【信息科学与控制工程】

民航客运量的多元线性回归分析

张　艳,苗　刚,李盈科

(新疆农业大学　数理学院,乌鲁木齐　830052)

摘要:为了对民航业务量做出准确地评估和预测,利用多元线性回归分析方法对民航客运量的变化趋势及成因建立了多元线性回归模型,并从国民收入、消费额、铁路客运量、民航航线里程、来华旅游入境人数等方面进行了多元线性回归分析。

关键词:回归分析;最小二乘法;回归方程;显著性检验;民航客运量

中图分类号:O212　　**文献标识码**:A　　**文章编号**:1006 - 0707(2012)08 - 0081 - 04

在实际问题中,常遇到研究一个随机变量与多个变量之间的相关关系,如,某产品的销售额不仅受到投入的广告费用的影响,还与产品价格、消费者收入状况、社会保障及其它可替代产品的价格等其他因素有关系。研究这种一个随机变量同多个变量之间关系的方法主要是多元回归分析法。

目前,我国国民收入实现了快速增长,民航业蓬勃发展,为了对民航业务量做出准确地评估和预测,民航客运量的变化趋势及成因成为航空公司关心的主要问题。影响我国民航客运量的因素,不仅有经济因素、政治因素,还有天气因素、季节因素,这些因素对我国民航客运量的变化影响程度各有不同,而这些因素的不同组合也会产生不同的效果。本文从国民收入、消费额、铁路客运量、民航航线里程、来华旅游入境人数等几个方面出发,运用多元回归分析法来研究其变化趋势及成因问题。

1　多元线性回归模型

1.1　多元线性回归模型的一般形式

设影响因变量 y 的自变量个数为 m 个,记为 $x_1,x_2,\cdots,x_m$,多元线性模型是指这些自变量对 y 的影响是线性的,即关系式

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \tag{1}$$

其中:$\beta_0,\beta_1,\beta_2,\cdots,\beta_m$ 是 $m+1$ 个未知参数,β_0 为常数项,$\beta_1,\beta_2,\cdots,\beta_m$ 称为回归系数;$x_1,x_2,\cdots,x_m$ 是 m 个可得到精确值并能够控制的一般变量,称为解释变量,称 y 为对自变量 $x_1,x_2,\cdots,x_m$ 的线性回归函数。当 $m=1$ 时,式(1)为一元线性线性回归模型,$m\geqslant 2$ 时,称为多元线性回归模型。ε 是随机误差,通常认为 $\varepsilon \sim N(0,\sigma^2)$。

在实际问题中,获得 n 组关于 $(x_{i1},x_{i2},\cdots,x_{im},y_i)$ $(i=1,2,\cdots,n)$ 观测数据,则

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1m} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2m} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{nm} + \varepsilon_n \\ \varepsilon_1,\varepsilon_2\cdots\varepsilon_n \overset{iid}{\sim} N(0,\sigma^2) \end{cases} \tag{2}$$

其中 $i=1,2,\cdots,n$,这个模型称为多元线性回归模型。

令

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

则上述数学模型的矩阵形式为

$$Y = X\beta + \varepsilon \tag{3}$$

其中 ε 是 n 维随机向量,它的各个分量相互独立同分布。

1.2　多元线性回归模型的基本假定

一般认为回归模型应满足以下几个基本假设:

1)解释变量 $x_1,x_2,\cdots,x_m$ 是随机变量,观测值 $(x_{i1},x_{i2},\cdots,x_{im})$ 为常数。

2)方差齐性及不相关的假定条件为[1]

$$\begin{cases} E(\varepsilon_i) = 0, i = 1,2,\cdots,n \\ COV(\varepsilon_i,\varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} (i,j = 1,2,\cdots,n) \end{cases} \tag{4}$$

这个称为高斯 - 马尔柯夫(Gauss-Markov)条件,简记为 G-M 条件。在此条件下,可以得到关于回归系数方程一些重要性质,比如,得到关于回归系数的最小二乘估计是回归系数的

收稿日期:2012 - 04 - 06

作者简介:张艳(1978—),女,主要从事数学科学研究。

最小方差线性无偏估计等[7]。

3）正态分布的假定条件为

$$\begin{cases}\varepsilon_i \sim N(0,\sigma^2), i = 1,2,\cdots,n \\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n \text{ 相互独立}\end{cases} \tag{5}$$

在此条件下可以得到关于回归系数的估计及 σ^2 估计的进一步的结果，比如，它们分别是回归系数及 σ^2 的最小方差无偏估计等，而且还可以作回归的显著性检验及区间估计[7]。

1.3 回归方程的显著性检验

1.3.1 回归系数的 t 检验

在多元线性回归问题中，回归方程显著并不能说明每个自变量对 y 的影响都显著，所以总想从回归方程中去除一些相关度比较低的变量得到其精简的回归方程。这时就需要对每个自变量进行显著性检验。

显然，若某个自变量 x_i 对 y 的作用不显著，那么在回归模型中，它的系数 β_i 就取值为 0。因此检验变量 x_i 是否显著，等价于检验假设

$$H_{0i}:\beta_i = 0 \quad i = 1,2,\cdots,m \tag{6}$$

如果接受原假设 H_{0i}，则 x_i 不显著；否则 x_i 是显著的。

可以知道[3]

$$\hat{\beta} \sim N(\beta,\sigma^2(X^TX)^{-1}) \tag{7}$$

记

$$(X^TX)^{-1} = (c_{ij}), i,j = 1,2,\cdots,m \tag{8}$$

于是有

$$\begin{aligned} &E(\hat{\beta}_i) = \beta_i, \mathrm{var}(\hat{\beta}_i) = c_{ij}\sigma^2 \\ &\hat{\beta}_i \sim N(\beta_i, c_{ij}\sigma^2) \end{aligned}, \ i = 0,1,2,\cdots,m \tag{9}$$

据此构造 t 统计量

$$t_i = \frac{\hat{\beta}_i}{\sqrt{c_{ii}}\hat{\sigma}} \tag{10}$$

其中

$$\hat{\sigma} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

是回归标准差。

当原假设 $H_{0i}:\beta_i = 0$ 成立时，式(10)构造的 t_i 统计量服从自由度为 $n-m-1$ 的 t 分布。给定显著性水平 α，查出双侧检验的临界值 $t_{\alpha/2}$。当 $|t_i| \geqslant t_{\alpha/2}$ 时拒绝原假设 $H_{0i}:\beta_i = 0$，认为 β_i 显著不为 0，认为 β_i 显著不为 0，自变量 x_i 对因变量 y 的线性效果显著；反之认为 β_i 为 0，自变量 x_i 对因变量 y 的线性效果不显著[4,8]。

1.3.2 回归系数的 F 检验

对多元线性回归方程的显著性检验就是看随机变量 x_1，$x_2,\cdots,x_m$ 从整体上对 y 是否有明显的影响。因此提出原假设

$$H_0:\beta_1 = \beta_2 = \cdots = \beta_m = 0$$

如果原假设被接受，则表明随机变量 y 与 $x_1,x_2,\cdots,x_m$ 之间的关系由线性回归模型表示不合适。一般用 F 检验来判别，为了建立对 H_0 进行检验的 F 统计量，用总离差平方和的分解式，即

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{12}$$

简写为

$$SST = SSR + SSE \tag{13}$$

此时用 F 检验统计量

$$F = \frac{SSR/m}{SSE/(n-m-1)} \tag{14}$$

在正态性假设下，当 $H_0:\beta_1 = \beta_2 = \cdots = \beta_m = 0$ 成立时，F 服从自由度为 $(m,n-m-1)$ 的 F 分布，于是可利用 F 统计量对回归方程的总体显著性进行检验。对于给定的数据，当 $i = 1,2,\ldots,n$，计算出 SSR 和 SSE，进而得到对应 F 的值，见一般列在下面的方差分析表中，再由给定的显著性水平 α，查 F 分布表，得到临界值 $F_\alpha(m,n-m-1)$。

表 1 方差分析表

方差来源	自由度	平方和	均方	F 值
回归	m	SSR	SSR/m	$\frac{SSR/m}{SSE/(n-m-1)}$
残差	$n-m-1$	SSE	$SSE/(n-m-1)$	
总和	$n-1$	SST		

当 $F > F_\alpha(m,n-m-1)$ 时，拒绝 H_0，认为在显著性水平 α 下，y 对 $x_1,x_2,\cdots,x_m$ 有显著的线性关系，也即回归方程的检验是显著的，就是接受“自变量全体对 y 有显著线性影响”这一结论犯错误的概率不超过 5%；反之，当 $F \leqslant F_\alpha(m,n-m-1)$ 时，接受 H_0，则认为回归方程不显著[8]。

1.4 置信区间和拟合优度

1.4.1 回归系数的置信区间

当有了参数向量 β 的估计 $\hat{\beta}$ 时，对于 $\hat{\beta}$ 与 β 的接近程度如何？这就需要构造 β_j 一个区间，以 $\hat{\beta}_j$ 为中心的区间，该区间以一定的概率包含 β_j，也就是作其对应的区间估计。可得

$$t_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{c_{ii}}\hat{\sigma}} \sim t(n-m-1) \tag{15}$$

按照一元线性回归系数区间估计的推导过程，可得 β_j 置信水平为 $1-\alpha$ 的置信区间为

$$(\hat{\beta}_i - t_{\alpha/2}\sqrt{c_{ii}}\hat{\sigma}, \hat{\beta}_i + t_{\alpha/2}\sqrt{c_{ii}}\hat{\sigma}) \tag{16}$$

1.4.2 拟合优度

拟合优度用于检验回归方程对样本观测值的吻合程度。在多元线性回归中，定义样本相关系数为

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{17}$$

样本决定系数 R^2 的取值在 $[0,1]$ 区间内，R^2 距离 1 越近，表明拟合的效果越好；R^2 距离 0 越近，表明拟合的效果越差。与 F 检验相比，R^2 检验可以更清楚直观的反映回归拟合的效果，但是并不能做为严格的显著性检验。称

$$R = \sqrt{R^2} = \sqrt{\frac{SSR}{SST}} \tag{18}$$

为 y 关于 $x_1,x_2,\cdots,x_m$ 的样本复相关系数[5]。

2　民航客运量模型的建立与求解

2.1　民航客运量模型的建立与求解

1）数据来源

以预测值 y 表示民航客运量(万人),x_1 表示国民收入总值(亿元),x_2 表示消费金额(亿元),x_3 表示铁路承载量(万人),x_4 表示民航航线距离(万公里),x_5 表示境外旅客人数(万人)。根据《2010 年统计摘要》获得 1995—2010 年统计数据,见表 2。

2）研究方法

建立 y 与各自变量 x_i,$1\leqslant i\leqslant 5$ 的多元线性回归模型如下

$$\begin{aligned} &y=\beta_0+\beta_i\sum_{i=1}^{5}x_i+\varepsilon\quad(i=1,2,3,4,5)\\ &\varepsilon\sim N(0,\sigma^2)\\ &E(\varepsilon_i)=0\quad(i=1,2,3,4,5)\end{aligned}\tag{19}$$

3）实证分析

利用原始数据资料,用 *SPSS* 软件计算相关阵,输出结果见表 3,并作相关分析。

从相关矩阵可以看出,y 与 x_1,x_2,x_3,x_4,x_5 相关系数都比较高,在 0.9 以上,说明所选自变量与 y 具有高度线性相关,用 y 与自变量 x_1,x_2,x_3,x_4,x_5 作多元线性回归是可以的。y 与 x_3 的相关系数 $r_{y3}=0.226$ 偏小,P 值 $=0.398$,x_3 是铁路客运量,这说明铁路客运量对民航客运量无显著影响。

4）计算结果

本例对原始数据作回归分析,并用 *SPSS* 软件计算,输出结果见表 4～6。

5）回归诊断

a. 回归方程为

$$\hat{y}=450.9+0.354x_1-0.561x_2-0.007\,3x_3+21.578x_4+0.435x_5\tag{20}$$

b. 复相关系数 $R=0.999$,决定系数 $R^2=0.988$,由相关系数来看回归方程高度显著。

c. 方差分析表中,$F=1\,128.303$,P 值 $=0.000$ 表明回归方程高度显著,说明 x_1,x_2,x_3,x_4,x_5 整体上对 y 有高度线性关系。

d. 回归系数的显著性检验。自变量 x_1,x_2,x_3,x_4,x_5 对 y 均有显著影响,其中 x_3 铁路客运量的 P 值 $=0.006$ 最大,可是仍然在 1% 的显著性水平上对 y 具有高度显著,这充分说明在多元回归分析中,不能仅凭相关系数的大小而决定变量的取舍。

6）回归应用

预测值的点估计为

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{10}+\hat{\beta}_2x_{20}+\cdots+\hat{\beta}_mx_{m0}\tag{21}$$

其精确置信区间的表达式较为复杂,也不可能用手工计算,可以仿照一元线性回归的情况用 SPSS 软件计算。其置信水平为 95% 的近似置信区间为

$$(\hat{y}-2\hat{\sigma},\hat{y}+2\hat{\sigma})\tag{22}$$

另外,x_2 的回归系数 -0.561 是负的,x_2 是消费额,负的回归系数显然是不合理的,其主要原因可能是由于自变量之间存在的共线性,因而回归方程式(2)还要在多重共线性部分作一步改进,或用其他消除共线性的方法重新建立回归方程,就不再讨论了。详见参考文献[5]。

表 2　各主要因素统计数据表

年份	y	x_1	x_2	x_3	x_4	x_5
1995	277	3 162	2 266	97 789	18	217
1996	358	4 020	2 634	10 3667	19	504
1997	412	4 426	3 037	110 645	23	684
1998	481	4 729	3 359	114 360	26	932
1999	534	5 110	3 665	119 906	28	951
2000	469	5 683	4 030	127 253	27	1 137
2001	665	6 782	4 686	132 424	31	1 542
2002	893	8 424	5 855	134 532	33	2 140
2003	1 196	9 431	6 662	130 295	39	2 738
2004	1 572	11 176	7 663	134 915	47	3 228
2005	1 730	14 086	9 646	147 174	45	3 803
2006	1 540	15 811	10 806	136 568	57	2 940
2007	1 992	17 261	11 596	114 854	61	3 295
2008	2 614	19 868	13 163	114 097	67	4 003
2009	3 463	24 268	15 582	119 632	100	3 974
2010	4 060	29 858	19 139	126 550	115	4 983

表3 相关阵表

		y	x_1	x_2	x_3	x_4	x_5
y	相关系数	1.000	0.988	0.984	0.226	0.986	0.923
	P 值	0.000	0.000	0.000	0.388	0.000	0.000
x_1	相关系数	0.989	1.000	0.999	0.257	0.983	0.929
	P 值	0.000	0.000	0.000	0.335	0.000	0.000
x_2	相关系数	0.984	0.999	1.000	0.288	0.977	0.941
	P 值	0.000	0.000	0.000	0.278	0.000	0.000
x_3	相关系数	0.226	0.257	0.288	1.000	0.212	0.503
	P 值	0.389	0.335	0.278	0.000	0.428	0.046
x_4	相关系数	0.986	0.983	0.977	0.212	1.000	0.881
	P 值	0.000	0.000	0.000	0.428	0.000	0.000
x_5	相关系数	0.923	0.929	0.941	0.503	0.881	1.000
	P 值	0.000	0.000	0.000	0.046	0.000	0.000

表4 常用统计量表

模型	相关系数	判定系数	调整的判定系数	回归估计的标准差
1	0.999[a]	0.998	0.997	49.4924

表5 方差分析表

离差平方	平方和	自由度	均方	统计量(F)	相伴概率(P 值)
回归	13 818 876.76	5	2 763 775.35	1 128.33	0.000
残差	24 494.98	10	2 449.49		
总和	13 843 371.75	15			

表6 回归系数分析

	非标准化系数		标准化系数	统计量	相伴概率
	B	Std. Error	Beta	(t)	(P 值)
Constant	450.910	178.079		2.531	0.030
x_1	0.354	0.085	2.446	4.151	0.002
x_2	−0.561	0.125	−2.484	−4.477	0.001
x_3	$-7.254E-03$	0.002	−0.082	−3.509	0.006
x_4	21.578	4.030	−0.030	5.353	0.000
x_5	0.435	0.052	−0.563	8.439	0.000

3 结束语

1)“国民收入”和“消费额”与民航客运量均具有正线性相关关系。这表明近年来我国国民收入的较快增长,乘飞机进行旅游和商务活动的比例就有所提高,这又进一步刺激了经济的发展。

2)“铁路客运量”与民航客运量呈一种线性负相关关系。这一点是显然的。

3)“民航航线里程”与民航客运量也呈一种线性正相关关系。这表明随着我国民航航线的增加,民航客运量也在不断的增加。

4)“入境旅游人数”与民航客运量呈一种线性正相关关系。这表明来华旅游入境人士生活条件基本上都很好,再加上路途遥远,他们就选择了飞机作为主要交通工具。

(下转第91页)

度误差小于 ±0.15 mm,响应时间小于 1.8 s,基本上没有超调,达到了一般工业应用场合的工作要求。

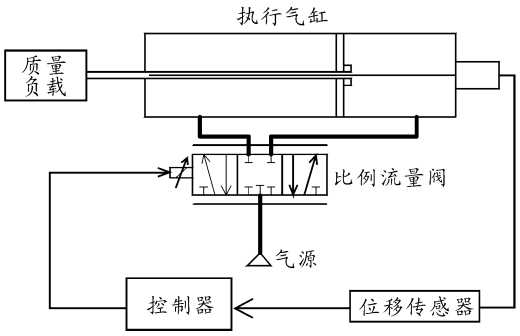

图 6 实验系统组成示意图

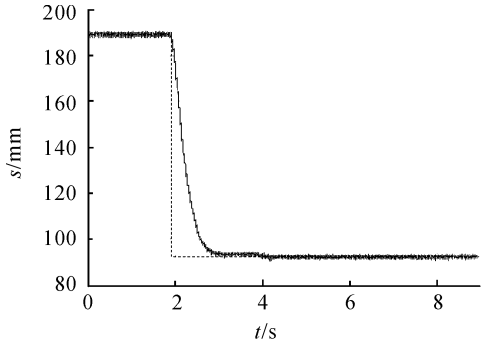

图 7 小幅值方波信号实验曲线

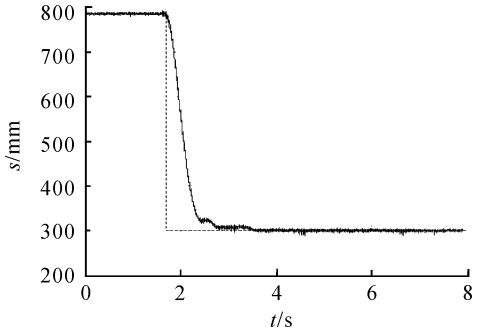

图 8 大幅值方波信号实验曲线

参考文献:

[1] 谢朝夕. 气动伺服定位系统的理论研究与应用[D]. 重庆:重庆大学, 2005.

[2] 闵为. 气动比例伺服系统控制算法及实验研究[D]. 重庆:重庆大学,2006.

[3] 鲍燕伟. 基于 DSP 的气动伺服系统的研究[J]. 液压与气动,2009(12):30 – 33.

[4] 鲍燕伟. 基于 DSP 气动伺服系统的智能模糊 PID 控制[J]. 液压与气动,2010(7):29 – 32

[5] 赵升奇. 气动位置控制系统建模及控制策略研究[D]. 长沙:湖南大学,2003.

[6] 余兵. 模糊控制及其在液压伺服系统中的应用[J]. 液压与气动,2006(10):56 – 64.

(责任编辑 周江川)

(上接第 84 页)

参考文献:

[1] 何晓群,刘文卿. 应用回归分析[M]. 北京:中国人民大学出社版,2001:18 – 19.

[2] 方开泰. 实用多元统计分析[M]. 上海:华东师范大学出版社,1989:87.

[3] 刘润幸. 利用 SPSS 进行多元线性回归分析[J]. 北京:中国公共卫生,2001(8):746 – 748.

[4] 陶勤南. 回归分析与回归设计[J]. 北京农业科学,1984(专集):1 – 76.

[5] 何晓群,刘文卿. 应用回归分析[M]. 北京:中国人民大学出社版,2001:76 – 77.

[6] 周复恭,黄运成. 应用线性回归分析[M]. 北京:中国人民大学出社版,1989:90.

[7] 马小光. 供电系统背景谐波电压辨识的研究[D]. 保定:华北电力大学,2007:24 – 27.

[8] 李伟. 保定地区电力市场需求预测分析研究[D]. 保定:华北电力大学,2003:20 – 30.

(责任编辑 杨继森)