An Analysis of Domestic Tourism Consumption Based on R Software

Hongyu Ding¹

Shanghai University, 200444, China

Abstract. R is a software system which can be used for data processing, calculation and mapping. The syntax of this language is superficially similar to C, but semantically it is functional programming language. It is widely used in statistical analysis. So this paper used it to analyze China domestic tourism consumption data during 1999-2015, and analyzed the main factors affecting the consumption level of domestic tourism in China from residents' disposable income, GDP, per capita consumption of tourists, tourists, the mileage of railways and the number of travel agencies. Finally it established and solved the multiple linear regression models, taking domestic tourism consumption as dependent variable and taking GDP, and per capita consumption of tourists, the number of tourists and the mileage of railways as dependent variables. The results show that there is significant positive correlation between domestic tourism consumption, GDP, per capita consumption of tourists, the number of tourists and the number of travel agencies.

1 Introduction

Liu Shenzhen's "Analysis of Domestic Tourism Consumption Based on Multiple Linear Regression Model" published on The Journal of Chongqing University of Technology (Natural Sciences) in June 2016, selecting data on domestic tourism consumption in China from 2003 to 2012, used a multiple linear regression model with domestic tourism consumption as the dependent variable and per capita disposable income, tourist per capita consumption, and number of travel agencies as independent variables to analyze factors affecting China's domestic tourism consumption. The paper concludes that: Domestic tourism consumption is positively correlated with disposable income of residents and per capita consumption of tourists, and negatively related to the number of travel agencies [1]. However, the results are inconsistent with the conclusions. Residents' disposable income and the number of travel agencies are not significant in this regression equation. Meanwhile, the article only makes a simple regression analysis of the variables without considering the problems of multicollinearity, time series between variables and etc. This paper holds the idea of verifying results and expands the selection time of variables, sorts out relevant literature from previous scholars, and increases the number of independent variables based on the preservation of the author's original variables, thereby performing a regression analysis of the equations.

2 Research review

¹ Corresponding author: suimosuishendhy@163.com

In recent years, with the continuous improvement of people's living standards, people have not been satisfied with the basic consumption, paid more attention to consumption at the spiritual level, which has led to an increase in domestic tourism consumption. Analyzing the major factors affecting tourism consumption has a very important significance for the development of tourism. It can not only provide a reference for tourism management departments to formulate tourism development plans, but also can be used to make short- and medium-term predictions of domestic tourism consumption and promote the healthy development of the tourism industry [2-6]. In studies related to domestic tourism consumption, Li Yunpeng (2005) found that China domestic tourism consumption of urban residents was mainly affected by current income and prices [7]. Sun Gennian and Xue Jia (2009) pointed out that the per capita disposable income was the basic factors affecting and restricting domestic tourism consumption. The view that income has a decisive effect on urban residents' domestic tourism consumption, have been extensively empirically tested [8]. In addition, Wu Chunlai and Gu Huimin (2003) believed that the changes in residents' income distribution structure had transformed the tourism market from a mass-visited tourism industry to a multi-faceted, heterogeneous structure [3]; Teng Li et al. (2004) found that the urban residents' tourism consumption differed in the types of travel and per capita tourism spending, and this difference was related to per capita income [9]. From the above, we can find that the per capita income of residents has an important role in residents' tourism consumption. This article selected it as one of the

explanatory variables. At the same time, with the continuous improvement of living standards, people are pursuing more spiritual and cultural consumption. Because of convenient transportation, more people are beginning to go abroad to experience different scenery and different customs. Weng Gangmin et al. (2007) found that there was the most closely related relationship between urban residents' travel rate and the factors of living standards, and per capita consumption of tourists was an important factor affecting tourism consumption income ^[10]. Wang Yu and Qin Yuanhao (2011) discussed the relationship between per capita tourism expenditure, GDP, free disposable income of urban residents, average labor compensation of urban residents, total amount of RMB deposits, length of railway, and free time ^[11]. They put these factors into the model and gradually returned. Finally, it was concluded that discretionary income, length of railway, and free time had a significant impact on tourism consumption. In domestic research, most scholars have used empirical analysis methods to select major factors for quantitative analysis in factors such as people's savings, per capita income, per capita disposable time, railway length, number of travel agencies, and national policies ^[12].

Based on the above literature review, we can draw the main factors and models that affecting tourism consumption. This article established and solved a multivariate linear regression model with the domestic tourism consumption (Y) as the cause, per capita disposable income of residents (X1), per capita consumption (X2) and number of travel agencies (X3), domestic tourists (X4), gross domestic product GDP (X5), and total railway kilometers (X6) as independent variables in the multivariate linear regression model.

3 Data sources

According to the availability, authority, and unified statistical principles of data, the data in this paper are all selected from the "China Statistical Yearbook" from 1999 to 2015. Through the review of previous literature, this paper selects the per capita disposable income of residents (X1), per capita consumption (X2) and number of travel agencies (X3), domestic tourists (X4), gross domestic product (GDP) GDP (X5), and total railway kilometers (X6) as independent variables in the multivariate linear regression model.

Table 1. Data for each variable in 1999~2015.

Year	Domestic tourism consumption / Billion Yuan (Y)	Per capita disposable income of residents / Yuan (X1)	Per capita consumption / Yuan (X2)	Number of travel agencies (X3)	Domestic tourists (ten thousand people) (X4)	GDP / Billion Yuan (X5)	Total railway kilometres / Ten thousand kilometres (X6)
1999	2831.92	5854	394	7326	71900	90564.4	6.74
2000	3175.54	6280	426.6	8993	74400	100280.1	6.87
2001	3522.37	6859.6	449.5	10532	78400	110863.1	7.01
2002	3878.36	7702.8	441.8	11552	87800	121717.4	7.19
2003	3442.27	8472.2	395.7	9751	87000	137422	7.3
2004	4710.71	9421.6	427.5	10888	110200	161840.2	7.44
2005	5285.86	10493	436.1	11828	121200	187318.9	7.54
2006	6229.7	11759.5	446.9	12751	139400	219438.5	7.71
2007	7770.6	13785.8	482.6	18943	161000	270232.3	7.8
2008	8749.3	15780.9	511	20110	171200	319515.5	7.97
2009	10183.69	17174.7	535.4	20399	190200	349081.4	8.55
2010	12579.77	19109.4	598.2	22784	210300	413030.3	9.12
2011	19305.39	21809.8	731	23690	264100	489300.6	9.32
2012	22706.22	24564.7	767.9	24944	295700	540367.4	9.76
2013	26276.12	26955	805.5	26054	326200	595244.4	10.31
2014	30311.86	28844	839.7	26650	361100	643974	11.18
2015	34195.05	31195	857	27621	400000	689052.1	12.1

Research model and theoretical hypothesis

In the selection of variables, the influence relationship between different variables has been combed. To make this model simple and clear, we use multiple linear regression model, combining the economic theory and the running law of tourism itself, the following equations are constructed:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \beta_5 * X_5 + \beta_6 * X_6 + \mu$$

Among them, Y is dependent variable, $X_1 \sim X_6$ is independent variable, $\beta_0 \sim \beta_6$ is correlation coefficient, and μ is random disturbance term.

4 Analysis process

4.1 Manufacturing time series

Figure 1. Time series of independent variables.

4.2 Stability analysis

The empirical judgment method shows that the ADF test value (T value) > 0.05 (critical value), does not pass the test. That means, the unit root is not stable, otherwise it is stable. According to the above results, the t value of the original data is more than 0.05, so the time series is not stable. After the first-order difference, only the value of the variable X3 and X4 is a stationary sequence, but because the other four variables are not stable, this paper carries out the second-order difference for all data and the second-order difference results are less than 0.05. The two order difference sequence is stationary.

Table 2. Stability analysis results of independent variables.

	ADF test	First-order difference	Second-order difference
X1	1	0.8081	0.00448

Table 3. Results of multicollinearity analysis among variables.

	Y	X1	X2	X3	X4	X5	X6
Y	1.0000000	0.9763805	0.9867182	0.9058634	0.9899242	0.9769787	0.9869175
X1	0.9763805	1.0000000	0.9734340	0.9683941	0.9959208	0.9992293	0.9817591
X2	0.9867182	0.9734340	1.0000000	0.9276144	0.9804547	0.9784178	0.9651946
X3	0.9058634	0.9683941	0.9276144	1.0000000	0.9484804	0.9695357	0.9187844
X4	0.9899242	0.9959208	0.9804547	0.9484804	1.0000000	0.9952837	0.9889577
X5	0.9769787	0.9992293	0.9784178	0.9695357	0.9952837	1.0000000	0.9794118
X6	0.9869175	0.9817591	0.9651946	0.9187844	0.9889577	0.9794118	1.0000000

4.5 Stepwise regression, screening main risk factors, and establishing a regression model

According to the value of AIC, X1 is eliminated first, and then the equation is constructed: $Y \sim X2 + X3 + X4 + X5 + X6$

Table 4. Regression results after eliminating variable X1.

R2	Adjusted R2	F value	p value	DW
0.998	0.997	1233	9.94E-15	1.5971

X2	0.9937	0.0569	0.0001
X3	0.8936	0.0126	2.13E-05
X4	1	0.4823	1.04E-05
X5	1	0.5943	0.00039
X6	1	0.832	0.001576

4.3 Co-integration test

The six variables in this paper are all unstable, but both of them are second order stationary. So the co-integration test is used to determine whether the regression equation composed of these six variables is stable or not. If it is stable, it shows that there is a stable relationship between these variables, and this research is significant. The first step is to regress the equation to generate residuals. In the second step, unit root test is performed on the residual sequence. The test result showed $t=0.001452 < 0.05$, so the residual sequence was stable after testing.

4.4 Multicollinearity analysis

In general, because of the limitation of economic data, general correlation among the explanatory variables existed in the design matrix. According to the criterion of correlation coefficient, when the absolute value of correlation coefficient is > 0.8 , it shows that there is significant linear correlation between the two variables. When the absolute value is less than 0.3, it is called low correlation. Others were moderately related. According to the calculation results, it can be found that the correlation coefficient between the factors in the equation is more than 0.9, showing a highly correlated trend. In order to guarantee the accuracy and the goodness of the model, the factors should be eliminated and then regressed step by step.

Table 5. Significance analysis results for each variable after eliminating variable X1.

	t value	Pr(> t)
Intercept term	-4.767	0.000583
X2	6.685	3.44E-05
X3	-2.065	0.063317
X4	4.356	0.001143
X5	-1.856	0.090355
X6	2.197	0.050364

According to the calculation results, the P values of X3(0.063317) and X5 (0.090355) are significant under the condition of less than 0.1, and not significant under the condition of less than 0.05. Usually the closer the DW value is to 2; the more unlikely the equation is autocorrelation. The DW value here is 1.5971,

the equation may have autocorrelation. Other results show that the regression results of the equation are better. The selection of

independent variables in this paper is relatively large, so this paper chooses to continue to eliminate variables for regression. Then the equation is $Y \sim X2 + X4 + X5 + X6$.

Table 6. Regression results after eliminating variable X3.

R2	Adjusted R2	F value	p value	DW
0.9975	0.9967	1211	1.59E-15	1.8995

Table 7. Significance analysis results of all variables after eliminating variable X3.

	t value	Pr(> t)
Intercept term	-6.212	4.51E-05
X2	6.835	1.81E-05
X4	4.901	0.000366
X5	-5.572	0.000121
X6	2.581	0.024059

According to the result of the calculation, the p value and t value of each variable are significant; R2 is close to 1, which shows good goodness of fit and F value. According to the results of look-up table, DW value is close to 2 in the corresponding interval. This paper holds that there is no self-correlation in the equation. According to the corresponding data, this paper does not eliminate the independent variable. The final expression of the equation is:

$$Y = -25930 + 31.33 * X_2 + 0.1072 * X_4 + -0.0446 * X_5 + 1749 * X_6$$

4.6 Heteroscedasticity test

In order to ensure that the estimators of regression parameters have good statistical properties, an important assumption of the classical linear regression model is that the random error term in the general regression function satisfies the same variance, which means they all have the same variance. If this assumption is not satisfied, which means the random error term has different variances. There is heteroscedasticity in the linear regression model. According to White Test, P value = 0.5416 > 0.05. According to white test, there is no heteroscedasticity in the model.

4.7 Sequence correlation test

If there is a correlation between the expected values of

the random error term, it is said that there is autocorrelation or sequence correlation between the random error terms, which is commonly seen in time series.

Table 8. Results of serial correlation test.

Variable	X2	X4	X5	X6
VIF	26.663	246.758	121.104	52.141

The results show that there are heavy multicollinearity between X2 and X6, and serious multicollinearity between X4 and X5. Because of the higher multicollinearity among the variables of the equation, the corresponding variables are removed, there is no multicollinearity between X2 and X3, but the results of X3 show that p = 0.565 and T = -0.589, the variable X3 is not significant, so the equation is not illustrative. Considering the significance of the variables and the better regression effect, there is only multicollinearity among the variables, and there is no variable X3 in the equation, so the article chooses no longer to eliminate the variables, and adopts the method of principal component analysis to further optimize the regression equation.

4.8 Principal component analysis

By orthogonal transformation, a set of variables that may be correlated are transformed into a set of linearly independent variables, which are called principal components.

According to the results of R studio calculation, the final number of principal components is 1, and the coefficients of each principal component are 0.25.A new equation, $F = 0.25 * (X2 + X4 + X5 + X6)$, is established based on the coefficients of the principal component. The results obtained from the regression of the new equation are as follow:

$$Y = -4490 + 0.03275 * (X2 + X4 + X5 + X6).$$

Table 9. Results of principal component analysis.

	PC1	h2	u2
X2	0.99	0.98	0.024
X4	1	1	0.0036
X5	1	0.99	0.0094
X6	0.99	0.98	0.0192

Table 10. Regression coefficient of Y and F.

	Intercept term	F
Correlation coefficient	-4.49E+03	1.31E-01
t value	-4.744	20.415
Pr(> t)	0.000261	2.34E-12

Table 11. Regression results of Y and F.

R2	Adjusted R2	F value	p value
0.9653	0.9629	416.8	2.34E-12

5 Conclusions

(1) With the aid of R statistical analysis software, the regression equation is established with the per capita consumption of tourists, domestic tourists, GDP of gross national product and the total number of railway kilometres. The results show that China's domestic tourism consumption is positively related to the per capita consumption of tourists, domestic tourists, gross national product (GDP), and the total number of railway kilometres. There is no significant relationship with the number of travel agencies, which is contradictory to the conclusions of Liu Zhenzhong's article.

(2) According to the established model and the result, it is not hard to find that gross national product (GNP) GDP is highly related to the disposable income of the residents. Although the article excludes the variable of residents' disposable income, GDP can partly reflect the size of the residents' disposable income. In recent years, with the rapid growth of China's economy, the disposable income of the residents and GDP are increasing, which lead to the improvement of transportation, communication and accommodation. It is more convenient for people to travel. Tourism is no longer the exclusive property of rich people.

(3) There is no significant correlation between number of travel agencies and domestic tourism consumption. With transportation convenience, the rapid development of OTA and UGC platforms and negative news of the travel agencies, more and more tourists prefer to travel by self-help rather than package tour. Tourists who join a package tour mainly go abroad, so the numbers of travel agencies have no significant impact on domestic tourism consumption. From the present development situation analysis, the travel agencies need to break through the models of traditional low price group, the zero pay groups and the shopping group, and seek the cooperation between the online OTA and the offline entities so as to seek a developing way.

(4) According to the analysis of previous surveys, tourists spend about 30% on transportation, which is enough to explain the importance of transportation and the impact on domestic tourism consumption. To develop domestic tourism, we should first develop transportation industry; especially increase railway mileage and road mileage. Many scenic spots are located in the poor areas, the lack of transportation facilities and other infrastructure restricted the development of tourism and the economy, so speeding up the construction of railways in these underdeveloped areas is of great importance. At the same time, the investment of highway and other transportation facilities will be increased to form a more unified national mechanism, which is conducive to the development of domestic tourism.

(5) Relevant policies and institutions need to be improved, so as to improve the overall service level of the tourism industry. With the improvement of the people's living standard, domestic tourism consumption

has been constantly increasing; tourism is more and more important. However, the overcharge problems of tour guide and the sky-high cost of scenic spots made tourists take a negative attitude towards some destinations. Tourism, a green industry, can not only promote economic, but also bring jobs. It is a way to develop and prosper in resource-poor areas. Therefore, the country needs more supervision system and perfect corresponding laws and regulations to promote sustainable and healthy development of the tourism industry.

References

1. Liu Zhenzhong. *Analysis of Domestic Tourism Consumption Based on Multiple Linear Regression Model* [J]. Journal of Chongqing University of Technology, 2016, 30 (6): 167-170.
2. Li Yiwei, Xia Lingen. *An Analysis of Tourism Consuming Structure of China's Urban Residents* [J]. Tourism Science, 2004, 18 (2): 30-32.
3. Gu Huimin, Wu Chunlai. *The Impact of the Evolution of Income Distribution Structure in China on the Domestic Tourism Consumption* [J]. Tourism Tribune, 2003, 18 (2): 19-23.
4. Li Yinlan, Fan Hong. *Discussion on Domestic Tourism Consumption Model* [J]. Journal of Chongqing Institute of Commerce, 2002 (3): 40-42.
5. Huang Xiujuan. *Positive Analysis of the Relationship between China's Domestic Touring Consumption and Residents' Income* [J]. Journal of Hebei university of Economics and Trade, 2004, 25 (5): 31.
6. Li Bingzhou, Yang Jian, Chen Xu. *A Study on Tourism Consumption of China's Resident* [J]. Soft Science, 2004, 18 (2): 58.
7. Li Yunpeng. *Research on Tourism Consumption of Domestic Urban Residents Based on Econometrics Model* [J]. Technoeconomics and Management Research, 2005 (6): 54-55.
8. Sun Gennian, Xue Jia. *Income-driven Model Research on China's Urban and Rural Resident Domestic Tourism* [J]. Commercial Research, 2009 (5): 13-16.
9. Teng Li, Wang Zheng, Cai Di. *An Analysis of the Difference of Tourism Demand among China's Urban Residents* [J]. Tourism Tribune, 2004, 19 (4): 9-13.
10. Xu Xiaona, Weng Gangmin. *Grey Relational Analysis on Influence Factors of Difference in Tourism Demand among Urban Residents* [J]. Resource Development and Market, 2007, 23 (1): 24.
11. Wang Xi, Qin Yuanhao. *Empirical Analysis on the Main Factors Influencing the Tourism Consumption of Urban Residents in China* [J]. China Market, 2011 (22): 131-133.
12. Ma Jiahui, Sun Peng. *An empirical Analysis of the Urban Residents in China Tourism Consumption Influencing Factors* [J]. Science & Technology and Economy, 2016, 29 (5): 60-64.