رگرسیون چندکی خطی

(تحقیق درس مدل های خطی) 1393/10/29

ارائه دهندگان

مریم حسن پور راحله موسوی

استاد مربوطه

دکتر ریحانه ریخته گران

در چه مواقعی از رگرسیون چندکی استفاده می کنیم؟

- برای بررسی ارتباط دو متغیر، معمولا رگرسیون حداقل مربعات به کارگرفته می شود. با رگرسیون حداقل مربعات ارتباط بین مت[illegible] و میانگین پاسخ شرطی را می توان برآورد کرد، اما این روش مثل بسیاری از روش های آماری دارا[illegible]
- رگرسیون حداقل مربعات در [illegible] فرض نرمال بودن داده ها، برای تخمین ارتباط بین متغیر کمکی ومتغیر پاسخ اعتبار کافی نخو[illegible]
- از طرفی رگرسیون حداقل مربعات تنها ارتباط بین متغیرهای کمکی و میانگین پاسخ [illegible]، درحالی که در بسیاری موارد هدف پیدا کردن ارتباط بین متغیرهای کمکی با سایر بخشهای توزیع به ویژه چندک های انتهایی توزیع می باشد.

یک روش ارزشمند در مواجهه با چنین مشکلاتی استفاده از رگرسیون چندکی است. (Quantile Regression)

رگرسیون چندکی یک روش آماری با قابلیت محاسبه و رسم منحنی های رگرسیونی متفاوت و منطبق با نقاط صدکی مختلف میباشد، که ضمن بیان تصویری کامل تر و جامع تر از داده ها، امکان سنجش ارتباط متغیرهای مستقل با چندک های مورد نظر متغیر وابسته را بدون نیاز به نرمال بودن داده ها و حتی در حضور نقاط دور افتاده فراهم میکند یعنی این رگرسیون نسبت به داده های دورافتاده نیرومند می باشد. از سوی دیگر برخلاف رگرسیون حداقل مربعات که روی میانگین شرطی یعنی پارامتر مکان متمرکز است، رگرسیون چندکی استراتژی منظمی را برای تعیین چگونگی تاثیر متغیرهای مستقل روی مکان و مقیاس و شکل توزیع پیشنهاد میکند.

به طور خلاصه رگرسیون چندکی مدلی است که به بیان چگونگی تاثیر متغیر های مستقل بر چندک های دلخواه متغیر پاسخ می پردازد.

مزایای استفاده از رگرسیون چندکی

- پیدا کردن ارتباط بین متغیرهای کمکی با سایر بخش های توزیع به ویژه چندکهای انتهایی توزیع میباشد.
- نیازی به شرط نرمال بودن ندارد
- نسبت به داده های دور افتاده نیرومند است
- چگونگی تاثیر متغیرهای مستقل روی مکان و مقیاس و شکل توزیع را نشان می دهد.

معایب

در ابتدای معرفی رگرسیون چندکی به دلیل نبود کامپیوتر و نرم افزار های آماری محاسبه رگرسیون چندکی بسیار سخت و خسته کننده بود که البته با آمدن نرم افزار های آماری این مشکل کاملا برطرف شد.

آشنایی با چندک ها

- اگر Y یک متغیر تصادفی با تابع توزیع $F(y) = p(Y \leq y)$ باشد τ امین چندک Y برابر است با

$$Q(\tau) = F^{-1}(\tau) = \inf\{y: F(y) \geq \tau\}$$

اگر $\{y_{i,i=1,2,\ldots,n}\}$ نمونه ای تصادفی از متغیر تصادفی Y با تابع توزیع F باشد چندک τ ام نمونه به صورت زیر محاسبه می‌شود:

$$\min\{\sum_{i\in\{i:\, y_i \geq u\}} \tau|y_i - u| + \sum_{i\in\{i:\, y_i < u\}} (1-\tau)|y_i - u|\}$$

$$= \min_u\{\sum_{i\in\{i:\, 1,\ldots,n\}} \rho_\tau(y_i - u)\}$$

$$\rho_\tau(u) = u\left(\tau - I_{(u<0)}\right), 0 < \tau < 1$$

مثال 1

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 $\tau = 0.5$

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

U=3

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

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 loss	36	29	24	21	20	21	24	29	36

رگرسیون چندکی (چندک های شرطی)

برخلاف رگرسیون حداقل مربعات که برای برآورد میانگین شرطی از مینیمم کردن مانده ها استفاده میکند، روش رگرسیون چندکی برای برآورد چندکهای شرطی از مینیمم کردن قدر مطلق موزون مانده های نامتقارن استفاده می کند.

فرض کنید τ امین چندک شرطی به صورت $Q_{Y|X}(\tau) = X\beta_\tau$ باشد، در این صورت برای برآورد β_τ از رابطه زیر استفاده می شود :

$$min_{\beta \in R^P}\left[\sum_{i\in\{i:y_i\geq x_i'\beta\}} \tau|y_i - x_i'\beta(\tau)| + \sum_{i\in\{i:y_i<x_i'\beta\}} (1-\tau)|y_i - x_i'\beta(\tau)|\right]$$

$$= \min_{\beta \in R^P}\sum_{i=1}^{n}\rho_\tau(y_i - x_i'\beta(\tau))$$

روش عددی

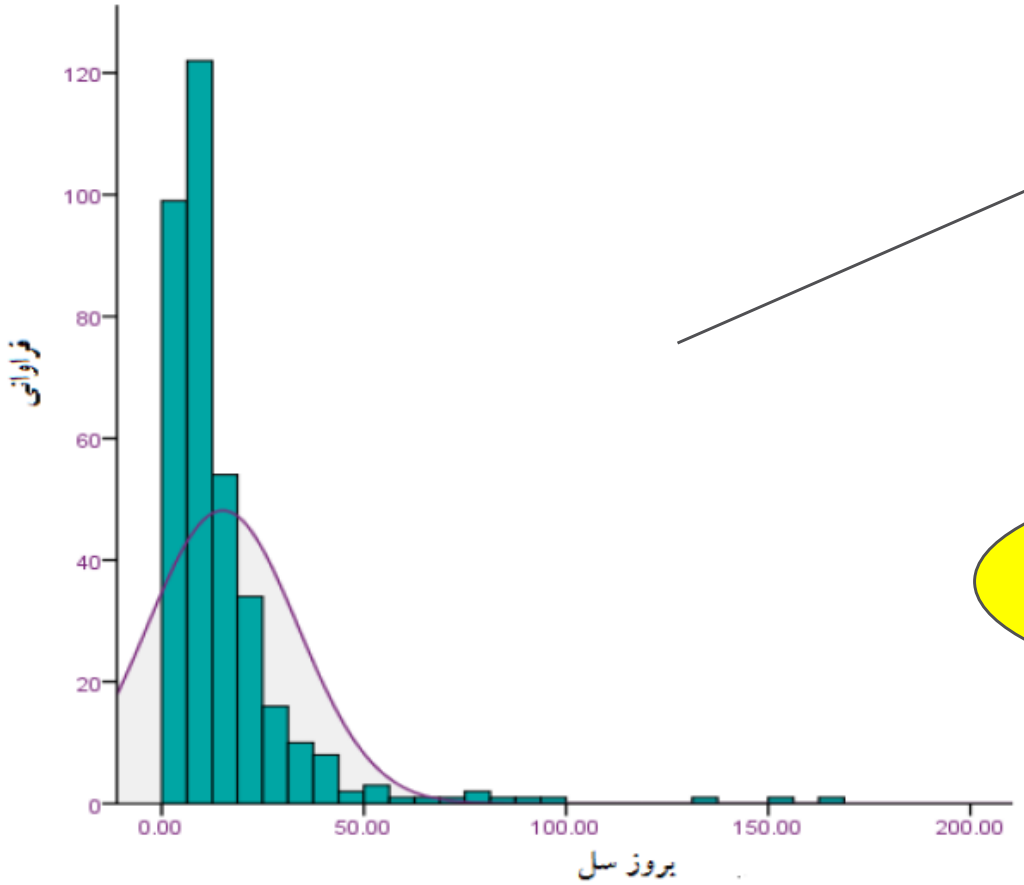
$$\rho_\tau(u) = u\left(\tau - I_{(u<0)}\right), 0 < \tau < 1$$

$$\widehat{\beta_\tau} = \arg\min \sum_{i=1}^{n} \rho_\tau (y_i - x\beta)_{\beta \epsilon R^k}$$

$$Q_{Y|X}(\tau) = X\beta_\tau$$

مثال 2

- در مطالعه ای اطلاعات مربوط به 11320 فرد مبتلا به بیماری سل در 359 شهرستان ایران تجزیه و تحلیل شد. هدف از این مطالعه بررسی اثر نرخ مهاجرت در میزان بروز بیماری سل بوده است. هیستوگرام مربوط به مشخصه میزان بروز سل به صورت زیر رسم شده است.



چوله به راست میزان بروز سل

دارای نقاط پرت حائز اهمیت

پس با توجه به این شرایط ترجیح می دهیم از رگرسیون چندکی استفاده کنیم

مدل رگرسیون چندکی برای صدک های 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95 به طور جداگانه برازش داده شد و ضرایب رگرسیون مربوط به هر صدک با استفاده از رابطه های تعریف شده برآورد شده اند. نتایج برازش مدل رگرسیونی چندکی در جدول روبرو نشان داده شده است.

$$Q_{Y|X}(\tau) = X\beta_\tau$$

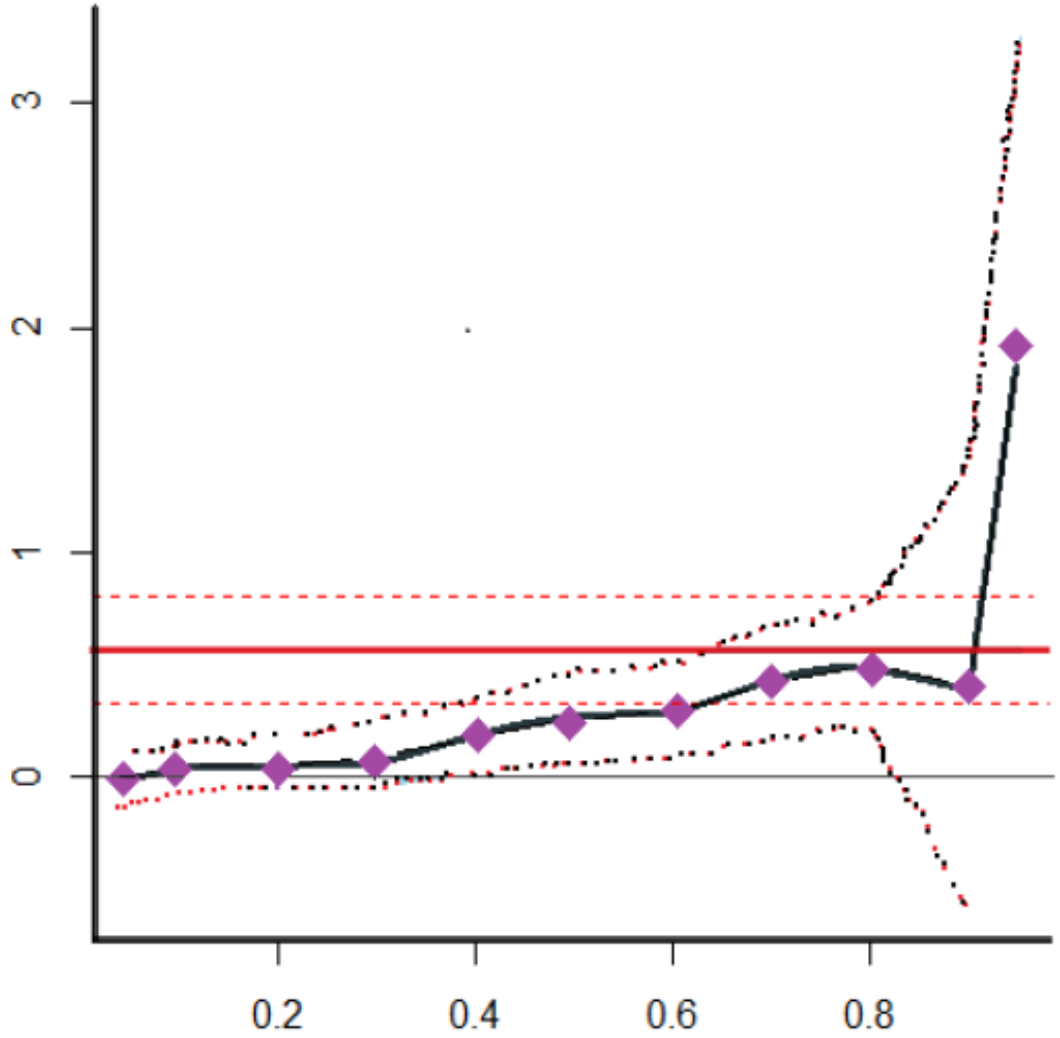
جدول ۱. نتایج برازش مدل رگرسیون چندکی برای بررسی ارتباط میزان بروز سل و نرخ مهاجرت

صدک	نرخ مهاجرت		
	$\hat{\beta}(\tau)$	خطای معیار	p
5	-۰/۰۰۵	۰/۰۷۲	۰/۹۴
10	۰/۰۵	۰/۰۷	۰/۴۴
20	۰/۰۵	۰/۰۷۱	۰/۴۵
30	۰/۰۸۹	۰/۰۹	۰/۲۶
40	۰/۱۷	۰/۰۹۸	۰/۰۷۸
50	۰/۲۶	۰/۱۲	۰/۰۳۳
60	۰/۲۹	۰/۱۲	۰/۰۱۷
70	۰/۴۲	۰/۱۵	۰/۰۰۶
80	۰/۵	۰/۱۵	۰/۰۰۱
90	۰/۴	۰/۵۷	۰/۴۸
95	۱/۹۱	۰/۸۳	۰/۰۲۲
رگرسیون حداقل مربعات	۰/۵۷	۰/۱۴	<۰/۰۰۱

مقدار $p < 0/05$ از نظر آماری معنی‌دار است.

نمودار ضرایب مدل رگرسیون چندکی

این نمودار برآوردهای نقطه ای ضرایب رگرسیون چندکی را برای هر چندک شرطی نشان میدهد.



صدک های سطوح بالاتر میزان بروز سل اثر پذیری بیشتری نسبت به نرخ مهاجرت دارد. در واقع به ازای یک واحد افزایش نرخ مهاجرت صدک های سطوح بالاتر بروز سل بیشتر از صدک های پایینی افزایش میابند.

مثال 3

در مطالعه ای هدف بررسی *نقش سن در سلامت روانی افراد در یک جامعه می باشد. این پژوهش رگرسیون چندکی رابطه سن با سلامت روان مردان و زنان را به طور جداگانه به گونه اي نشان می دهد که با رگرسیون میانگین نمایش پذیر نیست.*

- *در این مطالعه شاخص سلامت روان با پاسخ گویی افراد مورد نظر به سوالات مربوطه با نمراتی بین 28 تا 118 مشخص شده است به طوري که مقادیر پایین تر نشان دهنده سلامت روان بهتر می باشند.*

- قصد اصلی این مطالعه توصیف توزیع کل این شاخص با استفاده از نه چندك شرطی مختلف است. به عبارت دیگر اثر متغیر توضیحی سن را بر شکل توزیع شاخص سلامت روان بررسی می کنیم .

جدول زیر برآورد پارامترهای مدل های برازش داده شده در هر دو حالت رگرسیونهاي چندك از روش بیان شده و رگرسیونهاي میانگین از روش حداقل مربعات را فراهم می سازد. ستاره ها در این جدول نشان دهنده برآورد پارامترهایی هستند که در سطح 0/05 معنا دارند.

زنان	مردان	چندک
سن	سن	
۰/۰۵	−۰/۱	۰/۱
۰/۰۷	−۰/۰۴۷	۰/۲
۰/۱۵*	−۰/۰۵	۰/۳
۰/۱۱*	−۰/۰۴۵	۰/۴
۰/۱۵*	−۰/۰۳۸	۰/۵
۰/۱۴*	−۰/۰۲۲	۰/۶
۰/۲۱	۰/۰۷۳	۰/۷
۰/۲۱	۰/۰۴	۰/۸
۰/۱۶*	۰/۰۳۱	۰/۹
۰/۱۱۱*	۰/۰۳۶	OLS

$$Q_{\theta i} = \alpha_\theta + \beta_\theta x_i \qquad i = 1,..,n$$

$$\theta \sum_{i/\,y_i \le \alpha_\theta + \beta_\theta x_i} (y_i - \alpha_\theta - \beta_\theta x_i) + (1-\theta) \sum_{i/\,y_i > \alpha_\theta + \beta_\theta x_i} (y_i - \alpha_\theta - \beta_\theta x_i)$$

یافته هاي رگرسیون چندك رابطۀ سن را با وضع سلامت روانی براي زنان در تمامی موارد اظهار می دارد اما براي مردان چنین رابطه اي معنادار نیست

Archive

مثلا مدل رگرسیون چندک برای $\theta = 0.6$ برای زنان به صورت زیر برآورد شده:

$$Q_{i\cdot/۶} = ۳۱ + ۰/۱۴\text{AGE}$$

این الگو نشان می دهد که صدک 60 ام شاخص سلامت روان با افزایش یک سال سن 14 درصد افزایش می یابد. پس چندک 0/6 شاخص سلامت به طور مثال برای زنان 40 ساله 36/6 برآورد شده است.

$$Q_{\cdot/۶} = ۳۱ + ۰/۱۴ \times ۴۰ = ۳۶/۶$$

این امر آشکار می سازد که 60 درصد از زنان 40 ساله داراي شاخص سلامت روانی 36/6 و 40 درصد بالاي آن می باشند.

با ادامه همین محاسبه ها، این نکته به دست آمده که چندک 0/9 شاخص سلامت روان برای زنان 40 ساله، 53/4 می باشد.

پس می توان نتیجه گرفت که 30 درصد زنان 40 ساله دارای شاخص سلامت بین 36/6 تا 53/4 هستند.

شکل ۱: خطوط رگرسیون چندک تنظیمی برای زنان

- چنانکه ملاحظه می شود، خطوط پایین تر نسبت به خطوط بالاتر به یکدیگر نزدیک ترند این بدان معنا است که شکل توزیع متمایل به راست است .
- میزان اثر پذیری شاخص سلامت از سن زنان در صدک های بالاتر این شاخص بیشتر است.
- تحلیل ما نشان می دهد که نقش سن در سلامت روانی به طور آشکاری براي مردان و زنان متفاوت است به طوری که برای مردان این رابطه تقریبا در همه چندک ها معنادار نبود ولی برای زنان این رابطه مستقیم و برای چندک های پایینی نسبت به چندک های بالایی این میزان تاثیر پذیری کمتر بوده است.

شکل ۱: خطوط رگرسیون چندک تنظیمی برای زنان

- توجه داشته باشید رگرسیون رایج نمی تواند میزان اثر پذیری متغیر وابسته از متغیر توضیحی را در سطوح مختلف توزیع نشان دهد و فقط با پارامتر میانگین (مکان) مرتبط است در حالی که رگرسیون چندکی ما را قادر به آشنایی با شکل توزیع می نماید.

نرم افزار

برنامه زیر برای به دست آوردن رگرسیون چندکی در قالب یک مثال برای مدل های مورد نظر پیشنهاد شده است.

```
proc quantreg ci=sparsity/iid algorithm=interior(tolerance=1.e-4)
data=new;
class visit ed;
model weight = black married boy visit ed smoke
cigsper mom_age mom_age*mom_age
m_wtgain m_wtgain*m_wtgain /
quantile= 0.05 to 0.95 by 0.05
plot=quantplot;
run;
```

مراجع

- **Koenker, R.: Quantile regression for longitudinal data. J. Multivar. Anal. 91, 74–89 (2004)**
- **Koenker, R.: Quantile Regression. Cambridge University Press, NewYork (2005)**
- **Koenker, R., Bassett, G.: Regression quantiles. Econometrica 46, 33- 85(1978)**
- ارتباط بروز بیماری سل و مهاجرت با استفاده از مدل رگرسیون چندکی در ایران سال 1389 (فاطمه سروی، یدا...محرابی ، علیرضا ابدی، مهشید ناصحی)
- کاربرد رگرسیون چندك در تحلیل سلامت روانی (محمدتقی انصاري، محمد بامنی مقدم، علیرضاخوشگویانفرد، عزت اله سام آرام)
- **http://www.wikipedia.org/**

خیلی ممنون از توجه شما