Matlab 多元线性回归预测城市建成区面积

曹应举[1]，张永彬[1]，温继满[2]，汲　姣[1]，刘佳丽[1]

（1. 华北理工大学 矿业工程学院，河北 唐山 063210；2. 开滦安全技术培训中心，河北 唐山 063000）

摘要：以 2015 年全国 25 个省会城市的建成区面积为研究对象，选择影响城市建成区面积的经济社会影响因子，将最小二乘法与 Matlab 软件相结合，研究及预测了城市建成区面积的影响因素，同时较好地解决了运用传统的纯数学方法的计算量大、计算过程麻烦的问题。研究结果表明：建成区面积主要受生产总值、第二产业、工业、第三产业的影响，即经济指标的影响程度较大。

关键词：城市建成区面积；多元线性回归；Matlab 软件；最小二乘法

中图分类号：TU984.11^{+}3　**文献标志码**：A　**文章编号**：1672－349X(2018)03－0060－06

DOI：10.16160/j.cnki.tsxyxb.2018.03.013

Prediction of Urban Built-up Area by Matlab Multiple Linear Regression

CAO Ying-ju[1], ZHANG Yong-bin[1], WEN Ji-man[2], JI Jiao[1], LIU Jia-li[1]

(1. College of Mining Engineering, North China University of Science and Technology, Tangshan 063210, China; 2. Kailuan Safety Technology Training Center, Tangshan 063000, China)

Abstract: With the built-up areas of the 25 provincial capital cities in 2015 as the research object, the authors of this paper determine the economic and social factors affecting the urban built-up areas, combine Partial Least Square and Matlab, study and predict the influencing factors of the urban built-up area. In this way the problem of overmuch calculation and a troublesome calculation process with a traditional mathematical method can be solved. The research results show that the built-up areas are mainly affected by the gross value, secondary industries, industries, and the tertiary industries, ie, the impact of the economic indicators is greater.

Key Words: urban built-up area; multivariate linear regression; Matlab; Partial Least Square

0　引言

随着经济的发展，城市化速度加快，城市的面积在不断的扩张。为了了解城市的扩张动态，获取城市扩张的历史轨迹，许多学者对城市建成区面积的预测问题进行了研究：李爱民利用遥感技术对不同年份城市建成区多期遥感影像进行提取分类，分析了城市扩张的时空规律[1]；刘柯运用 BP 神经网络方法建立预测模型，使用多期数据作为学习样本和检验样本，对 2005 年北京市城市建成区面积进行了模拟预

作者简介：曹应举(1991－)，男，甘肃定西人，硕士研究生，主要从事 3S 技术与应用研究。

测[2]；雷波通过建立多元回归模型，选用 13 个社会和经济驱动因子，对福州城市建成区面积的扩张驱动进行了回归分析[3]。在模拟预测方面，多元回归模型是一种常用的预测模型，其预测精度高、应用领域广泛，例如，周永生等通过综合影响粮食产量的多种因素，将多元回归分析应用于粮食产量的预测中[4]；付倩娆通过对不同大气成分浓度数据进行分析，将多元回归分析应用在雾霾的预测中[5]；叶锋通过综合考虑经济技术等多种影响因子，将多元回归分析应用于油田产量的预测中[6]；韦浩采用不同方法对滑坡距离进行预测，得出结论是“与传统预测方法相比，多元回归分析法所建立的预测模型精度较高”[7]。但传统的用纯数学方法来实现多元线性回归方程求解的过程比较繁琐，为了简化计算难度，提高计算的速度，本文以 Matlab 为语言平台、以回归分析为数学统计方法，建立关于建成区面积与其影响因子的多元回归模型，并应用其进行城市建成区面积的预测。

1　线性回归模型

变量 Y 的值主要受影响因子 X（由 X_1，X_2，…，X_k 确定，可以表示为 X_1，X_2，…，X_k 的某个函数关系式：$Y=f(X_1,X_2,\cdots,X_k)$）和随机误差 ε 的影响，本研究将自变量写成如下形式：

$$Y=f(X_1,X_2,\cdots,X_k)\varepsilon。\tag{1}$$

其中，随机变量 Y 称为被解释变量或因变量；X_1，X_2，…，X_k 称为解释变量或自变量。$f(X_1,X_2,\cdots,X_k)$ 为一般变量 X_1，X_2，…，X_k 的确定性关系，ε 为随机误差。

当概率模型式(1)中回归函数为线性函数时，即有

$$Y=\beta_0+\beta_2X_1+\cdots\beta_kX_k+\varepsilon。\tag{2}$$

其中 β_0，β_1，β_2，…，β_k 是 $k+1$ 个未知参数，β_0 称为回归常数，β_1，β_2，…，β_k 称为回归系数，Y 称为被解释变量（因变量），而 X_1，X_2，…，X_k 是 k 个可以控制的一般变量，称为解释变量（自变量）。$k=1$ 时，为一元线性回归模型；$k\geqslant 2$ 时，称为多元线性回归模型。

线性回归模型的“线性”是针对未知参数 β_i（$i=0,1,2,\cdots,k$）而言的。对于回归解释变量的线性是非本质的，因为解释变量是非线性的，常可以通过变量的替换把它转化成线性的。

若（X_{i1}，X_{i2}，…，X_{ik}；Y_i），$i=1,2,\cdots,n$ 是式(2)中变量（X_1，X_2，…，X_k；Y）的一组观测值，则线性回归模型可表示为

$$Y_i=\beta_0+\beta_1X_{i1}+\cdots\beta_kX_{ik}+\varepsilon_i。\tag{3}$$

其中 $E(Y_i)=\beta_0+\beta_1X_{i1}+\cdots\beta_kX_{ik}$ 为多元回归方程组，其相应的矩阵表达式为

$$Y=x\beta。\tag{4}$$

$$\beta=\begin{bmatrix}\beta_0\\ \beta_1\\ M\\ \beta_k\end{bmatrix},Y=\begin{bmatrix}Y_0\\ Y_1\\ M\\ Y_k\end{bmatrix},x=\begin{bmatrix}1 & X_{11} & \cdots & X_{1k}\\ 1 & X_{21} & \cdots & X_{2k}\\ \cdots & \cdots & \cdots & \cdots\\ 1 & X_{n1} & \cdots & X_{nk}\end{bmatrix}。$$

其中回归系数的最小二乘估计为

$$\hat{\beta}=(x'x)^{-1}x'Y。\tag{5}$$

2　建成区面积案例分析

通过搜集《中国统计年鉴》(2016)和地方统计年鉴整理获得 25 个省会城市 2015 年的建成区面积。鉴于影响建成区面积有诸多因素，本文在参照已有研究成果的基础上，结合实际情况，构建了城市建成区面积的经济社会影响因子体系（如图 1 所示），并通过建立多元回归模型进行分析和检验，从而对未来的城市建成区面积进行预测。

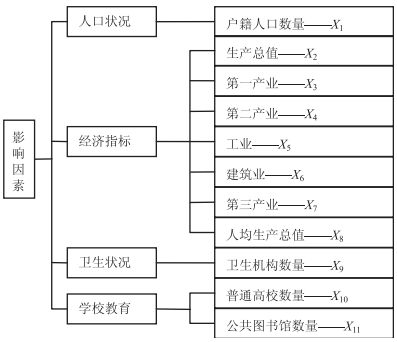

图 1　建成区面积影响因子

3　数据处理

Matlab 多元线性回归数据处理过程包括数据预处理、建模、分析及预测四大部分，其具体步骤如图 2 所示。

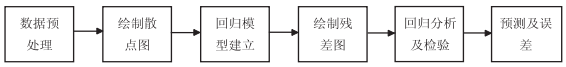

图 2 线性回归流程图

3.1 利用 *regress* 函数实现多元线性回归

在 Matlab 中使用命令 *regress* 实现多元线性回归，调用格式为：$b=regress(Y,X)$，$[b,bint,r,rint,stats]=regress(Y,X,alpha)$[8]，其中：$Y$ 表示一个 $n-1$ 的因变量数据矩阵；X 是 $n-p$ 矩阵，自变量 X 是一列具有相同行数，值是 1 的矩阵的组合；$alpha$ 为显著性水平（缺省时设定为 0.05）；输出向量 b 为回归系数最小二乘估计值；$bint$ 为 b 的置信区间；r，$rint$ 为残差及其置信区间。

stats 是用于检验回归模型的统计量，第一个是 R^2，其中 R 是相关系数；第二个是 F 统计量值；第三个是与统计量 F 对应的概率 P；第四个是 S^2，为误差方差估计值。R^2 越接近 1，说明回归方程越显著；$F>F_{1-alpha}(P,n-p-1)$ 时拒绝 H_0，F 越大，回归方程越显著；与 F 对应的概率 $P<alpha$ 时拒绝 H_0，回归模型成立。

3.2 原始数据归一化

为了方便数据处理，需要对样本中的各影响因子进行归一化处理。归一化是一种消除指标之间量纲影响的简化计算方法，常用的数据归一化方法有“最小－最大标准化”“Z-score 标准化”和“按小数定标标准化”等，本文采用的是“最小－最大标准化”方法对原数据进行线性变换。将 A 中的一个原始值 X 通过“最小－最大标准化”映射成在区间[0,1]中的值 X'，其形式如式(6)所示。

$$X'=\frac{X-X_{\min}}{X_{\max}-X_{\min}}。\tag{6}$$

其中：X' 为归一化后的数据，X 为原始数据，$X_{\min}$ 和 $X_{\max}$ 分别是 X 的最小值和最大值。

3.3 整体最小二乘(TLS)多元回归

通常认为数据矩阵 X 是给定的，不存在误差，但是如果数据矩阵 X 也存在误差或者扰动，那么最小二乘估计从统计观点看就不再是最优的，它将是有偏的，而且偏差的协方差将由于 X 的噪声误差的作用而增加。因此，当 X 也存在误差时，应该使用整体最小二乘进行回归[9]。

在方程 $\underset{n,1}{Y}=\underset{n,mn,1}{X\beta}$ 中，不仅向量 Y 中存在误差 V_Y，而且系数矩阵 X 中也含有误差 V_X，可用 TLS 方法求得参数 $\hat{\beta}$，在 TLS 中，考虑的是矩阵方程的求解。

$$(X+V_X)\hat{\beta}=Y+V_Y,\tag{7}$$

或

$$X\hat{\beta}=\hat{Y},(\hat{X}=X+V_X,\hat{Y}=Y+V_Y)。\tag{8}$$

其中，n 为观测个数，m 为参数个数，通常情况下 $n>m$，矩阵 X 的秩 $r(\underset{n,m}{X})=m<n$。一般也可将式(7)改写为

$$([X\quad Y]+[V_X\quad V_Y])\begin{bmatrix}\hat{\beta}\\-1\end{bmatrix}=0,\tag{9}$$

或等价为

$$(B+D)Z=0。\tag{10}$$

式中：$\underset{n,m+1}{B}=[\underset{n,m}{X},\underset{n,1}{Y}]$ 为增广矩阵，$D=[V_X,V_Y]$ 为误差矩阵，$\underset{m+1,1}{Z}=\begin{bmatrix}\hat{\beta}_{m,1}\\-1\end{bmatrix}$，求解式(10)的总体最小二乘方法可以表示为约束最优化问题：

$$\|D\|_F=\min。\tag{11}$$

其中 $\|D\|_F$ 是 D 的 $F(Frobenius)$ 范数。

求得 $\|D\|_F=\min$ 的问题称为 TLS 问题，若能找到式(7)的一个最小点 $[V_{X_0},V_{Y_0}]$，则任何满足 $(X+V_{X_0})\hat{\beta}=Y+V_{Y_0}$ 的 $\hat{\beta}$ 都称为 TLS 解。

3.4 相关性检验

若想使得所拟直线有实际意义，必须保证建成区面积（变量 Y）与其影响因子（自变量 X）存在线性相关性[10]，描述它们之间相关性系数的定义为：

$$\rho=\frac{\sigma_{XY}}{\sigma_X\sigma_Y},\tag{12}$$

其估值为：

$$\tilde{\rho}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}。\tag{13}$$

当 ρ 越接近 ±1 时，表明随机变量 Y 与 X 的相关性越密切，即所建立的线性模型和实际的

试验情况越接近。

在 Matlab 中使用 *corrcoef* 函数可以求两个序列的相关度，*corrcoef*(*X*,*Y*)表示序列 *X* 和序列 *Y* 的相关系数，得到的结果是一个 2 * 2 矩阵。

相关系数的大小所表示的意义通常如表 1 所示。

表 1　相关系数表示的意义

相关系数	相关程度
0.00～±0.30	微相关
±0.30～±0.50	实相关
±0.50～±0.80	显著相关
±0.80～±1.00	高度相关

依据相关系数表，编写相应程序代码，将相关程度在微相关范围内的影响因子（户籍人口数量 X_1，第一产业 X_3，人均生产总值 X_8，卫生机构数量 X_9，普通高校数量 X_{10}，公共图书馆数量 X_{11}）予以剔除，保留相关系数大于 0.3 的影响因子，剩余 5 个影响因子（生产总值 X_2，第二产业 X_4，工业 X_5，建筑业 X_6，第三产业 X_7）的相关程度显著，对因变量建成区面积 Y 的解释程度较高。

创建城市建成区面积 Y 和 5 个影响因子（生产总值 X_2，第二产业 X_4，工业 X_5，建筑业 X_6，第三产业 X_7）的一元线性回归预测方程，如表 2 所示。

表 2　一元线性回归预测方程

因变量	自变量	拟合方程	判定系数 R^2
建成区面积 Y	生产总值（X_2）	$Y=0.095\ 3+0.711\ 4X_2$	0.648 2
	第二产业（X_4）	$Y=0.118+0.7376X_4$	0.582 2
	工业（X_5）	$Y=0.148\ 7+0.707\ 7X_5$	0.527 1
	建筑业（X_6）	$Y=-0.125\ 5+0.859\ 2X_6$	0.579 6
	第三产业（X_7）	$Y=0.099\ 8+0.877X_7$	0.741 2

根据表 2 的拟合方程结果来看，建成区面积 Y 与各个影响因子 X 的拟合效果良好。对于 R^2 来说数值越大拟合效果越好，各个方程的判定系数最低为 0.527，总体来说判定系数比较高，所以建成区面积与各个影响因子拟合效果总体较好。为了进一步探讨建成区面积与各个影响因子之间的关系，以建成区面积 Y 为因变量，以影响因子 X_2-X_7 为自变量进行多元线性回归分析。

绘制因变量 Y 与自变量 X_2-X_7 之间的散点图，如图 3 所示。

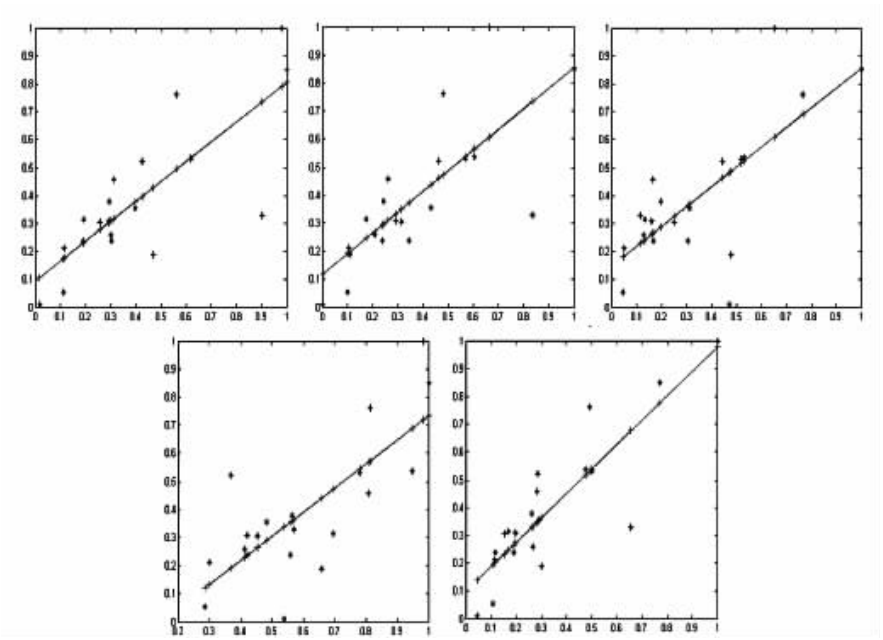

图 3　Y 与 X_2，X_4，X_5，X_6，X_7 的散点图

模型建立过程中绘制的残差图会有异常点而影响模型的正确性，需要对异常点进行剔除，利用 Matlab 编写简单的循环语句可以实现以上操作，剔除完成后绘制残差图如图 4 所示。

剔除残差后利用 *regress* 再次求解参数，结果如表 3 所示。

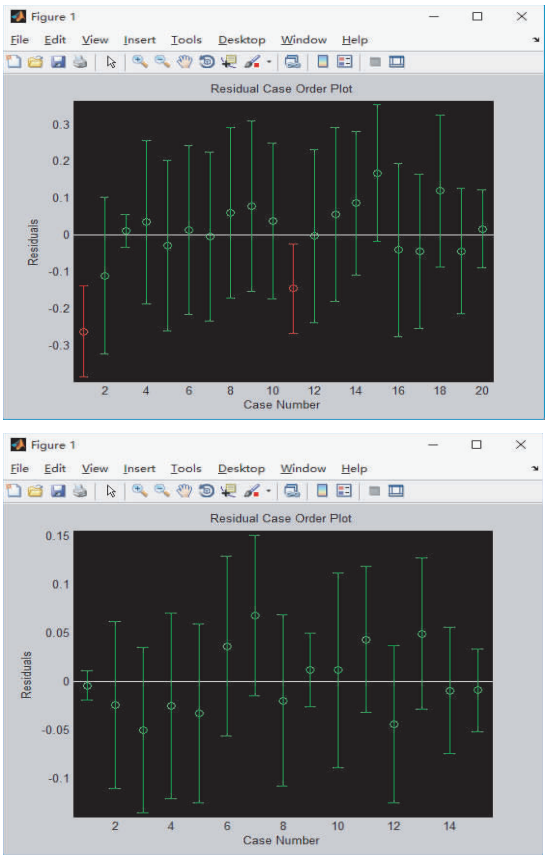

图 4　异常点剔除前后的残差图

表 3　参数求解

b	*bint*		*stats*	*F* 检验值	*t* 检验值
0.095 9	−0.002 9	0.194 8	0.979 9		2.195 7
−1.745 6	−2.256 4	−1.234 7	87.546 5		−7.729 3
0.518 5	0.256 3	0.780 7	0.000	87.546 5	4.473 7
0.388 3	0.224 9	0.551 6	0.002		5.376 8
0.172 6	−0.040 7	0.385 9			1.830 2
1.855 5	1.406 2	2.304 8			9.342 0

t 检验是逐一对参数的显著性进行检验。其原理是$|t(f)|>t_{\alpha/2}$时，接受 H_0，查表可得，$t_{\alpha/2}=2.262$，经检验，*t* 检验统计量小于$t_{\alpha/2}$的常数项以及自变量 X_6 显著性不明显，应予以剔除。*stats* 中第 1 个参数 R^2 是回归平方和与总离差平方和的比值，其值越大越好，该模型达到 0.979 9；第 2 个参数 *f* 统计量，越大越好，本模型为 87.546 5；第 3 个参数为 *P* 的显著性概率，应该小于 0.05，越接近 0 越好，本模型基本为 0；第 4 个参数为估计误差方差，本模型估计误差方差为 0.002，综上，*stats* 中 4 个参数充分说明回归方程显著，该回归模型成立。

因此，最终建立的多元线性回归方程式为：$Y=1.746X_2+0.519X_4+0.388X_5+1.856X_7$。

3.5　结果预测

对城市建成区面积进行回归分析的目的主要是进行预测和控制[11]。未来城市建成区的面积通过单一的往年建成区增减面积来预测难以定性和定量，所以在实际生活中，都是通过近几年的数据构建主要影响因子和建成区面积的方程，通过观测影响因子，来预测未来的城市建成区面积。

Matlab 自带程序 rstool 可以实现回归分析的控制预测功能，在本例中选择 5 个城市的建成区面积留做预测，选出未参与回归处理的 5 组数据，将其自变量影响因子 *X* 输入相应位置，Matlab 将自动计算预测结果，如图 5 所示。

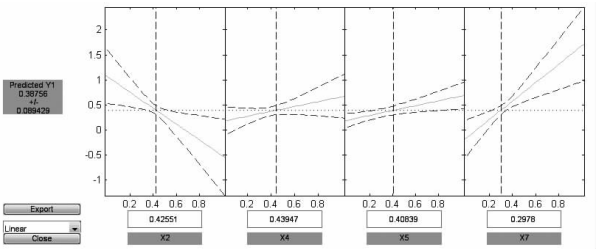

图 5　rstool 预测处理

将 Matlab 计算结果与经过归一化后的真实值进行比对，预测值在限差范围内，说明了回归方程的正确性，结果比对如表 4 所示。

表 4　实际值与预测值比对

实际值	预测区间	正确性
0.104	[0.078,0.246]	正确
0.000	[0.025,0.210]	不正确
0.166	[0.150,0.337]	正确
0.063	[0.045,0.221]	正确
0.285	[0.139,0.310]	正确

结果显示，第一、三、四、五组数据的实际值都在预测区间内，预测的可靠性高，第二组数据预测结果不理想，导致该结果的原因可能是由于各种因素之间产生了更为复杂的相互作用，使得城市建成区面积不再表现为线性关系中那种按比例的规则变化，而代之以不按比例、不规则的变化或突变。例如，城市区划调整使城市建成区面积和城市经济社会指标突然变化，这

必然使传统的基于线性假设的回归预测模型产生很大的误差。

4　小结

城市建成区面积是人口、经济、社会、环境等多因素综合影响的结果，从本研究中发现，2015 年所选 25 个省会城市的建成区面积受第三产业和国民生产总值的影响最大，第二产业和工业的影响次之。说明要想对城市进行合理建设使之健康发展，应该以大力调控第三产业、国民生产总值为出发点。

本文利用建成区面积作为研究变量，运用 Matlab 软件建立多元线性回归模型，利用最小二乘原理求解数学模型中的最优解，不仅简化和优化了繁琐的计算过程，而且通过检验证明所建立的模型计算结果精度较高，对城市建成区面积变化的预测有一定的参考价值。然而，由于在实际情况中城市建成区面积受复杂的多种因素影响，在建模过程中对因素考虑或选择不同，则会造成计算结果的多样性，因此，在研究中应力求完善。

参考文献：

[1] 李爱民. 基于遥感影像的城市建成区扩张与用地规模研究[D]. 郑州：解放军信息工程大学，2009.

[2] 刘柯. 基于主成分分析的 BP 神经网络在城市建成区面积预测中的应用——以北京市为例[J]. 地理科学进展，2007(6)：129－137.

[3] 雷波. BP 神经网络和多元回归模型在城市建成区面积预测中的应用比较——以福州市为例[J]. 城市发展研究，2008(1)：153－155.

[4] 周永生，肖玉欢，黄润生. 基于多元线性回归的广西粮食产量预测[J]. 南方农业学报，2011，42(9)：1165－1167.

[5] 付倩娆. 基于多元线性回归的雾霾预测方法研究[J]. 计算机科学，2016，43(S1)：526－528.

[6] 叶锋. 多元线性回归在经济技术产量预测中的应用[J]. 中外能源，2015，20(2)：45－48.

[7] 韦浩. 多元回归分析法在滑坡空间预测中的应用[D]. 西安：长安大学，2011.

[8] 张智星. MATLAB 程序设计与应用[M]. 北京：清华大学出版社，1993：56－65.

[9] 邱卫宁，陶本藻. 测量数据处理理论与方法[M]. 武汉：武汉大学出版社，2008：109－113.

[10] 刘大杰，陶本藻. 实用测量数据处理方法[M]. 北京：测绘出版社，2000：7－14.

[11] 王乐洋，朱建军. 回归分析、测量平差与大地测量反演[J]. 测绘通报，2007(2)：27－30.

（责任编校：李秀荣）

（上接第 59 页）

参考文献：

[1] 刘晶波，杜修力. 结构动力学[M]. 北京：机械工业出版社，2006：152－178.

[2] 葛楠，苏幼坡，王兴国，等. 竖向刚度对 FPS 滑移摩擦摆系统隔震性能影响研究[J]. 工程抗震与加固改造，2010，32(4)：20－25.

[3] 周锡元，阎维明，杨润林. 建筑结构的震动、减振和振动控制[J]. 建筑结构学报，2002，23(2)：2－11.

（责任编校：李秀荣）