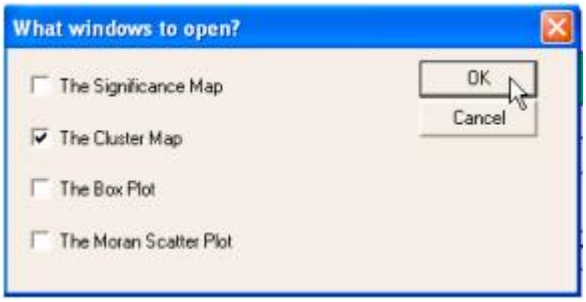


图 20.9 LISA 结果窗口中 cluster 地图选项

右击出现 Options 对话框，设置 Randomization 为 9999。运行几次序列，直到所示的模式稳定到图 20.10 右侧面板中的数字。注意：与原始比率聚集地图稍有不同，在图 20.10 左边面板中重新产生。

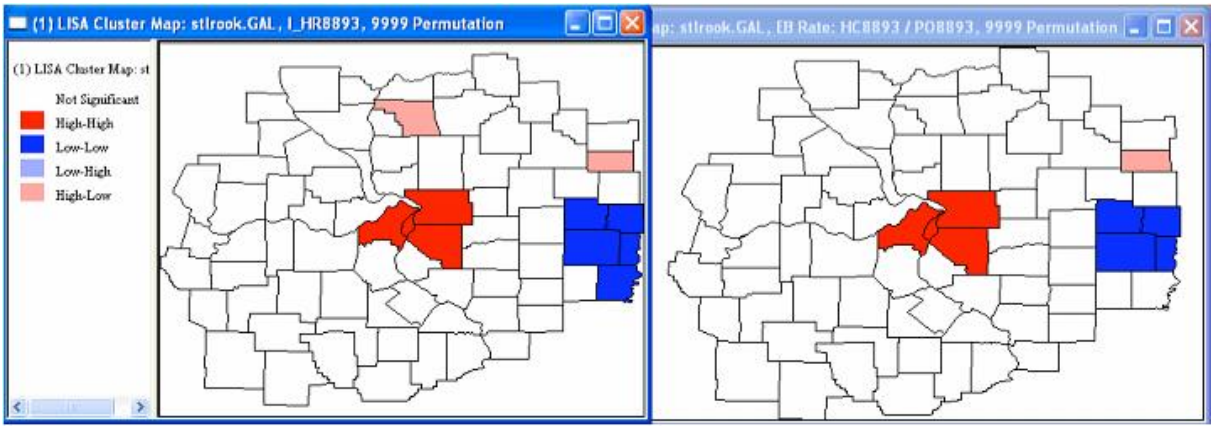


图 20.10 原始比率和 EB 调整后比率的 LISA 聚集地图

注意空间离群值 Morgan 县(在左边地图中点击该县，在表中找到它，位于第 9 行)。在地图中选择它的邻居，如图 20.11(用 shift-click 来添加选择)，在表格中激活被选中的县。考虑表格中 HR8893，HC8893 和 PO8893 的值，如图 20.12。注意 Morgan 县有相当高的杀人案发生率(4.581251)，但在组中并不是最高的(Sangamon 县为 6.029489，IL-FIPS 17167)。但更重要的是，组中最低的两个值，杀人案发生率为 0 的县，人口数据也是最少的(分别是 33911 和 35051)。EB 标准化将会拉高杀人案发生率，减小与 Morgan 县的差异，清除一个空间离群值的建议。

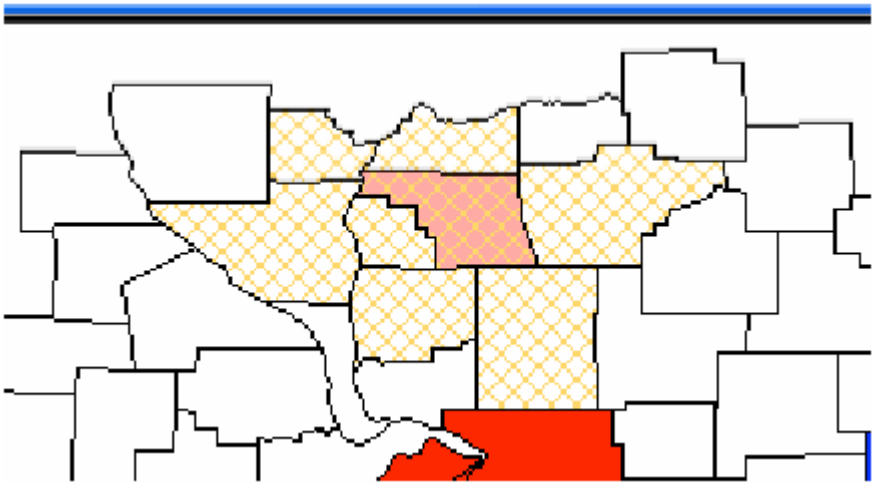


图 20.11 LISA 比率地图敏感性分析：邻居

20.5 练习

用练习 18 和 19 中的相同的数据集和比率变量来估计你对比率方差不稳定性的推断的敏感性。

练习 21　双变量空间自相关

21.1 目标

本练习关注 Moran 散点图和 LISA 地图在双变量空间自相关方面的扩展，其中时-空相关是一种特例。有关方法细节请见 Anselin 等(2002)。

在本练习末，你将知道：

- 创建和解释一幅双变量地图
- 建立一幅 Moran 散点图矩阵
- 解释不同形式的时-空联系
- 创建和解释双变量 LISA 地图

有关本操作的详细信息请参考 User's Guide 第 94-96 页。

21.2 双变量 Moran 散点图

载入数据集 OZ9799 文件(ozthies.shp，关键字为 STATION)，是包含了 30 个 Los Angles 空气质量监测点建立的 Thiessen 多边形文件。底图将如图 21.1 所示。如果你还没有 Thiessen 多边形，按照 6.3 部分的指导创建需要的文件。

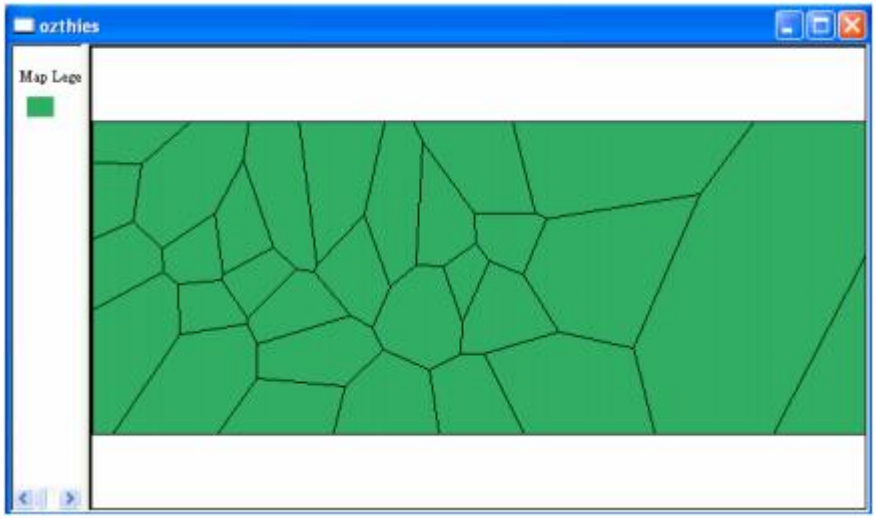


图 21.1 Los Angeles 监测点的 Thiessen 多边形底图

你还需要一个 Thiessen 多边形的 rook 邻接权重文件(ozrook.GAL)。如果需要的话，再创建一个权重文件(详细的指导见 15.2 部分)。从菜单 Space>Multivariate Moran 调用双变量散点图功能(图 21.2)，或点击相应的工具条按钮。出现一个变量设置对话框，如图 21.3 所示。注意有两栏变量。左边的一栏(Y)为空间滞后变量，右边的一栏(X)为非滞后变量。指定 A987(1998 年 7 月平均每 8 小时的 ozone 数据)为滞后变量，A988(1998 年 8 月平均 8 小时的 ozone 数据)为 X 变量，如图 21.3，点击 OK 将出现权重选择对话框，如图 21.4 所示。

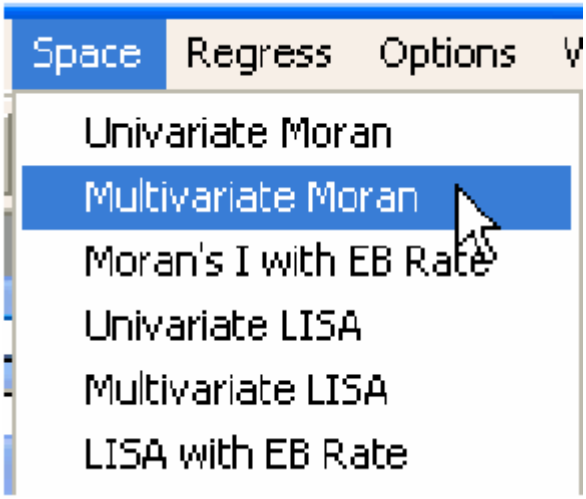


图 21.2 双变量 Moran 散点图功能

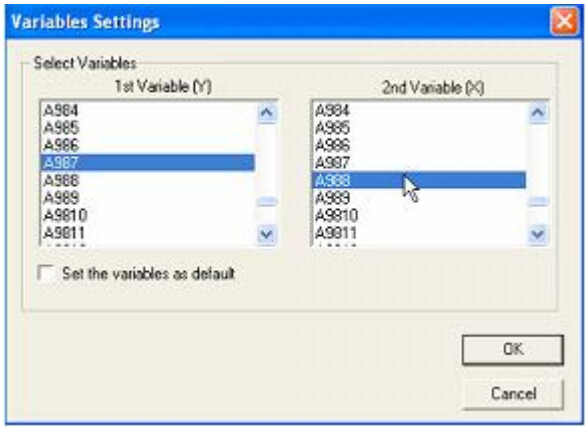


图 21.3 双变量 Moran 散点图变量选择

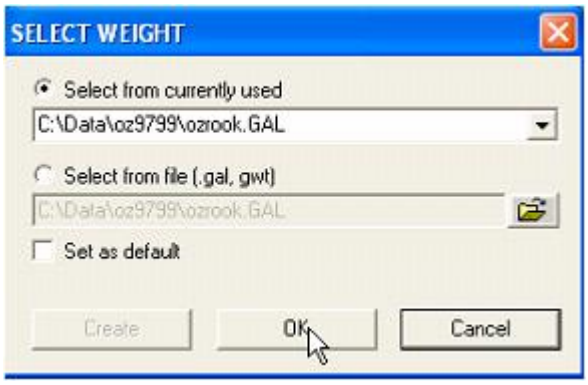


图 21.4 双变量 Moran 散点图空间权重文件选择

选择 ozrook.GAL 作为权重文件，点击 OK 产生双变量 Moran 散点图，如图 21.5。注意：这幅图的合理解释是关于产生的 Moran's I 用所有相邻位置的平均值评估一个位置的 X 变量(A988)值与其它变量(A987)相关程度。作为一幅散点图，所有的标准选项都被执行，如随机化(randomization)，随机化信封(envelopes)及保存中间过程变量。同样，作为一幅散点图，可以设置 Exclude Selected 选项来方便链接和刷光(见练习 18，关于这些标准选项的详细谈论)。

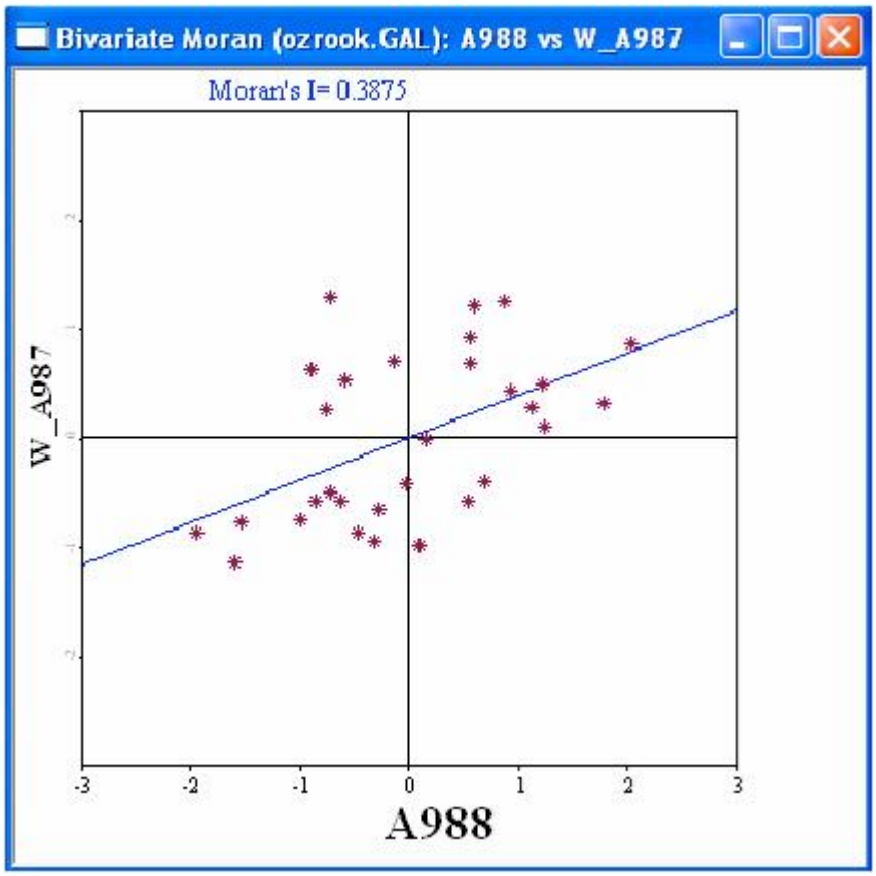


图 21.5 双变量 Moran 散点图：988 与 987 相邻的 ozone

21.2.1 时-空相关性

图 21.5 所用的例子是关于在两个不同时间点的同一变量(第 8 小时的 ozone 观测值)。时-空相关性被认为是普通双变量空间关系的一种特例。然而，图 21.5 所绘的关系并不是唯一的解释。将 X 变量和空间滞后变量相互转换也是合理的。

调用双变量 Moran 散点图功能，A988 为 Y 变量(空间滞后)，A987 为 X 变量。结果将如图 21.6。这幅图描绘了一个位置 1998 年 7 月的 ozone 与 1998 年 8 月邻居的平均值的相关性。你可以认为图 21.5 中的双变量相关性是关于向心扩散(从现在的相邻位置到未来的核心位置)，而图 21.6 所示的双变量相关性是向外扩散(从现在的核心到未来的相邻位置)。每一个都是时-空相关性的不同视角。

时-空相关性可以分解为一个纯空间自相关，如图 21.7 所示在的 Moran 散点图，和一个纯时间序列的相关性，如图 21.8 中的相关性图。

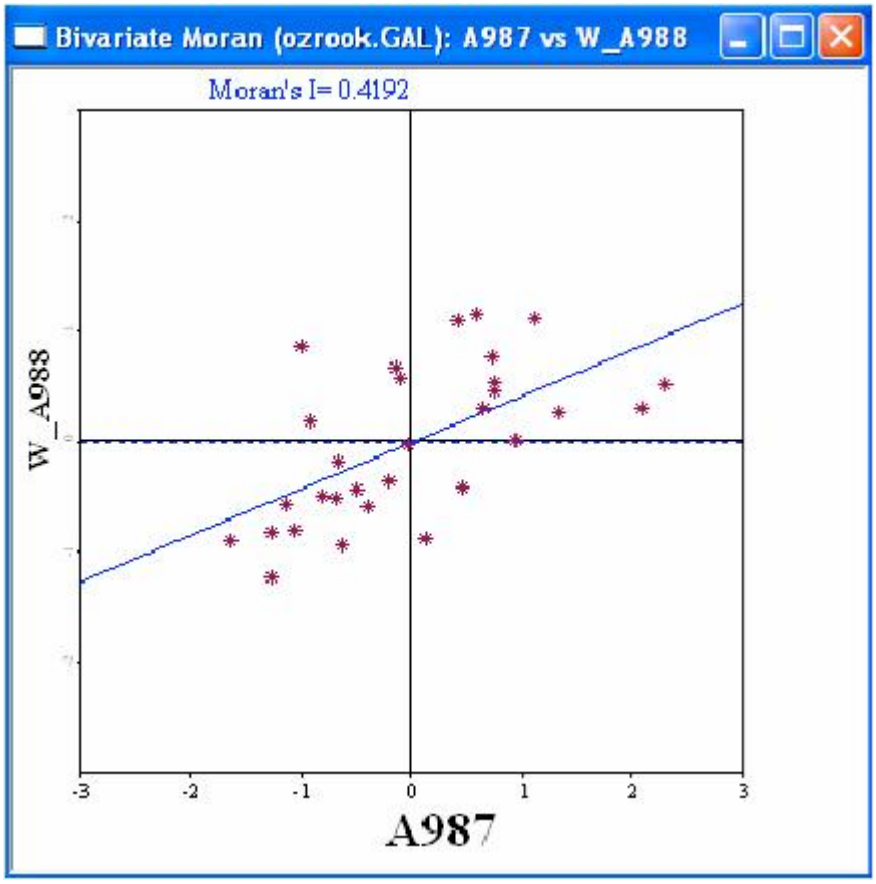


图 21.6 双变量 Moran 散点图：987 与 988 相邻的 ozone

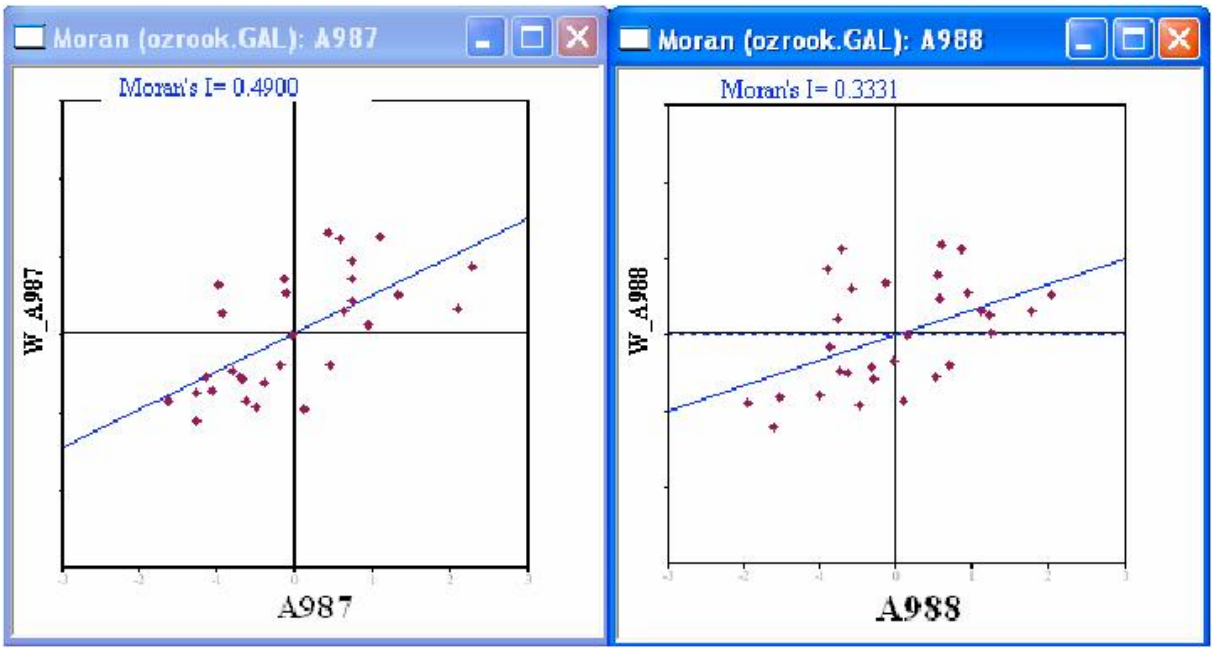


图 21.7 987 和 988 中 ozone 的空间自相关

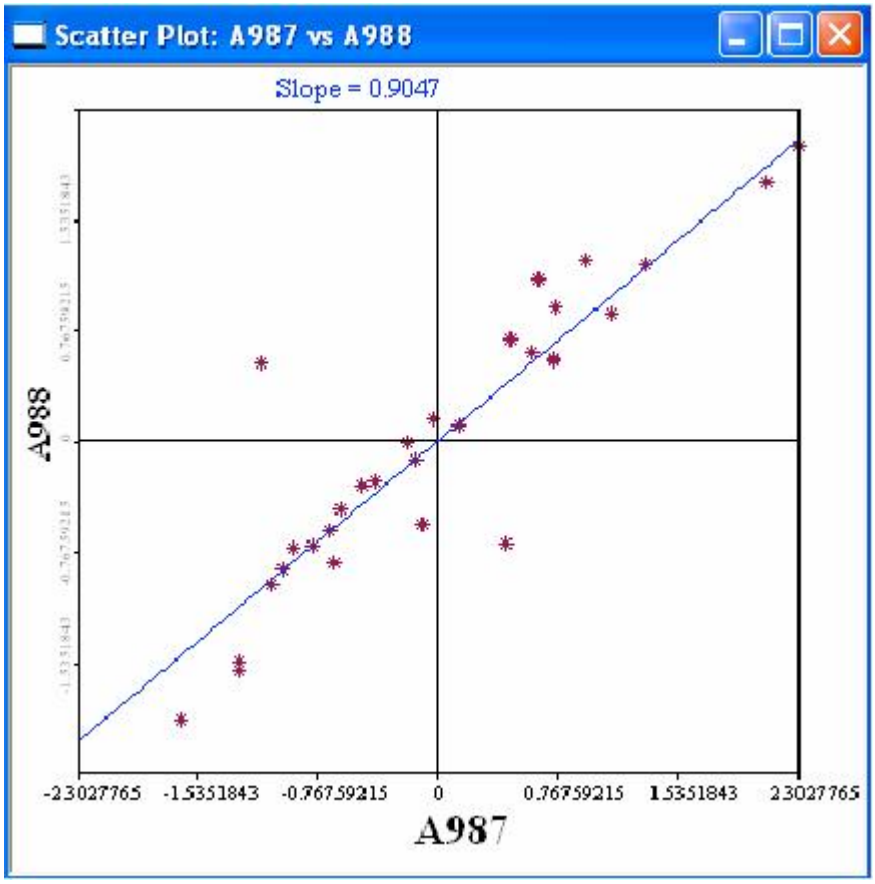


图 21.8 987 和 988 之间的 ozone 相关性

作为与 Moran's I 相似的可替代物，是利用时-空回归。因为空间滞后和原始变量是有关不同时期的，通过后者解释前者是合理的。但在纯横截面(cross-sectional)环境(context)中，这是无效的，是因为空间滞后的内生性(同时还有方程的偏心(bias))，但在时-空设置中，没有这一问题。

例如为 A987 创建空间滞后，利用练习 17 所述的方法。然后创建一幅以 A988 为 Y 变量，A987 空间滞后为 X 变量的散点图。结果将会如图 21.9。此图中的斜率(0.5749)是 ozone 变量的时-空回归系数。它可以与其它两个时-空 Moran's I 系数相比较。另外，它可以用通常的方式刷光或链接，对特别的高值的敏感性进行评估。

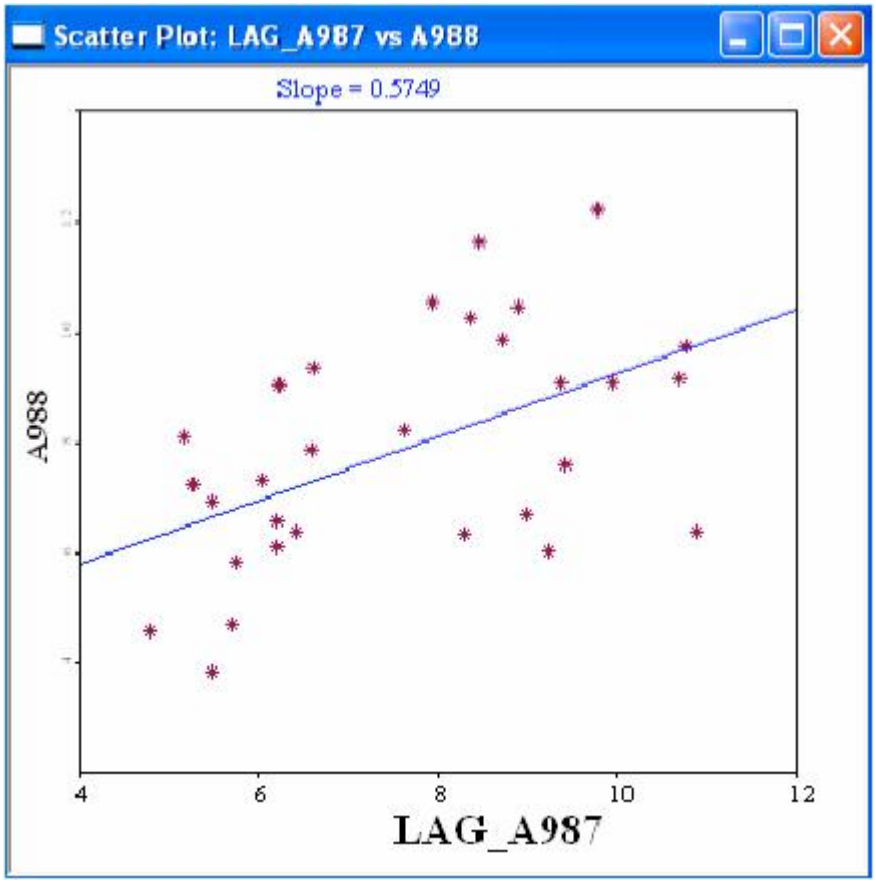


图 21.9 988 与 987 的 ozone 的时-空回归

21.3 Moran 散点图矩阵

由横截面(cross-sectional)的空间自相关提供的不同视角的结合与时-空相关系数，可以生成散点图矩阵(见 9.2 部分)。所谓的 Moran 散点图矩阵由主对角线(mail diagonal)上的 Moran 散点图和非对角线位置上的时-空图组成。

利用这两个刚建立的时-空散点图和横截面图，按排在一个矩阵中，如图 21.10。确保在每一幅图中 Exclude Selected 选项是处于 ON 状态(一次只能完成一个)。然后用刷光来探索空间和时-空相关性之间的不同标准(measures)的联系，验证有影响的观测值(位置)。

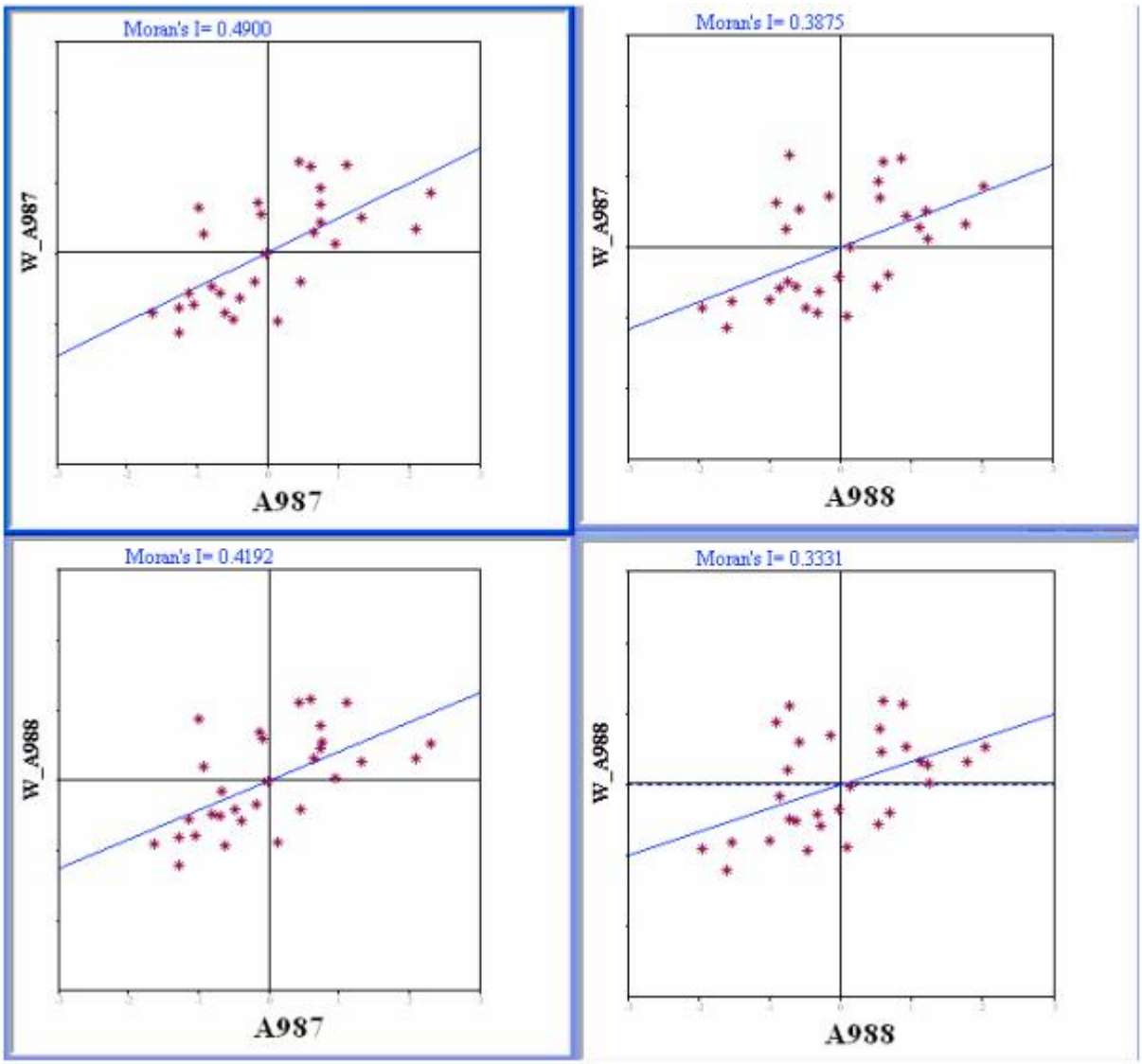


图 21.10 987 和 988 中 ozone 的 Moran 散点图

21.4 双变量 LISA 地图

双变量 LISA 是对两种又是的 LISA 功能的直接扩展，一个是位置，另一个是邻居平均值。从菜单调用这一功能 Space>Multivariate LISA，如图 21.11，或点击相应的工具条按钮。会出现与前面相同的变量设置对话框。用图 21.3 中所述的相同方法，选择 A987 为 Y 变量，A988 为 X 变量。注意：Y 变量是具有空间滞后的变量(邻居平均值)。如前，选择 ozrook.GAL 为空间权重文件(见图 21.4)。

在结果窗口对话框，选中 Cluster Map，如图 21.12。点击 OK 产生双变量 LISA 聚集地图，如图 21.13 所示。注意：这表示出一个位置上 ozone 在 1998 年 8 月和 1998 年 7 月其邻居平均值之间的空间相关性的局部模式。在变量设置对话框交换 Y 和 X 变量的选择，会创建一幅某一位置的 ozone 在 1998 年 7 月和 1998 年 8 月其邻居平均值之间相关性的 LISA 地图。与时-空散点图的解释是相似。将这两个双变量 LISA 地图与它们的横截面相对物(counterparts)相比较。

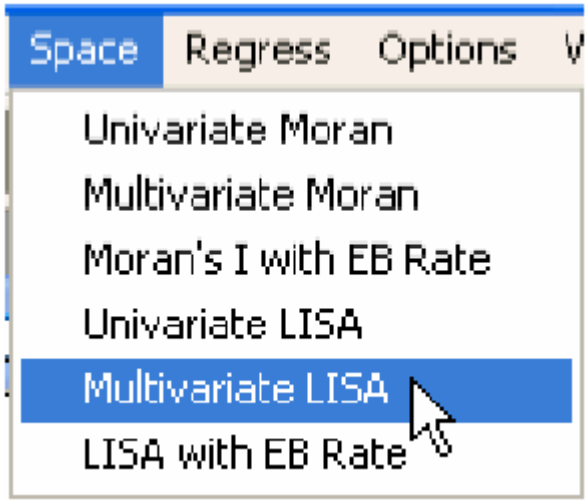


图 21.11 双变量 LISA 功能

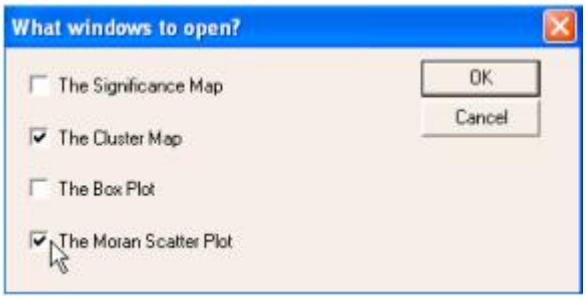


图 21.22 双变量 LISA 结果窗口选项

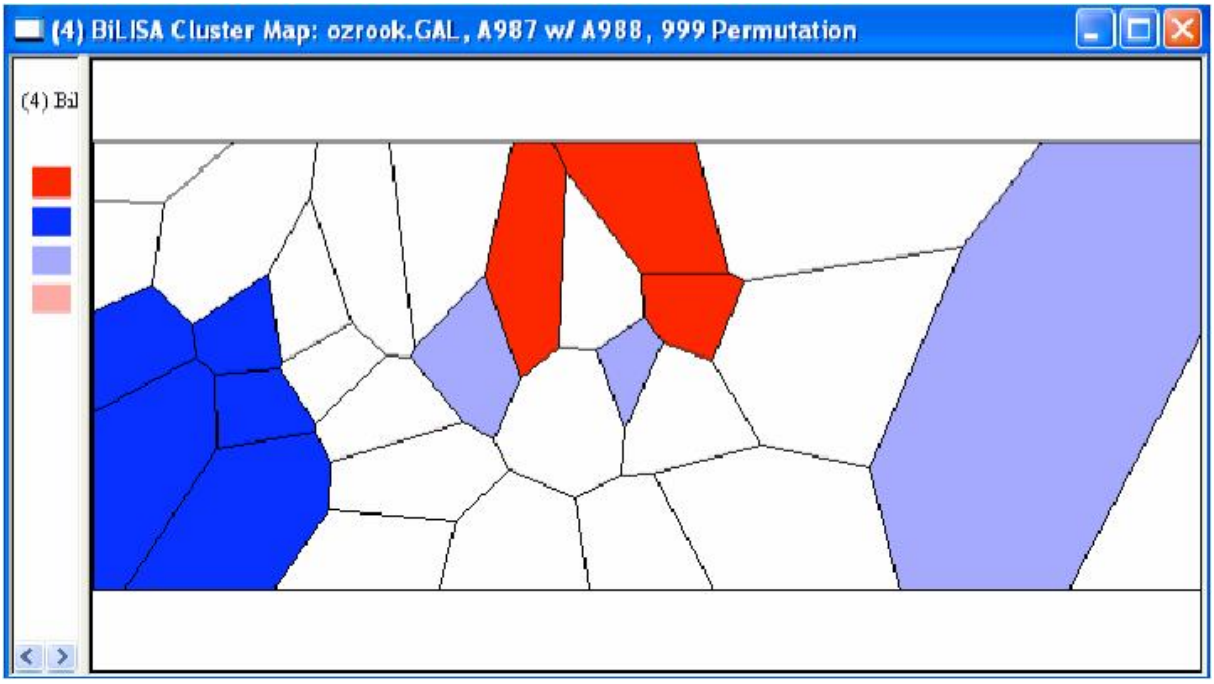


图 21.13 双变量 LISA 聚集地图：987 与 988 相邻的 ozone

21.5 练习

一些样本数据集包含了在不同位置适时(in time)观测的一些变量。除了 Los Angeles 的 ozone

数据集，还包括 St Louis 杀人案数据(stl_hom，关键字为 FIPSNO)，SIDSO 数据(ohlung.shp，关键字为 FIPSNO)。或者，分析不包含时间维的两个变量。

练习 22　回归基础

22.1 目标

该练习开始 GeoDa 中空间回归功能的综述，以一些基本概念为开始。多变量回归分析的方法背景可以参考许多计量经济学的文章，这里不再涉及。回归诊断及具体的空间模型留在练习 23 和 25 中论述。

在本练习末，你将知道：

- 建立线性回归模型说明
- 运行普通最小二乘估计(OLS)
- 保存 OLS 结果
- 添加 OLS 预测值和残差到数据表格中
- 创建带有预测值和残差的地图

有关本操作的详细信息请参考 Release Notes 第 45-56 页。

22.2 准备工作

在本练习中，你将使用 columbus.shp 样本数据集(关键字为 POLYID)中包含的经典 Columbus 相邻犯罪数据(Anselin 1998，第 188-190 页)。底图将如图 22.1 所示。

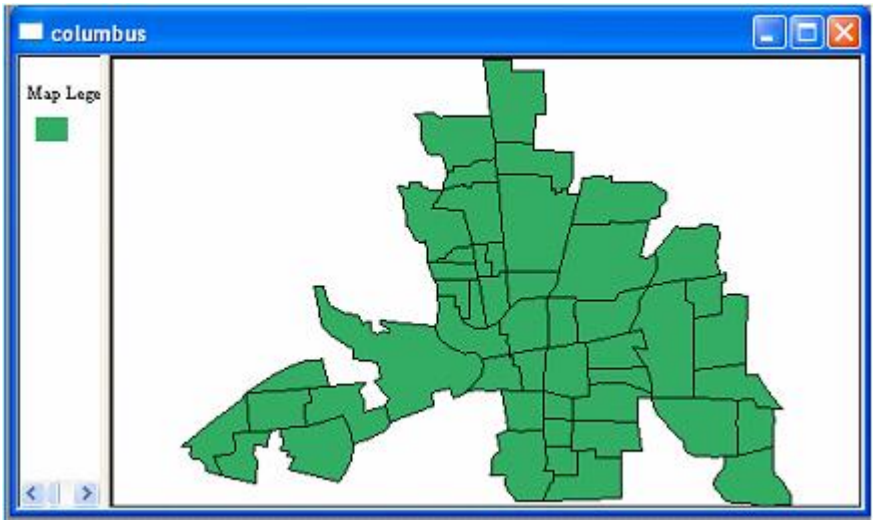


图 22.1 Columbus 相邻犯罪底图

在 GeoDa 中，可以在没有打开一个项目，而调用回归功能。这在分析大的数据集的时候非常有用，它可以链接和刷光数据表的过度负担。从 GeoDa 初始窗口开始回归功能(图 1.1)，选择 Methods>Regress，如图 22.2。或者，当项目被打开(载入一个文件后)，直接从主菜单条调用 Regress 命令，如图 22.3。出现默认的回归标题和输出对话框，如图 22.4 所示。

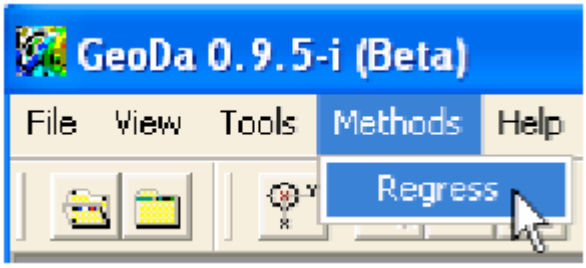


图 22.2 项目外回归

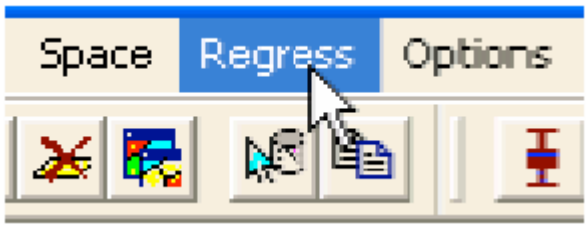


图 22.3 项目中的回归

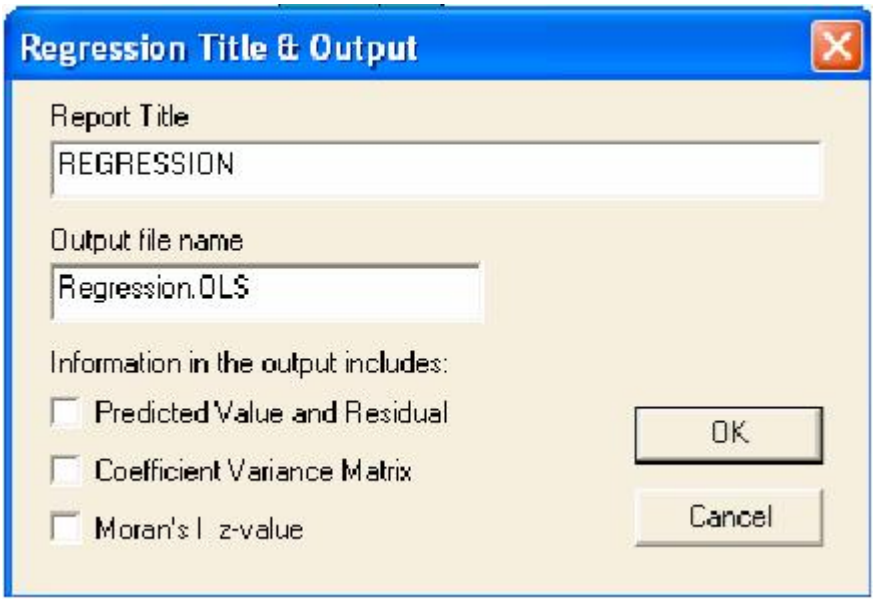


图 22.4 默认的回归标题和输出对话框

这个对话框有两个重要方面，输出文件名和输出选项。目前没有用到的 Report Title 可以忽略。在 Output file name 中指定的文件将在一个 rich 文本文件(RTF)中包含回归结果，保存在当前工作目录中。默认为 Regression.OLS，作为文件名没有什么意义。输入关于分析类型的的有提示作用的文件名，如 columbus.rtf，如图 22.5 和 22.6 所示。

对话框还包括一些多选按钮来指定长(long)输出选项。默认为不选，如图 22.5。可以选中各自的按钮创建长输出，如 Predicted Value and Residual 和 Coefficient Variance Matrix，如图 22.6。

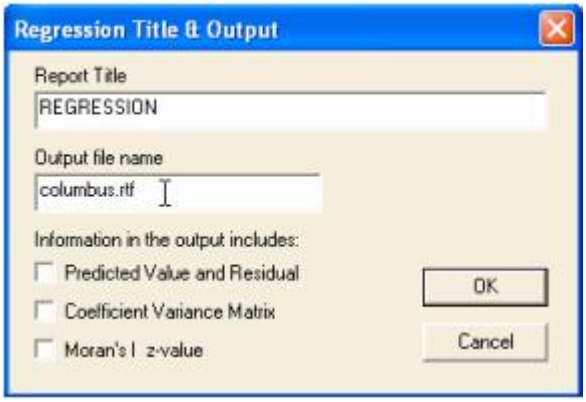


图 22.5 标准(短的)输出选项

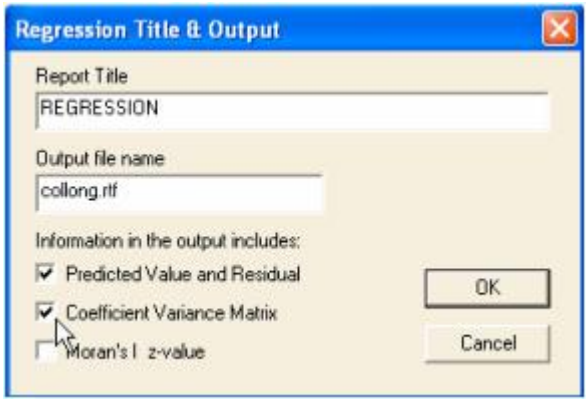


图 22.6 长输出选项

预测值及残差和选项应谨慎使用，特别是对大的数据集。它会向回归输出窗口和文件中添加两个矢量(vectors)，其长度等于观测点的数目。将很快导致失控，即使对中等大小的数据集。

相关系数变量矩阵不但提供估计变量(在对角线上)，也提供所有的协方差。这个矩阵可以用于

在 GeoDa 外执行模型相关系数的限制(customized)检验。如果并不关心这一点，可以不选这一项。同样，为了向表格中添加预测值及残差，长输出选项也不必选中，注意这一点非常重要(见 22.4.1 部分)。它们只影响列在输出窗口和文件中的内容。

在标题和输出对话框中点击 OK 按钮，出现回归模型指定对话框，如图 22.7 所示。

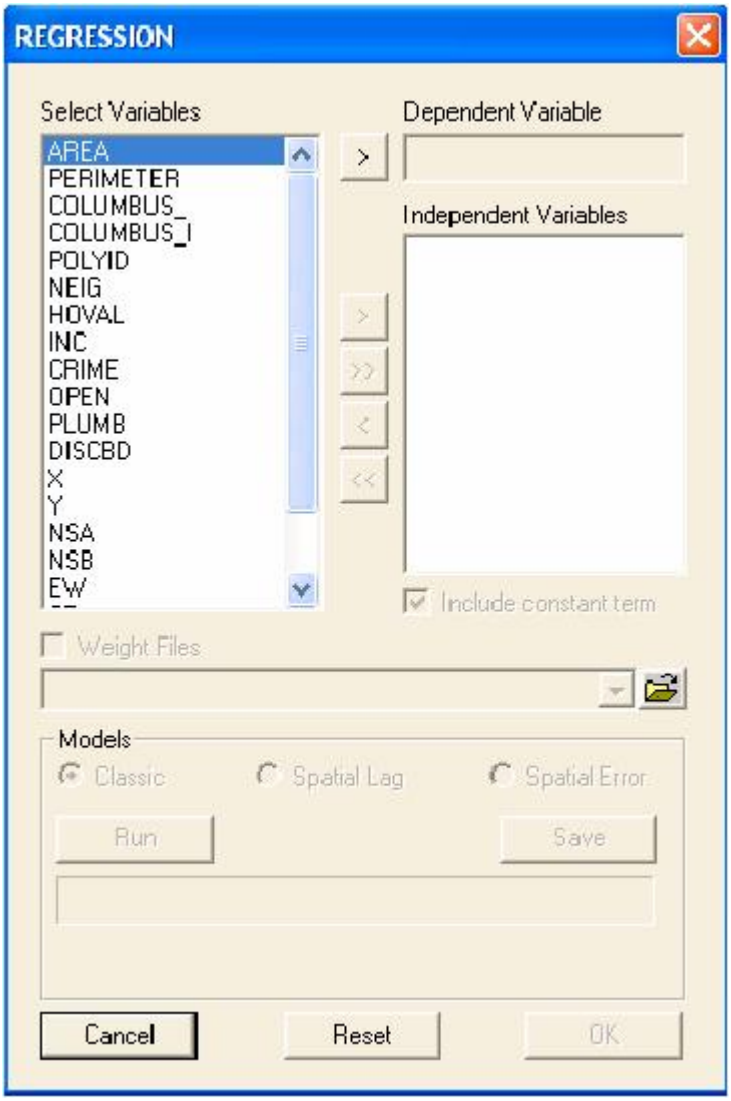


图 22.7 回归模型指定对话框

22.3 指定回归模型

图 22.7 所示的回归对话框中选择因变量和解释变量，也指定空间权重文件和所要执行的回归类型。现在，我们就将我们注意力只集中于 Classic 选项(默认)，不指定权重文件。

首先在 Select Variables 栏点击变量名 CRIME，选择 CRIME 为因变量，然后点击 Dependent Variable 旁的>按钮，如图 22.8。在选择一个因变量之间，只有一个 Cancer 按钮是可用的。直到选定一个 Independent Variable 才会改变这种情况。选择 Select Variable 栏中的 INC，点击>按钮将其移动到 Independent Variable 列中，如图 22.9 所示。为 HOVAL 重复这一过程，指定基本的回归模型，如图 22.10 所示。注意 Run 和 Reset 现在为可用。如果你想在运行之前重新指定模型，可以用 Reset。

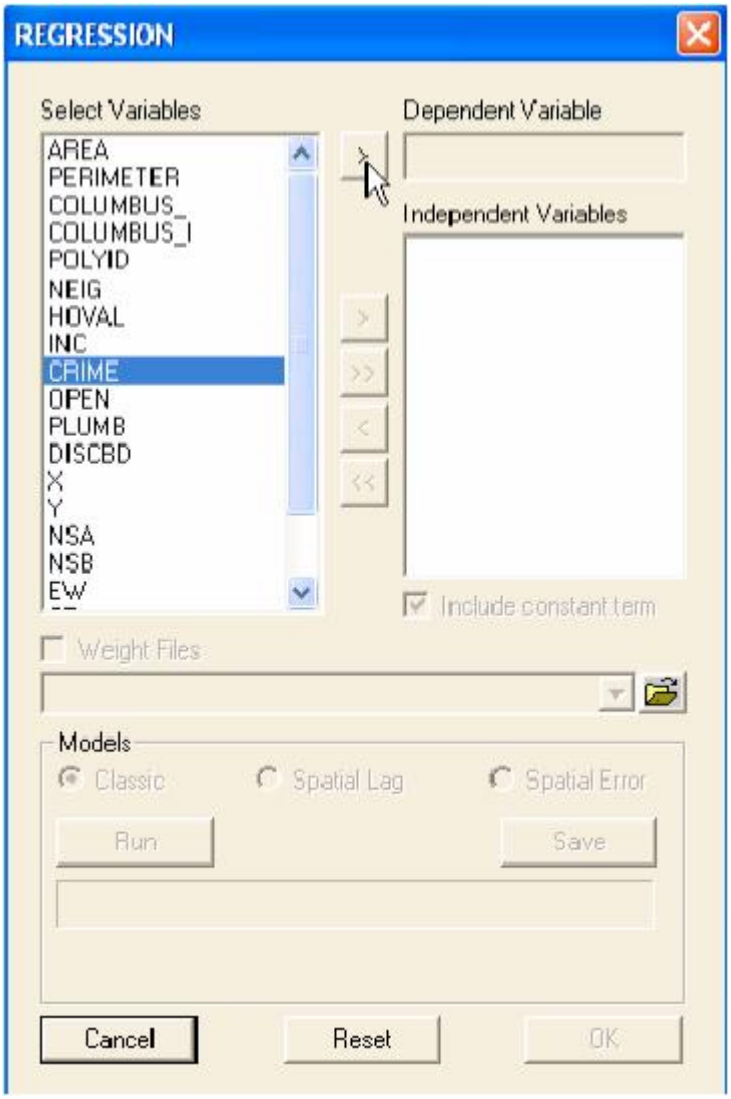

REGRESSION
Select Variables
AREA
PERIMETER
COLUMBUS_
COLUMBUS_I
POLYID
NEIG
HOVAL
INC
CRIME
OPEN
PLUMB
DISCBD
X
Y
NSA
NSB
EW
Dependent Variable
Independent Variables
Include constant term
Weight Files
Models
Classic
Spatial Lag
Spatial Error
Run
Save
Cancel
Reset
OK

图 22.8 选择因变量

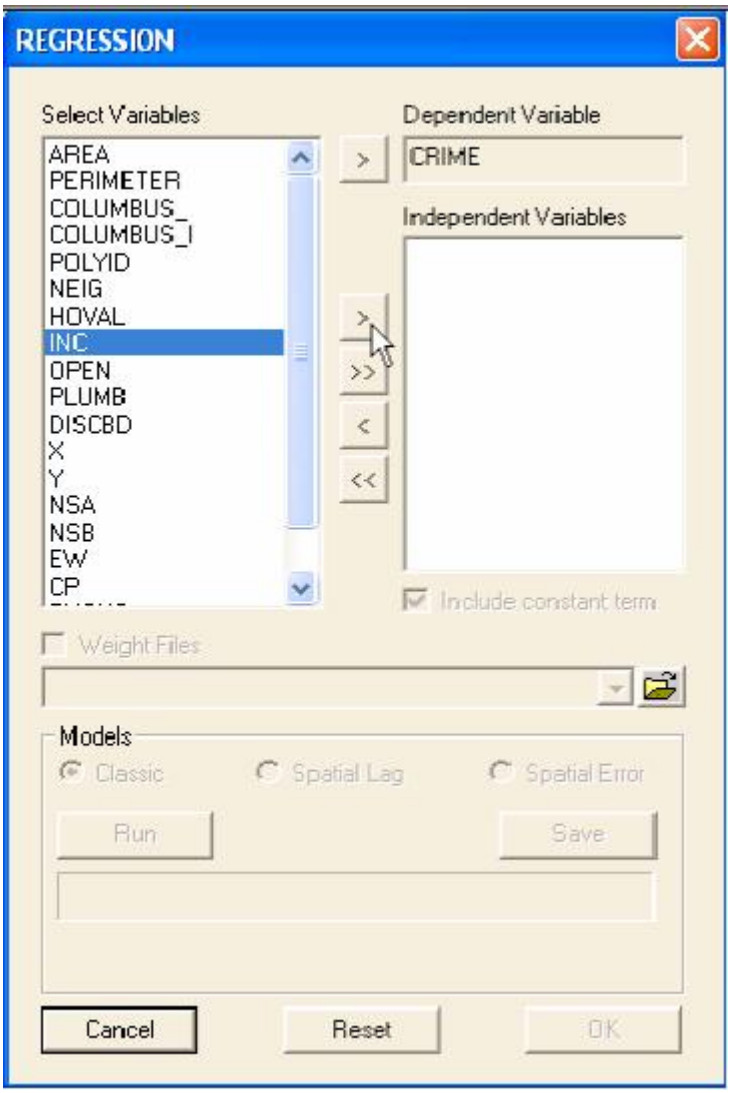

REGRESSION
Select Variables
AREA
PERIMETER
COLUMBUS_
COLUMBUS_I
POLYID
NEIG
HOVAL
INC
OPEN
PLUMB
DISCBD
X
Y
NSA
NSB
EW
CP
Dependent Variable
CRIME
Independent Variables
Include constant term
Weight Files
Models
Classic
Spatial Lag
Spatial Error
Run
Save
Cancel
Reset
OK

图 22.9 选择解释变量

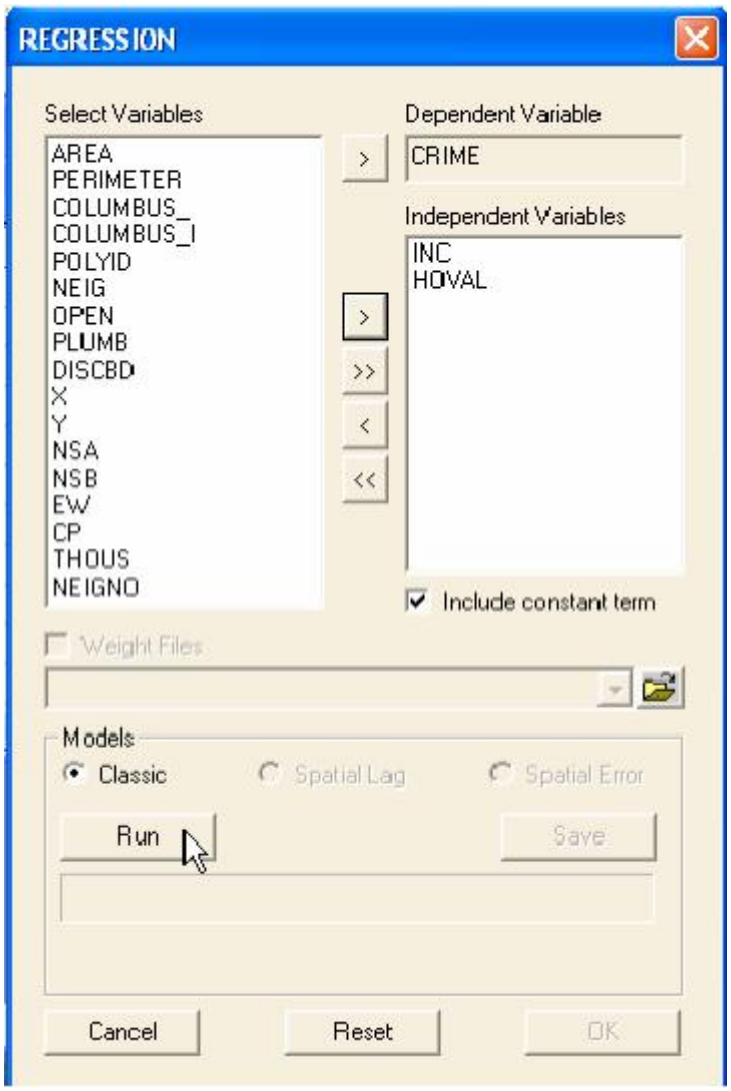


图 22.10 运行经典(OLS)回归

注意 Include constant term 是默认选项被选中。如果你有确定的理由在没有常数项(constant term)时运行模型，取消选中这一项。

点击 Run 按钮，运行回归。会出现一个进度条，显示估计过程已经完成(对 OLS 来讲，这一时间非常短)，图 22.11。

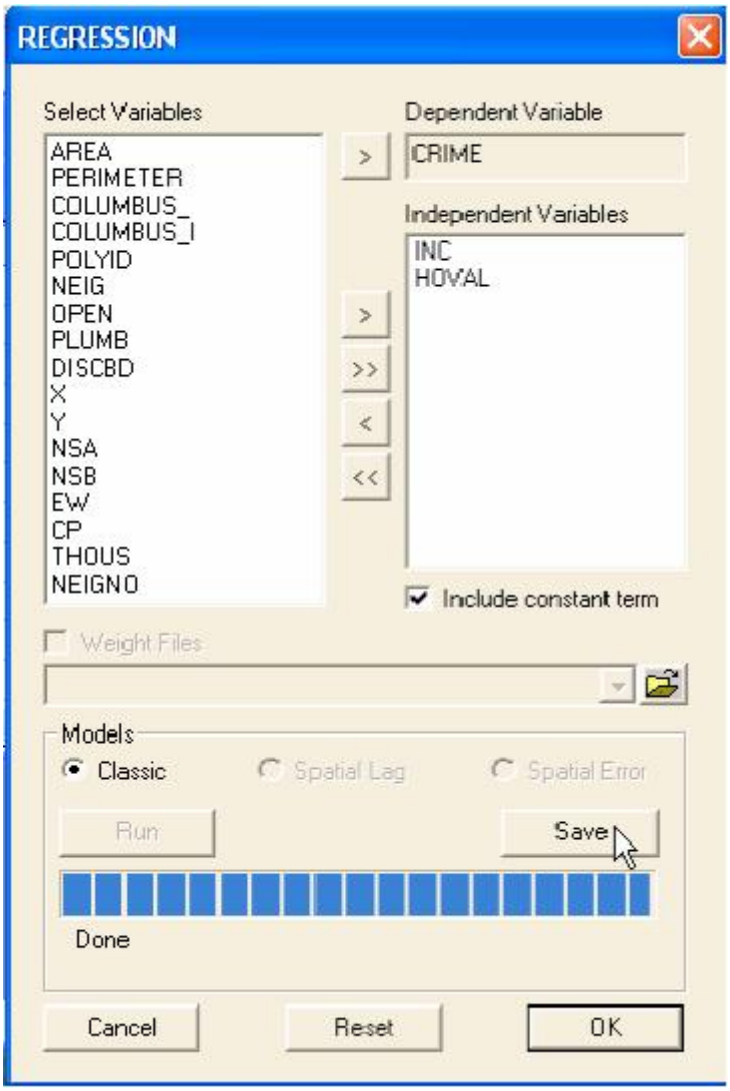


图 22.11 保存预测值及残差

这时可以点击 OK 按钮，会出现一个回归结果窗口。但是，如果你要将预测值及残差添加到数据表中，你必须首先选择 Save。在 GeoDa 的当前版本中，一旦打开了输出窗口，回归对话框会关闭，就不能再返回保存这些项目。

22.4 普通最小二乘回归

22.4.1 保存预测值及残差

如果你将预测值及/或残差添加到数据表中，不要选择 OK 按钮，而点击 Save，图 22.11。会产生一个对话框要求指定残差及/或预测值的变量名，如图 22.12。

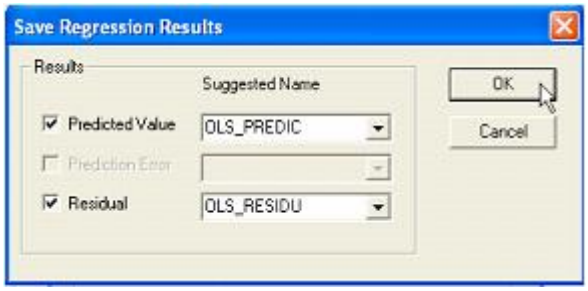


图 22.12 预测值及残差变量名对话框

NEIGNO	OLS_RESIDU	OLS_PREDIC
1005.000000	0.346542	15.379438
1001.000000	-3.694799	22.496553
1006.000000	-5.287394	35.914175
1002.000000	-19.985515	52.373275

图 22.13 添加到表格中的预测值及残差

在对话框中，你可以选中 Predicted Value 和/或 Residual 旁的多选框，保留默认的变量名(预测值为 OLS_PREDIC，残差为 OLS_RESIDU)，或用更有意义的名字来代替(重写默认值)。点击 OK，添加所选的栏到表格中。所有选中选项的结果如图 22.13。切记将文件以不同的名字保存，才能使新的变量成为 dbf 文件的永久组成部分。

22.4.2 回归结果

在回归变量对话框中点击 OK(图 22.14)，产生一个结果窗口，如图 22.15 所示。窗口的顶部包括一些模型的简要特征及契合度(measures of fit)。接着是一个变量名的列表，包括相关系数估计，标准差，T 统计和概率(拒绝 β=0 的 0 假设)。再后是模型诊断列表，有关讲座将在练习 23 中涉及。在窗口顶部所列的模型简要特征包括数据集名称(Columbus)，因变量(CRIME)，平均值(35.1288)和标准差(16.5605)。另外，列出了观测点的数量(49)，模型中的变量个数(包括常数项，值为 3)，自由度(46)。

标准输出结果的左栏为传统契合度(measures of fit)，包括 R2(0.55404)和调整 R2(0.532943)，残差平方和(6014.89)，残差方差和标准差估计，both with adjustment for a loss in degrees of freedom (sigma-square and S.E. of regression) as well as without (sigma-square ML and S.E. of regression ML)[1].

右栏列出的是关于 Null 假设的 F 统计量，所有回归系数都为 0(28.3856)，相关概率(9.3407e-009)。

This test statistic is included for completeness sake, since it typically will reject the null hypothesis and is therefore not that useful.

最后，这一栏包含了三个“量”(measures)，它们被包括在维持空间回归模型的可比性，将在练习 24 及 25 中论述。它们是对数似然值(log likelihood)(-187.377)，Akaike 信息标准(380.754)，Schwarz 标准(386.43)。这三个量(measures)是基于多变量正态假设和标准回归模型的对应似然功能。对数似然值越高，契合越好(在实线上高，所以负值是较好的)。对于信息标准，方向是正的，量越低，契合越好[2]。

当在回归标题对话框中选中了长输出选项，如图 22.6，在输出窗口中还包括另外的结果。它们是回归系数估计的完整协方差矩阵，和/或每一预测值的预测值及残差。这些结果列在诊断之后，如图 22.16 所示。变量名显示在协方差矩阵每栏的顶部 (此矩阵为对称，行与列相对应)。另外，对每一观测值，列出了预测的因变量，及预测值及残差(观测少于预测)。

[1] 两者的不同之处在于第一个的残差平方和被自然度(46)所除，第二个被观测点总数(49)所除。第二个值总会小于第一个，但对于大的数据集，差异可以忽略。

[2] AIC=－2L＋2K，L 指对数似然值，K 是模型中的参数数量，此处是 3。在 Columbus 例中：AIC＝－2×（－187.377）＋2×3＝380.754。SC＝－2L＋K×ln（N），ln 是自然对数。结果，在 Columbus 例中：SC＝－2（－187.377）＋3×3.892＝386.43。

collong.rtf

```
COEFFICIENTS VARIANCE MATRIX
    CONSTANT          INC        HOVAL
   22.424829    -0.942351    -0.161567
   -0.942351     0.111643    -0.017237
   -0.161567    -0.017237     0.010650


   OBS            CRIME        PREDICTED        RESIDUAL
     1          15.72598         15.37944         0.34654
     2          18.80175         22.49655        -3.69480
     3          30.62678         35.91417        -5.28739
     4          32.38776         52.37328       -19.98552
     5          50.73151         44.28396         6.44755
```

图 22.16 OLS 长输出窗口

22.4.3 回归输出文件

回归结果被写入一个当前工作路径的文件中，文件名是在标题对话框中指定的(图 22.6)。在本例中，文件为 columbus.rtf。文件为 rich 文本格式，已经在 Wordpad，Word 或其它文本编辑器中打开，你可以将结果剪切或粘贴到其它文档中。文件内容如图 22.17 所示。

注意当你试图在 Notepad 等简单文件编辑器中打开这个文件时，结果会如图 22.18，显示格式代码。

columbus.rtf - WordPad

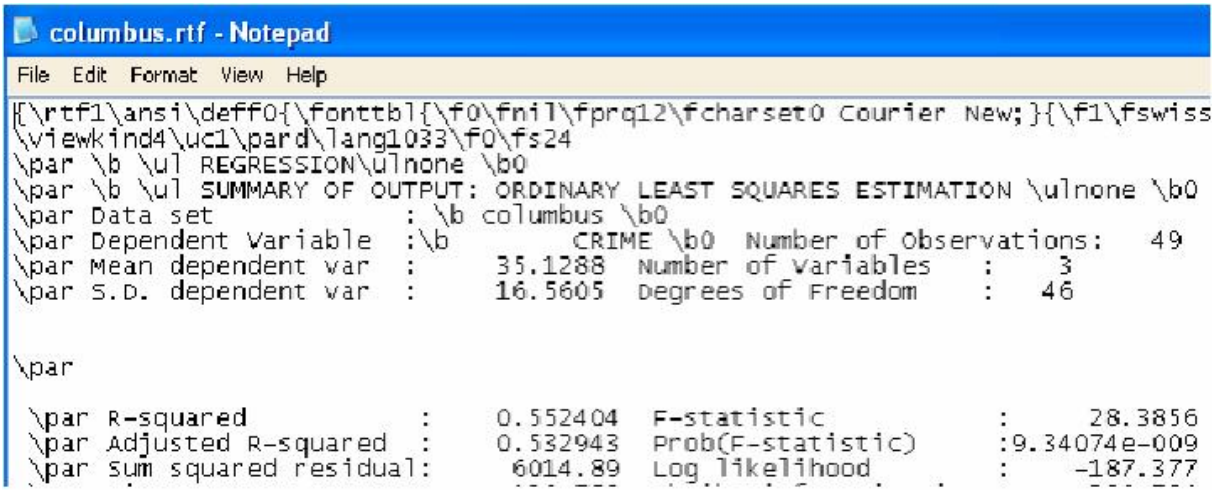
File Edit View Insert Format Help

Courier New 12 Western

```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set              : columbus
Dependent Variable    :       CRIME  Number of Observations:   49
Mean dependent var    :     35.1288  Number of Variables   :    3
S.D. dependent var    :     16.5605  Degrees of Freedom    :   46

R-squared             :    0.552404  F-statistic           :     28.3856
Adjusted R-squared    :    0.532943  Prob(F-statistic)     :9.34074e-009
Sum squared residual:       6014.89  Log likelihood        :    -187.377
Sigma-square          :     130.759  Akaike info criterion :     380.754
S.E. of regression    :      11.435  Schwarz criterion     :      386.43
Sigma-square ML       :     122.753
S.E of regression ML:       11.0794
-----------------------------------------------------------------------
    Variable     Coefficient      Std.Error     t-Statistic    Probability
-----------------------------------------------------------------------
    CONSTANT       68.61896       4.735486       14.49037      0.0000000
         INC      -1.597311      0.3341308      -4.780496      0.0000183
       HOVAL     -0.2739315      0.1031987      -2.654409      0.0108745
-----------------------------------------------------------------------
```

图 22.17 在 Wordpad 中 OLS rich 文本格式(rtf)输出文件


columbus.rtf - Notepad

File Edit Format View Help

```
{\rtf1\ansi\deff0{\fonttbl{\f0\fnil\fprq12\fcharset0 Courier New;}{\f1\fswiss
\viewkind4\uc1\pard\lang1033\f0\fs24
\par \b \ul REGRESSION\ulnone \b0
\par \b \ul SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION \ulnone \b0
\par Data set              : \b columbus \b0
\par Dependent Variable  :\b        CRIME \b0  Number of Observations:   49
\par Mean dependent var  :     35.1288  Number of Variables   :    3
\par S.D. dependent var  :     16.5605  Degrees of Freedom    :   46


\par

 \par R-squared           :    0.552404  F-statistic           :      28.3856
 \par Adjusted R-squared  :    0.532943  Prob(F-statistic)     :9.34074e-009
 \par Sum squared residual:     6014.89  Log likelihood        :     -187.377
```

图 22.18 Notepad 中打开的 OLS rich 文本格式(rtf)输出文件

22.5 预测值及残差地图

当预测值及回归残差被保存到数据表中，它们就可以在 GeoDa 任何的探索功能中使用，包括制图。地图，预测值地图和残差地图，对模式的可视化检查是非常有用的。在随机变化率的意义上，预测值地图被认为是 smoothed 地图，是由于因素，而不是包含在模型中的被清除(smooth out) 的因素。残差地图可能给出特定地区系统高值(预测值高于真实值)或低值(预测值低于真实值)的指示，这可能是空间自相关的证据(需要假设检验进行更严格的评估)。

例如，载入 Columbus 底图(图 22.1)，预测值的分位(六分位)地图可以制作了。从菜单选择 Map>Quantile，变量为 OLS_PREDIC(假设你已经将它添加到表格中，如 22.4.1)。然后，将分类数字由默认的 4 改为 6。结果将如图 22.19。

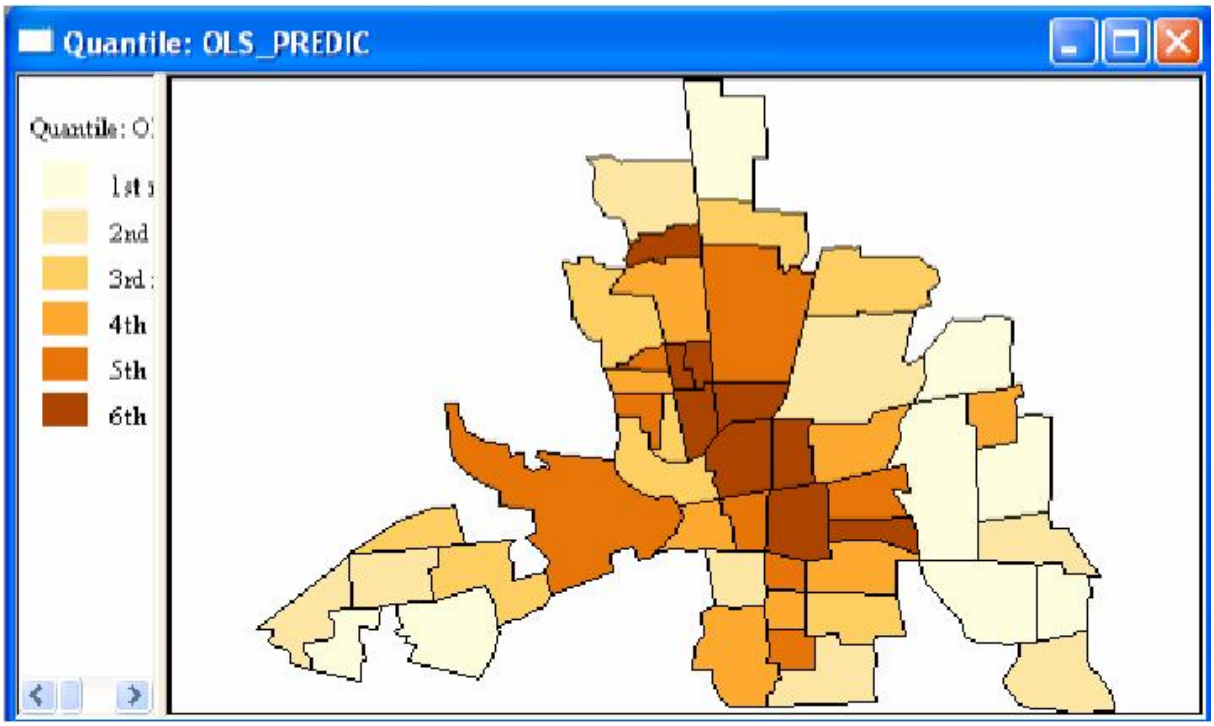


图 22.19 CRIME 回归的预测值的分位(6 类)地图

用 Map>St.Dev，选择 OLS_RESIDU，为残差产生一个标准差地图，如图 22.20。此图暗示相似颜色的区域趋向于具有相似的位置，指示具有正空间自相关关系(经 Moran's I 检验，残差空间自相关是正的且高度显著)。同样，它也显示出在偏远地区出现高预测值(负残差)，在核心地区出现低预测值(正残差)的趋势，说明由于空间政治制度不同而可能存在空间异质性。两个大的离群值(一个正的和一个负的)显示需要更大的关注。

22.6 练习

一些数据集中包括了允许已出版的空间回归研究的复制。包括从 Harrison 和 Rubinfeld()1978)(见 Gilley and Pace 1996，Pace and Gilley 1997)而来的 BOSTON 数据中的变量，POLICE 数据集中所需复制的 Kelejian and Robinson(1992)变量，Dubain(1992)的 BALTIMORE 数据中再运行回归的变量，Anselin(1993)举例的 NDVI 中的变量，Baller 等(2001)的 SOUTH 数据集的变量，Anselin 等(2004a)的 LASROSAS 数据集的变量。

例如，载入来自于 Kelejian and Robinson(1992)的 POLICE 数据集，再运行回归，以 POLICE 为因变量，TAX，INC，CRIME，UNEMP，OWN，COLLEGE，WHITE，COMMUTE 为解释变量。保存预测值和残差，比较观测的警用支出分位地图与预测值。创建一幅残差的标准差地图，“可视化”评估可能存在的模式。或者用上面所列的样本数据进行相似的练习。

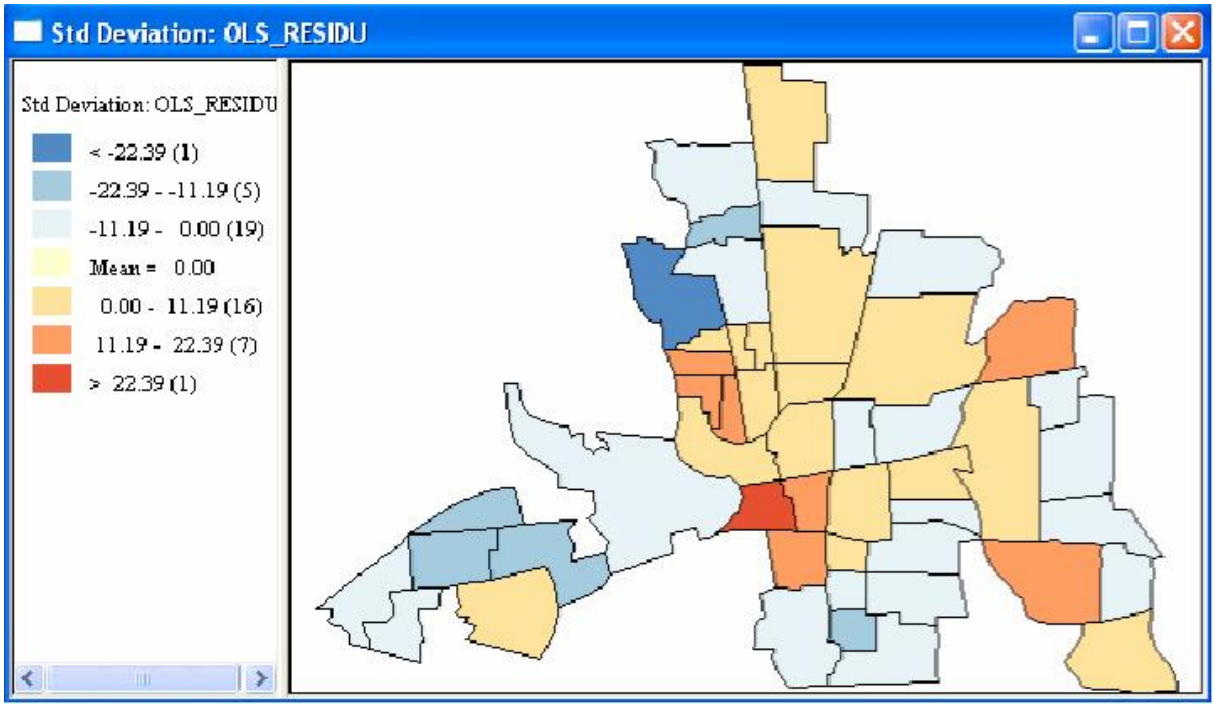


图 22.20 CRIME 回归残差的标准差地图

练习 23　回归诊断

23.1 目标

本练习继续谈论 GeoDa 的空间回归功能，现在关注回归诊断。标准回归诊断的方法背景，如多重共线性条件数量和正态及异方差性检验统计已经在许多经济计量学中论及，这里不再详细讨论。空间回归诊断的方法方面见 Anselin(1998)，Anselin and Bera(1998)，Anselin(2001)，及其它。

在本练习末，你将知道：

- 建立趋势面回归模式的详细说明
- 建立和解释回归诊断图
- 解释空间自相关的回归诊断
- 根据空间自相关诊断的结果，选择另外一个空间回归模型

有关这些操作的更详细的信息请参考 Release Notes 第 48－52 页。

23.2 准备工作

载入 Baltimore 样本数据集（baltim.shp，关键字为 STATION），其中有 211 个观测点的房屋售价和 hedonic 变量。底图如图 23.1。用这些点创建一个 Thiessen 多边形文件，名为 balthiesen.shp，关键字不变（按照 6.3 部分的指导）。结果应如图 23.2。最后，如果你还没有一个权重文件，就用 Thiessen 多边形创建一个 rook 邻接空间权重文件（baltrook.GAL），如图 23.3 所示（见 15.2 的指导）。

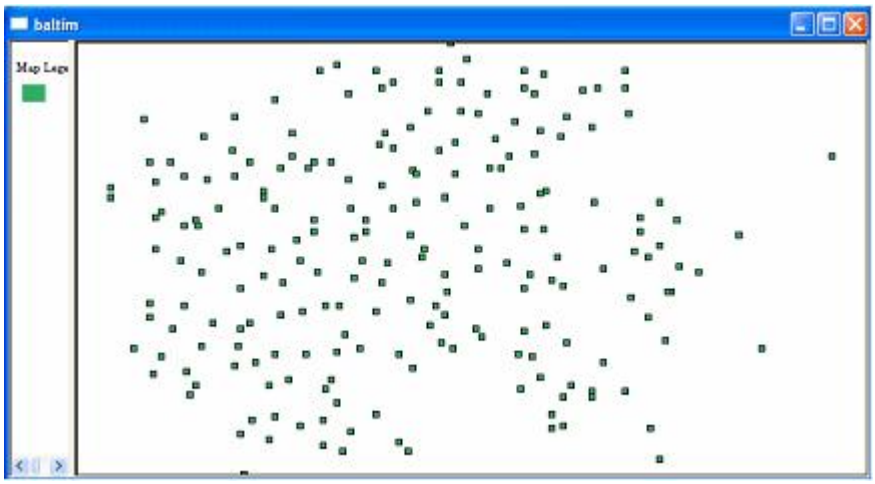
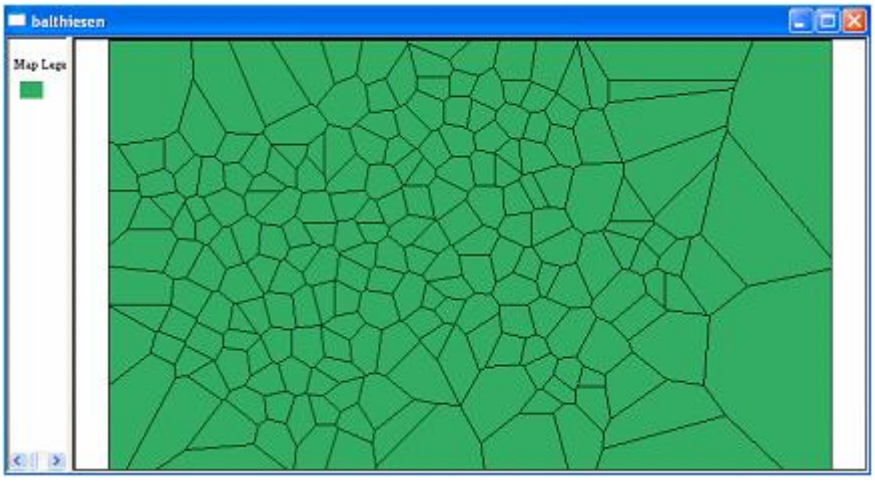


图 23.1　Baltimore 房屋销售点底图



23.2 Baltimore 房屋销售 Thiessen 多边形底图

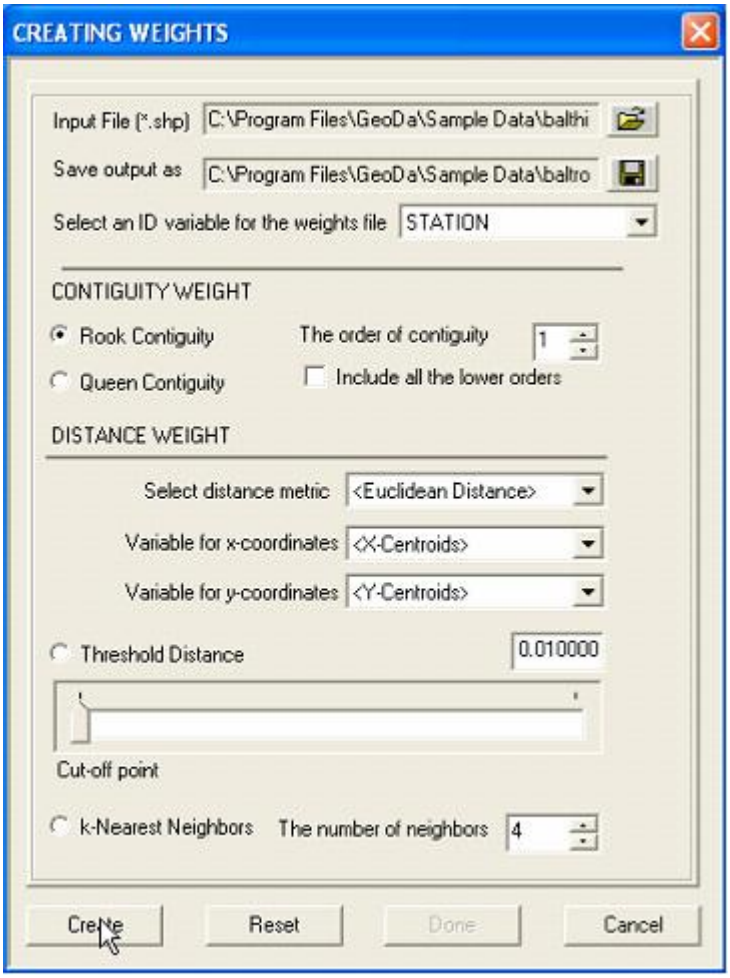


图 23.3　Baltimore Thiessen 多边形 Rook 邻接权重

23.3 趋势面回归

我们将分析房屋价格（PRICE 变量）趋势面回归模型的回归诊断。趋势面模型是详细说明，其中解释变量由观测点的 X 和 Y 坐标的多项式组成。我们将考虑线性（X 和 Y 为解释变量）和二次趋势面（X，Y，X^2，Y^2 和 XY 为解释变量）。

23.3.1 趋势面变量

趋势面回归中所需的解释变量只是数据集中的 X 和 Y。二次和叉积要用 Table 功能来建立（详细操作见 3.4 部分）。从 Table 菜单选择 Field Calculation，选中 Binary Operations 标签，计算平方坐标和叉积。例如，图 23.4 表明叉积计算的过程，结果为 XY，X 为 Variables-1，Y 为 Variables-2，Operators 为 MULTIPLY。点击 OK 将叉积添加到数据表中。最后，你将在数据表中有三个新的变量，如图 23.5。

图 23.4　趋势面变量计算

X	Y	X2	Y2	XY
907.000000	534.000000	822649.000000	285156.000000	484338.000000
922.000000	574.000000	850084.000000	329476.000000	529228.000000
920.000000	581.000000	846400.000000	337561.000000	534520.000000
923.000000	578.000000	851929.000000	334084.000000	533494.000000
918.000000	574.000000	842724.000000	329476.000000	526932.000000

图 23.5　将趋势面变量添加到数据表

23.3.2 线性趋势面

调用回归标题（regression title）对话框（在主菜单中选择 Regress），确保选中 Moran's I z-value 旁的多选框，如图 23.6。或者你可以指定一个输出文件名，而不是默认的，如图 23.6 中的 baltrend.

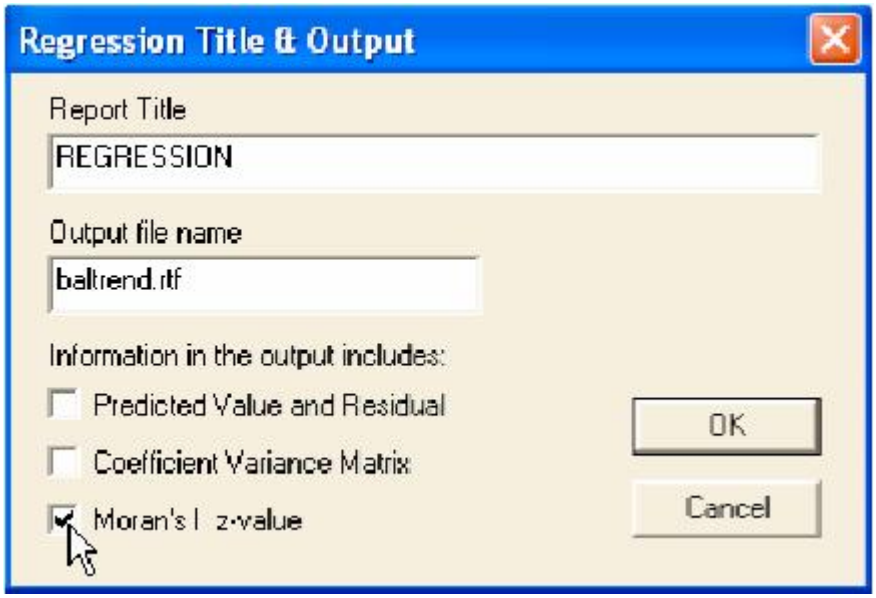


图 23.6　线性趋势面标题和输出设置

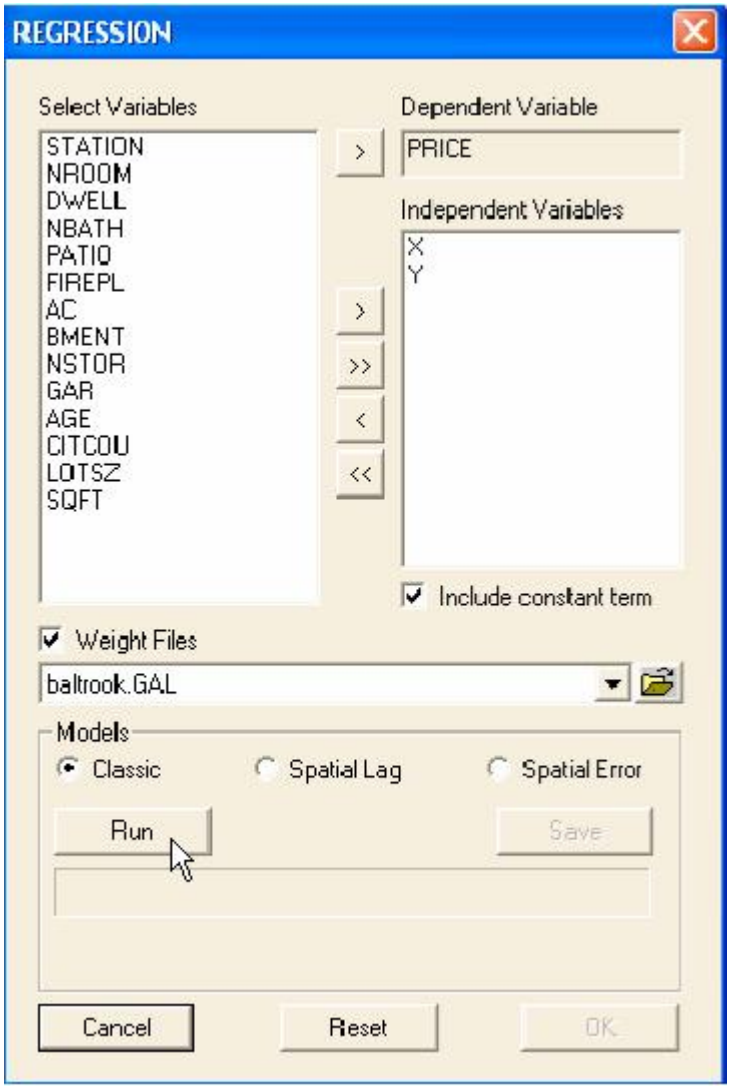


图 23.7　线性趋势面模型说明

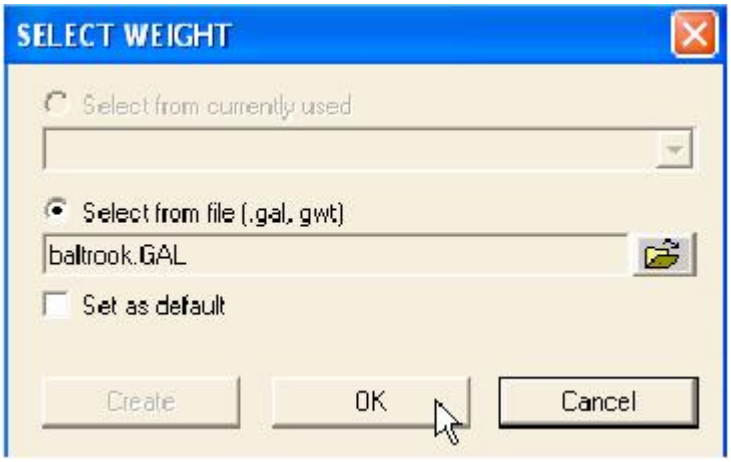


图 23.8　回归诊断的空间权重说明

点击 OK，产生回归说明对话框。选择 PRICE 作为因变量，选择 X 和 Y 作为自变量，如图 23.7。我们将关注空间自相关的诊断，所以要确保在运行回归前在对话框中选择了一个权重文件。例如图 23.7 中你刚刚建立的 baltrook.GAL。点击对话框中的打开文件按钮，在对话框中选择 baltrook.GAL

文件。然后，点击 Run，执行估计。在查看查实的回归结果之前，确信选择了 Save（*本程序中不可用*）按钮，为预测值和残差指定了变量名。它们会被添加到数据表中，也将在残差地图和其它诊断图中使用。在图 23.9 中，变量名分别为 OLS__PLIN 和 OLS__RLIN。最后，点击 OK，产生结果窗口，如图 23.10 所示。

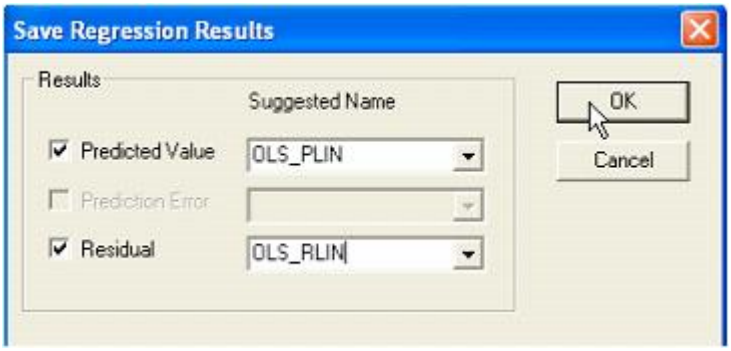
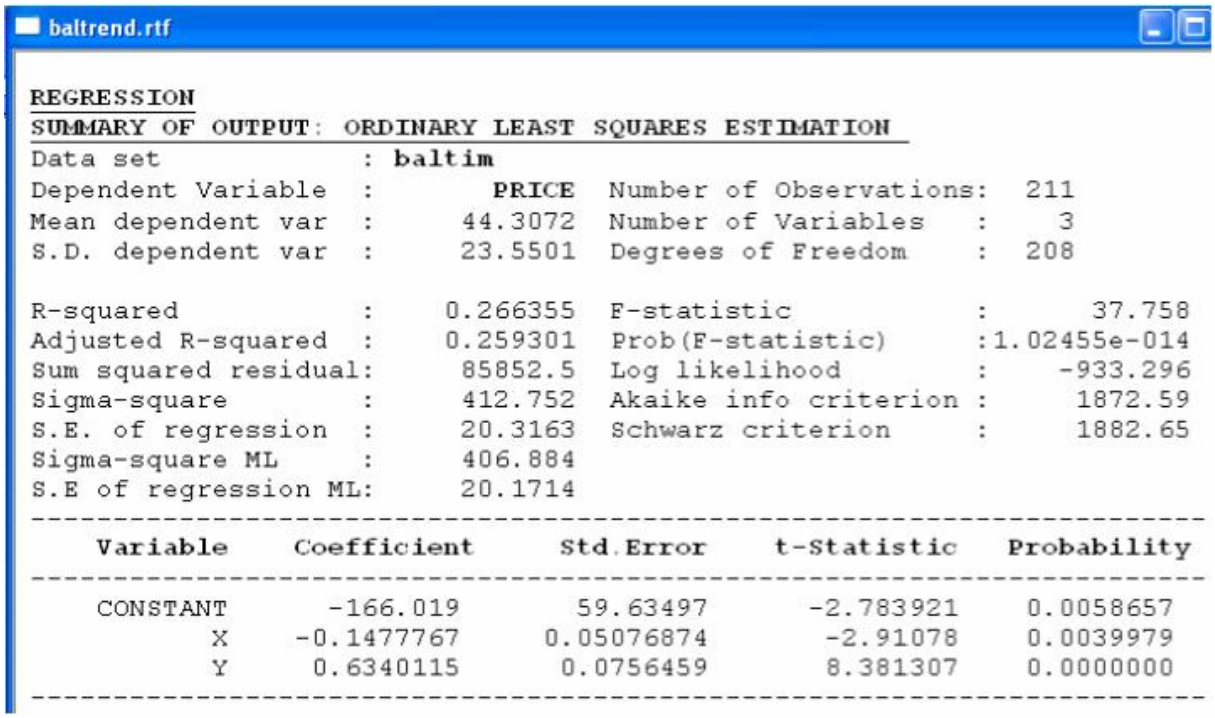


图 23.9　线性趋势面残差与预测值



baltrend.rtf

```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set              : baltim
Dependent Variable    :       PRICE  Number of Observations:  211
Mean dependent var    :     44.3072  Number of Variables   :    3
S.D. dependent var    :     23.5501  Degrees of Freedom    :  208

R-squared             :    0.266355  F-statistic           :      37.758
Adjusted R-squared    :    0.259301  Prob(F-statistic)     :1.02455e-014
Sum squared residual:      85852.5  Log likelihood        :    -933.296
Sigma-square          :     412.752  Akaike info criterion :     1872.59
S.E. of regression    :     20.3163  Schwarz criterion     :     1882.65
Sigma-square ML       :     406.884
S.E of regression ML:      20.1714
-----------------------------------------------------------------------
    Variable     Coefficient      Std.Error     t-Statistic   Probability
-----------------------------------------------------------------------
    CONSTANT       -166.019       59.63497      -2.783921     0.0058657
           X     -0.1477767     0.05076874       -2.91078     0.0039979
           Y      0.6340115      0.0756459       8.381307     0.0000000
-----------------------------------------------------------------------
```

图 23.10　线性趋势面模型结果

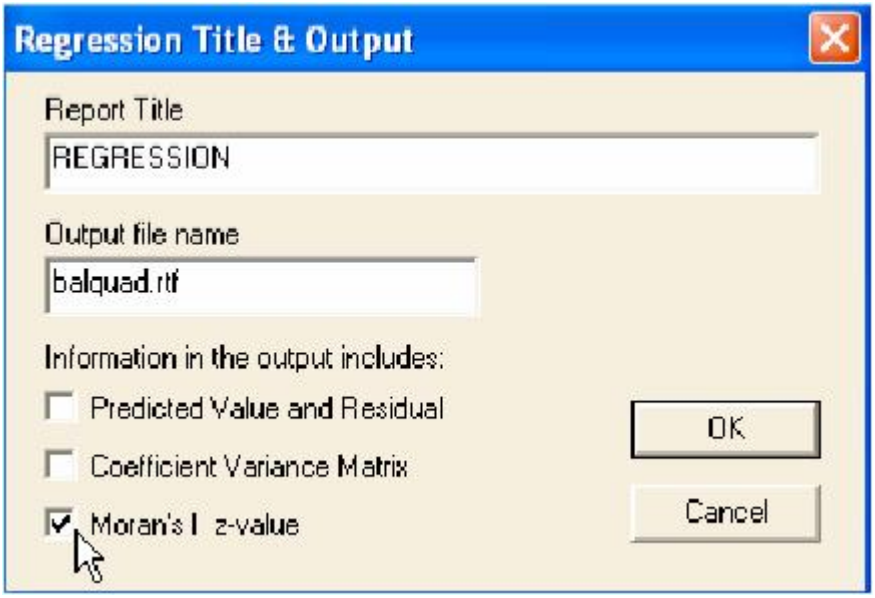


图 23.11　二次趋势面标题和输出设置

回归结果显示出一个较好的契合，调整 R2＝0.26。两个维都非常显著，但意义不同。X 变量为负相关，说明从西向东下降。相反，Y 变量为正相关，显示从南向北为上升。

23.3.3 二次趋势面

除了线性趋势，我们还要考虑二次趋势面模型的诊断。打开回归标题窗口（Regress），输入一个文件名（如 balquad.rft），确定选中了 Moran's I z-value，如图 23.11。点击 OK，产生图 23.12 中的回归说明对话框。

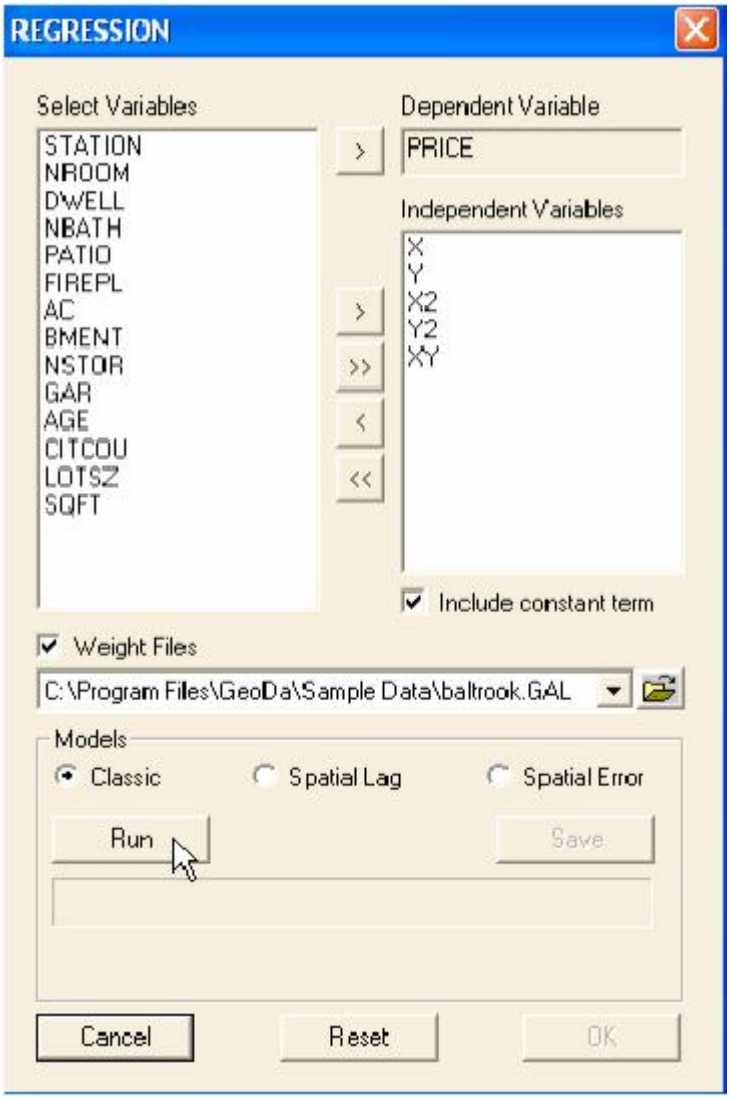


图 23.12　二次趋势面模型说明

对于线性趋势，指定 PRICE 为因变量。选择 X，Y，X2，Y2 和 XY 为自变量（见图 23.12）。确定选择了 baltroo.GAL 为权重文件，点击 Run 开始运行（estimation）。

如同线性趋势，指定 PRICE 为因变量。选择 X，Y，X2，Y2 和 XY 为自变量（见图 23.12）。确保选择了 baltrook.GAL 为权重文件，点击 Run 开始运行。

在查看回归结果前，选择 Save 来指定保存的预测值（OLS__PQUAD）和残差（OLS__RQUAD）变量名，如图 23.13 所示。在这个对话框点击 OK，将变量添加到数据表中，在回归对话框中点击 OK，产生结果窗口，如图 23.14 所示。

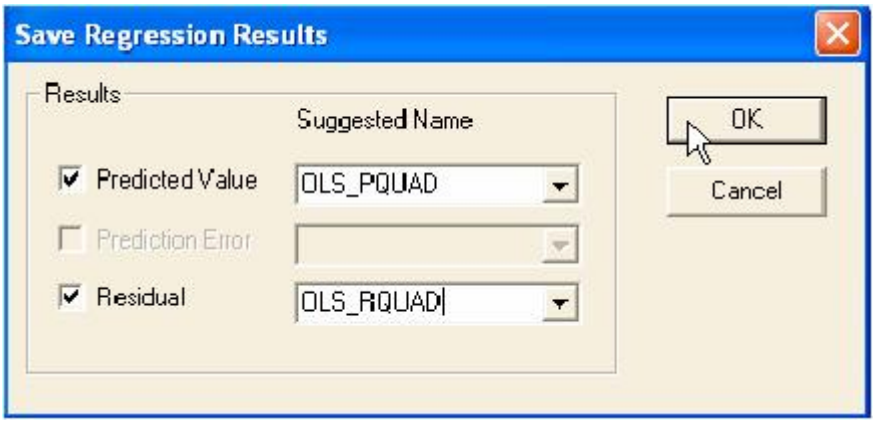


图 23.13　二次趋势面残差与预测值

baltquad.rtf

```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set              : baltim
Dependent Variable    :       PRICE  Number of Observations:  211
Mean dependent var    :     44.3072  Number of Variables   :    6
S.D. dependent var    :     23.5501  Degrees of Freedom    :  205

R-squared             :    0.451675  F-statistic           :     33.7732
Adjusted R-squared    :    0.438302  Prob(F-statistic)     :4.38088e-025
Sum squared residual:       64165.9  Log likelihood        :    -902.579
Sigma-square          :     313.005  Akaike info criterion :     1817.16
S.E. of regression    :     17.6919  Schwarz criterion     :     1837.27
Sigma-square ML       :     304.104
S.E of regression ML:       17.4386
--------------------------------------------------------------------------
    Variable     Coefficient      Std.Error     t-Statistic    Probability
--------------------------------------------------------------------------
    CONSTANT       10944.32       2105.653       5.197588      0.0000005
           X      -9.601599       3.124628      -3.072878      0.0024081
           Y      -24.36005       4.225294      -5.765291      0.0000000
          X2    0.006672922    0.001509604       4.420313      0.0000160
          Y2     0.02728535    0.003573074       7.636379      0.0000000
          XY   -0.005101239    0.002666842      -1.912839      0.0571623
--------------------------------------------------------------------------
```

图 23.14　二次趋势面模型输出结果

回归增加了平方及叉积后，调整 R2 提高到 0.44。但叉积 XY 表不显著，相反所有其它的系数都是显著的。模型暗示趋势面为一个碗形，低房价在中间，向四周逐渐升高。

为显示这种模式，用预测值创建一幅分位地图。确保活动窗口是 Thiessen 多边形底图（balthiessen.shp）。然后，从菜单中选择 Map>Quantile，选择 OLS__PQUAD 为变量，将分类数字改为 6。结果如图 23.15，较暗阴影与较高的房价相对应。

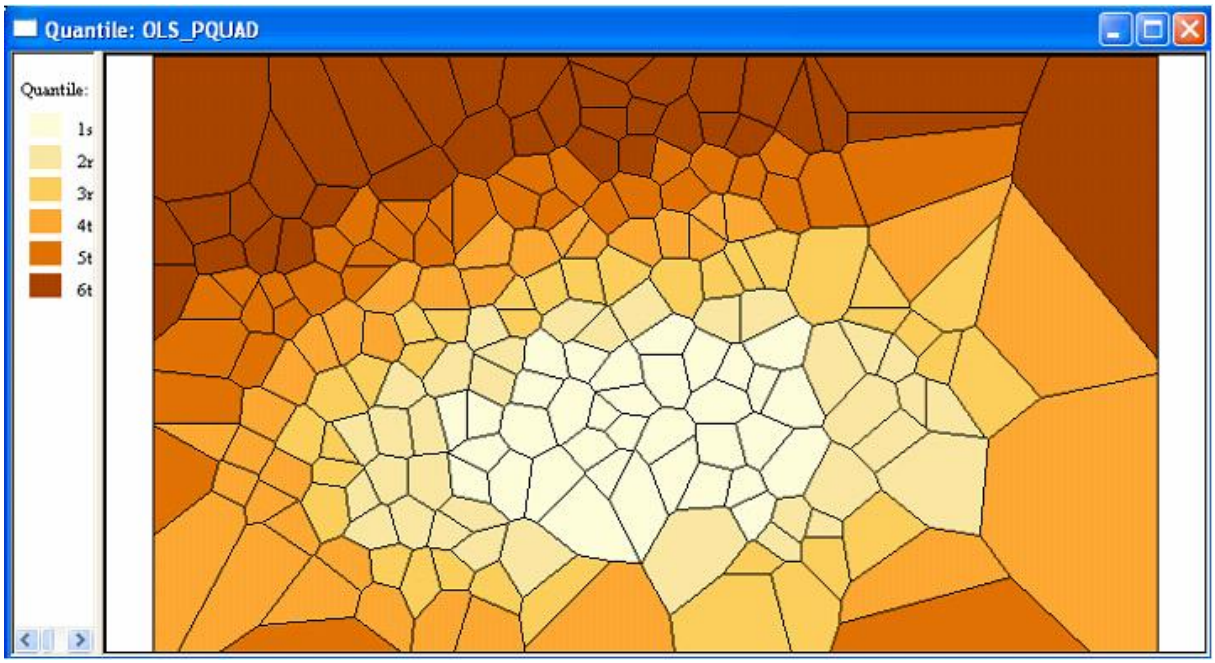


图 23.15　二次趋势面预测值地图

23.4 残差地图和残差图

当预测值和残差以新加变量被保存到数据表中，它们即可在 GeoDa 其它探索功能中使用。这对于建立诊断地图和绘图是有帮助的。在下面的例子中，我们将用二次趋势面回归的残差和预测值，OLS__PQUAD 和 OLS__RQUAD。也可以直接复制线性趋势模型的残差（OLS__RLIN）和预测值（OLS__PLIN）。

23.4.1 残差地图

最有用的残差地图可能就是标准差地图，因为它清晰显示高估或低估的模式，残差的大小，特别是超过了两倍标准差的值。选 Map>St.Dev，选择 OLS__RQUAD 作为变量。结果地图如图 23.16 所示。注意清楚表示出的高估值（负的残差，或蓝色调）和低估值（正的残差，棕色调）。这种“可视化的检察”暗示存在空间自相关，但在下定论前需要进行正式的检验。

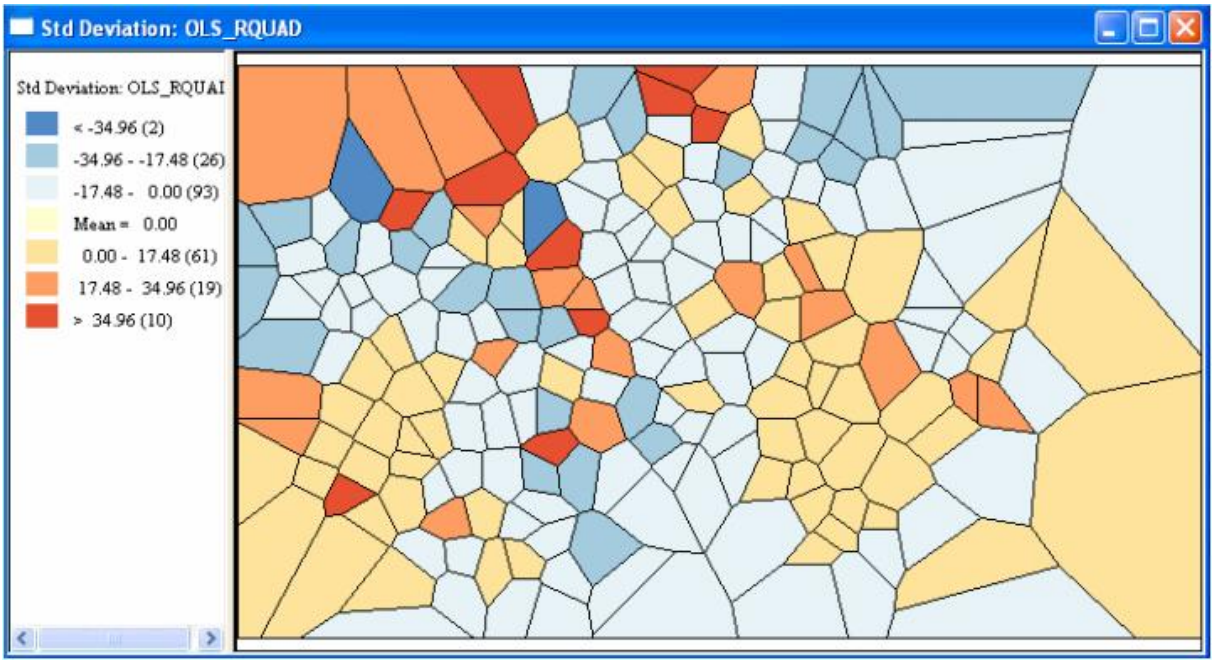


图 23.16　二次趋势面残差地图

还要注意一些残差非常大的地区（深棕色和深蓝色）。这本奇怪，因为“模型”中包括变量的位置，没有考虑房屋的其它独特的特征。离群值表明存在着仅用位置并不能以解释价格的交易。选择这些位置，链接到图表或地图中（如一些多变量 EDA 工具）可能可以解释哪一个变量应当包括在改进的回归解释（specification）。

23.4.2 模型查看图

模型残差的简单绘图经常是有启迪作用（revealing）的，因为它不能显示任何类型的图形（patterning）。GeoDa 现在还没有简单绘图功能，但可以用散点图来达到同样的目的。例如，绘制二次趋势面模型残差相对于一列观测数目的，如变量 STATION。

开始 Explore>Scatter Plot，选择 OLS__RQUAD 作为第一变量（Y 轴），STATION 作为第二变量（X 轴）。绘制结果如图 23.17。

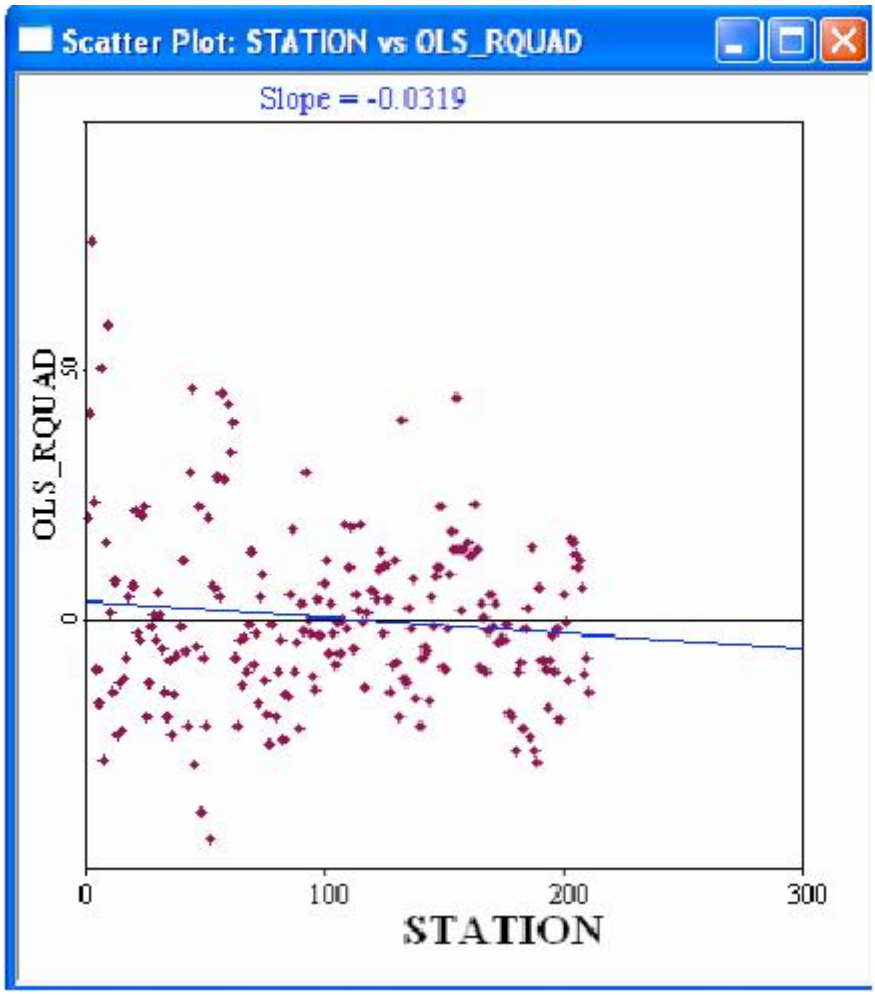


图 23.17　二次趋势面残差绘图

绘图确认了存在一些非常大的残差。在图中选择这些，链接一幅地图或其它统计图表（描述其它变量）可以显示与“忽略”变量的系统联系，优化模型。

另外要关注预测值相对于残差的图。这里的兴趣在于探测异方差性的模式，或残差方差与其它变量的变化。如前，选择 Explore>Scatter Plot，选 OLS__RQUAD 作为第一变量（Y 轴），OLS__PQUAD 为第二变量（X 轴）。结果如图 23.18。

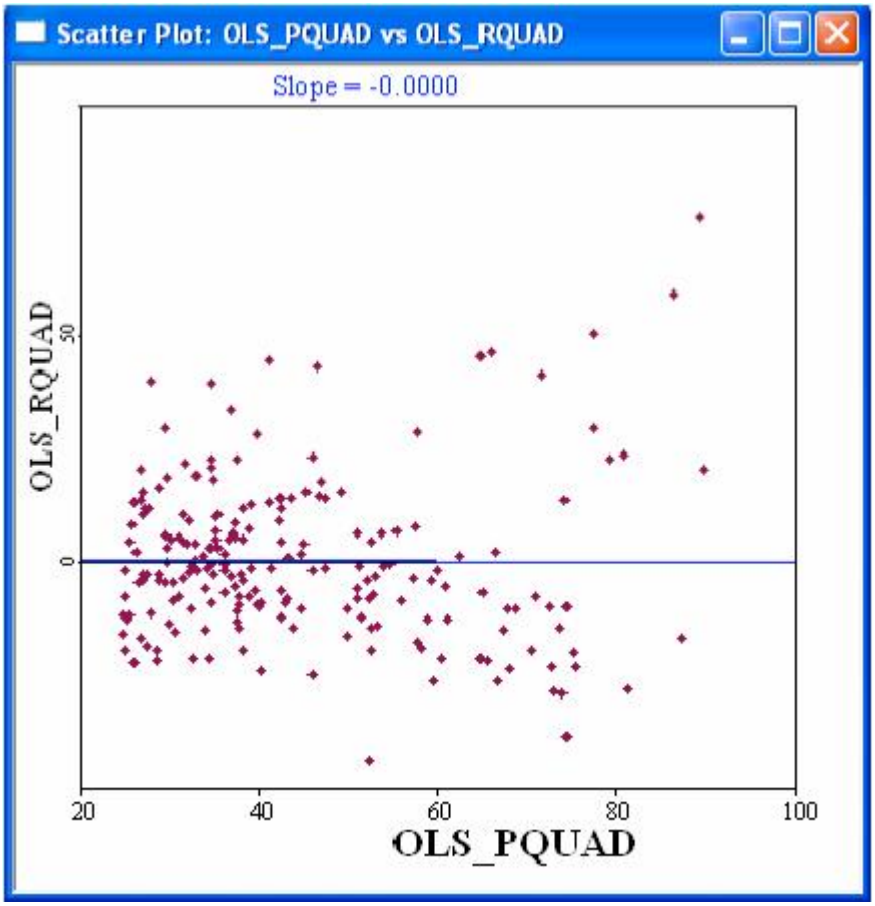


图 23.18　二次趋势面残差/契合度值图

在该图中，有人试图找出形似漏斗的模式，这种模式显示了观测值与残差传播之间的关系。在图 23.18 中有一些这样的证据，但不足以得出有力的结论。正式的异方差性的检验将需要执行“可视化检查”。

除了观测值，也可以选择其它变量为 X 轴，特别是当非常怀疑它们是“原因”异方差性时。经常地，这些变量与大小相关，如地区或总人口。

23.4.3 残差的 Moran 散点图

残差的空间模式可以通过 Moran 散点图来进行更正式的分析。在常用的方式中，选择从菜单选择 Space>Univariate Moran，选择 OLS__RQUAD 作为变量，baltrook.GAL 为空间权重文件。

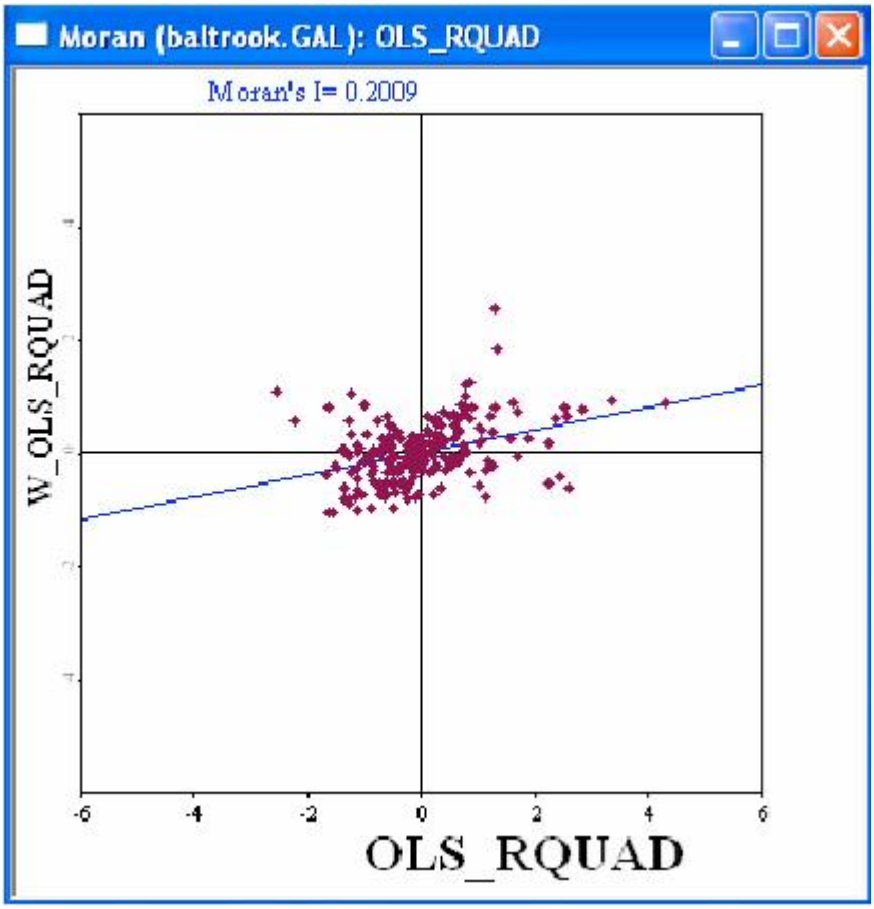


图 23.19　二次趋势面残差 Moran 散点图

结果将如图 23.19 所示，显示残差的 Moran's I=0.2009。注意这种度量已经足够解释了，它也可能进行链接和刷光。用序列方法评估其显著性是不合适的[1]。两样的原因，也不适于为残差建立 LISA 地图。

23.5 多生共线性，正态，异方差性

在回归结果窗口中诊断的第一部分由三个传统标准组成：多重共线性条件数，非正态检验（Jarque-Bera），三个异方差性诊断（Breusch-Pagan，Koenker-Bassett，and White）[2]。

线性和二次趋势面模型的结果分别列在图 23.20 和图 23.21 中。首先看多重共线性数目。它在本质上不是一个检验统计量，但是由于多重共线性，它一个显示回归结果稳定性问题的诊断（解释变量相互关系很强，提供不充分的独立信息）。典型地，超过 30 的数目问题的暗示。在趋势面回归中，这很常见，因为解释变量是变量一次和交叉乘积。在我们的例子中，线性模型的值为 90.8，但二次模型的值高达 7842.2。

误差正态性的 Jarque-Bera 检验的分布如同 x^2 统计量，自由度为 2。在这两个例子中，强烈建议误差的非正态性。就其本身而言，这并不是一个太严重的问题，因为在回归分析中的许多属性渐近均匀分布，没有假设正态。但是，对于“有限例子（或精确的）”推断，正态分布是必不可少的，当前的模型显然违反这一假设。

后面三个诊断是探测异方差性常用的统计量，如非常数误差方差。Breusch-Pagan 和

[1] 这是因为 OLS 残差在建立时已经是存在相互关系的，序列方法忽视这一事实。
[2] 有关这些诊断的方法细节，见中级经济计量学论文。

Koenker-Bassett 检验作为“random coefficients”检验来执行，它是假设异方差性的特定功能形式[1]。Koenker-Bassett 检验与 Breusch-Pagan 检验是同等重要的，除了残差是化残差的（studentized），例如，它们对非正态是稳定的。在每个趋势面说明中，这两个检验统计量都显示了异方差性问题。

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER    90.81299
TEST ON NORMALITY OF ERRORS
TEST                    DF          VALUE            PROB
Jarque-Bera              2           143.9515        0.0000000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF          VALUE            PROB
Breusch-Pagan test       2           42.40993        0.0000000
Koenker-Bassett test     2            16.1171        0.0003164
SPECIFICATION ROBUST TEST
TEST                    DF          VALUE            PROB
White                    5           35.39886        0.0000013
```

图 23.20　回归诊断—线性趋势面

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER    7842.225
TEST ON NORMALITY OF ERRORS
TEST                    DF          VALUE            PROB
Jarque-Bera              2           65.78718        0.0000000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF          VALUE            PROB
Breusch-Pagan test       5           86.58141        0.0000000
Koenker-Bassett test     5              43.695       0.0000000
SPECIFICATION ROBUST TEST
TEST                    DF          VALUE            PROB
White                   20            N/A            N/A
```

图 23.21　回归诊断—二次趋势面模型

White 检验称为“specification-robust”异方差性检验，因为它并不假设一个异方差性的特定功能形式工。相反，它近似一个模型中解释变量的所有平方和交叉乘积的大的概率范围。在一些例子中，当交叉乘积已经作为一个项目存在时会产生一个问题。二次趋势面的情况是其已经包括平方和 X 和 Y 变量的交叉乘积。在此情况下，存在多重共线性。目前，GeoDa 还不能校正这一项，报告 N/A，如图 23.21。在线性趋势面模型中，White 统计量是 35.4，它支持其它两个检验所提供的异方差性的证据。这一结果并不完全总是可信，因为 Breusch-Pagan and Koenker-Bassett 检验中所执行的随机系数假设并不合适。在此例中，White 检验可能并不显著，但其它两个可能显著。重要的是记住 White 检验与常用异方差性形式是相反的。

[1] 特别的，异方差性是解释性变量平方的函数。这是在 GeoDa 中执行的形式。在其它的计量经济学软件中，使用 X 值本身，而不是平方，这可能会给出稍有不同的结果。

23.6 空间自相关诊断

模型诊断的最后部分是对空间自相关的检验。共报告了 6 个统计量，如图 23.22 中对线性趋势面和图 23.23 中对二次趋势面。现在，所有检验都是用相同的权重矩阵来计算，这里使用了 baltrook.GAL，在诊断窗口中已列出。结果，如果你想要用一些空间权重检验残差，你需要重新运行回归分析[1]。

```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : baltrook.GAL  (row-standardized weights)
TEST                              MI/DF       VALUE           PROB
Moran's I (error)               0.360334     9.3483560       0.0000000
Lagrange Multiplier (lag)          1        74.6629386       0.0000000
Robust LM (lag)                    1         0.0469581       0.8284436
Lagrange Multiplier (error)        1        75.9419444       0.0000000
Robust LM (error)                  1         1.3259639       0.2495245
Lagrange Multiplier (SARMA)        2        75.9889025       0.0000000
========================= END OF REPORT ============================
```

图 23.22 空间自相关诊断－线性趋势面模型

```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : baltrook.GAL  (row-standardized weights)
TEST                              MI/DF       VALUE           PROB
Moran's I (error)               0.200899     5.8323965       0.0000000
Lagrange Multiplier (lag)          1        20.5454453       0.0000058
Robust LM (lag)                    1         1.2626832       0.2611438
Lagrange Multiplier (error)        1        23.6063131       0.0000012
Robust LM (error)                  1         4.3235510       0.0375884
Lagrange Multiplier (SARMA)        2        24.8689962       0.0000040
========================= END OF REPORT ============================
```

图 23.23 空间自相关诊断－二次趋势面模型

23.6.1 Moran's I

第一个统计量就是 Moran's I，与图 23.19 中 Moran 散点图的值是相同的（二次趋势面的值为 0.200899）。当在回归标题对话框中选中了 Moran's I 选项（如图 23.11），在诊断结果中会报告出 Z 值与相关的 P 值[2]。当没有选中些选项，只报告出这个统计量[3]。

在线性和二次趋势面模型中（Z 值分别是 9.35 和 5.83），Moran 统计量高度显著，暗示存在着空间自相关的问题。当 Moran's I 统计量有很强的探测模型中设定误差的能力（不仅是空间自相关），暗示应该用哪一个 specifiction 就没有多大用处了。最后，我们用 Lagrange 乘数检验统计量。

23.6.2 Lagrange 乘数检验统计量

诊断结果中给出了 5 个 Lagrange 乘数检验统计量。前两个（LM－Lag and Robust LM-Lag）适合空间滞后模型，可以二者选其一。然后两个（LM－Error and Robust LM-Error）是指空间误差模型，可以二者选其一。最后一个，LM－SARMA，是关于空间滞后和空间误差项的模型的高次选择。这个检验只是为了完全性，因为它在实践中并没有多大用处。更需说明的是，除了探测设计的更高次

[1] 在 GeoDa 未来的版本中可能会有所改变。
[2] Z 值是根据正态近似值，考虑存在残差的事实（而计算的）。详见 Anselin and Bera (1998)。
[3] 对于大的数据集，计算 Moran's I 的 Z 值所需的矩阵操作相当费时。结果，在 GeoDa 中统计量的推断是可选项。还有，在 23.6.3 中 specification 搜索 outlined 是根据 Lagrange 乘数统计量，而不是根据 Moran's I。

的选择，检验也有高次的一维选择。换句话说，当误差或滞后模型是合适的选择，蛤并不是需要的更高次的选择时，它将趋于显著。

所有的一维检验统计量服从 x^2 分布，自由度为 1（LM－SARMA 检验统计量自由度为 2）。为指导 specification search，检验统计量应该放以一定的顺序来考虑，将在 23.6.3 中详细说明。要记住的重要一项是当标准形式（LM－Lab or LM-LError）是显著的，只需考虑统计量的 Robust 形式（versions）。当都不显著时，Robust 形式的属性可能不再 hold[1]（有用、有效？）。

这两种趋势面模型，LM－Lag 和 LM-Error 统计量都是高度显著的，后者稍微更显著些。被两种 LM 检验统计量拒绝 0 假设是在实践中常会遇到的情况。注意 LM－SARMA 统计量是显著的。但是，它的值只是比一维（one-directionalc 统计量稍高，说明它可能选取（pick up）了一个单选，而不是真正的高次模型。在线性趋势面中（图 23.22），Robust 检验结果是反常的。当标准检验是显著的，Robust 形式就不显著，说明不存在空间自相关问题。这明显是错误的（见 Moran's I）。它显示其它的设定误差问题存在的，这使 Robust LM 检验统计量所依据的渐近结果失效。这并不奇怪，因为线性趋势 specification 非常简单。幸运的是，这种检验统计量结果在实践中很少用到。

更常见的结果是二次趋势面的，如图 23.23。Robust LM-Error 统计量是显著的（P＜0.04），但 Robust LM-Lag 统计量并不显著（p=0.26）。这表明应该估计（estimate）一个空间误差 specification。

23.6.3 空间回归模型选择决策规则

空间自相关检验统计量的阵列(array)一下子有些令人眼花缭乱，但有一个相关直觉（intuitive）的方法可以穿过（proceed through）空间自相关 specification 的结果。此过程概括于图 23.24。

[1] 技术细节见 Anselin et al. (1996) and Anselin and Florax (1995)。

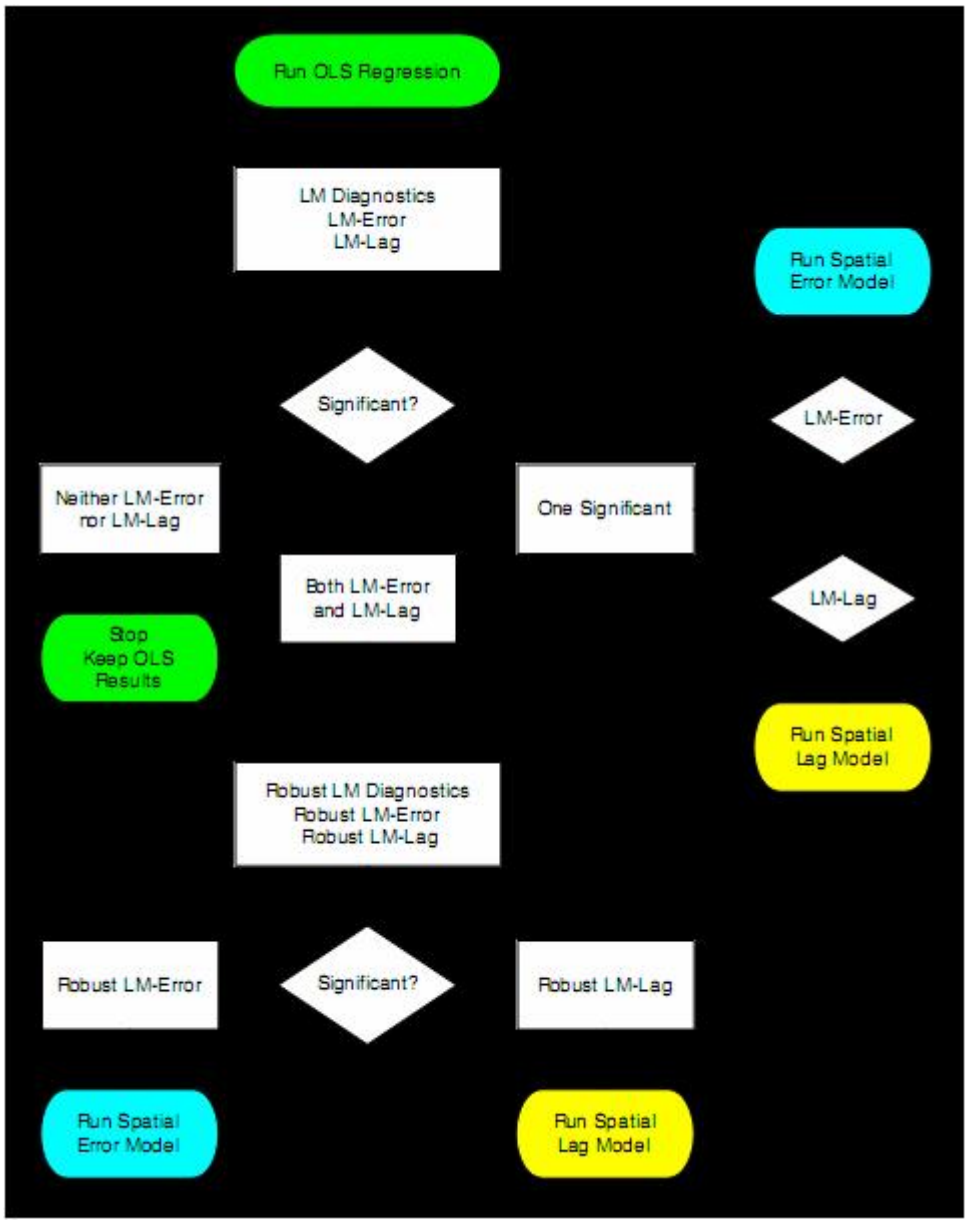


图 23.24　空间回归决策过程

从图表的顶部开始此过程，考虑标准（如不是 robust 形式）LM-Error 和 LM-Lag 检验统计量。如果两者都不能拒绝 0 假设，坚持（stick with）OLS 的结果。在本例是 Moran's I 检验统计量不能拒绝 0 假设也是可能的[1]。如果有一个 LM 检验统计量拒绝了 0 假设，但其它的没有拒绝，也可以直接得出结论：估算与拒绝了 0 假设的检验统计量相对应的另一个空间回归模型。所示，如果 LM-Error 拒绝了 0 假设，但 LM-Lag 没有，估算空间误差模型，反之亦然。

当两个 LM 检验统计量都拒绝了 0 假设，继续到图表的底部，考虑检验统计量的 Robust 形式。特别地，只会有一个是显著的（图 23.23），或一个的重要性比其它的更显著（如 $p<0.00000$ 与 $p<0.03$ 相比）。在这种情况下，容易决定：估算与最显著的统计量相对应的空间回归模型。两个都非常显著的情况是很少见的，估算检验统计量最大值的模型。但是，在这种情况下，需要注意，因为可能存在其它的设定误差源。要采取的一个明显措施就是考虑模型的不同空间权重的结果和/或改变基本（不是空间部分）的 specification。如图 23.22 所示，两个 Robust LM 检验统计量都不显著的情况是

[1] 如果是这样，比如在 Moran's I 和 LM 检验统计量所给的表征之间存在矛盾，可能是由于 Moran's I power against other alternatives than spatial autocorrelation, 如异方差性或非正态。

很少的。这些情况下，可能会出现更严重的设定误差的问题，需要首先 addressed。

23.7 练习

继续使用 22.6 部分的 POLICE 样本数据，考虑使用其它空间权重的诊断，如 rook 邻接，基于距离的邻接等等。将你的结果与 Kelejian and Robinson(1992)的结果相比较。

或者，考虑 Dubin(1992)使用的 BALTIMORE 的模型 specification，将其与简单趋势面的结果相比较。它的模型 specification 用 PRICE 为因变量，以下为解释变量：NROOM，DWELL，NBATH，PATIO，FIREPL，AC，BMENT，NSTOR，GAR，AGE，CITCOU，LOTSZ，SQFT 和二次趋势面[1]。在练习中除了使用 rook 空间权重，也考虑一下基于距离的权重。

[1] Dubin(1992)文章中的变量 STIME，并不在样本数据集中。

练习 24　空间滞后（Lag）模型

24.1 目标

该练习通过空间回归模型最大似然法考虑估计，模型中包括一个空间滞后自变量[1]。有关方法背景请见 Anselin(1998)和 Anselin and Bera（1998）。GeoDa 所用的特定运算法则见 Smirnov and Anselin(2001)。不象传统方法，用权重矩阵的特征值，这种方法非常适合在数据量非常大的情况下使用。

在本练习末，你将知道：

- 为空间滞后回归模型建立 specification
- 解释空间滞后模型中的估计结果
- 解释空间滞后模型中的拟合（fit）度量
- 解释空间滞后模型中的回归诊断
- 理解预测值和空间滞后模型中的残差的不同概念

相关操作的更多详细信息见 Release Notes 第 53－54 页。

24.2 准备工作

载入样本数据 south.shp，关键字为 FIPSNO，它是 U.S.南部 1412 个县杀人案发生率和相关变量的数据。底图如图 24.1。如果你还没有创建空间权重矩阵，一定要创建一个。本次应用中，我们将会用到累积一次和二次 rook 文件。选择 rook 标准，将 order of contiguity 改为 2，选中 Include all the lower orders(见 15.2 部分的指导)。将文件命名为 southrk12.GAL。

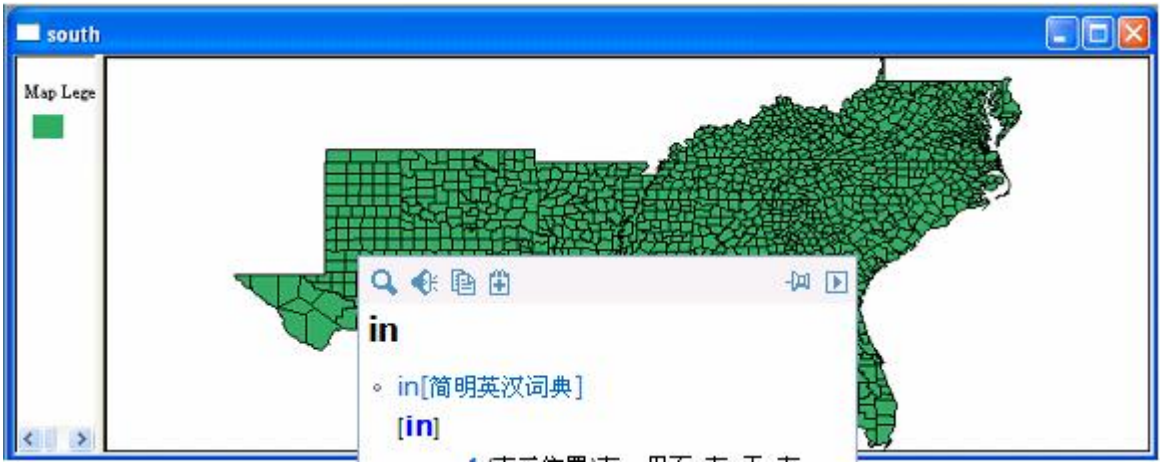


图 24.1　南部县杀人案底图

注意空间滞后模型的 ML 估计只对与对称邻接关系相一致的空间权重有作用。换句话说，只对 rook 和 queen 邻接有作用，还有距离束邻接，但不包括 K-nearest 邻居。

[1] 这个模型是 $y = \rho Wy + X\beta + \varepsilon$，y 是关于自变量观测值的向量，Wy 是权重矩阵 W 的空间滞后自变量，X 是解释变量的观测值的矩阵，ε 是 i.i.d.误差项向量，ρ 和 β 是参数。

24.2.1 OLS（Ordinary Least Squares，普通最小二乘）诊断

作为一个参考点，我们首先运行 Classic OLS 回归（按照 22.3 部分的指导），HR60 为自变量，RD60，PS60，MA60，DV60 和 UE60 为解释变量[1]。回归对话框应如图 24.2 所示。确信选中了 Weight Files 选项，指定 southrk12.GAL 为权重文件。

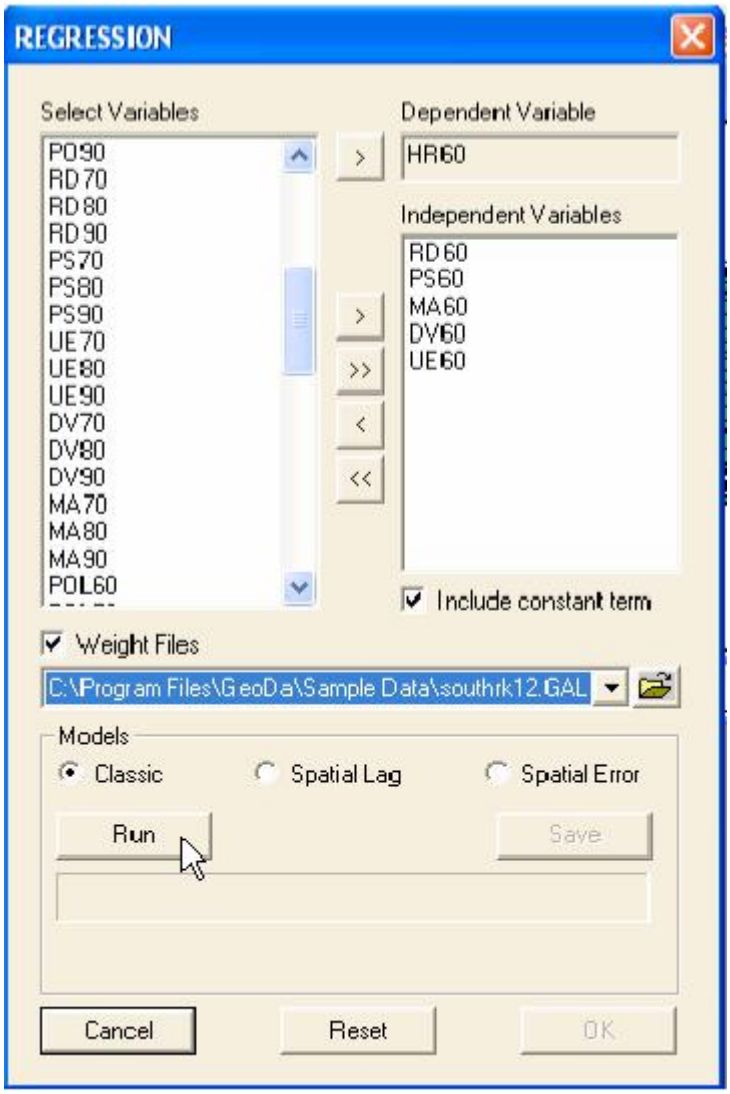


图 24.2　1960 年的杀人案经典回归

运行回归，点击 OK 查看结果。OLS 估计就列在图 24.3 中，诊断列在图 24.4 中。

[1] 复制了 Baller et al. (2001)的分析，但空间权重稍有不同。

```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : south
Dependent Variable  :        HR60  Number of Observations: 1412
Mean dependent var  :     7.29214  Number of Variables   :    6
S.D. dependent var  :     6.41874  Degrees of Freedom    : 1406

R-squared           :    0.103657  F-statistic           :     32.5192
Adjusted R-squared  :    0.100470  Prob(F-statistic)     :1.85631e-031
Sum squared residual:     52144.5  Log likelihood        :     -4551.5
Sigma-square        :     37.0872  Akaike info criterion :     9115.01
S.E. of regression  :     6.08992  Schwarz criterion     :     9146.52
Sigma-square ML     :     36.9296
S.E of regression ML:     6.07697
--------------------------------------------------------------------------
    Variable    Coefficient     Std.Error    t-Statistic   Probability
--------------------------------------------------------------------------
    CONSTANT      13.21547       1.124565       11.75163    0.0000000
        RD60      1.764484       0.198244       8.900568    0.0000000
        PS60      0.299302      0.2142573       1.396928    0.1626563
        MA60    -0.2752095     0.03806419      -7.230141    0.0000000
        DV60      1.179452       0.243517       4.843405    0.0000014
        UE60    -0.2918555     0.07117148      -4.100737    0.0000435
--------------------------------------------------------------------------
```

图 24.3　OLS 估计结果，1960 年杀人案回归

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   18.89912
TEST ON NORMALITY OF ERRORS
TEST                    DF          VALUE            PROB
Jarque-Bera              2           87427.87        0.0000000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF          VALUE            PROB
Breusch-Pagan test       5           599.4759        0.0000000
Koenker-Bassett test     5           30.10693        0.0000141
SPECIFICATION ROBUST TEST
TEST                    DF          VALUE            PROB
White                   20           197.0809        0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : southrk12.GAL  (row-standardized weights)
TEST                            MI/DF       VALUE           PROB
Moran's I (error)             0.136568      N/A             N/A
Lagrange Multiplier (lag)        1       222.5280524       0.0000000
Robust LM (lag)                  1        18.4455725       0.0000175
Lagrange Multiplier (error)      1       205.9505673       0.0000000
Robust LM (error)                1         1.8680875       0.1716943
Lagrange Multiplier (SARMA)      2       224.3961399       0.0000000
========================= END OF REPORT =========================
```

图 24.4　OLS 诊断，1960 年杀人案回归

模型的拟合并不是很理想，调整 R2＝0.10，但除 PS60 外其它都是高度显著，with the expected sign。为了与空间模型相比较，注意 Log Likelihood＝－4551.5，AIC＝115.01。

回归诊断显示出非常大的非正态和异方差性，以及高度空间自相关。按照 23.6.3 部分的步骤，我们可以得出结论：空间滞后模型是合适的替代。LM-Lag 和 LM-Error 都是显著的，但在 robust 形

式中，Robust LM-Lag 统计量是高度显著（p<0.000002），但 Robust LM-Error 统计量并不显著（p<0.17）。这就设定了空间滞后模型估计的阶段（stage）。

24.3 ML 估计诊断

空间滞后模型的 ML 估计以与回归相同的方式调用，从主菜单点击 Regress，或在载入一个项目前选择 Methods>Regress。标题和输出对话框与所有回归分析相同。要求指定输出文件名，如 southlag.rtf，如图 24.5 所示。点击 OK，产生熟悉的回归 specification 对话框。

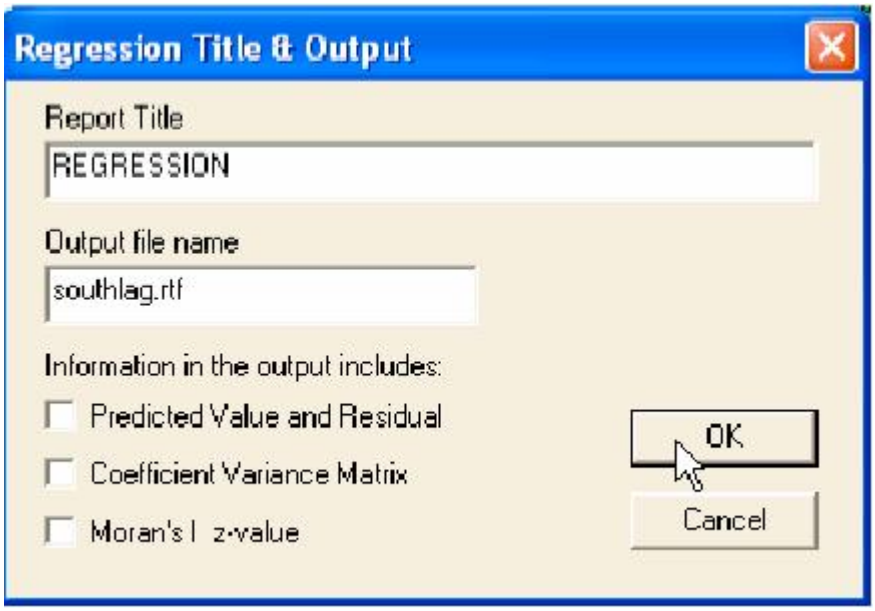


图 24.5　空间滞后回归标题和文件对话框

24.3.1 模型规范（specification）

图 24.6 中的对话框与前面的相同。输入与经典回归相同的自变量和解释变量，确定指定了空间权重文件，如图所示。不选默认的 Classic，而是选中单选项 Spatial Lag。调用如前的估计规则，点击 Run 按钮。

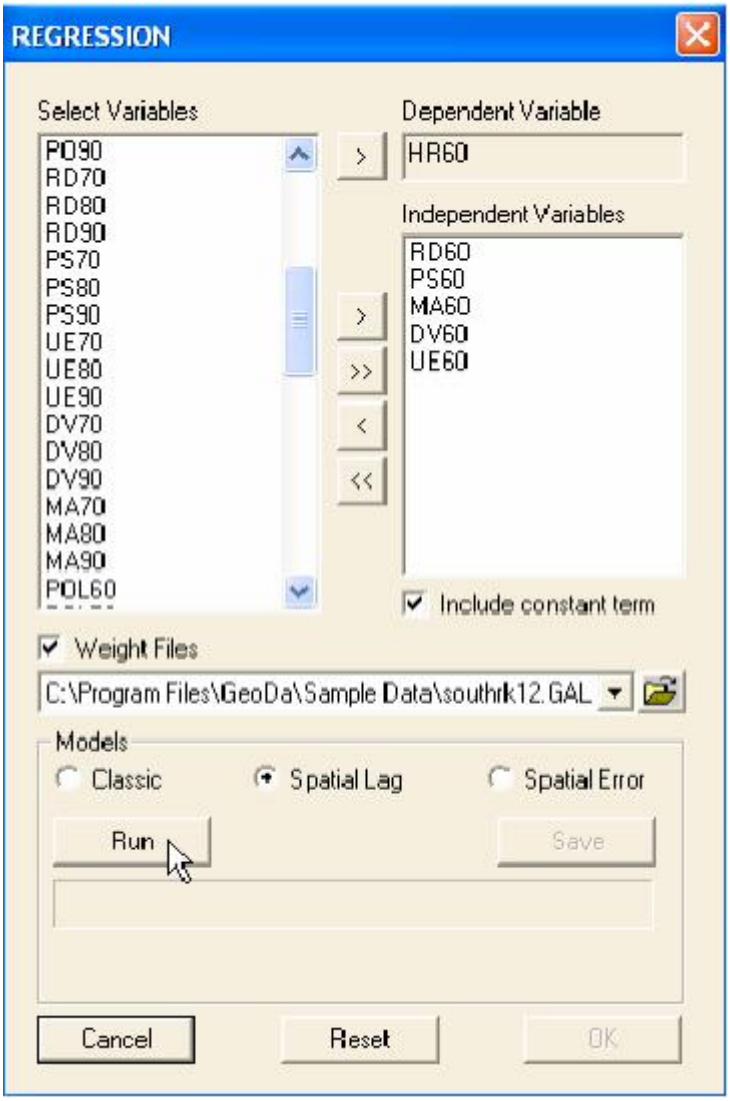


图 24.6　1960 年杀人案空间滞后回归 specification

估计完成后，图 24.7 所示及进度条显示，确定在选择 OK 前点击 Save 按钮。会产生一个指定变量名的对话框，将残差和预测值添加到数据表中。

如图 24.8 所示，在空间滞后模型中有三个选项。这些将在 24.4 部分来详细论述。现在，选择所有三个选项，保留变量默认的变量名，Predicted Value 名为 LAG_PREDIC，Prediction Error 名为 LAG_PREDERR，Residual 名为 LAG_RESIDU。点击 OK 回到回归，再选择 OK 产生估计结果和诊断。

24.3.2 估计结果

估计与拟合度列在图 24.9 中。First, a word of caution. 当它试图关注传统的度量时，如 R2，其在空间回归模型中并不适用。列在空间滞后输出中的值并不是一个真正的 R2，而是一个伪 R2，其不能直接与 OLS 结果所给的度量相比较。

合适的拟合度是 Log-Likelihood，AIC 和 SC。如果我们比较图 24.9 和图 24.3 中的值，我们会注意到 Log-Likelihood 由－4551.5（OLS）增长到－4488.97。弥补了新增变量的改进拟合度（空间滞后自变量），AIC（从 9115 到 8991.9）和 SC（从 9146.5 到 9028.7）都相对于 OLS 下降了，又一次表明空间滞后规范拟合度的增加。空间回归系数大约为 0.53，高度显著（p<0.0000000）。这对于象这样大小的数据集而言是不同寻常的，部分是由于方差所有的分析表达式（analytical expressions）的渐近的本性。

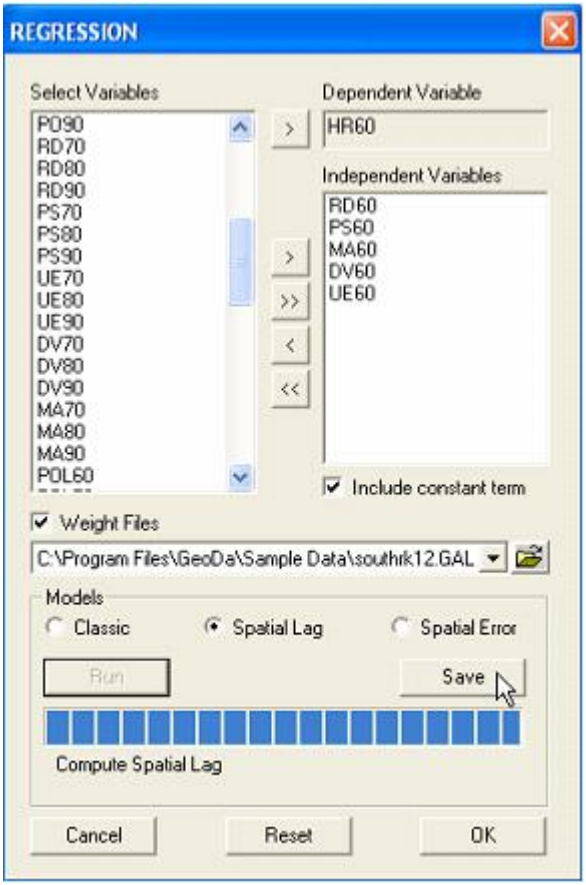


图 24.7　保存残差和预测值对话框

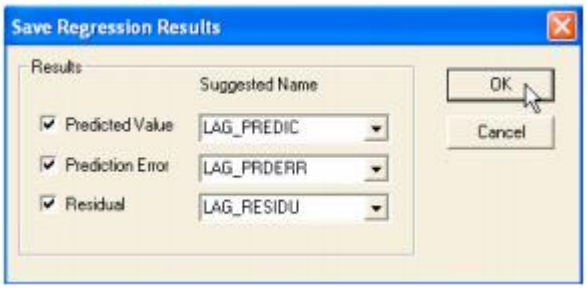


图 24.8　空间滞后预测值和残差变量名窗口

```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set              : south
Spatial Weight        : southrk12.GAL
Dependent Variable    :        HR60  Number of Observations: 1412
Mean dependent var    :     7.29214  Number of Variables   :    7
S.D. dependent var    :     6.41874  Degrees of Freedom    : 1405
Lag coeff.    (Rho)   :    0.532889

R-squared             :    0.197931  Log likelihood        :    -4488.97
Sq. Correlation       : -            Akaike info criterion :     8991.93
Sigma-square          :     33.0455  Schwarz criterion     :      9028.7
S.E of regression     :     5.74852

-----------------------------------------------------------------------
    Variable     Coefficient      Std.Error     z-value      Probability
-----------------------------------------------------------------------
      W_HR60       0.5328888     0.04566825       11.66869     0.0000000
    CONSTANT        6.574962       1.172724       5.606573     0.0000000
        RD60        1.100473      0.1963386       5.604976     0.0000000
        PS60     0.03791171      0.2026779       0.187054     0.8516183
        MA60     -0.1752564     0.03671206      -4.773809     0.0000018
        DV60      0.9352081      0.2303864       4.059302     0.0000492
        UE60     -0.1326599     0.06735334      -1.969612     0.0488827
-----------------------------------------------------------------------
```

图 24.9　ML 估计结果，空间滞后模型，HR60

空间滞后模型与经典规范的其它回归系数的显著性存在一些小的差异：PS60 显著性小于前面（p<0.85），更重要，UE60 的显著性从 p<0.00004 变化到 p<0.04。所有估计系数的重要性都受到影响，全部表现出绝对值的下降。在某种程度上，这些变量的解释能力真正是由于邻近位置。这是通过空间滞后自相关变量获得的。

24.3.3 诊断

只有有限的 ML 滞后估计的诊断被列出，如图 24.10 所示。第一是误差项的异方差性 Breusch-Pagan 检验。高度显著性的值为 697.9，表明异方差性仍然是一个严重的问题。第二个检验是空间自回归系数的渐近显著性检验的替代，它并不是关于保留空间自相关的检验。似然比例检验是三个经典 specification 检验之一，将 0 模型（经典回归 specification）与空间滞后模型[1]相比。125 的值确认了空间回归系数的强显著性。

```
REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE          PROB
Breusch-Pagan test                      5        697.9206      0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : southrk12.GAL
TEST                                    DF      VALUE          PROB
Likelihood Ratio Test                   1        125.0736      0.0000000
========================= END OF REPORT ==============================
```

图 24.10　诊断，空间滞后模型，HR60

这三个经典检验是近似相等的，但在有限的例子中的顺序是 W＞LR＞LM。在我们的例子中，Ward 检验 11.72＝136.9（大约），LR 检验是 125，LM-Lag 检验是 222.5，与期望的顺序并不相同。这可能表明其它的设定误差源可能使 ML 估算和检验统计的渐近属性失效。假如出现的模型拟合度相当差，高度的非正态和强的异方差性，这也并不奇怪。需要进一步的考虑，包括新的解释变量或包含的不同空间权重。

24.4 预测值和残差

在空间滞后模型中，必须区别用于进一步诊断检查的模型残差和预测误差。后者是观测值与预测值之差，是只考虑外因变量而得到的[2]。

这些结果之间的不同如图 24.11 和 24.12 所示。

[1] 另外两个经典检验是 Wald 检验，如渐近 t 值平方（或 z 值），及基于 OLS 残差的 LM-Lag 检验。

[2] 残差是模型误差项的估计，$(I-\hat{\rho}W)y-X\hat{\beta}$。预测值是 $\hat{y}=(I-\hat{\rho}W)^{-1}X\hat{\beta}$，预测误差是 $y-\hat{y}$。

	HR60
1	1.682864
2	4.607233
3	0.974132
4	0.876248
5	4.228385

图 24.11　观测值，HR60

LAG_RESIDU	LAG_PREDIC	LAG_PRDERR
-1.544858	3.261497	-1.578633
2.921503	2.848264	1.758969
-1.483714	2.859864	-1.885732
-2.150745	3.956595	-3.080347
-0.132526	3.877435	0.350950

图 24.12　空间滞后预测值和残差 HR60

为进一步显示其差异，为两个残差 LAG_RESIDU 和 LAG_PRDERR，创建一幅 Moran 散点图，权重文件为 southrk12.GAL（从菜单选择 Space>Univariate Moran，指定变量和权重文件）。结果将如图 24.13 和 24.14。

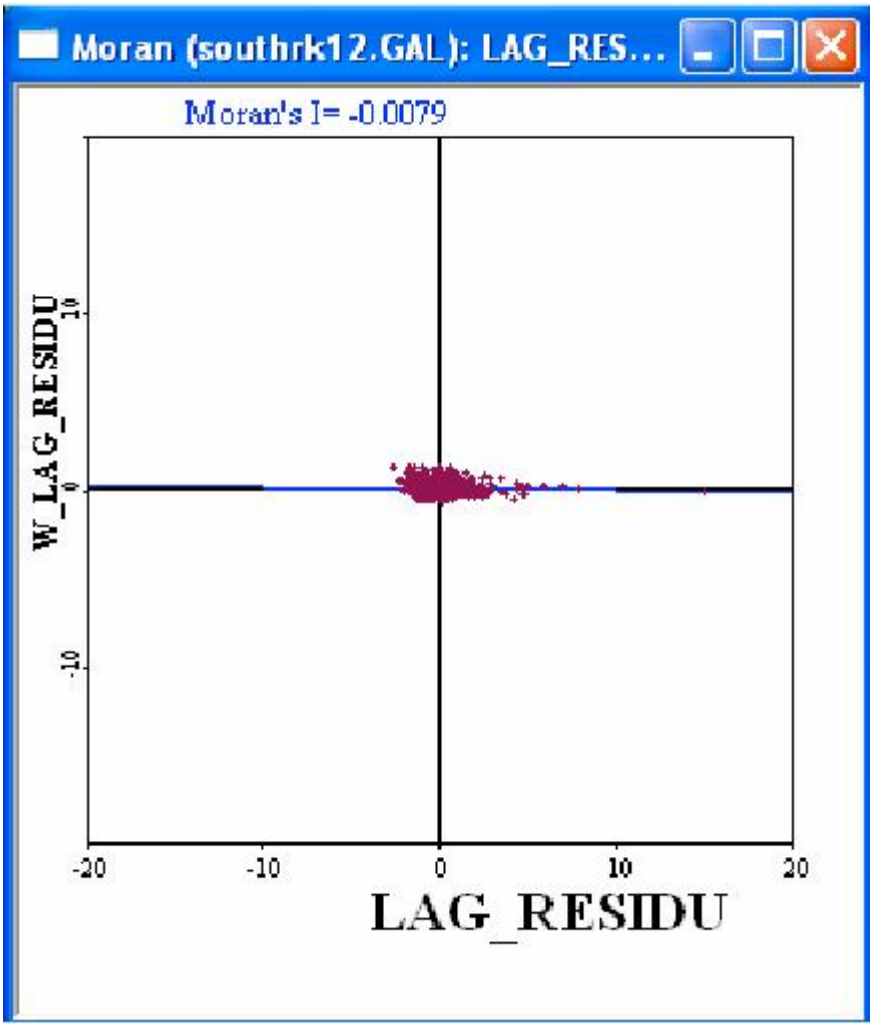


图 24.13　空间滞后残差的 Moran 散点图，HR60

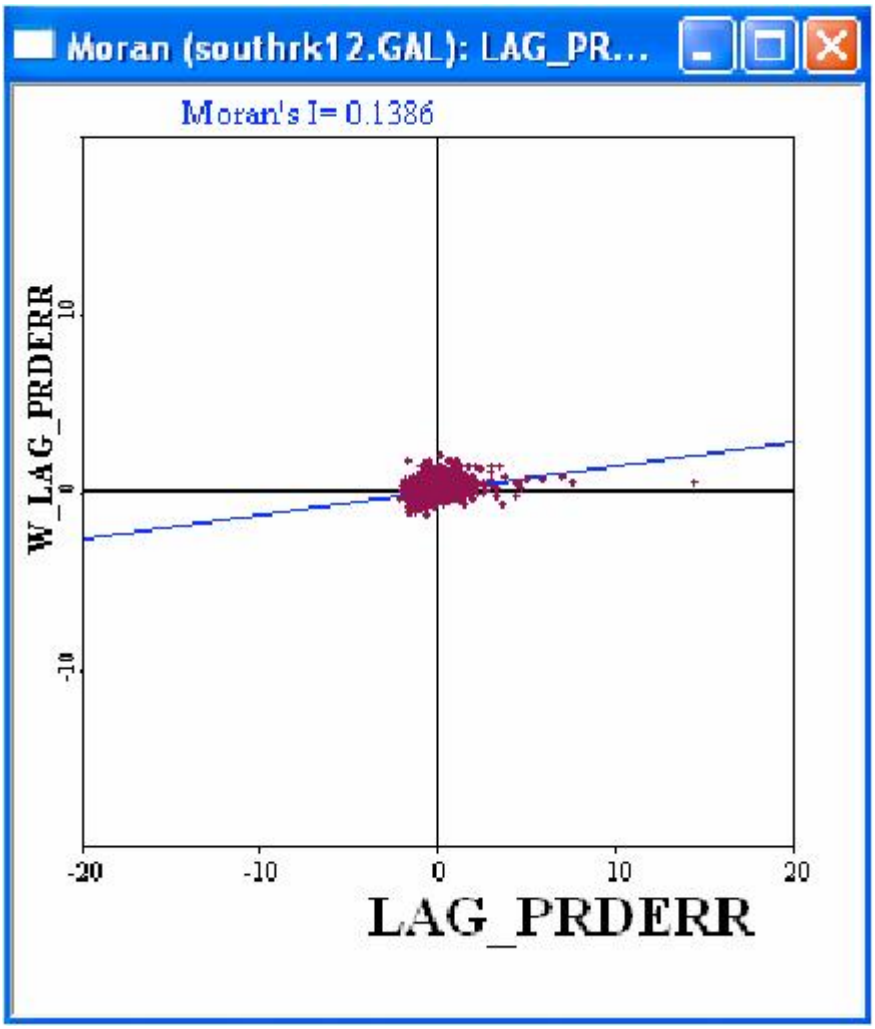


图 24.14　空间滞后预测误差散点图，HR60

对于 LAG_RESIDU，Morans I 统计量是-0.0079，几乎为零。这表明包括在模型中的空间滞后自变量项已经排除了空间自相关，如其应当的情况[1]。

作为比较，LAG_PRDERR 统计量为 0.1386，与原始 OLS 残差几乎相同。乍一看，这可能有点问题，但实际上，正如所预料的。预测误差是一个$(I-\rho W)^{-1}u$ 估计，或空间转换误差。结果，它们被建成为空间相关的。

24.5 练习

用 ML 为相同的模型和数据集估计空间滞后模型，但现在用一次 rook 空间权重文件。将结果与练习中的结果相比较。还有，查看 W，LR 和 LM 检验统计量。你所用的 specification 能够表示出什么意思？

或者，对你已经运行过的 OLS 回归，或不同年份的杀人案模型，查看诊断表明的滞后模型的位置。运行 ML 估计，将经典回归的结果与空间模型的结果进行比较。

[1] 注意这并不是一个正式的假设检验，但只是一个描述性统计。使用残差的序列方法并不适当。

练习 25　空间误差模型

25.1 目标

该练习通过空间回归模型的最大似然方法考虑估计，在模型中包括了空间自回归误差项。如练习 24，方法方面的论述见 Anselin(1998)和 Anselin and Bera(1998)。相同的运算法则也用于空间滞后模型，详见 Smirnov and Anselin(2001)。

在本练习末，你将知道：

- 为空间误差回归模型建立 specification
- 解释空间误差模型中的估计结果
- 解释空间误差模型中的拟合度
- 解释空间误差模型中的回归诊断
- 理解空间误差模型中预测值和残差的不同概念
- 将空间误差模型的结果与空间滞后模型的结果相比较

相关操作的更多详细信息请参考 Release Notes，第 55－56 页。

25.2 准备工作

我们继续使用前面练习中的样本数据，U.S.南部 1412 个县的杀人案发生率及相关变量（south.shp，关键字为 FIPSNO）。底图如图 24.1。如果你还没有创建权重文件，确保创建一个一次 rook 邻接空间权重矩阵，文件名为 southrk.GAL(详细指导见 15.2 部分)。

对于空间滞后模型的 ML 估计，这种方法只在被选择的权重文件对应于对称邻接关系时对空间误差起作用。它只对 rook 和 queen 邻接，及距离束邻接起作用，但对 k-nearest 邻居无作用。

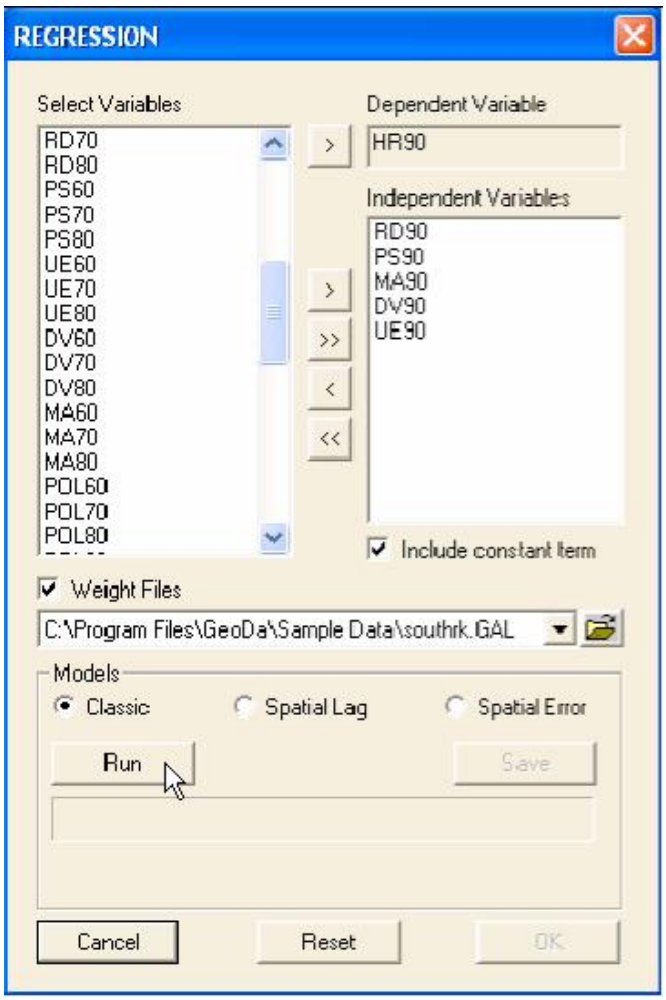
25.2.1 OLS 诊断

作为一个参考点，我们首先运行 Classic　OLS 回归（按照 22.3 部分的指导），但现在是 1990 年的杀人案发生率。指定 HR90 为自变量，解释变量为 RD90，PS90，MA90，DV90T　UE90。回归窗口将如图 25.1。确保选中了 Weight Files 选项，指定 southrk.GAL 为空间权重文件。

运行回归，点击 OK 查看结果。OLS 回归结果列在图 25.2 中，诊断列在图 25.3 中。

模型的拟合优于 1960 年，调整 R2＝0.31。系数的显著性也有些不同，PS90 现在非常显著且为正，但 MA90 并不显著。如前，为比较空间模型，注意 Log Likelihood=-4497.37，AIC=9006.74。

回归诊断显示出相当大的非正态性和异方差性，以及高度空间自相关。按照 23.6.3 部分的步骤，空间误差模型是一个合适的替代。LM-Lag 和 LM-Error 都是显著的。Robust 形式中，Robust LM-Error 统计量显著（p<0.02），但 Robust LM-Lag 统计量明显是不显著的（p<0.12）。这设定了空间误差估计的阶段。



25.1 1990 年杀人案经典回归

```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : south
Dependent Variable  :        HR90  Number of Observations: 1412
Mean dependent var  :     9.54929  Number of Variables   :    6
S.D. dependent var  :     7.03636  Degrees of Freedom    : 1406

R-squared           :    0.309157  F-statistic           :     125.839
Adjusted R-squared  :    0.306701  Prob(F-statistic)     :           0
Sum squared residual:     48295.8  Log likelihood        :    -4497.37
Sigma-square        :     34.3498  Akaike info criterion :     9006.74
S.E. of regression  :     5.86087  Schwarz criterion     :     9038.26
Sigma-square ML     :     34.2038
S.E of regression ML:     5.84841
-----------------------------------------------------------------------
    Variable     Coefficient      Std.Error     t-Statistic   Probability
-----------------------------------------------------------------------
    CONSTANT          8.9625       1.781333       5.031343     0.0000005
        RD90        4.587779      0.2145695       21.38132     0.0000000
        PS90        1.955893      0.2054007       9.522333     0.0000000
        MA90     -0.04948176     0.04890142      -1.011867     0.3117676
        DV90       0.4615939      0.1151724       4.007853     0.0000645
        UE90      -0.5244021     0.07002751      -7.488515     0.0000000
-----------------------------------------------------------------------
```

25.2　1990 年杀人案回归的 OLS 估计结果

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   30.86322
TEST ON NORMALITY OF ERRORS
TEST                    DF          VALUE            PROB
Jarque-Bera              2           2833.409        0.0000000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF          VALUE            PROB
Breusch-Pagan test       5           515.0796        0.0000000
Koenker-Bassett test     5           124.2749        0.0000000
SPECIFICATION ROBUST TEST
TEST                    DF          VALUE            PROB
White                   20           242.8053        0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : southrk.GAL  (row-standardized weights)
TEST                            MI/DF       VALUE           PROB
Moran's I (error)             0.121889      N/A             N/A
Lagrange Multiplier (lag)        1        50.5794169       0.0000000
Robust LM (lag)                  1         2.4456884       0.1178482
Lagrange Multiplier (error)      1        53.6507800       0.0000000
Robust LM (error)                1         5.5170515       0.0188320
Lagrange Multiplier (SARMA)      2        56.0964683       0.0000000
========================= END OF REPORT =========================
```

图 25.3　1990 年杀人案回归 OLS 诊断

25.3 ML 估计诊断

空间误差模型的 ML 估计的调用方法与经典回归和空间滞后模型（24.3）部分是相同的。在主菜单点击 Regress，或在载入一个项目前选择 Methods>Regress。标题和输出文件对话框与所有回归分析窗口是完全相同的。如前，这个窗口要求指定输出文件名，如 southerr.rtf。点击 OK，出现熟悉的回归 specification 对话框。

25.3.1 模型 specification

图 25.4 中所示的对话框与前面的相同。输入与经典回归相同的自变量和解释变量，确保指定了空间权重文件，如图所示。选中 Spatial Error 选项，而不是 Classic。通过点击 Run 调用前面提到的估计程序（routine）。

如前，在估计完成之后，由进度条显示，确保在选择 OK 前点击 Save 按钮。这会出现指定添加到数据表的残差和预测值中的变量名的对话框。

如图 25.5 所示，在空间误差模型中有三个选项。这些将在 25.4 中详细讲解。现在，选择所有的三个选项，保留变量的默认文件名，Predicted Value 为 ERR_PREDIC，Prediction Error 为 ERR_PREDRER，Residual 为 ERR_RESIDU。点击 OK 回到回归窗口，再选择 OK 出现估算结果和诊断。

25.3.2 估计结果

图 25.6 给出了估计和拟合度。ML Lag 估计中，列出的 R2 又是一个伪 R2，与 OLS 结果中所给的度量无法直接对比。合适的拟合度是 Log-Likelihood，AIC 和 SC。如果我们将图 25.6 中的值与图 25.2 中 OLS 的值相比，我们会注意到 Log-Likelihood 从－4497.37（OLS）上升到－4471.32。为补偿增加变量后提高的拟合度（空间滞后自变量），AIC（从 9006.74 到 8954.63）和 SC（从 9038.26

到 8986.15）都相对于 OLS 有所提高，又一次表明空间误差 specification 拟合度的提高。

空间自回归系数是 0.29，高度显著（p<0.0000000）。如在 OLS 例中，MA90 的系数不显著（p<0.75），但其它的系数是显著的。就绝对值，它们稍小于 OLS 的结果，除了 DV90，它的值从 0.462 增加到 0.499。

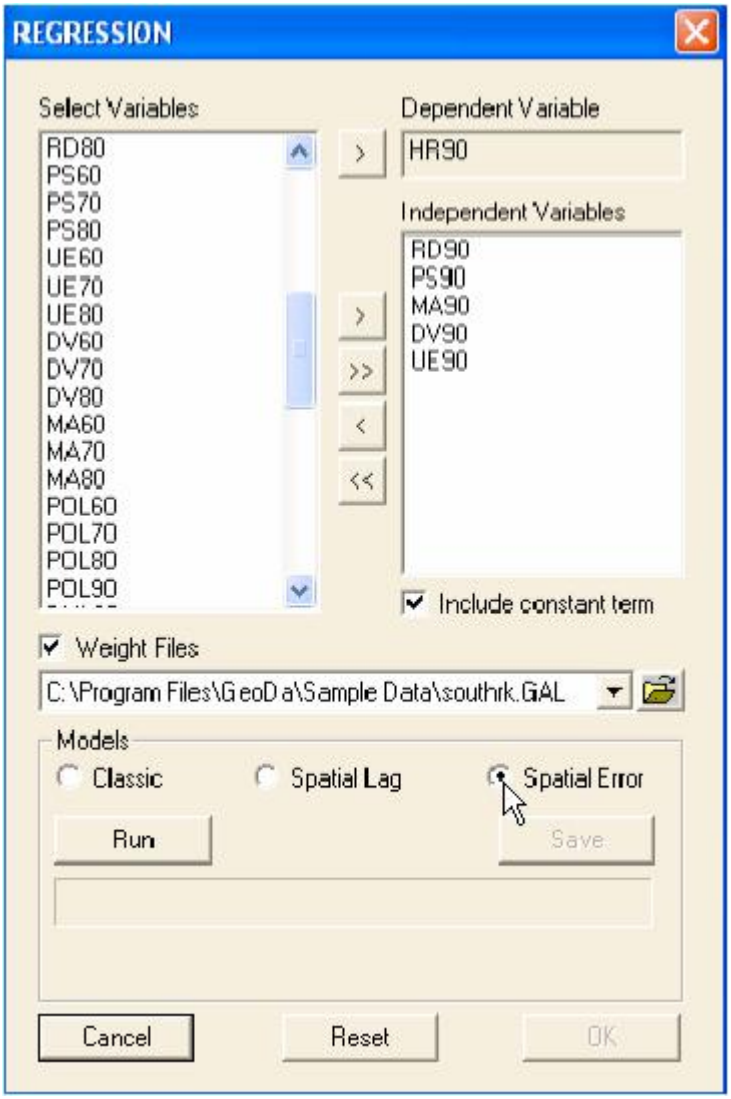


图 25.4　空间误差模型 specification 对话框

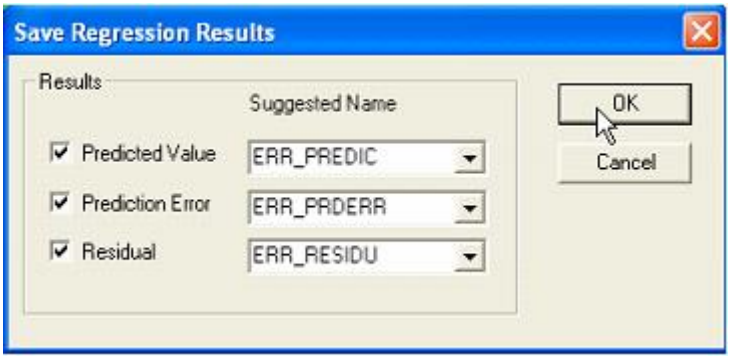


图 25.5　空间误差模型残差和预测值对话框

```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set              : south
Spatial Weight        : southrk.GAL
Dependent Variable    :        HR90  Number of Observations: 1412
Mean dependent var    :    9.549293  Number of Variables   :    6
S.D. dependent var    :    7.036358  Degree of Freedom     : 1406
Lag coeff. (Lambda)   :    0.291609

R-squared             :    0.345458  R-squared (BUSE)      : -
Sq. Correlation       : -            Log likelihood        :-4471.317119
Sigma-square          :   32.406602  Akaike info criterion :     8954.63
S.E of regression     :     5.69268  Schwarz criterion     : 8986.150813

----------------------------------------------------------------------
    Variable     Coefficient      Std.Error     z-value      Probability
----------------------------------------------------------------------
    CONSTANT       6.693515       1.958045      3.418469     0.0006298
        RD90       4.407397       0.237668      18.54434     0.0000000
        PS90       1.766328      0.2256524       7.82765     0.0000000
        MA90    -0.01663971     0.05298999    -0.3140161     0.7535089
        DV90      0.4991464      0.1249123      3.995975     0.0000645
        UE90     -0.3878414     0.07847802     -4.942039     0.0000008
      LAMBDA      0.2916094     0.03727543      7.823098     0.0000000
----------------------------------------------------------------------
```

图 25.6　空间误差模型 ML 估计结果，HR90

25.3.3 诊断

ML 误差估计所报告的两个模型诊断与 Lag specification 是相同的，是异方差性的 Breusch-Pagan 和关于空间自回归系数的 Likelihood Ratio 检验。它们列于图 25.7 中。两个诊断高度显著，表明在模型中的剩余 specification 问题。查看关于空间自回归误差系数 W，LR，LM 统计量的顺序，我们发现 $W=7.82^2=61.2$（图 25.6 中渐近 t 检验的 z 值的平方），LR=52.1，但 LM=53.7（见图 25.2）。如在练习 24 中的滞后模型，这与期望的顺序不同，在这一点上显示了不太令人满意的模型 specification。

```
REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE          PROB
Breusch-Pagan test                      5       549.2706       0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : southrk.GAL
TEST                                    DF      VALUE          PROB
Likelihood Ratio Test                   1       52.10949       0.0000000
========================= END OF REPORT ==============================
```

图 25.7　空间误差模型 ML 诊断，HR90

为进一步比较误差和滞后模型的结果，用与前面相同的 specification，如用 HR90 作为自变量（运行空间滞后模型的指导见 24.3 部分）运行 Spatial Lag 选项。估计结果列于图 25.8。

就系数等级（magnitude）、符号、显著性而言，这些结果与误差模型结果非常相似，进一步强调了区分两个空间模型的困难。误差模型中的 Log Likelihood(-4471)稍优于空间滞后模型（-4474）。相似的，误差模型的 AIC(8954.63)低于滞后模型（8963.84）。

两个模型结果的相似性与剩余 specification 问题的显示，表明优化模型可能会 in order。

```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set              : south
Spatial Weight        : southrk.GAL
Dependent Variable    :        HR90  Number of Observations: 1412
Mean dependent var    :     9.54929  Number of Variables   :    7
S.D. dependent var    :     7.03636  Degrees of Freedom    : 1405
Lag coeff.    (Rho)   :     0.22622

R-squared             :    0.337480  Log likelihood        :    -4474.92
Sq. Correlation       : -            Akaike info criterion :     8963.84
Sigma-square          :     32.8016  Schwarz criterion     :     9000.61
S.E of regression     :     5.72727

------------------------------------------------------------------------
    Variable    Coefficient     Std.Error    z-value      Probability
------------------------------------------------------------------------
      W_HR90     0.2262204      0.0335461      6.743568    0.0000000
    CONSTANT      5.100989       1.793351       2.84439    0.0044498
        RD90      4.030911        0.22555      17.87147    0.0000000
        PS90      1.786308      0.2018053      8.851641    0.0000000
        MA90   -0.01129424     0.04793461    -0.2356177    0.8137294
        DV90     0.4769045      0.1125612      4.236845    0.0000227
        UE90    -0.4393495     0.06870696     -6.394542    0.0000000
------------------------------------------------------------------------
```

图 25.8　空间滞后模型估计结果，HR90

25.4 预测值和残差

如在空间滞后模型中，在空间误差模型中必须用诊断检查和预测误差来区别模型残差。预测误差是观测值与预测值之差，与 y 的条件期望相对应，given X。[1]

这两个结果的不同显示在图 25.9 和 25.10 中。以 HR90 这开头的一列的值是前五个观测点的观测值（y）。ERR_RESIDU 是模型残差（^u），ERR_PREDIC 是预测值（^y），ERR_PRDERR 是预测误差（y－^y）。

	HR90
1	0.946083
2	1.234934
3	2.621009
4	4.461577
5	6.712736

图 25.9　观测值，HR90

[1]

ERR_RESIDU	ERR_PREDIC	ERR_PRDERR
-4.910911	6.856830	-5.910747
-1.545560	4.663617	-3.428683
-6.139540	9.625642	-7.004633
-1.092667	6.966123	-2.504546
-0.902924	8.042464	-1.329728

图 25.10　空间误差预测值和残差，HR90

对于空间滞后模型，为两个残差 ERR_RESIDU 和 ERR_PREDERR 创建散点图，以 southrk.GAL 为权重文件（从菜单选择 Space>Univariate Moran，指定变量和权重文件）。结果将如图 25.11 和 25.22 所示。

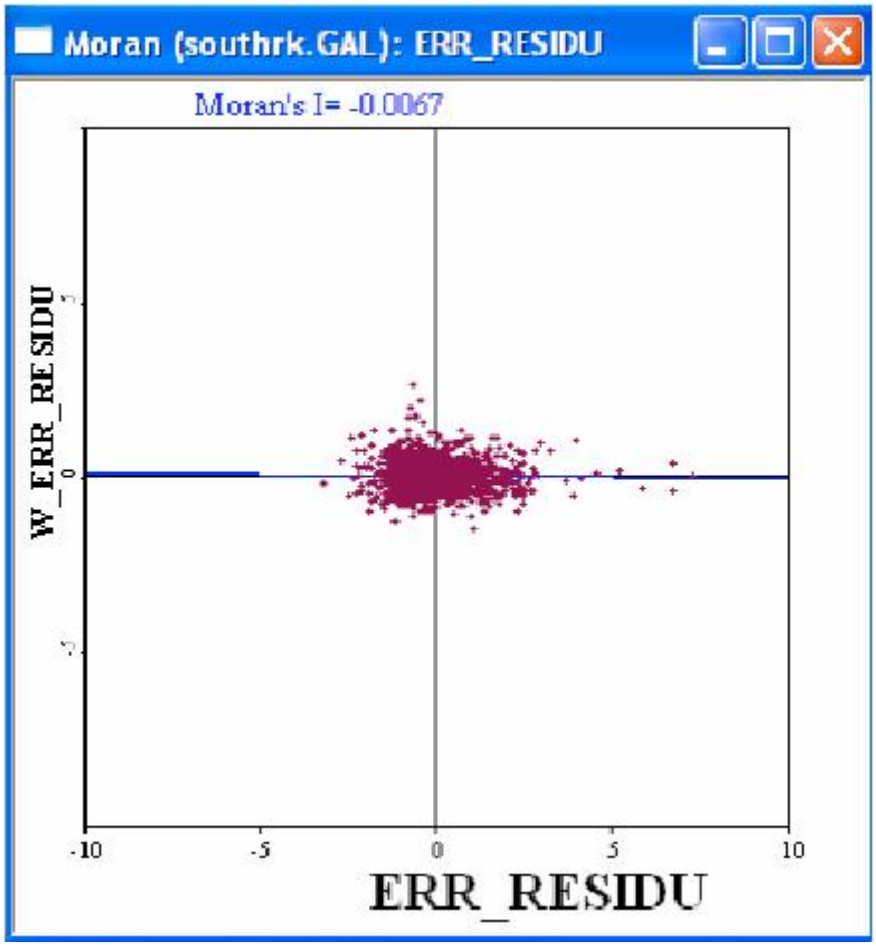


图 25.11　空间误差残差散点图，HR90