کاربرد رگرسیون چندک در شناسایی شکل توزیع رفاه مورد انتظار جوانان

محمد بامنی مقدم*، علی‌رضا خوش‌گویان فرد**

هدف اصلی این مقاله، آشنا کردن خواننده با کاربرد رگرسیون چندک در تحلیل داده‌هاست. رگرسیون چندک، رابطهٔ چندک دلخواهی از توزیع متغیر وابسته را با متغیرهای تشریحی از طریق مدل آماری تبیین می‌کند. در این مقاله، مدل رگرسیون چندک معرفی و به شیوهٔ برآورد پارامترها اشاره می‌شود؛ به قابلیت شناسایی شکل توزیع که مدل رگرسیون معمولی (میانگین شرطی) آن را دارا نیست، تأکید می‌شود؛ در پایان با یک مثال عددی از داده‌های رفاه که در آن براساس یک نمونهٔ تصادفی ۶۸۴ نفر از جوانان ۱۸ تا ۲۹ سال تهرانی انتخاب شده است، تلاش شده است کاربرد رگرسیون چندک برای رفاه تشریح شود و رابطهٔ رفاه مطلوب جوانان (متغیر وابسته مدل) با تعداد سال‌های تحصیل آنان (متغیر تشریحی مدل) تحت بررسی قرار گیرد.

کلید واژه‌ها: برآورد حداقل مربعات، حداقل قدر مطلق انحرافات، رگرسیون چندک، رفاه، معیار d کوک، میانگین شرطی

تاریخ دریافت مقاله: ۸۳/۷/۲۷ تاریخ پذیرش مقاله: ۸۳/۱۰/۲۸

مقدمه

ابعاد گوناگون پدیده‌ها در آمار، خود را قالب متغیرهای تصادفی نشان می‌دهند و مطالعه آن‌ها با تعیین توزیع (Distribution) آن‌ها میسر می‌شود. برای مثال، وزن نوزادان به عنوان

* دکترای آمار، عضو هیئت علمی دانشگاه علامه طباطبایی <bamenimoghadam@atu.ac.ir>

** فوق لیسانس آمار اقتصادی اجتماعی، مدیر هماهنگی پژوهش‌های مرکز تحقیقات صداوسیما

یک پدیده می‌تواند متغیری تصادفی محسوب شود و در صورت مشخص بودن توزیع آن، تعیین این که نوزادی دارای وزن طبیعی است، امکان‌پذیر می‌شود. توزیع متغیر تصادفی تمام اطلاعات لازم در خصوص آن را به دست می‌دهد، به طوری که می‌توان پدیده‌ها را تفسیر یا پیش‌بینی کرد. به عنوان مثال، با داشتن توزیع وزن نوزادان قادریم پیش‌بینی کنیم که تا چه اندازه ممکن است کودکانی کم‌وزن داشته باشیم.

در این راستا، معیارهای آماری گوناگونی وجود دارد که هر یک از توزیع متغیر تصادفی اطلاع متفاوتی در اختیار می‌گذارد. برای مثال، واریانس، از نحوهٔ پراکندگی و نما، از قلهٔ آن اطلاعی فراهم می‌کند. برخی از معیارها مانند برجستگی (Kurtosis: کشیدگی یا پخی) و چولگی (Skewness : کجی) نیز به شکل توزیع اختصاص دارند. شکل توزیع در متغیرهایی مانند درآمد یا هوش از اهمیت ویژه‌ای برخوردار است. مثلاً اگر توزیع درآمد در جامعه‌ای چولگی زیادی به چپ داشته باشد، حکایت از وجود افراد محرومی با درآمد بسیار کم دارد. هم‌چنین اگر توزیع نمرات هوش در دانشگاهی متقارن با دم‌هایی کوتاه باشد، نشان می‌دهد که در مجموع، هوش دانشجویان آن دانشگاه معمولی است و افراد باهوش و کم‌هوش در آن دانشگاه نسبتاً برابرند.

با توجه به آن‌چه گذشت دور از انتظار نیست اگر ادعا کنیم که استنباط‌های آماری (Statistical Inference) همگی اطلاعی از توزیع یک متغیر تصادفی به دست می‌دهند. برای مثال، آزمون F در تحلیل واریانس، مقایسه‌ای بین میانگین چند توزیع است؛ محاسبه ضریب همبستگی پیرسون (Pearson) و آزمون مربوط به آن اطلاعی از یک توزیع دو متغیره را فراهم می‌کند؛ مدل‌های رگرسیونی که در ادامه به آن می‌پردازیم نیز مانند دیگر روش‌ها برای بررسی خصوصیات خاصی از توزیع یک متغیر تصادفی به کار می‌روند.

مدل رگرسیون معمولی به تحلیل‌گر کمک می‌کند تا رابطهٔ میانگین توزیع متغیر تصادفی Y را با تعدادی متغیر تشریحی بررسی کند. برای روشن شدن مطلب، یک مدل رگرسیون خطی را تنها با یک متغیر تشریحی به این صورت در نظر می‌گیریم:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad \text{مدل ۱}$$

در این مدل رگرسیونی ε_iها، متغیرهای تصادفی، α و β، پارامترهای نامعلوم که باید برآورد شوند و سرانجام x_iها مقادیر معلومی از متغیر تشریحی هستند. در صورتی که $E(\varepsilon_i)=0$ باشد، آن‌گاه می‌توان مدل شمارهٔ ۱ را به صورت دوم بازنویسی کرد:

$$E(Y_i) = \alpha + \beta x_i \qquad \text{مدل۲:}$$

کمیت $E(Y_i)$ را میانگین شرطی (Conditional Mean) متغیر تصادفی Y می‌نامند و به همین دلیل آن را با $E(Y/x_i)$ نیز نشان می‌دهند. بنابراین، مدل شمارهٔ ۲ بیان می‌کند که میانگین‌های توزیع Y در سطوح مختلف متغیر تشریحی در امتداد یک خط راست قرار دارند. به عبارت دیگر، متغیر تصادفی Y در هر سطح از متغیر تشریحی دارای توزیعی است که میانگین‌های این توزیع‌ها روی یک خط راست جای گرفته‌اند. یکی از حالاتی که در آن چنین رابطه‌ای قطعاً برقرار می‌شود زمانی است که دو متغیر X و Y دارای توزیع نرمال دو متغیره باشند (باز هم صحبت از توزیع است!).

از آن‌جا که میانگین، یکی از معیارهای تمرکز است، آگاهی از آن به تنهایی نمی‌تواند اطلاعات کاملی از شکل توزیع به همراه داشته باشد. با توجه به این واقعیت، رگرسیون معمولی نیز ممکن است نتواند اطلاعات کافی دربارهٔ شکل توزیع متغیر تصادفی تحت مطالعه را ـ در سطوح مختلف متغیر تشریحی ـ به دست دهد. چندک‌ها (Quantiles) معیارهای دیگری برای توزیع هستند که «در کنار هم» می‌توانند شکل توزیع را جامع‌تر به تصویر بکشند. برای مثال، اگر دهک‌های توزیعی تقریباً دارای فاصله برابری از یکدیگر باشند، انتظار داریم توزیع «نسبتاً» هموار یا یک‌نواختی داشته باشیم. هم‌چنین اگر دهک‌های بالایی دارای فاصله زیاد و دهک‌های پایینی دارای فاصله کمی از یکدیگر باشند، توزیع به سمت راست چوله خواهد بود. اکنون اگر مانند رگرسیون معمولی که برای میانگین به کار می‌رود، یک شیوهٔ رگرسیونی برای چندک‌ها وجود داشته باشد، قادر خواهیم بود شکل توزیع را در سطوح مختلف متغیرهای تشریحی به

دست آوریم. این همان هدفی است که رگرسیون چندک دنبال می‌کند.

۱) معرفی رگرسیون چندک

همان‌طور که در بخش قبل اشاره شد، مدل رگرسیون معمولی، مانند مدل شمارهٔ ۲، برای میانگین شرطی برازش داده می‌شود. مدل رگرسیون چندک با ایده‌ای مشابه برای چندک‌های شرطی (Conditional Quantile) به کار می‌رود. مانند رگرسیون معمولی (میانگین)، کاربردهایی نظیر بررسی رابطه متغیرهای تشریحی با چندک‌ها و هم‌چنین پیش‌بینی آن‌ها برای این نوع از رگرسیون نیز امکان‌پذیر است. با وجود این، شاید مهم‌ترین کاربرد رگرسیون چندک شناسایی شکل توزیع متغیر وابسته مدل در سطوح گوناگون متغیرهای تشریحی باشد؛ این کار با برازش مدل‌های رگرسیونی متعدد، به ازای چندک‌های مختلف بر یک مجموعه‌داده، صورت می‌گیرد.

برای ارائهٔ تعریف دقیقی از مدل رگرسیون چندک $\theta \in (0,1)$ام، ابتدا حالت سادهٔ آن را در نظر می‌گیریم. مدل شمارهٔ ۱ را با شرط $\varepsilon_i \overset{iid}{\sim} F(\cdot)$ (تابع F به یک توزیع دلخواه اشاره دارد) در نظر بگیرید. هدف ما یافتن مدلی است که مثلاً رابطهٔ چندک اول (و نه میانگین) توزیع Y را با متغیر X نشان دهد. در این صورت، مدل برای چندک $\theta \in (0,1)$ام متغیر Y که با $Q_\theta(Y \mid x_i)$ نشان داده می‌شود، عبارت است از:

$$Q_\theta(Y \mid x_i) = \alpha + \beta x_i + F^{-1}(\theta)$$ مدل۳:

تابع فوق، به ازای $\theta \in (0,1)$های مختلف، دسته‌ای از خطوط موازی را به دست خواهد داد که دارای عرض از مبدأهای متفاوتی هستند. در صورتی که $F(\cdot)$ همان توزیع نرمال (یا هر توزیع متقارن دیگری) باشد، به ازای $\theta = 0.5$، مدل شمارهٔ ۳ همان مدل شمارهٔ ۲ خواهد بود، زیرا $F^{-1}(0.5) = 0$ ممکن است به یک تغییر مکان نیاز باشد. اکنون به تعریف کلی مدل رگرسیون چندک می‌پردازیم. برای این منظور، فرض کنید $Y_i = x_i'\beta_\theta + \varepsilon_{\theta i}$ و

$$Q_\theta(Y \mid x_i') = x_i'\beta_\theta \qquad i=1,\ldots,n \qquad \text{مدل ۴:}$$

که در آن $x_i' = (1, x_{i1},\ldots, x_{ik})$ و $\beta_\theta' = (\beta_0, \beta_1,\ldots,\beta_k)$ به ترتیب برداری از مقادیر معلوم و پارامترهای نامعلوم بوده و $\varepsilon_{\theta i}$ یک متغیر تصادفی مشاهده‌نشدنی است. هم‌چنین، $Q_\theta(Y \mid x_i)$ نمایان‌گر چندک شرطی $\theta \in (0,1)$ام توزیع Y است، بنابراین $Q_\theta(\varepsilon_\theta \mid x_i) = 0$. مدل شمارهٔ ۴ را با شرایط گفته شده، مدل رگرسیون خطی چندک θ ام می‌نامیم.

شیوهٔ برآورد پارامترهای مدل رگرسیون معمولی بر حداقل کردن مربع باقیمانده‌های (انحرافات) مدل مبتنی است که روش حداقل مربعات (Least Squares) نامیده می‌شود. در این روش، منحنی رگرسیونی به گونه‌ای برازش داده می‌شود که در مجموع، فاصلهٔ نقاط از آن به حداقل برسد. در رگرسیون چندک برخلاف رگرسیون معمولی از حداقل کردن مجموع قدر مطلق موزون برای برآورد پارامترهای مدل استفاده می‌شود که به آن روش حداقل قدر مطلق انحرافات (Least Absolute Deviations) یا LAD گفته می‌شود. گفتنی است که استفاده از روش LAD که در مدل رگرسیون چندک به کار می‌رود، دارای پیشینه‌ای طولانی است. در میانهٔ قرن هجدهم، بوسکویچ (Boscovich) یک مدل خطی دو متغیره را برای بررسی بیضوی بودن کره زمین از طریق کمینه کردن قدر مطلق خطاها به کار برد. به دنبال آن، لاپلاس (Laplace) برآورد ضریب زاویهٔ مدل رگرسیونی بوسکویچ را به طور دقیق معرفی و توزیع مجانبی آن را به دست آورد. ظاهراً، اجورث (F. Y. Edgeworth) اولین کسی است که مدل رگرسیونی میانه با چند متغیر تشریحی را در حالت کلی بررسی کرد. توسعهٔ رگرسیون میانه برای هر چندک دلخواه نیز به کوشش کانوکر و باست (Koenker and Bassett) در ۱۹۷۸ صورت گرفت. هدف آن‌ها برآورد بردار پارامترهای $\beta_\theta' = (\beta_0, \beta_1,\ldots,\beta_k)$ در مدل شمارهٔ ۴ بود که برای این منظور تابع زیانی که در پی می‌آید (قدر مطلق باقی‌مانده‌ها یا انحرافات موزون) نسبت به عناصر β_θ کمینه می‌شود:

$$\varphi_\theta(\beta_\theta) = \sum_i w(\theta) \mid y_i - x_i'\beta_\theta$$ مدل ۵:

در این تابع زیان $w(\theta) = \begin{cases} \theta & Y_i \le X'_i \beta_\theta \\ 1-\theta & Y_i > X'_i \beta_\theta \end{cases}$. موزون کردن قدر مطلق باقیمانده‌ها در تابع فوق باعث می‌شود تا خط برازشی به گونه‌ای باشد که $\theta \times 100$٪ داده‌ها تقریباً زیر آن و باقی آن‌ها بالای خط قرار گیرند. کمینه کردن رابطهٔ فوق و یافتن برآورد LAD پارامترها با استفاده از روش‌های برنامه‌ریزی خطی و از طریق بسته‌های نرم‌افزاری صورت می‌گیرد. در ادامه به ویژگی‌های برآورد LAD اشاره می‌شود.

۱-۱) ویژگی‌های حداقل قدر مطلق انحرافات (LAD)

الف) در حالت خاص که مدل تنها شامل عرض از مبدأ و $\theta=0.5$ است، کمینه کردن رابطهٔ ۵ منجر به کمینه کردن عبارت $\sum_i |y_i - \beta_0|$ می‌شود که در این صورت برآورد β_0 همان میانه داده‌ها خواهد بود.

ب) برخلاف روش حداقل مربعات، روش حداقل قدر مطلق انحرافات نسبت به داده‌های دور افتاده (Outliers) استوار (Robust) است. این ویژگی ناشی از آن است که برخلاف اهمیت اندازه باقی‌مانده‌ها در روش حداقل مربعات، در این روش تنها به علامت باقی‌مانده‌ها توجه نمی‌شود. بنابراین نه تعداد باقی‌مانده‌هایی که بیش‌تر (مثبت) یا کم‌تر (منفی) از چندک مورد نظرند و نه مقدار بزرگی آن‌ها در برآوردها اثرگذار است. پس، داده‌های دورافتاده که تأثیر خود را از طریق بزرگی باقی‌مانده‌ها نشان می‌دهند، نمی‌توانند برآوردهای LAD را متأثر سازند.

ج) شکل بسته‌ای برای برآورد پارامترهای این مدل وجود ندارد و از روش‌های عددی برای برآورد آن‌ها استفاده می‌شود. هم‌چنین، جواب‌های نهایی مدل رگرسیون چندک می‌تواند یکتا نباشد. البته یافتن جواب یکتا با انتخاب یک معیار مناسب امکان‌پذیر است. برای مثال، مسئلهٔ یافتن میانهٔ ۱۰ عدد را به یاد بیاورید که برای این منظور، میانگین عدد پنجم و ششم به عنوان میانه در نظر گرفته می‌شد. این در حالی است که کلیهٔ اعداد بین

عدد پنجم و ششم می‌توانند به عنوان میانه انتخاب شوند. در واقع، با یک قرارداد (معیار)، میانهٔ این اعداد به صورت منحصر به فردی تعیین می‌شود.

د) وقتی $\varepsilon_{\theta i}$ها متغیرهای تصادفی iid باشند، خطوط رگرسیونی به ازای چندک‌های مختلف، موازی خواهند بود (مدل شمارهٔ ۳ را به یاد بیاورید).

در رگرسیون چندک نیز مانند رگرسیون معمولی (میانگین شرطی) می‌توان استنباط کرد. پاول و بوچنسکی نشان داده‌اند که برآورد LAD پارامترها، سازگار و به طور مجانبی نرمال است (Powell, 1989., Buchinsky, 1998). موضوع اخیر را برای حالت خاص که $\varepsilon_{\theta i}$ها هم‌توزیع هستند در قضیه زیر بیان می‌کنیم.

۲-۱) قضیه

فرض کنید X یک ماتریس $n\times k$ باشد که سطر iام آن را x_i' تشکیل می‌دهد. هم‌چنین تابع چگالی و توزیع $\varepsilon_{\theta i}$ها به ترتیب f و F باشد. اگر $\lim_{n\to+\infty}[(XX')/n]=V$، $f\left[F^{-1}(\theta)\right]>0$، $g^2(\theta,F)=\frac{\theta(1-\theta)}{[f(F^{-1}(\theta)]^2}$ و $\beta_\theta^*=(F^{-1}(\theta)+\beta_0,\beta_1,...,\beta_k)'$ برآورد آن را با b_θ^* نشان دهیم، آن‌گاه

مدل ۶: $$\sqrt{n}(b_\theta^*-\beta_\theta^*)\xrightarrow{d}N[0,g^2(\theta,F)V]$$

اکنون با استفاده از این قضیه می‌توان دربارهٔ ضرایب مدل رگرسیون چندک استنباط کرد. بر این اساس، آمارهٔ آزمون فرضیهٔ صفر $H_0:A\beta_\theta=l$ که در آن A یک ماتریس معلوم با رتبهٔ کامل سطری و l یک بردار از مقادیر معلوم است، عبارت است از:

مدل ۷: $$\lambda=(Ab_\theta-l)'[A(XX')^{-1}A'](Ab_\theta-l)\hat{g}^{-2}$$

این بخش را با بررسی نحوهٔ شناسایی شکل توزیع و با استفاده از رگرسیون چندک پایان می‌دهیم. برای این منظور ابتدا مدل‌های متعددی به ازای چندک‌های مختلف بر داده‌ها برازش داده می‌شود. سپس، برای سطوحی از متغیرهای تشریحی که مدّ نظر است،

چندک‌ها براساس مدل‌های برازشی پیش‌بینی می‌شوند و از مقایسهٔ چندک‌های پیش‌بینی شده، به شکل توزیع در آن سطوح از متغیرهای تشریحی پی می‌بریم. برای مثال، فرض کنید نه دهک از توزیع متغیر وابسته در سطحی از متغیر تشریحی به صورت حالت اول، دوم یا سوم جدول شمارهٔ ۱ پیش‌بینی شده است.

جدول ۱

دهک	۰/۱	۰/۲	۰/۳	۰/۴	۰/۵	۰/۶	۰/۷	۰/۸	۰/۹
حالت اول	۳	۴	۶	۷	۸	۱۳	۱۹	۲۹	۳۷
حالت دوم	۳	۸	۱۱	۱۵	۱۹	۲۳	۲۷	۳۲	۳۷
حالت سوم	۳	۱۵	۲۲	۲۷	۲۹	۳۱	۳۲	۳۴	۳۷

الف) حالت اول به توزیع چوله به راستی اشاره دارد. به فاصلهٔ بیش‌تر میان دهک‌های بالایی در مقایسه با دهک‌های پایینی توجه کنید. این نشان می‌دهد که ۴۰ درصد از مقادیر بزرگ‌تر متغیر در فاصله وسیعی جای گرفته‌اند در حالی که ۶۰ درصد باقی‌مانده از مقادیر کوچک‌تر، در فاصلهٔ کوتاه‌تری قرار دارند.

ب) حالت دوم نمایان‌گر یک توزیع نسبتاً هموار است، زیرا فاصله دهک‌ها تقریباً برابر است.

ج) حالت سوم وضعیتی عکس حالت اول دارد؛ یعنی توزیع چوله به چپی را نشان می‌دهد.

آن‌چه از طریق پیش‌بینی چندک‌ها با کمک مدل به دست آمد ـ در حالتی که تنها یک متغیر تشریحی در مدل وجود داشته باشد، به کمک رسم خطوط مدل‌های برازشی نیز دست‌یافتنی است. در واقع، بررسی فاصلهٔ خطوط رگرسیونی چندک‌های مختلف نیز می‌تواند فاصلهٔ چندک‌ها را از یکدیگر در سطوح مختلف متغیر تشریحی نشان دهد.

۲) کاربرد عملی

این بخش به ارائهٔ یک مثال عددی از رگرسیون چندک اختصاص دارد. یازده مدل رگرسیون چندک همراه با رگرسیون معمولی برای بررسی رابطهٔ رفاه درخواستی افراد

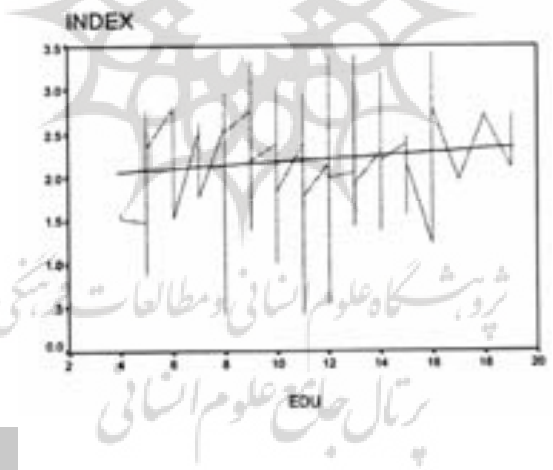
(متغیر وابستهٔ مدل) با تعداد سال‌های تحصیل آنان (متغیر تشریحی مدل) به کار می‌رود. رفاه مطلوب و تعداد سال‌های تحصیل به ترتیب با INDEX و EDU نشان داده خواهند شد. داده‌ها به یک نمونهٔ تصادفی ۶۸۴ نفری از جوانان ۱۸ تا ۲۹ سال تهرانی اختصاص دارد که با روش نمونه‌گیری خوشه‌ای سه مرحله‌ای (Three-stage Cluster Sampling) در سال ۱۳۸۱ گردآوری شده است. شایان ذکر است که رفاه درخواست شد. با شاخصی که حاصل از ترکیب ۳۳ سؤال یک پرسش‌نامه است، سنجیده شده است. گفتنی است مقدار این شاخص از ۰ تا ۴ تغییر می‌کند، به طوری که هرچه مقدار آن بیش‌تر می‌شود بر انتظارات بیش‌تری نیز دلالت دارد.

کار را با برازش یک مدل رگرسیون خطی معمولی با روش حداقل مربعات بر داده‌ها آغاز می‌کنیم. برآورد پارامترهای مدل (عرض از مبدأ و ضریب متغیر تعداد سال‌های تحصیل) در جدول شمارهٔ ۲ دیده می‌شود؛ بر این اساس مدل برازشی عبارت است از:

$$E(\hat{INDEX}_i) = 1.99 + 0.0189 EDU_i$$

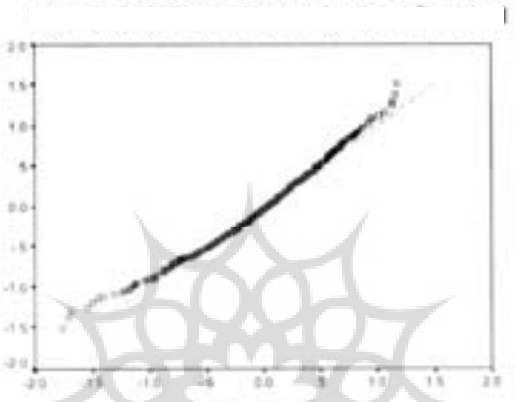
که در آن $E(\hat{INDEX}_i)$ برآورد میانگین توزیع رفاه درخواستی به ازای EDU_i سال تحصیل است. نمودار پراکنش (Scatter Plot) داده‌ها همراه با خط برازش داده شده در شکل شمارهٔ ۱ ارائه شده است.

شکل ۱: خط رگرسیونی برازشی و نقاط مشاهده شده



با توجه به مدل برازشی، معیار «d کوک» (Cook's D) به وجود تعداد زیادی دادهٔ دورافتاده اشاره داشت. هم‌چنین ترسیم نمودار Q-Q (Quantile-Quantile Plot) در شکل شمارهٔ ۲ برای بررسی نرمال بودن توزیع باقی‌مانده‌های مدل، انحراف از توزیع نرمال را نشان می‌دهد. آزمون کلموگرف ـ اسمیرنف (Kolmogorov Smirnov Test) نیز فرضیهٔ نرمال بودن این توزیع را رد می‌کند. تمام این موارد حکایت از نامناسب بودن مدل رگرسیون معمولی دارد.

شکل ۲: نمودار Q-Q برای باقی‌مانده‌های مدل رگرسیونی برازشی



اکنون به برازش مدل‌های رگرسیون چندک به ازای یازده چندک مختلف به کمک برنامه‌ای که در محیط IML از نرم‌افزار SAS نوشته شده است، می‌پردازیم (برنامه در پیوست ارائه شده است). این برنامه برای مدل ۳ و براساس الگوریتمی است که باست و کونکر در ۱۹۸۲ تهیه کرده‌اند (Bassett & Koenker, 1982). گفتنی است وقتی این چندک، همان میانه باشد، الگوریتم سریع‌تری از طرف مدسن و نیلسن پیشنهاد شده است که در SAS/IML در

قالب روال LAV (Routine) استفاده می‌شود (Madsen & Nielsen, 1993).

در مجموع، یازده مدل به ازای یازده چندک ۰/۱، ۰/۲، ۰/۲۵، ۰/۳، ۰/۴، ۰/۵، ۰/۶، ۰/۷، ۰/۷۵، ۰/۸، ۰/۹ بر داده‌ها برازش داده شد. برآورد پارامترهای این مدل‌ها در جدول شمارهٔ ۱ ارائه شده است. شکل شمارهٔ ۳ نیز خطوط برازش داده شده را نشان می‌دهد (خطوط از پایین به بالا مربوط به چندک‌های ۰/۱ تا ۰/۹ هستند).

بنابراین، مثلاً، مدل برازش داده شده برای چندک ۰/۲ عبارت است از

$$\hat{Q}_{i,0.2} = 1.58 + 0.0225 EDU_i$$

که در آن $\hat{Q}_{i,0.2}$ برآورد چندک ۰/۲ توزیع رفاه درخواستی، به ازای EDU_i سال تحصیل است (در شکل شمارهٔ ۳، خط دوم از پایین، مربوط به این مدل است). پس برآورد چندک ۰/۲ برای افراد با ۱۹ سال تحصیل برابر است با

$$2.0075 = 1.58 + 0.0225 \times 19$$

بنابراین می‌توان انتظار داشت که ۲۰ درصد از افراد با ۱۹ سال تحصیل، دارای شاخص انتظارات کم‌تر از ۲/۰۰۷۵ و ۸۰ درصد بیش از آن باشند. به همین ترتیب براساس مدل چندک ۰/۹، انتظار داریم ۹۰ درصد از افراد با ۱۹ سال تحصیل دارای شاخص انتظارات کم‌تر از ۲/۸۲۴۹ و ۱۰ درصد بیش از آن باشند. پس تقریباً ۷۰ درصد از این افراد دارای شاخصی بین ۲/۰۰۷۵ و ۲/۸۲۴۹ هستند. توجه کنید که چنین تحلیل‌هایی تنها با مدل‌های رگرسیون چندک قابل انجام است و مدل‌های رگرسیون معمولی چنین قابلیت‌هایی را ندارند. اکنون به یافته‌های حاصل از برازش این مدل‌ها می‌پردازیم:

۱-۲) یافته‌های حاصل از برازش

الف) شکل شمارهٔ ۱ حاکی از آن است که خط رگرسیون نمی‌تواند پیش‌بینی‌کنندهٔ مناسبی برای شاخص انتظارات باشد. زیرا با توجه به پراکندگی زیاد داده‌ها در برخی از سطوح تعداد سال‌های تحصیل، میانگین نمی‌تواند این شاخص را برای این سطوح به خوبی پیش‌بینی کند. برای مثال در این شکل افرادی را با ۸ یا ۱۲ سال تحصیل ملاحظه کنید.

ب) مثبت بودن شیب خطوط در شکل شمارهٔ ۳ یعنی ضریب تعداد سال‌های تحصیل، نشان‌دهندهٔ رابطهٔ مستقیم بین متغیر وابسته و تشریحی است. بنابراین، با افزایش تعداد سال‌های تحصیل، مقدار هر یک از یازده چندک شاخص انتظارات نیز افزایش می‌یابد. بر اساس چندک‌های برازشی، سال‌های تحصیل بر چندک‌های پایینی بیش‌تر از چندک‌های بالایی اثر دارد.

ج) برای افراد با تحصیلات بیش‌تر، فاصلهٔ کم چندک‌های بالایی (چندک ۰/۹، ۰/۸، ۰/۷۵ و ۰/۶) در مقایسه با چندک‌های پایینی نشان می‌دهد که فشردگی داده‌ها در بخش بالایی زیاد است. به عبارت دیگر، در سمت راست توزیع شرطی، فشردگی بیش‌تری در مقایسه با سمت چپ وجود دارد. بنابراین، با افزایش تعداد سال‌های تحصیل، توزیع شرطی شاخص مطلوب، چوله به چپ می‌شود.

د) می‌توان پیش‌بینی کرد که مثلاً ۷۰ درصد (از دهک ۰/۲ تا ۰/۹) افراد با ۱۹ سال تحصیل دارای شاخصی بین ۲/۰۰۷۵ و ۲/۸۲۴۹ هستند، در حالی که این فاصله برای

افرادی با ۴ سال تحصیل از ۱/۶۷ تا ۲/۷۱۸۴ است.

جدول ۱: مشخصات مدل‌های رگرسیونی

مدل		برآورد پارامترهای مدل		مجموع قدرمطلق موزون خطاها
		عرض از مبدأ	تعداد سال‌های تحصیل	
رگرسیون معمولی		۱/۹۹	۰/۰۱۸۹	—
رگرسیون چندک	۰/۱	۱/۴۲	۰/۰۱۲۹	۶۶
	۰/۲	۱/۵۸	۰/۰۲۲۵	۹۹
	۰/۲۵	۱/۶۵	۰/۰۲۵	۱۱۰
	۰/۳	۱/۷	۰/۰۲۵	۱۱۸
	۰/۴	۱/۸۷	۰/۰۲۵	۱۲۷
	۰/۵	۲/۰۹	۰/۰۱۲۹	۱۲۸
	۰/۶	۲/۱۶	۰/۰۱۶۶	۱۲۲
	۰/۷	۲/۳۴	۰/۰۱۱۳	۱۰۸
	۰/۷۵	۲/۴۷	۰/۰۰۶۳	۹۸
	۰/۸	۲/۴۷	۰/۰۱۲۵	۸۶
	۰/۹	۲/۶۹	۰/۰۰۷۱	۵۲

۳) نتیجه‌گیری

این مقاله نشان داد که رگرسیون چندک نه تنها می‌تواند جانشین مناسبی برای رگرسیون میانگین باشد (با جانشین کردن میانه به جای میانگین)، بلکه در برخی از حالات، اطلاعات بیش‌تری (شکل توزیع) را در مقایسه با رگرسیون میانگین در اختیار تحلیل‌گر قرار می‌دهد. در بخش قبل دیده شد که رگرسیون میانگین به سبب وجود داده‌های دورافتاده و انحراف از متعادل بودن و هم‌چنین پراکندگی زیاد متغیر پاسخ در برخی از سطوح متغیر تشریحی، از اعتبار لازم برخوردار نبود؛ در حالی که رگرسیون چندک، یافته‌های مفیدی را به دست داد.

References منابع

1. Bassett, G. and Koenker, R. (1982). **"An Empirical Quantile Function for Linear Models with iid Errors"**. Journal of American Statistical Association, Vol. 77, No.378: 407-415.
2. Buchinsky, M. (1998). **"Recent Advances in Quantil Regression Models: A Practical Guideline for Empirical Research"**. Journal of Human Resources, 33(1): 88-126.
3. Koenker, R. and Bessett, G. (1978). **"Reggression Quantiles".** Econometrica, 46: 33-50.
4. Madsen, K. And Nielsen, H. B. (1993). **"A Finite Smoothing Algorithm for Linear L1 Estimation"**. SIAM Journal Optimization, Vol. 3: 223-235.
5. Powell, J. (1989). **"Least Absolute Deviation Estimation of the Censored Regression Model"**, Journal of Econometrics, 25: 303-325.