基于偏回归的商品房价格预测模型

倪建威 高雅

（指导教师：张艳萍）

河北工程大学

一、案例背景概述

我国全面实行商品房代替福利分房的政策始于 1998 年，至今已有二十多个年头。房地产市场化改革加速了城市商品房的建设，不仅促进了城市的经济发展，也提高了人民的生活水平。房地产行业是否健康发展影响着国家宏观经济，同时关系着居民的生活，商品房价格越来越受到人们的关注。国家宏观调控已经开始了对房价的调整，房价影响因素、需求预测和风险评价对于房价高低的推断有着十分重要的意义。

2015 年降息 5 次、降准 3 次、营业税 5 改 2、二套房首付比例下调、公积金新政、房贷 6.15%直降至 4.9%，这些强有力措施给房地产市场打了一针强心剂，但这些政策对邯郸房地产市场的影响却不是很明显。从成交套数来看，邯郸市区 2015 年全年商品房住宅共成交 7016 套，其中上半年成交 3742 套，下半年 3274 套，较 2014 年环比下跌 49.80%，从成交面积来看，2015 年邯郸市新房成交面积 829540.78 ㎡，跌幅达 47.56%。2015 年央行降息等多项利好政策、金九银十市场活动等等并没有起到成效，2014 邯郸地产融资风波并未结束，购房者对本地房企期房不信任，邯郸楼市不容乐观。可见，邯郸房地产市场有其特殊性。

目前，国内的商品房市场研究主要分为价格影响因素研究和价格预测研究。目前国内外对房地产的研究主要集中于两个方面：房地产影响因素研究和房地产价格与宏观经济关系的实证研究。在房地产影响因素研究中，周文静 2014 年通过理论分析和实证研究考察货币供给对房地产价格的影响，通过建立计量经济模型对比了广义货币、准货币、人均可支配收入和 GDP 与房价间的关系，最后得出的结论是广义货币对房价的上涨贡献最大[1]；耿源 2014 年用格兰杰因果检验法剖析了北京市经济基本面与北京市房价的关系，通过测算得出了北京市 GDP 和人均可支配收入是北京市房价的格兰杰原因[2]；郑宁和陈立文 2018 年通过建立面板模型分析了货币供给及其他因素对房地产价格的影响，结果发现：货币供

应量和城镇居民可支配收入对房地产价格的正向影响较大，城镇人口数量对房地产价格的正向影响较小[3]。所以，对商品房价格产生影响的主要因素为经济整体因素、房地产供给因素、房地产需求因素。

在商品房价格预测方面，赵泰和迟建英通过建立灰色 GM（1，1）模型对青岛市商品房价格进行了预测分析[4]，王蕾和刘佳杰使用 ARIMA 模型对保定市商品房价格进行了预测分析[5]，这些模型在少样本、贫数据的问题上具有计算简单、便于推广的特点，但是并没有考虑到国家宏观调控对销售价格的影响。乔维德建立基于 BP 神经网络的某城市商品房价格预测模型[6]，取得了很好的效果，但是缺乏对实证研究的理论解释意义。可见，模型的选择要建立在具体数据上，同时对自变量与因变量之间复杂关系进行针对研究。

二、数据来源与指标选取

以邯郸市样本为例，将邯郸市常住人口（X_1），GDP（X_2），城镇居民人均消费性支出（X_3），户籍人口（X_4），房地产开发投资（X_5），城乡储蓄存款（X_6），登记结婚对数（X_7），城镇居民人均可支配收入（X_8），城市平均每人居住面积（X_9）九个因素指标作为自变量，将房地产销售面积（Y）设为房地产需求因变量，如表 1 所示。

表 1　房地产市场需求预测模型中的因变量和自变量

变量	名称	变量	名称
Y	邯郸市房地产销售面积（万平方米）	X_5	房地产开发投资（亿元）
X_1	常住人口（万人）	X_6	城乡储蓄存款（亿元）
X_2	GDP（亿元）	X_7	登记结婚对数（对）
X_3	城镇居民人均消费性支出（元）	X_8	城镇居民人均可支配收入（元）
X_4	户籍人口（万人）	X_9	城镇平均每人居住面积（平方米）

通过邯郸市 2016 年统计年鉴得到 2005-2015 年邯郸市房地产市场需求因变量和自变量相关数据，如表 2 所示。

表 2　2005-2015 年邯郸市房地产市场需求因变量与自变量数据表

年份	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
2005	105.5	865.6	1116.0	6281	871.4	27.75	692	69426	9233	25.7
2006	113.6	871.5	1354.8	7121	884.1	26.91	758.9	64828	10503	26.25
2007	96.4	875.9	1608.1	7587	896.4	47.14	817.6	74161	12583	27.24
2008	152.9	881.6	1906.4	8357	928.1	76.94	1022.0	87284	14457	24.82
2009	195.3	887.9	2015.3	8691	942.8	113.38	1213.4	92730	15961	25.15
2010	275.7	918.8	2361.6	9438	963.5	225.63	1373.1	104010	17562	27.22
2011	380.7	923.9	2789.0	11756	979.9	240.29	1569.1	102540	19322	28.91
2012	431.8	928.6	3024.3	12413	993.1	257.83	1790.8	107451	21740	29.69
2013	355.8	932.5	3061.5	12539	1012.0	305.04	2014.2	113712	20807	26.57
2014	364.5	937.4	3080.0	13048	1029.5	377.09	2251.7	104769	22699	26.15
2015	361.7	943.3	3145.4	14387	1049.7	341.25	2707.2	96572	24630	27.4

三、模型简介

（一）多元线性回归模型

在现实问题中，因变量的变化往往受多个因素影响，就要用多个自变量来解释因变量的改变，这就是多元回归。当多个自变量和因变量之间呈线性相关，对其回归就是多元线性回归。设 $x_1,x_2,\cdots,x_k$ 为自变量，y 为因变量，则回归模型为：

$$y=b_0+b_1x_1+\cdots+b_kx_k+\varepsilon$$

b_0 为常数项，$b_1,b_2,\cdots,b_k$ 为系数，b_1 为 $x_1,x_2,\cdots,x_k$ 固定值时，y 随 x_1 的单位改变量，即 x_1 对 y 的偏回归系数。

为检验方程的显著性，需要进行 F 检验，现提出如下假设：

$$H_0:b_1=b_2=\cdots=b_k=0$$

$$H_1:b_1,b_2,\cdots,b_k \text{不全为 } 0$$

经分析计算，由 F 分布定义，得检验统计量：

$$F=\frac{S_R/k}{S_E/(n-k-1)}\sim F(k,n-k-1)$$

S_R 为离差平方和；S_E 为残差平方和。若 $F\geq F_\alpha(k,n-k-1)$，则拒绝 H_0，

该回归显著。若$F < F_{\alpha}(k, n-k-1)$，则接受H_0，该回归不显著。

（二）偏最小二乘法回归模型

多重共线性是实际研究中时常发生的现象，偏最小二乘法可以更好的提取自变量与因变量之间的关系，即一个在这两个空间对协方差结构建模的隐变量方法。偏最小二乘回归适合当预测矩阵比观测的有更多变量，以及自变量中有多重共线性的时候，通过投影预测变量和观测变量到一个新空间来寻找一个线性回归模型。

考虑 p 个因变量$y_1, y_2, \cdots, y_p$和 m 个自变量$x_1, x_2, \ldots, x_m$的建模问题。偏最小二乘回归的基本做法是首先在自变量集中提出第一个成分t_1（t_1是$x_1, x_2, \ldots, x_m$的线性组合，且尽可能多地提取自变量集中的变异信息）；同时在因变量集中也提取第一成分u_1，并要求t_1和u_1相关程度达到最大。然后建立因变量集$y_1, y_2, \cdots, y_p$与t_1的回归方程，如果回归方程满足要求，则算法终止。否则继续第二对成分的提取，则达到满意精度为止。若最终对自变量集提取 r 个成分$t_1, t_2, \ldots, t_r$，偏最小二乘回归将通过建立因变量与$t_1, t_2, \ldots, t_r$的回归式，然后再表示为因变量与原自变量的回归方程式，即最小二乘法回归方程式。

假设因变量和自变量的 n 次标准化观测数据矩阵分别记为F_0和E_0，即：

$$F_0 = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix}$$

$$E_0 = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

偏最小二乘分析的具体建模步骤如下：

（1）分别提取两变量组的第一对成分，并使之相关性达到最大。

（2）假设从两组变量分别提出第一对t_1和u_1，t_1是自变量集$X = (x_1, x_2, \ldots, x_m)^T$的线性组合：$t_1 = w_{11}x_1 + \ldots + w_{1m}x_m = w_1^T X$，$u_1$是因变量集$u_1 = v_{11}y_1 + \ldots + v_{1p}y_p = v_1^T Y$。为了回归分析的需要，要求：

①t_1和u_1各自尽可能多的提取自变量和因变量信息；

②t_1和u_1的相关程度达到最大。

由两组变量集的标准化观测数据矩阵F_0和E_0，可以计算第一对成分的得分向量，记$\hat{t}_1$和$\hat{u}_1$：

$$\hat{t}_1 = E_0 w_1 = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} w_{11} \\ \vdots \\ w_{1m} \end{bmatrix} = \begin{bmatrix} t_{11} \\ \vdots \\ t_{1m} \end{bmatrix}$$

$$\hat{u}_1 = F_0 v_1 = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix} \begin{bmatrix} v_{11} \\ \vdots \\ v_{1p} \end{bmatrix} = \begin{bmatrix} u_{11} \\ \vdots \\ u_{1p} \end{bmatrix}$$

第一对成分t_1和u_1的协方差$Cov(t_1,u_1)$可用第一对成分的得分向量$\hat{t}_1$和$\hat{u}_1$的内积来计算。其数学表达形式为：

$$\max \ w_1^T E_0^T F_0 v_1$$

$$s.t. \begin{cases} w_1^T w = 1 \\ v_1^T v = 1 \end{cases}$$

利用拉格朗日乘数法将上述问题化为求单位向量w_1和v_1，使得$\theta_1 = w_1^T E_0^T F_0 v_1$取得最大值。问题的求解只需要求得$M = E_0^T F_0 F_0^T E_0$的特征值和特征向量，且$M$的最大特征值为$\theta_1^2$，相应的单位特征向量就是$w_1$，而$v_1$可由$w_1$计算得出$v_1 = \frac{1}{\theta_1} F_0^T E_0 w_1$。

（3）建立$y_1, y_2, \ldots, y_p$和$x_1, x_2, \ldots, x_m$对t_1的多元线性回归方程。

$$\begin{cases} E_0 = \hat{t}_1 \alpha_1^T + E_1 \\ F_0 = \hat{u}_1 \beta_1^T + F_1 \end{cases}$$

其中$\alpha_1 = (\alpha_{11}, \ldots, \alpha_{1m})^T$，$\beta_1 = (\beta_{11}, \ldots, \beta_{1p})^T$是回归方程中的参数。$E_1$和$F_1$是残差阵。$\alpha_1$和$\beta_1$的估计方法为最小二乘法。

（4）用残差阵E_1和F_1替代E_0和F_0重复（1）—（3）步骤，直到满足要求，算法停止。

四、统计建模

（一）建模准备

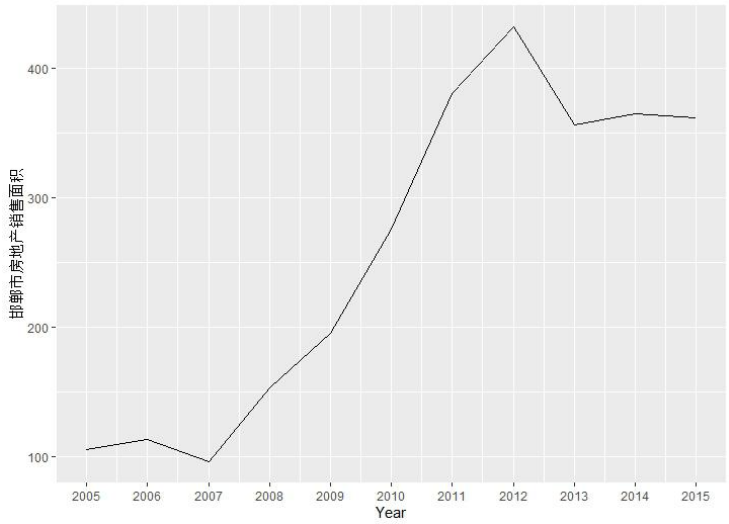

图 1　2005——2015 年邯郸市房地产销售面积折线图

通过图 1 可以发现 2005——2015 年邯郸市房地产销售面积总体呈上涨趋势，特别是 2007——2012 年涨幅最大，达到 335.37 万平方米。从图上可以看出，邯郸市房地产销售面积数据复杂多变，应该使用其他变量信息进行预测。

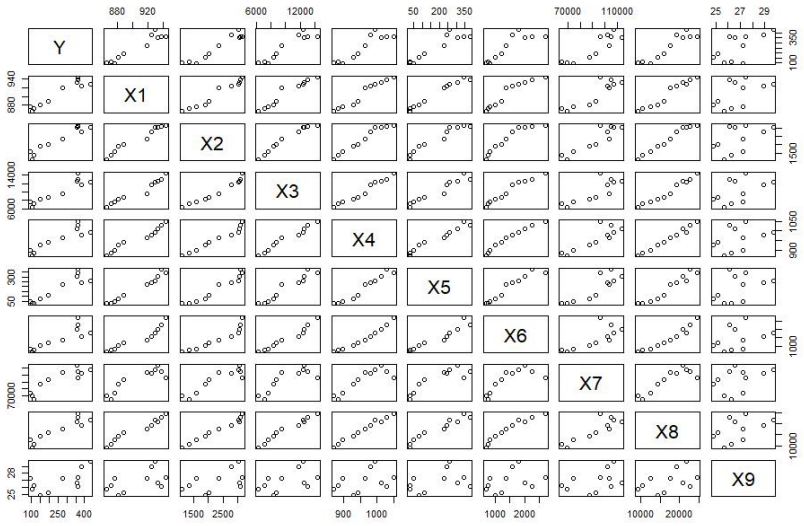

图 2　房地产销售面积与其他变量散点图阵

图 2 展示了房地产销售面积与其他变量之间的关系，从图上可以看出，房地产销售面积与自变量 X_1、X_2、X_3、X_4、X_5、X_6、X_7、X_8 可能存在线性相关关系，随着这些自变量的增长，房地产销售面积也随之增长。为了验证这些相关关系是否显著，使用皮尔逊相关系数进行检验。

表 3　因变量与自变量相关系数矩阵

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	1									
X_1	0.9526	1								
X_2	0.9583	0.9799	1							
X_3	0.9318	0.9695	0.9790	1						
X_4	0.9047	0.9730	0.9766	0.9816	1					
X_5	0.9203	0.9851	0.9628	0.9595	0.9755	1				
X_6	0.8563	0.9433	0.9319	0.9741	0.9809	0.9546	1			
X_7	0.8943	0.8854	0.9128	0.8193	0.8589	0.8653	0.7572	1		
X_8	0.9280	0.9726	0.9858	0.9831	0.9907	0.9616	0.9636	0.8680	1	
X_9	0.6467	0.5296	0.5116	0.4983	0.3792	0.4193	0.3533	0.4114	0.468	1

通过因变量与自变量之间的相关系数矩阵可以发现，因变量与自变量 X_1、X_2、X_3、X_4、X_5、X_6、X_7、X_8 确实存在显著的强线性相关关系 $(p<0.001)$。虽然因变量与 X_9 存在线性相关关系，但是相关系数的显著性较大 $(0.01<p<0.05)$，所以不把该自变量考虑进回归方程中。

（二）模型的建立

将自变量 X_1、X_2、X_3、X_4、X_5、X_6、X_7、X_8 放入回归方程中构建的模型为全变量模型，通过表 4 可以发现全变量模型整体效果良好，$R^2=0.9791$，

调整 $R^2=0.9303$，都非常接近于 1，F 统计量为 $20.07\left(p<0.05\right)$满足模型建立要求。但是每个变量的系数并没有通过显著性检验$\left(p>0.05\right)$，说明模型仍未达到最优。

表 4　全变量模型系数

变量	B	Std. Error	t	P—value
Intercept	1531.00	2677.0000	0.5720	0.6070
X_2	-0.48	0.4290	-1.1170	0.3450
X_3	0.16	0.0921	1.6950	0.1890
X_4	-2.95	3.2010	-0.9220	0.4250
X_5	0.73	0.4570	1.5890	0.2100
X_6	-0.23	0.2839	-0.8000	0.4820
X_7	0.01	0.0044	1.7490	0.1790
X_8	0.03	0.0240	1.3650	0.2660

注：R^2=0.9791，调整R^2=0.9303，F=20.07（p<0. 05）。

所以对模型进一步进行改进，选取相关系数比较大的前三个变量放入模型，同时为了更好描述个变量之间的复杂关系，加入平方项$\left(X_2^2\right)$与交叉项$\left(X_1X_2\right)$。具体模型参数如表 5 所示。

表 5　部分变量回归模型参数

	B	Std. Error	t	P—value
Intercept	-41960.0000	7680.0000	-5.4640	0.0028
X_1	50.0900	9.1130	5.4970	0.0027
X_2	16.2100	3.0000	5.4030	0.0029
X_3	0.0763	0.0254	3.0040	0.0300
X_2^2	0.0007	0.0001	5.8630	0.0021
X_2X_3	-0.0214	0.0040	-5.4070	0.0029

注：R^2=0.9914，调整R^2=0.9828，F=115.4$\left(p<0.05\right)$。

在部分变量建立的回归模型当中，模型的整体效果良好，R^2和调整R^2均达到0.98以上，F统计量为115.4$\left(p<0.05\right)$满足模型要求，同时模型中各个自变量的系数也通过了显著性检验$\left(p<0.05\right)$。优化后的模型由于全变量模型。回归方程数学表达式为：

$$Y=-41960+50.09X_1+16.21X_2+0.0007X_2^2-0.0214X_2X_3$$

表6　部分变量模型中各变量方差膨胀系数值

变量	VIF
X_1	2598.678
X_2	183129.2
X_3	176.6087
X_2^2	5152.453
X_2X_3	312394.6

对优化后的模型进行多重共线性的检验，各变量的方差膨胀系数（VIF）值如表6所示，变量X_2和X_2X_3的VIF值达到了183129.2、312394.6，这说明模型仍未达到最优，模型中各个变量之间存在较强的相关关系。为了克服多重共线性对模型的影响，采取偏最小二乘法对模型进行优化。

表7　偏最小二乘法模型选择

成分	RMSEP	方差解释率(%)
1comps	0.34	98.84
2comps	0.37	99.42
3comps	0.44	99.71
4comps	0.43	100.00
5comps	0.18	100.00

表7为在偏最小二乘法选择模型过程中指标值，综合考虑模型效果达到最优和防止模型过拟合两种情况，选择只包含一个成分的偏最小二乘法模型为最后的模型。模型的数学表达式为：

$$y' = 0.1928 \times x_1' + 0.1939 \times x_2' + 0.1886 \times x_3' + 0.1944 \times x_2^{2\prime} + 0.1942 \times x_2' x_3'$$

（三）模型的应用

表 8　模型预测

年份	商品房销售面积实际值	商品房销售面积预测值	残差	相对误差
2005	105.52	78.6178	26.90218	0.254949
2006	113.57	110.1813	3.388743	0.029838
2007	96.42	139.2465	42.82651	0.444166
2008	152.9	177.2355	24.33552	0.15916
2009	195.29	195.7053	0.415293	0.002127
2010	275.71	263.1177	12.59232	0.045672
2011	380.71	331.3141	49.39588	0.129747
2012	431.79	366.4546	65.33544	0.151313
2013	355.84	375.2010	19.36104	0.054409
2014	364.45	386.2048	21.7548	0.059692
2015	361.65	410.5714	48.92139	0.135273
2016	——	427.2187	——	——
2017	——	434.6562	——	——
2018	——	435.0129	——	——
2019	——	427.6685	——	——
2020	——	412.1714	——	——

根据得到的偏最小二乘回归模型式 2016-2020 年邯郸市商品房销售面积，如表 8 所示。从表 8 的预测结果分析可知，未来几年内邯郸市商品房销售面积呈现比较稳定的上升趋势，2016 年、2017 年和 2018 年商品房销售面积分别为 427.2187、434.6562、435.0129；2019 年和 2020 年商品房销售面积分别为 427.6685、412.1714，与前几年相比商品房销售面积稍微有些回落。

五、结论和建议

本文通过因变量商品房销售面积与九个自变量：常住人口，GDP，城镇居民人均消费性支出，户籍人口，房地产开发投资，城乡储蓄存款，登记结婚对数，城镇居民人均可支配收入，城市平均每人居住面积的 Pearson 相关系数矩阵，找

到影响商品房销售面积最大的三个主要因素：常住人口，GDP，城镇居民人均消费性支出，建立了多元回归模型。由于在社会经济中的时间序列数据具有较强的相关性，在进行回归时很可能出现多重共线性问题，通过共线性分析得到预测变量之间有很强的共线性，因此本文采用偏最小二乘方法消除预测变量之间的共线性，得到更稳定的回归系数估计值，根据建立的偏最小二乘模型预测模型预测出2016-2020年邯郸市商品房销售面积。

参考文献

[1] 周文静. 货币供应量对我国房地产价格的影响研究[D].云南财经大学,2014.

[2] 耿源. 北京市房价波动对城镇居民消费影响研究[D].首都经济贸易大学,2015.

[2] 郑宁,陈立文.对房地产价格影响的货币因素研究[J].价格理论与实践,2018(05):55-58.

[4] 赵泰,迟建英.灰色GM(1,1)模型在商品房销售价格预测中的应用[J].价值工程,2019,38(23):76-78.

[5] 王蕾,刘佳杰.基于ARIMA模型的保定市商品房价格预测研究[J].产业与科技论坛,2019,18(09):96-98.

[6] 乔维德.基于BP神经网络模型的商品房价格预测研究[J].常州工程职业技术学院高职研究,2020(01):35-42.

[7] 韩翔. 湖南省商品房价格分析及预测[D].安徽理工大学,2019.

[8] 戚明远. 网络搜索数据与商品住宅市场的相关性研究[D].华南理工大学,2019.

[9] 崔庆岳,赵国瑞.基于灰色GM(1,1)模型的商品房销售价格预测[J].哈尔滨商业大学学报(自然科学版),2019,35(02):253-256.

[10] 孙守瑄,吴言,潘亚诚,张红伟.基于灰色系统GM(2,1)模型的商品房价格分析及预测——以海南省主要城市为例[J].电脑知识与技术,2019,15(06):191-192+197.

附录：本案例所使用的 R 软件程序命令(部分)

```
setwd("G:/2020 其他事件薄/案例库建设/基于岭回归的商品房预测模型")
DATA=read.csv("DATA02.csv")
library(openxlsx)
library(Hmisc)
library(ggplot2)
DATA_2=DATA[,-c(1,11)]
model_001=lm(Y~.,DATA_2)
summary(model_001)
model_002=step(model_001,direction = 'both')
pairs(DATA[,-1])
ggplot_1=ggplot(data = DATA,aes(x=DATA$年份,y=DATA$邯郸市房地产销售面积))+geom_line()+
            scale_x_continuous('Year',breaks = seq(2005,2015,1))+
             scale_y_continuous('邯郸市房地产销售面积',breaks = seq(100,500,100))
ggplot_1
ggplot(DATA,aes(x=DATA$常住人口,y=DATA$邯郸市房地产销售面积))+geom_point(shape=19,size=2.5)+
   scale_x_continuous('常住人口')+scale_y_continuous('邯郸市房地产销售面积')
ggplot(DATA,aes(x=DATA$城镇居民人均消费支出,y=DATA$邯郸市房地产销售面积))+geom_point(shape=19,size=2.5)+
   scale_x_continuous('城镇居民人均消费支出')+scale_y_continuous('邯郸市房地产销售面积')
ggplot(DATA,aes(x=DATA$户籍人口,y=DATA$邯郸市房地产销售面积))+geom_point(shape=19,size=2.5)+
   scale_x_continuous('户籍人口')+scale_y_continuous('邯郸市房地产销售面积')
ggplot(DATA,aes(x=DATA$房地产开发投资,y=DATA$邯郸市房地产销售面积))+geom_point(shape=19,size=2.5)+
```

```
  scale_x_continuous('房地产开发投资')+scale_y_continuous('邯郸市房地产销售面积')
ggplot(DATA,aes(x=DATA$城乡储蓄贷款,y=DATA$邯郸市房地产销售面积))+geom_point(shape=19,size=2.5)+
  scale_x_continuous('城乡储蓄贷款')+scale_y_continuous('邯郸市房地产销售面积')
ggplot(DATA,aes(x=DATA$登记结婚对数,y=DATA$邯郸市房地产销售面积))+geom_point(shape=19,size=2.5)+
  scale_x_continuous('登记结婚对数')+scale_y_continuous('邯郸市房地产销售面积')
ggplot(DATA,aes(x=DATA$城镇居民人均可支配收入,y=DATA$邯郸市房地产销售面积))+geom_point(shape=19,size=2.5)+
  scale_x_continuous('城镇居民人均可支配收入')+scale_y_continuous('邯郸市房地产销售面积')
ggplot(DATA,aes(x=DATA$城市平均每人居住面积,y=DATA$邯郸市房地产销售面积))+geom_point(shape=19,size=2.5)+
  scale_x_continuous('城市平均每人居住面积')+scale_y_continuous('邯郸市房地产销售面积')
list_1=rcorr(as.matrix(DATA),type = 'pearson')
list_2=NULL
list_2[[1]]=list_1$r
list_2[[2]]=list_1$n
list_2[[3]]=list_1$P
write.xlsx(list_2,'result_1.xlsx')
colnames(DATA)=c('Year','Y','X1','X2','X3','X4','X5','X6','X7','X8','X9')
X2_2=DATA$X2^2
X1_X2=DATA$X1*DATA$X2
data_1=cbind(DATA,X2_2,X1_X2)
model_1=lm(data=data_1,Y~X1+X2+X3+X2_2+X1_X2)
model_1=lm(data=data_1,Y~X1+X2+X3+X2_2)
```

```
summary(model_1)
library(car)
var_02=vif(model_1)
model_1_step<-step(model_1,direction="both")
data_2=cbind(data_1[,c(2,3,4,5,12,13)])
data_mean=apply(data_2,2,mean)
data_std=apply(data_2,2,sd)
data_3=matrix(NA,11,6)
for (i in 1:6) {
   data_3[,i]=(data_2[,i]-data_mean[i])/data_std[i]
}
data_3=as.data.frame(data_3)
colnames(data_3)=colnames(data_2)
list_21=rcorr(as.matrix(data_3),type = 'pearson')
x.pr=princomp(x,cor=TRUE)
x.tezhengzhi=eigen(cor(x))
library(psych)
fa.parallel(list_21$r,fa="PC",n.iter=2,show.legend=FALSE,main="Screen plot with
parallel analysis")
KMO(list_21$r)########非正定
step(model_1,direction="both")
library(pls)
pls1<-plsr(Y~.,data=data_3,validation="LOO",jackknife=TRUE,method="widekernel
pls")
summary(pls1,what="all")
pls2<-plsr(Y~.,data=data_3,ncomp=1,validation="LOO",jackknife=TRUE)
coef(pls2)
summary(pls2)
```