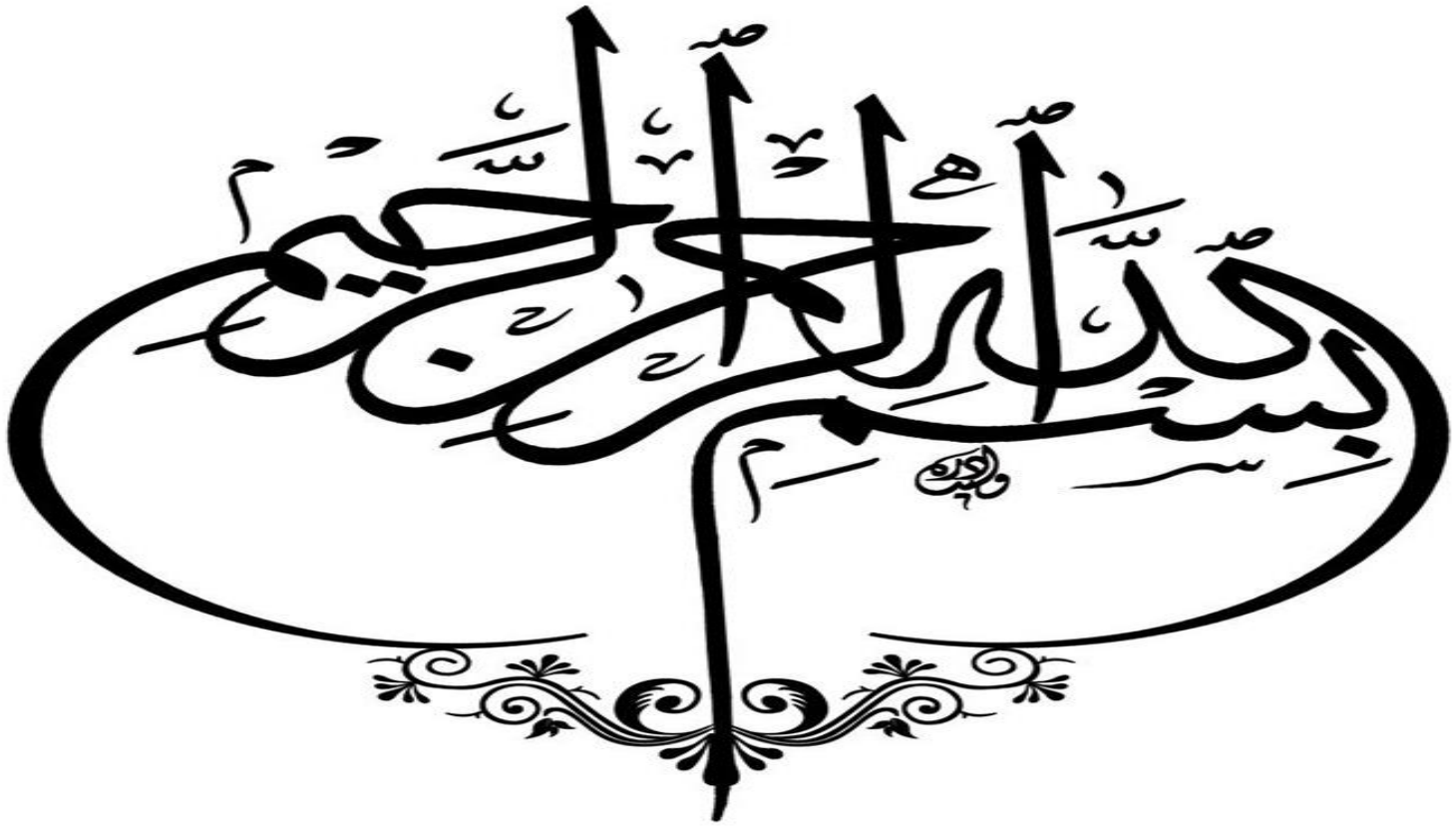

بسم الله الرحمن الرحيم

ارائه دهندگان:
مهسا حاتمی
زکیه فرجی

درس مدل های خطی(۱)

مقدمه:

برای اولین با رگرسیون چندکی را کنکر و باکس در سال ۱۹۷۸ معرفی کردند. پس از آن این روش به یک روش پر کاربرد و مهم برای مطالعه توزیع شرطی کامل متغیر پاسخ تبدیل شده است. رگرسیون چندکی توابع چندکی شرطی را به عنوان توابعی از متغیرهای پیش بینی کننده برآورد می کند.

هدف اصلی از بکارگیری رگرسیون چندکی ارائه مدلی است که امکان دخالت متغیرهای مستقل نه تنها در مرکز داده ها بلکه در تمام قسمت های توزیع به ویژه در دنباله های ابتدایی و انتهایی را فراهم کند.

از رگرسیون خطی ساده تا رگرسیون چندکی

رگرسیون برای مطالعه‌ی رابطه‌ی بین متغیر پاسخ Y و یک یا تعدادی متغیر پیش بینی کننده‌ی X به کار می رود. در رگرسیون خطی ساده میانگین شرطی $\mu(x)=E(Y \mid X=x)$ بر حسب X مدل بندی می شود. به عنوان مثال در رگرسیون ساده فرض می شود که

$$\mu(x)=\beta_0+\beta_1 x$$

خواص رگرسیون چندکی:

رگرسیون چندکی، چندک های شرطی $q_p(x)$ را بر حسب X مدل بندی می کند.
رگرسیون چندکی برای چندک های مختلف تصویر کاملتری نسبت به رگرسیون خطی ارائه می دهد.
مدل بندی چندک ها نسبت به مشاهدات پرت پایاتر از مدل بندی ساده است.
به علاوه تجزیه و تحلیل اثر متغیر پیش بینی کنند روی چندک های مختلف متغیر پاسخ تصویر آشکارتری از رابطه متغیر پاسخ با متغیر های پیش بینی کننده ارائه می دهد.

معروف ترین چندک میانه است. میانه جمعیت عددی است که توزیع را به ۲ قسمت مساوی تقسیم می کند.
به این معنی که برای متغیر تصادفی Y میانه عددی است مانند m که در شرط زیر صدق می کند.

$$p(y \geq m) \geq \frac{1}{2} \qquad p(y \leq m) \geq \frac{1}{2}$$

برای مثال توزیع متغیر حقوق کارگران را در نظر بگیرید. این توزیع چوله به راست است زیرا معمولا به طور نسبی تعداد کمی از افراد حقوق بالا دریافت می کنند بنابراین میانه معیار مناسب تری برای میزان تمرکز حقوق نسبت به میانگین ارائه می دهد.

مدل رگرسیون چندکی

مدل رگرسیون چندکی $y_i = x_i'\beta_\tau + \varepsilon_{\tau i}$ را در نظر بگیرید. بردار پرامتر β_τ معمولا از روش حداقل قدرمطلق انحرافات (LAD) انجام می‌گیرد. در مدل فوق x_i^τ ها برابرند با $x_i^\tau = (x_{i1}, x_{i1}, \dots, x_{i1})$ و β_τبرداری از پارامتر های نامعلوم است. ε یک متغیر تصادفی مشاهده نشده می باشد. فرض کنید $Q_\tau(\varepsilon_i \,|x_i) = 0$ ، چندک شرطی τ ام توزیع متغیر های مستقل X به صورت زیر تعریف می شود:

$$Q_\tau(y_i \,|x_i) = x_i'\beta_\tau$$

برآورد یابی β_τ از طریق مینیمم کردن عبارت زیر صورت می گیرد:

$$\beta(\tau) = \min[\sum \tau|y_i - x_i'\beta| + \sum(1-\tau)|y_i - x_i'\beta|] = min\sum_{i=1}^{n}\rho_\tau(y_i - x_i'\beta)$$

$$i\in\{i: y_i \geq x_i'\beta\}$$

که $\rho_\tau(u)$ تابع مقادیر مطلق شیب است و به صورت زیر تعریف می شود:

$$\rho_\tau(u) = u(\tau - I(u<0))$$

که در آن $I_A(u)$ تابع نشانگر معمولی روی مجوعه A است و $0 < \tau < 1$ چندک τ ام مقدار θ ای است که عبارت $E(\rho_\tau(y-\theta))$ را مینیمم کند.

مزایای رگرسیون چندکی (QR)

۱- اطلاعات بیشتری از توزیع شرطی متغیر پاسخ ارائه می دهد.
۲- نیازی به شرط نرمال بودن [redacted] ندارد
۳- نسبت به داده های دور افتاده (پرت) نیرومند است.
۴- چگونگی تاثیر متغیرهای مستقل روی مکان و مقیاس و شکل توزیع را نشان می دهد.
۵- در واقع بدون محدودیت ها و مفروضیات رگرسیون خطی، ورلی را می توان ارائه کرد که نسبت به داده های پرت استوارتر است و حتی اگر فرض ناهمسانی واریانس ها نیز وجود داشته باشد این مدل برآورد مناسبی از ضرائب رگرسیون را ارائه خواهد کرد.

معایب رگرسیون چندکی

۱- اشکال عمده رگرسیون چندکی این است که به منظور بدست آوردن ماتریس کواریانس مجانبی برآوردگرها نیاز به برآورد چگالی رگرسور که اغلب وقت گراست داریم.

۲- به دلیل نبودن نرم افزارهای آماری در گذشته امکان محاسبه ی دستی رگرسیون چندکی بسیار سخت بوده است که البته با اومدن نرم افزارهای آماری این مشکل حل شده است.

کاربرد رگرسیون چندکی:

رگرسیون چندکی در علم پزشکی، آنالیز بقا و اقتصاد کاربرد دارد.

مثال۱

فرض کنید y یک متغیر تصادفی گسسته باشد که مقادیر $1,2,\ldots,9$ را با احتمال مساوی اختیار می کند، قصد داریم میانه y را به دست آوریم. با استفاده از فرمول چندک نمونه ای ارائه شده داریم :

برای $\tau = 0.5$

$$L(u) = \min\left\{ \sum_{i \in \{i : y_i \geq u\}} 0.5\left|y_i - u\right| + \sum_{i \in \{i : y_i \prec u\}} (1-0.5)\left|y_i - u\right| \right\}$$

U=3 $$L(3) \propto \sum_{i=1}^{2} -(i-3) + \sum_{i=3}^{9} (i-3) = \left[(2+1)+(0+1+2+\ldots+6)\right] = 24$$

U=5 $$L(5) \propto \sum_{i=1}^{4} i + \sum_{i=0}^{4} i = 20$$

u	1	2	3	4	5	6	7	8	9
Expected value	36	29	24	21	20	21	24	29	36

مثال2

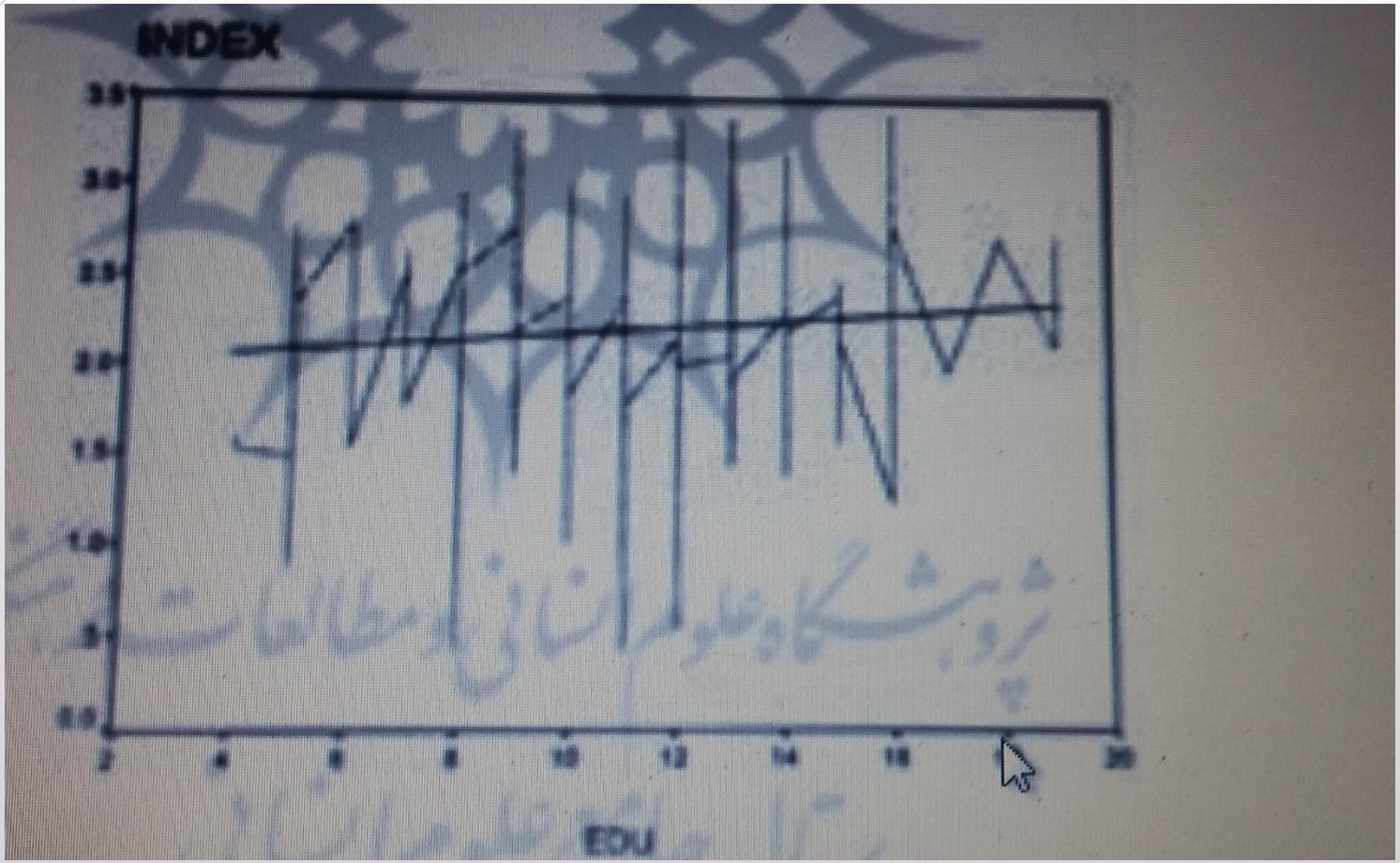
یازده مدل رگرسیون چندک همراه با رگرسیون معمولی برای بررسی رابطه‌ی رفاه درخواستی افراد با تعداد سالهای تحصیل انان به کار می رود.

رفاه مطلوب و تعداد سال های تحصیل به ترتیب با INDEX و EDU نشان خواهند شد. داده ها به یک نمونه تصادفی 684 نفری از جوانان 18 تا 29 سال تهرانی اختصاص دارد که با روش نمونه گیری خوشه ای سه مرحله ای در سال 1381 گرداوری شده است. شایان ذکر است که رفاه درخواست شد. با شاخصی که حاصل از ترکیب 33 سوال یک پرسشنامه است.، سنجیده شده است. گفتنی است مقدار این شاخص از 0 تا 4 تغییر میکند، به طوری که هر چه مقدار آن بیشتر می شود بر انتظارات بیشتری نیز دلالت دارد.

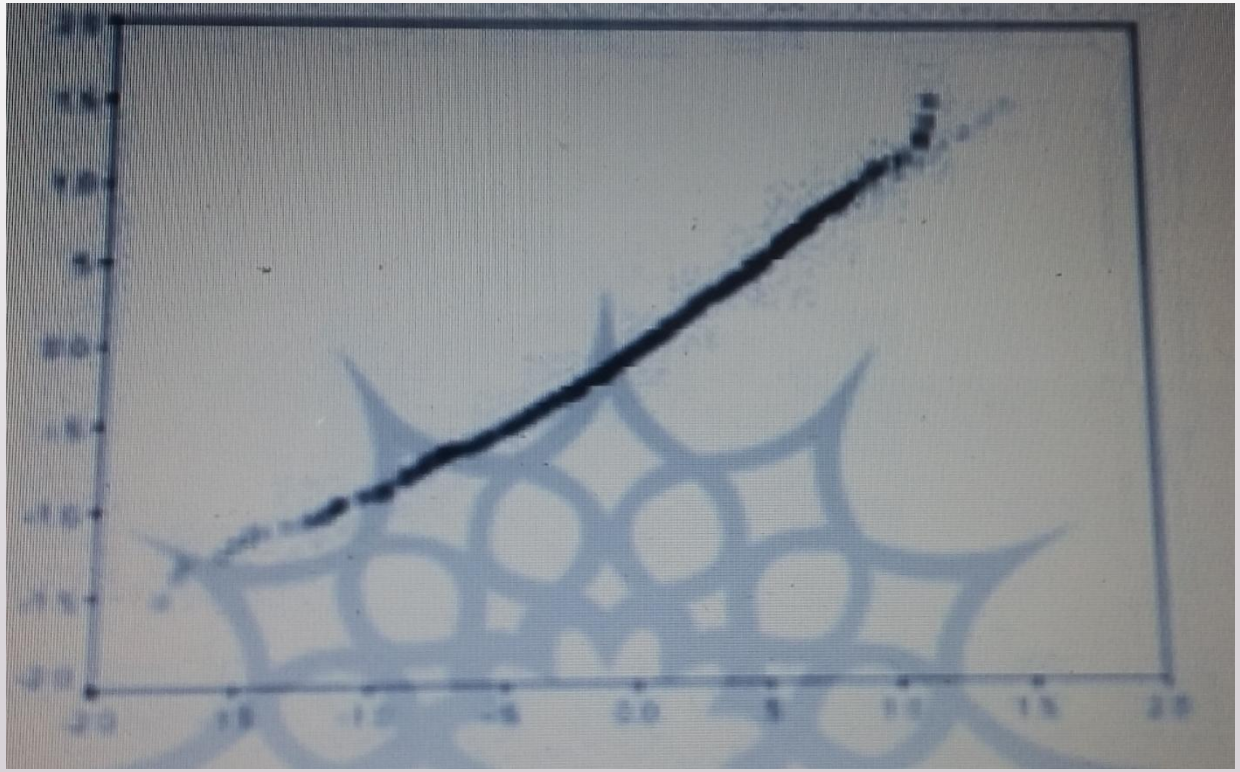
کار را با برازش یک مدل رگرسیون خطی معمولی با روش حداقل مربعات بر داده ها آغاز می کنیم.بر این اساس مدل برازشی عبارت است از:

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

که در آن E(INDEX) براورد میانگین توزیع رفاه درخواستی به ازای EDU سال تحصیل است. نمودار پراکنش داده ها همراه با خط برازش داده شده در شکل زیر آمده است.



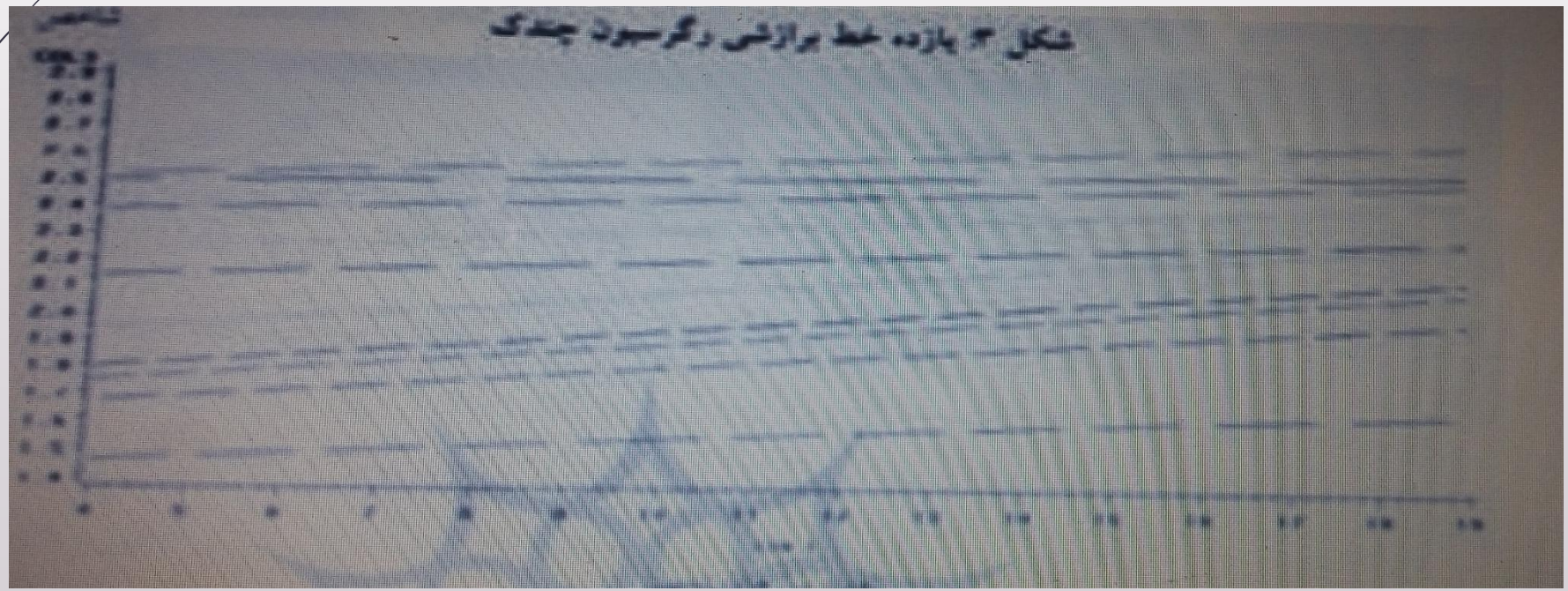
با توجه به مدل برازشی معیار d کوک به وجود تعداد زیادی داده ی دور افتاده اشاره داشت. هم چنین ترسیم نمودار Q-Q در شکل زیر برای بررسی نرمال بودن توزیع باقی مانده های مدل، انحراف از توزیع نرمال را نشان می دهد. آزمون کلموگرف-اسمیرنف نیز فرضیه ی نرمال بودن این توزیع را رد میکند. تمام این موارد حکایت از نامناسب بودن مدل رگرسیون معمولی دارد.



اکنون به برازش مدل های چندک به ازای یازده چندک مختلف به کمک برنامه ای که در محیط IML از نرم افزار SAS نوشته شده است می پردازیم.

این برنامه برای مدل 3 و براساس الگوریتمی است که باست و کونکر در ۱۹۸۲ تهیه کرده اند. گفتنی است وقتی این چندک، همان میانه باشد الگوریتم سریع تری از طرف مدسین و نیلسن پیشنهاد شده است که در SAS\IML درقالب روال LAV استفاده می شود.

در مجموع یازده چندک 0.1 و 0.2 و 0.25 و 0.3 و 0.4 و 0.5 و 0.6 و 0.7 و0.75 و 0.8 و0.9 بر داده ها برازش داده شد. شکل زیر نیز خطوط برازش داده شده را نشان می دهد .



بنابراین مثلا مدل برازش داده شده برای چندک 0.2 عبارت است از

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

که در آن $\hat{Q}_{i,0.2}$ براورد چندک 0.2 توزیع رفاه درخواستی به ازای EDU سال تحصیل است. پس براورد چندک 0.2 برای افراد با 19 سال تحصیل برابر است با

$$2.0075 = 1.58 + 0.0225 * 19$$

بنابراین می توان 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 90درصد از افراد با 19 سال تحصیل دارای شاخص انتظارات کمتراز ۲.۸۲۴۹ و 10 درصد بیشتر از ان باشند. پس تقریبا 70درصد از این افراد دارای شاخصی بین 2.0075 و 2.8249 هستند. توجه کنید که چنین تحلیل هایی تنها با مدل های رگرسیون چندک قابل انجام است و مدلهای رگرسیون معمولی چنین قابلیت هایی را ندارند.

یافته های حاصل از برازش:

- شکل اول حاکی از آن است که خط رگرسیون نمی تواند پیش بینی کننده ی مناسبی برای شاخص انتظارات باشد. زیرا باتوجه به پراکندگی زیاد داده ها در برخی از سطوح تعداد سال های تحصیل، میانگین نمی تواند این شاخص را برای این سطوح به خوبی پیش بینی کند. برای مثال در شکل اول افرادی را با 8 یا 12 سال تحصیل ملاحظه کنید.
- مثبت بودن شیب خطوط در شکل سوم یعنی ضریب تعداد سال های تحصیل، نشان دهنده ی رابطه مستقیم بین متغیر وابسته و تشریحی است. بنابراین با افزایش تعداد سالهای تحصیل مقدار هر یک از یازده چندک شاخص انتظارات نیز افزایش می یابد. بر اساس چندک های برازش ، سالهای تحصیل بر چندک های پایینی بیش تر از چندک های بالایی اثر دارد.
- برای افراد با تحصیلات بیشتر، فاصله ی کم چندک های بالایی در مقایسه با چندک های پایینی نشان می دهد که فشردگی داده ها 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 می شود.

- می توان پیش بینی کرد که مثلا 70 درصد افراد با 19 سال تحصیل دارای شاخصی بین 2.0075 و 2.8249 هستند، در حالی که این فاصله برای افرادی با ۴ سال تحصیل از 1.67 تا2.7184 است.

روش حداقل قدرمطلق انحرافات:

شیوه یبراورد پارمترهای مدل رگرسیون معمولی بر حداقل کردن مربع باقیمانده های 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 ی نقاط از آن به حداقل برسد. در رگرسیون چندک برخلاف رگرسیون معمولی از حداقل کردن مجموع قدرمطلق انحرافات یا LAD گفته می شود. گفتنی است که استفاده از روش LAD که در مدل رگرسیون چندک به کار می رود دارای سابقه طولانی است.

در میانه یقرن هجدهم بوسکویچ یک مدل خطی دو متغیره را برای بررسی بیضوی بودن کره زمین از طریق کمینه کردن قدر مطلق خطاها به کار برد. به دنبال ان لاپلاس براورد ضریب زاویه ی مدل رگرسیونی بوسکویچ را به طور دقیق معرفی و توزیع مجانبی انرا بدست اورد . توسعه رگرسیون میانه برای هر چندک دلخواه نیز به کوشش کانوکر و باست در 1978 صورت گرفت.
هدف انها براورد بردار پارامترهای مدل زیر بود که برای این منظور تابع زیانی که در پی می آید نسبت به عناصر β_θ کمینه می شود.

$$Q_\theta(Y \mid \mathrm{x}_i') = \mathrm{x}_i' \beta_\theta$$

$$\psi_\theta(\beta_\theta) = \sum_i \omega(\theta) \mid y_i - \mathrm{x}_i' \beta_\theta$$

در این تابع زیان

$$\omega(\theta) = \begin{cases} \theta & y_i \le \mathrm{x}_i' \beta_\theta \\ 1-\theta & y_i \succ \mathrm{x}_i' \beta_\theta \end{cases}$$

موزون کردن قدرمطلق باقیمانده ها در تابع فوق باعث میشود تا خط برازشس به گونه ای باشد که تتا در صد درصد داده ها تقریبا زیر آن و باقی آنها بالای خط قرار گیرند.

کمینه کردن رابطه قبل و یافتن براورد LAD پارامترها با استفاده از روشهای برنامه ریزی خطی و از طریق بسته های نرم افزاری صورت می گیرد.

ویژگی های حداقل قدرمطلق انحرافات

- در حالت خاص که مدل تنها شامل عرض از مبدا و تتای 0.5 است کمینه کردن رابطه ی

$$\psi_\theta(\beta_\theta) = \sum_i \omega(\theta) |y_i - \mathbf{x}_i' \beta_\theta$$

منجر به کمینه کردن عبارت $\sum_i |y_i - \beta_0|$ می شود که در این صورت براورد عرض از مبدا همان میانه داده ها خواهد بود.

- بر خلاف روش حداقل مربعات روش حداقل قدر مطلق انحرافات نسبت به داده های دورافتاده استوار است . این ویژگی ناشی از آن است که بر خلاف اهمیت اندازه باقی مانده ها در روش حداقل مربعات در این روش تنها به علامت باقی مانده ها توجه نمی شود. بنابراین نه تعداد باقی مانده هایی که بیشتر یا کمتر از چندک مورد مظرند و نه مقدار بزرگی آنها در براوردها اثرگذار است، پس داده های دور افتاده که تاثیر خود را از طریق بزرگی باقی مانده ها نشان می دهند نمی توانند براورد LAD را متاثر سازند.

- شکل بسته ای برای براورد پارامترهای این مدل وجود ندارد و از روش های عددی برای براورد آنها استفاده می شود . هم چنین جوابهای نهایی مدل رگرسیون چندک می تواند یکتا نباشد . البته یافتن جواب یکتا با انتخاب یک معیار مناسب امکان پذیر است.
- وقتی ε_{θ_i} متغیرهای تصادفی مستقل و هم توزیع باشند، خطوط رگرسیونی به تازای چندک های مختلف موازی خواهند بود.
- در رگرسیون چندکی نیز مانند رگرسیون معمولی می توان استنباط کرد. پاول و بوچنسکی نشان داده اند که براوردLAD پارامترها سازگار و به طور مجانبی نرمال است.

منابع:


با تشکر از توجه شما
HarfeTaze.com