اندیشه آماری، سال ۱۴۰۳، دوره ۲۹، شماره ۱، ص ۲۵ – ۱۷ https://doi.org/10.22034/jr$_{i}ss$.2024.729436

رگرسیون چندکی متغیر ابزاری با رویکرد حداقل فاصله و کاربرد آن در تحلیل داده‌های هزینه و درآمد خانوار ایرانی

سید احسان حسین زاده[۱] و موسی گلعلی زاده[۲]*

[۱,۲] دانشکده علوم ریاضی، دانشگاه تربیت مدرس، تهران، ایران

تاریخ دریافت: ۱۴۰۴/۰۲/۱۱
تاریخ پذیرش: ۱۴۰۴/۰۶/۳۰

چکیده

یکی از رایج‌ترین مدل‌های آماری برای بررسی روابط بین متغیرهای یک مطالعه علمی مدل‌های رگرسیون خطی مبتنی بر میانگین متغیر پاسخ است که بر پایه پذیره‌هایی مانند نرمال بودن توزیع خطا و ناهمبسته بودن متغیرهای تبیینی و خطای مدل بنا می‌شود. در صورت انحراف از پذیره نرمال بودن مولفه خطا، مدل رگرسیون چندکی به‌دلیل نیرومندبودن در برابر داده‌های دورافتاده، جایگزین مناسبی است. همچنین، در مواجهه با مشکل درون‌زایی ناشی از برقرار نبودن پذیره ناهمبسته بودن متغیر تبیینی و خطای مدل، استفاده از متغیرهای ابزاری ضروری است. از طرفی، روش حداقل فاصله که با کمینه کردن فاصله بین توزیع‌های مشاهده شده و برآورد متناظر آن پارامترها را برآورد می‌کند، ابزاری قدرتمند برای برازش مدل‌ها است. مقاله حاضر، روش ترکیبی رگرسیون چندکی متغیر ابزاری با حداقل فاصله را تشریح و جزئیات نظری مرتبط با آن را توضیح می‌دهد. همچنین، با استفاده از داده‌های هزینه و درآمد خانوارهای ایرانی در سال ۱۴۰۲، این روش با روش مرسوم رگرسیون چندکی مقایسه و توانایی آن در کنترل درون‌زایی و آشکارسازی ناهمگنی تشریح می‌شود.

واژه‌های کلیدی: رگرسیون چندکی، متغیر درون‌زا، متغیر ابزاری، حداقل فاصله، داده‌های هزینه و درآمد.

رده‌بندی ریاضی: 62J04، 62J08.

۱ مقدمه

در مطالعات علوم اجتماعی و اقتصادی، تحلیل روابط میان متغیرهای وابسته و تبیینی از جایگاه ویژه‌ای برخوردار است. در این میان، مدل‌های رگرسیونی ابزار مناسبی برای مدل‌سازی این روابط فراهم کرده‌اند. با این حال، در بسیاری از کاربردهای واقعی، برقراری برخی پذیره‌های کلیدی در مدل‌های کلاسیک رگرسیون، نظیر استقلال خطاها و ناهم‌بستگی میان متغیرهای تبیینی و خطای مدل، با چالش‌هایی همراه است. نقض این پذیره‌ها، به‌ویژه در قالب پدیده‌ی درون‌زایی، می‌تواند به برآوردهایی اریب و ناسازگار منجر شود kivi۹۵، imbang۹۴.

در مواجهه با چنین شرایطی، بهره‌گیری از متغیرهای ابزاری به‌عنوان یکی از رویکردهای شناخته‌شده در ادبیات اقتصادسنجی مطرح شده است. این متغیرها، که با متغیرهای تبیینی هم‌بسته ولی با خطای مدل ناهم‌بسته‌اند، امکان برآوردهای سازگار را فراهم می‌آورند. در همین راستا، روش‌های مختلفی برای استفاده از متغیرهای ابزاری توسعه یافته که از جمله‌ی آن‌ها می‌توان به روش دو مرحله‌ای کمترین توان‌های دوم اشاره کرد andrub۴۹. برای نمونه، کاراییِ رگرسیون چندکی برای آشکارسازی تفاوت‌های بین دهک‌ها را [۲] در بازار کار و [۳] در هزینه انرژی نشان داده‌اند. همچنین [۱] با به‌کارگیری مدل آمیخته رگرسیون

*نویسنده مسئول مقاله، golalizadeh@modares.ac.ir

چندکی بر داده‌های هزینه و درآمد خانوار ایرانی نشان دادند که ناهمگنی استانی و وجود داده‌های دورافتاده بر برآورد مخارج خانوار مؤثر است. [۱۴] اعتقاد دارد که رگرسیون چندکی با تمرکز بر چندک‌های مختلف توزیع متغیر وابسته، توانایی بیشتری در بازنمایی ناهمگنی موجود در داده‌ها نسبت به رگرسیون سنتی مبتنی بر میانگین دارد. به نظر می‌رسد در صورتی‌که این روش با چارچوب متغیرهای ابزاری تلفیق شود، امکان تحلیل دقیق‌تر اثر متغیرهای درون‌زا را در سطوح مختلف توزیع متغیر پاسخ فراهم می‌آورد. بر اساس این دیدگاه، در سال‌های اخیر، برآوردگر حداقل فاصله در رگرسیون چندکی متغیر ابزاری [1] (MD-IVQR)، به‌عنوان پاسخی به چالش‌های محاسباتی روش‌های کلاسیک در داده‌های پانلی با اثرات ثابت معرفی شده است. برای مطالعه درباره این موضوع می‌توان به [۷]، [۵] و [۱۸] مراجعه کرد. در مقالات مذکور، این رویکرد با بهره‌گیری از ساختار وزنی بر مبنای ماتریس‌های کوواریانس، برآوردی کارآمد و محاسبه‌پذیر ارائه می‌دهد و ادعا می‌شود که از نظر دقت و سهولت پیاده‌سازی، مزایای قابل‌توجهی دارد.

هدف اصلی مقاله حاضر تشریح رویکرد اخیر برای نمایش کارکرد مدل MD-IVQR است. برای این منظور پس از معرفی چارچوب نظری مدل MD-IVQR، عملکرد آن در یک مطالعه موردی بر اساس داده‌های هزینه و درآمد خانوارهای شهری ایران در سال ۱۴۰۲ بررسی خواهد شد. لذا، چیدمان مقاله نیز بر اساس این دو عنوان به ترتیب در بخش دوم و سوم تدوین شده است. بحث و نتیجه‌گیری نیز در انتهای مقاله خواهد آمد.

۲ رگرسیون چندکی متغیر ابزاری با حداقل فاصله

در این بخش، روشی برای برآورد مدل رگرسیون چندکی در حضور متغیرهای درون‌زا معرفی می‌شود که با عنوان MD-IVQR شناخته می‌شود. هدف اصلی این برآوردگر، کاهش اریبی ناشی از درون‌زایی متغیرها و ساده‌سازی محاسباتی در مدل‌های پانلی با اثرات ثابت است، به‌ویژه در شرایطی که تعداد نمونه در سطح پانل، n، زیاد باشد. از منظر ساختاربندی مدل، روش پیشنهادی با توسعه‌ی چارچوب IVQR کلاسیک صورت می‌گیرد؛ به این معنا که به‌جای برآورد مستقیم ضرایب مدل با حل یک مسئله پیچیده‌ی بهینه‌سازی برای کل داده‌ها، ابتدا داده‌ها بر اساس متغیرهای برون‌زا، X، به زیرگروه‌هایی همگن یا «سلول‌هایی» تقسیم می‌شوند. در هر سلول، برآوردگر IVQR به‌صورت جداگانه اعمال شده و تخمین‌های موضعی از پارامترهای ساختاری به‌دست می‌آید. سپس، با استفاده از رویکرد حداقل فاصله، این تخمین‌های سلولی به‌صورت میانگین وزنی ترکیب می‌شوند. بیان این نکته ضروری است که وزن‌دهی مبتنی بر وارون ماتریس کوواریانس هر سلول صورت می‌گیرد، که موجب افزایش کارایی آماری و پایداری عددی برآورد نهایی می‌شود. در نتیجه، برآوردگر MD-IVQR نه‌تنها از نظر محاسباتی مقیاس‌پذیرتر از روش IVQR مستقیم است، بلکه قابلیت شناسایی و مدل‌سازی ناهمگنی‌های موجود در پاسخ را در بخش‌های مختلف توزیع (مثلا چندک [2]) حفظ می‌کند. از منظر نظری نیز، این برآوردگر دارای خاصیت سازگاری است و در چارچوب تخمین‌های مبتنی بر روش حداقل فاصله، تحت پذیره‌های معین، دارای توزیع حدی نرمال خواهد بود. در ادامه ساختار نظری مدل مورد بحث تشریح می‌شود.

معمولا، مدل رگرسیون چندکی در حضور متغیر درون‌زا به صورت

$$y_{it} = D_{it}^T\alpha + X_{it}^T\beta + \eta_i + U_{it}, \quad i = 1, ..., n, \quad t = 1, ...k \qquad (۱)$$

نوشته می‌شود که در آن، n تعداد مشاهدات، k تعداد دفعات مشاهده، y_{it} متغیر پاسخ برای واحد i، D_{it} برداری از متغیرهای درون‌زا که با عوامل مشاهده نشده‌ی اثرگذار بر متغیر وابسته همبستگی دارد، X_i' برداری از متغیرهای برون‌زا، α و β بردار پارامترهای مورد نظر مدل، η_i اثر فردی (اثرات ثابت) و U_i جمله خطا است.

مدل (۱) برای سهولت در تحلیل‌های بعدی به‌صورت ماتریسی

$$\mathbf{y} = \mathbf{D}\alpha + \mathbf{X}\beta + \mathbf{Z}\eta + \mathbf{U}, \qquad (۲)$$

بازنویسی می‌شود که در آن، $\mathbf{y} = (y_i)$ یک ماتریس $n \times 1$-بعدی، $\mathbf{D} = (\mathbf{D}_i)$ یک ماتریس $n \times \dim(\alpha)$ -بعدی، $\mathbf{X} = (\mathbf{X}_i)$ یک ماتریس $n \times \dim(\beta)$ -بعدی، $\mathbf{Z} = I_n$ ، که I_n ماتریس همانی $n \times n$، $\eta = (\eta_1, ..., \eta_n)'$ برداری $n \times 1$ از اثرات ثابت فردی برای هر واحد مقطعی $i = (1, ..., n)$ است که به‌منظور مدل‌سازی ناهمگنی غیرمشاهده‌پذیر میان واحدها وارد مدل شده است و $\mathbf{U} = (U_i)$ یک ماتریس $n \times 1$ حاوی جملات خطا است. توجه شود که Z یک ماتریس تلاقی (وقوع) است که n مشاهده متمایز را در نمونه مشخص می‌کند.

فرض کنید چندک τ-ام جملهٔ خطا (U_i) برابر با صفر است. در نتیجه، مدل رگرسیون چندکی برای چندک τ-ام به‌صورت

$$Q_{y_i}(\tau|\mathbf{D}_i, \mathbf{X}_i, \mathbf{Z}_i) = \mathbf{D}_i'\alpha(\tau) + \mathbf{X}_i'\beta(\tau) + \mathbf{Z}_i'\eta(\tau), \qquad (۳)$$

[1] Minimum Distance Instrumental Variable Quantile Regression
[2] Quantile

خواهد بود که در آن Z_i، ستون i–ام ماتریس Z است که به‌عنوان یک بردار تکرارشونده عمل می‌کند تا اثر فردی، $\eta_i(\tau)$، را برای آن واحد در معادله رگرسیون چندکی لحاظ کند.

توجه شود که در رابطه (۲) متغیر درون‌زا، D_i، با جملهٔ خطا، U_i، همبستگی دارد. این همبستگی منجر به برآورد اریب[3] پارامترهای مدل می‌شود. برای غلبه بر مشکل سوءگیری ناشی از درون‌زایی، پژوهش‌گران از رویکرد متغیرهای ابزاری، به‌عنوان یک روش قدرتمند و کارآمد استفاده می‌کنند harlam۰۹.

اکنون به بررسی برآوردگر رگرسیون متغیر ابزاری با اثرات ثابت برای مدل رگرسیون چندکی با درون‌زایی پرداخته می‌شود که توسط [۹] پیشنهاد شده است. آنها نشان دادند که برآوردگر رگرسیون چندکی ابزاری با اثرات ثابت[4] (IVFEQR) برای پارامترهای مربوط به متغیرهای برون‌زا، β، از حل مسئله بهینه‌سازی

$$\min_{\eta,\beta,\gamma}\sum_{i=1}^{n}\rho_\tau(y_i-\mathbf{D}_i'\alpha(\tau)-\mathbf{X}_i'\beta(\tau)-\mathbf{Z}_i'\eta(\tau)-\gamma(\tau)), \quad (۴)$$

به دست می‌آید که در آن γ بردار ضرایب مربوط به متغیرهای ابزاری و $\rho_\tau(u)=u(\tau-I(u<0))$ تابع چک[5] برای چندک τ–ام و $I(.)$ تابع نشانگر[6] است. به طور خلاصه، برآوردگر مورد نظر به صورت

$$\hat{\Theta}=\arg\min_{\alpha\in A}\{[\hat{\gamma}(\alpha)]'\mathbf{A}\hat{\gamma}(\alpha)\},$$

تعریف می‌شود که A مجموعه فشرده پارامترها، $\mathbf{A}$ ماتریسی مثبت معین است و $\Theta=(\eta(\alpha),\beta(\alpha),\gamma(\alpha))$.

به‌دلیل وجود درون‌زایی، امکان حذف اثرات ثابت با تبدیل متغیرهای $\mathbf{X}_i$، $\mathbf{D}_i$ و y_i به انحراف از میانگین‌های فردی، به‌منظور کاهش تعداد پارامترهای برآوردی وجود ندارد. به‌عبارت دیگر، برای رگرسیون چندکی متغیرهای ابزاری، ناگزیر باید کل ساختار مدل را درگیر فرآیند برآورد کرد. این امر باعث می‌شود با افزایش تعداد پانل‌ها، n، و بُعد کل، p، شامل بعد متغیرهای ابزاری، $\dim(\omega)$، و بُعد پارامترها، $\dim(\alpha)$، تعداد پارامترهای نیازمند برآورد خیلی زیاد شود. [۱۴] نشان داد که با داشتن مجموعه‌ای مناسب از مقادیر $\{\alpha_j, j=1,,...,J\}$، برآوردگرهای به‌دست آمده از طریق فرمول (۴) اساساً با برآوردگرهای رگرسیون چندکی با اثرات ثابت[7] (FE–QR) مطابقت دارند. به‌عبارت دیگر، برای به‌دست آوردن برآوردگرهای IVFEQR ، نیاز به محاسبه مجموعه‌ای از برآوردگرهای FE-QR است. با این‌حال، برآوردگر FE-QR شامل بهینه‌سازی با تعداد زیادی از پارامترهای برآورد شونده است که این امر مسئله را از نظر محاسباتی سنگین و اغلب غیرقابل اجرا می‌کند.

با انگیزه رفع چالش‌های عملیاتی پیاده‌سازی برآوردگر FE-QR، [۷] برآوردگر MD-QR را برای اثرات ثابت ارائه کردند که پیاده‌سازی ساده‌تری دارد. آنها برآوردگر جدیدی را برای مدل (۱) معرفی کردند تا اریبی ناشی از درون‌زایی را کاهش داده و محاسبات پیچیده را ساده‌سازی نماید. برآوردگر پیشنهادی، ترکیبی وزنی از برآوردگرهای رگرسیون چندکی متغیر ابزاری[8] (IVQR) است که می‌توان آن‌ها را از طریق رگرسیون‌های چندکی با متغیرهای ابزاری معمولی به‌دست آورد. رویکرد برآورد پیشنهادی از نظر محاسباتی راحت و در بسیاری از کاربردهای رایج، ساده‌سازی شده است. این مزیت در مقایسه با روش‌های سنتی مانند IVFEQR که با چالش‌های محاسباتی سنگین روبرو هستند، قابل توجه است. با الهام از روش [۷] که برآوردگر کارآمد کمترین فاصله را پیشنهاد دادند، برآوردگر دیگری با هدف ساده‌سازی محاسبات و غلبه بر مشکلات روش‌های سنتی، طراحی شده است.

اکنون فرض کنید بردار پارامترهای مربوط به متغیرهای درون‌زا و برون‌زا با $\theta=(\alpha',\beta')'$ نشان داده شود. برآوردگر حداقل فاصله در چارچوب رگرسیون چندکی متغیر ابزاری، که با $\hat{\theta}_{\text{MD-IVQR}}$ نمایش داده می‌شود، به‌صورت

$$\hat{\theta}_{\text{MD-IVQR}}=\left(\sum_{i=1}^{n}V_i^{-1}\right)^{-1}\sum_{i=1}^{n}V_i^{-1}\hat{\theta}_i, \quad (۵)$$

نوشته می‌شود که در آن، $\hat{\theta}_i=(\hat{\alpha}_i',\hat{\beta}_i')'$ برآوردگر IVQR برای واحد i–ام و V_i ماتریس کوواریانس متناظر با $\hat{\theta}_i$ است. این برآوردگر در واقع یک میانگین وزنی از برآوردگرهای تکی IVQR برای هر واحد مقطعی است، به‌طوری‌که وزن هر واحد متناسب با دقت آماری آن تعیین می‌شود؛ یعنی، ماتریس وزن برابر وارون ماتریس کوواریانس آن واحد است. این ساختار اجازه می‌دهد که ناهمگنی آماری بین واحدها در فرآیند برآورد لحاظ شود و از آنجا که محاسبات آن به‌صورت جزء به جزء انجام می‌گیرد، از نظر محاسباتی نیز کارآمدتر از روش‌های کلاسیک مانند IVFEQR است.

نکته حائز اهمیت این است که دسترسی مستقیم به ماتریس‌های

[3] Biased Estimation
[4] Instrumental Variable Fixed Effect Quantile Regression
[5] Check Function
[6] Indicator Function
[7] Fixed Effect Quantile Regression
[8] Instrumental Variable Quantile Regression

کوواریانس V_i در (۵) در عمل میسر نیست. برای حل این مشکل از برآوردگرهای سازگار $\hat{V}_i$ به‌جای V_i استفاده می‌شود. بنابراین، برآوردگری که در عمل مفید فایده است به صورت

$$\hat{\theta}_{\text{MD-IVQR}} = \left(\sum_{i=1}^{n} \hat{V}_i^{-1}\right)^{-1} \sum_{i=1}^{n} \hat{V}_i^{-1}\hat{\theta}_i \qquad (۶)$$

خواهد بود. این برآوردگر را می‌توان به‌صورت صریح از حل یک مسئله بهینه‌سازی مربعی استخراج کرد. برای انجام این امر، فرض کنید:

$$\hat{\xi} = (\hat{\theta}_1', \hat{\theta}_2', \ldots, \hat{\theta}_n')' \in \mathbb{R}^{n\times d}, \quad \text{که}, \quad d = \dim(\theta)$$

با این شرط که

$$\hat{\xi} - 1_n \otimes \theta = 0,$$

که در آن 1_n بردار ستونی شامل n عدد یک است و $\otimes$ نشان‌دهنده ضرب کرونکر است. این رابطه بیانگر آن است که باید همه برآوردهای انفرادی ضرایب برابر یک مقدار مشترک θ شوند. از آنجا که این قید در عمل به‌طور کامل قابل تحقق نیست، محققین به‌جای آن، حل مسئله کمترین مربعات با وزن بهینه به‌صورت

$$\min_{\theta} \left(\hat{\xi} - 1_n \otimes \theta\right)' U \left(\hat{\xi} - 1_n \otimes \theta\right),$$

که U ماتریس وزن معین مثبت است، را برای برآورده شدن قید موردنظر پیشنهاد دادند. در شرایطی که پذیره استقلال بین واحدها برقرار باشد، می‌توان U را به‌صورت قطری یعنی $U = \text{diag}(U_1, U_2, \ldots, U_n)$ تعریف کرد. در نتیجه، صورت نهایی مسئله بهینه‌سازی به شکل

$$\min_{\theta} \sum_{i=1}^{n} \left(\hat{\theta}_i - \theta\right)' U_i \left(\hat{\theta}_i - \theta\right), \qquad (۷)$$

قابل بازنویسی است. در این حالت، جواب این مسئله، به صورت صریح، برابر است با:

$$\hat{\theta} = \left(\sum_{i=1}^{n} U_i\right)^{-1} \sum_{i=1}^{n} U_i\hat{\theta}_i. \qquad (۸)$$

این فرمول، ساختار برآوردگر $\hat{\theta}_{\text{MD-IVQR}}$ در (۶) را با انتخاب $U_i = \hat{V}_i^{-1}$ توجیه خواهد کرد. بنابراین، می‌توان نتیجه گرفت که برآوردگر پیشنهادی نه‌تنها از نظر تئوری با روش کمترین فاصله هم‌خوان است، بلکه از منظر محاسباتی نیز اجرای ساده‌تری نسبت به روش‌های پانلی کلاسیک دارد.

در این قسمت، بیان چند نکته مهم محاسباتی ضروری است. برآوردگر سازگار $\hat{V}_i$ معمولاً با استفاده از روش‌های نمونه‌گیری مجدد مانند بوت‌استرپ یا از طریق تقریب ماتریس اطلاعات فیشر برای برآوردگر $\hat{\theta}_i$ به دست می‌آید. با در نظر گرفتن الگوریتم بوت‌استرپ، نمونه‌های تصادفی متعددی از داده‌ها برای هر واحد i تولید می‌شوند و واریانس برآوردهای حاصل از آن نمونه‌ها به عنوان تقریب $\hat{V}_i$ مورد استفاده قرار می‌گیرد. این رویکرد از نظر محاسباتی انعطاف‌پذیر است و اتفاقا کارآمدی آن در مطالعات متعدد پژوهشی از جمله [۷] مورد بررسی قرارگرفت.

در عمل، اگر از الگوریتم بهینه‌سازی خطی برای محاسبه هر برآوردگر $\hat{\theta}_i$، استفاده شود، می‌توان ماتریس گرادیان تابع زیان و مشتقات درجه دوم را برای محاسبه برآورد $\hat{V}_i$ مورد استفاده قرار داد. در هر صورت، هدف از استفاده از $\hat{V}_i^{-1}$ به‌عنوان وزن این است که به برآوردگرهایی که نوسان بیشتری دارند (واریانس بالاتر)، وزن کمتری داده شود و بالعکس، تا بدین ترتیب برآوردگر نهایی پایدارتر، دقیق‌تر و از نظر آماری کاراتر باشد. برای مطالعه بیشتر در این زمینه به [۱۶]، [۱۷] و [۱۰] مراجعه شود.

سازگاری آماری $\hat{\theta}_{MD-IVQR}$ در صورتی برقرار است که پذیره‌هایی نظیر استقلال بین واحدها، کران‌دار بودن متغیرها و منظم‌بودن ماتریس‌های کوواریانس برقرار باشند. واضح است که با افزایش تعداد واحدهای پانل و ثابت‌ماندن ابعاد پارامترها، برآورد نهایی همگرا به مقدار واقعی می‌شود. این پذیره‌ها در متون کلاسیک از جمله [۱۵]، [۹] و [۷] مورد بحث قرار گرفتند.

توجه شود زمانی‌که U_i مقادیر مختلفی دارد، می‌توان تصویری از براوردگرهای کمینه فاصله که با M نشان داده می‌شود را بدست آورد. همان‌طور که توسط [۷] اشاره شد، وزن‌های بهینه در میان تمام برآوردگرها در M، وارون ماتریس کواریانس مجموعه چندک‌های رگرسیون است. این بدان معناست که برآوردگر تعریف شده در (۶)، کارآمدترین برآوردگر در M است. به عبارتی دیگر، $\hat{\theta}_{MD-IVQR}$ دارای کوچک‌ترین ماتریس کوواریانس مجانبی در میان تمام برآوردگرها در M است.

لازم به اشاره است که برای حصول برآوردگر سازگار و کارا از پارامترهای مدل، دو مرحله اصلی در چارچوب MD-IVQR دنبال می‌شود. در مرحله نخست، برآوردگرهای شیب رگرسیون چندکی برای هر بخش از داده‌ها (مثلاً برحسب چندک‌های مختلف یا خوشه‌های مقطعی) به‌صورت جداگانه محاسبه می‌شوند. این برآوردها به‌طور مستقل و با استفاده از روش‌های کلاسیک حداقل قدر مطلق انحرافات یا الگوریتم‌های مبتنی بر تابع زیان چندکی به دست می‌آیند. سپس، در مرحله دوم این برآوردهای اولیه به کمک یک ماتریس وزن‌دهی مناسب با یکدیگر ترکیب می‌شوند. وزن هر برآورد با استفاده از وارون ماتریس کوواریانس آن بخش تعیین می‌شود که نقش اساسی در بهبود دقت و

سازگاری نهایی ایفا می‌کند. برآوردگر نهایی که ترکیبی وزنی از نتایج چند بخش است، تحت شرایط منظم‌بودن مناسب و برقراری پذیره‌های اساسی همگرایی، از نظر آماری سازگار و به صورت مجانبی نرمال است. همچنین، برای محاسبه ماتریس کوواریانس هر برآوردگر تکی، می‌توان از روش‌های بوت‌استرپ تصادفی یا برآورد نااریب ماتریس اطلاعات استفاده کرد که توسط محققینی مانند [۷] و [۶] مدنظر قرار گرفت.

بنابراین می‌توان گفت که برآوردگر MD-IVQR نه تنها راهکاری محاسباتی ساده و قابل‌پیاده‌سازی برای مدل‌های اثر تصادفی با متغیرهای درون‌زا فراهم می‌آورد، بلکه از منظر نظری نیز دارای خواص مطلوبی مانند سازگاری آماری و کارایی مجانبی است که توجیه مناسبی را برای استفاده از آن در تحلیل‌های کاربردی در علوم اجتماعی، اقتصاد و داده‌های پیمایشی فراهم می‌سازد.

۳ تحلیل داده‌های هزینه و درآمد خانوار شهری

داده‌های این پژوهش مربوط به طرح آمارگیری هزینه و درآمد خانوارهای شهری است که در سال ۱۴۰۲ توسط مرکز آمار ایران اجرا شد. از آن بین، اطلاعات متنوع مرتبط با هزینه و درآمد ۸۶۲۵ خانوار ساکن مناطق شهری تهران مدنظر این تحقیق قرار گرفت. با توجه به بار محاسباتی بالای برآوردگر MD-IVQR، که به حل مسائل بهینه‌سازی پیچیده و انجام بوت‌استرپ نیاز دارد، اجرای الگوریتم روی کل نمونه با زمان اجرای بسیار طولانی و در مواردی با مشکلات همگرایی همراه می‌شد. از این رو، برای تحلیل‌های تجربی این بخش یک نمونه‌گیری تصادفی ساده بدون جایگزینی انجام شد و ۶۰۰ مشاهده انتخاب گردید. اندازه‌ی نمونه به‌گونه‌ای تعیین شد که ضمن کاستن از هزینه‌ی محاسباتی، شکل کلی توزیع متغیرها و شاخص‌های گرایش مرکزی و پراکندگی آنها طوری باشند که نتایج قابل تعمیم به کل جامعه‌ی در دسترس باشد.

طرح آمارگیری هزینه و درآمد خانوارهای شهری اجرا شده توسط مرکز آمار ایران در پایگاه داده مرکز وجود دارد. این پایگاه علاوه بر متغیرهای اقتصادی اصلی (هزینه و درآمد)، شامل مجموعه‌ای از شاخص‌های جمعیتی و ویژگی‌های مسکونی نیز است. بنابراین، امکان بررسی رابطه‌ی هزینه و درآمد با اتکا بر متغیرهای ابزاری و نیز تحلیل تفاوت‌ها در بخش‌های مختلف توزیع پاسخ وجود دارد. جدول ؟؟ یک شمای کلی از متغیرهای مورد استفاده در این مقاله را ارائه می‌کند. از منظر سیاست‌گذاری، خروجی‌های این مطالعه می‌تواند برای طراحی مداخلات هدفمند به منظور بهبود رفاه خانوارهای شهری و کاهش نابرابری مفید فایده باشد.

لازم به ذکر است برای تفسیر بهتر نتایج، متغیرهای پولی به میلیون تومان تغییر مقیاس داده شده است. البته انجام این عمل فقط مقیاس ضرایب را تغییر دادند و تاثیری بر استنباط نهایی ارائه شده در ادامه مقاله نداشتند. با توجه به مطالب ارائه شده در بخش نظری، در ادامه تحلیل حاضر، هزینه‌ی خانوار، Hazineh، به عنوان متغیر پاسخ، $\mathbf{Y}$، و درآمد، Daramad، به‌عنوان متغیر درون‌زا، $\mathbf{X}_1$، و سن سرپرست خانوار، Age، به‌عنوان متغیر برون‌زا، $\mathbf{X}_2$، در نظر گرفته شدند. لذا، مدل MD-IVQR پایه به‌صورت

$$Q_\tau(\mathbf{Y} \mid \mathbf{X}_1, \mathbf{X}_2) = \beta_0(\tau) + \beta_1(\tau)\mathbf{X}_1 + \beta_2(\tau)\mathbf{X}_2, \qquad (۹)$$

تعریف می‌شود که در آن $\beta_0(\tau)$ عرض از مبدأ، $\beta_1(\tau)$ ضریب درآمد و $\beta_2(\tau)$ ضریب سن به ازای چندک τام توزیع متغیر پاسخ هستند. به‌سبب در نظر گرفتن متغیر درآمد به عنوان متغیر درون‌زا، معادله‌ی مرحله‌ی اول به شکل

$$\mathbf{X}_1 = \alpha_0 + \alpha_1\mathbf{Z}_1 + \alpha_2\mathbf{Z}_2 + v, \qquad (۱۰)$$

نوشته می‌شود که در آن $\mathbf{Z}_1$ بُعد خانوار، FZ، و $\mathbf{Z}_2$ سطح زیربنای محل سکونت، GL، هستند و v معرف مولفه اغتشاش مرحله‌ی اول است. در این چارچوب، پارامترهای α_0، α_1 و α_2 به‌ترتیب معرف عرض از مبدأ و ضرایب متغیرهای FZ و GL هستند. این مشخصه‌سازی امکان شناسایی اثر علّی درآمد بر هزینه را در کل توزیع پاسخ، نه فقط حول میانگین هزینه خانوارها، فراهم می‌کند. با مشخص سازی تشریح شده، برازش مدل MD-IVQR برای داده مورد بحث و چندک‌های مدنظر این مقاله (چندک‌های ۰٫۲۵، ۰٫۵۰ و ۰٫۷۵ام)، منجر به برآورد پارامترها، خطای برآوردها و Pـ مقدارهای متناظر شد که در جدول ؟؟ آمده است. لازم به ذکر است متغیرهایی که واحد اندازه‌گیری آن‌ها پول بودند، پس از تبدیل واحد پول به میلیون تومان، بر اساس روش استانداردسازی مبتنی بر Z-Score به صورت استاندارد در آمدند. بقیه متغیرهای کمی به کمک آماره Z-Score به متغیرهای استاندارد تغییر یافتند. سعی شد تفسیرهای مرتبط با ضرایب بر اساس مقیاس اولیه متغیرها صورت گیرند.

بر اساس مقادیر جدول ؟؟، برآورد عرض از مبدأ، $\hat{\beta}_0$، به ازای چارک اول و دوم معنادار نیست. مقادیر P-value متناظر با آزمون فرضیه هر یک به ترتیب برابر اعداد ۰٫۸۲ و۰٫۸۶ است. اما، برآورد

در چارک سوم به‌طور معناداری مثبت است چرا که در اجرای آزمون فرضیه مربوطه مقدار P-value= ۰٫۰۰۰۲ بدست آمد. در مقابل، در برآورد ضریب درآمد ($\hat{\beta}_1$) به ازای چارک اول و سوم مقادیر P-value به ترتیب ۲٫۰۶e-۹ و ۱e-۳۶ حاصل شد که نشان از معناداری مثبت این برآورد دارد. اما این برآورد به ازای میانه (چندک ۰٫۵۰) معنادار نیست چون P-value = ۰٫۲۳ حاصل شد. این نشان می‌دهد که اثر درآمد بر هزینه در بخش‌های پایین و بالای توزیع هزینه قابل توجه است، اما در حول و حوش میانه توزیع هزینه، آن تأثیر آماری معناداری ندارد. با توجه به برآوردهای حاصل و از منظر کاربردی، می‌توان اظهار داشت که به شرط ثابت در نظر گرفتن عوامل دیگر، به‌ازای هر یک میلیون افزایش درآمد، مخارج خانوار با مدنظر قرار دادن چارک‌های اول، دوم و سوم توزیع هزینه به‌ترتیب حدود ۰٫۳۹، ۰٫۵ و ۰٫۴۳ میلیون تومان افزایش پیدا می‌کند. به زبانی دیگر، به‌ازای ۱۰ میلیون افزایش درآمد، شخص انتظار افزایش ۳٫۹، ۵ و ۴٫۳ میلیون تومان در مخارج آن خانوار به ترتیب در چارک اول، دوم و سوم دارد. البته، لازم است اظهار نظر به ازای میانه هزینه با احتیاط صورت گیرد چرا که برآورد متناظر برای ضریب درآمد از نظر آماری معنادار نشد.

اکنون راجع به آنچه برای متغیر سن در مدل MD-IVQR اتفاق افتاد بحث می‌شود. برآورد ضریب سن ($\hat{\beta}_2$) تنها به ازای مدل مبتنی بر میانه (چندک ۰٫۵۰) توزیع هزینه معنادار است (P-value = ۰٫۰۳)، در حالی که به ازای چارک اول و سوم توزیع هزینه این برآورد معنادار نیست چرا که P-valueهای متناظر برابر ۰٫۴۳ و ۰٫۰۷ بدست آمد. این نشان می‌دهد که افزایش سن سرپرست خانوار فقط در میانه توزیع هزینه منجر به افزایش مخارج خانوار مربوطه خواهد شد. با توجه به برآوردهای به دست آمده و از منظر اقتصادی و اگر خود را محدود به چارک میانی توزیع هزینه کنیم، به‌ازای هر ده سال افزایش سن، مخارج آن خانوار ۱۰٫۳ میلیون تومان افزایش می‌یابد به شرطی که شخص عوامل دیگر را در نظر نگیرد.

جمع بندی نهایی تحلیل داده‌های هزینه و درآمد خانوار شهری بر اساس مدل MD-IVQR این است که اثر درآمد سرپرست خانوار در بخش‌های پایین و بالای توزیع هزینه خانوار قابل توجه است، در حالی که اثر سن سرپرست خانوار تنها در میانه توزیع هزینه خانوار تا حدی معنادار است.

بخشی از علاقه‌ی این پژوهش بررسی رابطه‌ی سه‌گانه‌ی درآمد، سن و هزینه در تحلیل داده‌های هزینه و درآمد خانوار شهری بر اساس مدل MD-IVQR است. برای این منظور، آنچه از برازش این مدل حاصل شد به صورت شکل ؟؟ نشان داده شد. در این شکل، نقاط قرمز معرف مشاهدات این مجموعه داده بوده و سه سطح رنگی، صفحات برازش‌شده‌ی مدل MD-IVQR به ازای چارک‌های اول تا سوم را نشان می‌دهند. همانطور که ملاحظه می‌شود شیب صفحات در راستای محور درآمد مثبت است و با افزایش چندک توزیع هزینه این شیب تندتر می‌شود. همانگونه که بحث شد این به معنی آن است که در بخش‌های بالاتر توزیع هزینه، تغییرات درآمد اثر بیشتری بر مخارج خانوارها دارد. به علاوه، مشاهده می‌شود که شیب صفحات در امتداد محور سن کم است. به طور دقیق، شیب در چندک پایین تقریباً تخت (حتی اندکی رو به پایین) و در میانه و بالای توزیع ملایم و رو به بالا است. فاصله‌ی میان صفحات هم نشان می‌دهد که رابطه‌ی هزینه با متغیرهای تبیینی در سراسر توزیع یکنواخت نیست و با جابه‌جایی در چندک‌ها تغییر می‌کند. بنابراین می‌توان گفت مدل MD-IVQR با کنترل درون‌زایی و نمایش ناهمگنی اثرها در طول توزیع هزینه‌ها، تصویری مناسب‌تری از تاثیر درآمد (افزایشی در چندک‌های بالاتر) و سن (ضعیف در چندک‌های پایین و بالای توزیع هزینه و قوی در میانه توزیع) ارائه می‌دهد و برای تحلیل‌های اقتصادی با داده‌های ناهمگون ابزاری مناسب است.

در مجموع، آنچه که می‌توان از نمودار سه‌بعدی حاضر برداشت اینست که:

- اثر درآمد بر هزینه در هر سه چندک مثبت است؛ ولی از نظر آماری در چندک‌های ۰٫۲۵ و ۰٫۷۵ معنادار بوده ولی در میانه توزیع هزینه معنی‌دار نیست. به علاوه مقدار شیب در چندک‌های پایین و بالا تا حدود زیادی نزدیک به هم است.
- اثر سن در چندک‌های ۰٫۲۵ و ۰٫۷۵ معنی‌دار نیست، اما در میانه توزیع هزینه معنی‌دار است.
- شدت و معناداری اثرها در طول توزیع ثابت نیست بلکه آن‌ها به چندک مورد نظر توزیع هزینه وابسته هستند و این نشان از ناهمگنی اثرها در چندک‌های متفاوت دارد.

با هدف سنجش عملکرد و هم‌زمان داشتن سنجه‌های قابل‌مقایسه، خروجی مدل MD-IVQR با برازش یک مدل QR در همان سه چندک مقایسه شد و نتایج در جدول ؟؟ آمده است. بر اساس اعداد گزارش شده در این جدول، در چندک‌های پایین و بالای توزیع هزینه، مقادیر

[9] Root Mean Square Error

[10] Akaike Information Criterion

جذر میانگین توان دوم خطا[۹] (RMSE) و معیار اطلاع آکائیکه[۱۰] (AIC) برای مدل MD-IVQR کم‌تر از مدل QR است. همانطور که ملاحظه می‌شود اختلاف بسیار اندک و تاحدی به نفع مدل QR در چندک میانی توزیع هزینه وجود دارد. جمع‌بندی کلی این است که، در خصوص عملکرد، مدل پیشنهادی در مقایسه با مدل مرسوم چندکی در چندک‌های پایین و بالا از برتری قابل توجهی برخوردار است ولی در چندک میانی تفاوت قابل ملاحظه‌ای با مدل QR ندارد. در مقایسه بین این دو مدل، باید توجه داشت که بنا به معیار کمینه‌سازی، مدل QR مستقیماً سراغ تابع زیان چندکی می‌رود حال آن‌که MD-IVQR برای کنترل درون‌زایی طراحی شده و در انجام بهینه‌سازی به برآورده شدن محدویت‌های ناشی از درون‌زایی نیز توجه دارد. نکته کاربردی دیگر در برازش مدل‌های چندکی این است که تحقیقات راجع به معرفی معیارهای مناسب ارزیابی عملکرد مدل کماکان در جریان است و لذا در این تحقیق تنها به ارائه دو معیار RMSE و AIC که به صورت سنتی و دستی محاسبه شدند، بسنده شد. لازم به اشاره است نحوه محاسبه چنین معیارهایی در نرم‌افزارهای آماری به صورت مستقیم پیاده‌سازی نشد. به عنوان نمونه‌ای از تفاوت دیدگاه رگرسیون معمولی و رگرسیون چندکی در مسائل کاربردی می‌توان به فعالیت [۱۳] که پیشنهاد دادند بجای استفاده از معیار مرسوم R^2 از معیار $R^1(\tau)$ استفاده شود چرا که برخی مفاهیم رگرسیون معمولی را نمی‌توان به صورت مستقیم برای مدل رگرسیون چندکی به‌کار برد. همچنین، [۱۵] اعتقاد دارد که چون در رگرسیون چندکی هدف کمینه کردن تابع زیان چک[۱۱] (زیان پین‌بال[۱۲]) است نه مجموع مربعات یک کمیت، لذا، شاخص‌هایی مانند RMSE معیار ارزیابی مناسبی برای این روش نیست چرا که، در مقایسه با برازش رگرسیون معمولی با همان داده‌ها، آن الزاما مقدار کوچک‌تری نخواهد داشت.

توصیه ما در تحلیل داده‌های هزینه و درآمد خانوارهای شهری تهران در سال ۱۴۰۲ با کمک مدل‌های MD-IVQR و QR، این است که این دو مدل به‌صورت مکمل مورد استفاده قرار گیرند به این ترتیب که مدل QR برای توصیف الگوها و ناهمگنی چندکی، و مدل MD-IVQR برای کنترل درون‌زایی و سنجش اثرات ساختاری به کار گرفته شوند.

بحث و نتیجه‌گیری

در این مقاله رویکرد MD-IVQR برای مواجهه‌ی هم‌زمان با ناهمگنی چندکی و درون‌زایی به کار گرفته شد. مسائل نظری مرتبط با این مدل مطالعه و عملکرد آن در تحلیل داده‌های هزینه و درآمد خانوارهای شهری تهران (نمونه‌ی تصادفی ۶۰۰تایی از سال ۱۴۰۲) تشریح شد. توصیه کاربردی این است که این مدل به همراه مدل QR در تحلیل داده‌های ذکر شده مورد استفاده قرار گیرد.

تحلیل داده واقعی این پژوهش بر اساس حجم نمونه کم (نمونه‌ی ۶۰۰تایی) انجام شد که بهتر است در صورت وجود منابع محاسباتی قوی با حجم نمونه بیشتر صورت گیرد و مقایسه مدل‌ها بر اساس استنباط مجانبی (مثلا پیاده‌سازی الگوریتم بوت‌استرپ) انجام شود. به علاوه، در این پژوهش تنها دو شاخص FZ (بُعد خانوار) و GL (سطح زیربنای محل سکونت خانوار) مورد استفاده قرار گرفته و توصیه می‌شود در تحقیقات آینده متغیرهای بیشتری در مدل پیشنهادی این مقاله در نظر گرفته شود. محقق علاقه‌مند تحلیل مبتنی بر مدل معرفی شده در این مقاله را با متغیرهای بیشتر دنبال کند. توجه شود که تنظیمات محاسباتی MD-IVQR (تعداد سلول‌ها، میزان برش، تعداد بوت‌استرپ) می‌تواند بر دقت نتایج عددی تا حدی اثر بگذارد. بنابراین غنی‌تر کردن مجموعه‌ی ابزارها، آزمون حساسیت نسبت به تنظیمات محاسباتی و مقایسه با نسخه‌های پیشرفته‌تر (مثلاً اثرات ثابت گروهی یا جریمه‌گذاری) می‌تواند در آینده تحقیق مورد توجه قرار گیرد.

مراجع

[۱] بارانی، ع. و گلعلی‌زاده، م. (۱۳۹۷)، استفاده از رگرسیون آمیخته چندکی برای تحلیل هزینه و درآمد خانوارهای ایرانی، نشریه بررسی های آمار رسمی ایران، سال بیست و نهم، شماره ۲ (پیاپی ۹۳)، ۲۰۵–۱۸۵.

[۲] کشاورز حداد، غ. و نور اشرف الدین، م. (۱۳۹۴)، نرخ بازدهی خصوصی آموزش در مناطق شهری ایران: روش رگرسیون چندک، فصلنامه تحقیقات اقتصادی، ۵۰(۴)، ۱۰۱۶– ۹۸۹.

[11] Check Loss Function
[12] Pinball Loss

[۳] نوروزی شیروانه ده، ح.، مهرآرا، م. و اوریوئی، ع. (۱۴۰۰)، دهک‌های مختلف هزینه انرژی خانوار و عوامل مؤثر بر آن (مطالعه خرد مبتنی بر بودجه خانوار)، فصلنامه مطالعات اقتصاد انرژی، ۱۷ (۶۹)،۱۲۶-۹۵.

[4] Anderson, T. W. and Rubin, H. (1949), Estimation of Parameters of a Single Equation in a Complete System of Stochastic Equations, *Annals of Mathematical Statistics*, **20**, 46-63.

[5] Galvao, A.F, Gu, J. and Volgushev, S. (2020), On the Unbiased Asymptotic Normality of Quantile Regression with Fixed Effects, *Journal of Econometrics*, **218**(1), 178–215.

[6] Galvao, A. F., Parker, T. and Xiao, Z. (2024), Bootstrap Inference for Panel Data Quantile Regression, *Journal of Business and Economic Statistics*, **42**(2), 628-639.

[7] Galvao, A.F. and Wang, L (2015), Efficient Minimum Distance Estimator for Quantile Regression Fixed Effects Panel Data, *Journal of Multivariate Analysis*, **133**, 1-26.

[8] Gu, A. J. and Volgushev, S. (2019), Panel Data Quantile Regression with Grouped Fixed Effects, *Journal of Econometrics*, **213**, 68–91.

[9] Harding, M. and Lamarche, C. (2009), A Quantile Regression Approach for Estimating Panel Data Models using Instrumental Variables, *Economics Letters*, **104**(3), 133-135.

[10] Hsiao, C. (2003), *Analysis of Panel Data*, Cambridge University Press, New York.

[11] Imbens, G. W. and Angrist, J. D. (1994), Identification and Estimation of Local Average Treatment Effects, *Econometrica*, **62**, 467-475.

[12] Kivietd, J. F. (1995), On Bias, nConsistency and Efficiency of Various Estimators in Dynamic Panel Data Models, *Journal of Econometrics*, **68**, 53-78.

[13] Koenker, R. and Machado, J. A. F. (1999), Goodness of Fit and Related Inference Processes for Quantile Regression, *Journal of the American Statistical Association*, **94**, 1296-1310.

[14] Koenker, R. (2004), Quantile Regression for Longitudinal Data, *Journal of Multivariate Analysis*, **91**, 74-89.

[15] Koenker, R (2005), *Quantile Regression*, Cambridge University Press, New York.

[16] Rao, C. R. (1965), *Linear Statistical Inference and its Applications*, John Wiley and Sons, New York.

[17] Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley and Sons, New York.

[18] Tao, L., Tai, L., Qian, M. and Tian, M. (2023), A New Instrumental-Type Estimator for Quantile Regression Models, *Mathematics*, **11**(5), 3412.

Andisheye Amari, 2024, Vol. 29, No. 1
https://doi.org/10.22034/jr$_i$ss.2024.729436

Analysis of Minimum Distance Instrumental Variable Quantile Regression and Its Application on Analyzing Iranian Household Expenditure and Income Data

Hosseinzadeh , S. E.[1], **Golalizadeh, M.**[2*1,2] *Department of Statistics, Tarbiat Modares University, Tehran , Iran*

Abstract

One of the most common statistical models to study relationships among variables in a scientific investigation is the linear (mean-based) regression that rests on assumptions like normal errors and no correlation between regressors and the error term. When the normality of error component is questionable, quantile regression is a good choice because it is more robust to outliers. If endogeneity is a concern-i.e., regressors correlate with the error- instrumental variables come to play. One the other hand, minimum distance method which minimizes the gap between observed and model-implied distributions is powerful tool to fit the models. This paper lays out minimum-distance instrumental-variable quantile regression (MD-IVQR) and gives details on its relevant theoretical computation. Also, using Iranian household expenditure–income data from 1402, we compare MD-IVQR with conventional Quantile Regression method and show that it controls endogeneity and reveals distributional heterogeneity.

Keywords: Quantile regression, Endogeneity, Instrumental variables, Minimum distance, Household expenditure and income data.

Mathematics Subject Classification: 62J04, 62J08.

*Corresponding author, *golalizadeh@modares.ac.ir*