DOI:10.13203/j.whugis1990.03.010

第15卷 第3期 武汉测绘科技大学学报 Vol.15 No.3
1990年 9月 Journal of Wuhan Technical University of Surveying and Mapping Sep. 1990

# 近代回归分析在交通调查分析建模中的应用

杨 仁 王冬根

**摘 要**

本文以出行发生量模型的建立为例，较为系统地讨论了近代回归分析中的自变量选择及回归诊断方法在交通调查分析建模中的应用。分析了在获取调查数据后，如何借助回归自变量选择方法来选择最佳自变量子集，以确定简捷的回归模型；文中应用回归诊断方法，讨论了修正回归模型、探测错误的调查数据的方法，从而为建立简捷、高精度的交通模型打下了基础。最后，作者提出了应用自变量选择及回归诊断方法建立出行发行量模型的一般步骤。

**【关键词】** 交通规划；出行发生量模型；回归分析；回归自变量选择；回归诊断

## 1 引 言

目前，我国许多城市进行了或即将进行居民出行和货物运输的交通调查，通过对调查所获得的数据进行分析建模，可以了解城市交通现状，预测未来交通情况。交通调查数据往往是非常浩翰的，其中还不乏劣质的乃至错误的数据。如果不采用适当的数学方法来对这些数据进行全面的分析，取出反映交通情况本质的数据，就得不到正确的分析结果。多元线性回归分析是交通调查数据建模中应用较广的一种数学方法，目前多采用其中的逐步回归分析方法。但逐步回归不能对数据进行全面的分析，往往被某些不可靠数据左右结果。近代回归分析的发展出现了两个新的分支——回归自变量选择和回归诊断。回归自变量选择是从影响因变量的所有可能自变量中选择对因变量影响最大的最优自变量子集。回归诊断则是依据回归结果来考察所建立的回归模型是否真正反映了原始数据中因变量与自变量的关系，各组数据对回归模型的建立影响如何，进而考察这些影响较大的数据是否是错误的，如果是这样的话，就应剔除。因此，回归自变量选择及回归诊断为全面地分析参加回归计算的所有数据，以求得最佳回归结果提供了可能。为此在下文中作者将就这两种方法在交通调查分析建模中

收稿日期:1989—12—23

的应用进行系统的讨论。

## 2 回归自变量选择方法

自变量选择过程实质上是一个确定回归模型的过程，它的重要性不言而喻。建立出行发生量模型中选择自变量，传统的做法是用每一个可能影响出行发生量的自变量对交通区的出行量作直方图，以便找出影响较大的一些自变量来建立回归模型。这种做法存在很多缺点：首先，直方图不能定量地说明因、自变量之间的关系，只能是两者关系的直观反映；其次，自变量间往往是相关的，两个相关自变量可能都对因变量影响较大，如果将这两个变量都选入模型是不合适的。另外，这种做法对于复杂的交通调查数据是相当繁琐的。

这里将介绍的自变量选择方法，只要从所有与因变量有关的全部可能的自变量子集中，依据一定的准则，选择出一个最优子集。关于自变量的选择准则，实际应用中有以下六种：

（1）残差平方和最小准则（$RMS_q$）

$$RMS_q = \frac{RSS_q}{N-q} \tag{1}$$

（2）预测偏差的方差之和最小准则

$$(N+q)RMS_q = (N+q)\frac{RSS_q}{N-q} \tag{2}$$

（3）平均预测均方差准则

$$S_q = \frac{1}{N-q+1} \cdot \frac{RSS_q}{N-q} \tag{3}$$

（4）$C_p$准则

$$C_p = \frac{RSS_q}{\sigma^2} - (N-2q) \tag{4}$$

（5）预测平方和最小准则

$$PRESS = \sum_{i=1}^{n}\left(\frac{\delta_i}{1-h_{ii}}\right)^2 \tag{5}$$

（6）AIC准则

$$AIC = N\ln(RSS_q) + 2q \tag{6}$$

上式中，$RSS_q$为$q$个自变量（含常数项）进入模型时的残差平方和；$N$为样本总数；$q$为全体自变量中的任一自变量子集所含自变量的个数（含常数项）。

文［3］对上述各准则的出发点、计算公式的推导进行了详细的讨论，这里就不重复。从上可以看出，各准则统计量的计算，基本上都包含了残差平方和和自变量个数$q$。利用某一准则对一实际问题进行自变量选择，先要计算出所有自变量的全部可能子集$q$回归的残差平方和$RSS_q$，再利用该准则计算出相应于子集$q$的统计量值，这些值中最小的子集为该准则下的最优子集。这种选择方法计算量很大，必须借助计算程序来进行，作者利用扫描运算编制了有关程序，可以利用上述各准则从50个自变量中选取出最优子集。

实际应用中是否需要利用所有的准则来进行自变量选择呢？针对这个问题，作者用样本

容量和自变量个数不同的实际例子，对根据不同自变量选择准则得出的最优自变量子集进行了比较，结果发现，各选择准则尽管出发点不同，理论各异，但选择出的最优子集却相近，在此基础上作者还用逐步回归的选择方法对这些例子进行了分析，选择出的结果与上述六种准则有一定的差异。因此实际应用中可酌情选择一至两种准则，进行比较得出最佳结果。另外，作者通过分析、计算，还发现参加回归的强影响点、高杠杆点及异体点等，对自变量的选择有很大影响，如果这样的一些数据是由原始数据中的错误所致，显然会歪曲选择结果，在进行自变量选择前应将它们剔除。

根据前面的讨论，本文提出了进行自变量选择的一般步骤为：

（1）根据所研究问题的专业理论及经验，分析出可能对因变量有关的所有自变量，为自变量的选择提供专业理论依据。

（2）自变量的粗选，分析单个自变量与因变量的关系，剔除那些与因变量线性关系很小，而又无其它任何关系的自变量，同时对相关性较强的自变量进行取舍。

（3）用经过上述粗选后的自变量建立回归全模型，并进行回归诊断，以便确定出自变量、因变量间的正确关系，剔除原始数据错误所致的强影响点、异体点及高杠杆点。

（4）在上述工作的基础上，利用一至二种自变量选择准则选择出最优子集。

## 3 回归诊断方法

回归诊断讨论两方面的问题：（1）检验实际数据是否满足线性最小二乘回归（以下简称LS估计）的GM假定（本文称为LS估计的随机模型），如果不满足，如何进行修正？（2）分析对回归结果影响较大的强影响点、异体点及高杠杆点，探测出错误数据并剔除之。这两项工作都是以拟合残差为基本量，下面先就有关残差的问题进行讨论。

### 3.1 残差的讨论

回归问题

$$Y = X\beta + e \tag{7}$$

的LS估计为：

$$\hat{Y} = X\hat{\beta}, \qquad \hat{\beta} = (X^TX)^{-1}X^TY \tag{8}$$

式中，$X$，$Y$分别为因、自变量矩阵；$\beta$为回归系数矩阵；$e$为因变量值的误差。

定义 $\delta = Y - \hat{Y}$为残差。将（7）、（8）两式代入上式，有：

$$\delta = (I - X(X^TX)^{-1}X^T)Y \tag{9}$$

令 $H = X(X^TX)^{-1}X^T$

$H$称为帽子矩阵，它是各组自变量值在总体自变量空间中位置的度量。$H$阵对角线元素的大小，反映了各自变量值偏离自变量空间的距离。远离自变量空间（即$h_{ii}$较大者）的自变量值被定义为高杠杆点。结合（9）、（7）式，有：

$$\delta = (I - H)e \tag{10}$$

可以看出，残差$\delta$是误差$e$的一种估计。由于$e$是无法知道的，关于$e$的GM假定就可用$\delta$来检验。

### 3.2 正态性及方差齐性的检验与处理

这里要讨论的是用δ来检验实际数据是否满足GM假定。用概率统计学中的正态检验方法可以检验出δ是否服从正态分布，用δ对因变量拟合值作残差图，根据残差图的分析就可来判断δ是否满足方差齐性的条件。当δ较大或分布不合理时，上述的两次检验将不会通过，这时就要对回归模型进行修正。

这里将残差作如下分解：$\delta=\delta_1+\delta_2+\delta_3$。$\delta_1$为回归函数模型不正确时产生的误差；$\delta_2$为随机模型假设不满足时产生的误差；$\delta_3$为正常的随机误差。因此，作者认为回归模型的修正应从下面两方面入手：

（1）函数模型误差的修正。函数模型存在误差，说明因、自变量的关系不正确或是遗漏了重要的自变量，这就需要设法找出它们间的正确关系及遗漏的自变量。常用作偏残差图的方法来修正。

（2）随机模型误差的修正。这是由于随机模型假设不满足即正态性和方差齐性不满足时产生的。此时是在确认函数模型正确的前提下，从参加回归计算的数据入手，来进行修正。如果实际数据无法满足GM假定，就要考虑采用其它估计方法。如果只是由于少量数据引起实际数据不满足GM假定，就应考虑将这些数据检验出来，进行分析，这一问题就是下面将要讨论的离群数据的检验及处理。

### 3.3 离群数据的检验及处理

离群数据也称异常值或异体点，异体点定义为拟合残差明显大于其它点拟合残差的数据组。异体点的产生不外乎两种途径即函数模型误差的影响和其本身在采集过程中的误差。前者由于对实际观测值拟合的函数模型不正确，使部分值拟合残差较大，后者是数据采集中因变量值误差较大所致。基于这两种来源，发展了相应的检验异体点的方法：（1）认为异体点产生于函数模型的误差。设某一回归问题存在$k$个异体点，将回归模型修改为：

$$\begin{cases} Y_i=X_i\beta+\eta_i+e_i & i\in I=(i_1,\ i_2,\ \cdots,\ i_k) \\ Y_i=X_i\beta+e_i & i\notin I \end{cases} \tag{11}$$

式中，$\eta_i$为异体点$i$的异常值。将上式求解，求出$\eta_i$及残差平方和，依据它们可构造检验异体点的$F$统计量、$t$统计量。检验$k$个异体点的$F$统计量：

$$F=\frac{\delta_I^T(I_k-X_IS^{-1}X_I^T)^{-1}\delta_I/k}{\mathrm{RSS}/(n-p-k-1)}\sim F_\alpha(k,n-k-p-1) \tag{12}$$

式中，RSS为$k$个异体点不参加计算时的残差平方和；$\delta_I$，$X_I$为相应于$k$个异体点的残差阵、系数阵。

检验一个异体点的$t$统计量：

$$t_i=\frac{|\delta_i|}{\hat{\sigma}(i)\sqrt{1-h_{ii}}}\sim t_\alpha(n-p-2) \tag{13}$$

式中，$\hat{\sigma}(i)$为$i$点不参加计算时的单位权方差。

（2）认为异体点是由其因变量值有较大的误差而产生的。求解时依据各因变量值误差的大小给各组数据不同的权，进行加权最小二乘回归，经迭代多次后求得的残差较大的点即为异体点。下面以双平方估计法为例说明选权迭代法检验异体点的原理：

双平方估计法的影响函数为：

$$\sum_{i=1}^{n}\varphi(u_i)=0 \tag{14}$$

式中：

$$\begin{cases}\varphi(u_i)=u_i(1-u_i^2)^2 & |u_i|<1.0\\ \varphi(u_i)=0 & |u_i|>1.0\end{cases} \tag{15}$$

$$u_i=\frac{\delta_i}{k\cdot S_n} \tag{16}$$

上式中，$\delta_i$为 $i$ 点残差；$k$ 为调整系数；$S_n$为残差绝对值中位数。

令　$w_i=(1-u_i^2)^2$，定义$w_i$为 $i$ 点的权。结合（14）、（15）两式有：

$$\sum_{i=1}^{n}\varphi(u_i)=\sum_{i=1}^{n}u_iw_i=0 \tag{17}$$

（17）式是基于残差的影响函数，将之转化为基于系数估计值的影响函数，有：

$$\sum_{i=1}^{n}\frac{Y_i-x_i\beta}{k\cdot S_n}\cdot w_i\cdot x_i=0 \tag{18}$$

用矩阵形式表示上式：

$$X^TWX\hat{\beta}-X^TWY=0 \tag{19}$$

$$\hat{\beta}=(X^TWX)^{-1}X^TWY \tag{20}$$

由（20）式可知，双平方估计法实质上是一种带权的LS估计，只不过它的解是通过迭代而得。在迭代过程中，每次迭代的权都要作相应的修正，而残差较大的点的权为零，不参加计算。设第 $k$ 次迭代时，$i$ 点的残差为 $\delta_i^{(k)}$，若 $\delta_i>kS_n$，则第$k+1$ 次迭代时，$i$ 点不参加计算，此时求出的 $i$ 点残差 $\delta_i^{k+1}$为预测残差。若 $\delta_i^{k+1}$仍大于$kS_n$，则下一次迭代时该点仍不参加计算，如果迭代结束时 $i$ 点的残差 $\delta_i^{(n)}$仍大于$kS_n$，则 $i$ 点为异体点。

用上述方法检验出异体点后，仍需对它们进行详细的分析，如果异体点是由原始数据中的错误所致，就应将其剔除。

## 4　实例分析

1986年北京市进行了抽样率为 5 %的居民出行调查。本文收集了与建立出行发生量模型有关的调查资料，将上述方法进行了实际应用。为节省篇幅，下面仅介绍建模过程，不列出有关的计算结果。

### 4.1　自变量的初选

本文选取了包括各小区中不同年龄段的人口数、收入水平的分布、人口密度等23个与交通小区出行发生量有关的自变量。需要指出的是，由于手头资料有限，可能漏选了或重复选取了一些自变量，这里只是为了说明问题。

对全部151个小区的数据进行人工检查，发现60，61，81，141，142，146等六个小区的数据明显有错，并有八个代表方向的小区的数据不宜参加计算。将它们一并剔除，这样样本

容量为137。

通过计算因、自变量的相关系数矩阵，将与因变量相关系数很小(小于0.3) 的两个自变量剔除。另外，各年龄段人口数的七个自变量与总人口数这个自变量存在高度相关，也将这七个自变量剔除。经过初选后剩下14个自变量。

### 4.2 全模型的LS估计及诊断

用LS估计对上述处理后的数据进行回归，求出回归模型，模型拟合度为 0.772 2，可见拟合明显不足。用 $\chi^2$检验法对残差进行正态性检验，结果表明残差服从正态分布。用学生化残差对因变量拟合值作残差图，残差图显示出明显的方差非齐性，这正是模型拟合不足的反映，需进行修正。

### 4.3 函数模型的修正

对每个自变量作偏残差图寻找因、自变量间是否存在其它关系，偏残差图并未显示任何趋势。分析残差较大点的有关资料，发现它们中大多数点对应的交通小区有大型集散点，如体育场、馆、公园或市区中心、商业中心等。这些小区的出行发行量都较大，但产生的原因各异，难以用统一的量化变量来表示。为此，作者设置了一个定性自变量，当其取 1 则表示对应小区有大型集散点，取 0 时表示没有此类点。加入定性自变量后，回归模型拟合度为 0.872 0，较0.772 2有较大提高，说明这种处理是有效的。

### 4.4 异体点的检验与处理

经过上述处理后，仍有一部分点的残差较大，预示着异体点的存在。用双平方估计法，取 $k=3.0$，结果检测出 3，4，6，14等19个点为异体点。分析检测出的这些点的原始数据，发现第42，52，53等点的因变量观测值（即交通区出行发行量值）远远小于交通区人口数，而本次调查人均出行率为1.61次/人日，可以认为这些交通区的出行发行量的统计有误，63，110号点所对应的交通区处市郊，但它们的出行发生量却分别达到209 087人次与169 380人次，远大于各交通区的平均出行发生量73 865人次，说明这两个交通区的数据也有错误。上述分析说明检测出的异体点大都反映了原始数据的错误，说明检验的结果是正确的。

### 4.5 自变量的选择

利用$C_p$、AIC、PRESS三个准则对经上述修正后的模型及样本值进行自变量选择，结果选择出包括前面设置的定性自变量在内的人口密度、自行车拥有数等10个自变量的最优子集。

用LS估计对选出的自变量子集进行回归，得到最后的出行发生量模型。模型拟合度为0.96，较0.77有很大提高。其它各项检验都表明所建立的模型精度很高。

## 5 结 论

根据以上分析，作者认为，应用回归分析技术来建立出行发生量模型，应按以下步骤进行：

（1）分析影响出行发生的因素，为回归自变量的选择提供专业理论依据。

（2）将与因变量关系较小的自变量剔除，同时取舍相关性较强的自变量。

（3）对经上述处理后的变量进行回归，并进行回归诊断，修正模型误差，剔除原始数据中有错误的数据。

（4）利用自变量选择方法，寻找最佳自变量子集，以便提高回归系数的估计精度，降低模型使用费用。

（5）用LS估计方法对上述子集进行回归，得出最后的回归模型。

## 参 考 文 献


[1] 同济大学等. 城市道路与交通. 中国建筑工业出版社，1981.

[2] [澳] John Black. 城市交通规划. 人民交通出版社，1987.

[3] 陈希孺，王松桂. 近代回归分析-原理方法及应用. 安徽教育出版社，1987.

[4] 张启锐. 实用回归分析. 地质出版社，1988.

[5] [美] 塞伯 G A F. 线性回归分析. 科学出版社，1987.

[6] 何 伟. 客运模型系统的研究：[硕士学位论文]. 吉林工业大学，1987.

[7] Veress S A and Huang Youcai. Application of Robust grammetry. PE & RS. 1987(2).

[8] 浙江大学数学系. 概率论与数理统计. 人民教育出版社，1979.


# Studies on the Prediction Model of People Trip Flow in Urban Aera


*Yang Ren* *Wang Donggen*



**Abstract**

This thesis is to study the methods of establishing Trip Production Model by using regression analysis. Some problems on the selection of independent variables and regression diagnostics have been discussed. Two examples are used to compare alls election rules, and the results show consistance. The influences on the results deriving from the over influence points, high leverage points and outliers are studied. The general procedure for the selection of independent variables is shown in thesis, also the applicable methods of polynomial regression diagnostics, especially, the methods for detecting outliers. When dependent variable is affected by some element which couldn't be quantified, we can add the quanlitative variable to improve the results of regression. According to these studies, a method establishing Trip Production Model by using regression analysis has been suggested in this thesis, which was used to set up the trip production model of Beijing, and the result is satisfactory.



**[Key words]** transport planning; trip production model; regression analysis; independent variables selection; regression diagnostics