doi:10. 16055/j. issn. 1672-058X. 2020. 0006. 015

基于半参数测量误差模型的加利福尼亚房价数据研究

娄 文
(南京理工大学 理学院,南京 210094)

摘 要:对于加利福尼亚房价数据,房屋中位年龄可以看作是潜在的混淆变量,有可能影响其他协变量与响应变量之间的关系。如果忽略测量误差对变量的影响,并直接运用假设响应变量和协变量可以准确观测的经典半参数模型来拟合该数据,则可能会导致结果存在较大偏差,因此提出了利用单指标扭曲测量误差模型对该数据进行拟合。观察扭曲函数的拟合曲线后发现:中位房价、中位收入、总房间数、总卧室数和人口确实受到了以房屋中位年龄为混淆变量的乘积污染,这说明了选择的单指标扭曲测量误差模型相比于不考虑测量误差的半参数模型更适合加利福尼亚房价数据。

关键词:单指标模型;乘积扭曲测量误差;加利福尼亚房价数据
中图分类号:O212. 7 **文献标志码**:A **文章编号**:1672-058X(2020)06-0095-08

0 引 言

在经济领域,研究地区房价与各种影响因素之间的关系有着较为重要的意义。何静等[1]利用可加模型对北京市二手房房价数据进行分析并通过分位回归方法对模型进行估计;周尔民等[2]根据江西省 2005—2013 年的房价数据,采用逐步回归分析法,建立多个房价回归模型,并对模型进行诊断和检验;尹雯雯[3]研究了变系数误差模型的核实方法在波士顿住房数据上的应用。研究的加利福尼亚房价数据集包含 20 640 个观测样本,其中响应变量是中位房价(Median house value),协变量包括中位收入(Median income)、房屋中位年龄(Housing median age)、总房间数(Total rooms)、总卧室数(Total bedrooms)、人口(Population)、家庭(Households)、纬度(Latitude)和经度(Longitude)。

单指标模型(Single-index model)是分析经济型数据的有力工具,由 Friedman 和 Stuetzle(1981)[4]研究投影追踪回归问题时首次提出。单指标模型通过线性组合的方式将多维协变量转换为一维指标,并利用未知联系函数来保留协变量和响应变量之间的非线性关系,避免了多元非参数回归中经常出现的“维数祸根”问题。有关单指标模型统计推断和理论性质的研究成果十分丰富。Stoker[5]和 Härdle 等[6]提出了平均导数方法,并讨论了该方法与投影追踪回归之间的关系;Powell 等[7]对平均导数方法进行改进之后提出了密度加权平均导数估计方法,能够减少大量的计算成本;Li[8]提出了切片逆回归方法来估计正确的降维方向,并证明了它的 $\sqrt{n}$ 相合性;Ichimura[9]和 Härdle 等[10]研究了半参数最小二乘估计方法,该方法先假设指标系数已知,然后利用核估计方法来估计联系函数,此时联系函数的估计值可以看作是指标系数的函数,最后通过极小化残差平方和得到指标系数的估计;Xia 等[11]在研究多指标模型中降维空间的估计问题时,提出了最小平均方差估计,其优势在于不需要对联系函数进行欠光滑且能够保证参数的估计以较快的速度收敛;Delecroix 等[12]研究指标系数的半参数极大似然估计时发现其仍具有参数极大似然估计的重

收稿日期:2019-11-20;**修回日期**:2020-03-20.
作者简介:娄文(1994—),男,福建三明人,硕士研究生,从事非参数与半参数统计研究.

要性质,并证明了得到的估计量是渐近有效的;Wang 等[13]提出了估计方程估计法,同时证明了估计量是$\sqrt{n}$相合的;Peng 和 Huang[14]提出了惩罚最小二乘方法来估计单指标模型中的指标系数和联系函数,该方法的特点在于能够在参数估计的同时进行变量选择;Fan 等[15]研究了单指标模型的复合分位数回归问题,提出了基于单指标模型规范的投影方法,该方法能够兼顾模型适用性和统计精度,他们通过将半参数思想与变量选择方法相结合的手段,解决了高维情况下维数与统计精度之间的平衡问题;Dong 等[16]研究了单指标模型和部分线性单指标模型的估计,利用正交级数展开来估计未知可积的联系函数,接着通过剖面方法获得参数的估计量,并发现了单指标模型的估计量具有两倍的收敛速度,部分线性单指标模型的估计量具有三倍的收敛速度;Sun[17]讨论了如何利用单指标模型处理空间依赖型数据,提出了一种两阶段估计方法,先通过局部线性方法得到非参数部分的估计,再利用 GMM 方法估计参数部分,并给出了估计量渐近性质的证明和非参数函数置信带的构造方法。

在经济和医疗等领域,具有乘积扭曲结构的测量误差十分常见。Şentürk 和 Müller[18]在研究协变量和响应变量都含有乘积扭曲测量误差的线性回归模型时,提出了协变量调整回归(covariate - adjusted regression,CAR)。该方法通过建立回归系数与变系数回归模型之间的联系,消除了乘积扭曲测量误差给回归系数估计带来的影响,他们证明了利用协变量调整回归获得的回归系数的估计量具有相合性,并用此方法分析了血液透析患者的纤维蛋白原水平与其他血浆蛋白水平(如转铁蛋白水平、铜蓝蛋白水平和酸性糖蛋白水平等)之间的关系;Delaigle 等[19]进一步讨论了非参数协变量调整回归的相关问题,在弱化了一些有关变量和扭曲函数的假设条件后,给出了更为灵活的非参数估计量,能够在协变量和响应变量期望为 0 或扭曲函数不满足严格大于 0 的条件下对非参数部分进行估计。

本文利用单指标扭曲测量误差模型对加利福尼亚房价数据进行拟合。由于单指标模型可以通过部分线性单指标模型退化得到,因此我们利用 Zhang[20]提出的估计方法来进行模型估计。

1 模型介绍

参数回归模型最大的特点在于假设模型的结构是已知的,即响应变量和协变量之间的函数关系是已知的,仅有有限个参数未知。在这样的假设下,参数回归模型的估计问题就等同于这有限个未知参数的估计问题。因此,诸如线性模型和广义线性模型等参数回归模型的估计方法相对简单。参数回归模型对模型结构的假设除了给模型估计带来了便利,还提高了模型被错误识别的风险。如果模型与实际情况相符,那么做出的统计推断则有着较高的精度。一旦模型与实际情况偏差较大,获得的估计结果会很差。

非参数回归模型没有给出完全已知的模型结构,而是通过未知函数来构建 Y 与 X 之间的关系,所以适用的范围要比参数回归模型广泛。非参数回归模型在协变量的维数是一维的时候,得到的未知函数的估计精度较高,而当协变量的维数超过一维的时候,得到的未知函数的估计精度会随着维数的增大快速下降。这是因为诸如 N-W 核估计法(Nadaraya - Watson)、局部多项式估计法(Local Polynomial)和 B 样条估计法(B-Spline)等非参数估计方法(即光滑方法)的本质是局部光滑,只有确保某一点的领域内有着足够多的数据点,才能得到未知函数在该点较为精确的估计。然而,随着协变量维数的增大,一个局部领域内的样本个数占总的样本个数的比例会越来越小,局部光滑所需要的数据点个数成指数倍增加,这就是所说的“维数祸根”(curse of dimensionality)现象。

半参数模型在保留非参数回归模型优点的同时对协变量进行降维,较好地解决了“维数祸根”问题。该模型能够根据数据来确定模型的最终结构,能够很好地解释协变量与响应变量之间的影响关系,能够减小假设模型与真实模型存在偏离时的影响。经过不断地发展,半参数回归模型的形式也越来越丰富,包括部分线性模型、单指标模型、变系数模型和单指标变系数模型等,这些模型都已经广泛地应用于经济和医疗等领域。

在实际应用中,能够影响变量观测准确度的因素有很多,例如测量仪器自身的准确度不足产生的

误差,使用测量仪器观测时读数产生的误差和获取各个样本的外部环境条件存在差异产生的误差等。如果忽略这些影响因素,默认变量的观测值与其真实值之间不存在偏差,利用半参数回归模型对含有测量误差的变量进行统计推断,那么推断的结果将存在偏差,严重时可能与真实情况完全违背。目前,测量误差影响观测值的方式主要有两类:一类被称为可加结构的测量误差模型,顾名思义就是测量误差以加和的形式影响真实值的观测,如 $W=X+U$(W 是观测值,X 是真实值,U 是测量误差);另一类被称作乘积结构的测量误差模型,即测量误差以乘积的形式影响真实值的观测,如 $W=XU$(W,X,U 的含义同上)。

随着不断深入的研究,测量误差对于观测值的影响方式越来越复杂,简单的乘积结构的测量误差模型无法在某些复杂情况下进行有效的纠偏。因此,乘积结构的测量误差模型有了更为复杂的扩展形式,例如乘积扭曲结构的测量误差模型,$W=X\psi(U)$ (W,X的含义同上,U 是混淆变量,ψ 是未知扭曲函数),乘积单指标扭曲结构的测量误差模型,$W=X\psi(\boldsymbol{\theta}^{\mathrm{T}}U)$ (W,X,U,ψ 的含义同上,θ 为未知的指标系数)。在经济和医疗领域,诸多变量都具有乘积扭曲结构的测量误差。经济领域的房屋年龄和医疗领域的身体质量指数(BMI)等通常被视作混淆变量。

根据加利福尼亚房价数据的特点,房屋中位年龄可能作为混淆变量影响其他变量的观测结果。为了能够让模型尽可能地符合数据的实际情况,选择单指标扭曲测量误差模型对该数据进行拟合。

2　模型估计

单指标扭曲测量误差模型具有如下形式:

$$\begin{cases}\boldsymbol{Y}_i=g(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_i)+\varepsilon_i\\ \widetilde{Y}_i=\psi(U_i)\boldsymbol{Y}_i\\ \widetilde{X}_i=\boldsymbol{\varphi}(U_i)\boldsymbol{X}_i\end{cases},i=1,2,\cdots,n \qquad (1)$$

其中 $\boldsymbol{Y}_i$ 是一维响应变量,$\boldsymbol{X}_i\in\boldsymbol{R}^p$ 是协变量,$\boldsymbol{\beta}_0$ 是 p 维未知指标系数,$g(\cdot)$ 是未知单变量联系函数,ε_i 是模型误差,与 $\boldsymbol{X}_i$ 独立且满足均值为 0,方差为 σ^2,$\widetilde{Y}_i$ 和 $\widetilde{X}_i$ 分别为 Y_i 和 X_i 的观测值,Y_i 和 X_i 无法直接观测到,U_i 是 1 维混淆变量且与 (Y_i,X_i,ε_i) 独立,$\psi(\cdot)$ 是未知的扭曲函数,$\boldsymbol{\varphi}(\cdot)=\mathrm{diag}\{\varphi_1(\cdot),\cdots,\varphi_p(\cdot)\}$ 是未知扭曲函数矩阵。为了保证模型的可识别性,假定:

(1) $\|\boldsymbol{\beta}_0\|=1$ 且 $\boldsymbol{\beta}_0$ 的第一个非 0 分量是正的,其中 $\|\cdot\|$ 表示 Euclid 模。

(2) $E\{\psi(U)\}=1$, $E\{\varphi_r(U)\}=1$, $r=1,2,\cdots,p$。

假定式(1)是为了保证参数 $\boldsymbol{\beta}_0$ 的唯一性。假定式(2)确保了乘积扭曲测量误差问题的可识别性,即从均值的角度来看乘积测量误差对变量无影响。这是一般情况下测量误差问题都需要满足的假定条件,其思想类似于经典的加性测量误差问题 $W=X+u$ 中,假设 $E(u)=0$ 来保证可识别性。

根据 Zhang[20] 提出的估计方法,首先利用条件绝对均值校准方法来消除乘积扭曲测量误差对变量观测的影响。令 $m_{|\widetilde{Y}|}(u)=E[|\widetilde{Y}||U=u]$ 和 $m_{|\widetilde{X}_r|}(u)=E[|\widetilde{X}_r||U=u]$,其中 $r=1,\cdots,p$,对于 $\psi(\cdot)>0,\varphi_r(\cdot)>0$,可以得到:

$$m_{|\widetilde{Y}|}(u)=\psi(u)E(|Y|)=\psi(u)E(|\widetilde{Y}|)$$

$$m_{|\widetilde{X}_r|}(u)=\varphi_r(u)E(|X_r|)=\varphi_r(u)E(|\widetilde{X}_r|)$$

这意味着可以通过 $\psi(u)=\dfrac{m_{|\widetilde{Y}|}(u)}{E(|\widetilde{Y}|)}$ 和 $\varphi_r(u)=\dfrac{m_{|\widetilde{X}_r|}(u)}{E(|\widetilde{X}_r|)}$ 来获得扭曲函数。相比于在 $E(Y)\neq0$ 的假设下通过 $\psi(u)=\dfrac{E(\widetilde{Y}|U=u)}{E(Y)}$ 来获得扭曲函数,将观测变量的条件期望改为观测变量绝对值的条件期望,能够有效地处理 $E(Y)=0$ 的情况。这是因为该方法要保证分母不为 0,只需要保证 $E(|\widetilde{Y}|)\neq0$,而 $E(|\widetilde{Y}|)=0$ 意味着 $P(\widetilde{Y}=0)=1$,显然 $\widetilde{Y}$ 恒等于 0 的情况几乎不可能存在。利用 Nadaraya-Watson 核估计方法就能够得到 $\psi(u)$ 和 $\varphi_r(u)$ 的估计量

$$\hat{\psi}(u)=\frac{1}{n\hat{f}_U(u)\overline{|\widetilde{Y}|}}\sum_{i=1}^{n}K_{h_1}(U_i-u)|\widetilde{Y}_i|$$

$$\hat{\varphi}_r(u)=\frac{1}{n\hat{f}_U(u)\ \overline{|\widetilde{X}_r|}}\sum_{i=1}^{n}K_{h_1}(U_i-u)\,|\widetilde{X}_{ri}|$$

其中:

$$\hat{f}_U(u)=\frac{1}{n}\sum_{i=1}^{n}K_{h_1}(U_i-u)$$

$$\overline{|\widetilde{Y}|}=\frac{1}{n}\sum_{i=1}^{n}|\widetilde{Y}_i|$$

$$\overline{|\widetilde{X}_r|}=\frac{1}{n}\sum_{i=1}^{n}|\widetilde{X}_{ri}|$$

$$K_{h_1}(\cdot)=h_1^{-1}K(\cdot/h_1)\ ,K(\cdot)$$

是核函数,h_1 是带宽。将响应变量和协变量的观测值与其各自对应的扭曲函数估计值相除,获得了校准后的变量:

$$\hat{Y}_i=\frac{\widetilde{Y}_i}{\hat{\psi}(U_i)},\hat{X}_{ri}=\frac{\widetilde{X}_{ri}}{\hat{\varphi}_r(U_i)}$$

$\{(\hat{Y}_i,\hat{X}_i),i=1,2,\cdots,n\}$ 可以看作是变量真实值 Y 和 X_i 的估计。

利用条件绝对均值校准方法来对乘积扭曲测量误差进行纠偏可以看作是在对真实模型进行估计前的数据预处理。根据响应变量和协变量的观测值,采用核光滑来得到扭曲函数的估计量,再通过简单的相除运算得到响应变量和协变量真实值的估计,即校准后的响应变量和协变量。在进行模型估计的时候,使用校准后的响应变量和协变量代替观测到的响应变量和协变量。这样一来,就完成了对乘积扭曲测量误差的纠偏。

下面介绍根据校准后的响应变量和协变量进行最终模型估计的方法。令 $\mathfrak{B}=\{\boldsymbol{\beta}\in\boldsymbol{R}^p:\|\boldsymbol{\beta}\|=1$ 且第一个非0分量为正$\}$,则 $\boldsymbol{\beta}_0$ 为集合 $\mathfrak{B}$ 的内点。因此,只在集合 $\mathfrak{B}$ 中搜索想要的参数 $\boldsymbol{\beta}_0$。当联系函数 $g_0(\cdot)$ 已知时,根据校准后的响应变量和协变量 $\{(\hat{Y}_i,\hat{X}_i),i=1,\cdots,n\}$,可以通过极小化目标函数:

$$Q(\boldsymbol{\beta})\equiv\sum_{i=1}^{n}[\hat{Y}_i-g_0(\boldsymbol{\beta}^{\mathrm{T}}\hat{X}_i)]^2,\boldsymbol{\beta}\in\mathfrak{B}\quad(2)$$

来获得 $\boldsymbol{\beta}_0$ 的最小二乘估计。如果利用 Newton 算法来寻求 $Q(\boldsymbol{\beta})$ 的极小值,需要计算 $Q(\boldsymbol{\beta})$ 在 $\boldsymbol{\beta}$ 点处的导数。由于 $\|\boldsymbol{\beta}\|=1$ 表示 $\boldsymbol{\beta}$ 是单位球球面上的点,所以 $g_0(\boldsymbol{\beta}^{\mathrm{T}}X)$ 在 $\boldsymbol{\beta}$ 点处的导数不存在。在此情况下,可以使用“去一分量”方法,对 $\boldsymbol{\beta}$ 进行再参数化,然后在 Euclid 空间 $\boldsymbol{R}^{p-1}$ 的一个区域上寻找方向 $\boldsymbol{\beta}_0$。不失一般性,假设真实的参数 $\boldsymbol{\beta}_0$ 的第 r 个分量是正的。令 $\boldsymbol{\beta}^{(r)}=(\beta_1,\cdots,\beta_{r-1},\beta_{r+1},\cdots,\beta_p)^{\mathrm{T}}$ 为 $\boldsymbol{\beta}$ 去掉第 r 个分量 β_r 之后得到的 $p-1$ 维参数向量。这样得到的新参数 $\boldsymbol{\beta}_0^{(r)}$ 必然满足 $\|\boldsymbol{\beta}_0^{(r)}\|<1$,即 $\boldsymbol{\beta}_0^{(r)}$ 移动到了单位球的内部。从而有 $\boldsymbol{\beta}$ 在 $\boldsymbol{\beta}_0^{(r)}$ 的领域内无限可微。记 $\beta_r=(1-\|\boldsymbol{\beta}^{(r)}\|^2)^{\frac{1}{2}}$,那么有 $\boldsymbol{\beta}=\boldsymbol{\beta}(\boldsymbol{\beta}^{(r)})=(\beta_1,\cdots,\beta_{r-1},(1-\|\boldsymbol{\beta}^{(r)}\|^2)^{\frac{1}{2}},\beta_{r+1},\cdots,\beta_p)^{\mathrm{T}}$。

通过简单的计算,可以得到 $\boldsymbol{\beta}$ 关于 $\boldsymbol{\beta}^{(r)}$ 的 Jacobian 矩阵:

$$\boldsymbol{J}_{\boldsymbol{\beta}^{(r)}}=\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\beta}^{(r)}}=(\boldsymbol{\gamma}_1,\cdots,\boldsymbol{\gamma}_p)^{\mathrm{T}}$$

其中 $\boldsymbol{\gamma}_r=-(1-\|\boldsymbol{\beta}^{(r)}\|^2)^{-\frac{1}{2}}\boldsymbol{\beta}^{(r)}$,$\boldsymbol{\gamma}_s(1\leqslant s\leqslant p,s\neq r)$ 是第 s 个分量为1的 $p-1$ 维单位向量。此时,再考虑式(2)中极小化目标函数 $Q(\boldsymbol{\beta})$ 的问题。可以发现,原本的极小化问题等价于求解估计方程组:

$$\begin{cases}\sum_{i=1}^{n}[\hat{Y}_i-g_0(\boldsymbol{\beta}^{\mathrm{T}}\hat{X}_i)]g'_0(\boldsymbol{\beta}^{\mathrm{T}}\hat{X}_i)\boldsymbol{J}_{\boldsymbol{\beta}^{(r)}}^{\mathrm{T}}\hat{X}_i=0\\ \|\boldsymbol{\beta}\|-1=0\end{cases}$$

其中 $g'_0(\cdot)$ 为函数 $g_0(\cdot)$ 的导数。但是由于 $g_0(\cdot)$ 和 $g'_0(\cdot)$ 是未知的,所以需要分别使用两个估计量来替换它们。运用局部线性光滑方法来估计函数 $g_0(\cdot)$ 和 $g'_0(\cdot)$。对 t 的一个小的领域内的点 T,可以通过线性函数局部地逼近 $g_0(T)$,即

$$g(T)\approx g(t)+g'(t)(T-t)\equiv a+b(T-t)$$

在 $\boldsymbol{\beta}_0$ 固定的情况下,定义 $g(t)$ 和 $g'(t)$ 的局部线性估计量 $\hat{g}(t;\boldsymbol{\beta}_0)=\hat{a}$ 和 $\hat{g}'(t;\boldsymbol{\beta}_0)=\hat{b}$,其中 $\hat{a}$ 和 $\hat{b}$ 是极小化加权平方和:

$$\sum_{i=1}^{n}\{\hat{Y}_i-[a+b(\boldsymbol{\beta}_0^{\mathrm{T}}\hat{X}_i-t)]\}^2K_h(\boldsymbol{\beta}_0^{\mathrm{T}}\hat{X}_i-t)$$

得到的 a 和 b 的估计量,h 是带宽。

根据最小二乘理论,可以得到:

$$\hat{g}(t;\boldsymbol{\beta}_0)=\sum_{i=1}^{n}\frac{W_{ni}(t,\boldsymbol{\beta}_0)}{\sum_{j=1}^{n}W_{nj}(t,\boldsymbol{\beta}_0)}\hat{Y}_i$$

$$\hat{g}'(t;\boldsymbol{\beta}_0)=\sum_{i=1}^{n}\frac{\widetilde{W}_{ni}(t,\boldsymbol{\beta}_0)}{\sum_{j=1}^{n}W_{nj}(t,\boldsymbol{\beta}_0)}\hat{Y}_i$$

其中:

$$W_{ni}(t,\boldsymbol{\beta}_0)=[S_{n,2}(t;\boldsymbol{\beta}_0,h)-(\boldsymbol{\beta}_0^{\mathrm{T}}\hat{X}_i-t)S_{n,1}(t;\boldsymbol{\beta}_0,h)]K_h(\boldsymbol{\beta}_0^{\mathrm{T}}\hat{X}_i-t)\ \widetilde{W}_{ni}(t,\boldsymbol{\beta}_0)=[(\boldsymbol{\beta}_0^{\mathrm{T}}\hat{X}_i-t)S_{n,0}(t;\boldsymbol{\beta}_0,h_2)-$$

$$S_{n,1}(t;\boldsymbol{\beta}_0,h_2)]K_h(\boldsymbol{\beta}_0^{\mathrm{T}}\hat{X}_i-t)S_{n,l}(t;\boldsymbol{\beta}_0,h)=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\beta}_0^{\mathrm{T}}\hat{X}_i-t)^l K_h(\boldsymbol{\beta}_0^{\mathrm{T}}\hat{X}_i-t),l=0,1,2$$

通过求解方程组:

$$\begin{cases}\frac{1}{n}\sum_{i=1}^{n}[\hat{Y}_i-\hat{g}(\boldsymbol{\beta}^{\mathrm{T}}\hat{X}_i)]\hat{g}'(\boldsymbol{\beta}^{\mathrm{T}}\hat{X}_i)J_{\boldsymbol{\beta}^{(r)}}^{\mathrm{T}}\hat{X}_i=0\\ \|\boldsymbol{\beta}\|-1=0\end{cases}$$

可以得到参数 $\boldsymbol{\beta}_0$ 的估计量 $\hat{\boldsymbol{\beta}}$,从而,获得 $g(t)$ 的最终估计量 $\hat{g}^*(t)\equiv\hat{g}(t;\hat{\boldsymbol{\beta}})$。详细的过程可参考 Zhang(2019)[20]。

3　模型拟合

本文研究的加利福尼亚房价数据可在 http://lib.stat.cmu.edu/datasets/houses.zip 获得。运用单指标扭曲测量误差模型对该数据进行拟合,选取其中的中位房价(Median house value)、中位收入(Median income)、房屋中位年龄(Housing median age)、总房间数(Total rooms)、总卧室数(Total bedrooms)和人口(Population)这 6 个变量进行研究。各变量与其对应的符号表示如表 1 所示。

表 1　房价数据变量

Table 1　The variables of housing prices data

变量名称	变量意义
Y	中位房价
X_1	中位收入
X_2	总房间数
X_3	总卧室数
X_4	人口
U	房屋中位年龄

首先对表 1 的 6 个变量进行标准化处理,然后选取模型估计所需要的 3 个带宽 h,h_1 和 h_2。带宽 h_1 用于对扭曲函数进行估计,h 和 h_2 用于对未知函数 $g(\cdot)$ 和 $g'(\cdot)$ 进行局部线性估计。

依照 Silverman(1986)[21] 提出的拇指规则(Rule of thumb),选取 $h_1=n^{-1/3}\mathrm{SE}(U)$,其中 $\mathrm{SE}(U)=\left[\frac{1}{n}\sum_{i=1}^{n}(U_i-\bar{U})^2\right]^{1/2}$,$\bar{U}=\frac{1}{n}\sum_{i}^{n}U_i$。然后,使用 Cai 等[22]提出的多折叠交叉验证(MCV)准则来选择对 $g_0(\cdot)$ 进行局部线性估计的最优带宽。该方法计算相对简单并且能够较快得到最优带宽。设 m 和 L 是两个给定的正整数且满足 $n>mL$。首先使用长度为 $n-lm(l=1,2,\cdots,L)$ 的 L 个子序列估计未知函数向量,然后基于估计的模型计算下一节长度为 m 的一步预测误差。更确切地说,通过极小化平均均方(average mean squared,AMS)误差:

$$\mathrm{AMS}(h)=\sum_{l=1}^{L}\mathrm{AMS}_l(h)\tag{3}$$

来选择 h,其中

$$\mathrm{AMS}_l(h)=\frac{1}{m}\sum_{i=n-lm+1}^{n-lm+m}[\hat{Y}_i-\hat{g}_l^*(\hat{\boldsymbol{\beta}}_l^{\mathrm{T}}\hat{X}_i)]^2$$

$$l=1,\cdots,L$$

$\hat{g}_l^*(\cdot)$ 是在带宽为 $h[n/(n-lm)]^{1/5}$ 时得到的 $g_0(\cdot)$ 第 j 个分量的估计量,$\hat{\beta}_l$ 是 $\boldsymbol{\beta}_0$ 的估计量,它们都是利用校准后的样本 $\{(\hat{Y}_i,\hat{X}_i),1\leqslant i\leqslant n-lm\}$ 计算得到的。设 h_{opt} 是极小化式(3)得到的带宽,那么 h_{opt} 是估计 $g_0(\cdot)$ 的最优带宽。当计算估计量 $\hat{\boldsymbol{\beta}}$ 时,使用近似带宽

$$\hat{h}=h_{\mathrm{opt}}n^{-\frac{1}{20}}(\log n)^{-\frac{1}{2}},\hat{h}_2=h_{\mathrm{opt}}$$

因为这确保了满足最优渐近性质所需要的带宽有着正确的数量级,选取的结果为

$$h_1=0.305,h=0.145,h_2=0.430$$

根据上一节介绍的模型估计方法,给出估计模型式(1)中的未知指标系数 $\boldsymbol{\beta}_0$ 和未知联系函数 $g_0(\cdot)$ 的具体步骤:

第 1 步:利用条件绝对均值校准获取校准后的变量 $\{(\hat{Y}_i,\hat{X}_i),i=1,2,\cdots,n\}$。

第 2 步:设置一个 $\boldsymbol{\beta}$ 的初始值 $\boldsymbol{\beta}_{\mathrm{initial}}$ 且满足 $\|\boldsymbol{\beta}_{\mathrm{initial}}\|=1$。

第 3 步:对给定的 $\boldsymbol{\beta}$,通过局部线性光滑获得 $g_0(t)$ 和 $g'_0(t)$ 的估计量 $\hat{g}(t;\boldsymbol{\beta})$ 和 $\hat{g}'(t;\boldsymbol{\beta})$。

第 4 步:根据估计量 $\hat{g}(t;\boldsymbol{\beta})$,$\hat{g}'(t;\boldsymbol{\beta})$ 和求解估计方程组来搜寻 $\boldsymbol{\beta}$。

第 5 步:重复第 3 步和第 4 步直至收敛,得到估计值 $\hat{\boldsymbol{\beta}}$。

第 6 步:固定 $\boldsymbol{\beta}$ 为 $\hat{\boldsymbol{\beta}}$,重复第 3 步得到 $g(t)$ 的最

终估计$\hat{g}^*(t)=\hat{g}(t;\hat{\beta})$。

该算法的思路十分清晰:先利用条件绝对均值校准对受到乘积扭曲测量误差污染的变量进行纠偏,得到校准后的变量$\{(\hat{Y}_i,\hat{X}_i),i=1,\cdots,n\}$,然后通过局部线性光滑方法获得$g_0(t)$和$g'_0(t)$的估计量,接着求解估计方程组获得$\beta_0$的估计量$\hat{\beta}$,最后将$\boldsymbol{\beta}$固定为$\hat{\beta}$后再次对$g_0(t)$进行局部估计得到$\hat{g}^*(t)$。

这里有一点需要注意,那就是非线性优化的收敛速度对初始值较为敏感。在某些情况下,广义线性模型能够帮助我们获得$\boldsymbol{\beta}_0$的初始值。但是当联系函数为指数函数或者三角函数的时候,就不能再通过广义线性模型得到初始值。此时,可以采用切片逆回归方法或者最小平均方差方法来获得$\boldsymbol{\beta}_0$的初始值。

根据加利福尼亚房价数据,依照上述算法,计算单指标扭曲测量误差模型的估计结果,最终得到的扭曲函数$\psi(\cdot)$,$\varphi_1(\cdot)$,$\varphi_2(\cdot)$,$\varphi_3(\cdot)$和$\varphi_4(\cdot)$的估计结果如图1—图5。如果中位房价、中位收入、总房间数、总卧室数和人口不受到以房屋中位年龄为混淆变量的乘积污染,那么扭曲函数的估计曲线应该近似与直线$Y=1$平行且在该直线的附近。

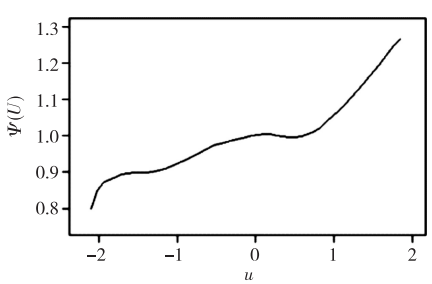

图1 $\psi(\cdot)$的估计

Fig. 1 The estimation of $\psi(\cdot)$

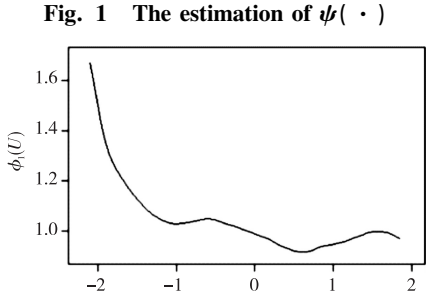

图2 $\varphi_1(\cdot)$的估计

Fig. 2 The estimation of $\varphi_1(\cdot)$

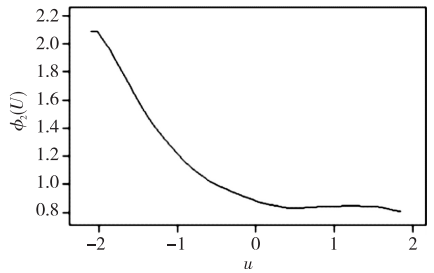

图3 $\varphi_2(\cdot)$的估计

Fig. 3 The estimation of $\varphi_2(\cdot)$

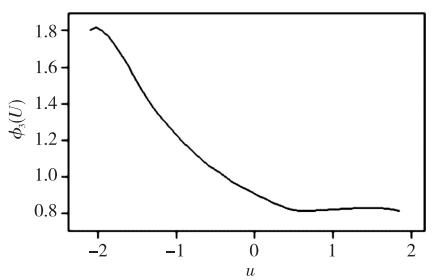

图4 $\varphi_3(\cdot)$的估计

Fig. 4 The estimation of $\varphi_3(\cdot)$

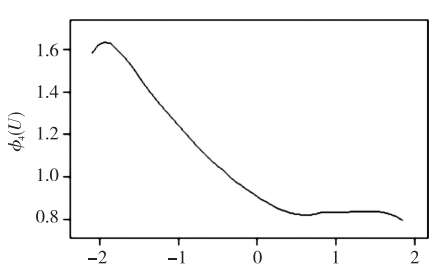

图5 $\varphi_4(\cdot)$的估计

Fig. 5 The estimation of $\varphi_4(\cdot)$

观察图1—图5发现5个扭曲函数的估计曲线既不平行于直线$Y=1$,也不在该直线的附近。这验证了一开始的想法,在一定程度上说明了中位房价、中位收入、总房间数、总卧室数和人口受到了以房屋中位年龄为混淆变量的乘积污染。

为了比较考虑扭曲测量误差和未考虑测量误差这两种方法得到的估计结果。分别使用者两种方法来估计模型的参数和联系函数。参数$\boldsymbol{\beta}_0$的估计结果如表2所示。两种方法都说明总的卧室数($\boldsymbol{X}_3$)对指标$\boldsymbol{\beta}_0^{\mathrm{T}}X$的影响最大,人口($X_4$)对指标$\boldsymbol{\beta}_0^{\mathrm{T}}X$的影响最小。不同在于考虑扭曲测量误差时中位收入($X_1$)对指标$\boldsymbol{\beta}_0^{\mathrm{T}}X$的影响比未考虑测量误差时的小。联系函数$g_0(\cdot)$的估计结果如图6所示,其中实线表示考虑扭曲测量误差时得到的$g_0(\cdot)$估计曲线,虚线表示未考虑测量误差时得到的$g_0(\cdot)$估计曲线。总的来看,未考虑测量误差的估计曲线

大部分位于考虑扭曲测量误差的估计曲线的下方。在刚开始很长的一段区间里 $g_0(\cdot)$ 的总体变化趋势是随着指标 $\boldsymbol{\beta}_0^{\mathrm{T}}X$ 的增长而增长,最后一小段区间里出现小幅度的减少后继续增长。

表 2　两种方法得到参数 $\boldsymbol{\beta}_0$ 的估计

Table 2　The estimation of $\boldsymbol{\beta}_0$ by two methods

方法	考虑扭曲测量误差	未考虑测量误差
$\hat{\beta}_1$	0.316 9	0.402 4
$\hat{\beta}_2$	0.507 0	0.489 3
$\hat{\beta}_3$	0.760 5	0.734 0
$\hat{\beta}_4$	0.253 5	0.244 6

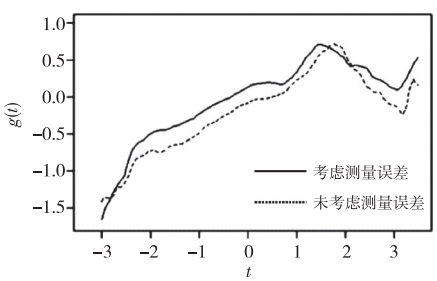

图 6　两种方法得到 $g_0(\cdot)$ 的估计

Fig. 6　The estimation of $g_0(\cdot)$ by two methods

4　结　论

经典半参数模型中大多数假设响应变量和协变量是可以准确观测的,这样能够简化模型。但是在实际应用中,数据存在测量误差的情况时有发生,尤其是在经济领域。忽略测量误差的影响,直接对模型进行估计会导致获得的结果存在偏差。针对加利福尼亚住房数据,选取房屋中位年龄作为混淆变量,采用单指标扭曲测量误差模型对该数据进行拟合。观察扭曲函数的拟合曲线后发现中位房价、中位收入、总房间数、总卧室数和人口均受到了以房屋中位年龄为混淆变量的乘积污染。这说明了所选择的单指标扭曲测量误差模型相比于不含测量误差的半参数模型更适合加利福尼亚住房数据。

参考文献(References):

[1] 何静, 熊巍, 田茂再. 可加模型的无交叉分位回归曲线与房价问题研究[J]. 数理统计与管理, 2015(4): 707—718

HE J, XIONG W, TIAN M Z. Non-Crossing Additive Quantile Curves and Its Applications to Housing Price [J]. Journal of Applied Statistics and Management, 2015(4): 707—718(in Chinese)

[2] 周尔民, 朱进, 王贵用. 房价影响因素模型的构建与实证分析——以江西省为例[J]. 兰州商学院学报, 2016(4):34—43

ZHOU E M, ZHU J, WANG G Y. Construction and Analysis on Influencing Factors Model of Housing Price: A Case Study of Jiangxi Province [J]. Journal of Lanzhou University of Finance and Economics, 2016(4):34—43 (in Chinese)

[3] 尹雯雯. 波士顿住房数据变系数误差模型的核实方法研究[J]. 重庆工商大学学报(自然科学版), 2018, 35(3): 26—29

YIN W W. Analysis on Varying Coefficient Measurement Errors Model with Boston Housing Data by Validation Method[J]. Journal of Chongqing Technology and Business University (Natural Scienle Edition), 2018, 35(3):26—29(in Chinese)

[4] FRIEDMAN J H, STUETZLE W. Projection Pursuit Regression [J]. Journal of the American Statistical Association, 1981, 76(376):817—823

[5] STOKER T M. Consistent Estimation of Scaled Coefficients [J]. Econometrica, 1986, 54(6): 1461—1481

[6] HÄRDLE W, STOKER T M. Smooth Multiple Regression by the Method of Average Derivatives [J]. Journal of the American Statistical Association, 1989, 84(408): 986—995

[7] POWELL J L, STOCK J H, STOKER T M, et al. Semiparametric Estimation of Index Coefficients [J]. Econometrica, 1989, 57(6): 1403—1430

[8] LI K C. Sliced Inverse Regression for Dimension Reduction [J]. Journal of the American Statistical Association, 1991, 86(414): 316—327

[9] ICHIMURA H. Semiparametric Least Squares (sls) and Weighted SLS Estimation of Single- Index Models [J]. Journal of Econometrics, 1993, 58(1-2):71—120

[10] HÄRDLE W, HALL P, ICHIMURA H, et al. Optimal Smoothing in Single-index Models [J]. Annals of Statistics, 1993, 21(1):157—178

[11] XIA Y, TONG H, LI W K, et al. An Adaptive Estimation of Dimension Reduction Space [J]. Journal of The Royal Statistical Society Series B-Statistical

Methodology,2002, 64(3):363—410
[12] DELECROIX M, HÄRDLE W, HRISTACH E. Efficient Estimation in Conditional Single-index Regression [J]. Journal of Multivariate Analysis,2003,86(2):213—226
[13] WANG J, XUE L, ZHU L,et al. Estimation for a Partial-Linear Single-Index Model[J]. Annals of Statistics,2010, 38(1):246—274
[14] PENG H, HUANG T. Penalized Least Squares for Single Index Models [J]. Journal of Statistical Planning and Inference, 2011, 141(4): 1362—1379
[15] FAN Y, HÄRDLE W K, WANG W, et al. Composite Quantile Regression for the Single-index Model [J]. SFB 649 Discussion Papers,(2013):10
[16] DONG C,GAO J,TJOSTHEIM D. Estimation for Single-index and Partially Linear Single-index Integrated Models[J]. The Annals of Statistics,2016,44(1):425—453
[17] SUN Y. Estimation of Single-index Model with Spatial Interaction [J]. Regional Science and Urban Economics, 2017, 62: 36—45
[18] ŞENTÜRK D, MULLER H G. Covariate-Adjusted Regression [J]. Biometrika, 2005, 92(1): 75—89
[19] DELAIGLE A, HALL P, ZHOU W X. Nonparametric Covariate-adjusted Regression [J]. The Annals of Statistics, 2016, 44(5): 2190—2220
[20] ZHANG J. Estimation and Variable Selection for Partial Linear Single-index Distortion Measurement Errors Models[J]. Statistical Papers,2019(3):1—27
[21] SILVERMAN B W. Density Estimation for Statistics and Data Analysis [M]. London: Chapman and Hall,1986
[22] CAI Z, FAN J, YAO Q. Functional-Coefficient Regression Models for Nonlinear Time Series[J]. Journal of the American Statistical Association, 2000, 95(451): 941—956

Study on California Housing Prices Data Based on Semi-parameter Measurement Errors Models

LOU Wen

(School of Science, Nanjing University of Science and Technology, Nanjing 210094,China)

Abstract:For California housing prices data, housing median age as a potential confounding variable may affect the relationship between other covariates and response variable. If we ignore the effect of measurement errors on the variables and directly apply the classic semi-parametric models assuming that the response variable and covariates can be accurately observed to fit the data, it may cause deviations in the results obtained. Therefore, we use single-index distortion measurement errors model to fit these data. After observing the fitted curve of the distortion function, we find that the confounding variable housing median age has a connection with the median house value, median income, total rooms, total bedrooms and population. This shows that the single-index distortion measurement errors model we choose is more suitable for California housing data than the semi-parametric models that do not consider measurement errors.

Key words:single-index models; multiplicative distortion measurement errors; California housing prices data

责任编辑:李翠薇

引用本文/Cite this paper:

娄文.基于半参数测量误差模型的加利福尼亚房价数据研究[J].重庆工商大学学报(自然科学版),2020,37(6):95—102

LOU W. Study on California Housing Prices Data Based on Semi-parameter Measurement Errors Models[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2020,37(6):95—102