doi:10.19920/j.cnki.jmsc.2025.09.003

偏线性分位数选择模型的估计与应用①

周亚虹[1,2,3],萧苹玥[1],金泽群[1*],忻 恺[4],姜帅帅[1]

(1. 上海财经大学经济学院,上海 200433;2. 上海财经大学数理经济学教育部重点实验室,上海 200433;3. 上海数学与交叉学科研究院,上海 200433;4. 上海大学经济学院,上海 200436)

摘要:本研究在分位数回归框架下探讨了偏线性分位数模型中存在样本选择时的模型识别与估计问题.为解决分位数回归中因样本选择问题引发的选择偏误,研究通过对结果方程和选择方程中不可观测扰动项的联合分布进行建模(如采用 Copula 函数)实现了有效修正.文中给出了模型识别的假设,结合分位数回归广义矩估计方法给出了分位数参数(函数)以及 Copula 参数的估计步骤,并且给出了估计量的一致性和渐近正态性.统计推断方面,研究采用非参数自助法估计标准误,并构建检验统计量来验证非线性设定的合理性.此外,基于控制函数方法,本研究建立了内生线性分位数选择模型与本研究模型之间的联系.通过蒙特卡洛模拟发现估计量在有限样本情形下表现良好.最后,本研究将估计方程和检验方法应用于分析女性教育回报率的实际问题,展现了本模型的实用价值.

关键词:分位数回归;样本选择;偏线性;Copula 函数

中图分类号:F064.1;F244 **文献标识码**:A **文章编号**:1007-9807(2025)09-0035-17

0 引 言

在经济数据中,由于个体存在自选择行为,研究者在搜集数据时经常会面临样本非随机的问题.以工资方程为例,只有当个体进入劳动力市场参与工作时,才能观测其工资数据.同时,个体是否参加工作是一个非随机的自选择过程,导致研究样本自身是非随机的.研究者如果直接使用可观测的数据会忽视样本选择问题,可能导致得到的实证结果存在偏误.目前,已经有很多文献研究如何在不同模型设定中修正样本选择问题所产生的偏误[1-6].经过几十年的研究,样本选择模型已成为处理非随机样本问题的有效方法,并在经济学及其他社会科学领域的研究中得到广泛应用[7-13].

然而,上述样本选择模型的文献均聚焦于均值回归模型的估计,针对分位数回归模型估计的文献却相对缺乏.分位数回归最早由 Koenker 和 Bassett[14] 提出,目前已成为分析解释变量对被解释变量异质性影响的重要方法[15-17].与均值回归框架仅能揭示平均因果效应不同,分位数回归可以更精确地刻画个体异质性(individual heterogeneity).以工资方程为例,研究者在均值回归框架下仅能识别教育对工资的平均回报率;然而分位数回归可以帮助识别教育回报率的异质性影响,即不同能力水平下教育回报率的差异.不仅如此,在金融学以及管理学中的市场营销策略分析等研究中,对研究对象所处分布位置的异质性因果分析也是目前研究的重点.例如在金融研究中,分位数回归可以用于描述股票市场风险在不同分布位置上的异质性影响,更好地评估和管理金融风

① 收稿日期:2022-11-15;修订日期:2023-06-26.
基金项目:国家自然科学基金资助项目(71833004;72333002;72342034;72173083;72103126);中央高校基本科研业务费资助专项基金(2023110077).
通讯作者:金泽群(1991—),男,上海人,博士,副教授. Email:polaris@163.sufe.edu.cn

险[18-20];在市场营销研究中,依据消费者所处分布位置,可以制定不同的营销策略,从而更有效地完成消费者目标定位,为精准营销提供有价值的指导[21]. 因此,本研究的模型可以用来刻画这一类异质性影响,对相关经济管理学实证分析起到一定的推动作用. 在已有文献中,分位数回归框架下研究样本选择模型的较少. 目前主流的研究范式由 Arellano 和 Bonhomme[22] 提出,其在一般性非参数模型设定下修正了选择偏差,并给出了线性设定情形下参数的估计方法.

本研究目标方程为偏线性设定时候分位数选择模型的估计问题. Arellano 和 Bonhomme[22] 在估计中采用线性分位数模型并给出一致估计量. 但是,如 Robinson[23] 所述, Arellano 和 Bonhomme[22] 中的线性设定往往面临严重的模型误设风险,从而可能导致估计结果不一致. 例如,在工资方程的研究中,研究者们普遍认为年龄对于工资的影响是非线性的,因而选择将年龄的二次项加入模型进行回归. 这种处理方式可以在一定程度上缓解模型误设风险,但仍然无法准确捕捉年龄的非线性影响. 与之相比,本研究提出的偏线性模型可以有效解决这一问题. 在实际中,本研究放松了年龄在函数结构上的依赖关系,允许其以任意形式来影响个人收入,从而更加精确地描述了年龄在回归方程中的非线性影响. 本研究实证部分得出结论,即年龄确实存在显著的非线性影响,因此直接采用 Arellano 和 Bonhomme[22] 的方法可能会导致估计结果不一致. 此外,国内外不少文献表明,变量的非线性影响在研究中十分常见[24-27]. 例如,黄潇和黄守军[28] 研究发现自雇对收入存在非线性影响;赵涛等[29] 基于门槛模型发现数字经济的高质量发展溢出效应呈现出“边际效应”递增的非线性变化趋势;Wolak[30] 研究了美国加利福尼亚州实行电价节点定价对电力需求的影响,并认为天然气的日均价格对电力需求存在非线性影响;Charnigo 等[31] 采用偏线性模型研究了帕金森患者病情评定影响因素,发现患者年龄对病情评定的影响是非线性的. 相反,完全非线性设定得到的估计量较为稳健,但估计结果不够精确,同时会遭受维数诅咒(curse of dimensionality)问题. 因此,本研究采用偏线性的模型设定,一方面降低了非参数估计的维数,减缓维数诅咒带来的问题;另一方面,在一定程度上降低了 Arellano 和 Bonhomme[22] 方法中可能存在的模型误设风险,提高了估计量的稳健性. 本研究的模型也可以用于分析数字化市场营销等潜在的相关问题. 例如,在分析网络营销对于乡村旅游经营效益的影响时,本研究的模型不仅可以分析网络营销的异质性影响,同时可以研究乡村旅游经营者个体、组织和区域特征的非线性影响,为决策者在制定策略时提供更有价值的指导.

在估计步骤中,本研究采用 Arellano 和 Bonhomme[22] 中提出的方法,通过对结果方程以及选择方程中扰动项的联合分布函数建模(例如 Copula 函数)来修正选择偏误,并结合核函数估计以及广义矩估计方法来估计模型中的参数. 由于本研究的估计步骤为多步估计,因此最终估计量的渐近协方差形式较为复杂,采用代入法估计会变得十分繁琐(需要额外增加多项条件期望的非参数估计,引入更多的窗宽参数). 因此,本研究将通过非参数自助法(nonparametric bootstrap)来估计渐近协方差并用于统计推断,避免了直接估计所带来额外的计算负担. 此外,本研究还给出了非线性设定的检验统计量,以此来说明在实际问题中采用本研究偏线性模型设定的合理性. 本研究还建立了内生情形下线性分位数选择模型与偏线性分位数选择模型之间的联系. 在线性设定下,进一步讨论了利用控制函数方法解决内生性,并且将带有内生性的线性模型转换成外生的偏线性模型. 这样可以采用本研究给出的估计步骤来进行估计.

1 模型和估计

1.1 模型设定

本研究主要研究如下分位数选择模型

$$Q_{Y^*}(\tau \mid X) = X_1' \beta_0(\tau) + g_0(X_2;\tau) \quad (1)$$

$$D = 1\{p_0(Z) \geqslant V\} \quad (2)$$

其中 Y^* 为潜在(*Latent*)因变量,$X = (X_1, X_2)$,D 以及 Z 是可观测变量,V 为不可观测扰动项,$1\{\cdot\}$ 表示取值为 $\{0,1\}$ 的指示函数(indicator function). 令 d_1, d_2 与 d_Z 分别表示 X_1, X_2 与 Z 的维数.

$$Y = Y^* \text{ 若 } D = 1 \tag{3}$$

因此本研究可以观测到的变量集合为(Y,D,Z,X)，并且当 $D=1$ 时本研究才能观测到潜变量 Y^*. 式(1) 是结果方程，这里设定为分位数回归模型，其中$Q_{Y^*}(\tau \mid X)$ 表示给定 X，潜在因变量Y^* 的 τ 分位数，g_0 是未知函数②. 式(1) 属于偏线性分位数回归模型，其中引入了未知函数g_0，允许变量X_2 之间任意相关，具有非线性效应(交互项，高次项等)，从而降低了模型误设的风险. 式(2) 为选择方程，其中$p_0(\cdot)$ 是未知函数. 因此选择方程是非参数设定. 式(3) 表示选择准则，仅当 $D=1$ 时，本研究才能观测到Y.

1.2　模型识别

不失一般性，本研究假设选择方程中的不可观测扰动项 V 服从标准均匀分布，即

$$V \sim \text{Uniform}(0,1) \tag{4}$$

实际上式(4) 并未对模型施加任何约束. 由于p_0 是任意未知函数，本研究可以对选择方程进行如下等价变换

$$\begin{aligned} D &= 1\{p_0(Z) \geqslant V\} \\ &= 1\{F_V(p_0(Z)) \geqslant F_V(V)\} \\ &= 1\{\tilde{p}(Z) \geqslant \hat{V}\} \end{aligned} \tag{5}$$

其中变换得到的新变量$\hat{V} = F_V(V)$ 服从标准均匀分布，且$\tilde{p}(Z) = F_V(p_0(Z))$ 仍是任意未知函数. 因此，若初始选择方程中的不可观测扰动项 V 不满足式(4)，本研究仅需通过式(5) 进行等价变换，并将$\hat{V}$ 替换原来的 V 即可.

为了更直观地描述假设和识别过程，本研究将式(1) 等价地改写成

$$Y^* = X_1' \beta_0(\varepsilon) + g_0(X_2;\varepsilon) \tag{6}$$

其中 ε 为不可观测扰动项，且满足③

$$\varepsilon \sim \text{Uniform}(0,1) \tag{7}$$

此外，令$q_0(X;\varepsilon) = X_1' \beta_0(\varepsilon) + g_0(X_2;\varepsilon)$，其中$q_0(\cdot;\tau)$ 关于τ 严格递增. 本研究假设如下独立性假设成立

$$(X,Z) \perp (\varepsilon,V) \tag{8}$$

基于式(2)、式(3) 以及式(6)、式(8)，本研究可得

$$\begin{aligned} &P(Y \leqslant X_1' \beta_0(\tau) + g_0(X_2;\tau) \mid D=1, \\ &\quad X = \boldsymbol{x}, Z = z) = P(Y^* \leqslant X_1' \beta_0(\tau) + \\ &\qquad g_0(X_2;\tau) \mid D=1, X=x, Z=z) \\ &\quad = P(\varepsilon \leqslant \tau \mid V \leqslant p_0(z), X=x, Z=z) \\ &\quad = P(\varepsilon \leqslant \tau \mid V \leqslant p_0(z)) \\ &\quad = C(\tau, p_0(z)) / p_0(z) \end{aligned} \tag{9}$$

其中 $C(\cdot,\cdot)$ 表示(ε,V) 的联合分布函数. 由于存在样本选择问题，即 ε 与 V 相关，因而 $C(\tau, p_0(z))/p_0(z) \neq \tau$. 式(9) 阐述了当模型中存在选择偏误时，潜变量Y^* 的条件分位点 τ 对应到可观测变量 Y 的条件分位点时发生了转动(rotation)，其映射关系为 $\tau \mapsto C(\tau, p_0(z))/p_0(z)$. 因此本研究可以推断，对于任意 $\tau \in (0,1)$，Y^* 的第 τ 条件分位数与 Y 给定 $D=1$ 的第 $C(\tau,p_0(z))/p_0(z)$ 条件分位数一致. 因此，假如本研究知道两者条件分位点之间的映射关系，便可利用可观测变量来识别$\beta_0(\tau)$ 与$g_0(X_2;\tau)$.

由于 ε 与 V 相关，因此直接针对联合分布函数进行建模较为困难. 根据 Sklar 定理，对于具有连续边缘分布的二元联合分布函数，存在唯一的 Copula 函数$C^*(\cdot,\cdot)$，使得

$$C(\cdot,\cdot) = C^*(F_\varepsilon(\cdot), F_V(\cdot)).$$

Copula 函数$C^*(\cdot,\cdot)$ 通过 Sklar 定理将联合分布函数和边缘分布函数联系起来，并且联合分布关于其相关性的性质完全由 Copula 函数所决定. 因此当 ε 与 V 的边缘分布已知的情况下，本研究可以通过 Copula 函数对其相关性的建模. 由于ε 与 V 分别服从标准均匀分布，因而 $C(\cdot,\cdot) = C^*(\cdot,\cdot)$. 所以在接下来的讨论中，本研究简称联合分布函数 $C(\cdot,\cdot)$ 为 Copula 函数. 参考 Arellano 和 Bonhomme[22] 的思路，假设 Copula 函数由有限维未知参数ρ_0 来决定.

$$C(\tau,p_0) = C(\tau,p_0;\rho_0) \tag{10}$$

其中 ρ_0 的维数为d_ρ. 式(10) 通过有限维未知参数 ρ_0 刻画扰动项之间的相关性，将一个复杂的非参数分布函数估计问题转化为简单的参数估计问题. 统计学相关文献已经提供了一系列便捷、简约

② 由于g_0 形式未知，因此 β_0 中不包含常数项.

③ 不失一般性，本研究可以假设给定 X 时，ε 服从条件标准均匀分布. 由于在本节中仅考虑独立情形，因此本研究在式(7) 中采用无条件表达形式.

的 Copula 函数设定形式，包括 Gaussian Copula，Frank Copula 等（具体可参考 Nelson[32] 和 Joe[33]）.

为了识别 $\beta_0(\tau)$，ρ_0 以及 $g_0(x_2;\tau)$，本研究给出如下假设条件.

假设 1.1 假设式(7)与式(8)成立，且 Z 中至少包含一个 X 没有的变量. 此外，X_1 中变量不存在共线性.

假设 1.2 $C(e,v;\rho_0)$ 表示 (ε,V) 的联合分布函数. 令 $0\leqslant c_1<c_2\leqslant 1$ 以及 $0\leqslant c_3<c_4\leqslant 1$，$C(e,v;\rho_0)$ 在勒贝格测度下绝对连续，其支撑域为 $[c_1,c_2]\times[c_3,c_4]$，且边缘分布服从标准均匀分布.

假设 1.3 条件累积分布函数 $F_{Y^*|X}(\cdot|x)$ 严格递增；$x_1'\beta_0(\tau)$ 与 $g_0(x_2;\tau)$ 关于 τ 严格递增；$C(e,v;\rho_0)$ 关于 e 严格递增.

假设 1.4 令 $\mathscr{Z}$ 表示 Z 的支撑域. $p_0(Z)\equiv P(D=1|Z)>0$ 依概率 1 成立.

假设 1.1 包含独立性假设，排除性假设(exclusion restriction)以及不完全共线性假设. 排除性假设是识别的关键假设. 假设 1.2 限制了不可观测扰动项 (ε,V) 的联合分布函数及其边缘分布. 假设 1.3 要求 Y^* 连续以及分位数回归中的单调性. 假设 1.4 是样本选择模型中的常用假设，限制了倾向得分函数 $p_0(z)$ 的支撑域，保证了式(9)右侧有界.

下述引理详细说明了对于 Copula 函数的限制条件，同时也是识别定理证明的基础. 令 $G(\tau,p_0;\rho_0)=C(\tau,p_0;\rho_0)/p_0$，$\mathscr{X}$ 表示 X 的支撑域. 同时令 $F^{-1}_{Y|D=1,X,Z}$ 与 G^{-1} 分别表示 $F_{Y|D=1,X,Z}$ 与 G 关于其第一个分量的逆函数. 假设 1.3 保证了逆函数的存在性.

引理 1.1 如果假设 1.1 ~ 假设 1.4 成立，则对于所有 $(x,z_1,z_2)\in\mathscr{X}\times\mathscr{Z}\times\mathscr{Z}$，

$$F_{Y|D=1,X,Z}(F^{-1}_{Y|D=1,X,Z}(\tau|x,z_2)|x,z_1)=G(G^{-1}(\tau,p_0(z_2);\rho_0),p_0(z_1);\rho_0) \quad (11)$$

引理 1.1 中 z_1 与 z_2 是支撑集 $\mathscr{Z}$ 中的任意取值④. 注意到当排除性假设不成立，即 Z 为 X 的子集时，给定 $X=x$ 的同时也给定了 $Z=z$，从而有 $z_1=z_2=z$. 这一情形令式(11)左右两侧恒等于 τ 并且等式恒成立. 此时式(11)将不提供任何信息，进而导致 $\beta_0(\tau)$，ρ_0 以及 $g_0(x_2;\tau)$ 无法识别.

基于假设 1.1 和假设 1.4，本研究只能部分识别 $\beta_0(\tau)$，ρ_0 以及 $g_0(x_2;\tau)$，例如仅能识别其属于某一区间. 根据式(9)可知，如果本研究能识别 Copula 函数，或 Copula 参数 ρ_0，本研究就能识别 $\beta_0(\tau)$ 与 $g_0(x_2;\tau)$. 虽然本研究无法直接通过引理 1.1 识别 ρ_0，但其为接下来的识别证明提供了重要信息. 本研究进一步给出下列两种简单情形并结合引理 1.1 来识别 $\beta_0(\tau)$，ρ_0 以及 $g_0(x_2;\tau)$，并且满足任意一种情形都能保证识别结果成立.

假设 1.5 令 P_Z 表示 $p_0(Z)$ 的支撑域. 以下两个条件之一满足.

(i) 存在 $z\in Z$，使得 $p_0(z)=1$.

(ii) $\mathscr{P}_Z$ 包含一个开区间，并且对于所有的 $\tau\in(0,1)$，函数 $p_0\mapsto C(\tau,p_0;\rho_0)/p_0$ 在单位区间内是实解析(real analytic)函数.

定理 1.1 如果假设 1.1 和假设 1.5 成立，则 $\beta_0(\tau)$，ρ_0 以及 $g_0(x_2;\tau)$ 可识别.

证明 本研究首先假设 1.5(i)成立. 假设 $z^*\in\mathscr{Z}$ 满足 $p_0(z^*)=1$. 首先注意到 $G^{-1}(\tau,1;\rho_0)=\tau$. 在式(11)中本研究令 $z_1=z$ 以及 $z_2=z^*$，则本研究有

$$G(\tau,p_0(z);\rho_0)=F_{Y|D=1,X,Z}(F^{-1}_{Y|D=1,X,Z}(\tau|x,z^*)|x,z)$$

因此 ρ_0 可识别. 根据式(9)可知 $\beta_0(\tau)$ 与 $g_0(x_2;\tau)$ 亦可识别.

假设 1.5(ii)成立. 假设存在另一个可能的参数 $\tilde{\rho}\neq\rho_0$ 满足式(1) ~ 式(3). 根据式(11)，对于所有 $(p_1,p_2)\in\mathscr{P}_Z\times\mathscr{P}_Z$，本研究可得

$$G(G^{-1}(\tau,p_2;\rho_0),p_1;\rho_0)-G(G^{-1}(\tau,p_2;\tilde{\rho}),p_1;\tilde{\rho})=0$$

假设 $\mathscr{P}_Z$ 包含的开区间表示为 (l_1,l_2). 则对于任意 $\tau\in(0,1)$，函数

$$(p_1,p_2)\mapsto G(G^{-1}(\tau,p_2;\rho_0),p_1;\rho_0)-G(G^{-1}(\tau,p_2;\tilde{\rho}),p_1;\tilde{\rho})$$

是实解析的，并且在 $(l_1,l_2)\times(l_1,l_2)$ 上等于 0. 基于实解析函数的性质，其在 $(0,1)\times(0,1)$ 上处处等于 0. 令 $p_2=1$. 则对于所有 $p_1\in(0,1)$ 以及任

④ 引理 1.1 的证明可向作者备索.

意 $\tau \in (0,1)$，

$$G(\tau,p_1;\rho_0) - G(\tau,p_1;\tilde{\rho}) = 0$$

当且仅当 $\tilde{\rho} = \rho_0$ 时成立. 因此 ρ_0 可识别，并且 $\beta_0(\tau)$ 与 $g_0(x_2;\tau)$ 亦可识别.

1.3　模型估计

本小节将给出 $\beta_0(\tau)$，ρ_0 以及 $g_0(x_2;\tau)$ 的具体估计步骤. 为了更清晰地陈述估计步骤，本研究先将式(9) 等价改写成下述条件期望形式

$$\mathrm{E}_{XZ}\left[D\left(\frac{1\{Y \leqslant X_1'\beta_0(\tau) + g_0(X_2;\tau)\} - \tau}{G(\tau,p_0(Z);\rho_0)} \right) \right] = 0 \quad (12)$$

其中 $G(\tau,p_0;\rho_0) = C(\tau,p_0;\rho_0)/p_0$，$\mathrm{E}_{XZ}$ 表示给定 (X,Z) 的条件期望. 令 $\boldsymbol{\xi} = (Y,D,X,Z)$，$W = (X_1,Z)$. 基于式(12)，本研究可以构造下列矩条件

$$\mathrm{E}[m_1(\xi;\beta_0,\rho_0,g_0,p_0,\tau) \mid X_2 = x_2] = 0 \quad (13)$$

$$\mathrm{E}\, m_2(\xi;\beta_0,\rho_0,g_0,p_0,\tau) = 0 \quad (14)$$

其中

$$m_1(\xi;\beta_0,\rho_0,g_0,p_0,\tau) = D\left(\frac{1\{Y \leqslant X_1'\beta_0(\tau) + g_0(X_2;\tau)\} - \tau}{G(\tau,p_0(Z);\rho_0)} \right),$$

$$m_2(\xi;\beta_0,\rho_0,g_0,p_0,\tau) = D\left(\frac{1\{Y \leqslant X_1'\beta_0(\tau) + g_0(X_2;\tau)\} - \tau}{G(\tau,p_0(Z);\rho_0)} \right)\phi(W).$$

在上述等式中，$\phi(\cdot)$ 为已知函数，维数为 d_ϕ. 在给定 β_0 和 ρ_0 之后，式(13) 可以看作给定 $\boldsymbol{X}_2 = \boldsymbol{x}_2$ 的条件分位数回归的一阶矩条件. 因此基于式(13)，在给定了 β_0 和 ρ_0 之后，本研究可以得到 g_0 的局部估计量[34, 35]. 之后，可以利用无条件形式的矩条件式(14) 来同时估计 β_0 和 ρ_0. 此时 $\phi(\cdot)$，β_0 与 ρ_0 的维数之间需满足不等式 $d_\phi \geqslant d_1 + d_\rho$. 具体的估计步骤可下.

步骤 1　首先本研究估计倾向得分函数 $p_0(z) \equiv P(D = 1 \mid Z = z)$. 文献中常见有两类估计方法，其一是假设函数由一组未知参数所决定，即 $p_0(z) \equiv p(z;\eta_0)$，并通过估计参数 η_0 来得到倾向得分函数的估计量，即 $\hat{p}(z) \equiv p(z;\hat{\eta})$. 另一类则是非参数方法，通过核函数来估计条件期望. 在本研究的模型中可以由

$$\hat{p}(z) = \sum_{i=1}^{n} D_i K_{h_Z}(Z_i - z) \Big/ \sum_{i=1}^{n} K_{h_Z}(Z_i - z)$$

估计得到，其中

$$K_{h_Z}(Z - z) = \frac{1}{h_Z^{d_Z}} k\left(\frac{Z - z}{h_z}\right) = \frac{1}{h_Z^{d_Z}} \prod_{s=1}^{d_Z} \bar{k}\left(\frac{Z_s - z_s}{h_Z}\right)$$

d_z 为 Z 的维数，$\bar{k}(\cdot)$ 为单变量核函数⑤. 参数估计方法的优势在于估计步骤简单，估计量的收敛速度快且大样本性质的推导简单，缺点是需要对函数进行参数化假设，面临模型误设风险. 非参数估计方法则恰恰相反，无需对函数形式进行假定，估计量相对稳健，代价是估计相对复杂，存在维数诅咒问题，且大样本性质的推导相对繁琐. 出于估计量的稳健性考虑，本研究采用非参数估计方法.

步骤 2　基于式(13)，本研究采用核函数方法来估计 g_0. 给定任意 b 与 ρ，对于 $\tau \in (0,1)$，本研究有

$$\breve{g}(x_2;b,\rho,\tau) = \underset{\alpha}{\operatorname{argmin}}\, S_{1n}(\alpha \mid x_2,b,\rho,\tau),$$

其中

$$S_{1n}(\alpha \mid x_2,b,\rho,\tau) = \frac{1}{n}\sum_{i=1}^{n} D_i(G(\tau,\hat{p}(Z_i);\rho) - 1\{Y_i \leqslant X_{1i}'b + \alpha\}) \times (Y_i - X_{1i}'b - \alpha) K_h(X_{2i} - \boldsymbol{x}_2),$$

K_h 为多变量核函数，形式为

$$K_h(X_2 - \boldsymbol{x}_2) = \frac{1}{h^{d_2}} k\left(\frac{X_2 - \boldsymbol{x}_2}{h}\right) = \frac{1}{h^{d_2}} \prod_{s=1}^{d_2} \bar{k}\left(\frac{X_{2s} - x_{2s}}{h}\right)$$

d_2 为 X_2 的维数. 由于目标函数 S_{1n} 中包含非参数形式的估计量 $\hat{p}$，并且目标函数 S_{1n} 关于 α 不可导，因而导致大样本性质分析变得十分复杂. 本研究参考 Horowitz[36] 中的思路，对目标函数 S_{1n} 中的指示函数 $1\{\cdot\}$ 进行光滑化. 假设 $\tilde{k}(s)$ 为密度核函数，并令

$$\Lambda(s) = \int_{-\infty}^{s} \tilde{k}(t)\,\mathrm{d}t$$

此外，令 $\tilde{h}$ 表示 $\Lambda(\cdot)$ 对应的窗宽. 因此 g_0 的实际估计量为

$$\hat{g}(x_2;b,\rho,\tau) = \underset{\alpha}{\operatorname{argmin}}\, \hat{S}_{1n}(\alpha \mid x_2,b,\rho,\tau),$$

⑤　为简化记号，本研究在正文中假设 (X, Z) 均为连续变量. 一般地，本研究允许 (X, Z) 为离散变量.

其中

$$\widehat{S}_{1n}(\alpha \mid x_2,b,\rho,\tau) = \frac{1}{n}\sum_{i=1}^{n} D_i \times \left(G(\tau,\widehat{p}(Z_i);\rho) - \Lambda\left(\frac{X_{1i}'b + \alpha - Y_i}{\tilde{h}}\right)\right) \times (Y_i - X_{1i}'b - \alpha) K_h(X_{2i} - x_2)$$

步骤 3 令

$$M_{2n}(b,\rho,g,p,\tau) = \frac{1}{n}\sum_{i=1}^{n} m_2(\boldsymbol{\xi}_i;b,\rho, g(\cdot;b,\rho,\tau),p(\cdot),\tau)$$

对于 $\tau \in (0,1)$，本研究可以通过求解下述最小化问题同时估计 β_0 与 ρ_0.

$$\min_{b,\rho} \| M_{2n}(b,\rho,\widehat{g},\widehat{p},\tau) \|,$$

其中 $\|\cdot\|$ 表示欧几里得范数，并得到估计量 $\widehat{\beta}(\tau)$ 和 $\widehat{\rho}(\tau)$. 由于 ρ_0 不随 τ 变化，本研究可以通过下述平均方法

$$\widehat{\rho} = \frac{1}{\#T}\sum_{\tau\in T}\widehat{\rho}(\tau)$$

来提高估计量的有效性，其中 T 表示分位数点 τ 取值的集合，$\#T$ 表示集合 T 中元素个数.

最后，本研究代入 $\widehat{\beta}(\tau)$ 和 $\widehat{\rho}$ 可得

$$\widehat{g}(x_2;\tau) = \widehat{g}(\boldsymbol{x}_2;\widehat{\beta}(\tau),\widehat{\rho},\tau).$$

2 大样本性质与统计推断

2.1 估计量渐近分布

本研究给出以下假设条件.

假设 2.1 $\{Y_i,D_i,Z_i,X_i\}_{i=1}^{n}$ 独立同分布.

假设 2.2 对于任意 $\tau \in (0,1)$，$\beta_0(\tau)$，ρ_0 属于紧空间 $\mathscr{B}\times\mathscr{L}_\rho$ 的内点，且 $\mathscr{B}\times\mathscr{L}_\rho \subset \mathscr{R}^{d_1}\times\mathscr{R}^{d_\rho}$. $\mathscr{X}_2$ 与 $\mathscr{Z}$ 分别表示 X_2 与 Z 的支撑域，$\mathscr{X}_2\times\mathscr{Z}\subset R^{d_2}\times R^{d_Z}$. 则 $\mathscr{X}_2$ 与 $\mathscr{Z}$ 凸且有界.

假设 2.3 $\bar{k},\hat{k}$ 为 L 阶核函数，并且它们的导数均有界. 窗宽 $h_Z,h,\tilde{h}$ 满足 $h_Z \propto n^{-\sigma_Z}$，$h\propto n^{-\sigma}$，$\tilde{h}\propto n^{-\tilde{\sigma}}$，并且

(i) $1/(4L) < \sigma_Z < 1/(2d_Z)$;

(ii) $1/(4L) < \sigma, 1/(4L) < \tilde{\sigma}$;

(iii) $d_2\sigma + \tilde{\sigma} < 1/2$;

(iv) $d_2\sigma + d_Z\sigma_Z < 1/2$.

假设 2.4 对于任意 $\tau\in(0,1)$，假设函数 $(g_0(\cdot;\tau),p_0(\cdot))$ 属于空间 $\mathscr{G}\times\mathscr{P}$. 此外，对于任意 $(g(\cdot;\tau),p(\cdot))\in\mathscr{G}\times\mathscr{P}$，$g(\cdot;\tau)$ 满足 κ_g 阶连续可微，且 $\kappa_g > d_2$ 及 $\kappa_g \geqslant L$；$p(\cdot)$ 满足 κ_p 阶连续可微，且 $\kappa_p > d_Z$ 及 $\kappa_p \geqslant L$；其中 d_2 与 d_Z 分别表示 X_2 与 Z 的维数.

假设 2.5 密度函数 $f_{DX_1X_2Z}(d,\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{z})$ 关于分量 $(x_2,\boldsymbol{z})$ 满足 L 阶连续可微，且导数一致有界并大于 0；令 $U = X_1'b + \alpha - Y$. 对于任意 $\tau\in(0,1)$，$(b,\alpha)\in\mathscr{B}\times\mathscr{G}$，密度函数 $f_{U\mid DX_2Z}(u\mid d,x_2,\boldsymbol{z})$ 关于分量 u 满足 L 阶连续可微，且导数一致有界；$G(\tau,p;\rho)$ 关于分量 τ 严格递增，关于分量 p 满足 L 阶连续可微，关于分量 ρ 满足 2 阶连续可微，且导数一致有界；$F_{Y\mid XZ}(\cdot\mid X,Z)$ 满足 Lipschitz 连续条件.

假设 2.6 定义

$$M_2(b,\rho,g,p,\tau) = \mathrm{E}m_2(\xi;b,\rho,g(\cdot;b,\rho,\tau),p(\cdot),\tau)$$

则目标函数 $\|M_2(b,\rho,g_0,p_0,\tau)\|$ 在 $(b,\rho) = (\beta_0(\tau),\rho_0)$ 处取到唯一最小值.

假设 2.7 令

$$\mathrm{H}_1(\alpha\mid x_2,\boldsymbol{b}) = \mathrm{E}(f_{YX_2\mid DX_1}(X_1'b + \alpha,x_2\mid D,X_1)D)$$

则对于任意 $\tau\in(0,1)$，$x_2\in X_2$，$b\in B$，$\rho\in\mathscr{D}$，$H_1(g_0(x_2;b,\rho,\tau)\mid x_2,b) > 0$.

定义 $M_2(b,\rho,g,p,\tau)$ 关于 (b,ρ) 的导数为

$$\boldsymbol{\Delta}_{1b}(b,\rho,g,p,\tau) = \frac{\partial}{\partial b'}M_2(b,\rho,g,p,\tau),$$

$$\boldsymbol{\Delta}_{1\rho}(b,\rho,g,p,\tau) = \frac{\partial}{\partial \rho'}M_2(b,\rho,g,p,\tau),$$

以及

$$\boldsymbol{\Delta}_1(b,\rho,g,p,\tau) = [\boldsymbol{\Delta}_{1b}(b,\rho,g,p,\tau), \boldsymbol{\Delta}_{1\rho}(b,\rho,g,p,\tau)].$$

则对于任意 $\tau\in(0,1)$，矩阵 $\boldsymbol{\Delta}_1(\tau) = \boldsymbol{\Delta}_1(\beta_0,\rho_0,g_0,p_0,\tau)$ 列满秩.

假设 2.1 为正则性假设. 假设 2.2 定义了估计量的参数空间以及变量所属空间. 假设 2.3 给出了核函数及其对应窗宽的选择标准. 假设 2.4 ~ 假设 2.5 包含了一系列光滑性假设. 其中 2.4 对于 $(g_0(\cdot;\tau),p_0(\cdot))$ 的光滑性假定需要满足两个条件，其一是随机等度连续(stochastic equicontinuity) 假设，即空间 $\mathscr{G}\times\mathscr{P}$ 的覆盖数需要满足 Chen 等[37] 中

条件 3.3，因此要求空间 $\mathscr{G}\times\mathscr{P}$ 中的函数足够光滑⑥；另一个条件则是在泰勒展开时候所需要的光滑性假设. 此外，假设 2.5 中对于 $F_{Y|XZ}(\cdot|X,Z)$ 的 Lipschitz 连续条件保证了目标函数 m_2 关于 (b,ρ,g,p) 满足局部的一致连续性质⑦. 假设 2.6 是一个全局识别条件，可由式(14)及定理 1.1 推出. 引入此假设仅为证明中表述更加清晰. 假设 2.7 是满秩假设，保证了渐近协方差矩阵可逆. 基于以上假设条件，可以证明如下定理成立.

定理 2.1　如果假设 1.1 ~ 假设 1.5 以及假设 2.1 ~ 假设 2.7 成立，对于任意 $\tau\in(0,1)$，

$$\sqrt{n}(\widehat{\beta}(\tau)-\beta_0(\tau))\rightarrow^{d}N(0,\boldsymbol{\Omega}_{\beta}(\tau)),$$

以及

$$\sqrt{n}(\widehat{\rho}-\rho_0)\rightarrow^{d}N(0,\boldsymbol{\Omega}_{\rho}).$$

定理 2.1 中渐近方差 $\boldsymbol{\Omega}_{\beta}(\tau)$ 与 $\boldsymbol{\Omega}_{\rho}$ 均是正定矩阵.

2.2　统计推断：自助法

为了进行统计推断，本研究需要估计渐近协方差 $\boldsymbol{\Omega}_{\beta}(\tau)$ 以及 $\boldsymbol{\Omega}_{\rho}$ 来构造估计量的置信区间. 如定理 2.1 所示，由于 $\beta_0(\tau)$，ρ_0 的估计步骤中包含 p_0 以及 g_0 的估计量，从而导致渐近协方差矩阵的形式异常复杂，使得直接估计渐近协方差矩阵 $\boldsymbol{\Omega}_{\beta}(\tau)$ 与 $\boldsymbol{\Omega}_{\rho}$ 变得非常困难. 为了避免直接估计协方差，文献中常用的一类方法是采用非参数自助法(nonparametric bootstrap)，通过重抽样来构造估计量的置信区间. 本研究参考 Chen 等[37] 中的相关结论来建立自助法的合理性(validity of bootstrap).

令 $\{\xi_i^*\}_{i=1}^n=\{Y_i^*,D_i^*,Z_i^*,X_i^*\}_{i=1}^n$ 表示通过有放回抽样方法从原始样本 $\{\xi_i\}_{i=1}^n$ 中随机生成的新样本，其概率测度定义为 P^*. 定义 $\widehat{p}^*$，$\widehat{g}^*$，$\widehat{\beta}^*$，$\widehat{\rho}^*$ 分别表示 p_0，g_0，β_0，ρ_0 基于重抽样的样本所得到的估计量⑧. 根据 Chen 等[37] 中定理 B，本研究基于如下定理建立自助法的合理性.

定理 2.2　如果假设 1.1 ~ 假设 1.5 以及假设 2.1 ~ 假设 2.7 成立，对于任意 $\tau\in(0,1)$，依概率测度 P^*，

$$\sqrt{n}(\widehat{\beta}^*(\tau)-\widehat{\beta}(\tau))\rightarrow^{d}N(0,\boldsymbol{\Omega}_{\beta}(\tau)),$$

以及

$$\sqrt{n}(\widehat{\rho}^*-\widehat{\rho})\rightarrow^{d}N(0,\boldsymbol{\Omega}_{\rho}).$$

定理 2.2 的结论说明了本研究可以利用自助法来构造 $\widehat{\beta}(\tau)$ 以及 $\widehat{\rho}$ 的置信区间⑨. 由于 $\widehat{\beta}^*(\tau)$ 以及 $\widehat{\rho}^*$ 的方差估计仅需不断重复 1.3 节中步骤 1 ~ 步骤 3，避免去逐项估计渐近协方差形式，因此实际应用非常方便. 基于定理 2.2，本研究通过自助法来构造置信区间的具体步骤如下.

步骤 2.1　从[a，1 - a]中选取一个有限集合 T 作为分位点的选择集合，其中 $a\in(0,1/2)$. 基于 1.3 节中步骤 1 - 3，对于任意 $\tau\in T$，估计 $\widehat{\beta}(\tau)$ 与 $\widehat{\rho}$.

步骤 2.2　设置迭代次数 B. 对于 $b=1,\cdots,B$，通过有放回抽样方法生成新样本 $\{\xi_{ib}^*\}_{i=1}^n$.

步骤 2.3　对于每一组新样本 $\{\xi_{ib}^*\}_{i=1}^n$ 以及任意 $\tau\in T$，计算 $\widehat{\beta}_b^*(\tau)$ 与 $\widehat{\rho}_b^*$.

步骤 2.4　对于任意 $\tau\in T$，定义 $\widehat{\sigma}_{\beta}(\tau)$ 与 $\widehat{\sigma}_{\rho}$ 分别等于 $\{\widehat{\beta}_b^*(\tau)\}_{b=1}^B$ 与 $\{\widehat{\rho}_b^*\}_{b=1}^B$ 的样本标准差. 构造 $\widehat{\beta}(\tau)$ 与 $\widehat{\rho}$ 的 $100(1-l)\%$ 置信水平的置信区间分别为

$$[\widehat{\beta}(\tau)\pm\widehat{\sigma}_{\beta}(\tau)\,\Phi^{-1}(1-l/2)]$$

以及

$$[\widehat{\rho}\pm\widehat{\sigma}_{\rho}\,\Phi^{-1}(1-l/2)]$$

其中 $\Phi(\cdot)$ 表示标准正态分布的累积分布函数.

2.3　非线性设定检验

虽然本研究提出的偏线性模型设定可以很好的平衡估计量的稳健性以及模型误设的风险，但如果偏线性部分 g 为线性时，本研究的估计量将会损失一定的效率. 因此在本小节中，本研究提出

---

⑥　与 Chen 等[37] 中的讨论类似，假设 2.4 中 (g,p) 所需的光滑程度与目标函数 m_2 的关于 (g,p) 的光滑程度负相关. 本研究中，由于 m_2 关于 (g,p) 欠光滑，因此需要提高 (g,p) 的光滑程度以满足随机等度连续假设.

⑦　任意函数 m 的局部的一致连续性质定义为对于任意正数序列 $\{\delta\}$ 满足 $\delta=o(1)$，常数 $K>0$ 以及参数 θ，以下不等式成立. $E\Big[\sup_{\|\theta-\theta\|\le K\delta}|m(\xi;\theta)-m(\xi;\theta)|^2\Big]\le K\delta$.

⑧　如 Hall 和 Horowitz[38]，Mammen 等[39]，Chen 等[37] 所述，$\widehat{\beta}^*$，$\widehat{\rho}^*$ 的估计与步骤 3 中略有不同，本研究需要对目标函数进行中心化. 具体形式可见 Chen 等[37] 中式(2). 而 $\widehat{p}^*$，$\widehat{g}^*$ 的估计无需变化，仅要求其一致收敛速度满足一定条件即可.

⑨　定理 2.1 ~ 定理 2.2 的详细证明可向作者备索.

一种简便的统计量来检验下述假设.

$H_0: \forall (x_2,\tau) \in X_2 \times (0,1), \exists \zeta(\cdot)$, 使得 $g_0(x_2;\tau) = x_2'\zeta(\tau)$.

那么与原假设H_0 对应的备择假设则是

$H_1: \forall (x_2,\tau) \in X_2 \times (0,1)$ 以及 $\zeta(\cdot)$,

$$g_0(x_2;\tau) \neq x_2'\zeta(\tau).$$

构造如下检验统计量

$$Y = n\int \frac{[g_0(x_2;\tau) - x_2'\zeta(\tau)]^2}{\omega(x_2,\tau)} \times \mathrm{d}x_2\mathrm{d}\tau \quad (15)$$

来检验原假设H_0,其中$\omega(\cdot,\cdot)$ 为权重函数. 当本研究的检验统计量 Y 接近0时,说明原假设成立. 根据 Arellano 和 Bonhomme[22],本研究可以得到 $\zeta(\tau)$ 的估计量$\hat{\zeta}(\tau)$. 在实际中为了计算简便,本研究可以选择$\omega(x_2,\tau) = f_{X_2}(x_2)1\{\tau \in T\}/\#T$作为权重函数,其中$f_{X_2}(\cdot)$ 表示X_2 的密度函数⑩. 代入估计量$\hat{g}(\cdot,\cdot)$ 以及$\hat{\zeta}(\cdot)$ 之后,本研究可以得到与式(15) 对应的可行检验统计量

$$\hat{Y} = \frac{1}{\#T}\sum_{\tau\in T}\sum_{i=1}^{n}[\hat{g}(X_{2i};\tau) - X_{2i}'\hat{\zeta}(\tau)]^2$$

由于直接建立检验统计量$\hat{\mathscr{Y}}$的渐近分布较为困难,因此本研究参考 Lu 和 White[40] 的思路,通过重抽样方法来模拟其渐近分布.

令 $s \equiv s_n$ 表示一个正数序列,当 $n \to \infty$,其满足 $s/n \to 0$ 以及 $s \to \infty$. 又令

$$\hat{\mathscr{H}}(x_2,\tau) = \hat{g}(x_2;\tau) - x_2'\hat{\zeta}(\tau)$$

下面给出构造可行检验统计量分布的具体步骤.

步骤 2.1 设置迭代次数 B. 对于 $k = 1,\cdots,B$,通过有放回抽样方法生成样本量为 s 的新样本 $\{Y_{ik},D_{ik},Z_{ik},X_{ik}\}_{i=1}^{s}$,其指标集定义为$N_k$.

步骤 2.2 对于 $k = 1,\cdots,B$,使用样本$\{Y_{ik}, D_{ik}, Z_{ik}, X_{ik}\}_{i=1}^{s}$ 计算 $\hat{\mathscr{H}}(x_2,\tau)$, 并定义为 $\hat{\mathscr{H}}_k(x_2,\tau)$.

步骤 2.3 检验统计量的分布可以由下式得到

$$\hat{F}_Y(t) = \frac{1}{B}\sum_{k=1}^{B} I\left\{\frac{1}{\#T}\sum_{\tau\in T}\sum_{i\in N_k}[\hat{\mathscr{H}}_k(X_{2i},\tau) - \hat{\mathscr{H}}(X_{2i},\tau)]^2 \leqslant t\right\}$$

本研究可以通过检验统计量所在分布的位置来检验原假设. 例如本研究设置显著性水平为5%. 那么如果$\hat{F}_Y(\hat{Y}) \geqslant 0.95$,即检验统计量的值落入拒绝域,则可以说明这种情况下本研究拒绝原假设,认为模型中存在非线性问题,反之亦然.

3 内生性

本研究在第 1 节中所讨论的估计步骤同样适用于处理内生性问题. 本研究考虑线性分位数回归模型式(2) ~ 式(3) 以及

$$Y^* = X_1'\beta_0(\tau) + e \quad (16)$$

并且X_1 与 e 相关,即

$$P(e \leqslant 0 \mid X) \neq \tau$$

此时模型中存在内生性问题. 解决模型内生性问题的方法有很多,文献中常见的一种方法是控制函数法[41, 42]. 假设$X_1 = ((X_1^E)', (X_1^I)')'$,其中$X_1^E$ 表示内生变量,X_1^I 为外生变量. 内生变量X_1^E 满足下述生成过程

$$X_1^E = E[X_1^E \mid R] + X_2 = \zeta(R) + X_2$$

其中 R 为可观测外生变量,X_2 表示不可观测扰动项. 注意到

$$Q_e(\tau \mid X_1,X_2) = Q_e(\tau \mid X_2) = g_0(X_2;\tau) \quad (17)$$

根据式(17),本研究可以将式(16) 变换成

$$Y^* = X_1'\beta_0(\tau) + g_0(X_2;\tau) + \tilde{\varepsilon} \quad (18)$$

其中新扰动项$\tilde{\varepsilon} = e - g_0(X_2;\tau)$ 满足

$$P(\tilde{\varepsilon} \leqslant 0 \mid X) = \tau$$

此时本研究将一个内生情形下的线性分位数回归模型转换成外生情形下的偏线性分位数回归模型. 由于内生性问题的存在,式(8) 中的无条件独立性假设不再成立. 参照 Arellano 和 Bonhomme[21],本研究提出下述条件独立性假设

$$Z \perp (\tilde{\varepsilon},V) \mid X \quad (19)$$

为了简化记号,假设 Z 包含所有 X 中的元素. 根据式(19),对式(9) 中的矩条件进行修改. 注意到

$$P(Y \leqslant X_1'\beta_0(\tau) + g_0(X_2;\tau) \mid D = 1, X = x, Z = z) = P(\tilde{\varepsilon} \leqslant 0 \mid V \leqslant p_0(z),$$

$$X = x, Z = z) = P(F_{\tilde{\varepsilon}\mid X}(\tilde{\varepsilon} \mid X) \leqslant F_{\tilde{\varepsilon}\mid X}(0) \mid V \leqslant p_0(z),$$

$$X = x, Z = z) = P(\varepsilon \leqslant \tau \mid V \leqslant p_0(z),$$

⑩ 根据 Lu 和 White[40] 可知,权重函数 $\omega(x_2,\tau)$ 满足条件$\int \omega(x_2,\tau)\mathrm{d}x_2\mathrm{d}\tau = 1$ 即可. 在模拟中本研究尝试更换不同的权重函数,发现检验结果对于权重函数的选择较为稳健,因此在模拟中只报告上述权重函数的检验结果.

$$X = x) = C_x(\tau, p_0(z)) / p_0(z) \tag{20}$$

其中 $\varepsilon = F_{\varepsilon|X}(\tilde{\varepsilon} \mid X)$ 服从条件标准均匀分布，$C_x(\cdot,\cdot)$ 表示 (ε, V) 给定 $X = x$ 的联合分布函数（或称条件 Copula 函数）. 由式(20) 可知，本研究可以采用与第 1 节中相似的步骤来进行估计. 同样地，本研究假设 Copula 函数由未知参数 ρ_0 决定.

$$C_x(\tau, p_0) = C_x(\tau, p_0; \rho_0) \tag{21}$$

令 $G_x(\tau, p_0; \rho_0) = C_x(\tau, p_0; \rho_0) / p_0$. 根据式(21)，本研究可以得到如下矩条件

$$\mathrm{E}[m_1^E(\xi; \beta_0, \rho_0, g_0, p_0, \tau) \mid X_2 = x_2] = 0,$$

$$\mathrm{E}\, m_2^E(\xi; \beta_0, \rho_0, g_0, p_0, \tau) = 0$$

其中

$$m_1^E(\xi; \beta_0, \rho_0, g_0, p_0, \tau) = D(1\{Y \leqslant X_1' \beta_0(\tau) + g_0(X_2; \tau)\} - G_X(\tau, p_0(Z), \rho_0)), m_2^E(\xi; \beta_0, \rho_0, g_0, p_0, \tau) = D(1\{Y \leqslant X_1' \beta_0(\tau) + g_0(X_2; \tau)\} - G_X(\tau, p_0(Z); \rho_0))\phi(W)$$

具体的估计步骤如下.

步骤 1　通过

$$\widehat{X}_{2i} = X_{1i}^E - \widehat{\zeta}(R_i)$$

估计 X_{2i}，其中

$$\widehat{\zeta}(r) = \sum_{i=1}^n X_{1i}^E K_{h_R}(R_i - r) \Big/ \sum_{i=1}^n K_{h_R}(R_i - r),$$

$$K_{h_R}(R_i - r) = \frac{1}{h_R^{d_R}} k\left(\frac{R - r}{h_R}\right) = \frac{1}{h_R^{d_R}} \prod_{s=1}^{d_R} \bar{k}\left(\frac{R_s - r_s}{h_R}\right)$$

式中 d_R 为 R 的维数.

步骤 2　令 $\widehat{Z}_i$ 表示 Z_i 中的 X_{2i} 替换成 $\widehat{X}_{2i}$. 因此 $p_0(z)$ 可由

$$\widehat{p^E}(z) = \sum_{i=1}^n D_i K_{h_Z}(\widehat{Z}_i - z) \Big/ \sum_{i=1}^n K_{h_Z}(\widehat{Z}_i - z)$$

估计得到.

步骤 3　给定任意 b 与 ρ，对于 $\tau \in (0,1)$，估计 $g_0(x_2; b, \rho, \tau)$. 与步骤 2 类似，本研究有

$$\widehat{g^E}(x_2; b, \rho, \tau) = \operatorname*{argmin}_{\alpha} \widehat{S}_{1n}^E(\alpha \mid x_2, b, \rho, \tau),$$

其中

$$\widehat{S}_{1n}^E(\alpha \mid x_2, b, \rho, \tau) = \frac{1}{n} \sum_{i=1}^n D_i \times \left(G_{X_i}(\tau, \widehat{p^E}(\widehat{Z}_i); \rho) - \Lambda\left(\frac{X_{1i}' b + \alpha - Y_i}{\tilde{h}}\right)\right) \times (Y_i - X_{1i}' b - \alpha) K_h(\widehat{X}_{2i} - x_2)$$

步骤 4　令

$$\widehat{M}_{2n}^E(b, \rho, g, p, \tau) = \frac{1}{n} \sum_{i=1}^n m_2^E(\widehat{\xi}_i; b, \rho, g(\cdot; b, \rho, \tau), p(\cdot), \tau)$$

其中 $\widehat{\xi}_i = (Y_i, D_i, \widehat{Z}_i)$. 对于 $\tau \in (0,1)$，本研究可以通过求解下述最小化问题同时估计 β_0 与 ρ_0.

$$\min_{b,\rho} \| \widehat{M}_{2n}^E(b, \rho, \widehat{g^E}, \widehat{p^E}, \tau) \|$$

并得到估计量 $\widehat{\beta^E}(\tau)$ 和 $\widehat{\rho^E}(\tau)$. 类似地，本研究可以做如下平均

$$\widehat{\rho^E} = \frac{1}{\#T} \sum_{\tau \in T} \widehat{\rho^E}(\tau)$$

来提高估计量的有效性.

最后，本研究代入 $\widehat{\beta^E}(\tau)$ 和 $\widehat{\rho^E}$ 可得

$$\widehat{g^E}(x_2; \tau) = \widehat{g^E}(x_2; \widehat{\beta^E}(\tau), \widehat{\rho^E}, \tau)$$

从上述估计步骤中可以发现，模型中存在内生性时，估计步骤的主要区别在于 X_{2i} 未知，因此本研究需要先对其进行估计. 其次，由于 X 与扰动项 (ε, V) 之间相关，因此内生情形下的 Copula 函数是控制 X 之后的条件分布函数，而不再是无条件分布函数形式.

4　模　拟

在本节中，通过一个简单的蒙特卡洛模拟，说明本研究在 1.3 节中提出的估计量具有良好的小样本性质，从而给出估计方法在实证中的可行性. 根据结果方程式(6) 和选择方程式(2) 的设定，本研究考虑两种不同的数据生成过程（data generating process，DGP），其一为线性模型设定

$$Y^* = X_1 + X_2 + \varepsilon, D = 1\{X_1 + X_2 + X_3 \geqslant V\}$$

与模型设定相对应，结果方程中的线性部分为 $X_1 \beta_0(\varepsilon) = X_1$，其中本研究令 $\beta_0(\varepsilon) = \beta_0 = 1$. 非线性部分为 $g_0(X_2, \varepsilon) = X_2 + \varepsilon$. 选择方程中 $p_0(Z) = p_0(X_1, X_2, X_3) = X_1 + X_2 + X_3$. 其二为偏线性模型设定

$$Y^* = X_1 + X_2 + \sin(X_2) + \varepsilon,$$

$$D = 1\{X_1 + X_2 + X_3 \geqslant V\}$$

与第一种 DGP 相比，本研究考虑了非线性的模型设定，其中 $g_0(X_2, \varepsilon) = X_2 + \sin(X_2) + \varepsilon$. 本研究令 X_2 服从标准正态分布，X_1 与 X_3 均服从二项分

布 $B(1,0.5)$，并且 X_1，X_2 与 X_3 之间互相独立. 随机扰动项 (ε,V) 服从联合正态分布，其相关系数 $\rho_0=0.5$. 此外，可观测随机变量 (X_1,X_2,X_3) 与随机扰动项 (ε,V) 之间互相独立.

本节中主要估计的参数是 β_0 与 ρ_0. 表 1 与表 2（表 3 与表 4）分别给出了两种 DGP $\beta_0(\tau)$ 与 ρ_0 估计量的模拟结果，包括偏误（bias），标准误（se）以及均方根误差（rmse）. Copula 函数分别选择了 Gaussian 以及 Frank Copula. Copula 函数的选择可以根据研究者的意图进行取舍. 如果研究者更关注实际应用便捷性，那么可以考虑文中提到的 Gaussian 或 Frank Copula；如果研究者倾向于估计结果的稳健性，那么可以采用更一般的 Copula 函数，如双参数 Copula，复合 Copula 等. 在模拟中，本研究选取支撑域在 $[-0.5,0.5]$ 上的均匀分布的概率密度函数作为核函数 $\bar{k}$ 与 $\tilde{k}$. 因此模拟中核函数满足 $L=2$. 此外，根据变量设定本研究可知 Z 与 X_2 中连续变量个数为均 1，因此由假设 2.3 可知窗宽的收敛速度应当满足

$$\begin{cases} 1/8 < \sigma_Z < 1/2, \\ 1/8 < \sigma, \widetilde{\sigma}, \\ \sigma + \max\{\sigma_Z, \widetilde{\sigma}\} < 1/2 \end{cases}$$

模拟中本研究分别选取 $h_Z = \mathrm{std}(Z) \times n^{-1/5}$，$h = \mathrm{std}(X_2) \times n^{-1/5}$ 以及 $\widetilde{h} = n^{-1/5}$，其中 $\sigma_Z = \sigma = \widetilde{\sigma} = 1/5$. 本研究选取的样本量为 200 和 500. 报告的分位数点 τ 从集合 中选取. 标准误通过非参数 bootstrap 方法重复 300 次模拟得到.

在估计步骤中本研究首先得到 ρ_0 关于不同 τ 时候的估计值，因此本研究通过平均方法得到最终 ρ_0 的估计量，并在对应表格的最后一行进行报告. 为了便于寻找估计结果，本研究分别对 $\beta_0(\tau)$ 与 ρ_0 的模拟结果用粗体表示. 通过表 1 ~ 表 4 可以发现本研究的估计量在两种模型设定下表现都比较好：1）在样本量增加的同时，估计量的偏差大致表现出减小的趋势；2）估计量的标准误在样本量增加的时候表现出明显的减小趋势，并且减小的比例与大样本性质分析中的结论基本一致；3）对于 ρ_0 的估计量而言，对于不同 τ 分位点的估计量进行平均有效地降低了 ρ_0 估计量的标准误，表明了本研究对 ρ_0 估计量进行平均处理这一方法是合理且必要的.

此外，本研究采用 2.3 节的方法来计算检验统计量与其分布，其中本研究选择 $s=n^{0.9}$ 以及 $B=100$，并且线性情形的参数估计量由 Arellano 和 Bonhomme[22] 方法得到. 本研究在各个表格最后一列给出了检验统计量的 p 值. 本研究发现在线性模型设定下（表 1、表 3），非线性设定无法通过检验（较大 p 值），因此无法拒绝原假设（模型为偏线性设定）. 反之，本研究可以发现在偏线性模型设定下（表 2、表 4），检验统计量的 p 值均小于 0.05，因此可以在 5% 的显著性水平下拒绝原假设，认为模型中存在非线性部分.

表 1 线性模型设定（Gaussian）

Table 1 Linear model specification (Gaussian)

$n=200$	β_0			ρ_0			p 值
τ	*bias*	*se*	*rmse*	*bias*	*se*	*rmse*	
0.1	−0.080	0.039	0.089	−0.255	0.710	0.754	
0.3	−0.051	0.031	0.060	−0.184	0.508	0.539	
0.5	−0.029	0.038	0.047	−0.064	0.372	0.377	
0.7	−0.009	0.037	0.038	0.111	0.258	0.281	
0.9	0.042	0.046	0.062	0.203	0.398	0.446	
				−0.038	0.449	0.479	0.65
$n=500$	β_0			ρ_0			p 值
τ	*bias*	*se*	*rmse*	*bias*	*se*	*rmse*	
0.1	−0.047	0.026	0.053	−0.305	0.636	0.704	
0.3	−0.024	0.026	0.035	−0.167	0.446	0.476	
0.5	−0.011	0.025	0.028	−0.090	0.341	0.352	
0.7	−0.001	0.024	0.024	0.114	0.203	0.232	
0.9	0.035	0.023	0.041	0.213	0.300	0.368	
				−0.047	0.385	0.428	0.45

表 2　偏线性模型设定(Gaussian)

Table 2 Partial linear model specification (Gaussian)

n = 200	β_0			ρ_0			p 值
τ	*bias*	*se*	*rmse*	*bias*	*se*	*rmse*	
0.1	−0.068	0.044	0.081	−0.184	0.671	0.690	
0.3	−0.032	0.037	0.049	−0.166	0.491	0.514	
0.5	−0.016	0.035	0.038	0.003	0.305	0.302	
0.7	0.007	0.032	0.032	0.042	0.277	0.277	
0.9	0.057	0.031	0.065	0.216	0.396	0.448	
				−0.018	0.428	0.446	0.03
n = 500	β_0			ρ_0			p 值
τ	*bias*	*se*	*rmse*	*bias*	*se*	*rmse*	
0.1	−0.052	0.024	0.057	−0.441	0.659	0.788	
0.3	−0.032	0.031	0.044	−0.161	0.521	0.541	
0.5	−0.023	0.027	0.035	0.002	0.284	0.284	
0.7	−0.006	0.023	0.024	0.084	0.167	0.185	
0.9	0.034	0.021	0.040	0.193	0.241	0.307	
				−0.065	0.374	0.420	0.02

表 3　线性模型设定(Frank)

Table 3 Linear model specification (Frank)

n = 200	β_0			ρ_0			p 值
τ	*bias*	*se*	*rmse*	*bias*	*se*	*rmse*	
0.1	−0.068	0.068	0.078	−0.107	0.497	0.506	
0.3	−0.025	0.025	0.040	−0.011	0.262	0.263	
0.5	0.003	0.030	0.031	0.020	0.296	0.298	
0.7	0.030	0.030	0.044	0.010	0.274	0.275	
0.9	0.074	0.074	0.084	0.001	0.466	0.466	
				−0.017	0.359	0.360	0.74
n = 500	β_0			ρ_0			p 值
τ	*bias*	*se*	*rmse*	*bias*	*se*	*rmse*	
0.1	−0.058	0.020	0.061	−0.026	0.308	0.310	
0.3	−0.023	0.019	0.030	−0.067	0.381	0.385	
0.5	−0.002	0.018	0.019	0.007	0.231	0.231	
0.7	0.020	0.018	0.027	0.037	0.118	0.123	
0.9	0.056	0.021	0.059	0.040	0.288	0.290	
				−0.002	0.265	0.267	0.70

表 4　偏线性模型设定(Frank)

Table 4 Partial linear model specification (Frank)

n = 200	β_0			ρ_0			p 值
τ	*bias*	*se*	*rmse*	*bias*	*se*	*rmse*	
0.1	−0.066	0.033	0.074	−0.158	0.554	0.574	
0.3	−0.029	0.028	0.041	−0.072	0.361	0.367	
0.5	−0.005	0.029	0.029	0.006	0.263	0.263	
0.7	0.021	0.030	0.036	0.020	0.276	0.277	
0.9	0.061	0.036	0.070	0.004	0.501	0.501	
				−0.040	0.391	0.395	0.01

续表 4

Table 4 Continues

n = 500	β_0			ρ_0			p 值
τ	*bias*	*se*	*rmse*	*bias*	*se*	*rmse*	
0.1	−0.054	0.019	0.057	−0.088	0.507	0.512	
0.3	−0.023	0.018	0.029	−0.102	0.154	0.185	
0.5	−0.005	0.017	0.017	0.021	0.274	0.275	
0.7	0.014	0.016	0.022	0.035	0.364	0.366	
0.9	0.048	0.018	0.051	0.082	0.348	0.371	
				−0.011	0.348	0.352	0.02

5 实证研究

为了说明本研究提出的估计量的实用价值，在本节中将估计量应用于研究教育对妇女劳动收入的影响，即教育回报率. 利用德国社会经济委员会(German Socioeconomic Panel，GSOEP) 2008 年的数据. 在剔除一些存在异常值的样本后，最终实证采用的数据由 1 341 名 21 岁 ~ 55 岁的已婚妇女组成. 表 5 给出了变量定义及描述性统计.

表 5 变量定义及描述性统计

Table 5 Definition and descriptive statistics of variables

变量	变量定义	均值	标准差
Lwage	妇女收入的对数	7.972	3.616
Employment	虚拟变量，妇女是否参加工作	0.824	0.381
Edu	妇女的受教育年数	12.656	2.518
Age	妇女的年龄	41.037	6.371
Lhuswage	丈夫收入的对数	10.536	0.904
Husworkhour	丈夫每年工作的千小时数	2.248	0.684
East	虚拟变量，地区是否属于东德	0.217	0.412
*Chld*16	家庭中 7 岁 ~ 16 岁孩子数量	0.720	0.878
*Chld*6	家庭中 6 岁以下孩子数量	0.265	0.441

在模型中，结果方程的因变量(Y) 是 *Lwage*，且仅当选择方程的因变量(D，是否参加工作) 等于 1 时，即 *Employment* = 1 时，*Lwage* 才可观测. 样本中约82% 的已婚妇女参加工作. 同时，数据中全部的丈夫都参加了工作. 因此，该数据非常适用于本研究的研究范畴. 此外，关于孩子数量的变量(*Chld*16，*Chld*6) 只出现于选择方程中，因此本研究选择其为排除性变量. 而其他变量会同时影响 *Employment* 和 *Lwage*. 在教育回报率问题中，采用分位数模型研究收入问题的必要性体现在以下两点：1) 收入是典型的偏态分布，均值回归与分位数回归会有所差异；2) 分位数回归有助于研究自变量对不同收入群体的异质性影响，并能探讨诸如收入差距等现实问题[43]. 本研究建立以下偏线性样本选择模型.

$$\begin{aligned} Lwage = {} & Edu \cdot \beta_{01}(\varepsilon) + Lhuswage \cdot \beta_{02}(\varepsilon) + \\ & Husworkhour \cdot \beta_{03}(\varepsilon) + East \cdot \beta_{04}(\varepsilon) + \\ & g_0(Age;\varepsilon) \end{aligned}$$

$$Employment = 1\left\{p_0\left(\begin{matrix} Edu, Lhuswage, Husworkhour, \\ East, Age, Chld16, Chld6 \end{matrix}\right) \geqslant V\right\}$$

已有的劳动经济学相关文献认为年龄对收入存在非线性影响，从而在经典模型中往往会加入年龄的平方项. 本研究的模型通过未知函数$g_0(\cdot;\cdot)$ 来刻画年龄对收入的影响，并包含经典模型的设定作为特例，能够更加精确地控制年龄的非线性影响.

直观上，本研究期望教育年限对妇女的劳动收入有正向的影响. 此外，如果丈夫的工作时间更长，那么妇女可能需要承担更多的家庭事务，所以对妇女的劳动收入有负向的影响. 如果丈夫的收入更高，那么家庭面临的收入压力可能更小，所以对妇女的劳动收入也可能有负向的影响.

在模型估计上，本研究选择了 $\tau = 0.25$，0.5，0.75 的分位数. 与模拟中的结果类似，在实证研究中估计结果在各个分位点上对于 Copula 函数的选择较为稳健，因此本研究报告 Gaussian Copula 的估计结果. 在核函数与窗宽的选择上，本研究使用了与模拟中相同的均匀分布以及窗宽. 同时，本研究也尝试了调整窗宽来检验估计量对窗宽的稳健性，发现估计结果没有明显的变化. 系数的标准误通过 500 次的非参数 bootstrap 估计得到.

此外，本研究还报告了 Heckman 样本选择的均值回归模型和 Arellano 和 Bonhomme[22] 的线性分位数回归模型的估计结果. 表6 汇报了结果方程中线性部分的系数，从中可以发现：1）教育对妇女的劳动收入有显著的正向影响，且数值以往的研究是一致的；其中，教育回报率呈现明显的异质性影响，在低分位点处教育回报率更高，这与一些现有文献（如程名望等[44]）的结果相吻合的；2）与线性分位数回归模型相比，本研究模型的结果在各个分位点处的系数更小. 因而在此数据中，年龄对收入可能呈现显著的非线性影响，从而简单地将年龄及平方项加入回归模型会导致教育对工资回报率被高估（不一致）；3）丈夫的劳动时间具有显著的负向影响，这与本研究的预期是一致的；4）丈夫的收入没有显著的影响，这可能是因为丈夫收入高的家庭中会拥有更多的孩子，从而家庭面临的收入压力更大，从而抵消了其收入效应，使得估计结果不显著.

表 6　实证估计：教育对妇女劳动收入的影响

Table 6 Empirical estimation: Effect of education on women's labor income

	均值回归模型	线性分位数回归（Arellano 和 Bonhomme[22]）			偏线性分位数回归（本研究模型）		
变量 \ 分位点		0.25	0.5	0.75	0.25	0.5	0.75
Edu	0.259***	0.151***	0.132***	0.010***	0.138***	0.094***	0.081***
	(0.046)	(0.017)	(0.019)	(0.012)	(0.021)	(0.012)	(0.011)
Lhuswage	−0.330**	−0.039	0.005	0.025	0.024	0.021	0.009
	(0.165)	(0.056)	(0.043)	(0.042)	(0.084)	(0.040)	(0.037)
Husworkhour	0.084	−0.184*	−0.195***	−0.135***	−0.177*	−0.166***	−0.148***
	(0.207)	(0.103)	(0.059)	(0.042)	(0.102)	(0.054)	(0.040)
East	0.087***	0.221	−0.030	−0.056	0.073	−0.027	0.022
	(0.028)	(0.137)	(0.088)	(0.046)	(0.112)	(0.071)	(0.065)

注： 1. 括号中为标准误；2. *、**、*** 分别表示 10%、5%、1% 的显著性水平.

进一步解释为何采用线性模型估计得到的教育回报率不一致. 图 1 展示了 $\tau = 0.25$，0.5，0.75 时年龄对妇女劳动收入的影响. 研究发现在不同分位点上，年龄对收入的影响呈现出相似的趋势. 此外，从估计结果的形状来看，年龄对收入的影响呈现出明显的非线性，并且与 U 型（或倒 U 型）有着较大差异. 因此在实证研究中若简单加入年龄（及其平方项）进行回归会导致模型误设，从而得到的估计量不一致.

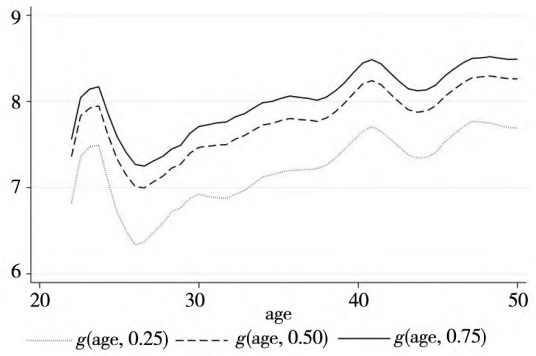

图 1　实证估计：年龄对妇女劳动收入的影响

Fig. 1 Empirical estimation: Effect of age on women's labor income

最后,本研究就非线性部分进行检验,结果呈现在表7. 在检验中选择有放回抽样的样本量 $s = n^{0.9} \approx 652$,模拟分布的迭代次数 $B = 100$. 线性设定下年龄(及其平方项)的系数由 Arellano 和 Bonhomme[22] 的方法得到. 通过估计得到检验统计量的 p 值为0.045,因此可以在5% 的显著性水平下拒绝原假设,即年龄(及其平方项)对女性收入是线性函数关系. 可以设想,如果研究者简单地将年龄(及其平方项)作为协变量进行建模将会面临模型误设,从而得到的估计量不一致(如表6中结论所示). 因此在本例子中本研究需要构建偏线性模型来降低模型误设的风险,得到一致估计量.

表7 实证估计:非线性设定检验

Table 7 Empirical estimation: Nonparametric specification test

检验变量	检验统计量	p 值
Age	171.3	0.045

综合上述结论可以发现,教育对女性工资回报率存在显著的异质性影响. 此外,通过回归结果以及假设检验结果发现,年龄对女性收入存在显著的非线性影响,因此采用本研究提出的偏线性分位数模型来研究女性教育回报率问题,进而能够得到更加稳健以及可靠的结论.

6 结束语

当分位数回归中存在样本选择问题时,本研究通过 Copula 函数对结果方程与选择方程中扰动项的联合分布进行建模,修正了选择偏差. 本研究还提出了三步估计方法来估计模型中的分位数参数(函数)以及 Copula 参数. 第一步是针对选择方程中的非参数回归,给出了倾向得分函数的非参数估计方法;第二步是给定线性部分的分位数参数和 Copula 参数后,对选择偏差进行修正,提供了非线性部分的函数估计方法;第三步则是通过广义矩估计方法来同时估计分位数参数和 Copula 参数. 此外,基于控制函数方法建立了内生情形下线性分位数选择模型与本研究模型之间的联系,同时给出了合理的估计方法. 本研究还给出了非线性部分的检验统计量及其分布的估计方法. 最后本文研究女性教育回报率的实际问题,将本研究模型与已有模型的估计结果进行比较,发现女性教育对收入存在显著的异质性影响,且年龄对于收入存在显著的非线性影响.

参 考 文献:

[1] Zhelonkin M, Genton M G, Ronchetti E. Robust inference in sample selection models[J]. Journal of the Royal Statistical Society, Series B (Statistical Methodology), 2016, 78(4): 805 – 827.

[2] Chen S, Zhou Y, Ji Y. Nonparametric identification and estimation of sample selection models under symmetry[J]. Journal of Econometrics, 2017, 202(2): 148 – 160.

[3] Semykina A, Wooldridge J M. Binary response panel data models with sample selection and self-selection[J]. Journal of Applied Econometrics, 2018, 33(2): 179 – 197.

[4] Dong Y. Regression discontinuity designs with sample selection[J]. Journal of Business & Economic Statistics, 2019, 37(1): 171 – 186.

[5] Bia M, Huber M, Lafférs L. Double machine learning for sample selection models[J]. Journal of Business & Economic Statistics, 2020, 42(3): 958 – 969.

[6] Bartalotti O, Kédagni D, Possebom V. Identifying marginal treatment effects in the presence of sample selection[J]. Journal of Econometrics, 2023, 234(2): 565 – 584.

[7] 李燕凌. 农村公共产品供给侧结构性改革:模式选择与绩效提升——基于5省93个样本村调查的实证分析[J]. 管理世界, 2016, (11): 81 – 95.
Li Yanling. Rural public goods supply-side structural reform: Model selection and performance enhancement: An empirical analysis based on surveys of 93 sample villages in 5 provinces[J]. Journal of Management World, 2016, (11): 81 – 95. (in Chinese)

[8]孙晓华，郭　旭，王　昀. 政府补贴，所有权性质与企业研发决策[J]. 管理科学学报，2017，20(6)：18－31.
Sun Xiaohua，Guo Xu，Wang Yun. Government subsidy，ownership，and firms' R&D decisions[J]. Journal of Management Sciences in China，2017，20(6)：18－31.（in Chinese）

[9]赵西亮. 教育、户籍转换与城乡教育收益率差异[J]. 经济研究，2017，52(12)：164－178.
Zhao Xiliang. Education，migration and the urban-rural difference in returns to education[J]. Economic Research Journal，2017，52(12)：164－178.（in Chinese）

[10]高　鸣，宋洪远. 补贴减少了粮食生产效率损失吗？——基于动态资产贫困理论的分析[J]. 管理世界，2017，(9)：85－100.
Gao Ming，Song Hongyuan. Did subsidies reduce the efficiency losses in grain production? An analysis based on dynamic asset poverty theory[J]. Journal of Management World，2017，(9)：85－100.（in Chinese）

[11]杜鹏程，徐　舒，吴明琴. 劳动保护与农民工福利改善——基于新《劳动合同法》的视角[J]. 经济研究，2018，53(3)：64－78.
Du Pengcheng，Xu Shu，Wu Mingqin. Labor protection and welfare improvement for rural migrant workers：The new labor contract law perspective[J]. Economic Research Journal，2018，53(3)：64－78.（in Chinese）

[12]鄢伟波，安　磊. 中国女性劳动供给为何降低：来自流动人口的证据[J]. 世界经济，2021，44(12)：104－130.
Yan Weibo，An Lei. Why has the women's labor supply in China declined? New evidence from the migrants[J]. The Journal of World Economy，2021，44(12)：104－130.（in Chinese）

[13]Harrison G W，Lau M I，Yoo H I. Risk Attitudes，sample selection，and attrition in a longitudinal field experiment[J]. The Review of Economics and Statistics，2020，102(3)：552－568.

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

[15]Chernozhukov V，Fernández-Val I，Kowalski A E. Quantile regression with censoring andendogeneity[J]. Journal of Econometrics，2015，186(1)：201－221.

[16]Wüthrich K. A closed-form estimator for quantile treatment effects with endogeneity[J]. Journal of Econometrics，2019，210(2)：219－235.

[17]Powell D. Quantile treatment effects in the presence of covariates[J]. The Review of Economics and Statistics，2020，102(5)：994－1005.

[18]熊和平，刘京军，杨伊君，等. 中国股票市场存在特质波动率之谜吗？——基于分位数回归模型的实证分析[J]. 管理科学学报，2018，21(12)：37－53.
Xiong Heping，Liu Jingjun，Yang Yijun，et al. Is there idiosyncratic volatility puzzle in Chinese stock markets：A quantile regression analysis[J]. Journal of Management Sciences in China，2018，21(12)：37－53.（in Chinese）

[19]凌爱凡，谢林利. 特异性尾部风险，混合尾部风险与资产定价——来自我国 A 股市场的证据[J]. 管理科学学报，2019，22(8)：71－87.
Ling Aifan，Xie Linli. Idiosyncratic tail risk，hybrid tail risk and asset pricing：Evidence from China's A-share market[J]. Journal of Management Sciences in China，2019，22(8)：71－87.（in Chinese）

[20]杨子晖，陈雨恬，张平淼. 股票与外汇市场尾部风险的跨市场传染研究[J]. 管理科学学报，2020，23(8)：54－77.
Yang Zihui，Chen Yutian，Zhang Pingmiao. Cross-market contagion effect on tail risks between stock markets and exchange markets[J]. Journal of Management Sciences in China，2020，23(8)：54－77.（in Chinese）

[21]张京京，刘同山，钟　真. 网络营销提升了乡村旅游经营效益吗？——来自第三次全国农业普查北京市调查的证据[J]. 中国农村经济，2022，38(3)：67－83.
Zhang Jingjing，Liu Tongshan，Zhong Zhen. Does online marketing improve operational benefits of rural tourism? Evidence from the Beijing agricultural census data[J]. Chinese Rural Economy，2022，38(3)：67－83.（in Chinese）

[22]Arellano M，Bonhomme S. Quantile selection models with an application to understanding changes in wage inequality[J]. Econometrica，2017，85(1)：1－28.

[23]Robinson P M. Root-n-consistent semiparametric regression[J]. Econometrica，1988，56(4)：931－954.

[24]夏南新. 国际金融市场波动非线性因果性和溢出效应[J]. 管理科学学报, 2016, 19(3): 64 - 76.
Xia Nanxin. Nonlinear causality and spillover effect of volatility of international finance market[J]. Journal of Management Sciences in China, 2016, 19(3): 64 - 76. (in Chinese)

[25]周晶淼, 赵宇哲, 武春友, 等. 绿色增长下的导向性技术创新选择研究[J]. 管理科学学报, 2018, 21(10): 61 - 73.
Zhou Jingmiao, Zhao Yuzhe, Wu Chunyou, et al. Selection of directed technological innovation from the perspective of green growth[J]. Journal of Management Sciences in China, 2018, 21(10): 61 - 73. (in Chinese)

[26]殷 红, 张 龙, 叶祥松. 美联储货币政策对人民币外汇市场压力的冲击[J]. 管理科学学报, 2020, 23(11): 87 - 102.
Ying Hong, Zhang Long, Ye Xiangsong. The shocks of federal reserve's monetary policy on RMB exchange market pressure[J]. Journal of Management Sciences in China, 2020, 23(11): 87 - 102. (in Chinese)

[27]Du K, Yu Y, Wei C. Climaticimpact on China's residential electricity consumption: Does the income level matter? [J] China Economic Review, 2020, 63: 101520.

[28]黄 潇, 黄守军. 流动人口自雇的决定机制及收入差异[J]. 管理科学学报, 2021, 24(6): 57 - 75.
Huang Xiao, Huang Shoujun. The determination mechanism and income difference of migrants' self-employment in China [J]. Journal of Management Sciences in China, 2021, 24(6): 57 - 75. (in Chinese)

[29]赵 涛, 张 智, 梁上坤. 数字经济, 创业活跃度与高质量发展[J]. 管理世界, 2020, (10): 65 - 76.
Zhao Tao, Zhang Zhi, Liang Shangkun. Digital economy, entrepreneurial activity and high-quality development[J]. Journal of Management World, 2020, (10): 65 - 76. (in Chinese)

[30]Wolak F A. Measuring the benefits of greater spatial granularity in short-term pricing in wholesale electricity markets[J]. American Economic Review, 2011, 101(3): 247 - 252.

[31]Charnigo R, Feng L, Srinivasan C. Nonparametric and semiparametric compound estimation in multiple covariates[J]. Journal of Multivariate Analysis, 2015, 141: 179 - 196.

[32]Nelsen R B. An Introduction to Copulas[M]. New York: Springer, 1999.

[33]Joe H. Multivariate Models and Dependence Concepts[M]. London: Chapman and Hall, 1997.

[34]Chaudhuri P. Nonparametric estimates of regression quantiles and their local Bahadur representation[J]. The Annals of statistics, 1991, 19(2): 760 - 777.

[35]Chaudhuri P, Doksum K, Samarov A. On average derivative quantile regression[J]. The Annals of Statistics, 1997, 25 (2): 715 - 744.

[36]Horowitz J L. A smoothed maximum score estimator for the binary response model[J]. Econometrica, 1992, 60(3): 505 - 531.

[37]Chen X, Linton O, Van Keilegom I. Estimation of semiparametric models when the criterion function is not smooth[J]. Econometrica, 2003, 71(5): 1591 - 1608.

[38]Hall P, Horowitz J L. Bootstrap critical values for tests based on generalized-method of moments estimators[J]. Econometrica, 1996, 64(4): 891 - 916.

[39]Mammen E, Rothe C, Schienle M. Semiparametric estimation with generated covariates[J]. Econometric Theory, 2016, 32(5): 1140 - 1177.

[40]Lu X, White H. Testing for treatment dependence of effects of a continuous treatment[J]. Econometric Theory, 2015, 31 (5): 1016 - 1053.

[41]Imbens G, Newey W K. Identification and estimation of triangular simultaneous equations models without additivity[J]. Econometrica, 2009, 77(5): 1481 - 1512.

[42]Escanciano J C, Jacho-Chávez D, Lewbel A. Identification and estimation of semiparametric two step models[J]. Quantitative Economics, 2016, 7(2): 561 - 589.

[43]高梦滔, 姚 洋. 农户收入差距的微观基础: 物质资本还是人力资本? [J]. 经济研究, 2006, (12): 71 - 80.
Gao Mengtao, Yao Yang. Which is the main reason for income inequality in rural China: Physical assets or human capital?

[J]. Economic Research Journal, 2006, (12): 71 – 80. (in Chinese)

[44]程名望, Jin Yanhong, 盖庆恩, 等. 农村减贫：应该更关注教育还是健康？——基于收入增长和差距缩小双重视角的实证[J]. 经济研究, 2014, 49(11): 130 – 144.

Cheng Mingwang, Jin Yanhong, Gai Qingen, et al. Focusing on education or health improvement for anit-poverty in rural China: Evidence from national household panel data[J]. Economic Research Journal, 2014,49(11): 130 – 144. (in Chinese)

Estimation and application of a partial linear quantile selection model

ZHOU Ya-hong[1, 2, 3], *XIAO Xin-yue*[1], *JIN Ze-qun*[1*], *XIN Kai*[4], *JIANG Shuai-shuai*[1]

1. School of Economics, Shanghai University of Finance and Economics, Shanghai 200433, China;
2. Key Laboratory of Mathematical Economics (SUFE), Ministry of Education, Shanghai 200433, China;
3. Shanghai Institute for Mathematics and Interdisciplinary Sciences, Shanghai 200433, China;
4. School of Economics, Shanghai University, Shanghai 200436, China

Abstract: This paper examines the identification and estimation of partially linear quantile regression models in the presence of sample selection. To address the selection bias arising in quantile regression, we model the joint distribution of the unobserved disturbances in the outcome and selection equations – using, for instance, a copula specification – to obtain consistent estimates. We set out the assumptions required for identification and, building on the generalized method of moments for quantile regression, develop estimation procedures for both the quantile parameters (or functions) and the copula parameters. We establish the consistency and asymptotic normality of the proposed estimators. For inference, we implement a nonparametric bootstrap to compute standard errors and construct test statistics to assess the validity of the nonlinear specification. Using a control function framework, we further clarify the relationship between the endogenous linear quantile selection model and our proposed specification. Monte Carlo simulations indicate that the estimators perform well in finite samples. Finally, we apply the methodology to estimate the returns to female education, demonstrating the model's practical relevance.

Key words: quantile regression; sample selection; partial linear model; Copula function