# کاربرد رگرسیون چندك در شناسایي شکل توزیع رفاه مورد انتظار جوانان

محمد بامني مقدم*، علي‌رضا خوشگویان فرد**

هدف اصلي این مقاله، آشنا کردن خواننده با کاربرد رگرسیون چندك در تحلیل داده هاست. رگرسیون چندك، رابطة چندك دلخواهي از توزیع متغیر وابسته را با متغیرهاي تشریحي از طریق مدل آماري تبیین مي کند. در این مقاله، مدل رگرسیون چندك معرفي و به شیوة برآورد پارامترها اشاره مي شود؛ به قابلیت شناسایي شکل توزیع که مدل رگرسیون معمولي (میانگین شرطي) آن را دارا نیست، تأکید مي‌شود؛ در پایان با یك مثال عددي از داده‌هاي رفاه که در آن براساس یك نمونة تصادفي 684 نفر از جوانان 18 تا 29 سال تهراني انتخاب شده است، تلاش شده است کاربرد رگرسیون چندك براي رفاه تشریح شود و رابطة رفاه مطلوب جوانان (متغیر وابسته مدل) با تعداد سال هاي تحصیل آنان (متغیر تشریحي مدل) تحت بررسي قرار گیرد.

* دکتراي آمار، عضو هیأت علمي دانشگاه علامه طباطبایي <bamenimoghadam@atu.ac.ir>

** فوق لیسانس آمار اقتصادي اجتماعي، مدیر هماهنگي پژوهش‌هاي مرکز تحقیقات صداوسیما

**كليد واژه ها: برآورد حداقل مربعات، حداقل قدر مطلق انحرافات، رگرسيون چندك، رفاه، معيار d كوك، ميانگين شرطي**

**تاريخ دريافت مقاله: 83/7/27 تاريخ پذيرش مقاله: 83/10/28**

**مقدمه**

ابعاد گوناگون پديده ها در آمار، خود را قالب متغيرهاي تصادفي نشان مي دهند و مطالعه آن‌ها با تعيين توزيع (Distribution) آن‌ها ميسر مي شود. براي مثال، وزن نوزادان به عنوان يك پديده مي تواند متغيري تصادفي محسوب شود و در صورت مشخص بودن توزيع آن، تعيين اين كه نوزادي داراي وزن طبيعي است، امكان‌پذير مي‌شود. توزيع متغير تصادفي تمام اطلاعات لازم در خصوص آن را به دست مي دهد، به طوري كه مي توان پديده ها را تفسير يا پيش بيني كرد. به عنوان مثال، با داشتن توزيع وزن نوزادان قادريم پيش‌بيني كنيم كه تا چ ه اندازه ممكن است كودكاني كم‌وزن داشته باشيم.

در اين راستا، معيارهاي آماري گوناگوني وجود دارد كه هر يك از توزيع متغير تصادفي اطلاع متفاوتي در اختيار مي گذارد. براي مثال، واريانس، از نحوة پراكندگي و نما، از قلة آن اطلاعي فراهم مي كند. برخي از معيارها مانند

برجستگي (Kurtosis: کشيدگي يا پخي) و چولگي (Skewness : کجي) نيز به شکل توزيع اختصاص دارند . شکل توزيع در متغيرهايي مانند درآمد يا هوش از اهميت ويژه اي برخوردار است . مثلاً اگر توزيع درآمد در جامعه اي چولگي زيادي به چپ داشته باشد، حکايت از وجود افراد محرومي با درآمد بسيار کم دارد. هم‌چنين اگر توزيع نمرات هوش در دانشگاهي متقارن با دم هايي کوتاه باشد، نشان مي‌دهد که در مجموع، هوش دانشجويان آن دانشگاه معمولي است و افراد باهوش و کم هوش در آن دانشگاه نسبتاً برابرند.

با توجه به آن‌چه گذشت دور از انتظار نيست اگر ادعا کنيم که استنباط هاي آماري (Statistical Inference) همگي اطلاعي از توزيع يك متغير تصادفي به دست مي‌دهند. براي مثال، آزمون F در تحليل واريانس، مقايسه‌اي بين ميانگين چند توزيع است؛ محاسبه ضريب همبستگي پيرسون (Pearson) و آزمون مربوط به آن اطلاعي از يك توزيع دو متغيره را فراهم مي‌کند؛ مدل هاي رگرسيوني که در ادامه به آن مي‌پردازيم نيز مانند ديگر روش ها براي بررسي خصوصيات خاصي از توزيع يك متغير تصادفي به کار مي‌روند.

مدل رگرسيون معمولي به تحليل گر کمك مي‌کند تا رابطة ميانگين توزيع متغير تصادفي Y را با

تعدادي متغير تشريحي بررسي كند. براي روشن شدن مطلب، يك مدل رگرسيون خطي را تنها با يك متغير تشريحي به اين صورت در نظر مي‌گيريم:

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

در اين مدل رگرسيوني $\varepsilon_i$ها، متغيرهاي تصادفي، $\alpha$ و $\beta$، پارامترهاي نامعلوم كه بايد برآورد شوند و $x_i$ها مقادير معلومي از متغير تشريحي هستند. در صورتي كه $E(\varepsilon_i)=0$ باشد، آن‌گاه مي‌توان مدل شمارة 1 را به صورت دوم بازنويسي كرد:

$$E(Y_i) = \alpha + \beta x_i \qquad \text{:مدل2}$$

كميت $E(Y_i)$ را ميانگين شرطي (Conditional Mean) متغير تصادفي Y مي‌نامند و به همين دليل آن را با $E(Y/x_i)$ نيز نشان مي دهند. بنابراين، مدل شمارة 2 بيان مي‌كند كه ميانگين هاي توزيع Y در سطوح مختلف متغير تشريحي در امتداد يك خط راست قرار دارند. به عبارت ديگر، متغير تصادفي Y در هر سطح از متغير تشريحي داراي توزيعي است كه ميانگين هاي اين توزيع ها روي يك خط راست جاي گرفته اند. يكي از حالاتي كه در آن چنين رابطه اي قطعاً برقرار مي‌شود زماني است كه دو متغير X و Y داراي توزيع نرمال دو متغيره باشند (باز هم صحبت از توزيع

است!).

از آن جا که میانگین، یکي از معیارهاي تمرکز است، آگاهي از آن به تنهایي نمي تواند اطلاعات کاملي از شکل توزیع به همر اه داشته باشد. با توجه به این واقعیت، رگرسیون معمولي نیز ممکن است نتواند اطلاعات کافي دربارة شکل توزیع متغیر تصادفيِ تحت مطالعه را ـ در سطوح مختلف متغیر تشریحي ـ به دست دهد. چندكها (Quantiles) معیارهاي دیگري براي توزیع هستند که «در کنار هم» مي‌توانند شکل توزیع را جامع‌تر به تصویر بکشند. براي مثال، اگر دهك هاي توزیعي تقریباً داراي فاصله برابري از یکدیگر باشند، انتظار داریم توزیع «نسبتاً» هموار یا یك‌نواختي داشته باشیم. هم‌چنین اگر دهك هاي بالایي داراي فاصله زیاد و دهك‌هاي پاییني داراي فاصله کمي از یکدیگر باشند، توزیع به سمت راست چوله خواهد بود. اکنون اگر مانند رگرسیون معمولي که براي میانگین به کار مي‌رود، یك شیوة رگرسیوني براي چندك‌ها وجود داشته باشد، قادر خواهیم بود شکل توزیع را در سطوح مختلف متغیرهاي تشریحي به دست آوریم. این همان هدفي است که رگرسیون چندك دنبال مي‌کند.

**1) معرفي رگرسیون چندك**

همان‌طور که در بخش قبل اشاره شد، مدل رگرسیون معمولي، مانند مدل شمارة 2، براي میانگین شرطي

برازش داده مي‌شود. مدل رگرسيون چندك با ايده اي مشابه براي چندك هاي شرطي (Conditional Quantile) به كار مي‌رود. مانند رگرسيون معمولي (ميانگين)، كاربردهايي نظير بررسي رابطه متغيرهاي تشريحي با چندك‌ها و هم‌چنين پيش‌بيني آن‌ها براي اين نوع از رگرسيون نيز امكان پذير است. با وجود اين، شايد مهم‌ترين كاربرد رگرسيون چندك شناسايي شكل توزيع متغير وابسته مدل در سطوح گوناگون متغيرهاي تشريحي باشد؛ اين كار با برازش مدل‌هاي رگرسيوني متعدد، به ازاي چندك هاي مختلف بر يك مجموعه‌داده، صورت مي‌گيرد.

براي ارائة تعريف دقيقي از مدل رگرسيون چندك ام $\theta \in (0,1)$، ابتدا حالت سادة آن را در نظر مي‌گيريم. مدل شمارة 1 را با شرط $\varepsilon_i \overset{iid}{\sim} F(\cdot)$ (تابع F به يك توزيع دلخواه اشاره دارد) در نظر بگيريد. هدف ما يافتن مدلي است كه مثلاً رابطة چندك اول (و نه ميانگين) توزيع Y را با متغير X نشان دهد. در اين صورت، مدل براي چندك ام $\theta \in (0,1)$ متغير Y كه با $Q_\theta(Y \mid x_i)$ نشان داده مي‌شود، عبارت است از:

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

تابع فوق، به ازاي هاي $\theta \in (0,1)$ مختلف، دسته‌اي از خطوط موازي را به دست خواهد داد كه داراي عرض

از مبدأهاي متفاوتي هستند . در صورتي كه $F(\cdot)$ همان توزيع نرمال (يا هر توزيع متقارن ديگري ) باشد، به ازاي $\theta = 0.5$، مدل شمارة 3 همان مدل شمارة 2 خواهد بود، زيرا $F^{-1}(0.5) = 0$ ممكن است به يك تغيير مكان نياز باشد . اكنون به تعريف كلي مدل رگرسيون چندك مي پردازيم. براي اين منظور ، فرض كنيد $Y_i = x_i'\beta_\theta + \varepsilon_{\theta i}$ و

$$Q_\theta(Y \mid x_i') = x_i'\beta_\theta \qquad \text{i=1,...,n} \qquad \text{: 4 مدل}$$

كه در آن $x_i' = (1, x_{i1}, ..., x_{ik})$ و $\beta_\theta' = (\beta_0, \beta_1, ..., \beta_k)$ به ترتيب برداري از مقادير معلوم و پارامترهاي نامعلوم بوده و $\varepsilon_{\theta i}$ يك متغير تصادفي مشاهده نشدني است. هم‌چنين، $Q_\theta(Y \mid x_i)$ نمايانگر چ ندك شرطي ام $\theta \in (0,1)$ توزيع Y است، بنابراين $Q_\theta(\varepsilon_\theta \mid x_i) = 0$. مدل شمارة 4 را با شرايط گفته شده، مدل رگرسيون خطي چندك θ ام مي‌ناميم.

شيوة برآورد پارامترهاي مدل رگرسيون معمولي بر حداقل كردن مربع باقيمانده هاي (انحرافات) مدل مبتني است كه ر وش حداقل مربعات (Least Squares) ناميده مي‌شود. در اين روش، منحني رگرسيوني به گونه‌اي برازش داده مي شود كه در مجموع، فاصلة نقاط از آن به حداقل برسد . در رگرسيون چندك برخلاف رگرسيون معمولي از حداقل كردن مجموع قدر مطلق موزون براي برآورد پارامترهاي مدل استفاده

مي‌شود كه به آن روش حداقل قدر مطلق انحرافات (Least Absolute Deviations) يا LAD گفته مي‌شود. گفتني است كه استفاده از روش LAD كه در مدل رگرسيون چندك به كار مي رود، داراي پيشينه اي طولاني است. در ميانة قرن هجدهم، بوسكويچ (Boscovich) يك مدل خطي دو متغيره را بر اي بررسي بيضوي بودن كره زمين از طريق كمينه كردن قدر مطلق خطاها به كار برد. به دنبال آن، لاپلاس (Laplace) برآورد ضريب زاوية مدل رگرسيوني بوسكويچ را به طور دقيق معرفي و توزيع مجانبي آن را به دست آورد. ظاهراً، اجورث (F. Y. Edgeworth) اولين كسي است كه مدل رگر سيوني ميانه با چند متغير تشريحي را در حالت كلي بررسي كرد. توسعة رگرسيون ميانه براي هر چندك دلخواه نيز به كوشش كانوكر و باست (Koenker and Bassett) در 1978 صورت گرفت. هدف آن‌ها برآورد بردار پارامترهاي $\beta'_\theta = (\beta_0, \beta_1, ..., \beta_k)$ در مدل شمارة 4 بود كه براي اين منظور تابع زياني كه در پي مي‌آيد (قدر مطلق باقي‌مانده‌ها يا انحرافات موزون ) نسبت به عناصر $\beta_\theta$ كمينه مي‌شود:

مدل 5:

$$\varphi_\theta(\beta_\theta) = \sum_i w(\theta) \mid y_i - x'_i\beta_\theta$$

در اين تابع زيان $w(\theta) = \begin{cases} \theta & Y_i \le X'_i\beta_\theta \\ 1-\theta & Y_i > X'_i\beta_\theta \end{cases}$.

موزون كردن قدر مطلق باقيمان ده‌ها در تابع فوق باعث مي شود تا خط برازشي به گونه اي باشد كه

$\theta^*100\%$ داده‌ها تقریباً زیر آن و باقي آن ها بالاي خط قرار گیرند. کمینه کردن رابطة فوق و یافتن برآورد LAD پارامترها با استفاده از روش هاي برنامه‌ریزي خطي و از طریق بسته هاي نرم افزاري صورت مي‌گیرد. در ادامه به ویژگي هاي برآورد LAD اشاره مي‌شود.

**1-1) ویژگي‌هاي حداقل قدر مطلق انحرافات (LAD)**

الف) در حالت خاص که مدل تنها شامل عرض از مبدأ و $\theta=0.5$ است، کمینه کردن رابطة 5 منجر به کمینه کردن عبارت $\sum|y_i-\beta_0|$ مي‌شود که در این صورت برآورد $\beta_0$ همان میانه داده‌ها خواهد بود.

ب) برخلاف روش حداقل مربعات، روش حداقل قدر مطلق انحرافات نسبت به داده هاي دور افتاده (Outliers) استوار (Robust) است. این ویژگي ناشي از آن است که برخلاف اهمیت اندازه باقي مانده‌ها در روش حداقل مربعات، در این روش تنها به علامت باقي‌مانده‌ها توجه نمي شود. بنابراین نه تعداد باقي‌مانده‌هایي که بیش‌تر (مثبت) یا کم‌تر (منفي) از چندك مورد نظرند و نه مقدار بزرگي آن ها در برآوردها اثرگذار است. پس، داده هاي دورافتاده که تأثیر خود را از طریق بزرگي باقي مانده‌ها نشان مي‌دهند، نمي‌توانند برآوردهاي LAD را متأثر سازند.

ج) شكل بسته اي براي برآورد پارامترهاي اين مدل وجود ندارد و از روش هاي عددي براي برآورد آن‌ها استفاده مي‌شود. هم‌چنين، جواب‌هاي نهايي مدل رگرسيون چندك مي‌تواند يكتا نباشد. البته يافتن جواب يكتا با انتخاب يك معيار مناسب امكان‌پذير است. براي مثال، مسئلة يافتن ميانة 10 عدد را به ياد بياوريد كه براي اين منظور، ميانگين عدد پنجم و ششم به عنوان ميانه در نظر گرفته مي شد. اين در حالي است كه كلية اعداد بين عدد پنجم و ششم مي‌توانند به عنوان ميانه انتخاب شوند. در واقع، با يك قرارداد (معيار)، ميانة اين اعداد به صورت منحصر به فردي تعيين مي‌شود.

د) وقتي $\varepsilon_{\theta i}$ها متغيرهاي تصادفي iid باشند، خطوط رگرسيوني به ازاي چندك هاي مختلف، موازي خواهند بود (مدل شمارة 3 را به ياد بياوريد).

در رگرسيون چندك نيز مانند رگرسيون معمولي (ميانگين شرطي) مي‌توان استنباط كرد. پاول و بوچنسكي نشان داده اند كه برآورد LAD پارامترها، سازگار و به طور مجانبي نرمال است (Powell, 1989., Buchinsky, 1998). موضوع اخير را براي حالت خاص كه $\varepsilon_{\theta i}$ها هم‌توزيع هستند در قضيه زير بيان مي‌كنيم.

**2-1) قضيه**

فرض كنيد X يك ماتريس $n\times k$ باشد كه سطر iام آن را $x_i'$ تشكيل مي‌دهد. هم‌چنين تابع چگالي و توزيع

$\varepsilon_{\theta i}$ها به ترتیب f و F باشد. اگر $\underline{\lim}_{n\to+\infty}[(XX')/n]=V$ ، $f\left[F^{-1}(\theta)\right]>0$ ، $g^2(\theta,F)=\frac{\theta(1-\theta)}{[f(F^{-1}(\theta)]^2}$ ، $\beta^*_\theta=(F^{-1}(\theta)+\beta_0,\beta_1,...,\beta_k)'$ و برآورد آن را با $b^*_\theta$ نشان دهیم، آنگاه

$$\sqrt{n}(b^*_\theta-\beta^*_\theta)\xrightarrow{d}N[0,g^2(\theta,F)V] \qquad \text{مدل 6:}$$

اکنون با استفاده از این قضیه مي توان دربارة ضرایب مدل رگرسیون چندك استنباط کرد . بر این اساس، آمارة آزمون فرضیة صفر $H_0:A\beta_\theta=l$ که در آن A یك ماتریس معلوم با رتبة کامل سطري و l یك بردار از مقادیر معلوم است، عبارت است از:

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

این بخش را با بررسي نحوة شناسایي شکل توزیع و با استفاده از رگرسیون چندك پایان مي دهیم. براي این منظور ابتدا مد لهاي متعددي به ازاي چندك‌هاي مختلف بر داده ها برازش داده مي شود. سپس، براي سطوحي از متغیرهاي تشریحي که مدّ نظر است، چندك ها براساس مدل هاي برازشي پیش بیني مي‌شوند و از مقایسة چندك هاي پیش‌بیني شده , به شکل توزیع در آن سطوح از متغیرهاي تشریحي پي مي‌بریم. براي مثال، فرض کنید نه دهك از توزیع متغیر وابسته در سطحي از متغیر تشریحي به صورت حالت اول، دوم یا سوم جدول شمارة 1 پیش‌بیني شده است.

**جدول 1**

| دهك | 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 توزيع چوله به راستي اشاره دارد. به فاصلة بيش تر ميان دهك هاي بالايي در مقايسه با دهك هاي پاييني توجه كنيد . اين نشان مي‌دهد كه 40 درصد از مقادير بزرگ تر متغير در فاصله وسيعي جاي گرفته اند در حالي كه 60 درصد باقي‌مانده از مقادير كوچك‌تر، در فاصلة كوتاه‌تري قرار دارند.

ب) حالت دوم نمايان گر يك توزيع نسبتاً هموار است، زيرا فاصله دهك‌ها تقريباً برابر است

ج) حالت سوم وضعيتي عكس حالت اول دارد ؛ يعني توزيع چوله به چپي را نشان مي‌دهد.

آن‌چه از طريق پيش بيني چندك ها با كمك مدل به دست آمد ـ در حالتي كه تنها يك متغير تشريحي در مدل وجود داشته باشد، به كمك رسم خطوط مدل هاي برازشي نيز دست يافتني است . در واقع، بررسي

فاصلة خطوط رگرسيوني چندك‌هاي مختلف نيز مي‌تواند فاصلة چندك‌ها را از يكديگر در سطوح مختلف متغير تشريحي نشان دهد.

**2) كاربرد عملي**

اين بخش به ارائة يك مثال عددي از رگرسيون چندك اختصاص دارد. يازده مدل رگرسيون چندك همراه با رگرسيون معمولي براي بررسي رابطة رفاه درخواستي افراد (متغير وابستة مدل) با تعداد سال هاي تحصيل آنان (متغير تشريحي مدل) به كار مي رود. رفاه مطلوب و تعداد س ال‌هاي تحصيل به ترتيب با INDEX و EDU نشان داده خواهند شد. داده‌ها به يك نمونة تصادفي 684 نفري از جوانان 18 تا 29 سال تهراني اختصاص دارد كه با روش نمونه‌گيري خوشه‌اي سه مرحله اي (Three-stage Cluster Sampling) در سال 1381 گردآوري شده است. شايان ذكر است كه رفاه درخواست شد. با شاخصي كه حاصل از تركيب 33 سؤال يك پرسشنامه است، سنجيده شده است. گفتني است مقدار اين شاخص از 0 تا 4 تغيير مي‌كند، به طوري كه هرچه مقدار آن بيش‌تر مي‌شود بر انتظارات بيش‌تري نيز دلالت دارد.

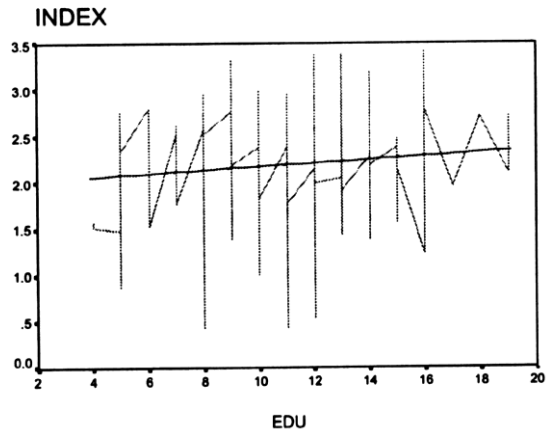
كار را با برازش يك مدل رگرسيون خطي معمولي با روش حداقل مربعات بر داده ها آغاز مي كنيم. برآورد پارامترهاي مدل (عرض از مبدأ و ضريب متغير تعداد سال هاي تحصيل) در جدول شمارة 2

ديده مي‌شود؛ بر اين اساس مدل برازشي عبارت است از:

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

كه در آن $E(\hat{INDEX}_i)$ برآورد ميانگين توزيع رفاه درخواستي به ازاي $EDU_i$ سال تحصيل است. نمودار پراكنش (Scatter Plot) داده‌ها همراه با خط برازش داده شده در شكل شمارة 1 ارائه شده است.

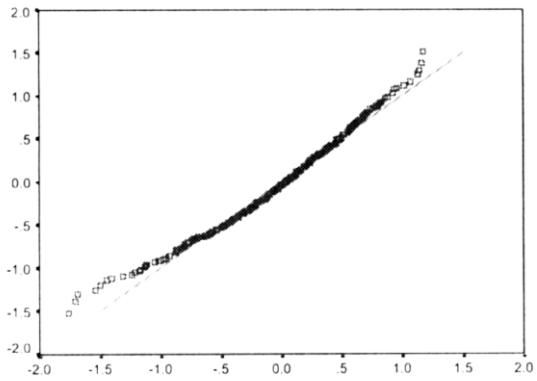
**شكل 1: خط رگرسيوني برازشي و نقاط مشاهده شده**




با توجه به وجود تعداد زيادي داده دورافتاده اشاره (Cook's D) » داشت. هم‌چنين ترسيم نمودار Q-Q (Quantile-Quantile Plot) در شكل شمارة 2 براي بررسي نرمال بودن توزيع باقي‌مانده‌هاي مدل، انحراف از توزيع نرمال را نشان مي‌دهد. آزمون كلموگرف ـ اسميرنف (Kolmogorov Smirnov Test) نيز فرضية نرمال بودن اين توزيع را رد

مي‌كند. تمام اين موارد حكايت از نامناسب بودن مدل رگرسيون معمولي دارد.

**شكل 2: نمودار Q-Q براي باقي‌مانده‌هاي مدل رگرسيوني برازشي**

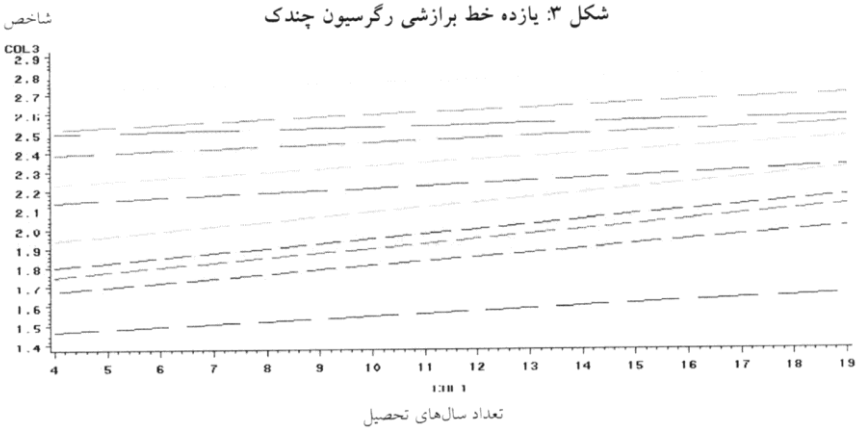



اكنون به برازش مدل هاي رگرسيون چندك به ازاي يازده چندك مختلف به كمك برنامه اي كه در محيط IML از نرم افزار SAS نوشته شده است، مي پردازيم (برنامه در پيوست ارائه شده است). اين برنامه براي مدل 3 و براساس الگوريتمي است كه باست و كونكر در 1982 تهيه كرده‌اند (Bassett & Koenker, 1982). گفتني است وقتي اين چندك، همان ميانه باشد، الگوريتم سريع‌تري از طرف مدسن و نيلسن پيشنهاد شده است كه در SAS/IML در قالب روال LAV (Routine)

استفاده مي‌شود (Madsen & Nielsen, 1993).

در مجموع، يازده مدل به ازاي يازده چندك 0/1، 0/2، 0/25، 0/3، 0/4، 0/5، 0/6، 0/7، 0/75، 0/8، 0/9 بر داده ها برازش داده شد. برآورد پارامترهاي اين مدل‌ها در جدول شمارة 1 ارائه شده است. شكل شمارة 3 نيز خطوط برازش داده شده را نشان مي‌دهد (خطوط از پايين به بالا مربوط به چندك‌هاي 0/1 تا 0/9 هستند).

شكل ۳: يازده خط برازشی رگرسيون چندک




بنابراين، مثلاً، مدل برازش داده شده براي چندك 0/2 عبارت است از

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

كه در آن $\hat{Q}_{i,0.2}$ برآورد چندك 0/2 توزيع رفاه

درخواستي، به ازاي $EDU_i$ سال تحصيل است (در شكل شمارة 3، خط دوم از پايين، مربوط به اين مدل است). پس برآورد چندك 0/2 براي افراد با 19 سال تحصيل برابر است با

$$2.0075=1.58+0.0225\times19$$

بنابراين مي‌توان انتظار داشت كه 20 درصد از افراد با 19 سال تحصيل، داراي شاخص انتظارات كم‌تر از 2/0075 و 80 درصد بيش از آن باشند. به همين ترتيب براساس مدل چندك 0/9، انتظار داريم 90 درصد از افراد با 19 سال تحصيل داراي شاخص انتظارات كم‌تر از 2/8249 و 10 درصد بيش از آن باشند. پس تقريباً 70 درصد از اين افراد داراي شاخصي بين 2/0075 و 2/8249 هستند. توجه كنيد كه چنين تحليل هايي تنها با مدل هاي رگرسيون چندك قابل انجام است و مدل هاي رگرسيون معمولي چنين قابليت‌هايي را ندارند. اكنون به يافته هاي حاصل از برازش اين مدل‌ها مي‌پردازيم:

**2-1) يافته‌هاي حاصل از برازش**

الف) شكل شمارة 1 حاكي از آن است كه خط رگرسيون نمي تواند پيش بيني‌كنندة مناسبي براي شاخص انتظارات باشد. زيرا با توجه به پراكندگي زياد داده ها در برخي از سطوح تعداد سال هاي

تحصيل، ميانگين نمي‌تواند اين شاخص را براي اين سطوح به خوبي پيش بيني كند. براي مثال در اين شكل افرادي را با 8 يا 12 سال تحصيل ملاحظه كنيد.

ب) مثبت بودن شيب خطوط در شكل شمارة 3 يعني ضريب تعداد سال هاي تحصيل، نشان دهندة رابطة مستقيم بين متغير وابسته و تشريحي است. بنابراين، با افزايش تعداد سال هاي تحصيل، مقدار هر يك از يازده چندك شاخص انتظارات نيز افزايش مي‌يابد. بر اساس چندك هاي برازشي، سال هاي تحصيل بر چندك هاي پايييني بيش‌تر از چندك هاي بالايي اثر دارد.

ج) براي افراد با تحصيلات بيش تر، فاصلة كم چندك‌هاي بالايي (چندك 0/9، 0/8، 0/75 و 0/6) در مقايسه با چندك‌هاي پايييني نشان مي‌دهد كه فشردگي داده‌ها در بخش بالايي زياد است. به عبارت ديگر، در سمت راست توزيع شرطي، فشردگي بيش تري در مقايسه با سمت چپ وجود دارد. بنابراين، با افزايش تعداد سال هاي تحصيل، توزيع شرطي شاخص مطلوب، چوله به چپ مي‌شود.

د) مي‌توان پيش‌بيني كرد كه مثلاً 70 درصد (از دهك 0/2 تا 0/9) افراد با 19 سال تحصيل داراي شاخصي بين 2/0075 و 2/8249 هستند، در حالي كه اين فاصله براي افرادي با 4 سال تحصيل از 1/67 تا 2/7184 است.

**جدول 1: مشخصات مدلهاي رگرسيوني**

| مدل | | برآورد پارامترهاي مدل | | مجموع قدرمطلق موزون خطاها |
|---|---|---|---|---|
| | | عرض از مبدأ | تعداد سالهاي تحصيل | |
| رگرسيون معمولي | | 1/99 | 0/0189 | ـــ |
| رگرسيون چندك | 0/1 | 1/42 | 0/0129 | 66 |
| | 0/2 | 1/58 | 0/0225 | 99 |
| | 0/25 | 1/65 | 0/025 | 110 |
| | 0/3 | 1/7 | 0/025 | 118 |
| | 0/4 | 1/87 | 0/025 | 127 |
| | 0/5 | 2/09 | 0/0129 | 128 |
| | 0/6 | 2/16 | 0/0166 | 122 |
| | 0/7 | 2/34 | 0/0113 | 108 |
| | 0/75 | 2/47 | 0/0063 | 98 |
| | 0/8 | 2/47 | 0/0125 | 86 |
| | 0/9 | 2/69 | 0/0071 | 52 |

## 3) نتيجه‌گيري

اين مقاله نشان داد كه رگرسيون چندك نه تنها مي‌تواند جانشين مناسبي براي رگرسيون ميانگين باشد (با جانشين كردن ميانه به جاي ميانگين )، بلكه در برخي از حالات، اطلاعات بيشتري (شكل توزيع) را در مقايسه با رگرسيون ميانگين در

اختيار تحليلگر قرار مي‌دهد. در بخش قبل ديده شد كه رگرسيون ميانگين به سبب وجود داده هاي دورافتاده و انحراف از متعادل بودن و هم چنين پراكندگي زياد متغير پاسخ در برخي از سطوح متغير تشريحي، از اعتبار لازم برخوردار نبود؛ در حالي كه رگرسيون چندك، يافته‌هاي مفيدي را به دست داد.

## منابع
## References


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.