

更多内容，请关注



4、地理加权回归

虾神daxialu



1

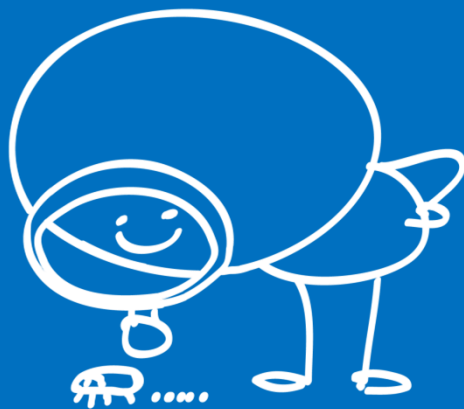
空间异质性

2

回归分析

3

地理加权回归





空间异质性

空间异质性

——地理学第二定律



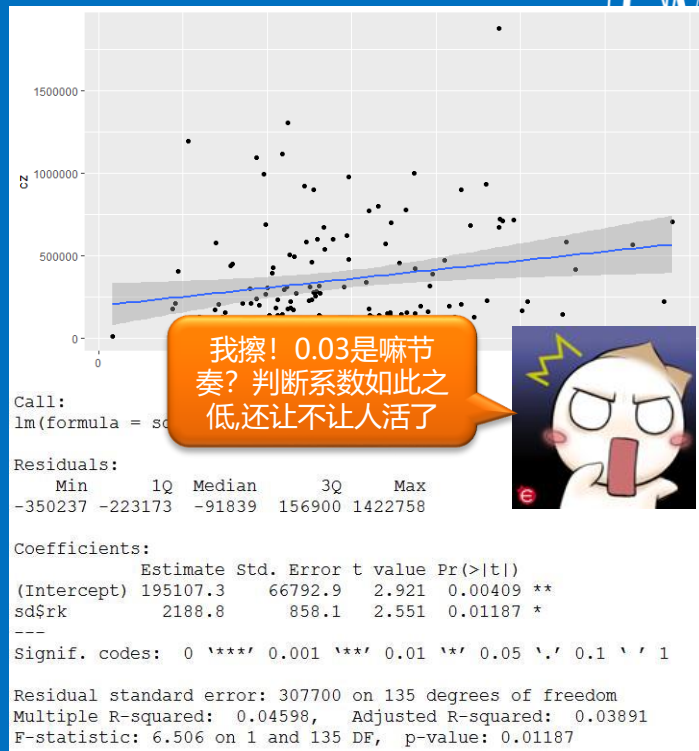
人多力量大



人口和财政收入的关系

- 山东省2015年统计年鉴：
 - 总财政收入：4868.2336亿
 - 总人口：9803万人

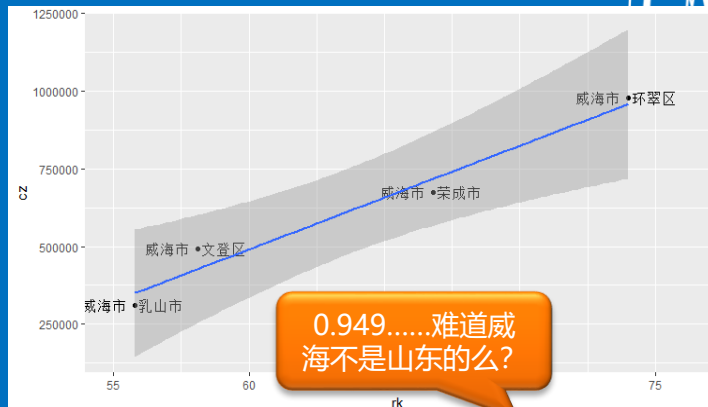
山东全国排名



人口和财政收入的关系

- 威海情况：
 - 总财政收入：245.1717亿
 - 总人口：254万

威海市排名情况



Call:
lm(formula = e_sd\$cz ~ e_sd\$rk)

Residuals:

1	2	3	4
18568	65609	-44440	-39737

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1512273	283578	-5.333	0.0334 *
e_sd\$rk	33376	4425	7.542	0.0171 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

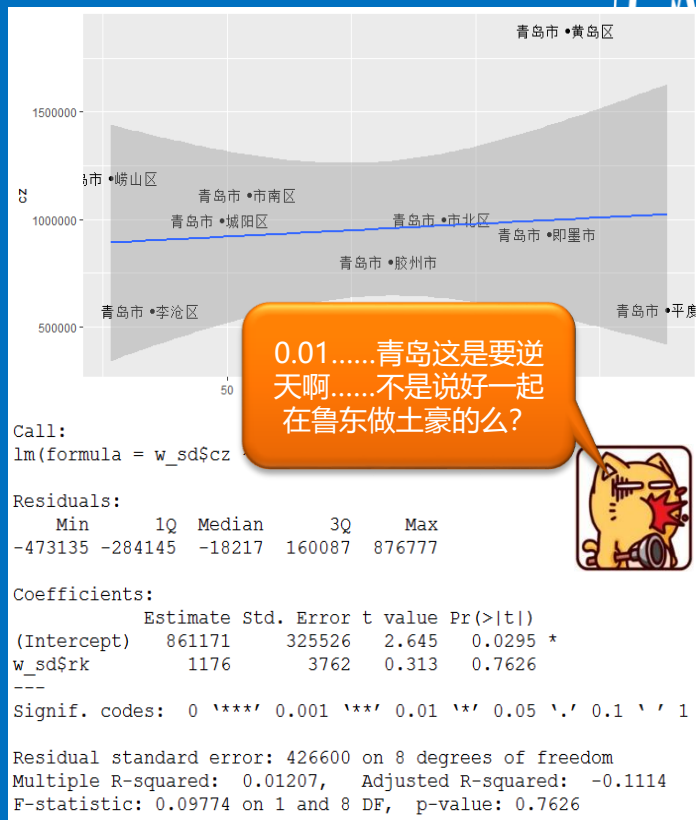
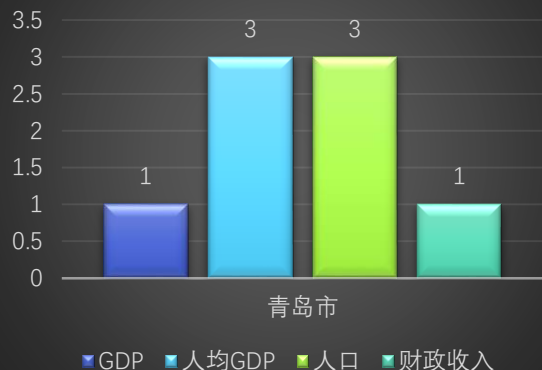
Residual standard error: 64040 on 2 degrees of freedom
Multiple R-squared: 0.966, Adjusted R-squared: 0.9491
F-statistic: 56.89 on 1 and 2 DF, p-value: 0.01713



人口与财政的关系

- 青岛情况：
 - 总财政收入：953.7903亿
 - 总人口：787万

青岛市全省排名情况



需要：因“地”制宜的算法



局部回归

- 空间非平稳性的解决方法之一：把研究区域根据某种指标，划分成若干个同质性的区域，然后分别进行回归。

来，山姆大叔，你不是号称世界第一强国么？我们来比比人均GDP

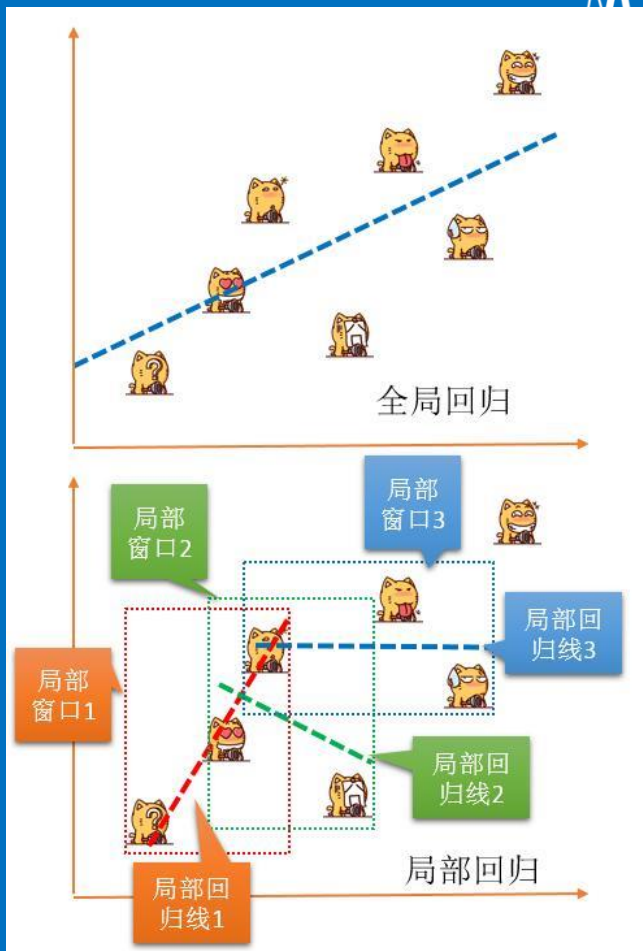


卢森堡
人口：57万

滚！



美国
人口：3.2亿



空间异质性如何衡量？

空间相关性有各种指数.....

- Join Count
- Moran' s I
- Geary's C

空间异质性的衡量方式？

空间异质性的应用模型？

如何计算空间异质性？

空间异质性
在分析中有
何种价值？

.....



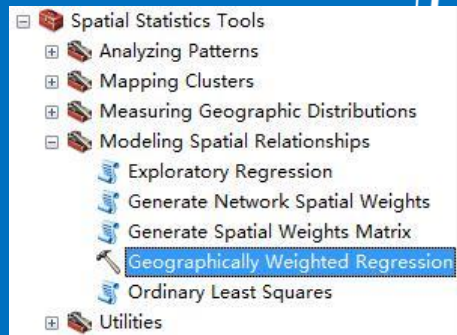
回归分析

——以统计学为世界观最直接的体现



ArcGIS与回归相关的工具集

- 空间统计工具箱
 - 空间关系建模
 - 探索性回归
 - 地理加权回归
 - 最小二乘法



许可信息

ArcGIS Desktop Basic: 需要 Spatial Analyst 或 Geostatistical Analyst

ArcGIS Desktop Standard: 需要 Spatial Analyst 或 Geostatistical Analyst

ArcGIS Desktop Advanced: 是

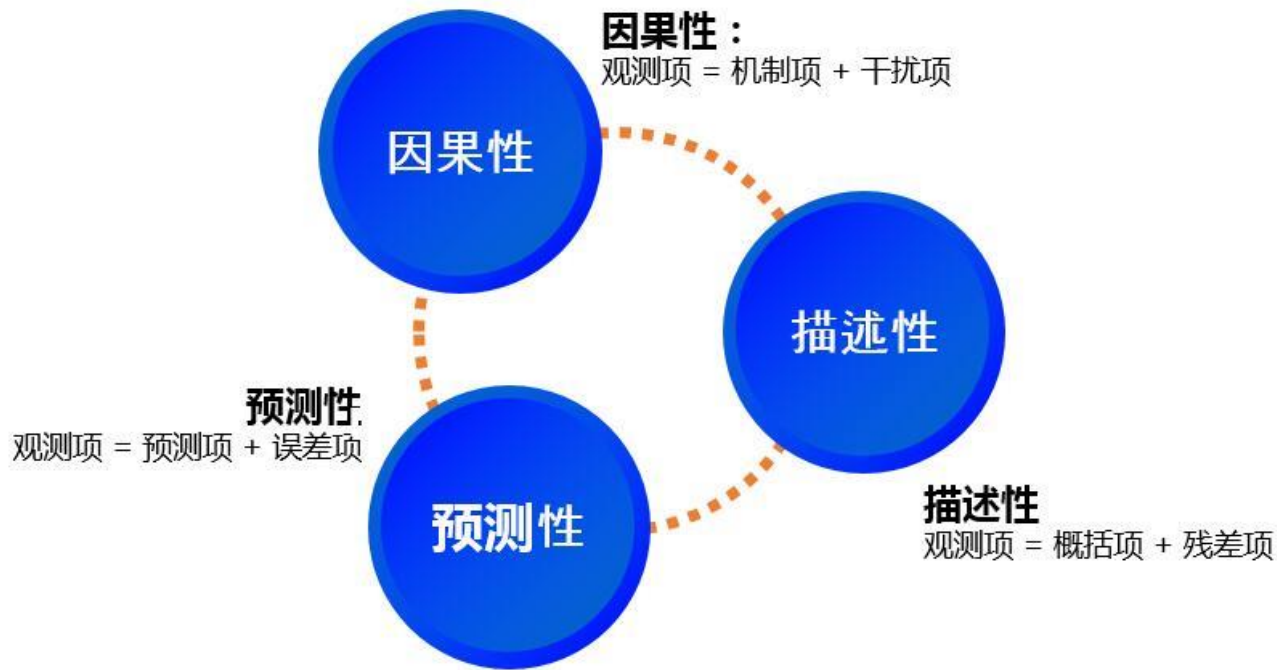
ArcGIS的空间统计工具的几点注意事项:

1. 尽量使用**英文版**
2. 数据最好符合计算机**变量命名法则**。(以字母开头的字母和数字组合)
3. Crack版的用户, 请不要使用ArcGIS 10.3x版本。
4. 360会误杀ArcGIS的一些功能文件。

为什么要做回归?

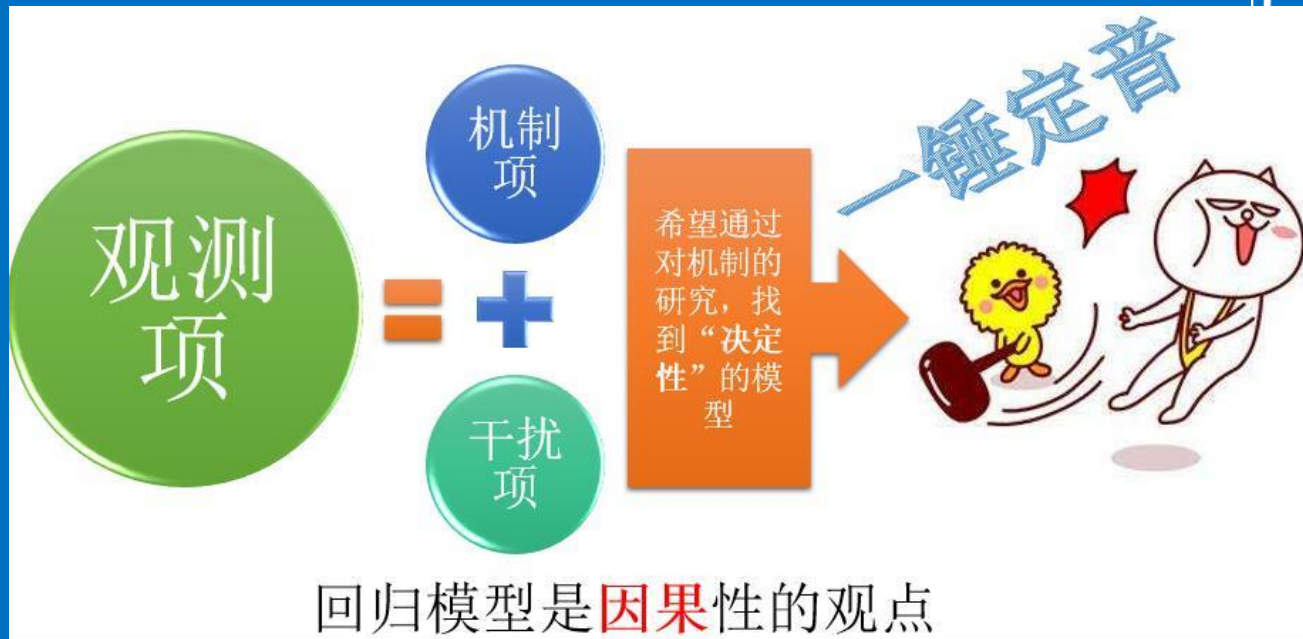


回归的三种思想



偏差：干扰、误差、残差……







回归模型是**预测**性的观点

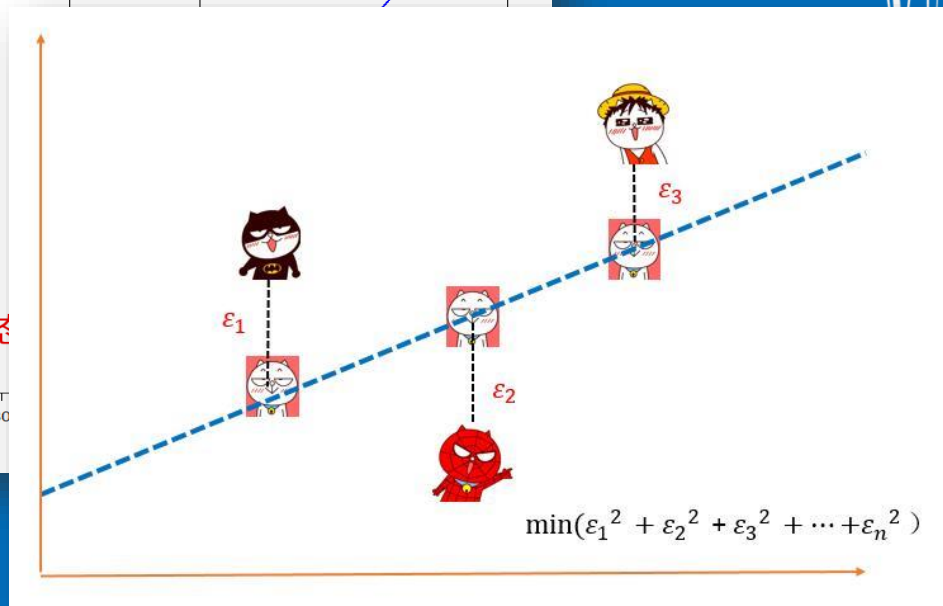
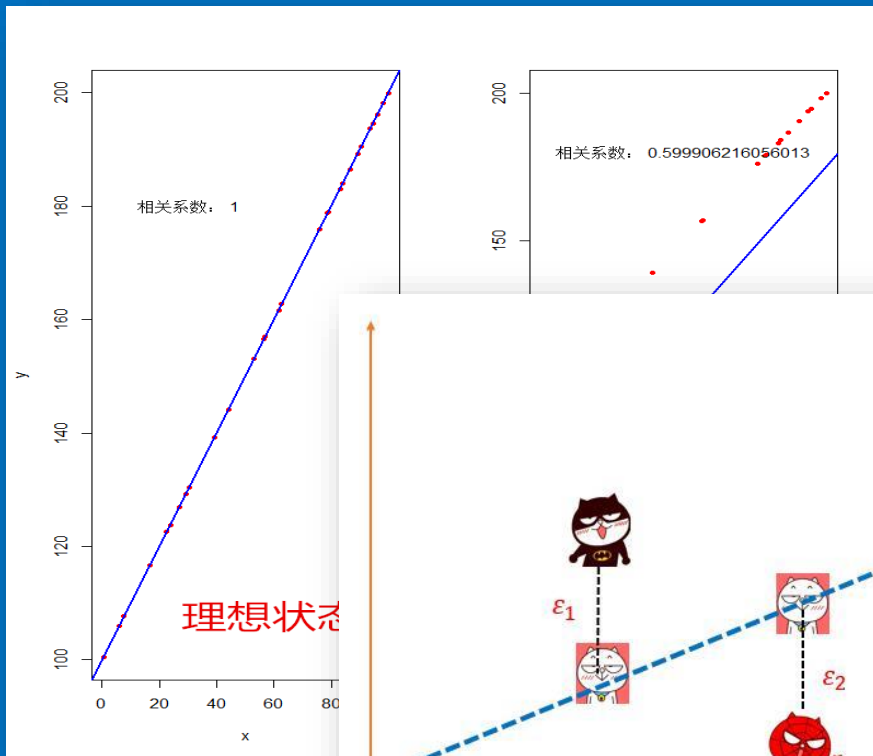




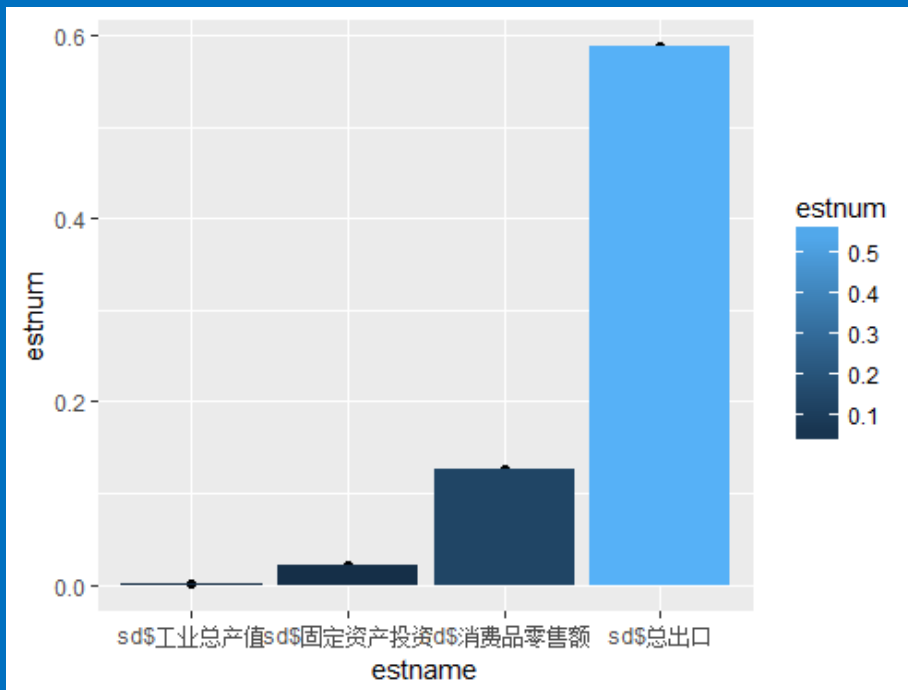
回归模型是描述性的观点



OLS: 最小二乘法回归



OLS分析山东经济数据实例



关键参数解读

Coefficients
贡献度

StdError
标准差

t-Statistic
T统计量

Probability
概率

Robust
稳健性

VIF
方差膨胀因子

R-Squared
R方系数

Joint F-Statistic
联合F统计量

Koenker (BP)
统计量

Jarque-Bera
Statistic
模型偏差评估



Coefficients : 贡献度

- 回归分析的系数代表了每个自变量对因变量的贡献度，系数的绝对值越大，表示该变量在模型里面贡献越大，也表示了该自变量与因变量的关系越紧密。

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5.477e+04	2.332e+04	-2.348	0.0204 *
sdv\$工业总产值	2.227e-03	2.072e-03	1.075	0.2843
sdv\$消费品零售额	1.257e-01	9.472e-03	13.274	< 2e-16 ***
sdv\$总出口	5.870e-01	7.750e-02	7.574	5.61e-12 ***
sdv\$固定资产投资	2.243e-02	9.194e-03	2.440	0.0160 *

- 另外这些系数的值表明了自变量与因变量的关系，比如S（总出口）的系数为0.58，则表示当总出口每增加一个单位，在其他自变量的值不发生改变的时候，因变量财政收入会增加0.58个单位。
- 而且这个系数也表示了自变量与因变量之间的关系类型，即它分为正向和负向，系数为正，表示正相关，系数为负，表示负相关。



StdError：标准差

- 回归的标准误是模型中随机扰动项（误差项）的标准差的估计值。它的平方误差项的方差的无偏估计量，实际上又叫做误差均方

残差的平方和 / (样本容量 - 待估参数的个数)

- 这个值越**小**，表示模型的预测越**准**。



T-Statistic：T统计量

- T统计量是假设检验的重要枢轴量，多用于两样本均值检验，回归模型系数显著性检验。

$$T\text{-Statistic} = \text{平均值} / \text{标准误}$$

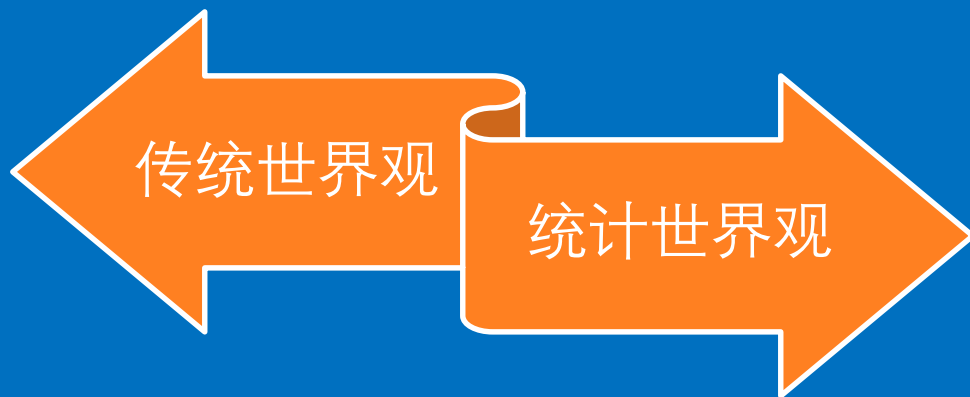
- 一般来来说，这个值表示，与P-value意义差不多，都是在验证零假设的情况下，模型的显著性，但是有些时候P-value会有一些问题，比如丢失一些信息。计算机里面进行统计验证的时候，T统计量越**大**，表示越**显著**。



Probability : 概率

- 概率的推理本身是统计学特有的世界观，只要概率不为零，一切皆可能。

——Stephen M. Stigler 《统计学七支柱》



Robust: 稳健性

- Robust有时候被直接成为鲁棒性
- 通常被称为稳健性检验，一般来说，就是通过修改（增添或者删除）变量值，看所关注解释变量的回归系数和结果是否稳健。
- Robust_SE：标准差的稳健性
- Robust_t：T统计量的稳定性
- Robust_Pr：概率的稳健性



VIF：方差膨胀因子

- 方差膨胀因子（Variance Inflation Factor, VIF）, 主要验证解释变量里面是否有**冗余变量**（即是否存在多重共线性）。一般来说, 只要VIF超过**7.5**, 就表示该变量有可能是冗余变量。



R-Squared : R方系数

- Multiple R-Squared: 多重R平方系数
- Adjusted R-Squared: 校正R平方系数
- 这个名词的解释在不同的书里面不同, 有的叫做“决定性系数”, 也有的直接叫“实际值和预测值之间的相关系数的平方”。
- 该术语用于衡量整个回归模型的性能, 通常它会与 Adjusted R-Squared (校正R平方系数) 一起用。
- 这两个系数的取值, 都在**0-1**之间, 可以转换为百分数, 通常指的是**自变量方程对因变量的解释能力**。比如等于0.8的时候, 表示回归方程能够解释80%的因变量的变化。
- 校正R平方系数, 通常要比多重R平方系数要**稍微低**一些, 因为这个系数的技术与数据的情况关系更强, 所以对模型的性能评估也更加准确一些。



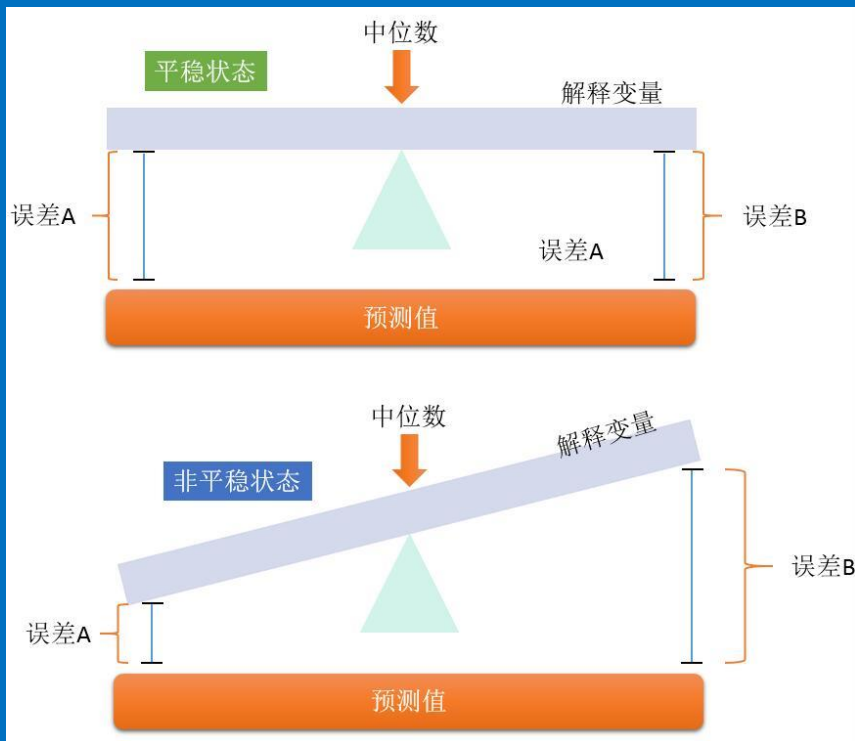
Joint F-Statistic: 联合F统计量

- Prob(>F) degrees of freedom F统计量的可信概率的自由度
- Joint Wald Statistic 联合卡方统计量
- Prob(>chi-squared) degrees of freedom 卡方统计量的可信概率的自由度
- 联合 F 统计量和联合卡方统计量均用于检验整个模型的统计显著性。



Koenker (BP) 统计量

- 这个统计量主要用于确定模型所使用的解释变量是否在位置空间和数据空间中都与因变量具有一致性（稳定性）。



Koenker (BP) 统计量与联合F统计量

- 当Koenker (BP) 统计量具有显著性的时候，联合卡方统计量决定模型的显著性。如果Koenker (BP) 统计量不具有显著性的时候，联合F统计量才有可信性：



Jarque-Bera Statistic: 模型偏差评估

- Jarque-Bera 统计量用于表示模型的残差（已观测/已知的因变量值 - 预测/估计值）是否呈现正态分布。P值表示了模型的残差是不是正态分布。



- 如果发现模型的残差非正态，则表示模型可能出现了偏差。



地理加权回归

——天若不生GWR，回归万古如长夜



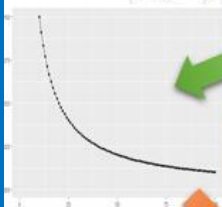
Stewart Fotheringham

- 姓名：Stewart Fotheringham
- 国籍：英裔美籍
- 简历：
 - 1976年毕业于英国阿伯丁大学地理系
 - 1980年获得加拿大McMaster大学博士学位
 - 1988年34岁时就成为美国纽约州立大学Buffalo分校(SUNY Buffalo)地理系正教授，专攻GIS与空间分析
 - 1994年开始发表一批有关GWR的研究论文
 - 2013年4月30日，Fotheringham教授当选美国科学院院士



GWR基本原理

① 在一个范围内，利用每个要素的位置，逐点测量空间距离，然后利用这个距离计算出一个连续的衰减函数。



② 利用这个衰减函数，带入并且计算每个要素值在局部回归方程里面的权重

$$\beta_k(u_i, v_i)$$

③ 得出最后的加权回归方程

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i$$

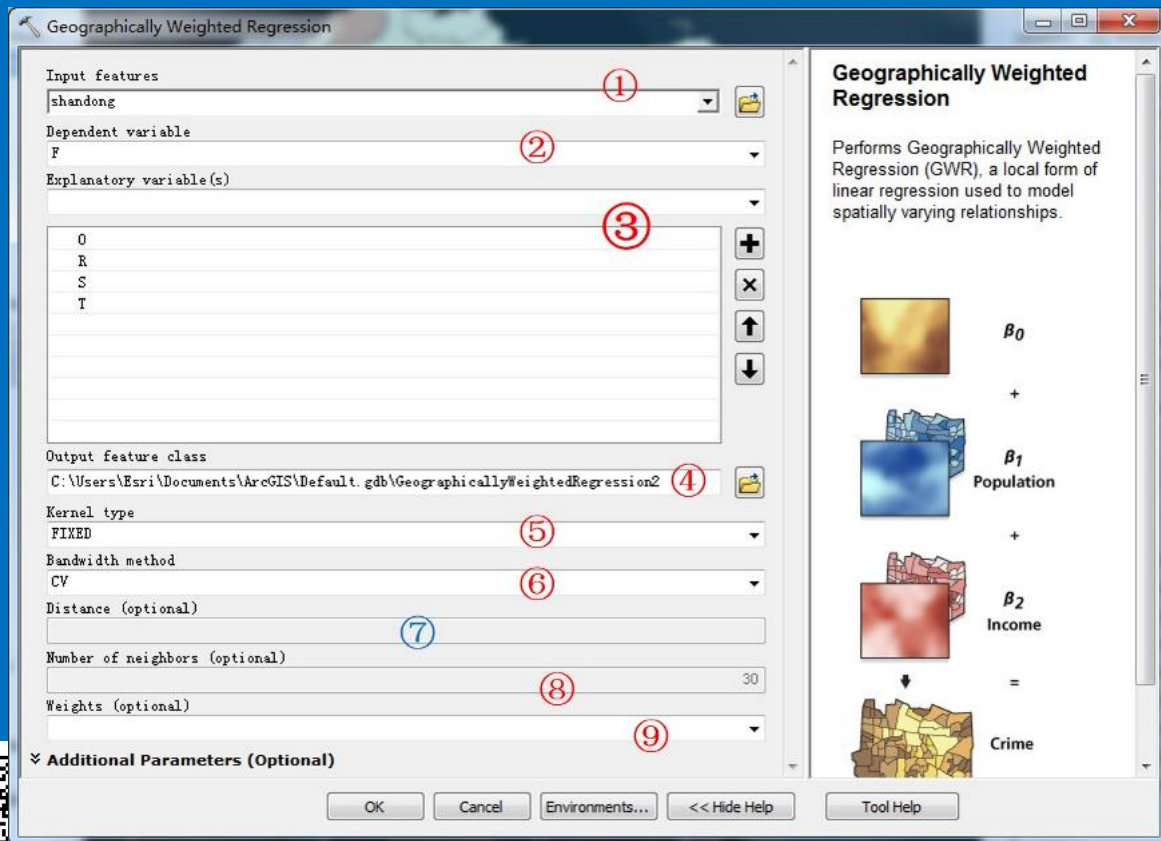




空间关系



ArcGIS的GWR工具



输入数据处理的提示

- 必须有用空间属性。
 - 避免出现只有属性没有空间信息的要素
- 属性数据必须要完整。
 - 避免出现空值。
 - 避免出现二值化数据。
- ArcGIS的GWR不支持同时回归多个图层
 - 将要分析的数据，合并为一个图层。
- 面要素和线要素，将使用质心作为空间位置。



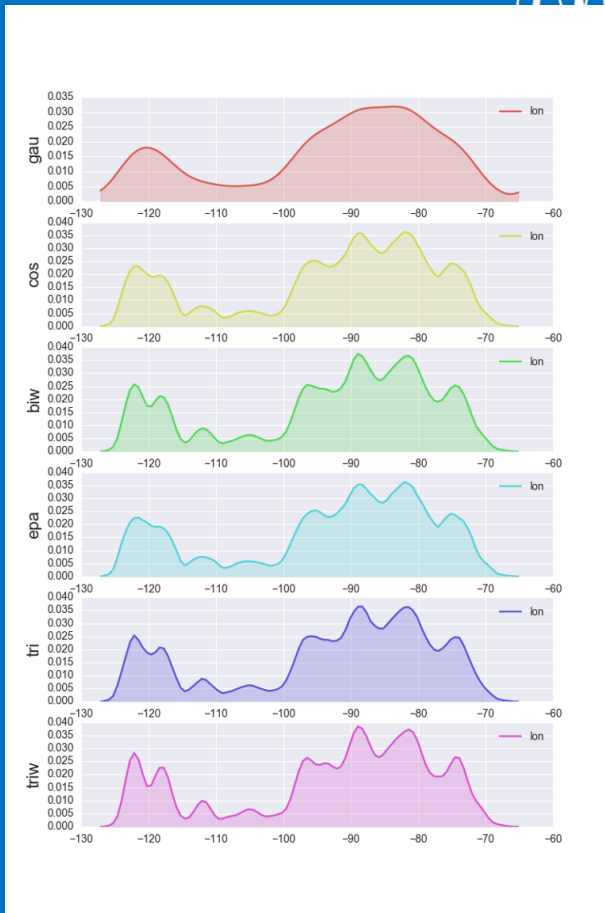
输出要素类

- 输出的要素类最好不要使用shapefile
 - shapefile 无法存储空值。因此，用来根据非 shapefile 输入创建 shapefile 的工具或其他程序可能会将空值存储为零或某些非常小的负数 ($-\text{DBL_MAX} = -1.7976931348623158\text{e}+308$)。
- ArcGIS推荐使用File GDB作为数据存储的主要载体。



核函数选项

- ArcGIS仅提供了高斯核函数
 - **FIXED**：固定距离法，也就是按照一定的距离来选择带宽，创建核表面
 - **ADAPTIVE**：自适应法。按照要素样本分布的疏密，来创建核表面，如果要素分布紧密，则核表面覆盖的范围小，反之则大。
 - 默认会使用固定方式，因为固定方式能够生成更加平滑的核表面。

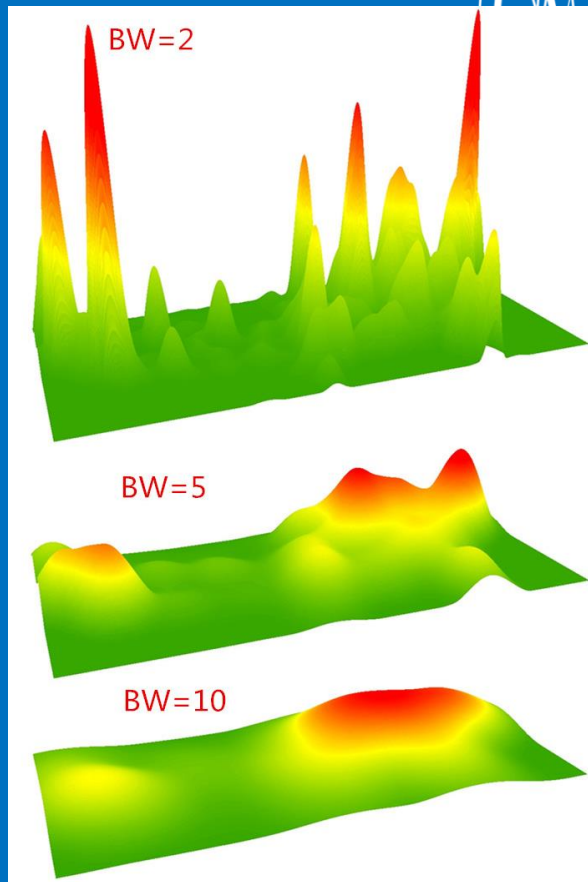


核带宽

- CV: 通过交叉验证法来决定最佳带宽。
- AIC: 通过最小信息准则来决定最佳带宽。
- BANDWIDTH_PARAMETER: 指定宽度或者临近要素数目的方法。

随着带宽的增大，曲面的曲率越来越平缓。那么可以得到下面的结论：

- 1、带宽越小，表面的曲率越大，越能突出不同区域之间的变化，揭露更多的细节情况。
- 2、带宽越大，表面曲率越小，生成的结果越平滑，结果更加抽象。



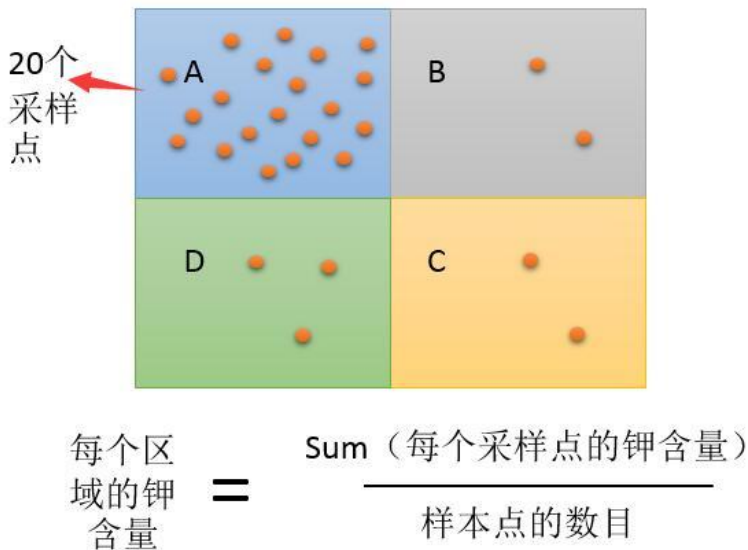
距离及临近要素的数目

- 当核带宽参数为BANDWIDTH_PARAMETER 的时候，这两个参数可用。
- **距离：**
 - 带宽距离单位，是要素类的空间参考中的单位，如果你是经纬度的话，这里设定的也是经纬度
 - 建议把数据投影为投影坐标系。
- **临近的要素数目：**
 - 如果核类型为自适应（ADAPTIVE），以及核带宽为BANDWIDTH_PARAMETER的时候，此参数才为可用
 - 默认是30，表示选择回归点周边的30个点作为核局部带宽中作为临近要素的点。




权重


- ArcGIS工具特有的一个参数，用来决定哪些数据在计算中发挥更大的作用。





扩展参数部分

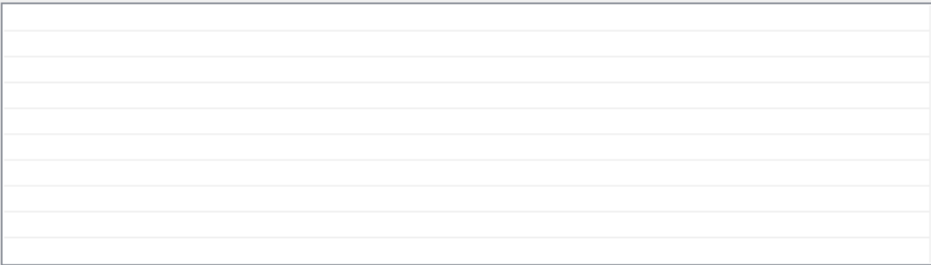
Additional Parameters (Optional)


Coefficient raster workspace (optional) 1 





Output cell size (optional) 2 

Prediction locations (optional) 3 

Prediction explanatory variable(s) (optional) 4 



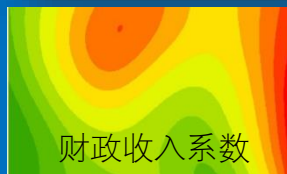
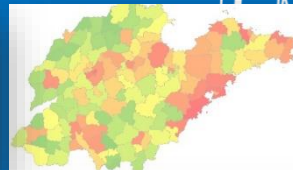
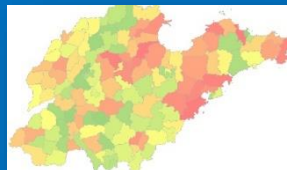
Output prediction feature class (optional) 5 

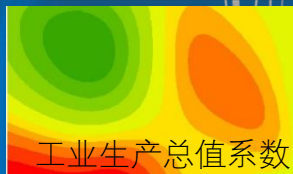


扩展参数的作用：栅格化系数

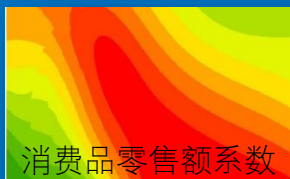
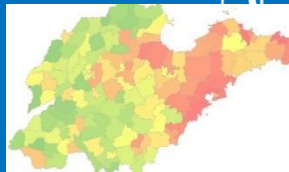
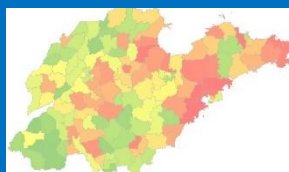
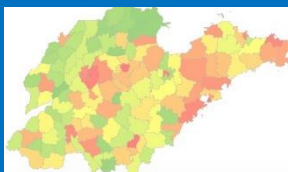
- 截距以及每个系数的栅格表面体现了空间异质性
 - 通过统计学上表示较小局部变化的较大全局变量可确定全局策略。
 - 通过统计学上表示较强局部变化的较大全局变量可确定局部策略。
 - 某些变量可能并不是在全局范围内各区域中均比较显著，因为在某些区域中，它们是正相关的关系，而在其他区域中它们则是负相关的关系。



财政收入系数



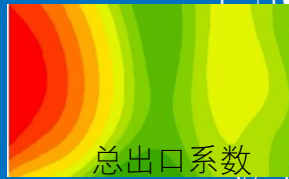
工业生产总值系数



消费品零售额系数



固定资产投资系数

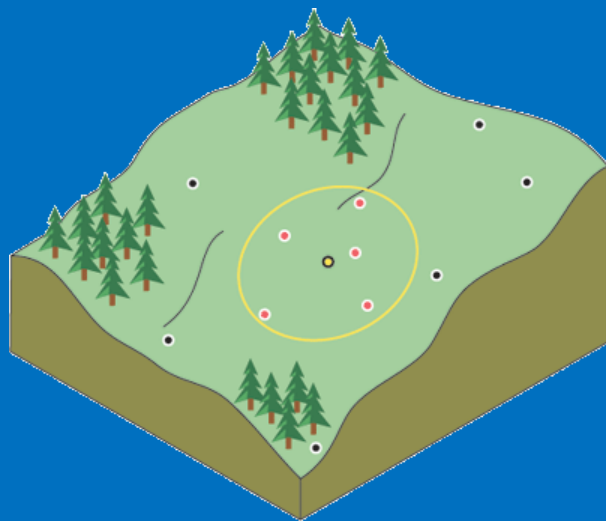


总出口系数



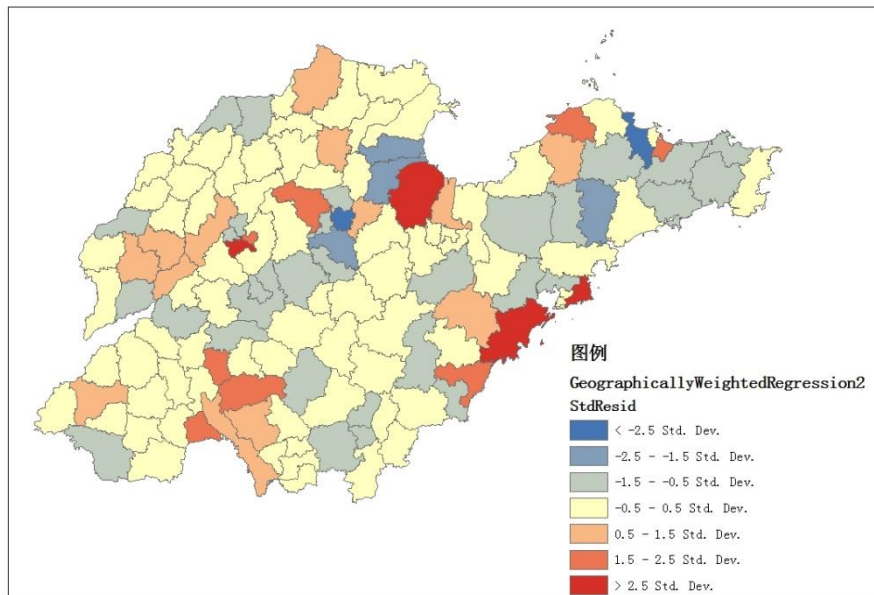
预测位置

- 主要用于采样数据的预测分析中。



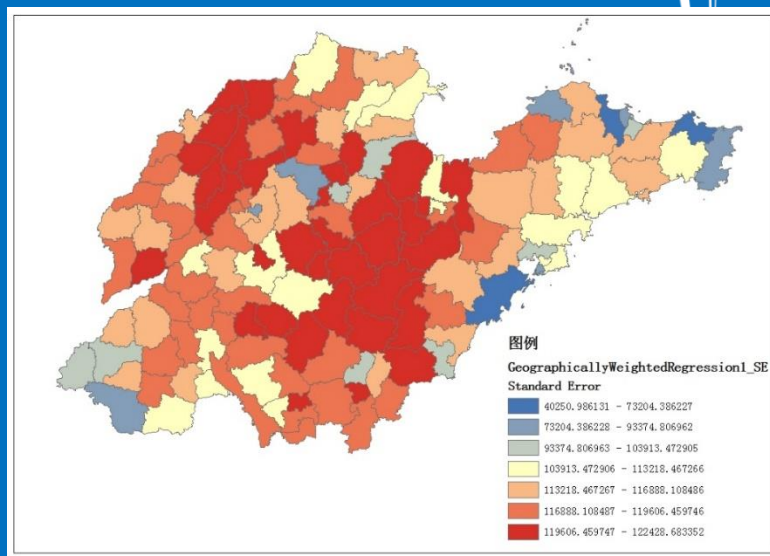
结果的解读（一）

- 标准化残差
 - 检查大于2.5倍标准差的区域，大于2.5倍标准差，表示这些区域的预测可能不可靠。



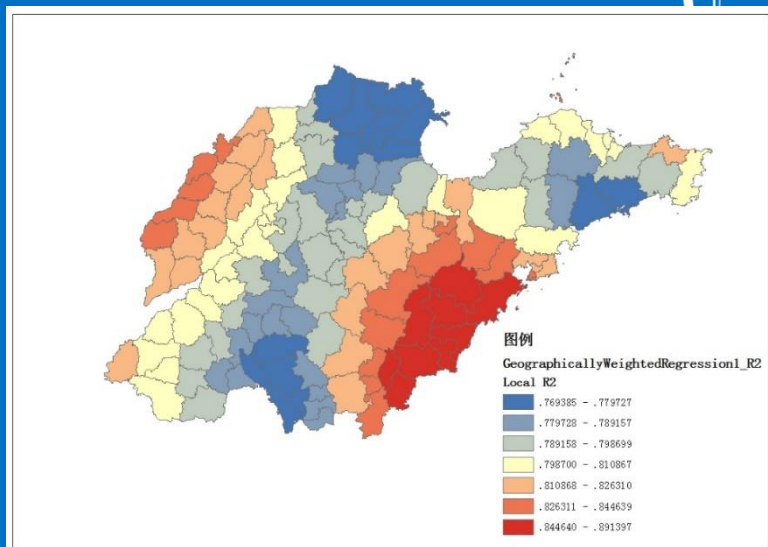
结果解读（二）

- 系数标准误（Standard Error）：
 - 衡量的是我们在用样本统计量去推断相应的总体参数时，一种估计的精度。
 - 用于衡量每个系数估计值的可靠性。标准误与实际系数值相比较小时，这些估计值的可信度会更高。较大标准误可能表示局部多重共线性存在问题。



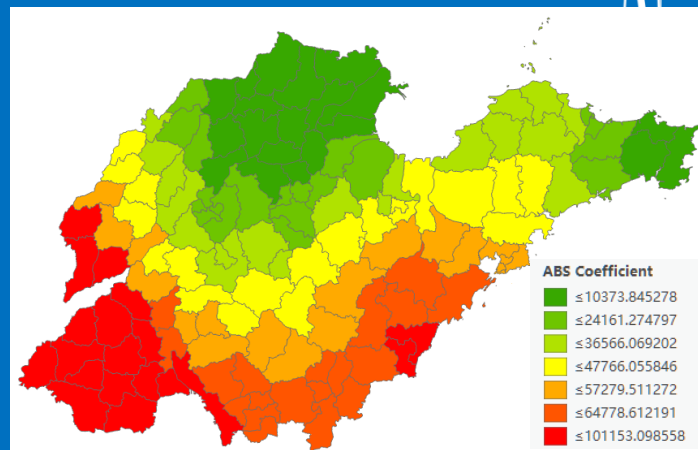
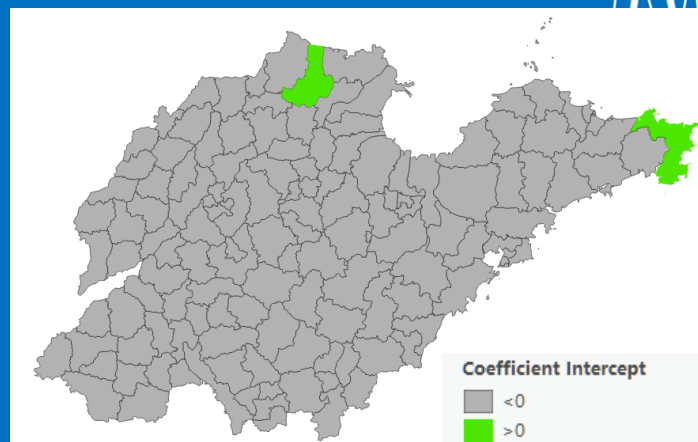
结果解读（三）

- Local R2:
 - 范围是 0.0 与 1.0 之间，表示局部回归模型与观测值的拟合程度。
 - 如果值非常低，则表示局部模型性能不佳。
 - 映射 Local R2 值以查看哪些位置 GWR 预测较准确和哪些位置不准确可为获知可能在回归模型中丢失的重要变量提供相关线索。



结果解读（四）

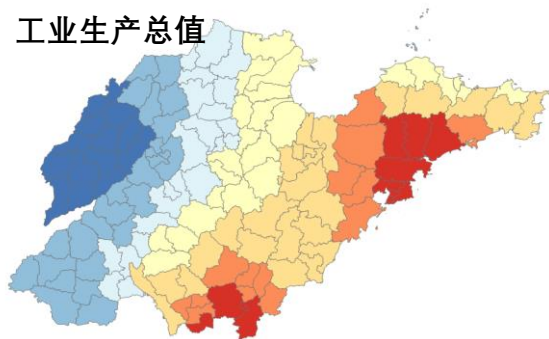
- 不同区域截距的情况
 - 截距，表示自变量在完全没有干扰项的情况下，对因变量的贡献。
 - 截距的正负，表示因自变量的正负相关。
 - 截距绝对值，表示了因变量对自变量的依赖程度。



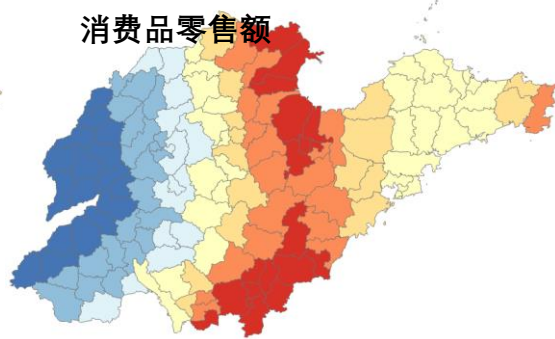
结果解读（五）

- 系数：不同区域各自变量对因变量的影响
 - 颜色越深，表示影响程度越大。

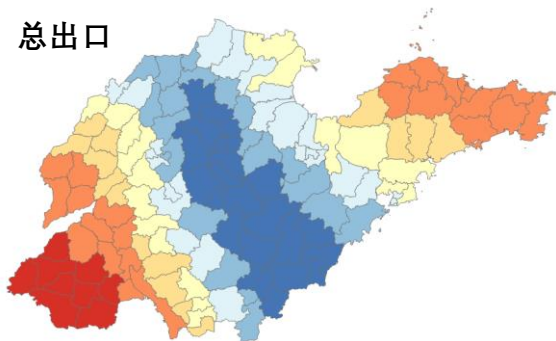
工业生产总值



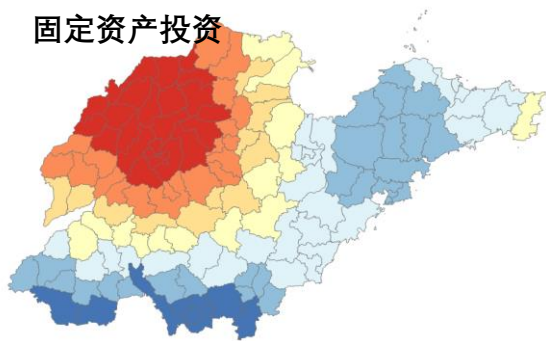
消费品零售额



总出口

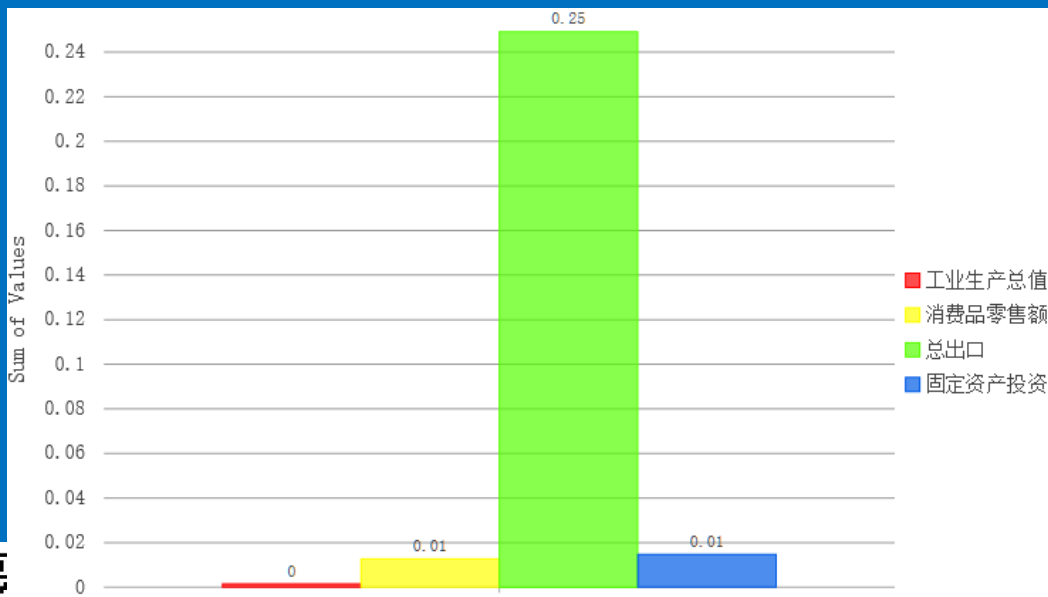


固定资产投资



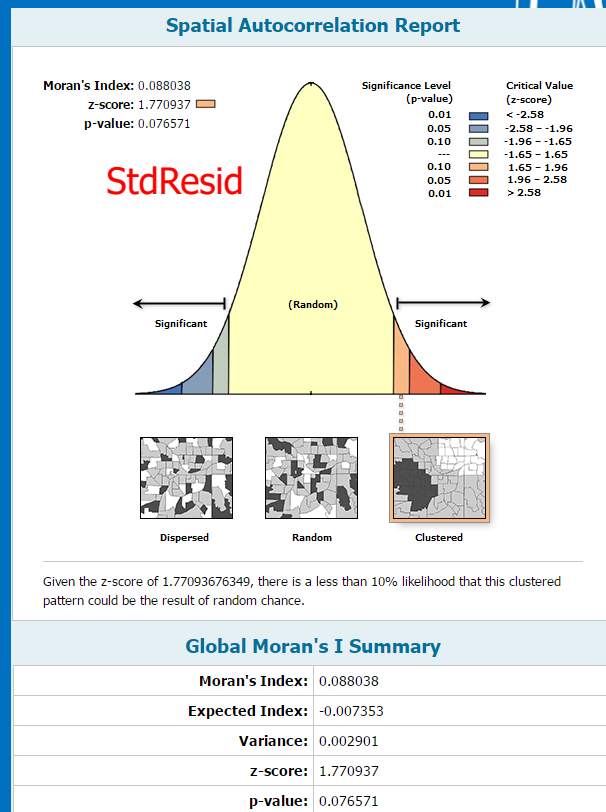
结果解读（五）

- 各系数之间的标准差
 - 标准差越大，表示空间差异性越大。
 - 标准差的解读，可以参考Z得分。



对各项结果进行扩展分析

- 全局空间自相关：
Moran's I
 - 出现聚集的趋势的，说明在分析的时候模型或者因子选择是有问题的。



Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]
Intercept	-54754.29553	23737.010973	-2.306706	0.022620*
O	0.002378	0.002095	1.135133	0.258374
R	0.125715	0.009578	13.125911	0.000000*
S	0.581584	0.078219	7.435333	0.000000*
T	0.022569	0.009285	2.430685	0.016405*

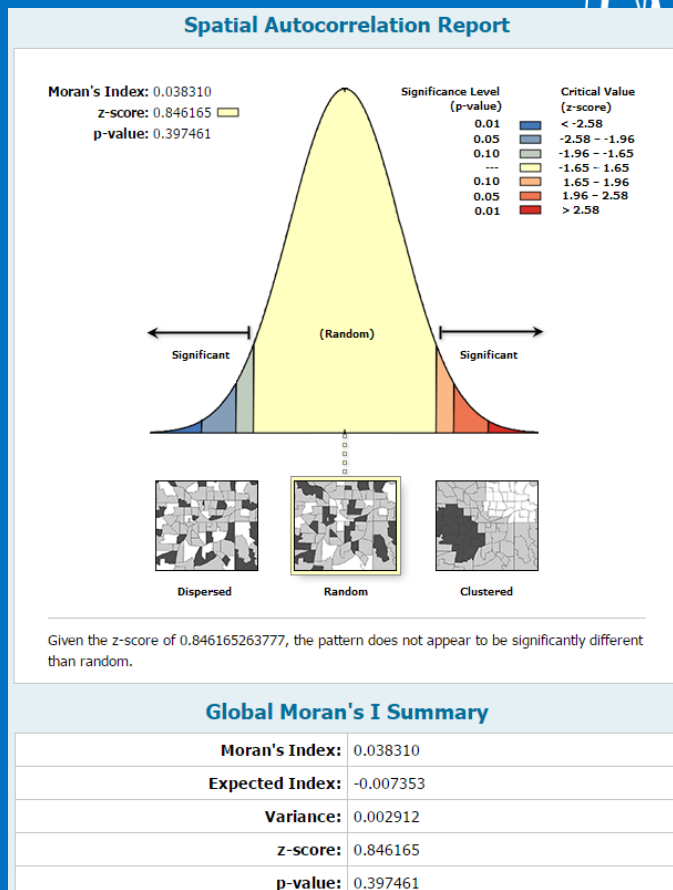


去掉(或者处理) 无法通过验证的因子

- Moran's I 结果为随机，表示通过统计检验。

多重共线性通用解决方法：

1. 保留重要解释变量，去掉次要或可替代解释变量
2. 用相对数变量替代绝对数变量
3. 差分法
4. 逐步回归分析
5. 主成份分析
6. 偏最小二乘回归
7. 岭回归
8. 增加样本容量



插播广告

- 所有的PPT、数据、文章、代码……均可以通过此公众号获取。

公众号：

虾神daxialu



谢谢！

