doi:10. 16055/j. issn. 1672-058X. 2022. 0002. 014

工具变量线性回归模型的指数平方损失估计

张 巍[1], 杨宜平[1,2]

(1. 重庆工商大学 数学与统计学院, 重庆 400067;
2. 重庆工商大学 经济社会应用统计重庆市重点实验室, 重庆 400067)

摘 要:在讨论协变量和响应变量关系时,常会遇到内生变量,已有关于内生变量的研究大多是在最小二乘目标函数的框架下讨论,然而该方法不具有稳健性,鉴于此,本文采用指数平方损失估计方法,构造模型中回归系数的稳健估计。为了克服内生变量对估计产生的偏差,利用工具变量消除协变量的内生性,再构造回归系数的指数平方损失估计;针对指数平方损失目标函数,提出选取有效的调节参数估计过程;在一些正则条件下,研究所提出估计的渐近正态性;模拟研究比较了朴素最小二乘估计、朴素 M 估计、朴素指数平方损失估计、基于工具变量的最小二乘估计、基于工具变量的 M 估计、基于工具变量的指数平方损失估计等 6 种估计方法,模拟结果表明:本文提出的基于工具变量的指数平方损失估计有效地消除了协变量的内生性,且具有较好的稳健性;最后,利用本文提出的方法分析了孪生双胞胎"收入-教育程度"的数据。

关键词:内生变量;工具变量; 线性模型; 指数平方损失
中图分类号:O212. 7 **文献标志码**:A **文章编号**:1672-058X(2022)02-0099-08

0 引 言

回归分析是研究各种现象之间数量关系的一种常用方法,其中最常见的回归模型是线性回归模型:

$$\boldsymbol{Y}_i=\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i, i=1,2,\cdots,n$$

其中,$\boldsymbol{Y}_i\in\mathbf{R}$ 是独立同分布的响应变量,$\boldsymbol{X}_i\in\boldsymbol{R}^p$ 是 p 维协变量,$\boldsymbol{\beta}$ 是未知的参数向量。线性回归模型的估计方法层出不穷,最经典的即为最小二乘法,但该方法对误差分布要求较为严苛,如零均值、同方差假定等。在实际应用场景中,最小二乘估计的效果并不理想。为了弥补最小二乘法的不足,Wang 等[1]提出了基于指数平方损失目标函数的估计方法。该方法不需要对模型误差分布作特定的限制,且估计的稳健性由调节参数 h 控制。该方法一经提出即受到了广泛的关注。Yu 等[2]讨论了半函数线性模型的指数平方损失估计,并指出如果随机误差服从重尾分布,该方法比最小二乘法更加有效;Jiang[3]将该方法应用于部分线性模型,并表示当数据集中存在离群点时,该方法得到的参数估计量标准差和均方误差皆优于现有的其他方法。

当前关于指数平方损失方法的研究,多数文献都假定协变量是外生变量,然而在实际应用中,协变量是内生变量的情况不在少数。这种情况下,如果将协变量视为外生变量进行估计,则得到的参数估计量将不再是无偏估计。为了消除内生性带来的影响,Ashenfelter[4]提出了倍差法,Thistlethwaite 等[5]提出了断点回归方法,Donald[6]研究了工具变量法。受

收稿日期:2021-04-14; **修回日期**:2021-05-18.
基金项目:重庆市社科规划委托项目(2019WT58);重庆市自然科学基金(CSTC2020JCYJ-MSXMX0006);2018 年重庆市《统计学》研究生导师团队(YDS183002).
作者简介:张巍(1996—),男,重庆丰都人,硕士研究生,从事非参数统计研究.
通讯作者:杨宜平(1981—),女,湖北荆州人,教授,从事非参数统计及数据分析研究. Email:yyp@ ctbu. edu. cn.

到 Yang 等[7]基于工具变量对含测量误差的线性模型进行参数估计的启发,本文基于工具变量的指数平方损失方法对含内生变量的线性模型进行参数估计。

首先给出了估计过程以及调节参数 h 的选取过程;进一步,在一些正则条件下,研究了估计的渐近性质,然后通过模拟研究,比较了不同误差分布、不同样本量下朴素 M 估计、朴素最小二乘估计、朴素指数平方损失估计以及基于工具变量的 M 估计、基于工具变量的最小二乘估计、基于工具变量的指数平方损失估计等 6 种方法的优劣;最后,利用提出的方法对孪生双胞胎"收入-教育程度"数据进行了实证分析。

1 方法与主要结果

考虑如下工具变量线性回归模型:

$$\begin{cases} Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i \\ \boldsymbol{X}_i = \boldsymbol{\Gamma}\boldsymbol{Z}_i + \boldsymbol{e}_i \end{cases} \quad i = 1,2,\cdots,n$$

其中,$\boldsymbol{X}_i$ 是 p 维内生变量,$\boldsymbol{\beta}$ 是 p 维未知向量,$\boldsymbol{Z}_i$ 是 q 维工具变量,满足 $cov(\boldsymbol{Z}_i,\varepsilon_i)=0$,$\boldsymbol{\Gamma}$ 是 $p\times q$ 维矩阵,$\varepsilon_i,\boldsymbol{e}_i$ 是随机误差。下面给出 $\boldsymbol{\beta}$ 的两阶段估计过程。

第一阶段,由于 $E(\boldsymbol{Z}_i\boldsymbol{e}_i)=0$,得到 $\boldsymbol{\Gamma}$ 的最小二乘估计:

$$\hat{\boldsymbol{\Gamma}} = \boldsymbol{X}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})^{-1}$$

其中,$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$ 是 $p\times n$ 维矩阵,$\boldsymbol{Z}=(Z_1,Z_2,\cdots,Z_n)$ 是 $q\times n$ 维矩阵。于是,得到 X_i 的估计量:

$$\hat{\boldsymbol{X}}_i = \hat{\boldsymbol{\Gamma}}\boldsymbol{Z}_i$$

第二阶段,用 $\hat{\boldsymbol{X}}_i$ 替换 $\boldsymbol{X}_i$,通过最大化目标函数:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\exp\left\{-\frac{(Y_i-\hat{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta})^2}{h}\right\}$$

可以获得 $\boldsymbol{\beta}$ 的指数平方损失估计,即

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$$

目标函数 $L(\boldsymbol{\beta})$ 中的 h 是调节参数,控制着估计的稳健性和有效性。对于较大的 h,有

$$1-\exp\left(-\frac{(Y_i-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})^2}{h}\right) \approx \frac{(Y_i-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})^2}{h} = \frac{\varepsilon_i^2}{h}$$

此时,该估计类似于极端情况的最小二乘估计。对于较小的 h,$|\varepsilon_i|$ 值越大,对估计的影响越小,因此,较小的 h 将限制离群值对该估计的影响,提高估计的稳健性。下面给出调节参数 h 的选择过程:

(1) 首先获得初始估计,一般用 M 估计得到 $\tilde{\boldsymbol{\beta}}$;

(2) 随后得到 $\tilde{\varepsilon}_i = Y_i - \hat{\boldsymbol{X}}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}$;

(3) 得到本文所提出估计的渐进方差估计为

$$\tilde{r}(h) = \frac{\tilde{G}(h)}{[\tilde{F}(h)]^2}$$

其中,$\tilde{G}(h)$ 和 $\tilde{F}(h)$ 分别为

$$\tilde{G}(h) = \frac{1}{n}\sum_{i=1}^{n}\{\varphi_h'(\tilde{\varepsilon}_i)\}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{4\tilde{\varepsilon}_i^2}{h^2}\exp\left(-\frac{2\tilde{\varepsilon}_i^2}{h}\right)$$

$$\tilde{F}(h) = \frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\tilde{\varepsilon}_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{4}{h^2}\left(\tilde{\varepsilon}_i^2-\frac{h}{2}\right)\exp\left(-\frac{\tilde{\varepsilon}_i^2}{h}\right)$$

(4) 得到 $r(h)$ 的估计值 $\tilde{r}(h)$ 后,利用网格点搜索法最小化 $\tilde{r}(h)$,从而得到 h_{opt}。本文中,选取 h 的网格点是 $h=0.5\tilde{\sigma}\times 1.02^j$,$j=1,2,\cdots,100$,其中 $\tilde{\sigma}=E(\tilde{\varepsilon}_i^2)=Var(\tilde{\varepsilon}_i)$。

2 渐近性质

本节研究 $\hat{\boldsymbol{\beta}}$ 的渐近性质,首先给出正则条件如下:

C(1) 对于所有的 i,随机变量 $\boldsymbol{Z}_i$ 是有界的,并且 $\Sigma_{\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}}=E[\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}]$ 的特征值是有界的,即 $\boldsymbol{\Sigma}_{\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}}$ 是正定矩阵。

C(2) 令 $\varphi_h(t)=\exp\left(-\frac{t^2}{h}\right)$,有

$$G(x,h) = E\{[\varphi_h'(\varepsilon)]^2 \mid \boldsymbol{X}=x\} = E\left\{\frac{4\varepsilon^2}{h^2}\exp\left(-\frac{2\varepsilon^2}{h}\right) \mid \boldsymbol{X}=x\right\}$$

$$F(x,h) = E\{\varphi_h''(\varepsilon) \mid \boldsymbol{X}=x\} = E\left\{\frac{4}{h^2}\left(\varepsilon^2-\frac{h}{2}\right)\exp\left(-\frac{\varepsilon^2}{h}\right) \mid \boldsymbol{X}=x\right\}$$

且 $G(x,h)$,$F(x,h)$ 关于 x 连续,$F(x,h)<0$。

C(3) $E[\varphi_h'(\varepsilon)\mid \boldsymbol{X}=x]=0$,$E[\varphi_h''(\varepsilon)\mid \boldsymbol{X}=x]$,$E[\varphi_h'''(\varepsilon)\mid \boldsymbol{X}=x]$ 关于 x 连续。

定理 1 如果条件 C(1),C(2),C(3)皆成立,

$\boldsymbol{\beta}_0$ 是 $\boldsymbol{\beta}$ 的真实值,则

$$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\xrightarrow{d}N(0,\boldsymbol{\Sigma}_{\mathrm{ESL}})$$

其中,

$$\boldsymbol{\Sigma}_{\mathrm{ESL}}=(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{ZZ^{\mathrm{T}}}\boldsymbol{\Gamma}^{\mathrm{T}})^{-1}\frac{E[\varphi_h'(\varepsilon)]^2}{\{E[\varphi_h''(\varepsilon)]\}^2}$$

证明　由于目标函数 $L(\boldsymbol{\beta})$ 在 $\hat{\boldsymbol{\beta}}$ 取到最大值,根据最值的一阶条件,有

$$\left.\frac{\mathrm{d}L(\boldsymbol{\beta})}{\mathrm{d}\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}=\frac{\mathrm{d}\sum_{i=1}^{n}\exp\left\{\frac{-(Y_i-\hat{\boldsymbol{X}}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2}{h}\right\}}{\mathrm{d}\beta}=$$

$$\sum_{i=1}^{n}\exp\left\{-\frac{(Y_i-\hat{\boldsymbol{X}}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2}{h}\right\}\frac{-2}{h}(Y_i-\hat{\boldsymbol{X}}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})(-\hat{\boldsymbol{X}}_i)=$$

$$\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\cdot\varphi_h'(Y_i-\hat{\boldsymbol{X}}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})=0\qquad(1)$$

记 $\varepsilon_i^{*}=Y_i-\hat{\boldsymbol{X}}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$,$\varepsilon_i^{**}=Y_i-\hat{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}_0$,并将式(1)在 ε^{**} 点泰勒展开,得到

$$\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\varphi_h'(\varepsilon_i^{*})=$$

$$\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\left\{\varphi_h'(\varepsilon_i^{**})+\varphi_h''(\varepsilon_i^{**})(\varepsilon_i^{*}-\varepsilon_i^{**})+\right.$$

$$\left.\frac{1}{2}\varphi_h'''(\xi_i)(\varepsilon_i^{*}-\varepsilon_i^{**})^2\right\}=$$

$$\sum_{i=1}^{n}\hat{\boldsymbol{X}}_i\left\{\varphi_h'(\varepsilon_i^{**})+\varphi_h''(\varepsilon_i^{**})[-\hat{\boldsymbol{X}}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)]+\right.$$

$$\left.\frac{1}{2}\varphi_h'''(\xi_i)[-\hat{\boldsymbol{X}}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)]^2\right\}=0$$

其中,ξ_i 位于 ε_i^{*} 与 ε_i^{**} 之间。经过简单的计算,可以得到:

$$\frac{1}{2}\varphi_h'''(\xi_i)[-\hat{\boldsymbol{X}}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)]^2=o_p$$

进而可以得到:

$$\frac{1}{n}\{\sum_{i=1}^{n}\varphi_h'(\varepsilon_i^{**})\hat{\boldsymbol{X}}_i-$$

$$\sum_{i=1}^{n}\varphi_h''(\varepsilon_i^{**})\hat{\boldsymbol{X}}_i\hat{\boldsymbol{X}}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}+o_p(1)=0$$

随之推出:

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i^{**})\hat{\boldsymbol{X}}_i\hat{\boldsymbol{X}}_i^{\mathrm{T}}\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_h'(\varepsilon_i^{**})+o_p(1)\qquad(2)$$

于式(2),首先考虑式(2)右边,将其在 ε_i 点泰勒展开,有

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_h'(\varepsilon_i^{**})\hat{\boldsymbol{X}}_i=$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\varphi_h'(\varepsilon_i)+\varphi_h''(\varepsilon_i)(\varepsilon_i^{**}-\varepsilon_i)+o_p(1)\}\hat{\boldsymbol{X}}_i=$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\varphi_h'(\varepsilon_i)+\varphi_h''(\varepsilon_i)(\varepsilon_i^{**}-\varepsilon_i)\}\times$$

$$(\hat{\boldsymbol{X}}_i-\boldsymbol{X}_i+\boldsymbol{X}_i)+o_p(1)=$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\varphi_h'(\varepsilon_i)+\varphi_h''(\varepsilon_i)(\boldsymbol{X}_i-\hat{\boldsymbol{X}}_i)^{\mathrm{T}}\boldsymbol{\beta}_0\}(\hat{\boldsymbol{X}}_i-\boldsymbol{X}_i)+$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\varphi_h'(\varepsilon_i)+\varphi_h''(\varepsilon_i)(\boldsymbol{X}_i-\hat{\boldsymbol{X}}_i)^{\mathrm{T}}\boldsymbol{\beta}_0\}\boldsymbol{X}_i+o_p(1)=$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\varphi_h'(\varepsilon_i)+\varphi_h''(\varepsilon_i)[(\boldsymbol{\Gamma}-\hat{\boldsymbol{\Gamma}})\boldsymbol{Z}_i+\boldsymbol{e}_i]^{\mathrm{T}}\boldsymbol{\beta}_0\}$$

$$[(\hat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})\boldsymbol{Z}_i]+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\varphi_h'(\varepsilon_i)+$$

$$\varphi_h''(\varepsilon_i)[(\boldsymbol{\Gamma}-\hat{\boldsymbol{\Gamma}})\boldsymbol{Z}_i+\boldsymbol{e}_i]^{\mathrm{T}}\boldsymbol{\beta}_0\}(\boldsymbol{\Gamma}\boldsymbol{Z}_i)+$$

$$o_p(1)\triangleq I_1+I_2+o_p(1)$$

根据正则条件 C(3)、最小二乘估计的收敛速度以及参考文献[8]引理 3 可知,$I_1=o_p(1)$,故只考虑 I_2。

$$I_2=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\varphi_h'(\varepsilon_i)+\varphi_h''(\varepsilon_i)[(\boldsymbol{\Gamma}-\hat{\boldsymbol{\Gamma}})\boldsymbol{Z}_i+$$

$$\boldsymbol{e}_i]^{\mathrm{T}}\boldsymbol{\beta}_0\}(\boldsymbol{\Gamma}\boldsymbol{Z}_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\varphi_h'(\varepsilon_i)+$$

$$\varphi_h''(\varepsilon_i)[\boldsymbol{\Gamma}\boldsymbol{Z}_i-(\boldsymbol{X}\boldsymbol{Z}^{\mathrm{T}})(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})^{-1}\boldsymbol{Z}_i+$$

$$\boldsymbol{e}_i]^{\mathrm{T}}\boldsymbol{\beta}_0\}(\boldsymbol{\Gamma}\boldsymbol{Z}_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_h'(\varepsilon_i)(\boldsymbol{\Gamma}\boldsymbol{Z}_i)+$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i)\{\boldsymbol{\Gamma}\boldsymbol{Z}_i-[(\boldsymbol{\Gamma}\boldsymbol{Z}+\boldsymbol{e})\boldsymbol{Z}^{\mathrm{T}}](\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})^{-1}\boldsymbol{Z}_i+$$

$$\boldsymbol{e}_i\}^{\mathrm{T}}\boldsymbol{\beta}_0(\boldsymbol{\Gamma}\boldsymbol{Z}_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_h'(\varepsilon_i)(\boldsymbol{I}\boldsymbol{Z}_i)+$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i)(\boldsymbol{\Gamma}\boldsymbol{Z}_i)\{\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{\beta}_0-[\boldsymbol{e}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})^{-1}\boldsymbol{Z}_i]^{\mathrm{T}}\boldsymbol{\beta}_0\}=$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_h'(\varepsilon_i)(\boldsymbol{\Gamma}\boldsymbol{Z}_i)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i)\{\boldsymbol{\Gamma}\boldsymbol{Z}_i\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{\beta}_0-$$

$$\boldsymbol{\Gamma}\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})^{-1}\boldsymbol{Z}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\beta}_0\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_h'(\varepsilon_i)(\boldsymbol{\Gamma}\boldsymbol{Z}_i)$$

再考虑式(2)的左边,有

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i^{**})\hat{X}_i\hat{X}_i^{\mathrm{T}} =$$

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i^{**})(\hat{\boldsymbol{X}}_i-\boldsymbol{X}_i+\boldsymbol{X}_i)(\hat{\boldsymbol{X}}_i-\boldsymbol{X}_i+\boldsymbol{X}_i)^{\mathrm{T}} =$$

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i^{**})[(\hat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})\boldsymbol{Z}_i+\boldsymbol{\Gamma}\boldsymbol{Z}_i][(\hat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})\boldsymbol{Z}_i+\boldsymbol{\Gamma}\boldsymbol{Z}_i]^{\mathrm{T}} = \frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i^{**})[(\hat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}(\hat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})^{\mathrm{T}}] +$$

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i^{**})[(\hat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Gamma}] +$$

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i^{**})[\boldsymbol{\Gamma}\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}(\hat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})^{\mathrm{T}}] +$$

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i^{**})(\boldsymbol{\Gamma}\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}) =$$

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i^{**})(\boldsymbol{\Gamma}\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}) + o_p(1) =$$

$$\frac{1}{n}\sum_{i=1}^{n}\{\varphi_h''(\varepsilon_i)+\varphi_h'''(\varepsilon_i)(\varepsilon_i^{**}-\varepsilon_i) + o_p(1)\}(\boldsymbol{\Gamma}\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}) + o_p(1) =$$

$$\frac{1}{n}\sum_{i=1}^{n}\{\varphi_h''(\varepsilon_i)+\varphi_h'''(\varepsilon_i)[(\boldsymbol{\Gamma}-\hat{\boldsymbol{\Gamma}})\boldsymbol{Z}_i+\boldsymbol{e}_i]^{\mathrm{T}}\boldsymbol{\beta}_0)\}(\boldsymbol{\Gamma}\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}) + o_p(1) =$$

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i)(\boldsymbol{\Gamma}\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}) +$$

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_h'''(\varepsilon_i)e_i^{\mathrm{T}}\boldsymbol{\beta}_0(\boldsymbol{\Gamma}\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}) +$$

$$o_p(1) = \frac{1}{n}\sum_{i=1}^{n}\varphi_h''(\varepsilon_i)(\boldsymbol{\Gamma}\boldsymbol{Z}_{ii}^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}) + o_p(1)$$

则可以推出:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \frac{(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}}\boldsymbol{\Gamma}^{\mathrm{T}})^{-1}}{E(\varphi_h''(\varepsilon_i))}\sum_{i=1}^{n}n^{-\frac{1}{2}}\varphi_h'(\varepsilon_i)(\boldsymbol{\Gamma}\boldsymbol{Z}_i) + o_p(1)$$

易知:

$$E\left(\sum_{i=1}^{n}n^{-\frac{1}{2}}\varphi_h'(\varepsilon_i)(\boldsymbol{\Gamma}\boldsymbol{Z}_i)\right)=0$$

且有

$$Var\left(\sum_{i=1}^{n}n^{-\frac{1}{2}}\varphi_h'(\varepsilon_i)(\boldsymbol{\Gamma}\boldsymbol{Z}_i)\right) = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}}\boldsymbol{\Gamma}^{\mathrm{T}}\{E[\varphi_h'(\varepsilon)]^2\}$$

再由中心极限定理,就完成了该定理的证明。

3 数值模拟

本节通过模拟研究评估所提出的 IVESL 估计量的有效性与准确性,作为比较,还计算了朴素 M 估计(nM)、朴素最小二乘估计(nLS)、朴素指数平方损失估计(nESL)、基于工具变量的 M 估计(IVM)、基于工具变量的最小二乘估计(IVLS)等 5 种方法的估计量。上面所指的朴素方法指不使用工具变量 Z_i,直接将 X_i 视为外生变量参与模型的估计方法。模拟数据来自下列模型:

$$\begin{cases}Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i\\ X_{i2} = \gamma Z_i + e_i, i = 1,2,\cdots,n\end{cases}$$

其中,$X_{i1} \sim N(0,1)$,$(\beta_1,\beta_2)^{\mathrm{T}} = (5,2)^{\mathrm{T}}$,$Z_i \sim N(1,1)$,$\gamma = 1$,$e_i \sim N(0,0.4^2)$,$\varepsilon_i = e_i + \sigma_i$。由此可见,$X_{i1}$ 是外生变量,X_{i2} 是内生变量。在本次模拟中,考虑 σ_i 的分布为正态分布、T 分布和柯西分布,样本容量 n=100,150,200,重复运行 1 000 次,比较不同误差分布情形下 6 种估计方法的均值、偏差和标准差,模拟结果见表 1—表 3。从表 1—表 3 可以看出:

(1) 3 种基于工具变量的估计方法优于 3 种朴素估计方法。由此可见,忽略内生变量直接采用 X 所获得的估计量是有偏的。

(2) 当 σ_i 服从正态分布时,3 种基于工具变量的估计方法所得估计量的偏差、标准差相差不大;当 σ_i 服从 T 分布或柯西分布时,IVLS 方法失去了稳健性,造成了过大的偏差和标准差,IVM 和 IVESL 方法依然稳健。大多数情况下,IVESL 方法略优于 IVM 方法,因此,本文提出的 IVESL 估计具有稳健性。

(3)样本容量 n 增大时,IVM 和 IVESL 估计量的偏差、标准差逐渐下降。

进一步,为了研究本文提出模型在高杠杆点存在的情况下是否依然有效,模拟了在 σ_i 服从正态分布的情况下,考虑 15%样本点的值为高杠杆点 X_{i1} = 3 的情况,模拟结果见表 4。从表 4 可以看出:3 种朴素方法以及 IVLS 方法的效果较差,不再适用,而 IVM,IVESL 方法效果较好,估计量仍然稳健,且 IVESL 估计量略优于 IVM 估计量。

因此,本文提出的 IVESL 方法不需要对模型误差分布作特定的假设,无论模型误差的分布是何种形式,都具有较好的性质,并且,IVESL 有效地处理了内生性问题,使得估计量仍然具有无偏性。

表 1　随机误差 $\sigma_i \sim N(0,0.4^2)$ 的数值模拟结果

Table 1　Numerical simulation results for the case of stochastic error $\sigma_i \sim N(0,0.4^2)$

n	方　法	测试 1			测试 2		
		平均值	标准差	偏　差	平均值	标准差	偏　差
100	nM	5.001 9	0.060 9	0.001 9	2.137 6	0.0551	0.137 6
	nLS	5.001 5	0.057 3	0.001 5	2.136 6	0.052 0	0.137 2
	nESL	5.001 6	0.056 8	0.001 6	2.073 1	0.038 0	0.073 1
	IVM	4.996 3	0.135 5	−0.003 6	2.004 4	0.109 1	0.004 4
	IVLS	4.995 3	0.125 2	−0.004 6	1.996 6	0.100 1	−0.003 3
	IVESL	4.999 5	0.132 7	−0.000 4	1.997 7	0.043 5	−0.002 2
150	nM	4.997 7	0.047 2	−0.002 8	2.137 3	0.045 7	0.137 3
	nLS	4.997 5	0.044 5	−0.002 4	2.137 4	0.042 2	0.137 4
	nESL	4.997 7	0.046 0	−0.002 3	2.073 6	0.032 0	0.073 6
	IVM	4.997 2	0.111 3	−0.002 7	1.997 5	0.090 9	−0.002 4
	IVLS	4.997 9	0.104 2	−0.002 2	1.997 2	0.082 5	−0.002 7
	IVESL	4.997 9	0.105 5	−0.002 0	1.999 0	0.035 7	−0.000 9
200	nM	4.998 5	0.039 1	−0.001 4	2.138 8	0.038 3	0.138 8
	nLS	4.998 2	0.037 2	−0.001 7	2.138 1	0.036 6	0.138 1
	nESL	4.998 6	0.038 2	−0.001 3	2.074 6	0.026 7	0.074 6
	IVM	4.997 1	0.096 9	−0.002 8	2.002 2	0.077 0	0.002 2
	IVLS	4.996 9	0.091 3	−0.003 0	2.001 3	0.068 2	0.001 3
	IVESL	4.996 8	0.091 5	−0.003 1	2.000 6	0.031 3	0.000 6

表 2　随机误差 $\sigma_i \sim 0.2T(2)$ 的数值模拟结果

Table 2　Numerical simulation results for the case of stochastic error $\sigma_i \sim 0.2T(2)$

n	方　法	测试 1			测试 2		
		平均值	标准差	偏　差	平均值	标准差	偏　差
100	nM	5.000 3	0.053 3	0.000 3	2.137 8	0.048 3	0.137 8
	nLS	5.002 0	0.115 8	0.002 0	2.136 7	0.070 3	0.136 7
	nESL	5.000 1	0.053 6	0.000 1	2.074 1	0.036 1	0.074 1
	IVM	4.969 4	0.122 5	−0.030 5	2.010 2	0.095 5	0.010 2
	IVLS	4.969 1	0.131 6	−0.030 8	2.002 6	0.145 0	0.002 6
	IVESL	4.957 7	0.116 1	−0.042 2	2.001 4	0.039 1	0.001 4
150	nM	5.000 9	0.043 4	0.000 9	2.138 3	0.039 7	0.138 3
	nLS	4.999 5	0.061 1	−0.000 4	2.138 8	0.059 2	0.138 8
	nESL	5.001 1	0.043 2	0.001 1	2.074 9	0.028 6	0.074 9
	IVM	4.982 0	0.109 0	−0.017 9	2.003 9	0.094 4	0.073 9
	IVLS	4.988 9	0.107 5	−0.011 0	1.996 7	0.081 2	−0.003 2
	IVESL	4.988 5	0.102 1	−0.011 4	1.999 5	0.034 1	−0.000 5
200	nM	4.999 2	0.035 4	−0.000 7	2.138 7	0.034 8	0.138 7
	nLS	4.999 8	0.047 7	−0.000 1	2.140 7	0.047 0	0.140 7
	nESL	4.999 7	0.035 2	−0.000 2	2.075 4	0.025 0	0.075 4
	IVM	5.001 2	0.077 5	0.001 2	1.992 4	0.088 2	−0.007 6
	IVLS	5.003 2	0.083 5	0.003 2	1.992 1	0.081 4	−0.007 8
	IVESL	5.005 2	0.075 9	0.005 2	2.001 3	0.033 7	0.001 3

表 3 随机误差 $\sigma_i \sim 0.2\mathrm{Cauchy}(2)$ 的数值模拟结果

Table 3 Numerical simulation results for the case of stochastic error $\sigma_i \sim 0.2\mathrm{Cauchy}(2)$

n	方 法	测试 1			测试 2		
		平均值	标准差	偏 差	平均值	标准差	偏 差
100	nM	4.998 6	0.063 6	−0.001 3	2.137 8	0.060 4	0.137 8
	nLS	4.806 0	5.086 8	−0.193 9	2.134 7	3.524 1	0.174 7
	nESL	5.000 6	0.063 1	0.000 6	2.073 0	0.044 5	0.073 0
	IVM	5.002 7	0.147 4	0.002 7	1.993 8	0.122 2	−0.006 1
	IVLS	4.470 2	2.353 3	−0.259 7	1.989 8	1.372 5	−0.010 1
	IVESL	4.996 1	0.140 1	−0.003 8	2.000 8	0.073 9	0.000 8
150	nM	4.999 7	0.050 1	−0.000 2	2.136 6	0.047 6	0.136 6
	nLS	4.856 0	17.101 5	−0.143 5	2.408 3	4.755 4	0.408 3
	nESL	4.999 2	0.052 2	0.000 7	2.073 4	0.035 5	0.073 4
	IVM	4.996 9	0.119 2	−0.003 0	1.995 6	0.100 7	−0.004 3
	IVLS	4.984 4	3.375 4	−0.015 5	1.954 1	2.346 6	−0.045 8
	IVESL	4.996 1	0.115 8	−0.003 8	1.998 2	0.051 9	−0.001 7
200	nM	5.000 6	0.043 6	0.000 6	2.137 5	0.040 1	0.137 5
	nLS	5.633 1	21.412 0	0.633 1	2.837 8	29.802 3	0.837 8
	nESL	5.000 4	0.044 7	0.000 4	2.074 7	0.030 1	0.074 7
	IVM	5.002 2	0.103 4	0.002 2	2.008 7	0.084 7	−0.008 7
	IVLS	4.912 0	2.425 2	−0.087 9	1.909 2	3.611 8	−0.090 7
	IVESL	5.001 8	0.101 2	0.001 8	2.001 6	0.044 9	0.001 6

表 4 随机误差 $\sigma_i \sim N(0,0.4^2)$ 且 15%样本点为高杠杆点 $X_{i1} = 3$ 的数值模拟结果

Table 4 Numerical simulation results for the case of $\sigma_i \sim N(0,0.4^2)$ and 15% high-leverage samples with $X_{i1} = 3$

n	方 法	测试 1			测试 2		
		平均值	标准差	偏 差	平均值	标准差	偏 差
100	nM	5.001 4	0.061 7	0.001 4	2.073 0	0.052 8	0.073 0
	nLS	2.005 5	0.384 8	−2.994 4	1.644 2	0.427 7	−0.355 7
	nESL	4.995 9	0.061 3	−0.004 0	2.072 3	0.041 6	0.072 3
	IVM	5.010 8	0.162 5	0.010 8	2.002 9	0.061 5	0.002 9
	IVLS	2.009 8	0.323 3	−2.997 1	1.543 0	0.268 6	−0.456 9
	IVESL	4.990 0	0.158 1	−0.009 9	2.000 8	0.065 5	0.000 8
150	nM	4.998 0	0.048 8	−0.001 9	2.073 3	0.044 5	0.073 3
	nLS	2.012 7	0.298 0	−2.987 2	1.674 8	0.228 7	−0.073 3
	nESL	4.992 5	0.049 1	−0.007 4	2.072 4	0.031 4	0.072 4
	IVM	5.000 3	0.113 9	0.000 3	2.001 0	0.050 4	0.001 0
	IVLS	2.034 0	0.207 1	−2.965 9	1.560 9	0.328 1	−0.439 0
	IVESL	5.000 9	0.114 6	0.000 9	1.999 4	0.051 6	−0.0005
200	nM	4.998 7	0.042 5	−0.001 2	2.079 4	0.029 2	0.074 9
	nLS	1.985 8	0.245 5	−3.041 1	1.665 8	0.178 0	−0.334 1
	nESL	4.993 2	0.044 0	−0.006 7	2.074 1	0.038 2	0.074 1
	IVM	4.999 2	0.091 2	−0.000 7	2.001 1	0.044 0	0.001 1
	IVLS	2.016 0	0.224 2	−2.983 9	1.524 6	0.185 0	−0.453 7
	IVESL	4.984 7	0.093 2	−0.015 2	2.000 0	0.038 1	0.000 0

4　实例分析

本节用提出的方法对“收入-教育程度”数据进行实证分析。该数据来源于 Ashehfelter 和 Krueger[9] 关于同卵双胞胎教育回报率的调查。在这项调查中,包含了 149 对同卵双胞胎的样本。Ashehfelter 和 Krueger 使用均值回归模型调查基因遗传对采访到的双胞胎收入与受教育程度的影响。如果用传统方式来量化受教育程度,则该变量会存在内生性,由此导致估计量产生偏差。因此,工具变量的引入可以较好地解决这个问题,构造下列工具变量线性回归模型:

$$\begin{cases}\log(w_1)-\log(w_2)=\beta_0+\beta_1(E_{2,2}-E_{1,1})+\varepsilon_i\\ E_{2,2}-E_{1,1}=\gamma(E_{2,1}-E_{1,2})+E_i,i=1,2,\cdots,n\end{cases}$$

其中,w_1 是孪生长子的报告收入,w_2 是孪生次子的报告收入,$E_{1,1}$ 是孪生长子报告的所受学校教育年数,$E_{2,2}$ 是孪生次子报告的所受学校教育年数。文献[9]分析该数据时,认为每对双胞胎受教育程度之差,即 $E_{2,2}-E_{1,1}$ 是内生变量,为了消除内生性,采用 $E_{2,1}-E_{1,2}$ 作为双胞胎受教育程度之差的工具变量,其中,$E_{1,2}$ 是孪生长子报告的孪生次子所受学校教育年数,$E_{2,1}$ 是孪生次子报告的孪生长子所受的学校教育年数。图 1 呈现了响应变量 Y 的直方图与密度函数曲线,显然,响应变量在右端有显著的重尾效应,根据 Kolmogorov-Smirnov 检验得到的 P 值远小于 0.000 1,因此,与最小二乘法相比,采用 IVESL 方法分析该数据更加合理。为了对比,利用第 3 节模拟研究的其余 5 种方法也分析了该数据,计算结果见表 5。

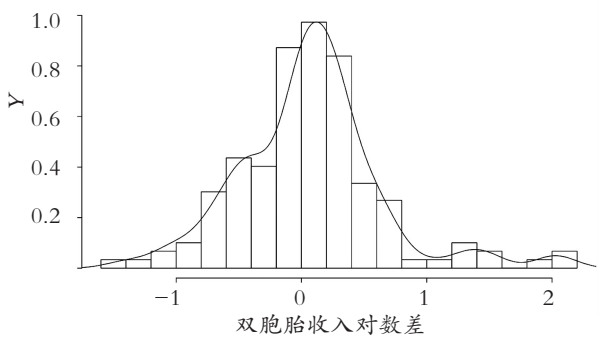

图 1　收入-教育程度数据中响应变量 Y 的柱形图和密度曲线图

Fig. 1　The histogram and density curve of the response variable Y for the wage and schooling data

表 5　收入-教育程度数据拟合结果

Table 5　Fitting results of the wage and schooling data

方　法	参数估计值 $\hat{\beta}_1$	标准差
nM	0.078 5	0.022 9
nLS	0.092 3	0.022 8
nESL	0.089 02	0.022 8
IVM	0.165	0.040 7
IVLS	0.156 6	0.040 3
IVESL	0.160 5	0.039 9

表 5 给出了 nM,nLS,nESL,IVM,IVLS,IVESL 方法下的参数估计值 $\hat{\beta}_1$ 以及其标准差。以上 6 种方法都指出收入与受教育程度呈正相关,并且 IVESL 和 IVLS 方法得到的估计值约为 0.16,这与 Ashehfelter 和 Krueger[9] 所得到的结果类似。由表 5 可知,忽略协变量内生性使用朴素方法得到的估计量在 0.07—0.09 附近,较 Ashehfelter 和 Krueger[9] 所得到的结果小了许多,而基于工具变量方法得到的估计量则避免了上述问题,与 Ashehfelter 和 Krueger[9] 所得结果更为相近。再者,由于响应变量不服从正态分布,故 IVM,IVESL 等稳健估计方法结果更加合理,与 IVM 相比,IVESL 给出了较小的标准差。

参考文献(References):

[1] WANG X Q, JIANG Y L, HUANG M, et al. Robust variable selection with exponential squared loss[J]. Journal of the American Statistical Association, 2013, 108(502): 632—643.

[2] YU P, ZHU Z Y, ZHANG Z Z. Robust exponential squared loss-based estimation in semi-functional linear regression models[J]. Computational Statistics, 2019, 34(2): 503—525.

[3] JIANG Y L. Robust estimation in partially linear regression models[J]. Taylor & Francis, 2015, 42(11): 2497—2508.

[4] ASHENFELTER O C. Estimating the effect of training programs on earnings[J]. The Review of Economics and Statistics, 1978, 60(1): 47—57.

[5] THISTLETHWAITE D L, CAMPBELL D T. Regression-discontinuity analysis: an alternative to the ex post facto experiment[J]. Journal of Educational Psychology, 1960, 51(6): 309—317.

[6] DONALD S G, NEWEY W K. Choosing the number of

instruments[J]. Econometrica, 2001, 69(5):1161—1191.
[7] YANG W M, YANG Y P. Composite quantile regression estimation of linear error-in-variable models using instrumental variables[J]. Metrika,2020,83(1):1—16.
[8] 杨宜平.协变量随机缺失下线性模型的经验似然推断及其应用[J].数理统计与管理,2011,30(4):655—663.
YANG Yi-ping. Empirical likelihood for linear models with covariate data missing at random[J]. Journal of Applied Statistics and Management,2011,30(4):655—663.
[9] ASHENFELTER O, KRUEGER A. Estimates of the economic return to schooling from a new sample of twins [J]. The American Economic Review,1994,84(5):1157—1173.
[10] SONG Y, JIAN L, LIN L. Robust exponential squared loss-based variable selection for high-dimensional single-index varying-coefficient model[J]. Journal of Computational and Applied Mathematics,2016, 308:330—345.

Exponential Squared Loss Estimation of Linear Regression Models Using Instrumental Variables

ZHANG Wei[1], YANG Yi-ping[1,2]

(1. School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China;
2. Chongqing Key Laboratory of Social Economy and Applied Statistics, Chongqing Technology and Business University, Chongqing 400067, China)

Abstract: Endogenous variables are often encountered when discussing the relationship between covariates and response variables. Most of the existing researches on endogenous variables are discussed in the framework of least squares objective function. However, this method is not robust. In this paper, the exponential squared loss estimation method is used to construct the robust estimation of regression coefficient in the model. In order to overcome the bias of endogenous variables on the estimation, instrumental variables are used to eliminate the endogeneity of covariates, and then the exponential squared loss estimation of regression coefficients is constructed. For the exponential squared loss objective function, the estimation process of selecting effective adjustment parameters is proposed. Under some regular conditions, the asymptotic normality of the proposed estimator is studied. In the simulation study, six estimation methods are compared, including the naive least squares estimation, the naive M estimation, the naive exponential squared loss estimation, the least squares estimation using instrumental variables, the M estimation using instrumental variables, and the exponential squared loss estimation using instrumental variables. The simulation results show that the proposed method can effectively eliminate the endogeneity of covariates, and has good robustness. Finally, the wage and schooling data collected from the survey of identical twins is analyzed by our proposed method.

Key words: endogenous variable; instrumental variables; linear model; exponential squared loss

责任编辑:李翠薇

引用本文/Cite this paper:
张巍,杨宜平.工具变量线性回归模型的指数平方损失估计[J].重庆工商大学学报(自然科学版),2022,39(2):99—106.
ZHANG Wei, YANG Yi-ping. Exponential squared loss estimation of linear models using instrumental variables[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(2):99—106.