Analysis of the Influencing Factors of the Multi-Linear Regression Model based on R Language on the Total Cost of Domestic Tourism

Youwei Ran¹,a, Xiufang Yuan¹,b

¹School of mathematics and statistics, Sichuan University of Science & Engineering, Yibin 644000, China.

a³youwei99@163.com, byxiufang1990@163.com

Abstract
With the development of the country, China has entered a decisive year to build an all-round well-off society. Therefore, in order to achieve the goal of an all-round well-off society, we should pay more attention to the contribution of tourism to the national economy and it also plays an important role in the development of the rural economy. China’s territory is vast and culturally diverse, so the development of tourism is full of great potential. Against the background of tourism, this paper shows the process of multiple regression modelling and the analysis method of regression model using R language to influence the total cost of tourism in China.

Keywords
R language; Multivariate linear regression model; Tourism; All subsets regression.

1. Introduction
China is still the largest developing country in the world. People are full of the yearning for a better cultural life. Tourism can not only meet people’s growing needs for a better cultural life, but also promote the country’s economic development. So with the development of society, tourism increasingly shows its important position in the national economy. We should follow General Secretary Xi’s green ecological civilization thought that Lucid waters and lush mountains are invaluable assets, combine it with General Secretary’s important strategic thought of poverty alleviation and development, and strive to develop tourism⁴. In this paper, we will use the R language to model the explanatory variables that affect the total cost of domestic tourism. Although the R language is difficult to get started when you have a certain mastery of R language and the corresponding package, its rich, powerful and flexible functions will greatly help you improve your ability to process and analyze data. And R, as a free, open source and powerful drawing software, can help you get more comprehensive visualized data. There are many functions to establish linear regression models in R. In this paper, the \( \text{lm}() \) function in the R-based installation package is mainly used to establish a linear regression model.

2. The selection of index
2.1. Choose variables
\( y \): The total cost of domestic tourism (Unit: Hundred million yuan)  
\( x_1 \): Number of domestic tourists (Unit: million)  
\( x_2 \): Passenger turnover (Unit: Hundred million person kilometers)  
\( x_3 \): Major public health emergencies
The left of "~" is the interpreted variable and its right is the explanatory variable. The predicted variable is separated by " + " symbols in R. The interactive interpretation variables are separated by " : " . " : " represents all variables except the interpreted variables.

The original model established by R in this paper is model1 <- lm(y ~ ., data = mydata). The original fitted model name is model 1, and mydata is the data imported into R in Table 1. The import format of is a data box. The detailed results of the fitting model can be displayed with

Table 1 Data table

<table>
<thead>
<tr>
<th>Year</th>
<th>y</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
<th>x6</th>
<th>x7</th>
<th>x8</th>
<th>x9</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>3176</td>
<td>744</td>
<td>12261</td>
<td>0</td>
<td>687.9</td>
<td>9.9</td>
<td>434</td>
<td>6256</td>
<td>45906</td>
<td>380</td>
</tr>
<tr>
<td>2001</td>
<td>3522</td>
<td>784</td>
<td>13155</td>
<td>0</td>
<td>650.3</td>
<td>8.5</td>
<td>437</td>
<td>6824</td>
<td>48064</td>
<td>450</td>
</tr>
<tr>
<td>2002</td>
<td>3878</td>
<td>878</td>
<td>14126</td>
<td>0</td>
<td>59.1</td>
<td>6.9</td>
<td>433.5</td>
<td>7652</td>
<td>50212</td>
<td>552</td>
</tr>
<tr>
<td>2003</td>
<td>3442</td>
<td>870</td>
<td>13811</td>
<td>24.6</td>
<td>1676.9</td>
<td>6.8</td>
<td>438.7</td>
<td>8406</td>
<td>52376</td>
<td>686</td>
</tr>
<tr>
<td>2004</td>
<td>4711</td>
<td>1102</td>
<td>16309</td>
<td>0</td>
<td>305</td>
<td>5.1</td>
<td>455.8</td>
<td>9335</td>
<td>54283</td>
<td>820</td>
</tr>
<tr>
<td>2005</td>
<td>5286</td>
<td>1212</td>
<td>17467</td>
<td>0</td>
<td>742.1</td>
<td>5.8</td>
<td>464</td>
<td>10382</td>
<td>56212</td>
<td>950</td>
</tr>
<tr>
<td>2006</td>
<td>6230</td>
<td>1394</td>
<td>19197</td>
<td>0</td>
<td>199.5</td>
<td>6.3</td>
<td>471</td>
<td>11620</td>
<td>58288</td>
<td>1070</td>
</tr>
<tr>
<td>2007</td>
<td>7771</td>
<td>1610</td>
<td>21593</td>
<td>0</td>
<td>434.2</td>
<td>5.1</td>
<td>493.6</td>
<td>13603</td>
<td>60633</td>
<td>1172</td>
</tr>
<tr>
<td>2008</td>
<td>8749</td>
<td>1712</td>
<td>23197</td>
<td>0</td>
<td>156720</td>
<td>21.4</td>
<td>522.7</td>
<td>15549</td>
<td>62403</td>
<td>1328</td>
</tr>
<tr>
<td>2009</td>
<td>10184</td>
<td>1902</td>
<td>24839</td>
<td>31.5</td>
<td>445.5</td>
<td>4.1</td>
<td>519</td>
<td>16901</td>
<td>64512</td>
<td>1380</td>
</tr>
<tr>
<td>2010</td>
<td>12580</td>
<td>2103</td>
<td>27894</td>
<td>0</td>
<td>3521.9</td>
<td>8</td>
<td>536.1</td>
<td>18779</td>
<td>66978</td>
<td>1447</td>
</tr>
<tr>
<td>2011</td>
<td>19305</td>
<td>2641</td>
<td>30984</td>
<td>0</td>
<td>78659.3</td>
<td>40.5</td>
<td>565</td>
<td>21427</td>
<td>69070</td>
<td>1501</td>
</tr>
<tr>
<td>2012</td>
<td>22706</td>
<td>2957</td>
<td>33383</td>
<td>0</td>
<td>968.7</td>
<td>4.9</td>
<td>579.7</td>
<td>24127</td>
<td>71182</td>
<td>1539</td>
</tr>
<tr>
<td>2013</td>
<td>26276</td>
<td>3262</td>
<td>27572</td>
<td>12.1</td>
<td>10738.5</td>
<td>6.3</td>
<td>594.8</td>
<td>26467</td>
<td>73111</td>
<td>1578</td>
</tr>
<tr>
<td>2014</td>
<td>30312</td>
<td>36111</td>
<td>28647</td>
<td>0</td>
<td>3363.7</td>
<td>3.4</td>
<td>606.7</td>
<td>28844</td>
<td>74916</td>
<td>1620</td>
</tr>
<tr>
<td>2015</td>
<td>34195</td>
<td>4000</td>
<td>30059</td>
<td>0</td>
<td>1687.1</td>
<td>2.5</td>
<td>615.2</td>
<td>31195</td>
<td>77116</td>
<td>1660</td>
</tr>
<tr>
<td>2016</td>
<td>39390</td>
<td>4440</td>
<td>31258</td>
<td>0</td>
<td>599.5</td>
<td>4.5</td>
<td>627.5</td>
<td>33615</td>
<td>79298</td>
<td>1702</td>
</tr>
<tr>
<td>2017</td>
<td>45661</td>
<td>5001</td>
<td>32813</td>
<td>0</td>
<td>1246.4</td>
<td>2.5</td>
<td>637.5</td>
<td>36396</td>
<td>81347</td>
<td>1747</td>
</tr>
<tr>
<td>2018</td>
<td>51278</td>
<td>5539</td>
<td>34218</td>
<td>0</td>
<td>232.1</td>
<td>2</td>
<td>650.9</td>
<td>39251</td>
<td>83137</td>
<td>1784</td>
</tr>
<tr>
<td>2019</td>
<td>57251</td>
<td>6010</td>
<td>35349</td>
<td>0</td>
<td>420.2</td>
<td>2.3</td>
<td>669.8</td>
<td>42359</td>
<td>84843</td>
<td>1836</td>
</tr>
</tbody>
</table>
the function \textit{summary()} . We use it to test the fitting effect of \textit{model 1}. As can be seen from Figure 1 (including code), many explanatory variables have not passed the T test, indicating that there may be multiple colinear problems in the interpretation variables.

\begin{verbatim}
> modell<-lm(y~.,data = mydata)
> summary(modell)

Call:
  lm(formula = y ~ . , data = mydata)

Residuals:
     Min      1Q  Median      3Q     Max
-393.00 -140.86  -12.63  151.04  404.30

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.864e+04  1.502e+04  -1.241 0.24284
 x1            1.026e+01  1.553e+00    6.605 6.04e-05 ***
 x2            -1.424e-02  5.778e-02   -0.246 0.81035
 x3            3.734e+00  1.177e+01    3.137 0.75750
 x4            1.387e-03  4.321e-03   -0.321 0.75486
 x5            1.631e+01  1.894e+01    0.861 0.40945
 x6            9.080e+00  2.157e+01    4.211 0.68266
 x7            9.377e+02  4.522e+02   2.027 0.83989
 x8            3.154e+01  2.393e+01    1.318 0.21684
 x9            -1.226e+01  3.177e+00   -3.860 0.00316 **

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

Residual standard error: 331.6 on 10 degrees of freedom
Multiple R-squared:  0.9998,  Adjusted R-squared:  0.9996
F-statistic: 5970 on 9 and 10 DF,  p-value: < 2.2e-16
\end{verbatim}

Fig.1 model1 Parameter diagram

3.2. Multiple colinear

The function \textit{cov()} in the \textit{car} package can be used to test the correlation coefficients of the interpretable variable and the interpreted variable, which is coded as: \textit{cov( model 1)}. There is a clear linear relationship between the interpreted variable and interpretable variable by testing. Multiple collinearity can be detected by Variance Inflation Factor. The square root of \textit{VIF} indicates the extent to which the confidence interval of variable regression parameters can expand into a predictive variable unrelated to the model\textsuperscript{[3]}. The \textit{vif()} function in the \textit{car} package provides the \textit{VIF} value. Under the general principle, \textit{vif} \textgreater{} 2 indicates the existence of multiple colinear problems. As can be seen from Figure 2 (including running code) excepting for \textit{x}_{3}, all other explanatory variables have multiple colinear problems.

\begin{verbatim}
> library(car)
> vif(modell)
 x1         x2         x3         x4         x5         x6         x7         x8         x9
1179.548517 35.574563 1.902014 4.657439 4.751367 530.7363538 4802.300349 1482.225837
386.795873
> sort(vif(modell))[>2]
 x1 x2 x3 x4 x5 x6 x7 x8 x9
 TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE

Fig.2 Multiple colinear test graph
\end{verbatim}

3.3. All subsets regression

All subsets regression means that all possible models will be tested. We can choose to show all possible results or we can also show good models of n different subset sizes (one, two or more
predictive variables). Because there are serious multi-collinear problems in model1. The variable on model 1 will be deleted by the total subset regression method to further optimise the model. The all subsets regression can be realized by the `regsubsets()` function in the `leaps` package. Selecting better model by criteria such as R², adjustment R² or Mallows Cp statistics[4]. Using all subsets regression method with code

```r
library(leaps)
leaps <- regsubsets(y~., data = mydata)
subsets(leaps, statistic = cp, main = regression)
abline(1,1,lty = 2, col = "red")
```

to model 1. In Figure 3, you can see seven best models based on Mallows Cp statistics for different subset sizes. The better models is closer to the straight line whose intercept term and slope are both 1. Selecting the new model2 from Figure 3.

Using methods in 3.1 and 3.2 test model 2. x₈ did not pass the T test, because residents at different levels are affected by different levels of education investment [5]. So x₇ (urban residents disposable income), x₈ (urban population) and x₉ (social education level) may interact.

![Fig.3 All subset regression diagram](image)

3.4. Interaction terms in multivariate linear regression

Sometimes it may be better to use the interaction of variables than variables to play alone in the model [6]. Because the explanation variable x₈ did not pass the T test, removed it from model 2, and then added the Interaction terms of x₇, x₈ and x₉ in model 2. Optimize model 2 and then name the new model model3. The code for building a new model is

```r
model 3 <- (y~x1 + x5 + x7 + x9 + x7:x8:x9, data = sdmymdata)
```

sdymdata in the code is the standardized data set of mydata. The standardized code is

```r
sdymdata <- scale(mydata).
```

In Figure 4, all the explanatory variables in the model have passed the T test, and the regression model has also passed the F test. It shows that the overall coefficients in the new regression model 3 are relatively significant.
3.5. Test the statistical assumptions of the regression model

A large number of methods for testing statistical assumptions in regression analysis are provided in R-based installation. The most common method is to use the `plot()` function to generate four graphics to evaluate the modelling of the objects returned by the `lm()` function. The `par()` function can specify the arrangement of the generated graphics. Use the following code to test the regression model’s final statistical assumptions. The code is

```r
par(mfrow = c(2,2))
plot(model3)
```

You can see that the points in the statistical hypothesis test graph (Residuals vs Fitted) are evenly distributed at both ends of the red line, which shows that the linear hypothesis is satisfied. Most of the points in the statistical hypothesis test graph (Normal Q-Q) fall on the dotted line, indicating that the normal hypothesis is satisfied. The random distribution of points around the horizontal line in the statistical hypothesis test graph (Scale-Location) shows that the variance invariant hypothesis is satisfied. There are outliers in the statistical hypothesis test graph (Residuals vs Leverage) that may have an impact on the model and the final model basically satisfies the statistical hypothesis. The final standardized model is

\[
\hat{y} = 0.756x_1 + 0.012x_5 + 0.464x_7 - 0.280x_9 + 0.018x_7 \times x_9 \times x_9 - 0.003.
\]
4. Conclusions and Recommendations

We see that the contribution of the number of domestic tourists to the regression equation is the greatest. First, due to the large population base in China, and second, as described in China’s basic social contradictions, people’s yearning for a good material and cultural life is also gradually increasing. So the contribution rate of domestic tourists to the total cost of domestic tourism is very large. It’s normal. In order to achieve the greatest satisfaction of consumers' psychological expectations, we can promote the improvement of the leisure industry through the progress of tourism, expand the connotation of tourism elements. We should grasp the psychological needs of the people and strive to develop the tourism culture recognized by the domestic people.

With the gradual strength of the country and the improvement of social education, people’s disposable income has also been raised. After satisfying their daily food, clothing, housing and transportation, people will yearn for tourism, which shows the great potential of tourism. We should therefore pay more attention to the role of tourism in the national economy.

The R language is open source software with powerful packets and top-level mapping functions. We can completely build multiple models for optimization and comparison through the functions in other packages.

Acknowledgements

This paper was financially supported by fundation: Key Laboratory of Bridge Nondestructive Testing and Engineering Calculation(2019QYY03).

References


