

کاربرد رگرسیون چندک در شناسایی شکل توزیع رفاه مورد انتظار جوانان

✚ بامنی مقدم*، علی‌رضا خوشگویان فرد**

هدف اصلی این مقاله، آشنا کردن خواننده با کاربرد رگرسیون چندک در تحلیل داده‌هاست. رگرسیون چندک، رابطه 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 شیوه 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 نمونه تصادفی 684 نفر از جوانان 18 تا 29 سال تهرانی انتخاب شده است، تلاش شده است کاربرد رگرسیون چندک برای رفاه تشریح شود و رابطه رفاه مطلوب جوانان (متغیر وابسته مدل) با تعداد سال‌های تحصیل آنان (متغیر تشریحی مدل) تحت بررسی قرار گیرد.

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

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

کلید واژه‌ها: برآورد حداقل مربعات، حداقل قدر مطلق

انحرافات، رگرسیون چندک، رفاه،

معیار d کوک، میانگین شرطی

تاریخ دریافت مقاله: 83/7/27 تاریخ پذیرش مقاله:

83/10/28

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 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 نحوه پراکندگی و نما، از قله 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 ویژه ای برخوردار است. مثلاً اگر توزیع درآمد در جامعه ای 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 دم هایی کوتاه باشد، نشان می‌دهد که در مجموع، هوش دانشجویان آن دانشگاه معمولی است و افراد باهوش و کم هوش در آن دانشگاه نسبتاً برابرند.

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

مدل رگرسیون معمولی به تحلیل گر کمک می‌کند تا رابطه 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:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad \text{مدل 1}$$

در این مدل رگرسیونی ε_i ها، متغیرهای تصادفی، α و β ، پارامترهای نامعلوم که باید برآورد شوند و سرانجام x_i ها مقادیر معلومی از متغیر تشریحی هستند. در صورتی که $E(\varepsilon_i) = 0$ باشد، آنگاه می‌توان مدل شماره 1 را به صورت دوم بازنویسی کرد:

$$E(Y_i) = \alpha + \beta x_i \quad \text{مدل 2:}$$

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 شماره 2 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 آن جا که میانگین، یکی از معیارهای تمرکز است، آگاهی از آن به تنهایی نمی تواند اطلاعات کاملی از شکل توزیع به همراه داشته باشد. با توجه به این واقعیت، رگرسیون معمولی نیز ممکن است نتواند اطلاعات کافی درباره شکل توزیع متغیر تصادفی تحت مطالعه را - در سطوح مختلف متغیر تشریحی - به دست دهد. چندکها (Quantiles) معیارهای دیگری برای توزیع هستند که «در کنار هم» می توانند شکل توزیع را جامع تر به تصویر بکشند. برای مثال، اگر دهک های توزیعی تقریباً دارای فاصله برابری از یکدیگر باشند، انتظار داریم توزیع «نسبتاً» هموار یا یک نواختی داشته باشیم. همچنین اگر دهک های بالایی دارای فاصله زیاد و دهک های پایینی دارای فاصله کمی از یکدیگر باشند، توزیع به سمت راست چوله خواهد بود. اکنون اگر مانند رگرسیون معمولی که برای میانگین به کار می رود، یک شیوه 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 می کند.

1) معرفی رگرسیون چندک

همان طور که در بخش قبل اشاره شد، مدل رگرسیون معمولی، مانند مدل شماره 2، برای میانگین شرطی

برازش داده می‌شود. مدل رگرسیون چندک با ایده ای مشابه برای چندک های 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 امکان پذیر 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 چندک های مختلف بر یک مجموعه داده، صورت می‌گیرد.

برای ارائه تعریف دقیقی از مدل رگرسیون چندک $\theta \in (0,1)$ ام، ابتدا حالت ساده آن را در نظر می‌گیریم. مدل شماره 1 را با شرط $\varepsilon_i \sim F(\cdot)$ (تابع F به یک توزیع دلخواه اشاره دارد) در نظر بگیرید. هدف ما یافتن مدلی است که مثلاً رابطه 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 $Q_\theta(Y|x_i)$ نشان داده می‌شود، عبارت است از:

$$Q_\theta(Y|x_i) = \alpha + \beta x_i + F^{-1}(\theta) \quad \text{مدل 3:}$$

تابع فوق، به ازای $\theta \in (0,1)$ های مختلف، دسته ای از خطوط موازی را به دست خواهد داد که دارای عرض

از مبداهای متفاوتی هستند . در صورتی که $F(\cdot)$ همان توزیع نرمال (یا هر توزیع متقارن دیگری) باشد، به ازای $\theta=0.5$ ، مدل شماره 3 همان مدل شماره 2 خواهد بود، زیرا $F^{-1}(0.5)=0$ ممکن است به یک تغییر مکان نیاز باشد . اکنون به تعریف کلی مدل رگرسیون چندک می پردازیم. برای این منظور، فرض کنید $Y_i = x'_i \beta_\theta + \varepsilon_{\theta i}$ و

$$Q_\theta(Y|x'_i) = x'_i \beta_\theta \quad i=1, \dots, n \quad \text{مدل 4:}$$

که در آن $x'_i = (1, x_{i1}, \dots, x_{ik})$ و $\beta'_\theta = (\beta_0, \beta_1, \dots, \beta_k)$ به ترتیب برداری از مقادیر معلوم و پارامترهای نامعلوم بوده و $\varepsilon_{\theta i}$ یک متغیر تصادفی مشاهده نشدنی است. همچنین، $Q_\theta(Y|x_i)$ نمایانگر چنک شرطی $\theta \in (0,1)$ ام توزیع Y است، بنابراین $Q_\theta(\varepsilon_\theta|x_i) = 0$. مدل شماره 4 را با شرایط گفته شده، مدل رگرسیون خطی چندک θ ام می نامیم.

شیوه برآورد پارامترهای مدل رگرسیون معمولی بر حداقل کردن مربع باقیمانده های (انحرافات) مدل مبتنی است که روش حداقل مربعات (Least Squares) نامیده می شود. در این روش، منحنی رگرسیونی به گونه ای برازش داده می شود که در مجموع، فاصله 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 پیشینه ای طولانی است. در میانه قرن هجدهم، بوسکوویچ (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 زاویه مدل رگرسیونی بوسکوویچ را به طور دقیق معرفی و توزیع جانبی آن را به دست آورد. ظاهراً، اچورث (F. Y. Edgeworth) اولین کسی است که مدل رگرسیونی میانه با چند متغیر تشریحی را در حالت کلی بررسی کرد. توسعه رگرسیون میانه برای هر چندک دلخواه نیز به کوشش کانوکر و باست (Koenker and Bassett) در 1978 صورت گرفت. هدف آن‌ها برآورد بردار پارامترهای $\beta'_\theta = (\beta_0, \beta_1, \dots, \beta_k)$ در مدل شماره 4 بود که برای این منظور تابع زیانی که در پی می‌آید (قدر مطلق باقی‌مانده‌ها یا انحرافات موزون) نسبت به عناصر β_θ کمینه می‌شود:

$$\varphi_\theta(\beta_\theta) = \sum_i w(\theta) |y_i - x'_i \beta_\theta| \quad \text{مدل 5:}$$

$$w(\theta) = \begin{cases} \theta & Y_i \leq X'_i \beta_0 \\ 1-\theta & Y_i > X'_i \beta_0 \end{cases}$$
 در این تابع زیان موزون کردن قدر مطلق باقیمانده‌ها در تابع فوق باعث می‌شود تا خط برازشی به گونه ای باشد که

$\theta * 100\%$ داده‌ها تقریباً زیر آن و باقی آن‌ها بالای خط قرار گیرند. کمینه کردن رابطه 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.

1-1) ویژگی‌های حداقل قدر مطلق انحرافات (LAD)

الف) در حالت خاص که مدل تنها شامل عرض از مبدأ و $\theta = 0.5$ است، کمینه کردن رابطه 5 منجر به کمینه کردن عبارت $\sum |y_i - \beta_0|$ 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 بسته ای برای برآورد پارامترهای این مدل وجود ندارد و از روش های عددی برای برآورد آنها استفاده می شود. همچنین، جواب های نهایی مدل رگرسیون چندک می تواند یکتا نباشد. البته یافتن جواب یکتا با انتخاب یک معیار مناسب امکان پذیر است. برای مثال، مسئله یافتن میانه 10 عدد را به یاد بیاورید که برای این منظور، میانگین عدد پنجم و ششم به عنوان میانه در نظر گرفته می شد. این در حالی است که کلیه اعداد بین عدد پنجم و ششم می توانند به عنوان میانه انتخاب شوند. در واقع، با یک قرارداد (معیار)، میانه این اعداد به صورت منحصر به فردی تعیین می شود.

د) وقتی ε_{θ_i} ها متغیرهای تصادفی iid باشند، خطوط رگرسیونی به ازای چندک های مختلف، موازی خواهند بود (مدل شماره 3 را به یاد بیاورید). در رگرسیون چندک نیز مانند رگرسیون معمولی (میانگین شرطی) می توان استنباط کرد. پاول و بوچنسکی نشان داده اند که برآورد LAD پارامترها، سازگار و به طور مجانبی نرمال است (Powell, 1989., Buchinsky, 1998). موضوع اخیر را برای حالت خاص که ε_{θ_i} ها هم توزیع هستند در قضیه زیر بیان می کنیم.

1-2) قضیه

فرض کنید X یک ماتریس $n \times k$ باشد که سطر i ام آن را x'_i تشکیل می دهد. همچنین تابع چگالی و توزیع

ε_{θ_i} ها به ترتیب f و F باشد. اگر $\lim_{n \rightarrow \infty} [(XX')/n] = V$ ، $\beta_{\theta}^* = (F^{-1}(\theta) + \beta_0, \beta_1, \dots, \beta_k)'$ ، $g^2(\theta, F) = \frac{\theta(1-\theta)}{[f(F^{-1}(\theta))]^2}$ ، $f[F^{-1}(\theta)] > 0$ و برآورد آن را با b_{θ}^* نشان دهیم، آنگاه

$$\sqrt{n}(b_{\theta}^* - \beta_{\theta}^*) \xrightarrow{d} N[0, g^2(\theta, F)V] \quad \text{مدل 6:}$$

اکنون با استفاده از این قضیه می توان درباره ضرایب مدل رگرسیون چندک استنباط کرد . بر این اساس، آماره آزمون فرضیه صفر $H_0: A\beta_{\theta} = l$ که در آن A یک ماتریس معلوم با رتبه کامل سطری و l یک بردار از مقادیر معلوم است، عبارت است از:

$$\lambda = (Ab_{\theta} - l)' [A(XX')^{-1}A'] (Ab_{\theta} - l) \hat{g}^{-2} \quad \text{مدل 7:}$$

این بخش را با بررسی نحوه شناسایی شکل توزیع و با استفاده از رگرسیون چندک پایان می دهیم . برای این منظور ابتدا مدلهای متعددی به ازای چندکهای مختلف بر داده ها برازش داده می شود . سپس، برای سطوحی از متغیرهای تشریحی که مد نظر است، چندکها براساس مدل های برازشی پیش بینی میشوند و از مقایسه چندک های پیش بینی شده ، به شکل توزیع در آن سطوح از متغیرهای تشریحی پی می بریم . برای مثال، فرض کنید نه دهک از توزیع متغیر وابسته در سطحی از متغیر تشریحی به صورت حالت اول، دوم یا سوم جدول شماره 1 پیش بینی شده است.

جدول 1

دهک	0/1	0/2	0/3	0/4	0/5	0/6	0/7	0/8	0/9
حالت اول	3	4	6	7	8	13	19	29	37
حالت دوم	3	8	11	15	19	23	27	32	37
حالت سوم	3	15	22	27	29	31	32	34	37

الف) حالت اول به توزیع چوله به راستی اشاره دارد. به فاصله بیش تر میان دهک های بالایی در مقایسه با دهک های پایینی توجه کنید. این نشان می دهد که 40 درصد از مقادیر بزرگ تر متغیر در فاصله وسیعی جای گرفته اند در حالی که 60 درصد باقی مانده از مقادیر کوچکتر، در فاصله کوتاهتری قرار دارند.

ب) حالت دوم نمایان گر 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 می دهد.

آنچه از طریق پیش بینی چندک ها با کمک مدل به دست آمد - در حالتی که تنها یک متغیر تشریحی در مدل وجود داشته باشد، به کمک رسم خطوط مدل های برازشی نیز دست یافتنی است. در واقع، بررسی

فاصله خطوط رگرسیونی چندک‌های مختلف نیز می‌تواند فاصله چندک‌ها را از یکدیگر در سطوح مختلف متغیر تشریحی نشان دهد.

2) کاربرد عملی

این بخش به ارائه یک مثال عددی از رگرسیون چندک اختصاص دارد. یازده مدل رگرسیون چندک همراه با رگرسیون معمولی برای بررسی رابطه رفاه درخواستی افراد (متغیر وابسته 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 نمونه تصادفی 684 نفری از جوانان 18 تا 29 سال تهرانی اختصاص دارد که با روش نمونه‌گیری خوشه‌ای سه مرحله‌ای (Three-stage Cluster Sampling) در سال 1381 گردآوری شده است. شایان ذکر است که رفاه درخواست شد. با شاخصی که حاصل از ترکیب 33 سؤال یک پرسشنامه است، سنجیده شده است. گفتنی است مقدار این شاخص از 0 تا 4 تغییر می‌کند، به طوری که هرچه مقدار آن بیشتر می‌شود بر انتظارات بیشتری نیز دلالت دارد.

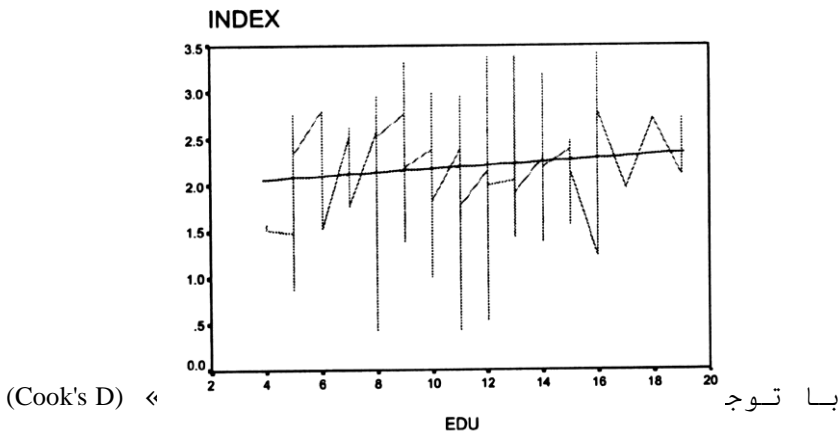
کار را با برازش یک مدل رگرسیون خطی معمولی با روش حداقل مربعات بر داده‌ها آغاز می‌کنیم. برآورد پارامترهای مدل (عرض از مبدأ و ضریب متغیر تعداد سال‌های تحصیل) در جدول شماره 2

دیده می‌شود؛ بر این اساس مدل برآزشی عبارت است از:

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

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 برآزش داده شده در شکل شماره 1 ارائه شده است.

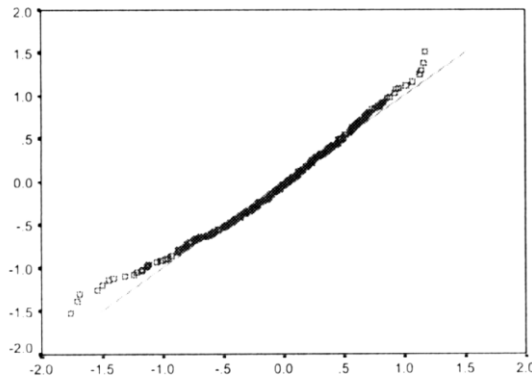
شکل 1: خط رگرسیونی برآزشی و نقاط مشاهده شده



به وجود تعداد زیادی داده دورافتاده اشاره داشت. همچنین ترسیم نمودار Q-Q (Quantile-Quantile Plot) در شکل شماره 2 برای بررسی نرمال بودن توزیع باقی‌مانده‌های مدل، انحراف از توزیع نرمال را نشان می‌دهد. آزمون کلموگروف - اسمیرنوف (Kolmogorov Smirnov Test) نیز فرضیه نرمال بودن این توزیع را رد

می‌کند. تمام این موارد حکایت از نامناسب بودن مدل رگرسیون معمولی دارد.

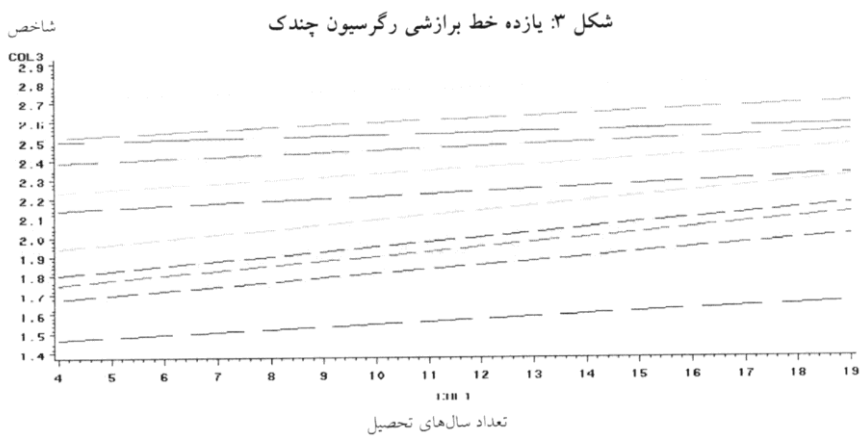
شکل 2: نمودار Q-Q برای باقی‌مانده‌های مدل رگرسیونی برازشی



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 3 و براساس الگوریتمی است که باست و کونکر در 1982 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 چنک 0/1، 0/2، 0/25، 0/3، 0/4، 0/5، 0/6، 0/7، 0/75، 0/8، 0/9 بر داده‌ها برازش داده شد. برآورد پارامترهای این مدل‌ها در جدول شماره 1 ارائه شده است. شکل شماره 3 نیز خطوط برازش داده شده را نشان می‌دهد (خطوط از پایین به بالا مربوط به چنک‌های 0/1 تا 0/9 هستند).



بنابراین، مثلاً، مدل برازش داده شده برای چنک 0/2 عبارت است از

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

که در آن $\hat{Q}_{i,0.2}$ برآورد چنک 0/2 توزیع رفاه

درخواستی، به ازای EDU_i سال تحصیل است (در شکل شماره 3، خط دوم از پایین، مربوط به این مدل است). پس برآورد چندک $0/2$ برای افراد با 19 سال تحصیل برابر است با

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

بنابراین می‌توان انتظار داشت که 20 درصد از افراد با 19 سال تحصیل، دارای شاخص انتظارات کمتر از $2/0075$ و 80 درصد بیش از آن باشند. به همین ترتیب براساس مدل چندک $0/9$ ، انتظار داریم 90 درصد از افراد با 19 سال تحصیل دارای شاخص انتظارات کمتر از $2/8249$ و 10 درصد بیش از آن باشند. پس تقریباً 70 درصد از این افراد دارای شاخصی بین $2/0075$ و $2/8249$ 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 برآزش این مدل‌ها می‌پردازیم:

1-2) یافته‌های حاصل از برآزش

الف) شکل شماره 1 حاکی از آن است که خط رگرسیون نمی‌تواند پیش‌بینی‌کننده 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 پیش بینی کند. برای مثال در این شکل افرادی را با 8 یا 12 سال تحصیل ملاحظه کنید. ب) مثبت بودن شیب خطوط در شکل شماره 3 یعنی ضریب تعداد سال های تحصیل، نشان دهنده رابطه مستقیم بین متغیر وابسته و تشریحی است. بنابراین، با افزایش تعداد سال های تحصیل، مقدار هر یک از یازده چنک شاخص انتظارات نیز افزایش می‌یابد. بر اساس چنک های برآزشی، سال های تحصیل بر چنک های پایینی بیش تر از چنک های بالایی اثر دارد.

ج) برای افراد با تحصیلات بیش تر، فاصله کم چنک های بالایی (چنک 0/9، 0/8، 0/75 و 0/6) در مقایسه با چنک های پایینی نشان می‌دهد که فشردگی داده ها در بخش بالایی زیاد است. به عبارت دیگر، در سمت راست توزیع شرطی، فشردگی بیش تری در مقایسه با سمت چپ وجود دارد. بنابراین، با افزایش تعداد سال های تحصیل، توزیع شرطی شاخص مطلوب، چوله به چپ می‌شود.

د) می‌توان پیش بینی کرد که مثلاً 70 درصد (از دهک 0/2 تا 0/9) افراد با 19 سال تحصیل دارای شاخصی بین 2/0075 و 2/8249 هستند، در حالی که این فاصله برای افرادی با 4 سال تحصیل از 1/67 تا 2/7184 است.

جدول 1: مشخصات مدل‌های رگرسیونی

مجموع قدرمطلق موزون خطاه ۱	برآورد پارامترهای مدل		مدل
	تعداد سال‌های تحصیل	عرض از مبدأ	
—	0/0189	1/99	رگرسیون معمولی
66	0/0129	1/42	0/1
99	0/0225	1/58	0/2
110	0/025	1/65	0/25
118	0/025	1/7	0/3
127	0/025	1/87	0/4
128	0/0129	2/09	0/5
122	0/0166	2/16	0/6
108	0/0113	2/34	0/7
98	0/0063	2/47	0/75
86	0/0125	2/47	0/8
52	0/0071	2/69	0/9

(3) 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 داده های دورافتاده و انحراف از متعادل بودن و هم چنین 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.

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.