Operations Research and Fuzziology 运筹与模糊学, 2024, 14(1), 267-278
Published Online February 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.141026

Hans汉斯

基于多元统计分析方法对我国民航客运量的研究

张宇星，何引红

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2023年11月26日；录用日期：2023年12月15日；发布日期：2024年2月22日

摘　要

本文以1978~1993年间我国民航客运量作为研究对象，选择“国民收入”，“消费总额”，“铁路客运量”，“民航航线里程”，“来中国旅行乘客数”5个指标。运用线性回归、主成分分析、聚类分析等多种多元统计分析方法，对影响我国民航客运量的主要因素进行分析，并探讨我国民航客运量与其它因素之间的具体函数关系。得到了以下结论：首先，假设检验的结果表明，铁路客运量、民航航线里程、来中国旅客数量是影响民航旅客运量的三个主要因素，其中民航航线里程与来中国旅行乘客数对我国民航客运量有显著的正向影响，而铁路客运量对我国民航客运量有显著的负向影响；另外，本文比较了由线性回归、岭回归与主成分回归三种方法拟合得到的回归方程，结果显示，岭回归法与主成分回归法所建立的回归模型对我国民航客运量的拟合程度更好，且更符合实际情况。

关键词

线性回归，主成分分析，聚类分析

Research on Passenger Volume of Civil Aviation in China Based on Multivariate Statistical Analysis Method

Yuxing Zhang, Yinhong He

School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Nov. 26th, 2023; accepted: Dec. 15th, 2023; published: Feb. 22nd, 2024

Abstract

This paper took the passenger volume of China’s civil aviation during 1978~1993 as the research

文章引用: 张宇星, 何引红. 基于多元统计分析方法对我国民航客运量的研究[J]. 运筹与模糊学, 2024, 14(1): 267-278.
DOI: 10.12677/orf.2024.141026

object, and selected five indicators: "national income", "total consumption", "railway passenger volume", "civil aviation route mileage" and "the number of passengers traveling to China". Using linear regression, principal component analysis, cluster analysis and other multivariate statistical analysis methods, this paper analyzed the main factors affecting China's civil aviation passenger volume and discussed the specific functional relationship between China's civil aviation passenger volume and other factors. The following conclusions were obtained: First, the results of hypothesis test showed that the railway passenger volume, the civil aviation route mileage, and the number of passengers coming to China were the three main factors affecting the air route volume. The civil aviation route mileage and the number of passengers coming to China had a significant positive impact on the passenger volume of China's civil aviation, while the railway passenger volume had a significant negative impact on passenger volume of China's civil aviation. In addition, the regression equations fitted by linear regression, ridge regression and principal component regression were compared in this paper. The results showed that the regression model built by ridge regression and principal component regression fitted passenger volume of China's civil aviation better and was more suitable for the practical situation.

Keywords

Linear Regression, Principal Component Analysis, Cluster Analysis

1. 引言

近几年，中国民用航空业经历了巨大变革，从 2012 至 2016 年，仅 5 年时间，我国民用航空就以年均 7 个左右的速度增加，线路增速超过 50%，横跨海内外 362 个大都市，民用航空客流量超过 5 亿，引领中国进入“云端时代”。伴随着一系列国家政策的制定，航空运输作为一种现代的交通方式，在人们的生活中起到了举足轻重的作用[1]。

中国的民用航空在持续地探索与实践中，已经获得了令人振奋的成绩，民用航空在国民经济中的地位也变得日益突出，在人口流动，信息交流，区域经济建设，以及交通运输等方面都发挥着不可忽视的作用[2]。在我国民用航空业的发展中，旅客运输问题是一个不容忽视的问题，而对旅客运输能力的分析更是一个值得关注的问题。在国外，关于民航旅客数量的研究起步较早，取得的结果也更为丰富，但在国内研究在这一点上却相对缺乏。此外，近年来，国际上正在经受着疫情的考验，我国的民用航空也面临着不小的挑战。因此，无论是从充实学术理论，还是从提升我国的民用航空发展水平的角度来讲，对民用航空客运量以及它的影响因素进行研究都具有十分重要的意义[3]。

多年以来，我国的学术界也曾就航空旅客流量进行过相关的研究。李在林利用灰色关联分析法从国内生产总值、最终消费额、铁路客运量、民航航线里程、来华旅游这五个重要因子入手进行研究，得出了 GDP 对民航旅客的重要程度最高的结论[4]。李忠虎和他的团队从国民文化程度入手进行研究，研究表明，高校毕业生数量和民航旅客数量之间存在着明显的正相关关系，皮尔逊相关系数达到 0.994，并据此构造了一个具有较强关联度的变量：40 年内的滚动累积普通本科和专科毕业生数量，之后通过回归分析，对今后两年内对民航旅客数量进行了预报[5]。熊崇俊等人已经发现，当考虑到许多个因子的顺序对于航空旅客数量和旅客周转量的影响时，灰色关联理论可以很好地表达出各个因子的相对比较优势，并且可

以将这些因子的作用程度进行定量化[6]。根据灰度关联系数表的结果可以发现，对民航客运周转量的影响最大的是外贸总额和人均消费支出。杨浩然、Guillaume Burghouwt 等人选用从 2007 到 2013 年的数据，对 138 条铁路和民航存在竞争的线路进行了面板数据分析，计算出在铁路进入市场之后，铁路运行的时间、铁路运行的次数、铁路的费用等因素对于中国民航客流的具体影响[7]。他们的研究发现，当高速铁路参与到市场中时，平均可以使飞机的需求量降低 27%，也就是说，有 27%的民用飞机使用者将会转而使用高速铁路。

在国外，对民用航空客流量的预测以及对其产生的影响进行了大量的研究，比如，Farzin Nourzadeh 等利用人工神经网络对 2020 年伊朗的国际航班旅客数量进行了预测[8]。他们利用不同的训练算法，利用 11 个指标对与伊朗在一些方面状况类似的国家进行分类，然后利用不同的训练算法对这些国家的航班旅客数进行了预测。最终，通过权重平均数和与其它国家在指标内的相似度，对入境伊朗的客流量进行了估算。并且他们在试验误差的基础上，为各个国家选取了相应的训练算法，并得到了一个具有 99%准确率的置信区间。Volodymyr Bilotkach 等人发现亚洲航空市场已成为世界上增长最快的航空市场，其低价航空公司的数目也是该地区最多的。他们使用根据预定数据计算出的国际航空乘客数量，在亚洲 30 个主要机场找到了低价位航空公司对国际航空乘客流量的影响[9]。研究结果发现，低价航空公司对国际航空客流量有绝对正面的影响，而低价航空公司的网路订票服务也是市场集中度的主要影响因素，说明低价航空公司对亚洲国际航空客流量有很大的贡献。

通过上述对国内外研究现状的阐述可以发现，不管是通过传统的统计方法，还是通过统计算法，在寻求民航客运量的影响因素，以及对民航客运量的预测上，都有很多的研究成果。但是，以上研究选择的自变量较多，若自变量间存在多重共线性关系，将会对研究结果造成不同程度的影响，因此，需要寻找合适的方法，探索各自变量与自变量间的函数关系。本文以 1978~1993 年间我国民航客运量作为研究对象，选择“国民收入”，“消费总额”，“铁路客运量”，“民航航线里程”，“来中国旅行乘客数”5 个指标，运用线性回归，主成分分析，聚类分析等方法，对我国民航客运量的主要影响因素进行分析，并探讨其与其它因素之间的具体函数关系。

2. 研究方法

2.1. 多元回归分析

2.1.1. 多元线性回归模型

用线性回归方法构建预测方程是一种行之有效的方法。其基本过程是：通过实验和调研，对自变量和因变量进行多次观测；然后，确定经验公式的所属类别，建立了相应的数学模型，给出了待估计的参数；在此基础上，对待估计的参数进行了拟合，并进行统计分析。

一般的，我们称

$$y = b_0 + b_1 x_1 + \cdots + b_m x_m + \varepsilon \tag{1}$$

为多元线性回归模型，其中 $E(\varepsilon)=0$，$D(\varepsilon)=\sigma^2$；$b_0, b_1, \cdots, b_m$，σ^2 是未知参数。b_0 称为常数项或截距，$x_1, \cdots, x_m$ 是自变量，y 是因变量，另外本文还要求模型满足 Gauss-Markov 条件，得到整体线性回归模型为

$$\begin{cases} y_1 = b_0 + b_1 x_{11} + \ldots + b_m x_{1m} + \varepsilon_1 \\ \cdots \\ \cdots \\ y_n = b_0 + b_1 x_{n1} + \ldots + b_m x_{nm} + \varepsilon_n \end{cases}, \begin{cases} E_{\varepsilon i} = 0 \quad i = 1, 2, \cdots, n \\ Var \begin{bmatrix} \varepsilon_1 \\ \cdots \\ \varepsilon_n \end{bmatrix} = \sigma^2 I \end{cases} \tag{2}$$

2.1.2. 线性关系显著性 F 检验

线性关系显著性 F 检验即要检验 $H_0: b_i = 0, i = 1,2,\cdots,m$。

显然，SST = SSR + SSE，因变量的发散程度可以用总离差 SST 反映；由回归引起的分散性可以用回归平方和 SSR 反映，误差变量的分散性则由 SSE 反映。为此可以选择 SSR/SSE 为统计量，又由于 SSR，SSE 独立，且它们与 σ^2 的商分别服从 $\chi^2(m)$ 和 $\chi^2(n-m-1)$，因此得到 F 统计量为

$$F = \frac{\frac{SSR}{m}}{\frac{SSE}{n-m-1}} \sim F(m, n-m-1) \tag{3}$$

只需计算 F 的值，当 F 的值大于临界值时，拒绝 H_0。

还可以用复相关系数(也称为决定系数)的平方来检验回归模型的线性关系显著性：

$$R^2 = 1 - \frac{SSE}{SST} \tag{4}$$

当复相关系数的平方较大时，回归模型的线性关系显著。

2.1.3. 单个自变量显著性 t 检验

一个好的模型应该是所有自变量都有效的。如果 x_i 的系数 b_i 为零或绝对值很小，那么 x_i 则是无作用的。为此对每个 i 要检验

$$H_{0i}: b_i = 0 \tag{5}$$

并且若 $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$ 则 SSE 与 $\hat{\beta}$ 独立，从而得到 t 统计量

$$t_i = \frac{\left(\hat{b}_i - b_i\right)}{STDERR\left(\hat{b}_i\right)} \sim t(n-m-1) \tag{6}$$

当 H_{0i} 成立时，统计量 $t_i = \frac{\hat{b}_i}{STDERR\left(\hat{b}_i\right)}$，由 t_i 服从自由度为 $n-m-1$ 的 t 分布知道，如果 t_i 绝对值很大，大于临界值时，则应当拒绝 H_{0i}。

2.1.4. 预报

做预报是回归分析的重要目的之一。当 $\hat{b}_0, \hat{b}_1, \cdots, \hat{b}_m$ 得到后，就有了回归方程

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_m x_m \tag{7}$$

若再给定自变量变量的值 $u = \left(x_1^o, \cdots, x_m^o\right)'$，就可得到预报值

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1^o + \cdots + \hat{b}_m x_m^o \tag{8}$$

求预报值的置信区间的理论介绍较为复杂，因此不过多介绍，这里仅介绍置信区间的计算方法，设自变量的值为 $u = \left(x_1^o, \cdots, x_m^o\right)'$，令 $\boldsymbol{ug} = \left(1, x_1^o, \cdots, x_m^o\right)'$，$v = \boldsymbol{ug}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{ug}$，则概率为 $1-\alpha$ 的预测区间端点为

$$\hat{b}_0 + \hat{b}_1 x_1^o + \cdots + \hat{b}_m x_m^o \pm t_{1-\frac{\alpha}{2}}(n-m-1)\left(\hat{\sigma}^2(1+v)\right)^{\frac{1}{2}} \tag{9}$$

2.1.5. 多重共线性分析

在线性回归模型中，如果自变量(包括常数项)间有一种线性或者接近一种线性的关系，我们将这种关系称为共线性或者多重共线性。共线性可以表示为：一个自变量是其它自变量的线性组合，或者它是其

它自变量的线性组合。有三种常见的多重共线诊断方法：条件指数法、方差膨胀因子法、方差比例法。

1) 条件指数法

首先把矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 标准化，即做矩阵 $\boldsymbol{A}$，$\boldsymbol{A}$ 对角线上元素等于 $\boldsymbol{X}'\boldsymbol{X}$ 对角线，矩阵 $\boldsymbol{A}$ 其余元素为零。$B=\boldsymbol{A}^{-1/2}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A}^{-1/2}$ 称为标准化的 $\boldsymbol{X}'\boldsymbol{X}$。

若标准化的若标准化的 $\boldsymbol{X}'\boldsymbol{X}$ 有 k 个逼近于零的特征值，那么，预测因子中存在 k 个共线性关系。估计多重共线性的经验法则是：$1\le k\le 10$ 预示自变量间多重共线性较弱；$10\le k\le 100$ 预示解释变量间存在较强多重共线性；$100<k$ 预示解释变量间存在高度的多重共线性。

2) 方差膨胀因子法

设 R_i 为 x_i 对其余 $p-1$ 个自变量的复相关相关系数(也称为决定系数)，那么 $VIF_i=\dfrac{1}{\left(1-R_i^2\right)},i=1,2,\cdots,p$ 就称为方差膨胀因子。

由观测数据对每个解释变量 x_i 计算其方差膨胀因子 VIF_i，用 VIF_i 来估计多重共线性的经验方法是：如果 $VIF_i>10$，则表示第 i 个解释变量的多重共线性是高度显著的。

3) 方差比例法

对每一个被解释的变量(含常数项)，求出每一个主成分在总方差中所占的百分比，称为方差比例。当条件指数较大，并且同时对应的两个以上的方差比例超过 50%时，则认为各变量之间有显著相关关系。

2.1.6. 岭回归模型

如果自变量之间有多重共线性关系，那么就可以将那些无关紧要的自变量剔除掉，但是如果有些时候又不愿意将预测因子去掉，那么可用岭回归模型，它的基本原理是：多重共线性使 $\left|\boldsymbol{X}'\boldsymbol{X}\right|$ 等于零或近似于零，从而由参数估计公式 $\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$ 造成方差过大，如果改用岭估计 $\hat{\boldsymbol{\beta}}(c)=\left(\boldsymbol{X}'\boldsymbol{X}+c\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$，则可以避免这种情况，其中 c 是(0, 1)中的某个值，称为岭参数。

2.2. 主成分分析

2.2.1. 主成分分析原理

当存在若干个随机变量时，寻求它们的少量线性组合(即主成分)，用以解释这些随机变量，是很有必要的。主成分分析的数学模型是：对于随机向量 $\boldsymbol{X}$，选一些常数向量 e_i，用 $c_i'\boldsymbol{X}$ 尽可能多反映随机向量 $\boldsymbol{X}$ 的主要信息。也即 $D\left(c_i'\boldsymbol{X}\right)$ 尽量大。但是 c_i 的模可以无限增大，从而使 $D\left(c_i'\boldsymbol{X}\right)$ 无限变大，这是我们不希望的；于是限定 c_i 模的大小，而改变 c_i 各分量的比例，使 $D\left(c_i'\boldsymbol{X}\right)$ 最大；通常取 c_i 的模为 1 最方便。

设随机向量 $\boldsymbol{X}=\left(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_p\right)'$ 二阶矩存在，若常数向量 c_1，在条件 $\left\|c\right\|=1$ 下使 $D\left(c'\boldsymbol{X}\right)$ 最大，则称 $Y_1=c'\boldsymbol{X}$ 是 $\boldsymbol{X}$ 的第一主成分或第一主分量。由定义可见，Y_1 尽可能多地反映原来 p 个随机变量变化的信息。但是一个主成分往往不能完全反映随机向量特色，必须建立其它主成分，它们也应当最能反映随机向量变化，而且他们应当与第一主成分不相关(不包含 Y_1 的信息)。

2.2.2. 主成分回归

在回归分析中，常遇到自变量存在多重共线性问题，即自变量的观测值存在线性相关，或近似线性相关。这时设计矩阵满足 $\left|\boldsymbol{X}'\boldsymbol{X}\right|\approx 0$，之前已指出，用公式 $\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$ 估计参数会造成较大方差。选取彼此正交的的主成分，用少量主成分作回归，再将主成分化为原始变量，这样得到的回归方程就不存在较大方差了。

具体来说，主成分回归，是运用主成分分析的方法，从 $m+1$ 个自变量中，选择出彼此无关的头 q 个主成分；在此基础上，以 q 个主成分为自变量进行回归分析；再保持因变量不变，用这 q 个主成分作为自变量作回归；最后把所得回归结果作变量代换，转化成原来因变量与原来自变量的关系。

2.3. 聚类分析

本文选用的聚类方法为系统聚类法中的类平均法。系统聚类法的基本思想是首先将 n 个样品各自作为一类，并规定样品之间的距离和类与类之间的距离，然后将距离最近的两类合并成一个新类，并求出新类与其他类间的间距；重复进行两个最近类的合并，每次减少一类，直到所有的样本都合并为一个类别。

类平均法有两种定义，本文选用的定义方法是把类与类之间的距离定义为所有样品对之间的平均距离，即定义 G_K 和 G_L 之间的距离为

$$D_{KL}=\frac{1}{n_K n_L}\sum_{i\in G_K,j\in G_L} d_{ij} \tag{10}$$

其中 n_K 和 n_L 分别为类 G_K 和 G_L 的样品个数， d_{ij} 为 G_K 中样品 i 与 G_L 中样品 j 之间的距离[10]。

3. 实证分析

3.1. 假设检验

本文以 1978~1993 年间我国民航客运量作为因变量，选择“国民收入”，“消费总额”，“铁路客运量”，“民航航线里程”，“来中国旅行乘客数”5 个指标作为自变量进行假设检验得到如下结果。

表 1 给出了方差分析表的上半部分，指出了各种平方和的来源和自由度等，从第 5 列中可以看出 F 值为 281.65，而第 6 列则是自由度为 5，9 的 F 分布随机变量大于 281.65 的概率，如果这个概率小于 0.0001，那就说明 F 的值大于 0.9999 分位点，也就说我国民航客运量与五个之变量之间存在显著的线性关系。

Table 1. Part of the analysis of variance table
表 1. 部分方差分析表

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8,513,633	1,702,727	281.65	<0.0001
Error	9	54,410	6045.58881		
Corrected Total	14	8,568,044			

Table 2. Parameter estimation table
表 2. 参数估计表

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	539.53044	329.75800	1.64	0.1362
X_1	1	0.18331	0.14918	1.23	0.2503
X_2	1	−0.22901	0.21008	−1.09	0.3040
X_3	1	−0.00880	0.00385	−2.29	0.0482
X_4	1	18.30704	6.22727	2.94	0.0165
X_5	1	0.28746	0.06099	4.71	0.0011

从表 2 中可以看出各变量的参数估计值，标准误，t 值等，第 6 列指出自由度为 9 时 t 分布大于这些 t 值的概率，若概率小于 0.05 则表示变量的作用显著，概率小于 0.01 表明变量的作用高度显著，因此可以得出铁路客运量 x_3，民航航线里程数 x_4 的作用是显著的，来中国旅游人数 x_5 的作用是高度显著的，国

民收入 x_1，消费总额 x_2 和截距的作用是不显著的。但是，在回归模型下，消费总额的数值是负的，这与实际情况不符：消费总额越高，就越有可能选择航空出行；而线性回归公式的系数是负值，意为当消费总额越大时，坐飞机的人越少，说明线性回归公式虽然拟合度较好，但并不符合实际，仍需进一步的回归诊断。

3.2. 线性回归预测

表 3 中给出了 15 次预测的参数估计值，其中第 2 列是因变量观测值，第 3 列是因变量预报值，第 5 列和第 6 列分别是预报值 95%置信区间下限和上限。从表中最后一列可以看出观测值和预测值之间存在巨大的差距，最大值达到了−176.524，置信区间上限和下限之间的差距也较大，这也体现了线性回归拟合的方程不符合实际，需要进一步诊断改进。

Table 3. Linear regression prediction table
表 3. 线性回归预测表

Obs	Variable	Dependent Value	Predicted Mean Predict	Std Erro 95% CL Predict		Residual
1	231	266.5075	52.7956	53.9013	479.1138	−35.5075
2	298	304.5910	43.1997	103.3759	505.8061	−6.5910
3	343	346.1419	30.8940	156.8759	535.4079	−3.1419
4	401	405.1773	37.3940	210.0028	600.3518	−4.1773
5	445	395.2868	37.9217	199.5919	590.9816	49.7132
6	391	397.4693	33.5299	205.9213	589.0173	−6.4693
7	554	556.1622	39.9316	358.4320	753.8924	−2.1622
8	744	742.7242	32.3814	552.1900	933.2583	1.2758
9	997	1003	49.8460	794.1096	1212	−6.0405
10	1310	1281	45.3146	1077	1484	29.2942
11	1442	1367	61.2551	1143	1591	75.2098
12	1283	1460	44.4053	1257	1662	−176.5240
13	1660	1551	64.0190	1323	1779	108.9636
14	2178	2208	62.6757	1983	2434	−30.4958
15	2886	2879	75.9807	2633	3125	6.6528

3.3. 多重共线性检验

从之前的介绍中可知，当方差膨胀因子大于 10 时，变量便存在高度的多重共线性。从表 4 中可以看出，国民收入 x_1，消费额 x_2，民航航线里程数 x_4，来华旅游人数 x_5 的方差膨胀因子都大于 10，国民收入 x_1，消费额 x_2 的方差膨胀因子甚至都超过了 1000，说明这些变量的共线性是很显著的。

当条件指数较大，并且同时对应的两个以上的方差比例超过 50%时，就判定这些变量间存在相关性。而从表 5 中最后一行可见条件数 246.65073 远大于 30，因而变量之间确实存在高度多重共线性。而 246.65073 对应的方差比例中国民收入 x_1 为 0.99597，消费总额 x_2 为 0.99132 都远大于 50%，因而国民收入 x_1，消费总额 x_2 是高度相关的，需要利用其他解决这个问题。

Table 4. Parameter estimation table and expansion factor table
表 4. 参数估计表和膨胀因子表

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	539.53044	329.75800	1.64	0.1362	0
X_1	1	0.18331	0.14918	1.23	0.2503	1521.45442
X_2	1	−0.22901	0.21008	−1.09	0.3040	1276.66982
X_3	1	−0.00880	0.00385	−2.29	0.0482	4.43432
X_4	1	18.30704	6.22727	2.94	0.0165	30.91690
X_5	1	0.28746	0.06099	4.71	0.0011	10.43019

Table 5. Conditional index table and variance scale table
表 5. 条件指数表和方差比例表

Number	Eigenvalue	Condition Index	Var Porp Intercept	Var Porp x_1	Var Porp x_2	Var Porp x_3	Var Porp x_4	Var Porp x_5
1	5.61163	1.00000	0.00010	0.00001	0.00001	0.00007	0.00021	0.00079
2	0.32297	4.16837	0.00421	0.00006	0.00006	0.00247	0.00089	0.00566
3	0.05594	10.01589	0.00171	0.00018	0.00010	0.00133	0.02127	0.32584
4	0.00694	28.43029	0.00377	0.00353	0.00645	0.00035	0.59353	0.11816
5	0.00243	48.08208	0.66520	0.00024	0.00206	0.51300	0.01544	0.52017
6	0.00009	246.65073	0.32501	0.99597	0.99132	0.48278	0.36866	0.02937

3.4. 岭回归

岭回归方法可以解决上述研究中存在的多重共线性问题，因此选用岭回归方法并探究变量之间更准确的函数关系。

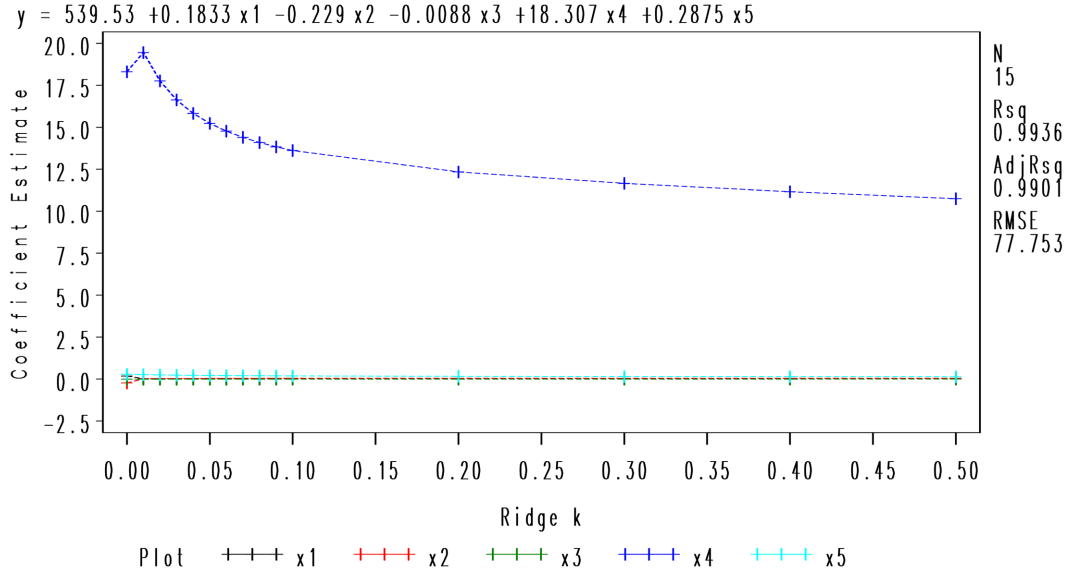

Figure 1. Diagram of the estimated parameters for ridge regression
图 1. 岭回归估计参数图

如图 1 所示，在岭参数值大于 0.02 时，曲线都趋于平缓，因此，取岭参数值为 0.02。在表 6 中岭参数栏中，查找其中岭参数值为 0.02 的两行，即第 6 行和第 7 行，第 6 行给出了 VIF 的值，第 7 行给出了参数估计值，从而得出了岭回归方程为

$$y = 637.053 + 0.03x_1 + 0.01x_2 - 0.01x_3 + 17.757x_4 + 0.249x_5 \tag{11}$$

这时 x_2 的系数为正，与实际情况相符，说明岭回归的结果比线性回归的结果更好。

Table 6. Part of the ridge estimate table
表 6. 部分岭估计表

Obs_	_TYPE_	_RIDGE_	_RMSE_	Intercept	x_1	x_2	x_3	x_4	x_5	y
1	PARMS	.	77.753	539.530	0.18	−0.23	−0.009	18.307	0.288	−1
2	RIDGEVIF	0.00	.	.	1521.45	1276.67	4.434	30.917	10.430	−1
3	RIDGE	0.00	77.753	539.530	0.18	−0.23	−0.009	18.307	0.288	−1
4	RIDGEVIF	0.01	.	.	5.09	6.68	2.080	11.469	7.624	−1
5	RIDGE	0.01	85.018	679.513	0.03	0.01	−0.011	19.455	0.268	−1
6	RIDGEVIF	0.02	.	.	2.61	3.74	1.832	7.690	6.023	−1
7	RIDGE	0.02	88.680	637.053	0.03	0.03	−0.010	17.757	0.249	−1
8	RIDGEVIF	0.03	.	.	1.74	2.53	1.645	5.577	4.918	−1
9	RIDGE	0.03	92.041	600.150	0.03	0.03	−0.010	16.627	0.235	−1
10	RIDGEVIF	0.04	.	.	1.27	1.86	1.499	4.264	4.110	−1

3.5. 主成分分析

在表 7 中，第 2 列显示了各因子的特征值，在第 5 列中，前 2 个特征值在总变差中所占比例为 99.3%，说明仅用 2 个主成分就可以对所有变化进行解释。

Table 7. Eigenvalue table of sample covariance matrix
表 7. 样本协差阵的特征值表

	Eigenvalue	Difference	Proportion	Cumulative
1	130,006,787	75,189,247	0.7029	0.7029
2	54,817,540	54,704,841	0.2964	0.9993
3	112,699	99,518	0.0006	0.9999
4	13,181	13,171	0.0001	1.0000
5	10		0.0000	1.0000

从表 8 可以看出，第一，二主成分分别为：

$$y_1 = 0.2742x_1 + 0.1861x_2 + 0.9404x_3 + 0.0009x_4 + 0.0764x_5 \tag{12}$$

$$y_2 = 0.7916x_1 + 0.4975x_2 - 0.338x_3 + 0.0028x_4 + 0.1073x_5 \tag{13}$$

由于第一主成分中铁路客运量 x_3 是较大正数，说明了我国民航客运量减少的主要因素，主要是受到铁路客运量的影响。第二主成分铁路客运量 x_3 的系数为负，而其他变量的系数为正，也反应了各个变量对民航客运量的影响是有差异的。

Table 8. Eigenvector table of sample covariance matrix
表 8. 样本协差阵的特征向量表

	Prin1	Prin2	Prin3	Prin4	Prin5
x_1	0.274242	0.791638	−0.117693	−0.532930	−0.015306
x_2	0.186068	0.497511	−0.051062	0.845533	0.018115
x_3	0.940388	−0.338026	−0.036101	−0.010223	−0.000213
x_4	0.000881	0.002828	−0.002780	−0.023444	0.999717
x_5	0.076411	0.107337	0.991075	−0.020162	0.001912

3.6. 主成分回归

主成分回归同样可以解决多重共线性问题，这是因为，从自变量中选择出来的主成分都是相互正交的，因此，只需要用少量的主成分对其进行回归，再将其转换为原始变量，得出的回归方程的方差便不会很大。通过表 9 中的最后一行结果可知主成分回归方程为：

$$y = 702.573 + 0.0388x_1 + 0.0597x_2 - 0.0106x_3 + 9.9057x_4 + 0.2185x_5 \tag{14}$$

其中消费额 x_2 的系数为正，与实际相符，也说明主成分回归的结果优于线性回归结果。

Table 9. Part of the principal component regression table
表 9. 部分主成分回归表

Obs	_TYPE_	_PCOMIT_	_RMSE_	Intercept	x_1	x_2	x_3	x_4	x_5	y
1	PARMS	.	75.9488	622.327	0.1343	−0.1572	−0.0097	18.4435	0.2928	−1
2	IPC	2	96.1468	702.573	0.0388	0.0597	−0.0106	9.9057	0.2185	−1

3.7. 聚类回归

Figure 2. Diagram of the cluster regression pedigree
图 2. 聚类回归谱系图

由图 2 的谱系图可以看出，若把这些变量作为自变量，年份作为因变量进行聚类分析，大致可以分为三类，1978~1983 为一类，1984~1988 为一类，1989~1993 为一类。因此可以得到结论，我国民航的发

展是与时间息息相关的，在改革开放以后，我们国家的经济得到了很大的发展，给我国的民航业的发展带来了更好的社会条件。

4. 结论与讨论

本文以 1978~1993 年间我国民航客运量作为研究对象，选择“国民收入”，“消费总额”，“铁路客运量”，“民航航线里程”，“来中国旅行乘客数”5 个指标进行研究。

首先由假设检验的结果得到，铁路客运量，民航航线里程数，来中国旅行乘客数这三个变量对民航客运量有显著性影响，其中民航航线里程与来中国旅行乘客数对我国民航客运量有显著的正向影响，而铁路客运量对我国民航客运量有显著的负向影响，这也与李在林的研究结果相似[4]。

但是，在线性回归模型下，消费总额的数值是负的，这与实际情况不符：消费总额越高，就越有可能选择航空出行；而线性回归公式的系数是负值，意为当消费总额越大时，坐飞机的人越少，说明线性回归公式虽然拟合度较好，但并不符合实际，仍需进一步的回归诊断。在李丽华的研究中同样出现了类似的问题，其线性回归模型显示铁路客运量的系数为正，意为当铁路客运量增加时，我国民航客运量也会增加[1]。但本文研究显示，铁路客运量的增加是降低我国民航客运量的显著因素，因此李丽华的研究结果同样说明了线性回归模型具有一定的局限性。

于是本文利用多重共线性检验进行回归诊断分析发现，国民收入，消费总额等变量之间都存在严重的多重共线性，而多重共线性的存在影响了线性回归模型的拟合效果，因此需要选择其他的方法拟合回归方程。本文选用的是利用岭回归和主成分回归两种方法进行拟合，结果显示这两种方法拟合得到的回归方程更符合实际情况，优于线性回归方程。

最后，本文利用聚类分析方法进行研究发现，我国民航事业的发展是随着时间阶段飞速变化的，这也与实际情况相符，说明国家的不断发展也是一项重要的影响因素。

与其他研究相比，本文的创新点在于使用了更多的多元统计分析方法研究我国民航客运量与五个指标之间具体的函数关系，并利用岭回归和主成分回归方法避免了线性回归中多重共线性问题带来的影响，同时对三种回归方法进行比较，得到了更准确，更符合实际结果的函数表达式，为之后的研究提供了参考。

基金项目

资助项目：国家自然科学基金(批准号：32000778)。

参考文献

[1] 李丽华. 我国民航客运量影响因素研究[J]. 纳税, 2018, 12(22): 245.

[2] You, H., Yang, J., Xue, B., Xiao, X., Xia, J.C., Jin, C. and Li, X. (2021) Spatial Evolution of Population Change in Northeast China during 1992-2018. *Science of the Total Environment*, **776**, Article ID: 146023. https://doi.org/10.1016/j.scitotenv.2021.146023

[3] Tang, X. and Deng, G. (2016) Prediction of Civil Aviation Passenger Transportation Based on ARIMA Model. *Open Journal of Statistics*, **6**, 824-834. https://doi.org/10.4236/ojs.2016.65068

[4] 李在林. 民航客运需求影响因素的灰色关联分析[J]. 经济视角(中旬), 2011(7):186.

[5] 李忠虎, 何苗. 民航客运量与国民受教育水平相关性研究[J]. 四川文理学院学报, 2020, 30(2): 76-81.

[6] 熊崇俊, 宁宣熙, 潘颖莉. 基于灰色关联理论的民航客运影响因素研究[J]. 统计与决策, 2006(1): 52-53.

[7] Yang, H., Burghouwt, G., Wang, J., Boonekamp, T. and Dijst, M. (2018) The Implications of High-Speed Railways on Air Passenger Flows in China. *Applied Geography*, **97**, 1-9. https://doi.org/10.1016/j.apgeog.2018.05.006

[8] Nourzadeh, F., Ebrahimnejad, S., Khalili-Damghani, K. and Hafezalkotob, A. (2020) Forecasting the International Air Passengers of Iran Using an Artificial Neural Network. *International Journal of Industrial and Systems Engineering*, **34**, 562-581. https://doi.org/10.1504/IJISE.2020.106089

[9] Bilotkach, V., Kawata, K., Kim, T.S., Park, J.K., Purwandono, P. and Yoshida, Y. (2019) Quantifying the Impact of Low-Cost Carriers on International Air Passenger Movements to and from Major Airports in Asia. *International Journal of Industrial Organization*, **62**, 28-57. https://doi.org/10.1016/j.ijindorg.2018.03.012

[10] 吴诚欧, 秦伟良. 近代实用多元统计分析[M]. 北京: 气象出版社, 2007.