geodetector

2022/10/8

## 地理探测器值与线性回归模型的等价性

### 单因子探测

地理探测器主要是分析类型量（分类变量）的影响作用（解释力），假设，可以是连续变量，也可以是离散变量，但是一定得是定量数据；可以是定性数据（分类变量），也可以是离散型的定量数据，若是连续型的定量数据，则需进行适当的离散化。

归根结底，这里是要将按照进行分组（统计学术语中一般称为“层（strata）”），然后比较各层方差之和与总方差的差异性。

然后，我们经过分析，发现这里地理探测器的值本质上就是**同样处理后**的对进行回归，得到的回归模型的。

要证明证明这一点，首先假设是分类变量，那么设得到的结果是什么？答：得到的结果是若干条与轴平行的**线段**，每条线段对应轴坐标为对应分类那一组的组均值。

分类变量对进行回归，要先将转变为**虚拟变量**，这一过程相当于编码。假设有四个分类（、、、），那我们可以进行如下编码：

| 分类 |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 0 | 0 |
|  | 1 | 0 | 0 |
|  | 0 | 1 | 0 |
|  | 0 | 0 | 1 |

个分类只需引入个虚拟变量，此处取，，即为类别，取，，即为类别。

然后我们就可以，构建回归模型：

这里有两条路可以走：

一种方法是待定系数法，假设里面的、、、都已经求出来了，就是、、、，然后可知：

对于类别，，，，

两边求期望，显然有

的意义就是类别的组均值。

同理，对于类别，，，，

对于类别，，，，

对于类别，，，，

另一种方法是通过整理样本构建，最上面先排个类别，即行，然后依次是个类别……

然后最小二乘法求解，得到的最终公式是：

由此，

于是，

将不同组别的虚拟变量值代入后，显然也是，**对于只包含一个分类变量的回归方程，，函数拟合值就是不同组别的组均值**。这里是4个类别，个类别同理。

明确这点后，我们自然可以计算，与值计算保持一致，假设我们自变量有个类别，或者说层。

代表**未被模型解释变化**，代表**数据差异变化**。我们以表示第个类别的组均值。

可以进行一下实证模拟：

使用地理探测器，

factor\_detector("incidence", "elevation", CollectData)

## [[1]]  
## q-statistic p-value  
## elevation 0.6067087 0.04080407

使用lm函数，将转变为因子变量，对进行回归，

fit\_1 <- lm(incidence ~ factor(elevation), data = CollectData)  
summary(fit\_1)

##   
## Call:  
## lm(formula = incidence ~ factor(elevation), data = CollectData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.50588 -0.13811 -0.02889 0.07412 1.29189   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.45588 0.02874 224.628 < 2e-16 \*\*\*  
## factor(elevation)2 -0.28477 0.08407 -3.388 0.000868 \*\*\*  
## factor(elevation)3 -0.19777 0.04842 -4.085 6.66e-05 \*\*\*  
## factor(elevation)4 0.16548 0.05813 2.847 0.004936 \*\*   
## factor(elevation)5 -0.54699 0.06282 -8.707 2.08e-15 \*\*\*  
## factor(elevation)6 0.43275 0.05813 7.445 4.02e-12 \*\*\*  
## factor(elevation)7 -0.66588 0.08407 -7.921 2.46e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.237 on 178 degrees of freedom  
## Multiple R-squared: 0.6067, Adjusted R-squared: 0.5935   
## F-statistic: 45.77 on 6 and 178 DF, p-value: < 2.2e-16

### 交互因子探测

地理探测器交互因子探测是识别不同风险因子之间的交互作用，即评估因子和共同作用时是否会增加或减弱对因变量的解释力，或这些因子对的影响是相互独立的。在算法上的实际操作是，将数据按两个分类变量和进行交叉分组，若有三个分类，有两个分类，交叉之下，就有了六个分类。然后，对这种新的分类方式进行如上所述的单因子探测，比较这种新分类是否比两个原始分类有更强的解释性。

而这种交叉分类的处理方式，也是lm函数对于两个分类变量交互的处理方式，因此，这种情况下，地理探测器值与线性回归模型依然是等价的。以实证数据验证一下：

使用地理探测器，

interaction\_detector("incidence", c("soiltype", "elevation"), CollectData)

## [,1] [,2] [,3]   
## [1,] "soiltype" "elevation" "0.663523698335635"  
## [2,] "soiltype" "soiltype" "0.385716842809428"  
## [3,] "elevation" "elevation" "0.606708709727727"

使用lm函数，将和转变为因子变量，对进行回归，

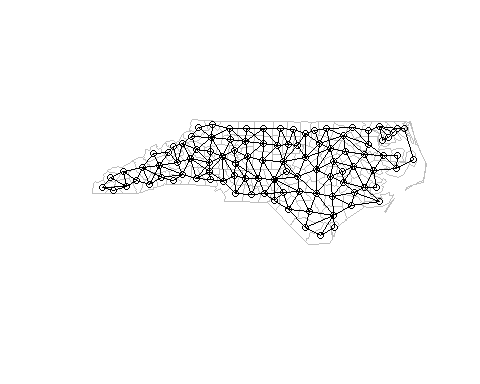
fit\_2 <- lm(incidence ~ factor(soiltype) \* factor(elevation), data = CollectData)  
summary(fit\_2)

##   
## Call:  
## lm(formula = incidence ~ factor(soiltype) \* factor(elevation),   
## data = CollectData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.45000 -0.13625 -0.02231 0.06167 1.20375   
##   
## Coefficients: (17 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.42412 0.03881 165.514 < 2e-16 \*\*\*  
## factor(soiltype)2 0.31588 0.08484 3.723 0.000268 \*\*\*  
## factor(soiltype)3 -0.10566 0.07380 -1.432 0.154111   
## factor(soiltype)4 -0.09917 0.14609 -0.679 0.498193   
## factor(soiltype)5 0.05755 0.07599 0.757 0.449934   
## factor(elevation)2 -0.30167 0.10764 -2.803 0.005668 \*\*   
## factor(elevation)3 -0.22181 0.07380 -3.006 0.003060 \*\*   
## factor(elevation)4 0.10833 0.23556 0.460 0.646187   
## factor(elevation)5 -0.44078 0.13631 -3.234 0.001473 \*\*   
## factor(elevation)6 0.14833 0.17285 0.858 0.392041   
## factor(elevation)7 -0.53495 0.16894 -3.167 0.001834 \*\*   
## factor(soiltype)2:factor(elevation)2 NA NA NA NA   
## factor(soiltype)3:factor(elevation)2 0.12321 0.20282 0.607 0.544367   
## factor(soiltype)4:factor(elevation)2 NA NA NA NA   
## factor(soiltype)5:factor(elevation)2 NA NA NA NA   
## factor(soiltype)2:factor(elevation)3 NA NA NA NA   
## factor(soiltype)3:factor(elevation)3 0.07585 0.12565 0.604 0.546909   
## factor(soiltype)4:factor(elevation)3 NA NA NA NA   
## factor(soiltype)5:factor(elevation)3 0.08639 0.11365 0.760 0.448219   
## factor(soiltype)2:factor(elevation)4 -0.31833 0.27974 -1.138 0.256760   
## factor(soiltype)3:factor(elevation)4 0.21154 0.24955 0.848 0.397823   
## factor(soiltype)4:factor(elevation)4 NA NA NA NA   
## factor(soiltype)5:factor(elevation)4 NA NA NA NA   
## factor(soiltype)2:factor(elevation)5 NA NA NA NA   
## factor(soiltype)3:factor(elevation)5 NA NA NA NA   
## factor(soiltype)4:factor(elevation)5 NA NA NA NA   
## factor(soiltype)5:factor(elevation)5 -0.10755 0.19980 -0.538 0.591104   
## factor(soiltype)2:factor(elevation)6 NA NA NA NA   
## factor(soiltype)3:factor(elevation)6 0.44771 0.19073 2.347 0.020083 \*   
## factor(soiltype)4:factor(elevation)6 NA NA NA NA   
## factor(soiltype)5:factor(elevation)6 NA NA NA NA   
## factor(soiltype)2:factor(elevation)7 NA NA NA NA   
## factor(soiltype)3:factor(elevation)7 NA NA NA NA   
## factor(soiltype)4:factor(elevation)7 NA NA NA NA   
## factor(soiltype)5:factor(elevation)7 NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2263 on 167 degrees of freedom  
## Multiple R-squared: 0.6635, Adjusted R-squared: 0.6293   
## F-statistic: 19.37 on 17 and 167 DF, p-value: < 2.2e-16

## 空间异质性与自相关性并存情形下地理探测器值的有效性

前文已经论述了OLS回归得到的R2等价于：=OLS-R2。从值的计算公式来看，组内方差和总方差计算的隐含假设为y是随机分布，无显著的整体或者层内空间自相关或联系效应。因为统计量和R2都是测度贡献度的，我们考虑在y有不同空间自相关情形下，和R2的差异关系。 以dpdata包内置的北卡罗来纳州（North Carolina）的拓扑结构构建邻接矩阵

library(geodetector)  
library(spdep)  
library(spatialreg)  
nc <- st\_read(system.file("shapes/sids.shp", package="spData")[1], quiet=TRUE)  
st\_crs(nc) <- "+proj=longlat +datum=NAD27"  
row.names(nc) <- as.character(nc$FIPSNO)  
# neighbourhood structure and weights matrix  
nb\_nc <- spdep::poly2nb(nc)  
lisw\_nc <- spdep::nb2listw(nb\_nc, style = "W", zero.policy = TRUE)  
plot(st\_geometry(nc),border = 'grey')  
plot( st\_centroid(nc$geometry), add=TRUE )  
plot(lisw\_nc, st\_centroid(nc$geometry), add= TRUE )



### 3.1 单因子

构建一个具有空间滞后形式的空间自回归数据产生过程（spatial lag model），为三类别的分类变量.

可以得到的产生过程：

不失一般性地将模拟参数设置为：

在模拟实验中，由于具有空间效应，并且真实的空间效应为,那么理论上需要解释的部分为，真实则为

其中：为预设，,和为OLS估计。

每个模拟100次，取均值，观察*q*相对于真实R2的偏差