基尼加权回归分析:概念、方法及应用

戴平生

内容提要:普通最小二乘法是进行回归分析最常用的基本方法,但该方法要求满足若干经典假设,对于小样本或在与收入相关回归分析的参数估计中易受奇异值、高收入群体的影响。本文试图利用基尼加权回归弥补以上不足。基尼加权回归可分为参数方法与非参数方法两类,参数方法基于样本残差的基尼平均差最小原则对参数进行估计;非参数方法则是直接由两点间的斜率加权得到。基尼加权回归分析可以进行参数假设检验并定义拟合优度,其中的假设检验在实际应用中采用 Jackknife 重抽样方法估计方差。文中提出的样本拓展基尼平均差算法,弥补了现有算法对样本数据只能提供近似计算的不足,极大地简化了相应的计算公式。本文利用我国 2015 年省域截面数据、1994—2015 年总量时间序列数据分别讨论入境旅游收入对收入基尼系数的影响,发现使用基尼加权回归的结果不仅符合理论预期,而且可以通过不平等厌恶参数的变化反映入境旅游收入对不同群体收入公平性的影响。

关键词:基尼加权回归;基尼平均差;参数估计;非参数估计

DOI: 10.19343/j.cnki.11-1302/c.2018.09.009

中图分类号: O212 文献标识码: A 文章编号: 1002-4565(2018)09-0103-12

Weighted Gini Regression: Concept, Method and its Application

Dai Pingsheng

Abstract: The method of ordinary least squares (OLS) is one of the most common one for regression analysis. OLS relies on several classical assumptions, and estimators are affected easily by extreme values, high income groups in regression analysis with related to income or small sample size. This paper promotes the weighted Gini regression as an alternative way, which consists of parameter estimating and non-parameter estimating. Parameter estimator is based on minimum of Gini mean difference of sample residues; non-parameter estimator comes from weighted value of slopes. Hypothesis test and R-squared calculating are carried in weighted Gini regression, resampling Jackknife technology is used to estimate variance for hypothesis test. It promotes a new algorithm of sample extend Gini mean difference, which can cover the shortage of approximate treatment of sample data. It discusses about how inbound tourism receipt influences income Gini coefficients by using 2015 provincial cross-sectional data and 1994—2015 total time series data in China. The results from weighted Gini regression line with expectations of relationship between variables, and they can reflect effects of inbound tourism receipt on income equity of different groups by changing inequality preference.

Key words: Weighted Gini regression; Gini's mean difference; Parameter Estimation; Non-parameter estimation

一、引言

回归分析从本质上说就是利用样本点拟合回归曲线,使得回归曲线与这些样本点尽可能地"近"。回归曲线通常可以通过若干参数的估计来确定,普通最小二乘法(OLS)是根据残差平方和

最小准则来得到参数估计，而最小绝对偏差法(LAD) 则是按残差绝对值之和最小准则对参数进行估计[1]。考虑到数据中某些异常点的影响，又发展出加权最小二乘法、LASSO 回归和分位数回归等分析工具[2][3]，它们都可以看作是以残差为基础的某一目标函数最小化的参数估计方法。以样本对应残差计算的组数据基尼平均差作为目标函数，对目标函数最小化的参数求解过程，我们称之为基尼加权回归分析。

基尼系数是一个在社会经济领域广泛应用于测度收入分配或资源配置均衡性的统计指标，它是基尼平均差与洛伦兹曲线碰撞相互结合的产物。基尼平均差最早用于测度数据的分散程度，基尼系数仅仅是基尼平均差的一种应用形式。Oklin 和 Yitzhaki[4] 提出了基尼回归分析的概念，首先分析了传统最小二乘法可能出现的问题，认为古典线性回归模型通常基于两个基本假设: 一是自变量与因变量的条件期望之间客观上存在线性关系，二是误差项应满足独立同正态分布且与自变量不相关。这样在样本容量小的情况下，不仅需要对误差项的正态分布假设进行有效分析，而且残差平方和对极端值的过度反应会使得 OLS 估计量对异常点极为敏感。因此他们提出了两种克服敏感性的方法，第一种是使用测度残差离散程度的替代方案。如果该测度方案对异常点较不敏感，那么回归系数也将会更为正常。于是他们把基尼平均差作为替代方案。第二种是使用由成对样本点定义的斜率进行加权平均构造稳健的回归系数估计。他们把前者称为基尼回归分析的最小化方法，后者称为加权平均法，并导出相关性质。无论是误差项的基尼平均差还是成对样本点的斜率都具有 U 统计量的特征，他们利用 U 统计量渐近正态的性质还对回归参数的估计进行显著性检验。

Yitzhaki[5] 从福利经济学政策评估角度认为关于消费的线性支出系统其回归系数的 OLS 估计严重依赖于高收入群体，占样本 10% 的最高收入群体对回归系数 OLS 估计量的贡献可能会高于其余 90% 人口的贡献。这一问题正在困扰着福利经济学的研究，因为政策倾向于对低收入群体的需求给予更大的权重。然而在回归参数的 OLS 估计中，低收入群体被忽略。如果他们的消费行为具有异质性，即恩格尔曲线是非线性的(消费收入弹性不等于常数) ，那么穷人将无法影响 OLS 估计的参数。这意味着不得不对 OLS 参数估计的正确性进行特定的检验。因此，基于经济学与统计学的综合考量，他提出引入拓展基尼平均差通过改变不平等厌恶参数减小高收入群体的相应权重。他在探讨参数的 OLS 估计量与基尼回归分析中斜率加权法的等价形式过程中，通过实证得出了以上结论和推断，同时还验证了拓展基尼平均差在不同不平等厌恶参数下确实取得了预期的效果。

Schechtman 等人[6] 在多元基尼回归分析的加权平均法研究中取得了重大的进展。Oklin 和 Yitzhaki[4] 的研究奠定了基尼回归分析的基础，Yitzhaki[5] 的贡献则是引入了拓展基尼平均差，两者都以简单一元基尼回归分析为主要的研究对象。通过对因变量条件期望函数关于各个自变量微分依据复合函数的求导法则获得的线性表达式，Schechtman 等人找到了多元基尼回归分析加权平均法的窍门。类似于多元线性回归模型参数 OLS 估计的正则方程，他们给出多元基尼回归分析加权平均法的参数估计定义，并得到了参数估计的若干性质，还利用 U 统计量性质从理论上给出了这些估计量的标准差从而解决对应参数的显著性问题。但实际应用中要从渐近正态分布中计算各参数估计量的标准差却不是一件容易的事情，幸好他们想到了具有良好属性的 Jackknife 方法，即利用样本重抽样获得估计量的标准差。

基尼回归分析在理论和应用研究方面都取得了很大的进步，但已有研究还存在一些不足。一是基尼回归分析的最小化方法，现有研究并没有表明如何通过最小化误差项的基尼平均差获得参数估计。本文尝试从某一给定的初始值出发，用网格搜寻法获得。二是对样本基尼平均差的计算，现有研究都采用连续分布下基尼平均差算法，当样本为组数据时它只能得到近似值，对于拓展基尼平均差更是如此。为弥补这一缺陷本文采用一种适用于个体数据和组数据拓展基尼平均差的新算

法。由于新算法适用于组数据方便进行参数的加权回归估计,我们称参数求解过程为基尼加权回归分析。本文的余下部分这样安排: 第二部分介绍基尼平均差及其拓展,第三部分是基尼加权线性回归分析,第四部分是拟合优度和假设检验,第五部分为应用举例。

二、基尼平均差及其拓展

基尼平均差的提出最初是用于测度个体数据的波动性①。本文主要讨论组数据,将个体数据作为它的特例。

(一) 基尼平均差的计算和性质

定义 1: 设有 n 个单元各自的平均收入分别为 $x_1, x_2, \cdots, x_n$(不妨假定已经按从小到大排列),相应的人口份额为 $p_1, p_2, \cdots, p_n$,则基尼平均差 Γ 可以定义为:

$$\Gamma = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|p_i p_j \tag{1}$$

显然基尼平均差具有非负性。若用 X 表示收入随机变量,$\mu = \mathrm{E}(X)$ 表示收入期望,那么基尼平均差与基尼系数满足等式 $\Gamma = \mu G$,G 表示基尼系数。与基尼系数对应的洛伦兹曲线就可以通过平面上的点(F_i, L_i)来描述,其中 $F_i = p_1 + p_2 + \cdots + p_i$,$L_i = (x_1p_1 + x_2p_2 + \cdots + x_ip_i)/\mu$,并记 $F_0 = 0$,$L_0 = 0$。F_i、L_i 分别被称为至第 i 个单元的累计人口份额、累计收入份额$(i = 0, 1, \cdots, n)$。

由式(1)可以得到基尼平均差的基本计算公式:

$$\Gamma = \mu\sum_{i=1}^{n}(F_iL_{i-1} - F_{i-1}L_i) \tag{2}$$

这是因为:

$$\Gamma = \sum_{i=1}^{n}\sum_{j=1}^{i}(x_i - x_j)p_ip_j = \sum_{i=1}^{n}\left(p_ix_i\sum_{j=1}^{i}p_j - p_i\sum_{j=1}^{i}x_jp_j\right) = \mu\sum_{i=1}^{n}\{(L_i - L_{i-1})F_i - (F_i - F_{i-1})L_i\}$$

$$= \mu\sum_{i=1}^{n}(F_iL_{i-1} - F_{i-1}L_i)$$

基尼平均差还有其他多种算法,这里给出两种算法的相关定理。

定理 1: 基尼平均差等于收入赤字的线性组合。

证明: 将式(2)稍加变形可以得到:

$$\Gamma = \mu\sum_{i=1}^{n}(F_iL_{i-1} - F_{i-1}L_i) = \mu\sum_{i=1}^{n}(F_i - L_i)F_{i-1} - \mu\sum_{i=1}^{n}(F_{i-1} - L_{i-1})F_i = \mu\sum_{i=1}^{n}(F_{i-1} - L_{i-1})(F_i -$$

$$F_{i-2}) = \mu\sum_{i=1}^{n}(F_{i-1} - L_{i-1})(p_i + p_{i-1}) \tag{3}$$

洛伦兹曲线上每一点的横、纵坐标即累计人口份额与累计收入份额两者之差 $Fi-Li$ 称为收入赤字式,式(3)表明基尼平均差是收入赤字的线性组合。

定理 2: 基尼平均差等于各项收入的线性组合。

证明: 将式(2)稍加变形还可以得到:

$$\Gamma = \mu\sum_{i=1}^{n}(F_iL_{i-1} - F_{i-1}L_i) = \mu\sum_{i=1}^{n}(L_i - L_{i-1})(F_i + F_{i-1} - 1) = \sum_{i=1}^{n}x_i(F_i + F_{i-1} - 1)p_i$$

$$= \sum_{i=1}^{n}x_i\omega_ip_i = Cov(X, \omega) \quad \omega_i = F_i + F_{i-1} - 1 \tag{4}$$

① 基尼教授 1912 年给出的计算公式为$\frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|$,即 n 个不同个体两两收入差距的平均值。

其中分布列$\{\omega_i\}$是收入分布F_i的函数$(i=0,1,\cdots,n)$，满足$\omega_1 p_1+\cdots+\omega_n p_n=0$[7]。式(4)又称为基尼平均差的协方差算法①，它不同于 Lerman 和 Yitzhaki 的$Cov(X, F(x))$算法[8]，就离散数据而言后者只是式(4)在个体数据下的特例。基尼平均差的概念可以推广到连续型随机变量，协方差算法是进行基尼平均差理论与应用研究的重要工具。

(二)拓展基尼平均差的概念和性质

Yitzhaki 拓展基尼系数的定义首先是从连续型拓展基尼平均差切入的[9]。通过考察其做法可以得出:

$$\Gamma(\nu)=\int_a^b[1-F(x)]dx-\int_a^b[1-F(x)]^{\nu}dx \tag{5}$$

其中$F(x)$为随机变量X的分布函数，ν称为不平等厌恶系数，当$0<\nu<1$时称为偏好不平等，当$\nu=1$时称为不平等中性，当$\nu>1$时称为厌恶不平等。式(5)称为拓展基尼平均差，不平等厌恶参数$\nu=2$时它就是前面定义的基尼平均差。Yitzhaki 根据洛伦兹曲线的定义并利用分部积分，得到它的一个等价形式:

$$\Gamma(\nu)=\mu\int_0^1(F-L)\nu(\nu-1)[1-F(\eta)]^{\nu-2}dF \tag{6}$$

式(6)的拓展基尼平均差可以看作是对收入赤字的加权积分，它与 Kakwani 给出的贫困测度指数是一致的[10]。

利用 Chotikapanich 和 Griffiths 的拓展基尼系数离散化公式及基尼平均差与基尼系数的关系[11]，可以定义离散型随机变量的拓展基尼平均差。

定义 2: 不平等参数下离散数据的基尼平均差$\Gamma(\nu)$为$(\nu>0)$:

$$\Gamma(\nu)=\sum_{i=1}^{n}p_i x_i\left(1+\frac{(1-F_i)^{\nu}-(1-F_{i-1})^{\nu}}{p_i}\right) \tag{7}$$

将式(7)中的大括号部分记为ω_i，容易证明$p_1\omega_1+\cdots+p_n\omega_n=0$。因此可以得到拓展基尼平均差的协方差表达式:

$$\Gamma(\nu)=Cov(X,\omega),\quad \omega_i=1+[(1-F_i)^{\nu}-(1-F_{i-1})^{\nu}]/p_i,\ i=1,2,\cdots,n \tag{8}$$

由式(7)通过恒等变形可以得到拓展基尼平均差关于洛伦兹曲线中收入赤字的表达式:

$$\Gamma(\nu)=\mu\sum_{i=1}^{n-1}(F_i-L_i)\Delta\omega_i\quad \Delta\omega_i=\omega_{i+1}-\omega_i \tag{9}$$

这是因为:

$$\Gamma(\nu)=\mu\sum_{i=1}^{n}(L_i-L_{i-1})\omega_i=\mu\sum_{i=1}^{n}(L_i-F_i+F_{i-1}-L_{i-1}+p_i)\omega_i=\mu\sum_{i=1}^{n}(F_{i-1}-L_{i-1})\omega_i-$$

$$\mu\sum_{i=1}^{n}(F_i-L_i)\omega_i=\mu\sum_{i=1}^{n-1}(F_i-L_i)\Delta\omega_i\qquad \Delta\omega_i=\omega_{i+1}-\omega_i$$

利用分部积分公式，由式(5)容易推出收入连续分布条件下拓展基尼平均差公式:

$$\Gamma(\nu)=-\nu Cov(X,(1-F(x))^{\nu-1}) \tag{10}$$

式(10)是 Schechtman 等人[12]给出的收入连续分布下的拓展基尼平均差的定义②。他们默认该公式也同样适用于离散数据。当$\nu=2$时，由式(10)可以得到$\Gamma=2Cov(X, F)$，它与式(4)结果

① 这里的符号 Cov 与通常的协方差定义不同，可以看作是双变量二维分布列仅主对角线上不为 0 的特例，即双变量之间具有一一映射关系，Yitzhaki(1996)将其对应于(x_i,y_i)的双变量分布列。如不特别指出，本文的双变量协方差都隐含存在这种一一映射。

② 他们将不平等厌恶参数$\nu-1$直接换为ν，因而取值范围$\nu>-1$。

显然是不同的,在利用样本计算基尼平均差时,只有在等概率条件下两者才能相等。尽管部分学者注意到两种算法的差异[12][13],但没有引起相关研究者的足够重视,原因是没人给出离散数据下拓展基尼平均差的对应算法。因此在多数情况下式(10)只是作为离散数据的近似算式。可以验证,式(7)在 $\nu=2$ 情形下的结果是式(4),称式(4)为式(7)的普通形式。

三、基尼加权线性回归分析

对多元线性模型的参数估计,下面将采用斜率权数法和最小基尼平均差方法进行讨论。

(一) 斜率权数法

1. 一元线性模型。

Olkin 和 Yitzhaki 对一元线性模型的参数估计使用了斜率权数法[14]。设一元线性模型为:

$$Y = a + bX + \varepsilon \tag{11}$$

其中 X 为自变量、Y 为因变量,ε 为误差扰动项。

为了估计模型参数获得了 n 个观测值:(x_1, y_1),(x_2, y_2),…,(x_n, y_n),他们给出了斜率权数法的一般形式:

$$\hat{b} = \sum_{i,j} w_{ij} m_{ij}, \quad m_{ij} = \frac{y_i - y_j}{x_i - x_j}, \quad \sum_{ij} w_{ij} = 1$$

斜率权数法被认为是一种非参数方法,但不少参数估计的结果都可以用它来表达。Yitzhaki[5]将 n 个观测值按 x 的取值递增排序,通过相当复杂的运算给出了式(11)参数 b 最小二乘估计以相邻两点的斜率权数法表达式:

$$\hat{b}_{OLS} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n-1} w_i b_i,\ w_i > 0,\ \sum_{i=1}^{n-1} w_i = 1,\ b_i = \frac{\Delta y_i}{\Delta x_i},$$

$$w_i = \frac{\left(\sum_{j=i}^{n-1} i(n-j)\Delta x_j + \sum_{j=1}^{i-1} j(n-i)\Delta x_j\right)\Delta x_i}{\sum_{k=1}^{n-1}\left(\sum_{j=k}^{n-1} k(n-j)\Delta x_j + \sum_{j=1}^{k-1} j(n-k)\Delta x_j\right)\Delta x_k} \tag{12}$$

利用本文介绍的基尼平均差方法,可以极大地简化以上演算过程。不同于个体数据的处理,组数据可以假设 n 个观测值相应的概率为 $p_1, p_2, \cdots, p_n$。通过加权最小二乘法可以得到参数 b 的估计:

$$\hat{b}_{OLS} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)p_i}{\sum_{i=1}^{n}(x_i - \mu_x)^2 p_i}$$

接下来分别对分子、分母进行处理:

$$\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)p_i = \sum_{i=1}^{n}(x_i p_i - \mu_x p_i)y_i = \mu_x \sum_{i=1}^{n}\left[L_i - L_{i-1}) - (F_i - F_{i-1})\right]y_i =$$

$$\mu_x\left[\sum_{i=1}^{n}(F_{i-1} - L_{i-1})y_i - \sum_{i=1}^{n}(F_i - L_i)y_i\right] = \mu_x \sum_{i=1}^{n-1}(F_i - L_i)\Delta y_i$$

用 x_i 代替 y_i,$i=1,\cdots,n$,可以得到分母关于收入赤字的表达式。于是有:

$$b_i = \frac{\Delta y_i}{\Delta x_i}, \quad w_i = \frac{(F_i - L_i)\Delta x_i}{\sum_{i=1}^{n-1}(F_i - L_i)\Delta x_i} \tag{13}$$

式(13)适用于组数据,当 n 个观测值等概率时就是个体数据的结果。该公式推导简单且便于记忆,其中 F_i、L_i 分别是 X 的累计人口份额(分布函数)和累计收入份额。

类似于式(11)参数 b 的最小二乘估计(OLS),Yitzhaki[5]给出了拓展基尼回归系数的定义($\nu=$

2 为基尼回归系数)：

$$\hat{b}_{OLS} = \frac{Cov(Y,X)}{Cov(X,X)},\ \hat{b}(\nu) = \frac{Cov(Y,[1-F(X)]^{\nu-1})}{Cov(X,[1-F(X)]^{\nu-1})}$$

并给出了相应的斜率权数法表达式：

$$b_i = \frac{\Delta y_i}{\Delta x_i},\ w_i = \frac{[n^{\nu-1}(n-i)-(n-i)^{\nu}]\Delta x_i}{\sum_{k=1}^{n-1}[n^{\nu-1}(n-k)-(n-k)^{\nu}]\Delta x_k} \tag{14}$$

利用本文提出的拓展基尼平均差式(8)的算法，可以得到以下的结果：

$$\hat{b}(\nu) = \frac{Cov(Y,\omega)}{Cov(X,\omega)},\ b_i = \frac{\Delta y_i}{\Delta x_i},\ w_i = \frac{[1-F_i-(1-F_i)^{\nu}]\Delta x_i}{\sum_{k=1}^{n-1}[1-F_k-(1-F_k)^{\nu}]\Delta x_k} \tag{15}$$

其中利用了以下恒等式：

$$Cov(Y,\omega) = \sum_{i=1}^{n-1}[1-F_i-(1-F_i)^{\nu}]\Delta y_i$$

证明：直接由式(8)的协方差定义和 ω_i 的表达式进行恒等变形：

$$Cov(Y,\omega) = \sum_{i=1}^{n}[p_i+(1-F_i)^{\nu}-(1-F_{i-1})^{\nu}]y_i = \sum_{i=1}^{n}([F_i+(1-F_i)^{\nu}]-[F_{i-1}+(1-F_{i-1})^{\nu}])y_i = y_n-y_1-\sum_{i=1}^{n-1}[F_i+(1-F_i)^{\nu}]\Delta y_i = \sum_{i=1}^{n-1}[-F_i-(1-F_i)^{\nu}]\Delta y_i$$

式(15)适用于组数据，当 n 个观测值等概率时就是个体数据的结果。对于个体数据该结果与式(14)完全一致，说明拓展基尼回归系数的连续分布与离散分布定义的不同造成两者的差异只是出现在组数据上。在获得参数 b 的估计之后，采用回归方程经过均值点估计参数 a，多元情形的截距项也是如此。

2. 多元线性模型。

Schechtman 等人[6]尝试对多元线性模型采用斜率权数法来估计参数。设多元线性模型为：

$$Y = \alpha+\beta_1X_1+\cdots+\beta_kX_k+\varepsilon \tag{16}$$

经过各变量的均值中心化先除去了参数 α，他们给出了与参数最小二乘估计方法类似的正则方程：$V'X\beta = V'Y$，其中 V 用本文的组数据公式可以表达为：

$$V' = (\omega_{ij})_{k\times n},\ \omega_{ij} = F_j(x_i)+F_{j-1}(x_i)-1 \tag{17}$$

这样参数向量 β 的估计可以由下式给出：

$$\hat{\beta}_{k\times1} = (Cov(\omega_i,x_j))^{-1}_{k\times k}(Cov(\omega_i,y_j))_{k\times1} \tag{18}$$

其中 ω_j 表示由自变量 x_j 分布函数根据式(4)计算的分布列，即分布函数的函数。当 $k=1$ 时，由式(18)可以得到式(13)的结果。当 $k>2$ 时，由式(18)可以得到 β 各分量估计都可以表述为因变量 y 与各自变量 x_i 斜率的线性组合，即斜率权数法的解。对于拓展基尼平均差，将式(17)的计算公式用式(8)代替也可以得到类似结果，且允许不同自变量取不同的不平等厌恶参数[6]。由于本文协方差公式将不平等厌恶参数 ν 内化在组合系数 ω 中，计算处理十分方便。根据正则方程容易得到 $V'\varepsilon = V'Y-V'X\beta = 0$，即各个解释变量的组合系数与误差项的协方差满足：

$$Cov(e,\omega(x_i)) = 0\quad i=1,\cdots,k \tag{19}$$

(二)最小基尼平均差法

1. 一元线性模型。

对于式(11)，假设 b 就是在 n 个观测值给定条件下使模型残差 e 分布列基尼平均差最小的参数估计，即：

$$\Gamma(e) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|e_i-e_j|p_ip_j = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i-y_j-b(x_i-x_j)|p_ip_j$$

于是利用绝对值不等式有:

$$0 \leqslant \Gamma(e) \leqslant \frac{1}{2}\sum_{i,j}|y_i - y_j|p_ip_j + \frac{1}{2}|b|\sum_{i,j}|x_i - x_j|p_ip_j = \Gamma(y) + |b|\Gamma(x)$$

它表明残差基尼平均差的有界性。不妨再设 n 个残差已按从小到大排列 $e_1 \leqslant e_2 \leqslant \cdots \leqslant e_n$,类似于式(4)的推导可得①:

$$\Gamma(e) = \sum_{i>j}(e_i - e_j)p_ip_j = \sum_{i>j}(y_i - y_j)p_ip_j - b\sum_{i>j}(x_i - x_j)p_ip_j = Cov(y,\omega(e)) - bCov(x,\omega(e))$$

其中 $\omega(e)$ 表示由残差分布计算的分布函数的函数。根据极值条件上式两边对 b 求导应为 0,可以得到相应的正则方程:

$$Cov(x,\omega(e_G)) = 0 \tag{20}$$

其中基尼平均差最小时残差记为 e_G,显然由式(20)无法得到 b 的解析表达式。

为了求解使残差基尼平均差最小的参数估计 b,一般情况下可以从 b 的最小二乘解出发,采取网格搜寻法获得参数 b 的估计(假定曲面或曲线是光滑的)。

由式(19)可以得到一元线性模型的两个性质:

$$Cov(\hat{y},\omega(e_G)) = Cov(a + bx,\omega(e_G)) = bCov(x,\omega(e_G)) = 0$$

$$Cov(y,\omega(e_G)) = \Gamma(e_G)$$

2. 多元线性模型。

对于式(16),对应残差的基尼平均差为:

$$\Gamma(e) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|e_i - e_j|p_ip_j = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j - \beta_1(x_{1i} - x_{1j}) - \cdots - \beta_k(x_{ki} - x_{kj})|p_ip_j$$

由绝对不等式性质可以得到:

$$0 \leqslant \Gamma(e) \leqslant \Gamma(y) + |\beta_1|\Gamma(x_1) + \cdots + |\beta_k|\Gamma(x_k)$$

关于偏导数等于 0 得到正则方程:

$$Cov(x_i,\omega(e_G)) = 0,\ i = 1,\cdots,k \tag{21}$$

一般情况下也是从 β 的最小二乘解出发,采取网格搜寻获得参数 β 的估计,为了加快搜索的收敛速度可以通过诸如梯度法等一些改进方法实施。

四、拟合优度与假设检验

(一)总体的一致估计量

设 n 个观测值 (x_i, y_i) 按 x 从小到大排列 $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$,作为一元线性模型斜率权数法的估计之一,有以下表达式:

$$\hat{b} = \sum_{i>j}w_{ij}m_{ij},\ m_{ij} = \frac{y_i - y_j}{x_i - x_j},\ w_{ij} = (x_i - x_j)/\sum_{i>j}(x_i - x_j)$$

它实际就是 Olkin 和 Yitzhaki [13] 定义的基尼回归系数,记 b_{NG},有:

$$b_{NG} = \sum_{i>j}(y_i - y_j)/\sum_{i>j}(x_i - x_j) = Cov(y,\omega(x))/Cov(x,\omega(x))$$

Schechtman 和 Yitzhaki [14] 发现上式第一个等号右侧的分子、分母两个和式都可以用 U - 统计量来表达,因此样本基尼回归系数就是总体基尼回归系数 $Cov(Y,F(X))/Cov(X,F(X))$ 的一致估计,对于基尼加权回归也是如此。而且利用 U - 统计量还可以进行相关参数的假设检验。

① 当不按变量值从小到大排序时,(F_i,L_i) 就不构成洛伦兹曲线,称为集中曲线。但同样可以得到类似的协方差公式,只是不再具有非负性。

(二) 基尼加权线性回归的拟合优度

Olkin 和 Yitzhaki 给出了多元线性模型基尼回归拟合优度的定义:

$$GR = 1 - \left(\frac{Cov(e,\omega(e))}{Cov(y,\omega(y))}\right)^2$$

类似于最小二乘法,斜率权数法估计参数通常不会使回归误差的基尼平均差大于总基尼平均差。因此,拟合优度 GR 的值在 0 ~ 1 之间,拓展基尼平均差也是如此。

(三) 斜率参数的假设检验

基尼回归系数对应于一元线性模型的残差还满足以下关系:

$$Cov(e_{NG},\omega(x)) = Cov(y - a - b_{NG}x,\omega(x)) = Cov(y,\omega(x)) - b_{NG}Cov(x,\omega(x)) = 0$$

类似于 Schechtman 和 Yitzhaki[14] 的处理,可以得到总体 $Cov(\varepsilon,F_X(X))$ 和 $Cov(X,F_\varepsilon(\varepsilon))$ 的 U - 统计量估计分别为 $Cov(e,\omega(x))$、$Cov(x,\omega(e))$,由于 U - 统计量的渐近正态性质①,利用它们能够检验式(19) 和式(20) 的显著性水平,从而判断线性关系中解释变量的显著性。

五、应用举例

近年来各级政府大力发展旅游业,作为推动地方产业升级和减贫增收的重要举措。关于旅游业特别是入境旅游在缩小收入差距方面的作用,长期以来受到了不少管理者和学者的关注。理论界存在着两种观点,一种是认为入境旅游增加外汇收入和就业机会,促进外商投资,能够缩小城乡收入差距;另一种则认为入境旅游与外商投资关系密切,而外商投资为方便产品的销售和服务更倾向于经济发达地区,从而扩大地区间的收入差距。两种观点看似矛盾,但实际上城乡收入差距与地区间的收入差距是两个完全不同的概念。我国的城乡收入差距是二元户籍制度的产物,与城乡劳动者所从事的主要职业有关;地区间的收入差距如省间收入差距,则与各省份的资源禀赋和发展机遇等因素相关联。国内外现有的实证研究主要是针对地区间的收入差距,形成了以下三种结论:入境旅游扩大了地区间的收入差距、入境旅游缩小了地区间的收入差距、入境旅游开始扩大了收入差距但一定规模之后缩小收入差距。

本文的应用举例分别采用了 2015 年省域截面数据,以及 1994—2015 年 31 个省市总量时间序列数据。数据以收入基尼系数为因变量,入境旅游的外汇收入为主要解释变量,同时引入了财政预算支出、外商投资额和国内旅游收入三个控制变量。省域截面数据中收入基尼系数使用各省份的城镇、农村人均可支配收入和人口数计算得到,入境旅游外汇收入(亿元,通过美元年平均汇率换算)、地区财政预算支出(亿元)、外商投资额(亿元,通过美元年平均汇率换算) 取自《中国统计年鉴 2016》,各省国内旅游收入取自 31 个省市 2016 年的统计年鉴。总量时间序列中收入基尼系数采用各省 *GDP* 作为宏观收入结合人口数计算得到,各省的 *GDP*、年末人口数,以及总量的中央和地方财政预算支出(F)、外商直接投资(FDI)、入境旅游外汇收入(TR)、美元年平均汇率取自相应年度的《中国统计年鉴》,同时以 1978 年为基期利用商品零售价格指数(RPI) 对各年财政支出、外商直接投资进行缩减,居民消费价格指数(CPI) 对入境、国内旅游收入进行缩减。表 1 给出了相应收入基尼系数的计算结果,作为随后基尼加权回归分析中因变量的基础数据。

其中测度城乡收入差距的基尼系数计算公式为:

$$G_i = \frac{(x_i - y_i)\,q_{ci}q_{ni}}{(q_{ci}x_i + q_{ni}y_i)(q_{ci} + q_{ni})} \quad i = 1,2,\cdots,31$$

① 其中在 x 取值未排序的情况下 $Cov(e,\omega(x_i)) = \sum_{i<j}[(e_j - e_i)I_{\{x_i<x_j\}} + (e_i - e_j)I_{\{x_i>x_j\}}]p_ip_j$,$I$ 为示性函数。分布列平方的数学期望就是 e 的方差,U 统计量方差估计参见 Olkin 和 Yitzhaki(1992)。

表 1　　各省城乡收入差距测度(2015 年)和省间收入基尼系数(1994 –2015 年)

省份	基尼系数	省份	基尼系数	省份	基尼系数	年度	基尼系数	年度	基尼系数
北　京	0. 0777	安　徽	0. 2125	四　川	0. 2229	1994	0. 2597	2005	0. 2716
天　津	0. 0715	福　建	0. 1755	贵　州	0. 2867	1995	0. 2569	2006	0. 2682
河　北	0. 2007	江　西	0. 2012	云　南	0. 2765	1996	0. 2537	2007	0. 2653
山　西	0. 2195	山　东	0. 1938	西　藏	0. 2651	1997	0. 2578	2008	0. 2569
内蒙古	0. 2088	河　南	0. 2065	陕　西	0. 2414	1998	0. 2624	2009	0. 2446
辽　宁	0. 1684	湖　北	0. 1821	甘　肃	0. 2907	1999	0. 2685	2010	0. 2409
吉　林	0. 1781	湖　南	0. 2222	青　海	0. 2548	2000	0. 2577	2011	0. 2267
黑龙江	0. 1689	广　东	0. 1639	宁　夏	0. 2208	2001	0. 2770	2012	0. 2153
上　海	0. 0655	广　西	0. 2421	新　疆	0. 2416	2002	0. 2694	2013	0. 2077
江　苏	0. 1544	海　南	0. 1977			2003	0. 2764	2014	0. 2045
浙　江	0. 1412	重　庆	0. 1924			2004	0. 2597	2015	0. 2033

这里分别用 x_i、y_i 表示第 i 个省份城镇、农村居民人均可支配收入，q_{ci}、q_{ni} 表示第 i 个省份城镇、农村年末人口。该公式表明，城乡收入基尼系数是城镇与农村人均可支配收入差值的乘数。尽管省域截面数据的城乡收入基尼系数与 31 个省市总量数据的地区间的收入基尼系数计算方法有所不同，但模型都使用式(11)或式(16)的线性形式。

(一)省域数据的截面回归分析

采用最小二乘法将城乡收入基尼系数关于入境旅游收入进行了一元线性回归(模型 I)，发现入境旅游收入能够显著解释收入差距，参数估计值为负表明入境旅游有助于缩小城乡收入差距。其中 G 表示各省份城乡收入基尼系数。

模型 I: $G = \alpha + \beta_1 TR + \varepsilon$

再加入入境旅游收入的平方作为解释变量，参数估计显著不为 0 且提高拟合优度，说明入境收入与城乡收入差距具有 U 型关系。当依次增加财政支出、外商直接投资和国内旅游收入等控制变量，调整 R^2 下降说明不宜加入这些变量。于是得到一元二次模型

模型 II: $G = \alpha + \beta_1 TR + \beta_2 TR^2 + \varepsilon$

其中 Y、TR 分别表示城乡收入基尼系数和入境旅游收入。然后以参数的 OLS 估计为初始值，利用网格搜寻法确定相应模型参数的误差项最小基尼平均差估计。估计结果见表 2，表中用 MG 表示最小基尼平均差方法，LAD 表示最小绝对偏差法(下同)。

表 2　　省域截面数据(2015 年)模型的加权参数估计

	模型 I			模型 II		
	MG	OLS	LAD	MG	OLS	LAD
常数项	0. 2056	0. 2541	0. 2560	0. 2258	0. 2577	0. 2566
	(0. 0000)	(0. 0000)	(0. 0000)	(0. 0000)	(0. 0000)	(0. 0000)
TR	-5.1282×10^{-5}	-4.9788×10^{-4}	-5.2239×10^{-4}	-2.8515×10^{-4}	-7.6196×10^{-4}	-7.3854×10^{-4}
	(0. 0335)	(0. 0000)	(0. 0006)	(0. 0000)	(0. 0000)	(0. 0004)
TR^2				2.0706×10^{-7}	6.4469×10^{-7}	5.8864×10^{-7}
				(0. 0000)	(0. 0216)	(0. 0286)
拟合度	0. 2025	0. 5558	0. 2528	0. 3790	0. 6333	0. 3490

注: MG 方法的常数项都由回归方程过均值点确定。括号内的数据为反映系数显著性的截尾概率即 p 值，其中 MG 方法通过 Jackknife 重抽样得到，OLS 和 LAD 方法的显著性结果直接由 Eviews8. 0 软件得出。下表同。

通常就 OLS 法而言截面数据线性回归的拟合优度不高，但在本例中还是较为理想的。从一元一次回归模型的结果看，三种方法对斜率参数的估计符号一致，具有相同的经济意义即入境旅游收入能够缩小城乡收入差距。一元二次回归模型的参数估计结果也进一步揭示了入境旅游与城乡居民收入差距长期趋势的倒 U 型关系: 入境旅游收入在相当长一段时间内能够缩小收入差距，但超

过一定数值后会转而推动收入差距的扩大。如有一些学者认为香港地区前些年由于入境旅游人数的持续高涨，在一定程度上推高了香港的房租价格和部分基本生活用品的费用，扩大了穷人与富人的收入差距。

（二）总量数据回归分析

对 31 个省市总量数据模型的参数估计，根据考察对象和控制变量的重要性采用 OLS 法进行模型变量选择。当仅用入境旅游收入对地区间的收入差距（因变量）线性回归时，发现拟合优度偏低（时间序列拟合优度通常较高）且系数为负不符合预期，加入入境旅游收入的平方项后拟合优度虽然有所提高，但斜率参数的估计都不显著，于是剔除平方项加入了国内旅游总收入变量，得到了效果较为理想、仅含两类旅游收入的简单模型

模型 III：$G = \alpha + \beta_1 TR + \beta_2 DTR + \varepsilon$

对三个控制变量的选择主要考虑国内旅游收入与入境旅游收入具有较强的关联性，财政支出作为政府进行二次分配的重要手段具有平衡地区间收入差距的功能，外商直接投资通常被认为对省间收入差距也具有重要影响。通过控制变量的相关性分析，发现国内旅游收入与财政支出的相关系数为 0.9866，为了避免方程的共线性由拟合优度准则在收入控制变量模型中仅保留财政支出。于是得到模型表达式

模型 IV：$G = \alpha + \beta_1 TR + \beta_2 TR^2 + \beta_3 F + \beta_4 FDI + \varepsilon$

表 3 列出了模型 III 与模型 IV 参数估计的结果。

表 3　　总量数据模型（1994 – 2015 年）的参数估计

	两类收入简单模型（模型 III）			收入控制变量模型（模型 IV）		
	MG	OLS	LAD	MG	OLS	LAD
常数项	0.2818	0.2621	0.2563	0.3463	0.2895	0.2893
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
TR	1.0137×10^{-4}	4.3014×10^{-5}	5.0499×10^{-5}	4.1793×10^{-4}	1.7734×10^{-4}	1.6943×10^{-4}
	(0.0455)	(0.0086)	(0.0452)	(0.0000)	(0.0000)	(0.0006)
TR^2				-1.9256×10^{-7}	-8.1711×10^{-8}	-7.5778×10^{-8}
				(0.0390)	(0.0005)	(0.0097)
DTR	-4.4532×10^{-5}	-1.8897×10^{-5}	-1.9430×10^{-5}			
	(0.0256)	(0.0000)	(0.0001)			
F				-6.0062×10^{-6}	-2.5486×10^{-6}	-2.5054×10^{-6}
				(0.0000)	(0.0000)	(0.0000)
FDI				-1.0741×10^{-4}	-4.5578×10^{-5}	-4.4584×10^{-5}
				(0.0000)	(0.0058)	(0.0358)
拟合度	0.8854	0.8842	0.6373	0.9645	0.9551	0.8163

从旅游两类收入模型看，入境旅游倾向于扩大地区间的收入差距，符合理论预期；国内旅游则缩小地区间的收入差距。从收入控制变量模型看，财政支出和外商直接投资都倾向于缩小地区间的收入差距。而且入境旅游与地区间的收入差距具有倒 U 型关系，即长期趋势表现为随着入境旅游收入的增长，入境旅游会从扩大地区间的收入差距向缩小地区间的收入差距转化。

（三）拓展基尼回归分析

斜率权数法在省际入境旅游对城乡居民收入差距（2015 年的截面数据）的作用，以及入境旅游对各省市地区间收入差距的影响，通过拓展基尼回归一并分析（基尼平均差仅仅是 $\nu = 2$ 的情形）。由于斜率权数法对于一元线性回归模型斜率的解并没有固定的形式，但为了方便起见本文将讨论限制在正则方程式（18）的框架之下。如由模型 IV，取不平等厌恶参数 $\nu = 0.5$，根据式（18）可以计算出四个解释变量系数的斜率权数法估计：

$$\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\\ \hat{\beta}_3\\ \hat{\beta}_4\end{pmatrix}=\begin{pmatrix}-164.2322 & -226160.7141 & -6098.5802 & -91.9005\\ -164.2732 & -226204.0077 & -6143.0821 & -92.3249\\ -158.4991 & -219416.3175 & -7887.0325 & -100.8000\\ -70.9674 & -70423.8538 & -4437.2203 & -117.6472\end{pmatrix}^{-1}\begin{pmatrix}9.2823\times10^{-3}\\ 9.4090\times10^{-3}\\ 1.4223\times10^{-2}\\ 9.7928\times10^{-3}\end{pmatrix}$$

$$=\begin{pmatrix}6.7351\times10^{-4}\\ -4.1884\times10^{-7}\\ -1.2244\times10^{-6}\\ -1.9261\times10^{-4}\end{pmatrix}$$

其中逆矩阵中对角线上的元素为各个解释变量的拓展基尼平均差,它们都等于负值是因为 $\nu=0.5<1$ 对应于偏好不平等①。对于模型 I、模型 II 和模型 III 本文也做了类似的处理②。

估计结果显示,在使用基尼平均差($\nu=2$)的情况下即总体而言,各个解释变量系数参数估计的符号都与理论预期一致,说明在对收入差距的作用方向上效果与 MG、OLS 和 LAD 估计方法相同。然而拓展基尼平均差在不同的不平等厌恶参数取值下,也可能对解释变量的系数估计带来一些意想不到的变化并带来新的经济学分析视角。

模型 I 中入境旅游收入的系数估计是最为稳健的,无论是赋予入境旅游高收入省份更大的权重($\nu=0.5$,偏好不平等),还是赋予低收入省份更大的权重($\nu>1$,厌恶不平等),都能够保持缩小城乡收入差距的作用方向不变,力度大小也相对均衡。

模型 II 中入境旅游对城乡居民收入的影响,随着不平等厌恶参数的增加从现期(一次项)缩小差距向扩大差距转变,而远期(二次项)则相反。这种变化表明,对于从入境旅游中获得高收入的省份($\nu=0.5$),现期入境旅游缩小城乡居民收入差距的力度最大,而远期则体现 U 型关系;而对获得较少入境旅游收入的省份($\nu=8$),入境旅游收入对城乡居民收入差距的影响远期呈现倒 U 型关系,而现期则扩大城乡居民的收入差距。

模型 III 与模型 I 类似,入境旅游和国内旅游对地区间收入差距的影响都十分稳健。无论对入境旅游高收入的省份,还是对入境旅游低收入的省份,入境旅游都扩大地区间的收入差距,而且力度随着入境旅游收入的减小而有所增强;国内旅游的作用则恰好与入境旅游相反。

模型 IV 中入境旅游无论现期还是远期对地区间的收入差距影响都保持作用方向一致,大小平稳变化,与地区间的收入差距呈现倒 U 型关系;财政支出对地区间收入差距的作用方向一致,力度大小随着对较低财政支出的省份赋予较大权重而有所增强,说明对财政支出较低省份更多的扶持可以增大缩小地区收入差距的效果。外商直接投资随着不平等厌恶参数的增大,不仅缩小地区间收入差距的作用效果逐渐减弱,而且从缩小差距向扩大差距转变。说明对于较多获得外商直接投资省份构成的群体($\nu=0.5$),外商直接投资可以缩小群体内部的收入差距;但是对于较少获得外商直接投资的省份($\nu=8$),则倾向于扩大他们的收入差距。

通过前面的理论推导和以上应用举例,可以发现基尼加权回归方法的一般性应用条件:一是要求自变量与因变量的条件期望之间客观上存在线性关系,二是误差项应满足相互独立且与自变量不相关(误差项并不要求满足同正态分布)。该方法更适合于与福利经济学相关的应用情形,当然

① Yitzaki(1983)证明了拓展基尼平均差是不平等厌恶参数 ν 的增函数,由定义当 $\nu=1$ 为不平等中性时可得 $\Gamma(1)=0$。因此当 $0<\nu<1$ 为偏好不平等时 $\Gamma(\nu)<0$,当 $\nu>1$ 为厌恶不平等时 $\Gamma(\nu)>0$。

② 限于篇幅,本文未给出估计结果,有兴趣读者可向作者索取。

并不局限于此。

参考文献

[1] Bassett G Jr, Koenker R. Asymptotic Theory of Least Absolute Error Regression [J]. Journal of American Statistical Association, 1978, 73(363):618 - 622.

[2] Tibshiran R. Regression Shrinkage and Selection via the Lasso [J]. Journal of Royal Statistical Society, Series B, 58(1):267 - 288.

[3] Koenker R, Bassett G Jr. Regression Quantiles [J]. Econometrica, 1978, 46(1):33 - 50.

[4] Olkin I, Yitzhaki S. Gini Regression Analysis [J]. International Statistical Review, 1992, 60(2):185 - 196.

[5] Yitzhaki S. On using Linear Regressions in Welfare Economics [J]. Journal of Business & Economic Statistics, 1996, 14(4):478 - 486.

[6] Schechtman E, Yitzhaki S, Artsev Y. Who does not Respond in the Household Expenditure Survey: An Exercise in Extended Gini Regressions [J]. Journal of Business & Economic Statistics, 2008, 26(3):329 - 344.

[7] 戴平生. 拓展基尼系数及其应用的拓展研究[J]. 统计研究, 2013(9): 69 - 78.

[8] Lerman R I, Yitzhaki S. A Note on the Calculation and Interpretation of Gini Index [J]. Economics Letter, 1984, 15(3 - 4):363 - 368.

[9] Yitzhaki S. On an Extension of the Gini Inequality Index [J]. International Economic Review, 1983, 24(3): 716 - 728.

[10] Kakwani N C. Income Inequality and Poverty. Methods of Estimation and Policy Applications [M]. Oxford: Oxford University Press, 1980.

[11] Chotikapanich D, Griffiths W. On Calculation of the Extended Gini Coefficient [J]. Review of Income and Wealth, 2001, 47(4): 541 - 547.

[12] Lerman R I, Yitzhaki S. Improving the Accuracy of Estimates of Gini Coefficients [J]. Journal of Econometrics, 1989, 42(1):43 - 47.

[13] Kakwani N C, Wagstaff A, Doorslaer E V. Socioeconomic Inequalities in Health: Measurement, Computation, and Statistical Inference [J]. Journal of Econometrics, 1997, 77(1):87 - 103.

[14] Schechtman E, Yitzhaki S. A Measure of Association Based on Gini's Mean Difference [J]. Communication in Statistics-Theory and Methods, 1987, 16(1): 207 - 231.

作者简介

戴平生,男,2004 年毕业于厦门大学计划统计系,获经济学(统计)专业博士学位,现为厦门大学经济学院统计系教授、博士生导师,美国康涅狄格大学统计系访问教授(2018 年 1 月至 7 月),教育部计量经济学重点实验室(厦门大学)兼职研究员。研究方向为数量经济学、经济统计。

(责任编辑:倪立行)