Linear Regression Model of Railway Passenger Traffic Volume and Its Influencing Factors

Yue Zhang1, *
1Dept of Mathematics, University College London, London, UK
*yue_zhang.20@ucl.ac.uk

Abstract: The railway is one of the most crucial transportation modes in China and plenty of people choose to travel by railway. Finding the influencing factors of passenger traffic volume of the railway can help us to develop the railway construction. Substantial research has probed factors influencing the railway passenger volume and has identified several influencing factors. In this study, we try to find a long-term relationship between passenger traffic volume of the railway and six influencing factors, which are total population, employed persons, number of national railway passenger coaches owned, GDP, and household consumption expenditure. In order to find long term relationship models, we collected the data of passenger traffic volume of the railway and other six influencing factors from 1978 to 2019. By developing several multiple linear regression models, we can see that all of these six factors can impact the passenger traffic volume of the railway and the model of six factors can describe the relationship well. After simplifying the model with six factors, we find two more models that can describe the relationship with less influencing factors, which are $y = 1.059 \times 10^5 - 1.773 \times 10^2 x_1 + 2.729 \times 10^{-3} x_2 + \epsilon$ and $\log(y) = 11.5 - 3.309 \times 10^{-5} x_1 + 1.435 \times 10^{-6} x_5 + \epsilon$. Using the diagnostic plots to analyze these two models, we find that $\log(y) = 11.5 - 3.309 \times 10^{-5} x_1 + 1.435 \times 10^{-6} x_5 + \epsilon$ is better. According to the model, GDP and traffic volume are two main influencing factors on the railway passenger volume. We use this model to predict the railway passengers in 2019 and compare the predict value with the data. The error rate is 5.4688% and we believe that this model can also be used to predict the railway passenger volume.

1. Introduction

China is a country with a large population and a vast territory. Railway transportation has the advantages of fast speed, large transportation volume, low cost, and strong adaptability compared with waterway, highway, pipeline, and air transportation. Railway plays a vital role in the transportation system. After years of development, the speed of high-speed railway in China has reached 350km per hour. An increasing number of people choose the railway when they travel.

Although the railway passenger carrying capacity and the speed is increasing, it is still hard to buy a ticket during the spring festival. By developing a model between influencing factors and the passenger traffic volume of the railway, we can know the relationship between them and predict the traffic volume when necessary.

According to Nan, influencing factors of the passenger traffic volume includes economic characteristics, population characteristics, geographical factors, and cultural tourism [1].

Through the structure interpret model, Guo indicated that the main influencing factors are population, household consumption, transportation mode, transportation cost, and the industry and commerce development [2].
Zhang applied a multiple linear regression model to analyze the relationship between passenger traffic volume of the railway and the factors including GDP, natural population growth rate, railway mileage, household consumption level, and domestic tourists. The data was from 1991 to 2011. GDP, household consumption level, and population have a more significant impact compared with other factors [3]. One can refer to [4-10] and references therein for more details.

This paper will use the data from 1978 to 2019 and analyze to find the main impact factors of passenger traffic volume of the railway. In order to clearly describe the relationship between passenger traffic volume of the railway and other influencing factors, we will develop a multiple linear regression model by given data and use the model to do prediction. Moreover, we will find the main factor that can influence the passenger traffic volume of the railway.

2. Data and Method

2.1 Data

According to Nan’s research [1], the influencing factors include population, national policy, economic development level, and railway construction. We choose six influencing factors: Passenger Traffic (10000 persons), Total Population (year-end) (10000 persons), Employed Persons (10000 persons), Number of National Railway Passenger Coaches Owned (coach), GDP (100 million yuan) and Household Consumption Expenditure (yuan). In order to find a long-term relationship between the passenger traffic volume of the railway and these six factors, we select the data from 1978 to 2019. However, we only use the data from 1978 to 2018 to develop the model, as we didn’t find the data of the Numbers of coaches owned in 2019. Table 1 and Table 2 show the data we will use.

| Year | Railway passengers | Passenger Traffic | Total Population | Employed Persons |
|------|--------------------|-------------------|-----------------|------------------|
| 1978 | 81491              | 253993            | 96259           | 40152            |
| 1979 | 86389              | 289665            | 97542           | 41024            |
| 1980 | 92204              | 341785            | 98705           | 42361            |
| 1981 | 95219              | 384763            | 100072          | 43725            |
| 1982 | 99922              | 428964            | 101654          | 45295            |
| 1983 | 106044             | 470614            | 103008          | 46436            |
| 1984 | 113353             | 530217            | 104357          | 48197            |
| 1985 | 112110             | 620206            | 105851          | 49873            |
| 1986 | 108579             | 688211            | 107507          | 51282            |
| 1987 | 112479             | 746422            | 109300          | 52783            |
| 1988 | 122645             | 809592            | 111026          | 54334            |
| 1989 | 113804.6           | 791373.6          | 112704          | 55329            |
| 1990 | 95712              | 772682            | 114333          | 64749            |
| 1991 | 95080              | 806048            | 115823          | 65491            |
| 1992 | 99693              | 860855            | 117171          | 66152            |
| 1993 | 105458             | 996634            | 118517          | 66808            |
| 1994 | 108738             | 1092882           | 119850          | 67455            |
| 1995 | 102745             | 1172596           | 121121          | 68065            |
| 1996 | 94797              | 1245357           | 122389          | 68950            |
| 1997 | 93308              | 1326094           | 123626          | 69820            |
| 1998 | 95085              | 1378717           | 124761          | 70637            |
| 1999 | 100164             | 1394413           | 125786          | 71394            |
| 2000 | 105073             | 1478573           | 126743          | 72085            |
| 2001 | 105155             | 1534122           | 127627          | 72979            |
In Table 1, the data collected from National Bureau of Statistics of China. Railway passengers represent Passenger Traffic of Railways (10000 persons), Passenger Traffic represents Passenger Traffic (10000 persons), Total population means Total Population (year-end) (10000 persons) and Employed Persons describes Employed Persons (10000 persons).

Table 2. Railway passengers and influencing factors (Part 2)

| Year | Railway Passengers | Number of Coaches Owned | Household | GDP Consumption |
|------|--------------------|-------------------------|-----------|-----------------|
| 1978 | 81491              | 15029                   | 3678.7    | 183             |
| 1979 | 86389              | 15355                   | 4100.5    | 206             |
| 1980 | 92204              | 16367                   | 4587.6    | 237             |
| 1981 | 95219              | 17075                   | 4935.8    | 263             |
| 1982 | 99922              | 17788                   | 5373.4    | 282             |
| 1983 | 106044             | 18759                   | 6020.9    | 313             |
| 1984 | 113353             | 19682                   | 7278.5    | 354             |
| 1985 | 112110             | 20872                   | 9098.9    | 437             |
| 1986 | 108579             | 22138                   | 10376.2   | 492             |
| 1987 | 112479             | 23474                   | 12174.6   | 553             |
| 1988 | 122645             | 24917                   | 15180.4   | 678             |
| 1989 | 113804.6           | 26304                   | 17179.7   | 779             |
| 1990 | 95712              | 27261                   | 18872.9   | 825             |
| 1991 | 95080              | 27612                   | 22005.6   | 910             |
| 1992 | 99693              | 28464                   | 27194.5   | 1051            |
| 1993 | 105458             | 29395                   | 35673.2   | 1324            |
| 1994 | 108738             | 31018                   | 48637.5   | 1789            |
| 1995 | 102745             | 32404                   | 61339.9   | 2317            |
| 1996 | 94797              | 34516                   | 71813.6   | 2749            |
| 1997 | 93308              | 35171                   | 79715     | 2959            |
| 1998 | 95085              | 35204                   | 85195.5   | 3107            |
| 1999 | 100164             | 35317                   | 90564.4   | 3327            |
A multiple linear regression model is a model with many regressor variables, which can be used to find the relationship between a response variable and regressor variables. The multiple linear regression model is

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \]  

(1)

where the parameters \( \beta_i \), \( i = 0, 1, \ldots, k \) are known as the regression coefficients, and \( \epsilon \) is a random error component. We assume the error satisfies a normal distribution with mean zero and the errors are independent [4].

First, we need to use the sample data to get an estimate of the parameters. We use the least-squares method to estimate the regression coefficients. Let RSS be the sum of square residuals.

\[ RSS = \sum (y_i - \hat{y}_i)^2 \]  

(2)

\[ RSS = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \hat{\beta}_2 x_{2,i} - \ldots - \hat{\beta}_k x_{k,i})^2 \]  

(3)

We need to minimize RSS, so we use partial differential equations to get the minimum value of RSS.

\[ \frac{\partial RSS}{\partial \hat{\beta}_0} = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \hat{\beta}_2 x_{2,i} - \ldots - \hat{\beta}_k x_{k,i}) (-1) = 0 \]  

\[ \frac{\partial RSS}{\partial \hat{\beta}_1} = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \hat{\beta}_2 x_{2,i} - \ldots - \hat{\beta}_k x_{k,i}) (-x_{1,i}) = 0 \]  

\[ \vdots \]  

\[ \frac{\partial RSS}{\partial \hat{\beta}_k} = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \hat{\beta}_2 x_{2,i} - \ldots - \hat{\beta}_k x_{k,i}) (-x_{k,i}) = 0 \]  

(4)

Then we can get \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k \). Next, we need to do the hypothesis test on the parameters. We need to check whether there is a relationship between \( x_i \) and \( y \), which means we need to check whether all of the regression coefficients \( \beta_i \) equals to zero. We establish a hypothesis test,

\[ H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \]
\[ H_1 : \beta_i \neq 0, \text{for some } i \in \{1, \ldots, k\} \]  
\hspace{1cm} \text{(5)}

We use F-statistics to find the result. \( TSS = \sum(y_i - \bar{y})^2 \) and \( F = \frac{(TSS - RSS)/k}{RSS/(n-k-1)} \).

Let \( \alpha = 0.05 \), when the p-value is greater than 0.05, \( H_0 \) is true.

3. Results and Discussion

Since we don’t have the data of Numbers of Coaches owned in 2019, we use the data from 1978 to 2018 to develop the model. First, we want to know the relationship between railway passenger volume and each factor. Scatter plot Matrix can help us to consider the relationship between pairwise variables in a series of 2 dimension plots. Here is the Scatter plot Matrix. From Figure 1, we can see the relationship between each two variables.

![Figure 1. Scatterplot Matrix](image)

From the Scatter plot matrix, we can see that railway passenger has clear linear relationship with several factors, such as total population, employed population and GDP. The relationship between railway passenger and GDP is like a straight line, so GDP might be a main factor compared with other influencing factors.

Since we want to find a formula to describe the relationship, we develop a linear regression model. In this case, we have six factors and we use

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon \]  
\hspace{1cm} \text{(6)}

where \( x_i \) represents Passenger Traffic, Total population, Employed Persons, Number Coaches Owned represents, GDP and Household Consumption, and \( y \) is Railway passengers. Figure 2 below shows the result.
From the figure above we can see that the adjusted R-squared is 0.9692, and the p-value of F-statistics is $2.2 \times 10^{-16}$. R-squared and p-value of F-statistics can show how well the model represents the given data. R-squared can help us know whether our model is good. If the R-squared is larger, the model is better. Since the p-value of F-statistics is smaller than 0.05, we can say the model is significant. This is a good model to describe the relationship between the factors and the passenger traffic of the railway. Since there are too many variables in this model, it is complicate to use this model to predict. We want a simpler model. In order to simplify our model, we need to reduce the factors. We use both the forward linear regression method and the backward linear regression method to get the most important factors.

Call:
\[ \text{lm(formula} = y \sim x1 + x2 + x3 + x4 + x5 + x6) \]

Residuals:

|       | Min   | 1Q    | Median | 3Q    | Max  |
|-------|-------|-------|--------|-------|------|
|       | -22688 | -6922 |        | 66    | 7969 | 21111|

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 1.200e+05 | 1.168e+05 | 1.028   | 0.31133  |
| x1             | -1.993e-02 | 7.077e-03 | -2.816  | 0.00805 **|
| x2             | 2.579e-02  | 1.707e+00 | 0.015   | 0.98803  |
| x3             | -1.387e+00 | 1.285e+00 | -1.079  | 0.28814  |
| x4             | 2.850e+00  | 1.675e+00 | 1.701   | 0.09800  |
| x5             | 4.609e-01  | 2.321e-01 | 1.986   | 0.05518  |
| x6             | -1.166e+01 | 1.820e+01 | -1.144  | 0.26077  |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11110 on 34 degrees of freedom
Multiple R-squared: 0.9738,  Adjusted R-squared: 0.9692
F-statistic: 210.6 on 6 and 34 DF,  p-value: < 2.2e-16

Figure 2. Results of \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon \)

We want a simpler model. In order to simplify our model, we need to reduce the factors. We use both the forward linear regression method and the backward linear regression method to get the most important factors.

Call:
\[ \text{lm(formula} = y \sim x5 + x1) \]

Residuals:

|       | Min   | 1Q    | Median | 3Q    | Max  |
|-------|-------|-------|--------|-------|------|
|       | -20926 | -6977 | -1654  | 7479  | 26942|

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 1.059e+05 | 3.430e+03 | 30.880  | < 2e-16 ***|
| x5             | 2.729e-01  | 9.100e-03 | 29.990  | < 2e-16 ***|
| x1             | -1.773e-02 | 2.658e-03 | -6.671  | 6.87e-08 ***|

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11230 on 38 degrees of freedom
Multiple R-squared: 0.97,  Adjusted R-squared: 0.9685
F-statistic: 615.3 on 2 and 38 DF,  p-value: < 2.2e-16

Figure 3. Results of the forward linear regression method
Figure 3 and Figure 4 indicate that the forward linear regression method and the backward linear regression method leads to the same result. The most critical factors are $x_1$ and $x_5$, which are Passenger Traffic and GDP, so we use these two factors to develop a model

$$y = 1.059 \times 10^5 - 1.773 \times 10^{-2} x_1 + 2.729 \times 10^{-1} x_5 + \epsilon.$$ (7)

Although the adjusted R-squared (0.9685) and the p-value of F-statistics ($2.2 \times 10^{-16}$) shows that this model is significant and represents the given data well, we still want to know how poorly the model describes those data and whether this model is acceptable. Residual can indicate whether a model is acceptable. Diagnostic plots show residual in different ways. We use diagnostic plots to test whether this model is acceptable.

The residual vs fitted plot describes whether railway passengers and these two factors has nonlinear relationship. If residuals spread equally around a horizontal line, then we can say the variables don’t have nonlinear relationship. The Normal Q-Q plot can help us check whether residuals satisfy normal distribution, if residuals are distributed normally, then we can see a straight line in the Normal Q-Q plot. The Scale-Location plot shows whether the standardized residuals spread equally around a horizontal line. The Residual vs Leverage tells us which data is special. We need to focus on the spots that are outside the dashed lines. These spots may influence our regression.

Figure 5 is the diagnostic plots of $y$. The residuals vs. fitted plot shows the red line is not approximately horizontal at zero, but from the Residual vs Leverage plot we know that there are no influential case in our model. Now, we try to optimize the model to get a better one.
One of the most popular ways to optimize the model is using logarithmic model. We take the logarithm of railway passenger and then develop a model.

Call:
`lm(formula = log(y) ~ x1 + x5)`

Residuals:

|       | Min     | 1Q    | Median | 3Q     | Max    |
|-------|---------|-------|--------|--------|--------|
|       | -0.187564 | -0.048594 | -0.005612 | 0.047133 | 0.223113 |

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 1.150e+01 | 2.727e-02  | 421.745 | <2e-16 *** |
| x1             | -3.309e-08 | 2.113e-08 | -1.566 | 0.126    |
| x5             | 1.435e-06  | 7.234e-08  | 19.842 | <2e-16 *** |

---

Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08929 on 38 degrees of freedom
Multiple R-squared: 0.9445,  Adjusted R-squared: 0.9415
F-statistic: 323.1 on 2 and 38 DF,  p-value: < 2.2e-16
Figure 6 shows the result of model log(y). R-squared value is 0.9415, which is smaller than the R-squared value of the previous model $y = 1.059 \times 10^5 - 1.773 \times 10^{-2} x_1 + 2.729 \times 10^{-1} x_5 + \epsilon$.

In order to understand this model better, we get the diagnostic plots of the model $\log(y) = 11.5 - 3.309 \times 10^{-8} x_1 + 1.435 \times 10^{-6} x_5 + \epsilon$.

The residuals vs. fitted plot and the scale-location plot in Figure 7 show the red line is approximately horizontal at zero. The Normal Q-Q plot indicates residuals follow a straight line well. In residuals vs. leverage, as all cases are well inside the Cooks distance lines, we say there are no influential cases. According to Figure 5 and Figure 7, $\log(y) = 11.5 - 3.309 \times 10^{-8} x_1 + 1.435 \times 10^{-6} x_5 + \epsilon$ (8) is a better regression model compared with $y = 1.059 \times 10^5 - 1.773 \times 10^{-2} x_1 + 2.729 \times 10^{-1} x_5 + \epsilon$.

We use the log(y) model to predict the passenger traffic volume of the railway in 2019. $\log(y) = 11.5 - 3.309 \times 10^{-8} x_1 + 1.435 \times 10^{-6} x_5 + \epsilon$ $\log(y) = 11.5 - 3.309 \times 10^{-8} \times 1760435.7 + 1.435 \times 10^{-6} \times 990865.1$ $\log(y) = 12.86364$ $y = 386018.2781$

Table 1 shows that the true value of passenger traffic volume of the railway’s passenger traffic volume in 2019 is 366002.3, and the error rate is 5.4688%.

4. Conclusion

We develop several multiple linear regression models which can describe the relationship between passenger traffic volume of the railway and passenger traffic volume, total population, employed persons, number of national railway passenger coaches owned, GDP, and household consumption expenditure. By reducing variables and optimizing the model, we get a logarithmic model, which has less variables and can represent the given data better. From the model we used, we know that these six factors influence
the passenger traffic volume of the railway, and there are two most essential factors, passenger traffic volume, and GDP. The \( \log(y) \) model can be used to predict, but we cannot test the accuracy of prediction due to the lack of data. By analyzing the data and developing the model, we find the primary and the secondary influencing factors. The increasing passenger traffic volume of railway mainly depends on the increase in GDP and traffic volume. Total population, employed persons, the number of national railway passenger coaches owned, and household consumption expenditure can also increase passenger traffic volume of the railway. Finding the relationship between passenger traffic volume of the railway and the factors can let us control the passenger traffic volume of the railway more easily.

References
[1] Nan, J. L. (2016). The Analysis of Influencing Factor about Railway Passenger Forecast. China Transportation Review, Vol. 38, No. 2, 35-40.
[2] Guo, Z. Z. (2007). Hierarchy Analysis on the Railway Passenger Flow Factors. Journal of Transportation Engineering and Information, Vol. 5, No. 4, 68-88.
[3] Zhang, X. J. (2014). Empirical Analysis on the Influencing Factors of Railway Passenger Transport Volume in China. Modern Business Trade Industry, No. 9, 5-7.
[4] Montgomery, D. C. Peck, E. A. & Vining, G. G. (2012). Introduction to Linear Regression Analysis, Fifth Edition, John Wiley & Sons, Inc..
[5] James, G. Witten, D. Hastie, T. & Tibshirani, R. (2013). An Introduction to Statistical Learning, Springer.
[6] Johnson, R.A and Wichern, D.W. (2003). Applied Multivariate Statistics Analysis. Prentice Hall.
[7] Krzanowski, W.J. (1988). Principles of Multivariate Analysis: A User's Perspective.
[8] Sharma, S. (1996). Applied Multivariate Techniques.
[9] Chatfield, C. and Collins, A.J. (1980). Introduction to Multivariate Analysis.
[10] Manly, B.F.J. (1986). Multivariate Statistical Methods: a Primer.