Advances in Social Sciences 社会科学前沿, 2017, 6(2), 151-160
Published Online February 2017 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2017.62020

Hans汉斯

Regression Analysis and Prediction of Highway Passenger Volume

Dandan Shen

Shanghai Maritime University, Shanghai
Email: 1174618783@qq.com

Received: Feb. 1st, 2017; accepted: Feb. 19th, 2017; published: Feb. 22nd, 2017

Abstract

With the development of economy and the increase of living standard, China has paid more attention on the infrastructure of highway transportation. This paper applies the related data of the passenger capacity of the highway transportation from 1981 to 2015 to analyze the factors influenced the passenger capacity of the highway transportation. It takes the population, GDP, agricultural GDP and the civil car ownership as the independent variables, and the passenger capacity of the highway transportation as the dependent variable to establish the multivariate regression model with MATLAB. At the same time, the rationality of the regress model is also analyzed in this essay. We have fitted various factors of the multivariate regress model in consideration of the complexity, verified and improved the multivariate regress model with stepwise regression in order to enhance the scientificity and accuracy of the model. In the final step, the passenger capacity of the highway transportation from 2011 to 2015 has been calculated by the regress model of univariate and multivariate, and a comparison has been made between the calculated date and the actual data which aims to analyze the difference and errors.

Keywords

Passenger Capacity of the Highway Traffic, Regression Model, Stepwise Regression, The Civil Car Ownership

公路客运量的回归分析和研究预测

沈丹丹

上海海事大学，上海
Email: 1174618783@qq.com

收稿日期：2017年2月1日；录用日期：2017年2月19日；发布日期：2017年2月22日

文章引用: 沈丹丹. 公路客运量的回归分析和研究预测[J]. 社会科学前沿, 2017, 6(2): 151-160.
https://doi.org/10.12677/ass.2017.62020

摘　要

随着我国经济的发展、国民生活水平的日益提高，我国对公路交通基础建设也越加重视。本文利用我国1981~2015年客运量各项有关数据，对公路客运量影响因素进行分析，以总人口、国内生产总值、农业生产总值以及民用载客汽车拥有量为自变量，公路客运量作为因变量。运用MATLAB建立多元回归模型，并对回归模型的合理性进行了分析。其中多元回归模型综合考虑了影响客运量的众多因素，对这些因素进行模型拟合，并用逐步回归法对模型进行了检验和改进，大大提高了模型的科学性和准确性。最后运用所得的回归模型拟合2011~2015公路客运量，并与实际客运量数据进行对比，分析误差。

关键词

公路客运量，多元回归模型，逐步回归，民用载客汽车拥有量

1. 引言

公路运输是国民经济的基础性和服务性产业，是合理配置资源、提高经济运行质量和效率的重要基础，具有基础性和先导性的作用。客运量是衡量公路运输发展程度的重要指标，可用于反映社会经济发展现状和人民生活水平。目前，客运量预测方法已达 300 多种，归纳起来大体分为定性预测和定量预测两类。常用的定量预测方法有指数平滑法、回归分析法、弹性系数法、灰色系统法、组合法等；定性预测方法有运输市场调查法、德尔菲法和类推法等[1]。

本文主要采用回归分析法对公路客运量进行回归分析和预测。世界上任何事物的产生和发展都是由一定的原因引出一定的结果。当一个变量(因变量)同其它一些因素(自变量)之间存在着某种因果关系的时候，我们就可以按照一定的方式建立反映这些关系的数学模型，然后根据自变量在未来的变化来计算因变量的变化，这就是因果关系预测。建立因果关系预测常采用的方法就是回归分析法，该方法是利用过去的历史资料，从中分析找出事物发展的内在联系，确定事物的自变量和应变量，以及它们之间的相关关系，建立数学方程式，一般称其为回归方程[2]。

2. 变量的选取

公路客运量主要受到经济发展水平、经济结构、人口及其构成、居民收入与消费水平、旅游业发展状况、运输网络结构等因素的影响。本文主要选取总人口 x_1、国内生产总值 x_2、工农业总产值 x_3、民用载客汽车拥有量 x_4 作为自变量，公路客运量 y 作为因变量建立模型。

3. 数据

1981~2015 年我国公路客运量、总人口、国内生产总值、工农业总产值、民用载客汽车拥有量数据见表 1。

4. 公路客运量与各自变量的多元回归模型

社会经济现象是复杂的，通常一种社会经济现象与许多种现象相联系。一种社会经济现象与多种现象相联系的最简单形式，是一个被解释变量与多个解释变量的线性关系[3] [4]。

Table 1. Data of highway passenger volume, total population and gross domestic product, gross output value of industry and agriculture, civil passenger car ownership in 1981-2015
表 1. 1981~2015 年我国公路客运量、总人口、国内生产总值、工农业总产值、民用载客汽车拥有量数据

	公路客运量 (万人) y	总人口 (万人) x_1	国内生产总值 (亿元) x_2	工农业总产值 (万元) x_3	民用载客汽车拥有量(万辆) x_4
1981 年	261559	100072	4891.56	1635.87	40.57
1982 年	300610	101654	5323.35	1865.3	44.18
1983 年	336965	103008	5962.65	2074.47	47.78
1984 年	390336	104357	7208.05	2380.15	56.28
1985 年	476486	105851	9016.04	2506.39	79.45
1986 年	544259	107507	10275.18	2771.75	96.61
1987 年	593682	109300	12058.62	3160.49	111.46
1988 年	650473	111026	15042.82	3666.89	130.38
1989 年	644508	112704	16992.32	4100.58	146.43
1990 年	648085	114333	18667.82	4954.26	162.19
1991 年	682681	115823	21781.5	5146.43	185.24
1992 年	731774	117171	26923.48	5588.02	226.16
1993 年	860719	118517	35333.92	6605.14	285.98
1994 年	953940	119850	48197.86	9169.22	349.74
1995 年	1040810	121121	60793.73	11884.63	417.9
1996 年	1122110	122389	71176.59	13539.75	488.02
1997 年	1204583	123626	78973.03	13852.54	580.56
1998 年	1257332	124761	84402.28	14241.88	654.83
1999 年	1269004	125786	89677.05	14106.22	740.23
2000 年	1347392	126743	99214.55	13873.59	853.73
2001 年	1402798	127627	109655.17	14462.79	993.96
2002 年	1475257	128453	120332.69	14931.54	1202.37
2003 年	1464335	129227	135822.76	14870.11	1478.81
2004 年	1624526	129988	159878.34	18138.36	1735.91
2005 年	1697381	130756	184937.37	19613.37	2132.46
2006 年	1860487	131448	216314.43	21522.28	2619.57
2007 年	2050680	132129	265810.31	24658.17	3195.99
2008 年	2682114	132802	314045.43	28044.15	3838.92
2009 年	2779081	133450	340902.81	30777.48	4845.09
2010 年	3052738	134091	401512.8	36941.11	6124.13
2011 年	3286220	134735	473104.05	41988.64	7478.37
2012 年	3557010	135404	519470.1	46940.46	8943.01
2013 年	3853463	136072	590422.4	51497.37	10561.78
2014 年	3908198	136782	643974.	60165.7	14598.11
2015 年	3619097	137462	685505.8	62918.7	16284.45

注：本表数据来自于《中国统计年鉴》，Wind 资讯

基本原理：多元线性回归原理[5]。

设 y 是一个可观测的随机变量，它受到 $m(m>0)$ 个随机变量因素 $x_1,x_2,\cdots,x_m$ 和随机误差 ε 的影响。若 y 与 $x_1,x_2,\cdots,x_m$ 有如下线性关系：

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m+\varepsilon \tag{1}$$

该模型即为多元线性回归模型，其中 $\beta_0,\beta_1,\beta_2,\cdots,\beta_m$ 是固定的未知参数，称为回归系数； ε 是均值为 0、方差为 $\sigma^2(\sigma>0)$ 的随机变量； y 称为被解释变量； $x_1,x_2,\cdots,x_m$ 被称为解释变量。

对于总体 $(x_1,x_2,\cdots,x_m;y)$ 的 n 组观测值 $(x_{i1},x_{i2},\cdots,x_{im};y_i)(i=1,2,\cdots,n;n>p)$ 应满足上述线性关系，即

$$\begin{cases}y_1=\beta_0+\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_mx_{1m}+\varepsilon_1\\ y_2=\beta_0+\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_mx_{2m}+\varepsilon_2\\ \cdots\\ y_n=\beta_0+\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_mx_{nm}+\varepsilon_n\end{cases} \tag{2}$$

其中 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 相互独立，且设 $\varepsilon_i\sim N(0,\sigma^2)(i=1,2,\cdots,n)$，记

$$y=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix},\ x=\begin{pmatrix}1 & x_{11} & x_{12} & \cdots x_{1m}\\ 1 & x_{21} & x_{22} & \cdots x_{2m}\\ \vdots & \vdots & \vdots & \cdots \vdots\\ 1 & x_{n1} & x_{n2} & \cdots x_{nm}\end{pmatrix},\ \beta=\begin{pmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_n\end{pmatrix},\ \varepsilon=\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}$$

则可用矩阵形式表示为[6]：

$$y=x\beta+\varepsilon$$

其中 y 称为观测向量； x 称为设计矩阵； β 称为待估计向量； ε 是不可观测的 n 维随机向量，它的分量相互独立，假定 $\varepsilon_i\sim N(0,\sigma^2I_n)$。

4.1. 多元线性回归模型

为了确定公路客运量与总人口、国内生产总值、工业生产总值、民用载客汽车拥有量之间的关系，首先建立四元线性回归模型[7]。

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\varepsilon \tag{3}$$

在 MATLAB 中对公路客运量 y 和总人口 x_1、国内生产总值 x_2、工业生产总值 x_3、民用载客汽车拥有量 x_4 进行拟合可得回归系数、系数置信区间与统计量见表 2。

Table 2. Coefficients, confidence intervals and statistics of regression models
表 2. 回归模型的系数、系数置信区间与统计量

回归系数	回归系数估计值	回归系数置信区间
β_0	576102.7200	[−2530217.7516　3682423.1915]
β_1	−1.9164	[−32.0219　28.1891]
β_2	15.1025	[8.8806　21.3224]
β_3	2.4743	[−74.3436　79.2922]
β_4	606.9620	[−916.5908　−297.3332]
$R^2=0.9336$，$F=98.3696$，$p<0.0001$，$s^2=6.0898\times10^{10}$		

因此回归模型为

$$\hat{y} = 576102.72 - 1.9164x_1 + 15.1025x_2 + 2.4743x_3 + 606.9620x_4 \quad (4)$$

回归模型中的各系数经济学意义解释：$\beta_1 = -1.9164$ 表示在其他条件不变的情况下，总人口每增加 1 万人公路客运总量会减少 1.9164 万人，这与事实不符合说明模型不是最优模型，有待改进；$\beta_2 = 15.1025$ 表示在其他条件不变的情况下，国内生产总值每增加 1 亿元公路客运量增加 15.1025 万人；$\beta_3 = 2.4743$ 表示在其他条件不变的情况下，工农业总产值每增加 1 万元公路客运总量增加 2.4743 万人；$\beta_4 = 606.9620$ 表示在其他条件不变的情况下，民用载客汽车拥有量每增加 1 万辆公路客运总量增加 606.9620 万人。

回归模型(4)的可决系数 $R^2 = 0.9336$， $p = 0.0001 < 0.05$，因此建立的回归模型有意义。

但由于 β_1 和 β_3 的置信区间包含零点，所以所建立的回归模型不是最优模型，下面对模型进行改进。

4.2. 模型的进一步改进

下面对模型进行进一步改进。

得到图形如图 1 所示，发现有两个异常点，剔除异常点后，重新建模。

仍有异常点继续剔除，直到没有异常点为止。剔除过程如图 2~5。

删除异常点后，由残差图 5 可得此时没有异常点，改进回归模型系数、系数置信区间与统计量见表 3。

故改进后的多元回归模型为：

$$\hat{y} = -1984102.4262 + 22.5837x_1 + 7.3229x_2 - 3.5319x_3 - 250.4032x_4 \quad (5)$$

将表 2 与表 3 加以比较，可以发现，可决系数从 0.9336 提高到 0.9954，F 统计量从 98.3696 提高到 1137.7580，删除异常点后的模型每个参数的置信区间进一步缩小，由此可知改进后的模型显著性提高。但是 β_3、β_4 的置信区间仍包含零故模型不是最优模型，再对模型的做进一步改进。

4.3. 改进后的模型与参考文献结果的比较

利用参考文献中的逐步回归法对模型检验然后再与参考文献中的结果作比较。

逐步回归基本原理：在逐步回归中，每当向模型中加入一个变量后，就对原来模型中的变量在新模型下再进行一次向后剔除的检查，直至所有已经在模型中的变量都不能被剔除，而且所有在模型外的变量都不能被加入，过程就终止[8]。

逐步回归模型的基本形式为 $y = k_0 + k_1x_1 + k_2x_2 + \cdots + k_mx_m + \varepsilon$。其中，$x_1, x_2, \cdots x_m$ 为需求影响因子，y 为需求量，$k_0, k_1, \cdots, k_m$ 为回归参数，ε 为误差变量，表示除 $x_1, x_2, \cdots x_m$ 对 y 的线性影响之外，其他随机因素对的影响[9]。

由表 3 知 β_3、β_4 的置信区间包含 0，下面采用逐步回归从各个自变量中挑选变量，建立更优回归模型，逐步回归过程见图 6~8。

由图 8 最后得到回归方程(蓝色行是被保留的有效行，红色行表示被剔除的变量)：

$$y = 375078 + 14.9214x_2 - 590.095x_4 \quad (6)$$

回归方程中录用了原始变量 x_2 和 x_4。

图 8 中显示了模型参数分别为 $R^2 = 0.9335$，修正 $R_\alpha^2 = 0.92909$， $F = 210.637$，与显著性概率相关的 $p = 2.18885e-018 < 0.05$，残差均方 RMSE = 238490 (在逐步回归中，均方残差逐渐减小)。

综上所述，相关参考文献中总人口是唯一的有效变量，其它 3 个变量即国内生产总值、工农业总产值、客车保有量被剔除。而本文中的逐步回归剔除了农业生产总值和总人口拥有量，选取的变量是国内

Table 3. Coefficients, confidence intervals and statistics of improved regression model
表 3. 改进回归模型的系数、系数置信区间与统计量

回归系数	回归系数估计值	回归系数置信区间
β_0	−1984102.4262	[−2492522.9698 −1475681.8826]
β_1	22.5837	[17.6682 27.4991]
β_2	7.3229	[0.9590 13.6868]
β_3	−3.5319	[−30.3505 23.2866]
β_4	−250.4032	[−599.3087 98.5024]

$R^2 = 0.9954$, $F = 1137.7580$, $p < 0.0001$, $s^2 = 1.2130 \times 10^9$

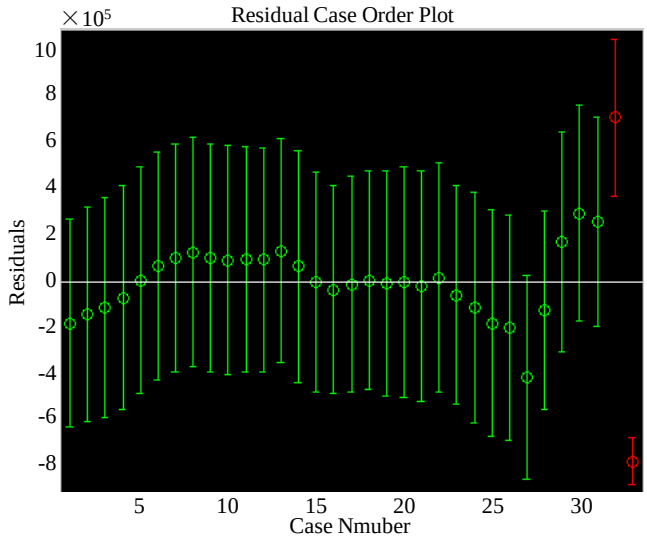

Figure 1. Schematic diagram of residual error
图 1. 残差示意图

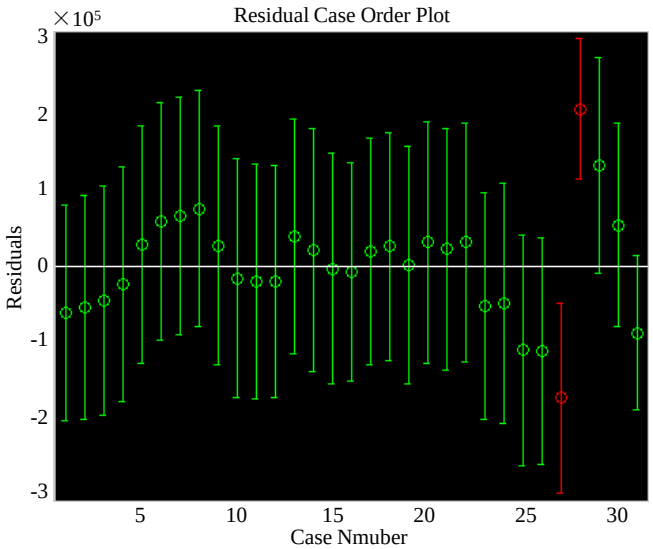

Figure 2. Sketch map of residual error I
图 2. 剔除异常点后的残差示意图 1

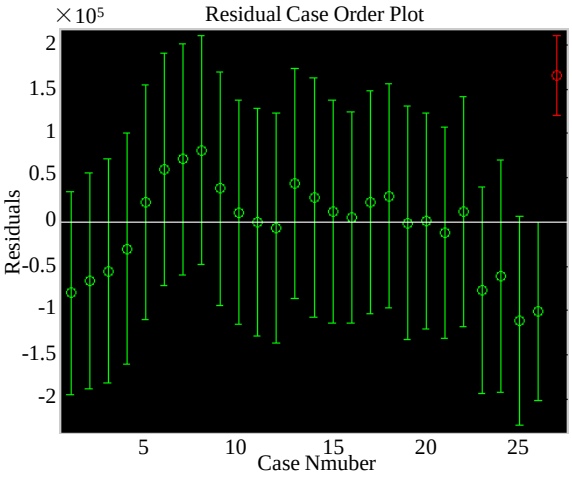

Figure 3. Sketch map of residual error II
图 3. 剔除异常点后的残差示意图 2

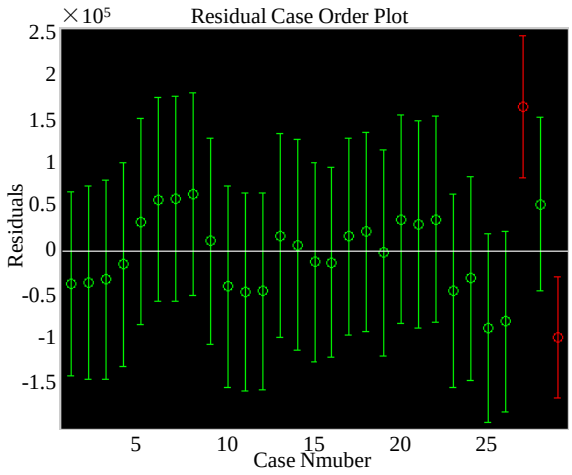

Figure 4. Sketch map of residual error III
图 4. 剔除异常点后的残差示意图 3

Figure 5. Sketch map of residual error IV
图 5. 剔除异常点后的残差示意图 4

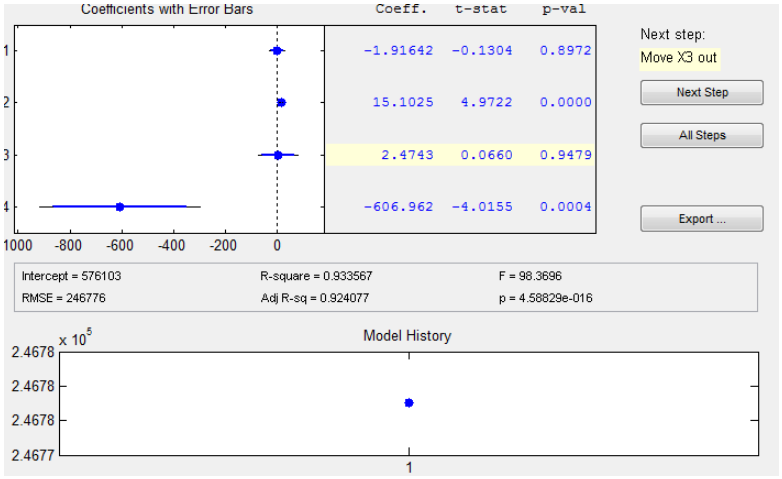

Figure 6. Stepwise regression process I
图 6. 逐步回归过程之一

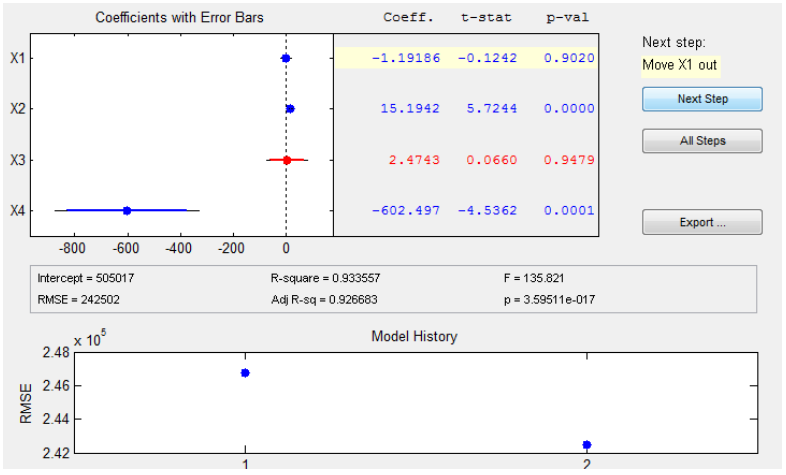

Figure 7. Stepwise regression process II
图 7. 逐步回归过程之二

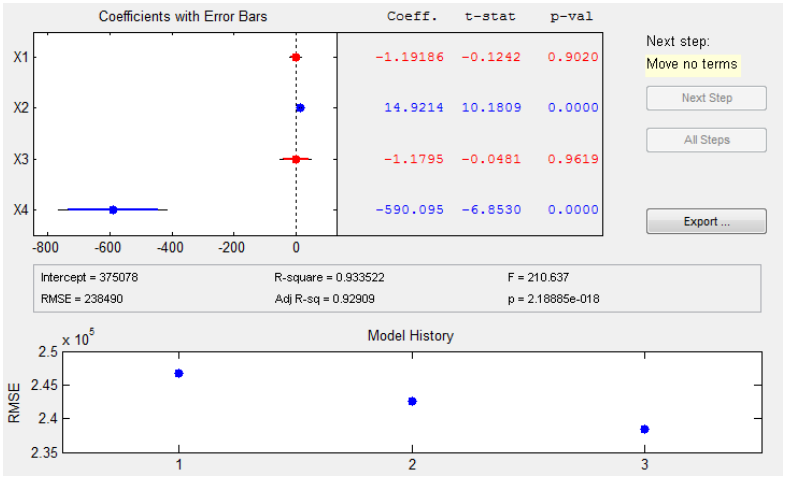

Figure 8. Stepwise regression process III
图 8. 逐步回归过程之三

Table 4. Five year value of each index
表 4. 各项指标五年数值

年份	公路客运量(万人)	国内生产总值(亿元)	民用载客汽车(万辆)
2011	3286220	473104.05	7478.37
2012	3557010	519470.1	8943.01
2013	3853463	590422.4	10561.78
2014	3908198	643974	14598.11
2015	3619097	685505.8	16284.45

Table 5. Compared with the original data values, the fitted value of highway passenger transport volume in recent five years
表 5. 近五年公路客运量拟合值与原数据对比

年份	2011	2012	2013	2014	2015
拟合值	3021504	3349073.7	3952553.2	3849074	3630591
原数值	3286220	3557010	3853463	3908198	3619097
相对误差	9.87610	0.06208	0.02506	0.01536	0.00316

生产总值和民用载客汽车。参考文献中选取的是总人口，这是因为我国是一个人口多大国，所以这项变量与公路客运量紧密相连。

5. 模拟实验

为了验证模型的合理性，选取 2011~2015 年的国内生产总值、民用载客汽车对公路客运量做实证分析，数据如下表 4。

5.1. 多元线性回归方程实证分析

将近五年的国内生产总值，民用载客汽车数值带入方程 $y=375078+14.9214x_2-590.095x_4$ 中，得出 2011~2015 年的客运量，结果与原数据作对比，如表 5。

5.2. 预测结果的分析

二元一次模型 $y=375078+14.9214x_2-590.095x_4$ 的拟合值与实际值相对误差较小，说明该模型预测公路客运量较为精确。这是因为载客汽车的多少对公路客运总量有着直接的影响，也进一步说明了该模型的实际应用意义。综上可选择民用载客汽车与公路客运量的回归模型来预测公路客运量[10]。

参考文献 (References)

[1] 何晓群, 刘文卿. 应用回归分析[M]. 北京: 中国人民大学出版社, 2011: 1-15.
[2] 景滨杰. 回归分析法在经济预测中的应用浅析[J]. 山西经济管理干部学院学报, 2004, 12(3): 32-34.
[3] 李柏年, 吴礼斌. MATLAB 数据分析方法[M]. 北京: 机械工业出版社, 2012: 33-80.
[4] 乔向明. 2003~2005 年我国公路客运量预测分析[J]. 山东交通学院学报, 2003, 11(1): 26-29.
[5] 庞皓. 计量经济学[M]. 北京: 科学出版社, 2010: 72-103.
[6] 王松桂, 陈敏, 陈立苹. 线性统计模型——线性回归与方差分析[M]. 北京: 高等教育出版社, 1999.
[7] 高惠璇. 应用多元统计分析[M]. 北京: 北京大学出版社, 2005: 22-45.

[8] 胡明伟, 史其信. 探讨回归分析在交通工程应用中的若干问题[J]. 公路交通科技, 2009, 19(1): 68-71.
[9] 杨巍, 张莉莉. 逐步回归分析在经济林产品需求预测中的应用[J]. 林业经济研究报告, 2009(8): 74-76.
[10] 贾俊平, 何晓群, 金勇进. 统计学[M]. 北京: 中国人民大学出版社, 2012: 265-317.