 |
| 1.000000 | 0.1257 | 0.1150 | 0.1181 | 0.1172 | 0.1053 |

Let $0.01 \leq K \leq 1$ and increase the step size to 0.01. By analyzing Table 6, it can be seen that as the K value increases, the VIF value becomes smaller and smaller, and when the K value increases to 0.1 to 0.3, The ridge regression estimates have higher stability, and the collinear effects between the variables are basically eliminated. It is reasonable to use the least squares method to estimate the value.

TABLE VII STANDARDIZED REGRESSION COEFFICIENTS, STANDARD ERRORS, VIF VALUES AT K = 0.200000

| | Standardized regression coefficient | Standard error | VIF value |
|---|---|---|---|
| X1 | 0.1872 | 0.0117 | 0.8998 |
| X2 | 0.0264 | 0.0760 | 0.8160 |
| X3 | 0.0537 | 0.0260 | 0.9334 |
| X4 | 0.2843 | 0.0210 | 0.6765 |
| X5 | 0.0371 | 1.0015 | 0.6803 |

The adjusted fitness R2=97.44%, and the fitness of the demand prediction model meets the requirements. The adjusted ridge regression equation is: Y=-1424.79+0.30X1++0.89X2+0.42X3+0.29X4+2.54X5

Using the ridge regression equation obtained above to predict the user demand for the next year, the results show that the regression curve has a good fit, which indicates that the reliability of the predicted value is high.

Table 8 shows the predicted and real value statistics of the campus network users in the past 7 years. According to the analysis, the accuracy of the prediction model is high, and the prediction accuracy is over 90%. The error between the predicted value and the true value is better. Small, the forecast results basically meet the expected requirements, and have high practical and reference value.

TABLE VIII PREDICTED AND ACTUAL VALUES OF CAMPUS NETWORK USER REQUIREMENTS

| years | Predictive value | Actual value | Accuracy |
|---|---|---|---|
| 2011 | 21552 | 23149 | 0.931 |
| 2012 | 21996 | 23301 | 0.944 |
| 2013 | 21618 | 23195 | 0.932 |
| 2014 | 22601 | 23841 | 0.948 |
| 2015 | 21117 | 23028 | 0.917 |
| 2016 | 21319 | 23123 | 0.922 |
| 2017 | 22300 | 23449 | 0.951 |

Based on the analysis of the factors affecting user needs, this paper establishes a multivariate linear regression model based on five factors: business accessibility, business response time, business interruption rate, service quality and business usability. The collinearity problem is introduced. Therefore, the ridge regression analysis method is introduced to eliminate the multicollinearity existing between the variables. The empirical results show that the obtained campus network user demand forecasting model has a good fitting effect and has certain theoretical significance and reference value.

## REFERENCES

[1] Statistical Report on the Development of China's Internet, http://www.cac.gov.cn/2018-01/31/c_1122347026.htm.

[2] Tian Jiule, Zhao Wei. A method for calculating the similarity of words based on the synonym word forest [J]. Journal of Jilin University (Information Science Edition) 2010, 28(6): 602-608.

[3] Wen Bin, He Tingting, Luo Le, et al. Research on text sentiment classification based on semantic understanding [J]. Computer Science, 2010, 37(6): 261-264.

[4] Ma Li, Liu Xiao, Gong Yulong. Semantic-based microblog short text orientation analysis [J]. Computer Application Research, 2016, 33(10): 2914-2918.