**回归模型视角下地理探测器方法的解释与拓展**

## 1地理探测器值与线性回归模型的等价性

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，则需进行的离散化。

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。

从回归模型视角分析对的解释度，我们发现，将分类变量转化为多个虚拟变量后，**无论单因子还是多因子探测，地理探测器的统计量均等价于的线性回归的。**

### 1.1单因子探测等价性证明

**线性回归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期望，并且期望在虚拟变量线性回归中恒等于组均值。**

假设有四个分类（、、、），将单个分类变量转化为多个虚拟变量，这一过程相当于编码：

| 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，，即为参照类别对应的虚拟变量，取，，即为类别对应虚拟变量，以此类推。构建回归模型：

设回归估计的系数为、、、，

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，，，，

对于具有*n*个类别的情况，恒有：

**因此线性回归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期望，即为组期望。**

进一步地可以证明**期望在虚拟变量线性回归中恒等于组均值**，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：

可得，

于是，

将不同组别的虚拟变量值代入后，显然是是组均值，**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，我们很容易可以推导出等价于，假设的自变量有个类别，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。

### 1.2多因子探测等价性证明

以双因子探测为例，假设和各有3个分类：1、2、3。将和转化为虚拟变量后（以对应小写表示），需要构建如下与地理探测器等价模型：

其中包含了单因子作用和双因子交互作用，对两边同求期望：

对于，，根据上述编码规则，和均为0，于是

对于，，根据上述编码规则，和，其余均为0，于是

同理，对于，，

对于，，

对于，，

同样可以断言**的估计值为交互分组后的组期望，且等于交互分组后的组均值**，到这里，同1.1可以推倒出等价于。

### 1.3实证数据验证

利用地理探测器自带的数据集CollectData，进行实证验证：

### 1.3.1单因子

使用lm函数，将转变为虚拟变量，对进行回归，得到lm结果

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

地理探测器结果：

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

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

### 1.3.2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。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对于两个分类变量交互的处理方式，这种情况下，地理探测器值与线性回归模型依然是等价的。

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

地理探测器结果：

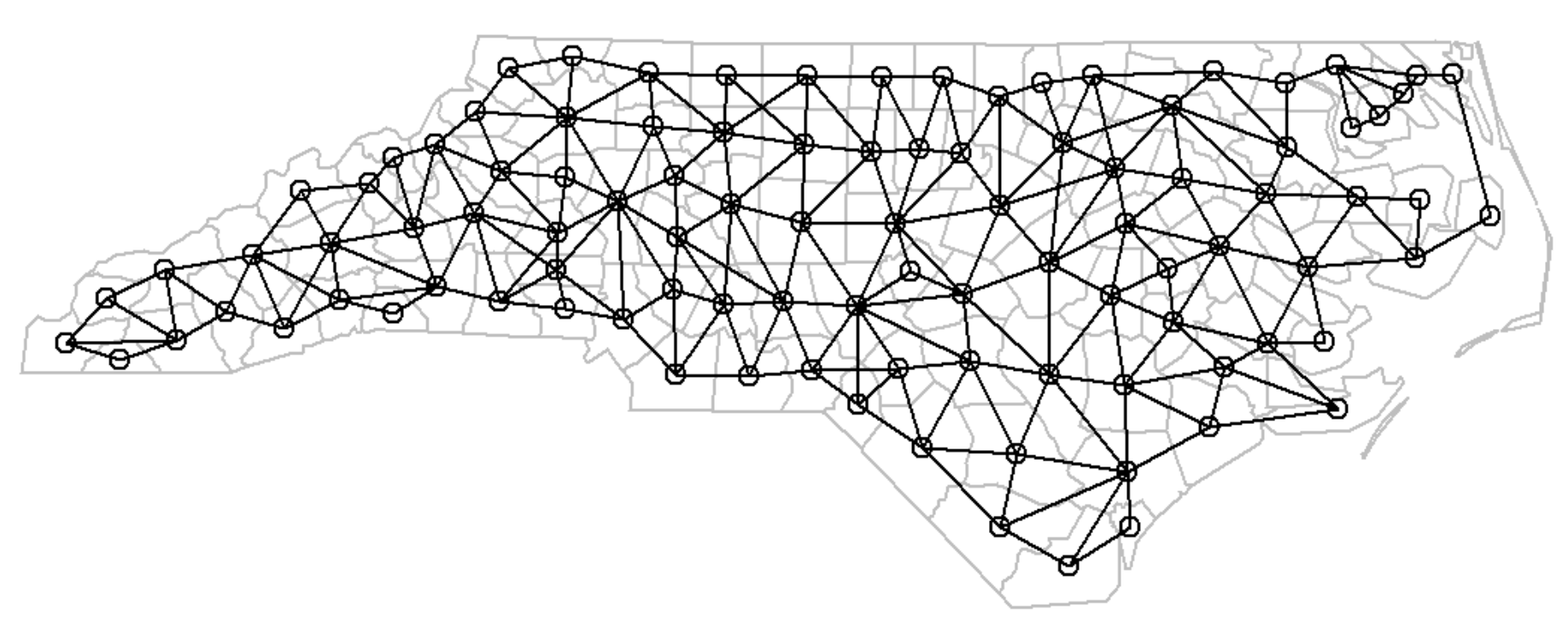
interaction\_detector("incidence", c("soiltype", "elevation"), CollectData)

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

## 2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*R*2等价于：=OLS-*R*2。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。统计量和*R*2都是测度贡献度的，我们考虑在具有不同程度空间自相关情形下，和*R*2的差异关系。

以spdata包内置的北卡罗来纳州（North Carolina）的拓扑结构构建邻接矩阵



### 2.1单因子模拟

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，真实对的解释度 则为

其中：为预设，,和为OLS估计。

每个模拟1000次，观察*q*相对于真实*R*2的偏差。这里我们采用相对于真实解释度的偏差百分比作为衡量标准：

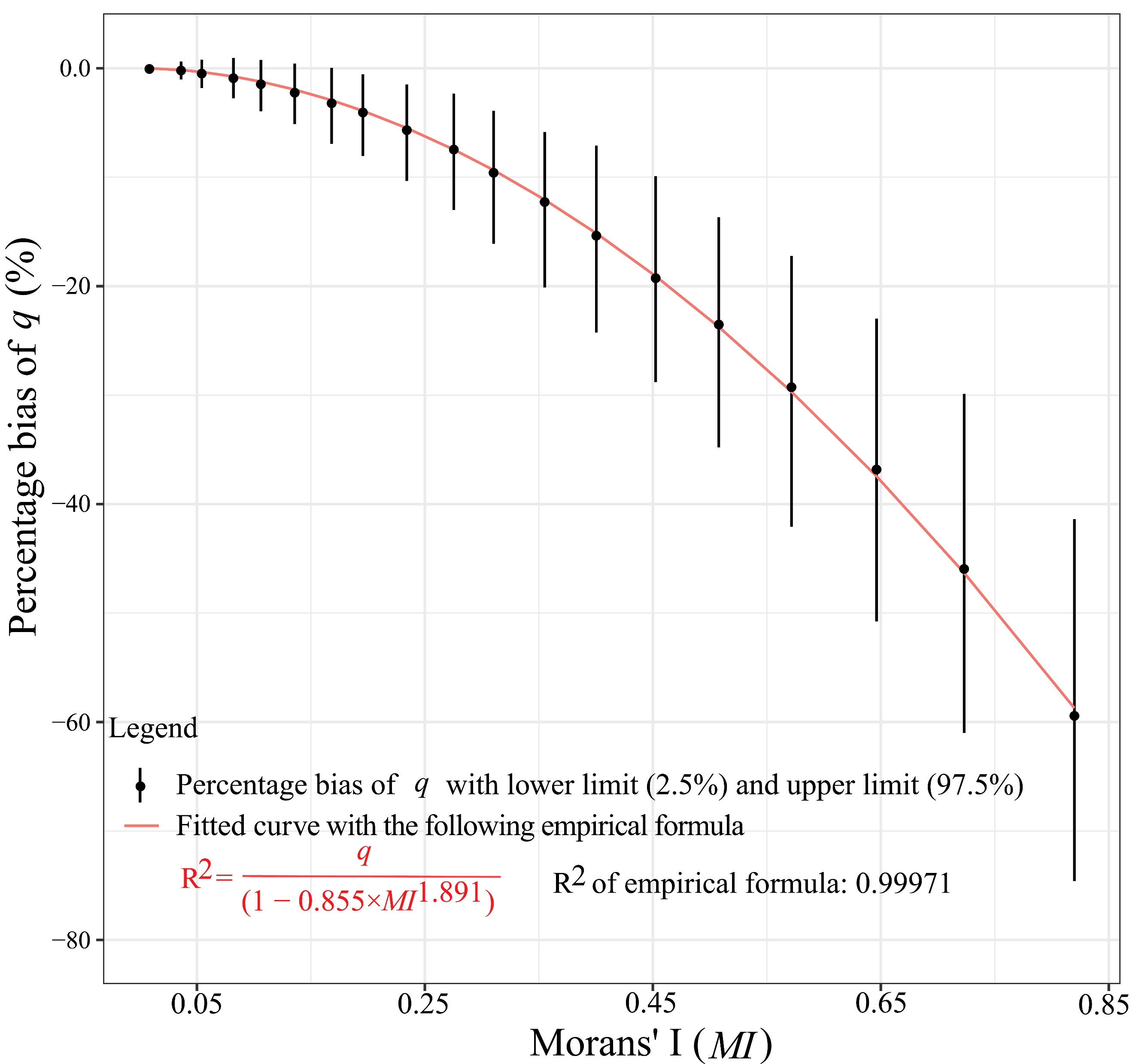
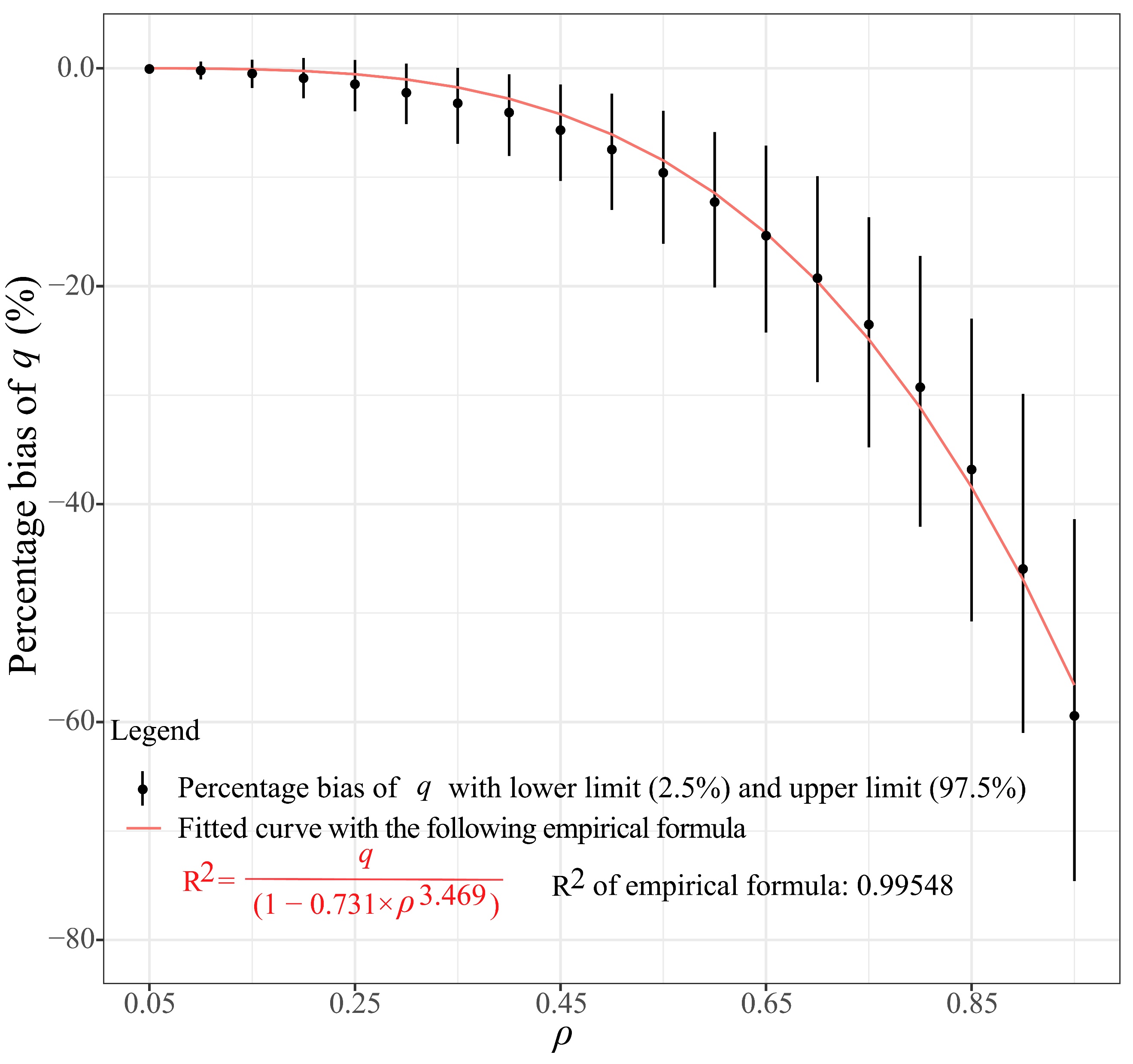


图1 与*q*偏差百分比的经验关系（左）；与*q*偏差百分比的经验关系（右）

由于空间效应*q*存在明显的低估现象，且随着和的增大（空间自相关性的增大），*q*低估程度越大，偏差越大，这种低估与空间自相关强度具有明显的指数关系。**造成*q*低估的原因在于，*q*本质上等价于虚拟变量OLS回归给出的因子解释度，它忽略了数据内在的关联性，导致*q*在空间自相关性存在的情况下是有偏的。因为具有空间自相关性的数据违背了OLS回归的基础假设：样本之间完全独立，所以OLS的估计系数存在偏差，其给出的因子解释度也存在偏差， 最终导致等价的*q*也存在偏差。**

### 2.2多因子交互模拟

假设和各有3个分类：1、2、3。将和转化为虚拟变量后（以各自对应的小写字母表示），需要构建如下空间自回归模型：.

可以得到的产生过程：

不失一般性地将模拟参数设置为：

每个模拟1000次，

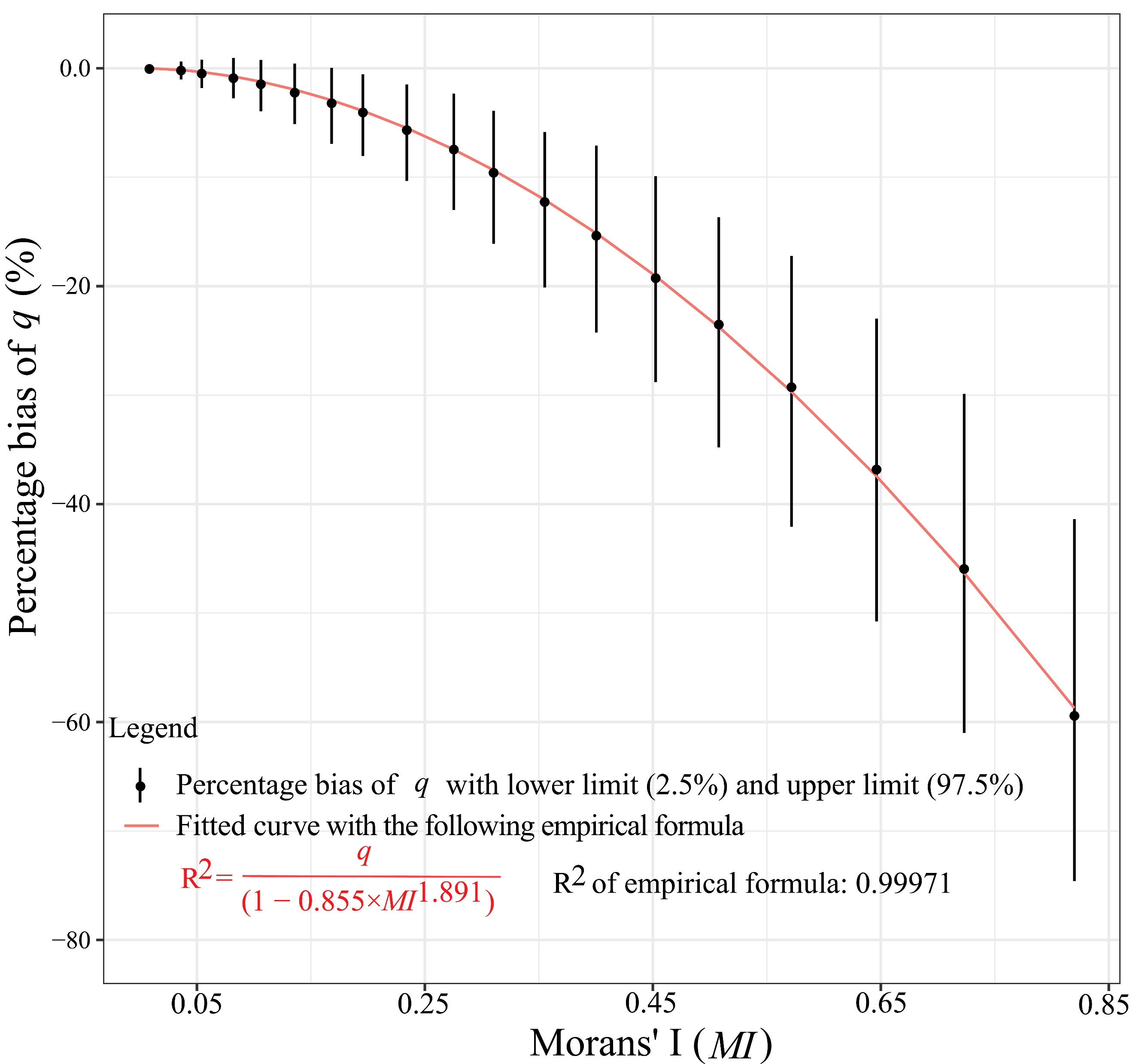
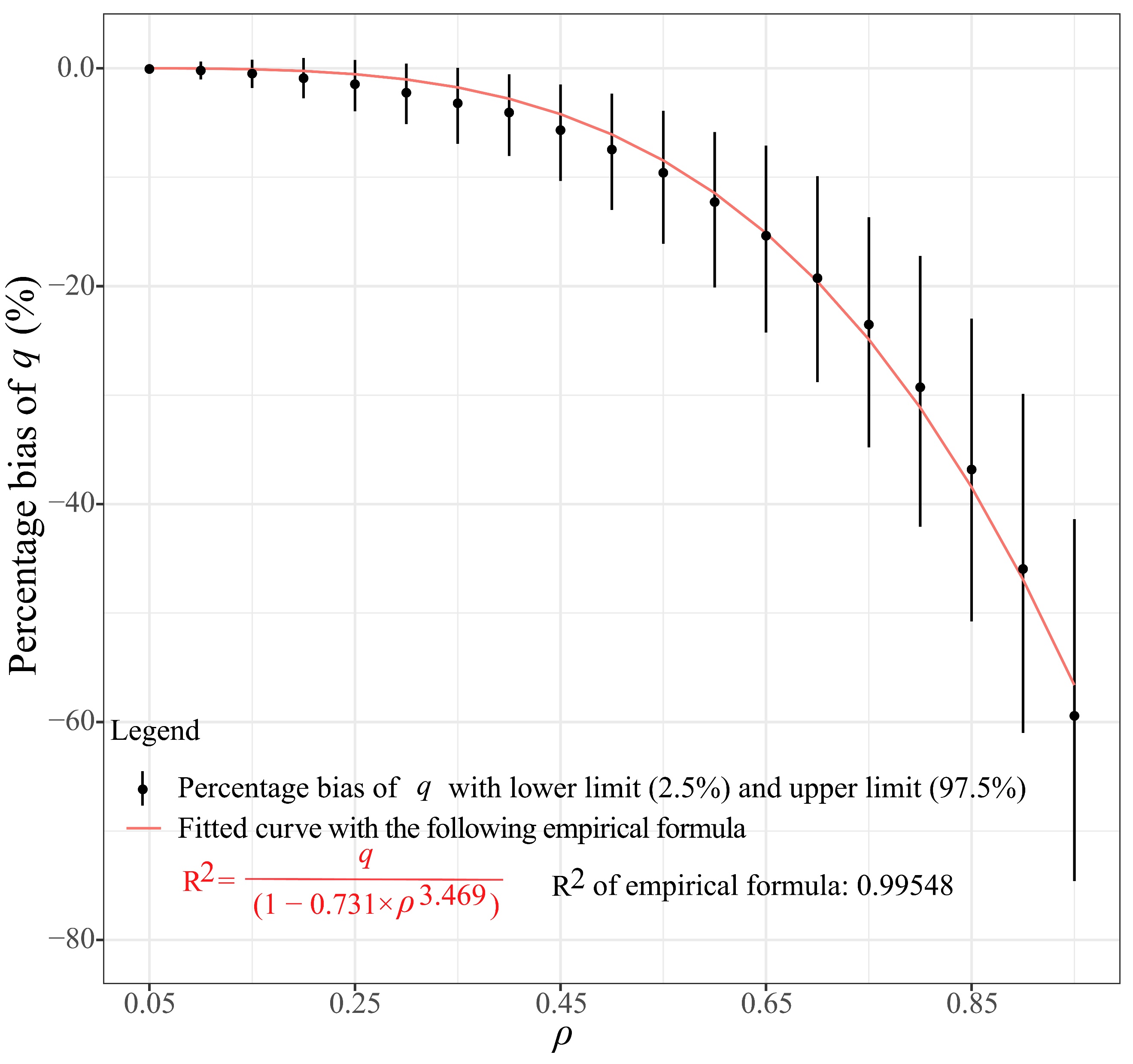
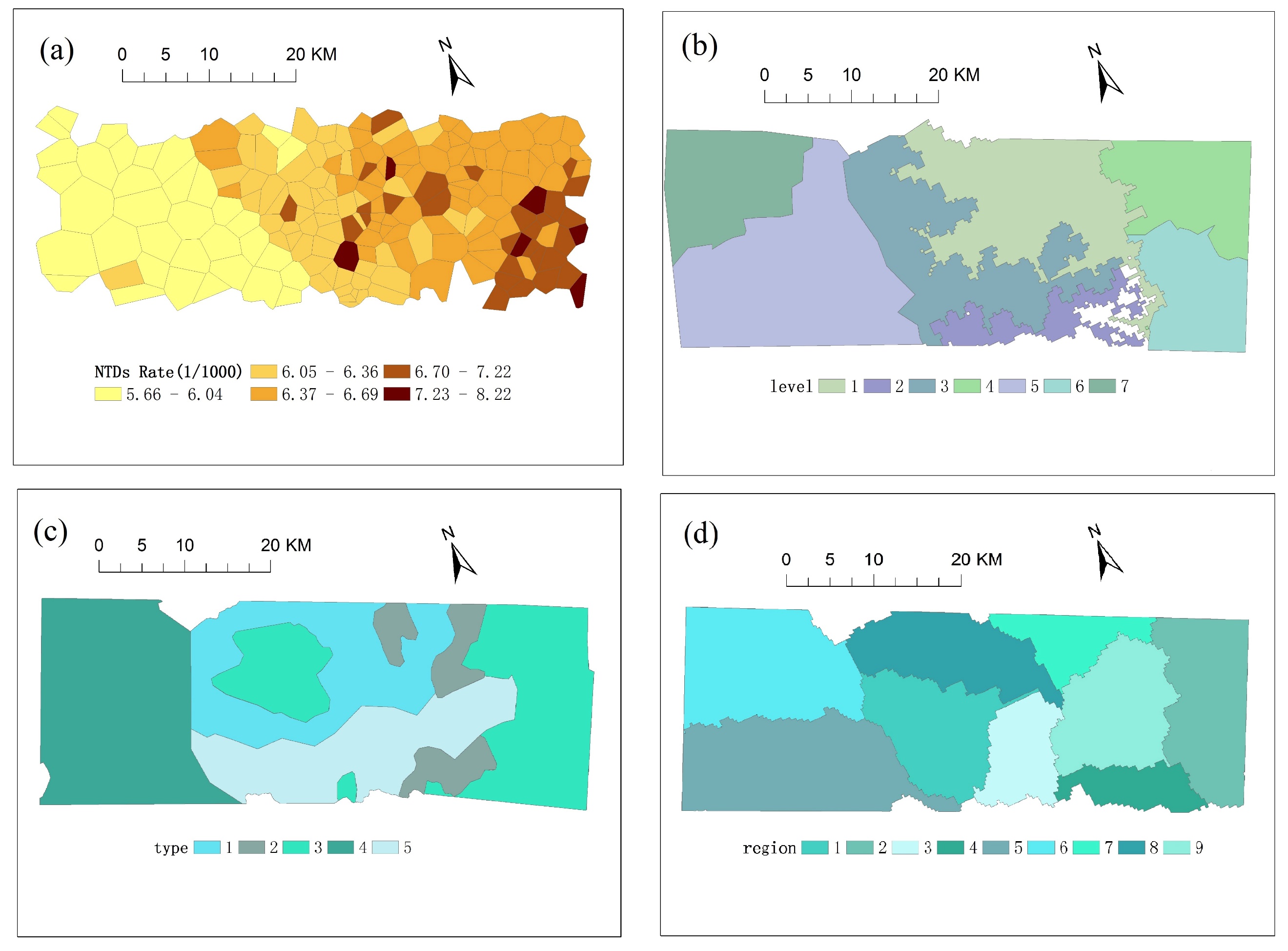


图2 因子交互情景下： 与*q*偏差百分比的经验关系（左）；与*q*偏差百分比的经验关系（右）

在双因子交互情景中结果与上文一致，*q*仍然由于空间效应存在明显的低估现象，且随着和的增大，*q*低估程度越大，偏差越大，**这种低估同样源于*q*的隐藏假设：y是随机分布设。**

### 2.3应用空间计量模型探测因子解释度

采用R包spatialreg提供空间滞后模型估计算法（lagsarlm）探测geodetector中自带数据的因子解释度。可以看到被解释变量*incidence*具有明显的空间依赖性。这意味着地理探测器给出的*q*值可能存在低估。



### 2.3.1单因子探测

*q*由于空间自相关性的存在（Morans’I = 0.6915）出现低估现象。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *q* | SAR.R2 | Spatial dependence | SAR.category | SAR.coefficients | SAR*.p*.value |
| 0.6067 | 0.6323 | Morans’I = 0.6915 | Intercept | 3.5131 | 0.0000 |
| factor(elevation)2 | -0.1751 | 0.0231 |
| factor(elevation)3 | -0.1028 | 0.0226 |
| factor(elevation)4 | 0.0974 | 0.0685 |
| factor(elevation)5 | -0.3057 | 0.0000 |
| factor(elevation)6 | 0.2631 | 0.0000 |
| factor(elevation)7 | -0.3646 | 0.0001 |

### 2.3.2多因子探测

*q*由于空间自相关性的存在（Morans’I = 0.6915）仍然出现低估现象。**应用空间计量模型探测因子解释度的优势在于不仅可以给出更为准确的（无偏的）因子解释度（SAR.R2），且可以通过shapley value计算方法探查单因子和多因子交互的边际贡献，且不局限于双因子探测，可以同时对*n*个因子进行探测，这将会在下一部分中介绍。**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *q* | SAR.R2 | Spatial dependence | SAR.category | SAR.coefficients | SAR*.p*.value |
| 0.6635 | 0.7393 (TC) | Morans’I = 0.6915 | (Intercept) | 3.992137 | 0.0000 |
| factor(elevation)2 | 0.271396 | 0.0004 |
| factor(elevation)3 | -0.074644 | 0.2552 |
| factor(elevation)4 | -0.051228 | 0.6945 |
| factor(elevation)5 | 0.033604 | 0.6190 |
| factor(soiltype)2 | -0.197883 | 0.0413 |
| factor(soiltype)3 | -0.126493 | 0.0599 |
| factor(soiltype)4 | 0.10883 | 0.6038 |
| factor(soiltype)5 | -0.283657 | 0.0233 |
| factor(soiltype)6 | 0.114838 | 0.4580 |
| factor(soiltype)7 | -0.337702 | 0.0292 |
| factor(elevation)3:factor(soiltype)2 | 0.098814 | 0.5838 |
| factor(elevation)3:factor(soiltype)3 | 0.065556 | 0.5575 |
| factor(elevation)5:factor(soiltype)3 | 0.051749 | 0.6085 |
| factor(elevation)2:factor(soiltype)4 | -0.301245 | 0.2263 |
| factor(elevation)3:factor(soiltype)4 | 0.111851 | 0.6141 |
| factor(elevation)5:factor(soiltype)5 | -0.044745 | 0.8012 |
| factor(elevation)3:factor(soiltype)6 | 0.296856 | 0.0809 |

## 3变量贡献度的更加直接的计算方法

### 3.1无空间自相关性

**无空间自相关效应前提下：OLS的LMG（shapley value）变量贡献度的计算方法可以直接计算每个分类变量及交互项的贡献度，并且每个单变量和交互变量的边际贡献的和等于*R*2，因此我们可以直接对shapley value给出边际贡献进行比较，并且确定不同变量具体的边际贡献差额。**

三因子（*soiltype, watershed, elevation*）OLS的LMG （shapley value**）**：

当然地理探测器交互效应分析的模式也可以利用更为精细的变量贡献度进行分析，**但是只能给出单因子和双因子解释度，并未给出多三因子解释度，且各解释度之间是非系统化的。更为重要的是，既然我们已经确定这三个变量都对被解释变量具有相当的解释度，那么忽略任一因子的解释度，只两两交互给出的*q*统计量相对于整体模型而言是有偏的。**

三因子地理探测器结果：

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

地理探测器可以比较交互因子与单因子、单因子之间的q统计量大小关系确定因子对被解释变量的解释度大小，但是这种数量关系的只被认为存在大小关系，具体的数额不能通过q统计量确定。shapley value给出的**单变量和交互变量的边际贡献的和等于*R*2，因此我们可以直接对边际贡献进行比较，并且确定不同变量之间具体的边际贡献差额。**

### 3.1有空间自相关性

有空间自相关效应的情境下，也可以计算shapley value，由于三因子线性模型由于具有更为完备的分区，已经基本达到了真实的可解释度，所以考虑空间效应后总解释度并未得到明显提升，但是依据shapley value变量进入法则，由于单因子时空间分区并不够完备，导致OLS的估计存在偏差，所以OLS回归的单因子边际贡献是有偏的。结合上文的论述，**我们可以确定：考虑空间效应的shapley value，即使总解释度并未明显提高，但其给出的单因子和因子交互的边际贡献更为可靠。**