چکیده

نظیر رگرسیون خطی کلاسیک که با کمینه کردن مجموع مربعات باقی مانده ها، تابع میانگین شرطی را پیش بینی می کند، رگرسیون چندک قادر است تابع میانه ی شرطی و تعداد زیادی از توابع چندک شرطی را پیش بینی نماید. هنگامی که مشاهدات سانسور شده باشند، رگرسیون چندک با استفاده از شیوه ی دوباره وزن دار کردن مشاهدات، تعمیم داده شده و برآورد حاصل از این روش در برابر نقاط موثر، دور افتاده در متغیر پاسخ یا دور افتاده در متغیر توضیحی، استوار است. هدف اصلی این پایان نامه ارائه ی برآوردهای مدل رگرسیون چندک سانسور شده و مقایسه روش های برآورد با استفاده از مطالعه شبیه سازی می باشد . در پایان کاربرد روش های مذکور برای داده های واقعی مورد بحث و بررسی قرار می گیرد

واژه های کلیدی:

رگرسیون چندک، رگرسیون چندک ژرفا، مشاهدات سانسور شده و استواری

فصل اول

تعاریف و مفاهیم اولیه

۱ -۱ مقدمه

مدل رگرسیون چندک برای نخستین بار در سال ۱۹۷۸ توسط کانکر- باست[۱] بیان شد. در سال های اخیر رگرسیون چندک، از نظر تئوری و کاربردی و تجربی بسیار مورد توجه قرار گرفته است، به گونه ای که در زمینه های متعددی مانند رگرسیون ناپارامتری، رگرسیون غیر خطی، سری زمانی و غیره راه یافته است . برآورد ارائه شده ی کانکر و باست برای مدل رگرسیون چندک در برابر نقاط موثر حساس بوده و در نتیجه خط برازش داده شده را منحرف می کند، لذا روسو و هوبرت در سال ۱۹۹۹ با معرفی روش ژرفترین رگرسیون، برآورد استوار مدل رگرسیون چندک را ارائه دادند. در بسیاری از حوزه ها مانند طب و اقتصاد با مشاهداتی سروکار داریم که قابل اندازه گیری نبوده و فقط حد پایینی از آن مد نظر است . چنین مشاهداتی را مشاهدات سانسور شده [۲] می نامیم. پورت نوی در سال ۲۰۰۳ مدل رگرسیون چندک کانکر وباست را بر اساس مشاهدا ت سانسور شده تعمیم داد. اما با وجود تاثیر پذیری برآورد پورت نوی در برابر نقاط موثر، پارک و هوانگ در سال ۲۰۰۳ برآورد استوار مدل رگرسیون میانه سانسور شده بیان نمودند.

ساختار این پایان نامه بدین صورت است: در ادامه این فصل، پس از آشنایی با بعضی عملیات های ریا ضی و مفاهیم کلیدی و اصطلاحات آماری به کار برده شده در پایان نامه، به معرفی رگرسیون چندک و رگرسیون چندک استوار [۳] و معیارهای ارزیابی استواری برآوردها می پردازیم . در فصل سوم با معرفی مشاهدات سانسور شده و انواع آن، رگرسیون چندک سانسور شده و رگرسیون چندک استوار سانسورشده را معرفی می کنیم و در نهایت در فصل چهارم با استفاده از داده های شبیه سازی شده به مقایسه روش های برآورد می پردازیم

---

[۱] Koenker-Bassett
[۲] Censored Observation
[۳] Depth Quantile Regression

۱-۱-۱ تعریف :

دورافتاده ها[۱]، مشاهداتی هستند که از الگوی اکثریت داده ها پیروی نمی کنند . داده دورافتاده ، مشاهداتی هستند که باقی مانده ی آن ها از نظر قدرمطلق، بسیار بزرگتر از باقی مانده های سایر مشاهدات است و شاید به فاصله ی بیشتر از سه یا چهار برابر انحراف از میانگین مانده ها قرار دارد.

۱-۱-۲ تعریف :

نقطه ی داده ای که مقدار منتهی الیهی برای یکی از متغیرهای توضیحی را داراست و هیچ تبعیتی از بقیه داده های مدل نمی کند یک **نقطه ی نافذ**[۲] نامیده می شود، نقاط نافذ ، اثری نامناسب در برآوردهای ضرایب رگرسیونی دارند.

۱-۱-۳ تعریف :

استواری[۳]، خاصیتی از برآورد می‌باشد، که مقاومت و تحت تأثیر قرار نگرفتن برآورد را در اثر نقاط دور افتاده و نقاط نافذ بیان می‌کند. معیارهایی نظیر، تابع تأثیر [۴]و نقطه‌ی فروریزش[۵] برای سنجش میزان استواری برآوردها معرفی شده است.

---

۱ Outlier
۲ Leverage Point
۳ Robustness
۴ Influence Function
۵ Breakdown point

۱-۱-۴ تعریف :

فرض کنید $x_1, x_2, \dots, x_n$ یک نمونه تصادفی n تایی از جامعه ای که توزیعش به پارامتر مجهول θ بستگی دارد و $T_1, T_2, \dots, T_m$ برآوردهای ناریب و مستقل از هم $\gamma(\theta)$ با واریانس های یکسان باشند . کلاس برآوردهای زیر را در نظر گیرید.

$$D = \{T: T = \sum_{i=1}^{m} \alpha_i T_i \ , \alpha_i \in [0,1], \ \sum_{i=1}^{m} \alpha_i = 1, \ V(T_i) = \delta^2 , COV(T_i, T_j) = 0 , V_i \neq j\}$$

کلاس D، کلاس برآوردهای ناریب خطی $\gamma(\theta)$ بر مبنای $T_1, T_2, \dots, T_m$ نامیده می‌شود. می گوییم $T^* \in D$

بهترین برآورد خطی ناریب (BLUE) [1]پارامتر $\gamma(\theta)$ در کلاسD است، اگر برای هر $T \in D$و هر $\theta \in \Theta$

$$V_\theta(T^*) \leq V_\theta(T)$$

۱-۱-۵ تعریف:

مقداری از x، که به ازای آن$f(x)$ دارای بیشترین مقدار مثلا M باشد، را تابع $\boldsymbol{arg\,max_x(f(x))}$ می گوییم.

$$arg\,\max_x(f(x)) := \{x | \forall y : f(y) \leq f(x) = M\}$$

۱-۱-۶ تعریف :

تابع $\boldsymbol{arg\,min_x(f(x))}$ مشابه تابع $\text{argmax}_x(f(x))$ می‌باشد ولی با این تفاوت که در آن m کمترین مقدار $f(x)$ است و به زبان ریاضی:

[1] Best Linear Unbiased Estimation

$$arg\min_{x}\big(f(x)\big) := \{x|\forall y : f(y) \geq f(x) = m\}$$

۱-۱-۷ تعریف :

سوپریمم مجموعه‌ای مانند X، کوچکترین کران بالای مجموعه X است. مثلاً اگر $X = \{x|\ ۰ < x < ۱۰۰\}$ آنگاه $sup(X) = ۱۰۰$. تابع سوپریمم به این دلیل که در بعضی مواقع، ماکسیمم وجود ندارد، مفید است.

۱-۱-۸ تعریف :

اینفیمم مجموعه‌ای مانند X، بزرگترین کران پایین مجموعه‌ی X است. در مثال ارائه شده ی تعریف قبل $inf(X) = ۰$ تابع اینفیمم به این دلیل که در بعضی مواقع، مینیمم وجود ندارد، مفید است.

۱-۱-۹ تعریف :

برای تعریف تابع $O(.)$ **"بزرگ در ریاضیات"**، فرض کنید توابع g و f روی یک حوزه تعریف یکسان و احتمالاً در یک فاصله نامتناهی تعریف شده باشد. گیریم $z_۰$ نقطه‌ای از یک فاصله یا کران بالا یا پایین این فاصله باشد، نیاز داریم به ازای همه ی zها در همسایگی $z_۰$ و $z \neq z_۰$، داشته باشیم $g(z) \neq ۰$. اگر مقدار ثابتی مانند M وجود داشته باشد، به طوریکه وقتی $z \to z_۰$ ، داشته باشیم:

$$|f(z)| \leq M|g(z)|$$

آنگاه می‌گوییم:

$$f(z)=O(g(z))$$

۱-۱-۱۰ تعریف :

با توجه به توابع g وf در تعریف قبل، تابع (.) "o" به این صورت است که، اگر $\lim_{z\to z.}\frac{f(z)}{g(z)}=0$ آنگاه می‌گوییم که $f(z)=o(g(z))$

۱-۱-۱۱ تعریف :

برای متغیر تصادفی X ، پارامتر Q_τ را **چندک** [۱] τ ام برای $F(x)$ یا برای X ، می نامند . هرگاه تقریباً 100τدرصد داده‌ها کوچکتر یا مساوی آن باشد، به عبارت دیگر نامساوی دوطرفه‌ی زیر برقرار باشد:

$$P(X<Q_\tau)\le\tau\le P(X\le Q_\tau)$$

نامساوی دوطرفه‌ی بالا، بدین معنی است که مقدار احتمال در فاصله ی باز $(-\infty,Q_\tau)$ حداکثر τ و در فاصله‌ی نیم با $(-\infty,Q_\tau]$ حداقل τ است و از لحاظ هندسی اگر از نقطه‌ی Q_τ، خطی به موازات محور y ها رسم کنیم، مساحت زیر منحنی فراوانی که در سمت چپ این خط قرار دارد برابر τ است.

۱-۱-۱۲ تعریف :

مدل ریاضی رگرسیون خطی $Y_i=\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik}+u_i \quad x_{i1}=1,\ i=1,2,\dots,n$ را در نظر می گیریم که u_i ها دارای تابع توزیع $F(.)$ هستند. صورت ماتریسی این عبارت، $Y=X\beta+\varepsilon$ است.که در آن $u=(u_1,u_2\dots.,u_n)'$ بردار خطا و $\beta=(\beta_1,\beta_2,\dots,\beta_k)'$ بردار پارامترهای مجهول وX ماتریس $n\times k$ ی معلوم و دارای رتبه ی کامل ستونی و $\acute{Y}=(Y_1,Y_2\dots.,Y_k)$ بردار مشاهدات متغیر وابسته است . اگر$\{x_t:t=1,2,\dots,n\}$ نشان دهنده ی بردارهای تشکیل دهنده ی سطرهای ماتریس X و $\{u_t:t=$

[۱] Quantile

$\{1,2,\ldots,n\}$ نشان دهنده ی اعداد تصادفی فرآیند رگرسیونی $u_t = Y_t - x_t'\beta$ با تابع توزیع $F(.)$ باشد. آنگاه برای $0 < \tau < 1$، برآورد τ امین **چندک رگرسیونی**[1] (β_τ) به صورت زیر تعریف می شود.

$$\hat{\beta}_\tau = min\left[\tau \sum_{t\in\{t:Y_t \geq x_t'\beta_\tau\}}(Y_t - x_t'\beta_\tau) + (1-\tau)\sum_{t\in\{t:Y_{t<x_t'\beta_\tau}\}}(Y_t - x_t'\beta_\tau)\right]$$

۱-۱-۱۳ تعریف :

میزان عدم تقارن منحنی فراوانی را **چولگی**[2] می‌نامند. فرض کنید $\bar{x}$ میانگین، m میانه، M نما، S انحراف استاندارد. هرکدام از فرمول‌های زیر را می‌توان به عنوان معیار چولگی به کار برد:

ضریب چولگی اول پیرسن $b_1 = \frac{\bar{x}-M}{s}$

ضریب چولگی دوم پیرسن $b_2 = \frac{3(\bar{x}-m)}{s}$

بر حسب این که b_1 , b_2 مثبت یا منفی باشند، منحنی، چوله به راست یا چوله به چپ می‌باشد. به طوری که اگر b_1 , $b_2 > 0$ منحنی به راست چوله و اگر b_1 , $b_2 < 0$ منحنی به چپ چوله خواهد بود.

۱-۱-۱۴ تعریف :

در رگرسیون چندک برای برآورد پارامترهای مدل از روش **کمترین قدر مطلق انحرافات** (LAD)[3] استفاده می شود. برای این منظور، تابع زیان $Q_\tau(.)$ که برابر با قدر مطلق باقی مانده ها یا انحرافات موزون است، نسبت به عناصر β_τ کمینه می شود.

$$Q_\tau(\beta_\tau) = \sum_i w(\tau)|Y_i - x_i'\beta_\tau|$$

---

[1] Quantile Regression
[2] skewness
[3] Least Absolute Deviations

$$w(\tau) = \begin{cases} ۱ - \tau & ; \quad Y_i \leq x_i' \beta_\tau \\ \tau & ; \quad Y_i > x_i' \beta_\tau \end{cases}$$

۱-۱-۱۵ تعریف :

ژرفای رگرسیون[۱]، عدد صحیحی بین صفر و n که خاصیتی از برازش می باشد که از آن به صورت رتبه ی خط تعبیر می کنیم. ژرفای رگرسیون معیاری برای مقایسه ی خطوط است. از این جهت که خط ژرفتر برازش بهتری برای داده هاست. بنابراین ژرفای رگرسیون β کیفیت برازش را اندازه گیری کرده و میزان تعادل مجموعه داده $Z_n \subseteq \Re^۲$ در اطراف خط برازش شده به وسیله ی برآوردβ را اندازه می گیرد.

۱-۱-۱۶ تعریف :

در بسیاری از مواقع، داده های حیاتی در هنگام رخداد پدیده، قابل اطمینان نیستند. چنین داده هایی را **داده های سانسور شده**[۲] می نامیم. این مجموعه داده ها، ممکن است در طول مطالعه به طور کامل شرکت نداشته و یا تا پایان مطالعه از کار نیافتاده باشند. باید توجه داشت که داده سانسور شده را باید با استفاده از آخرین اطلاع از واحدها دقیقاً ثبت کرد تا در تحلیل داده ها مورد استفاده قرار گیرد.

۱-۱-۱۷ تعریف :

فرض کنید بخواهیم مشتق جهتی یا سویی تابع f را در نقطه ی (x_0, y_0) در جهت بردار یکه ی u بدست آوریم. **مشتق جهتی یا سویی** برابر با حاصلضرب نقطه ای گرادیان f (∇f) در نقطه ی (x_0, y_0)ضرب در بردار u یعنی $\nabla f . u$

---

[۱] Regression Depth
[۲] Censored Data

فصل دوم

معرفی رگرسیون چندک و

رگرسیون چندک ژرفا

۲-۱ مقدمه ای بر رگرسیون

در تجزیه وتحلیل داده ها اغلب، هدف آگاهی از رابطه بین یک متغیر وابسته با یک یا چند متغیر مستقل است. در واقع ممکن است وابستگی ساده ای بین متغیرها وجود داشته باشد و یا این که رابطه ای تابعی وجود دارد که درک یا توصیف آن با اصول ساده،کمی پیچیده است. مایلیم که این وابستگی را به وسیله ی تابعی ریاضی، نظیر یک تابع چند جمله ای از متغیرهای مناسب که تابع واقع ی را تقریب بزند، بیان کنیم . با بررسی چنین تابعی قادریم که رابطه ی واقعی بین متغیرها را شناخته و اثرات جدا یا توأم متغیرها را ارزیابی کنیم . همچنین بررسی ماهیت و مدل بندی چنین روابطی را می توان با استفاده از مدل های رگرسیون خطی مورد آزمون قرار داد . در مدل های رگرسیون با دو نوع متغیر، متغیرهای پیشگو[۱] یا توضیحی یا مستقل و متغیر پاسخ [۲]یا وابسته، سروکار داریم که فرض شده متغیرهای پیشگو مقید به تغییر تصادفی نیستند. اما متغیر پاسخ تصادفی است. چون از نظر کاربردی، اگر چنین نباشد، شیوه ی برازش بسیار پیچیده خواهد بود. ەمچنین رابطه ای میان متغیرهای پیشگو و متغیر پاسخ رابطه ی ریاضی نبوده، بلکه رابطه ای آماری است. بدین معنی که به ازای هرمقدار x، مقدار y به طور کامل مشخص نیست و تعیین مقدار آن با مقداری خطا همراه است . معمولاً به دلیل ساده بودن و وجود مبانی نظری مبتنی بر تقریب روابط غیر خطی به روابط خطی، رابطه ی میان متغیر پاسخ و متغیرهای توضیحی خطی در نظر گرفته می شود . منظور از بیان خطی بودن مدل، مدل خطی بودن نسبت به پارامترهاست . در رگرسیون خطی معمولاً برای برآورد پارامترها از شیوه ی کمترین مربعات خطا [۳] استفاده می شود. که اولاً در کاربردهای معمولی تقریباً روش ساده ای است . ثانیاً برآورد ها بنا به قضیه ی گاوس- مارکف بهترین برآورد نا اریب خطی (BLUE) می باشند (رنچر[۴] ۱۹۹۸) . ثالثاً انطباق آن ها به برآوردهای دیگر، نظیر برآورد ماکسیمم

---

[۱] Predictor variable
[۲] Response variable
[۳] Least Squares Method
[۴] Rencher

راستنمایی[1](MLE) به مطلوبیت آنها می افزاید. مدل رگرسیون معمولی به تحلیل گر کمک می کند تا رابطه ی میانگین توزیع متغ یر تصادفی Y را با تعدادی متغیر توضیحی بررسی کند . برای روشن شدن مطلب، مدل رگرسیون خطی ساده ی زیر را در نظر می گیریم .

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (۱-۲)$$

که در آن β_0, β_1 و $\varepsilon_i \sim N(0, \sigma^2)$ به ترتیب، پارامتر ومقادیر مشاهده نشدنی هستند. اگر $E(\varepsilon_i) = 0$ آنگاه مدل (۲-۱) را می توان به صورت زیر بازنویسی کرد:

$$E(Y) = \beta_0 + \beta_1 x_i \qquad (۲-۲)$$

کمیت $E(Y_i)$ میانگین شرطی متغیر تصادفی Y است که آن را با $E(Y|x_i)$ نشان میدهند. رابطه (۲-۲) بیان می کند که متغیرهای تصادفی Y در هر سطحی از متغیر توضیحی دارای توزیعی است که میانگین های این توزیع روی خط راست جای گرفته اند.

۲-۲ چندک های جامعه و نمونه

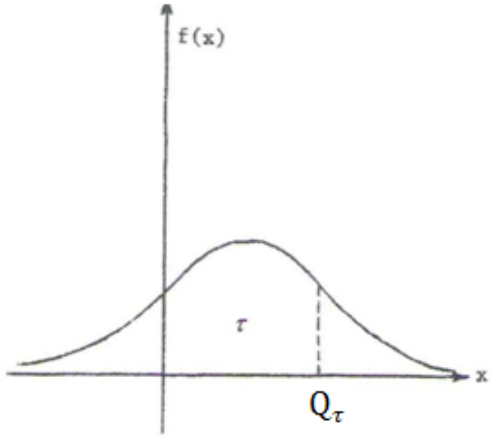
اصطلاح چندک مترادف با درصد و نقاط درصدی است، بطوریکه میانه معروف ترین مثال از چندک هاست. فرض کنید، متغیر تصادفی X دارای توزیع تجمعی $F(x)$ باشد، پارامتر Q_τ را چندک مرتبه τ برای $F(x)$ یا برای X، می نامیم هر گاه نامساوی دو طرفه زیررا داشته باشیم :

$$P(X < Q_\tau) \leq \tau \leq P(X \leq Q_\tau) \qquad 0<\tau<1 \qquad (۳-۲)$$

معنی این نامساوی دو طرفه این است که مقدار احتمال در فاصله باز $(-\infty ، Q_\tau)$ حداکثر τ و در فاصلهٔ نیم باز $(-\infty ، Q_\tau]$ حداقل τ می باشد. از لحاظ هندسی اگر از نقطه‌ی Q_τ، خطی به موازات محور Y ها رسم کنیم،

[1] Maximum Likelihood Estimator

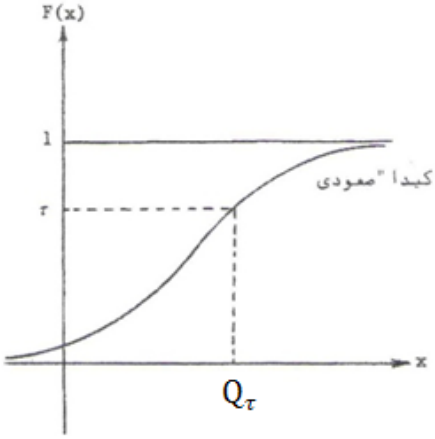
مساحت زیر منحنی فراوانی که در سمت چپ این خط قرار دارد برابر τ است. این مطلب را در شکل (۲-۱) می توان دید.



شکل (۲-۱)

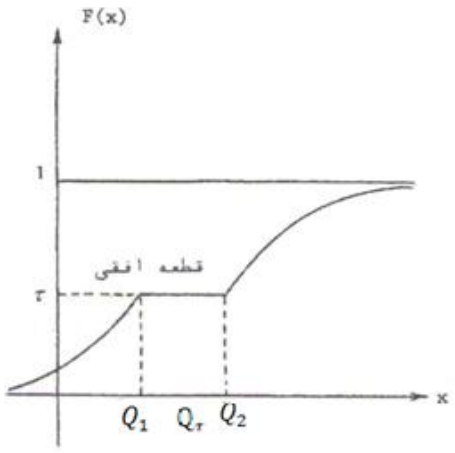
اینک حالات خاص زیر را در نظر می گیریم:

الف: اگر $F(x)$ پیوسته و اکیداً صعودی باشد، یعنی نمودار آن دارای خطوط افقی یا جهشی نباشد، آنگاه نامساوی (۲-۳) تبدیل به تساوی $F(Q_\tau) = \tau$ شده و Q_τ پاسخ یکتای معادله $F(Q_\tau) = P(X \leq Q_\tau)$ یا $\int_{-\infty}^{Q_\tau} f(x)dx = \tau$ می باشد و این مطلب را در شکل (۲-۲) می توان دید.



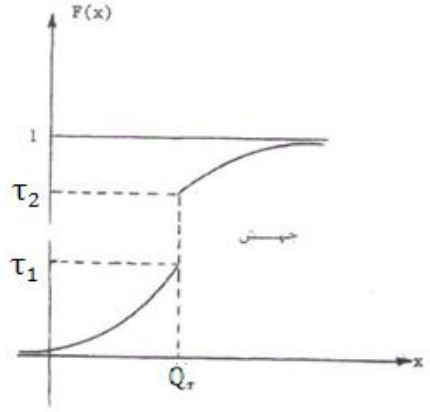
شکل (۲-۲)

ب: اگر نمودار $F(x)$ محتوی یک یا چند قطعه خط افقی باشد، ممکن است Q_τ برای بعضی از مقادیر τ یکتا نباشد. این مطلب را در شکل (۲-۳) می توان دید.



شکل (۲-۳)

ج: اگر $F(x)$ در یک یا چند نقطه دارای جهش باشد ممکن است، Q_τ برای بعضی از مقادیر τ یکسان باشد که در شکل(۲-۴) نمایش داده شد.



شکل (۲-۴)

عدد Q_τ که در آن ۱> τ >۰، است، چندک مرتبه τ می نامند، هر گاه تقریباً τ۱۰۰ درصد داده ها کوچکتر از آن باشند. در حقیقت $Q_{۰/۵}$ ، همان میانه است. به زبان هندسی اگر از نقطه Q_τ خطی به موازات محور y ها رسم کنیم مساحت زیر منحنی فراوانی که در سمت چپ این خط قرار دارد برابر τ واحد مربع می باشد.

چندکهای معروف عبارتند از:

الف: چارکهای [۱] اول تا سوم که به ازای ۰/۷۵, ۰/۵, ۰/۲۵= τ ، به دست می آیند و آنها را با $Q_{۰/۲۵}$ و $Q_{۰/۵}$ $Q_{۰/۷۵}$ و $Q\textbf{Error! Bookmark not defined.}$ نشان می دهند.

ب: دهک ها به ازای ۰/۹ , ... , ۰/۱= τ ، به دست می آیند و آنها را با $D_۹$, ... , $D_۱$ نشان می دهند.

ج: صدکها که به ازای ۹۹/ ۰,.... , ۰/۰۱= τ ، به دست می آیند و آنها را با $P_{۹۹}$, ... , $P_۱$ نشان می دهند.

می دانیم که میانه می تواند به عنوان مقدار میانی یا میانگین دو مقدار میانی یک مجموعه با داده های ترتیبی تعریف شود. معمولاً میانه نمونه را به عنوان برآورد ی از میانه ی جامعه m، در نظرمی گیرند . مخصوصاً برای متغیرهای تصادفی پیوسته، m جواب معادله ی $F(m)=\frac{۱}{۲}$ که در آن $F(y)=P(Y\leq y)$ تابع توزیع تجمعی متغیر تصادفی Y است. عدد Q_τ را به ازای $۰<\tau<۱$ چندک مرتبه ی τ ام می نامند، هرگاه τ۱۰۰ درصد داده ها کوچکتر از آن باشد. مثلاً $Q_{۰/۱۵}$ را چندک مرتبه ی ۰/۱۵ می گویند هر گاه تقریباً ۱۵ درصد داده ها کوچکتر از $Q_{۰/۱۵}$ باشد. از آن جا که میانگین، یکی از معیارهای تمرکز است، آگاهی از آن به تنهایی اطلاعات کاملی را از شکل توزیع بیان نمی کند . بنابراین رگرسیون کلاسیک ابزار مناسبی برای بیان شکل توزیع متغیر مورد مطالعه در سطوح مختلف متغیر توضیحی نمی باشد. چندک ها معیار دیگری برای توزیع هستند که در کنار هم می توانند شکل توزیع را به تصویر کشند. برای مثال اگر دهک [۲] های توزیعی دارای فاصله یکسانی از

[۱] Quartile
[۲] Deciles

یکدیگر باشند، انتظار داریم توزیع نسبتاً هموار یا یکنواختی داشته باشیم. همچنین اگر در توزیعی دهک های بالایی دارای فاصله زیاد و دهک های پایینی دارای فاصله کمی از یکدیگر باشند، توزیع به سمت راست چوله خواهد بود. هدفی که رگرسیون چندک دنبال می کند ، ارائه ی شیوه ی رگرسیونی برای چندک ها است که با استفاده از آن قادر خواهیم بود، شکل توزیع را در سطوح مختلف متغیرهای توضیحی به دست آوریم.

۲-۳ معرفی رگرسیون چندک

مدل رگرسیون چندک با ایده ای مشابه رگرسیون کلاسیک به برآورد چندک های شرطی و بررسی رابطه ی متغیرهای توضیحی با چندک ها می پردازد . با وجود این مهمترین کاربرد رگرسیون چندک، شناسایی شکل توزیع متغیر وابسته مدل در سطوح گوناگون متغیرهای توضیحی باشد که این کار با برازش مدل رگرسیونی به ازای چندک های مختلف برای مجموعه ای از داده ها انجام می شود. برای بیان تعریف دقیقی از مدل رگرسیون چندک τ ام، به ازای $\tau \in (۰,۱)$ ابتدا حالت ساده آن را در نظر می گیریم . مدل به ازای τ های مختلف، دسته ای از خطوط موازی که دارای عرض از مبدأ متفاوتی هستند، ارائه می دهد. به طور کلی مدل رگرسیون چندک به صورت زیر تعریف می شود:

$$Y_i = x_i{}'\beta_\tau + \varepsilon_{\tau i} \qquad (۴-۲)$$

$$Q_\tau\left(Y\middle|x_i{}'\right) = x_i{}'\beta_\tau \qquad (۵-۲)$$

که در آن $x_i{}' = (۱, x_{i۱}, x_{i۲}, \ldots\ldots, x_{ik})$ و $\beta_\tau{}' = \left(\beta_۰(\tau), \beta_۱(\tau), \ldots., \beta_k(\tau)\right)$ به ترتیب بردارهایی از مقادیر معلوم و پارامترهای نا معلو م بوده و $\varepsilon_{\tau i}$ یک متغیر تصادفی مشاهده نشدنی است . همچنین $Q_\tau(Y|x_i)$ نشان دهنده ی چندک شرطی τ ام توزیع Y است. با شرط $Q_\tau(\varepsilon_\tau|x_i) = ۰$ مدل (۲-۵) را مدل

رگرسیون خطی چندک τ ام می نامند. همان طور که در رگرسیون کلاسیک پارامتر های مدل را با کمینه [1] کردن مربع باقی مانده ها برآورد می کنند ، به گونه ای که فاصله نقاط از خط برازش داده شده حداقل شود . رگرسیون چندی نیز با کمینه کردن مجموع قدر مطلق موزون، پارامتر ها را برآورد می کند که این روش برآورد کردن را روش کمترین قدرمطلق انحرافات(LAD) می نامند.

در عبارت زیر تابع زیان$Q_\tau(.)$ برابر با قدر مطلق باقی مانده ها یا انحرافات موزون است که نسبت به عناصر β_pکمینه می شود. که در آن :

$$Q_\tau(\beta_\tau) = \sum_i W(\tau)\ |Y_i - x_i{}'\beta_\tau| \qquad (۶-۲)$$

$$W(\tau) = \begin{cases} ۱ - \tau & ; \quad Y_i \le x_i{}'\beta_\tau \\ \tau & ; \quad Y_i > x_i{}'\beta_\tau \end{cases}$$

موزون کردن قدر مطلق باقی مانده ها در این تابع باعث می شود تا خط برازش داده شده به گونه ای باشد که τ ۱۰۰ درصد داده ها تقریباً بالای خط و بقیه پایین آن قرار گیرند.

۲-۴ روش های برآورد در رگرسیون

در رگرسیون کلاسیک پارامترهای مدل را از طریق کمینه کردن تابع زیان درجه دوم $r(t) = t^۲$ برآورد می کنند، یعنی با توجه به داده های $\{(x_i, Y_i)\ ;\ i = ۱,۲, \dots, n\}$ برآورد پارامترهای مدل با کمینه کردن رابطه ی زیر به دست می آید.

$$\sum_{i=۱}^{n} r(Y_i - x_i{}'\beta) = \sum_{i=۱}^{n} (Y_i - x_i{}'\beta)^۲ \qquad (۷-۲)$$

[1] Minimum

در رگرسیون چندک نیز برای برآورد کردن پارامتر ه۱، از روش کمترین قدرمطلق انحرافات (LAD) که تابع زیان را کمینه کند، استفاده می شود. در اینجا حالت خاصی از رگرسیون چندک رگرسیون میانه[1] را بررسی می کنیم. روش برآورد کردن پارامتر مدل رگرسیون میانه از طریق کمینه کردن$\sum_{i=1}^{n}|(Y_t - x_t'\beta)|$ می باشد که این روش را رگرسیون L_1 می نامند. در رگرسیون میانه تابع $\rho_{0/5(t)}=0/5|t|$ زیان در رگرسیون میانه تعریف می شود، در واقع

$$\rho_{0/5(t)} = 0/5\ t\ I(t \geq 0) - (1 - 0/5)tI(t < 0) \qquad (2\text{-}8)$$

که در آن $I(.)$ تابع نشانگر می باشد . این تعریف را می توان با جایگزین کردن عدد ۰/۵ با عددی مانند $\tau \in (0,1)$ تعمیم داد. در نتیجه تابع

$$\rho_{\tau(t)} = \tau\ t\ I(t \geq 0) - (1 - \tau)tI(t < 0) \qquad (2\text{-}9)$$

که به آن تابع بازبینی [2] می گویند و به صورت $\rho_{\tau(t)} = t\big(\tau - I(t < 0)\big)$ نیز نمایش می دهند.

برآورد (LAD) پارامترها نسبت به برآورد کمترین مربعات دارای مزیت های زیر است:

الف) در حالت خاص که مدل تنها شامل عرض از مبدأ بوده و $\tau = 0/5$ می باشد، کمینه کردن رابطه ی (۲-۸) برابر به‌کمینه کردن عبارت $\sum_i|Y_i - \beta_0|$ می شود که در این صورت برآورد همان میانه ی داده ها خواهد بود.

ب) برای برآورد پارامتر ها به روش حداقل قدرمطلق انحرافات، از روش های عددی استفاده می شود. همچنین جواب های نهایی مدل رگرسیون چندک می توانند یکتا نباشد. البته با به کارگرفتن معیار های مناسب می توان به جواب یکتا دست یافت. (کانکر و باست ۱۹۷۸)

---

[1] Median Regression
[2] Check Function

ج) یکی از پیش فرض های رگرسیون این است که توزیع شرطی Y به ازای x های مختلف دارای واریانس برابری باشند. در عمل ممکن است از این پیش فرض تخطی شود، که به آن ناهمواریانسی گویند. یک وظیفه مهم برای تحلیل گر داده ها ، تشخیص وجود ناهمواریانسی است . مثلا اگر خط برازش داده شده رگرسیون چندک به ازای چندک های مختلف نمودارهای موازی بدهد، تحلیل گر عدم وجود ناهمواریانسی را نتیجه می گیرد. بنابراین نمودارهای چندک ابزار توصیفی مهمی برای تشخیص ناه‌م‌واریانسی می باشند. همچنین وقتی $\varepsilon_{\tau i}$ ها متغیرهای تصادفی مستقل و هم توزیع باشند، خطوط رگرسیونی در مدل (۲-۶) به ازای چندک های مختلف موازی خواهد بود.

د) در رگرسیون چندک نیز مانند رگرسیون معمولی، می توان به استنباط آماری پرداخت. به عنوان مثال پاول (۱۹۸۹) و بوچسکی (۱۹۹۸) نشان داده اند که برآورد LAD پارامترها سازگار و به طور مجانبی نرمال است.

۲- ۵ چندک ها

همان طور که روش های رگرسیون خطی کلاسیک، بر مبنای کمینه کردن مجموع مربعات باقی مانده ها است و برآورد توابع میانگین شرطی مدل ها را امکان پذیر می سازد، روش های رگرسیون چندک هم مکانیزمی را برای برآورد مدل های توابع میانه ی شرطی و تعداد زیادی از توابع چندک شرطی دیگر پیشنهاد می کند. این توابع با بیان شکل توزیع، مفسر خوبی برای شرح نقاط پایینی یا بالایی به همان خوبی نقاط میانی هستند. در نتیجه می توان برآورد رگرسیون چندک را به عنوان تعمیم طبیعی برآورد کمترین مربعات کلاسیک مدل های میانگین شرطی، برای چندک شرطی مد نظر قرار داد . حالت خاص این نوع برآور دگر، رگرسیون میانه می باشد . این برآورد مجموع قدرمطلق خطاها را کمینه می کند. سایر توابع چندک شرطی، با کمینه کردن یک مجموع موزون نا متقارن از قدر مطلق خطاهای برآورد، به دست می آیند.

۲-۵-۱ چندک های نمونه ای

اگر بخواهیم اطلاعات بیشتری را در مورد یک مجموعه از مشاهدات ارائه دهیم، ساده ترین راه حل محاسبه ی چندک های نمونه ای است. در این بخش هدف معرفی ابتدایی چندک های نمونه ای است که در مرتب شدن مشاهدات نقش بسزایی دارد . اگر $\{Y_t: t = ۱,۲, \dots, n\}$ یک نمونه تصادفی از متغیرهای تصادفی Y با تابع توزیع $F(.)$ باشد، آنگاه τ امین چندک نمونه ای (μ_τ) به ازای $۰ < \tau < ۱$ جوابی است که از کمینه کردن عبارت زیر به دست می آید.

$$min\left[\tau \sum_{t\in\{t:Y_t\geq\mu_\tau\}}(Y_t-\mu_\tau)+(۱-\tau)\sum_{t\in\{t:Y_t<\mu_\tau\}}(Y_t-\mu_\tau)\right] \quad (۱۰-۲)$$

۲-۵-۲ چندک های رگرسیونی

مدل ریاضی رگرسیون خطی زیر را در نظر می گیریم :

$$Y_i = \beta_۱ x_{i۱} + \beta_۲ x_{i۲} + \cdots + \beta_k x_{ik} + u_i \qquad x_{i۱} = ۱,\ i = ۱,۲, \dots, n \quad (۱۱-۲)$$

هستند. صورت ماتریسی رابطه ای (۲-۱۱) عبارت است از : $F(.)$ ها دارای تابع توزیع u_i که

$$Y = X\beta + u \quad (۱۲-۲)$$

که در آن $u = (u_۱, u_۲, \dots, u_n)'$ بردار خطا و $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$ بردار پارامترهای مجهول و X ماتریس $n \times k$ ی معلوم و دارای رتبه ی کامل ستونی و $Y = (Y_۱, Y_۲, \dots, Y_n)$ بردار مشاهدات متغیر وابسته است. اگر $\{x_t : t = ۱,۲, \dots, n\}$نشان دهنده ی بردارهای تشکیل دهنده ی سطرهای ماتریس X و $\{u_t: t = ۱,۲, \dots, n\}$ نشان دهنده ی اعداد تصادفی فرآیند رگرسیونی $u_t = Y_t - x_t'\beta$ با تابع