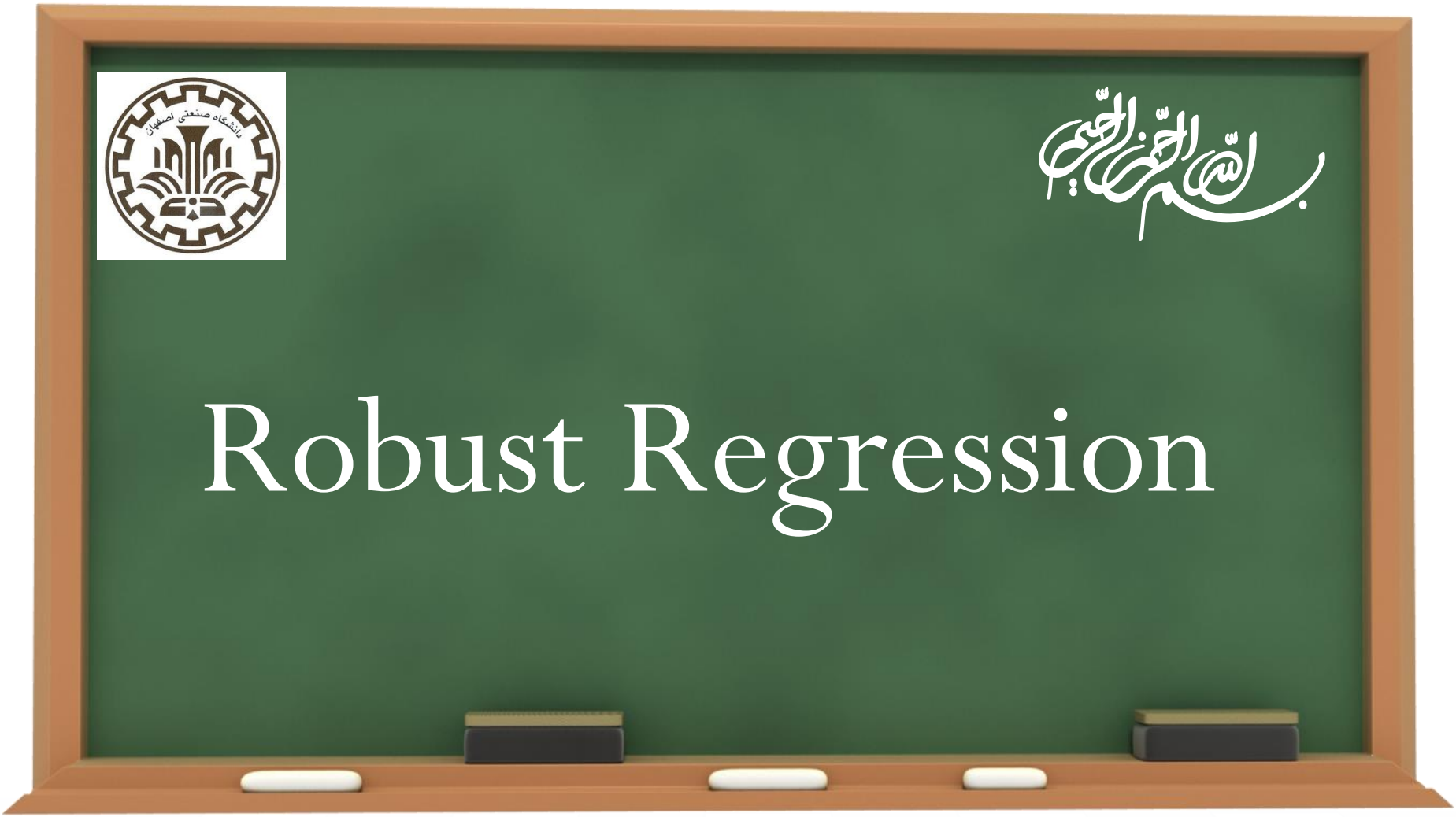

بسم الله الرحمن الرحیم
Robust Regression

نام درس : مدل خطی

-ارائه دهندگان :

فاطمه افیونیان

مهرنوش شوشتری

نیاز برای برآورد نیرومند

- در مدل رگرسیون خطی $Y=X\beta+\varepsilon$ مشاهدات و خطاها دارای توزیع نرمال هستند در نتیجه می توان از روش حداقل مربعات برای براورد β استفاده کرد. در حالیکه اگرمشاهدات، توزیع غیر نرمال داشته باشند به خصوص که شاخه انتهایی منحنی طویل تر یا ضخیم تر از نرمال باشد ممکن است روش حداقل مربعات مناسب نباشد.
- شاخه های ضخیم توزیع معمولا نقاط دورافتاده ایجاد می کند که این نقاط دورافتاده بر براورد حداقل مربعات تاثیر می گذارد .در تاثیر گذاری، دورافتاده ها برازش حداقل مربعات را تا حد زیادی به سمت خود می کشند و در نتیجه تعیین و تشخیص این دور افتاده ها مشکل می شود زیرا باقیمانده های مربوط به آن ها بطور ساختگی و مصنوعی کوچک هستند.
- روش رگرسیون نیرومند برای کاهش اثر مشاهداتی به کار می رود که اگر روش حداقل مربعات به کار گرفته شود، تاثیرگذاری بالایی خواهند داشت یعنی روش نیرومند باقیمانده های مرتبط با دورافتاده های بزرگ را کنار می گذارد که باعث می شود تشخیص دورافتاده ها خیلی ساده تر شود.
- یک رگرسیون نیرومند علاوه بر حساس نبودن نسبت به دورافتاده ها، زمانیکه مشاهدات توزیع نرمال دارد، دارای کارایی ۹۵- ۹۰ درصد نسبت به روش برآورد حداقل مربعات می باشد.

نقاط دور افتاده (Out Lier)

نقاطی هستند که $|\hat{\varepsilon}_i|$ آن ها بزرگ باشد.

مشکلات به وجود آمده از داشتن نقاط دور افتاده:

باعث افزایش SSE می شود:

$\widehat{\varsigma^2} = \frac{SSE}{n-p} \rightarrow$ MSE = MSE $\sum_{i=1}^{n} \hat{\varepsilon}_i \rightarrow$ = SSE

باعث افزایش MSE می شود:

$$T_j = \frac{\widehat{\beta_j} - 0}{\sqrt{MSE\ C_{jj}}}$$

افزایش MSE باعث کم شدن مقدار T_j می شود:

$H_0 \rightarrow t_{\alpha/2}$ n-p < $|T_j|$ If رد

در نهایت شانس رد شدن فرض H_0 کمتر می شود.

افزایش MSE باعث می شود که بعضی از متغیرهای توضیحی مهم به اشتباه حذف شوند، همچنین باعث افزایش طول فاصله اطمینان می شود در نتیجه دقت کم می شود، طول فاصله پیش بینی و فاصله اطمینان تک تک ضرایب رگرسیونی نیز افزایش می یابد.

راه های تشخیص داده های دور افتاده:

1-باقیمانده های استاندارد نشده ($\hat{\varepsilon}_i$) :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

2-باقیمانده های استاندارد شده (d_i):

$$d_i = \frac{\hat{\varepsilon}_i}{\sqrt{MSE}}$$

3-باقیمانده های استیودنت شده (r_i):

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{MSE\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{sxx}\right)}}$$

راه حل برخورد با اینگونه مشاهدات(داده های پرت):

1-حذف مرحله به مرحله داده های دور افتاده

2-استفاده از توزیع های دم کلفت تر برای برآورد ML پارامترها

3-استفاده از روش کمترین قدرمطلق خطا برای برآورد پارامترها

نقاط اهرم گون(High Leverge)

اگر بردار مشاهدات متغیرهای توضیحی داده i ام دور از سایر مشاهدات باشد، داده i ام اهرم گون است.

روش تشخیص داده های اهرم گون:

$h_{ii} > \frac{2p}{n}$; p=k+1 , n= تعداد مشاهدات

دراین صورت مشاهده i ام اهرم گون است.

نقاط اهرمی خوب: نقاطی که ازطرح کلی داده ها پیروی می کنند ودرجهت X پرت هستند.$|r_i| < 2$

نقاط اهرمی بد: نقاطی که ازطرح کلی داده ها پیروی نمی کنند ودرجهت Y پرت هستند.$|r_i| > 2$

داده های تاثیرگذار(Influential):

داده هایی هستند که خصوصیات کلیدی مدل($\hat{\beta}$ و $\widehat{\beta_j}$و σ^2و $\widehat{y_i}$) را تحت تاثیر خود قرار می دهند.

راه های تشخیص داده های تاثیرگذار:

اگر خصوصیت کلیدی مدنظر ما $\hat{\beta}$ باشد از آماره کوک برای تشخیص داده های تاثیرگذار استفاده می کنیم:

$$D_i=\frac{(\widehat{\beta_i}-\hat{\beta})' x' x(\widehat{\beta_i}-\hat{\beta})}{p\ MSE}$$

If $D_i>1$: مشاهده i ام تاثیرگذار


برآوردگرM
برآوردگر LTS
رگرسیون نرم-L1
برآوردگرهای رگرسیون نیرومند
برآوردگرLMS
برآوردگرGM
برآوردگر R
برآوردگرMM

رگرسیون نرم-L1

مینیمم کردن مجموع قدرمطلق های خطاها، معمولا مسئله رگرسیون نرم-L1 نامیده می شود. حداقل مربعات، مسئله رگرسیون نرم-L2 می باشد.

مسئله رگرسیون نرم- L1 حالت خاصی از رگرسیون نرم- L_p می باشد که در آن پارامترهای مدل طوری انتخاب می شوند که $(1\leq p\leq 2)$ $\sum_{i=1}^{n}|\varepsilon_i|^p$ مینیمم شود.

P=1.5یک انتخاب قراردادی است که منجر به براورد هایی می شود که وقتی خطاها نرمال نیستند از براورد حداقل مربعات بهتر است .وقتی هم که خطا ها نرمال باشند به کارگیریp=1.5 برآوردگرهایی را باکارایی به اندازه % 90حداقل مربعات می دهد.اما نمی توان این روش رابه عنوان جایگزین برای حداقل مربعات در نظر گرفت زیرا به شدت تحت تاثیرمشاهده پرت درجهت x است.

برآوردگر M

متداول ترین روش برآورد رگرسیون نیرومند، روش برآورد M است
در حالت کلی می توان دسته ای از برآورد گرهای نیرومند را تعریف کرد که یک تابع ρ از باقی مانده ها را مینیمم می کند:

$$min \sum_{i=1}^{n} P(e_i) = min \sum_{i=1}^{n} P(y_i - x_i' \beta)$$

که در آن x_i' نشان دهنده i- امین سطر x می باشد. اینگونه برآوردگر یک برآوردگر-M نامیده می شودکه در آن M به عنوان ماکزیمم درستنمایی است یعنی تابع ρ به تابع درستنمایی(برای یک انتخاب مناسب توزیع خطا) بستگی دارد.

برای مثال اگر روش حداقل مربعات به کار برده شود نتیجه گرفته می شود که توزیع خطا نرمال است. دراین صورت:

$$P(z) = \frac{1}{2} z^2 \quad -\infty < z < \infty$$

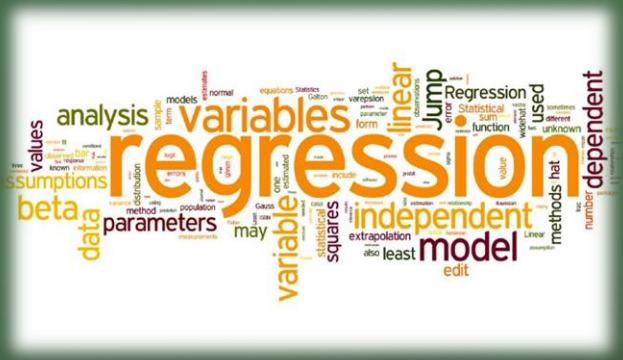
تابع p باید ویژگی های زیر را داشته باشد:

$p(0) = 0$

متقارن است: $p(e) = p(-e)$

همواره نامنفی است: $p(e) \geq 0$

یکنوا است: if $e_1 \geq e_2$ Then $p(e_1) \geq p(e_2)$



برآوردگر-M لزوما پایا مقیاس نیست (یعنی اگر باقیمانده های $y_i - \acute{x_i}\beta$ در یک ثابت ضرب شوند جواب جدید ممکن است مانند جواب قدیم نباشد). برای بدست آوردن شکل جدید پایا- مقیاس از این برآوردگر معمولا معادله زیر را حل می کنیم:

$$min \sum_{i=1}^{n} P(\frac{e_i}{s}) = min \sum_{i=1}^{n} P(\frac{y_{i-x_i'\beta}}{s})$$

که در آن s یک برآوردگر نیرومند مقیاس است. یک جواب عمومی برای s عبارتست از:

$$s = \frac{median|e_i - median(e_i)|}{0.675}$$

اگر n بزرگ باشد و توزیع خطا نرمال باشد ثابت 0.6745 ، S را تقریبا به یک برآوردگر نااریب تبدیل می کند.

برای مینیمم کردن مشتقات جزئی مرتبه اول P نسبت به (j=0,1,...,k) β_j را مساوی صفر قرار می دهیم بدین ترتیب یک شرط لازم برای بدست آوردن مینیمم حاصل می شود . با این کار دستگاهی از P= k+1 معادله تشکیل می گردد که عبارتند از :

$$\sum_{i=1}^{n} x_{ij}\, \psi\left(\frac{e_i}{s}\right) = 0$$

که در آن $\psi = \rho'$ و x_{ij} عبارتست از i امین مشاهده از j امین متغیر رگرسیونی و $x_{i0} =$ ۱ می باشد. در حالت کلی تابع ψ غیر خطی و بایستی با روش تکرار حل شود. اگر تکنیک های متعدد غیر خطی بتواند بکار گرفته شود ،حداقل مربعات موزون بطور وسیعی مورد استفاده قرار می گیرد. این روش معمولا به بیتون و توکی [۱۹۷۴] نسبت داده می شود. برای اینکه بطور مکرر حداقل مربعات دوباره موزون شده را بکار ببریم فرض کنیم که یک برآورد اولیه $\widehat{\beta_0}$ در دسترس باشد و s یک بردار مقیاس است.

در اینصورت P= k+1 معادله مذکور بصورت زیر نوشته میشود:

$$\sum_{i=1}^{n} x_{ij}\,\psi\left(\frac{e_i}{s}\right) =$$

$$\sum_{i=1}^{n} \frac{x_{ij}\psi\{[(y_i - x_i'\beta)/s]/(y_i - x_i'\beta)/s\}\{(y_i - x_i'\beta)\}}{s} = 0$$

به طوری که:

$$\sum_{i=1}^{n} x_{ij}\, w_{i0}(y_i - x_i'\beta) = 0$$

که در آن :

$$w_{i0} = \begin{cases} \dfrac{\psi[(y_i - x_i'\widehat{\beta_0})/s]}{(y_i - x_i'\widehat{\beta_0})/s} & y_i \neq x_i'\widehat{\beta_0} \\ 1 & y_i = x_i'\widehat{\beta_0} \end{cases}$$

با نماد ماتریسی بصورت زیر نوشته می شود:

$$x'w_0x\beta = x'w_0y$$

که در آن w_0 یک ماتریس قطری n×n از "وزن ها" با اعضای قطر w_{n0} ,...., w_{20} , w_{10} که درصفحه قبل مشخص گردیده اند. معادله بالارا به عنوان معادلات حداقل مربعات نرمال معمولی می شناسیم. در نتیجه برآورد قدم اول عبارتست از:

$$\widehat{\beta_1} = (x'w_0x)^{-1}x'w_0y$$

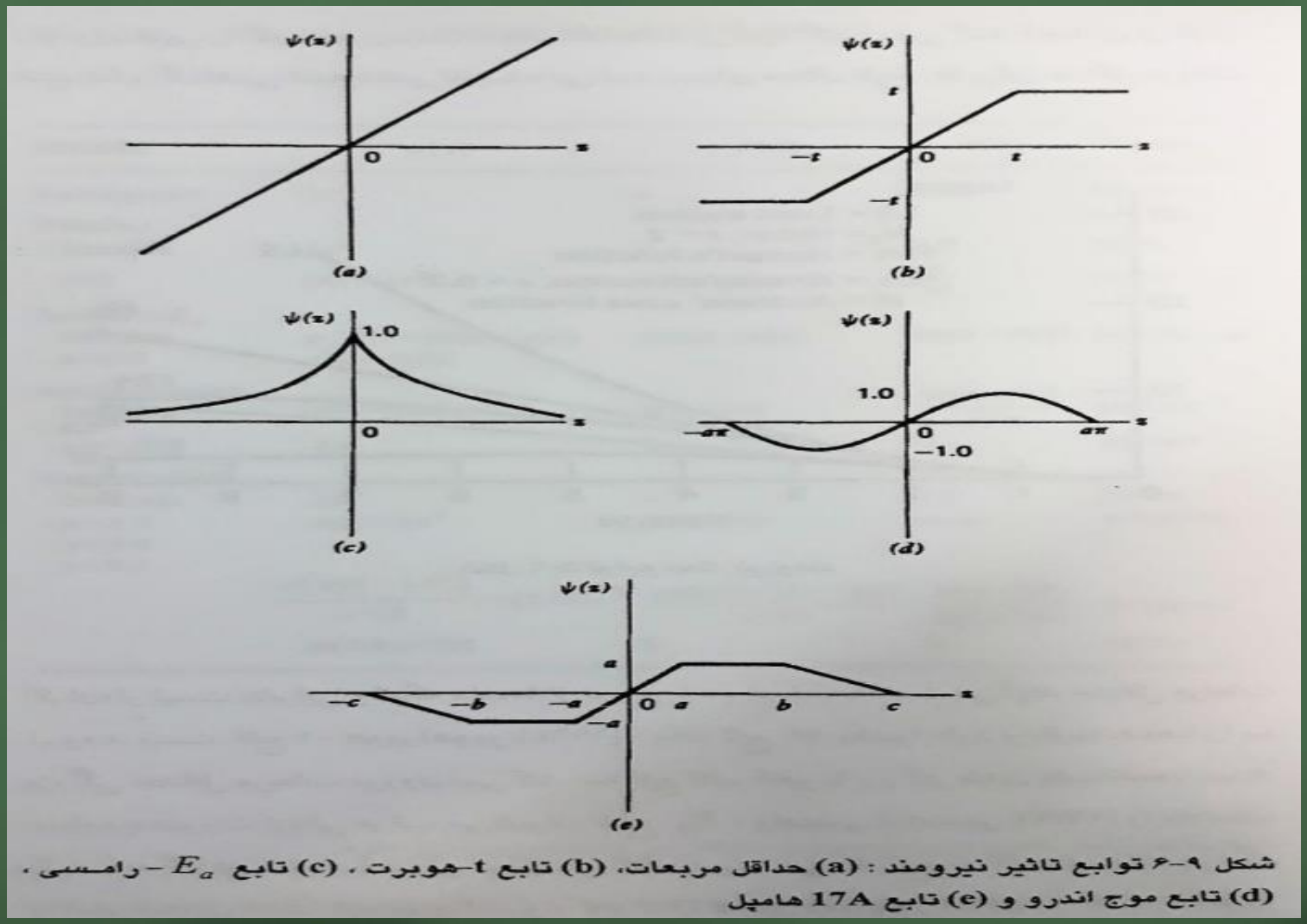
در قدم بعدی وزن ها را مجددا محاسبه می کنیم اما $\widehat{\beta_1}$ را به جای $\widehat{\beta_0}$ به کار می بریم.درحالت کلی داریم:

$$\widehat{\beta_{j+1}} = (x'w_jx)^{-1}x'w_jy$$

دراین شکل تعدادی توابع محک نیرومند متداول نشان داده شده است:

Criterion	$\rho(z)$	$\psi(z)$	$w(z)$	Range
Least squares	$\frac{1}{2}z^2$	z	1.0	$\lvert z\rvert<\infty$
Huber's t function	$\frac{1}{2}z^2$	z	1.0	$\lvert z\rvert\leq t$
$t=2$	$\lvert z\rvert t-\frac{1}{2}t^2$	$t\,\mathrm{sign}(z)$	$\frac{t}{\lvert z\rvert}$	$\lvert z\rvert>t$
Ramsay's E_a function $a=0.3$	$a^{-2}[1-\exp(-a\lvert z\rvert)\cdot(1+a\lvert z\rvert)]$	$z\exp(-a\lvert z\rvert)$	$\exp(-a\lvert z\rvert)$	$\lvert z\rvert<\infty$
Andrew's wave function	$a[1-\cos(z/a)]$	$\sin(z/a)$	$\frac{\sin(z/a)}{z/a}$	$\lvert z\rvert\leq a\pi$
$a=1.339$	$2a$	0	0	$\lvert z\rvert>a\pi$
Hampel's 17A function	$\frac{1}{2}z^2$	z	1.0	$\lvert z\rvert\leq a$
$a=1.7$	$a\lvert z\rvert-\frac{1}{2}a^2$	$a\,\mathrm{sign}(z)$	$a/\lvert z\rvert$	$a<\lvert z\rvert\leq b$
$b=3.4$				
$c=8.5$	$\frac{a(c\lvert z\rvert-\frac{1}{2}z^2)}{c-b}-(7/6)a^2$	$\frac{a\,\mathrm{sign}(z)(c-\lvert z\rvert)}{c-b}$	$\frac{a(c-\lvert z\rvert)}{\lvert z\rvert(c-b)}$	$b<\lvert z\rvert\leq c$
	$a(b+c-a)$	0	0	$\lvert z\rvert>c$

دراین شکل نیز نمودار توابع محک نیرومند ارائه شده است:



شکل ۹-۶ توابع تاثیر نیرومند : (a) حداقل مربعات، (b) تابع t-هوبرت ، (c) تابع E_a - رامسی ، (d) تابع موج اندرو و (e) تابع 17A هامپل

برآوردگرR

روش برآورد-R بر اساس رتبه ها است.دراین روش جانشین کردن یک عامل رتبه در تابع حداقل مربعات $S(\beta) = \sum_{i=1}^{n}(y_i - \acute{x_i}\beta)^2$ به وسیله رتبه هایش را مورد توجه قرار می دهیم. اگر R_i رتبه y_i - $\acute{x_i}$ β باشد می خواهیم

$\sum_{i=1}^{n} (y_i - \acute{x_i}\beta)R_i$ مینیمم شود. رتبه ها را که اعداد صحیح هستند با تابع زیان a(i) ; (i=1,2,...,n) جانشین می کنیم بنابراین تابع هدف برابر می شود با:

$$min_\beta \sum_{i=1}^{n} (y_i - x_i'\beta)a(R_i)$$

برآوردگرR ازنظر محاسباتی ازبرآوردگرM مشکل تر است اما درشرایط واقعی به طورمجانبی با برآوردگرM معادل است.

برآوردگر LTS

برآوردگر کمترین مربعات پیراسته:

$$\hat{\beta} = \sum_{i=1}^{n} e_{(i)}^2$$

که درآن:

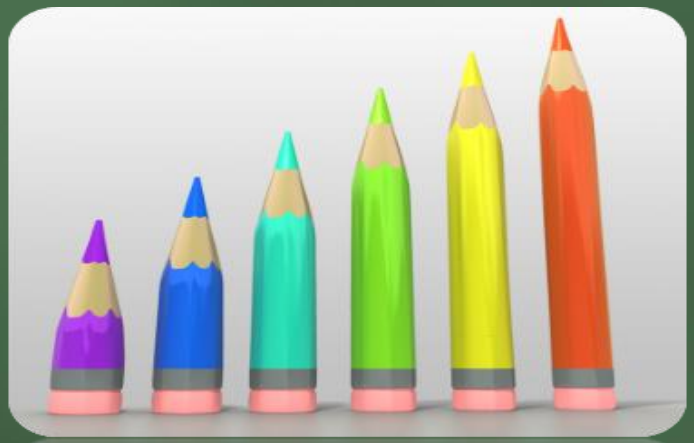
$$e_{(1)}^2 \leq e_{(2)}^2 \leq \ldots \leq e_{(n)}^2$$

دراین رابطه بزرگترین عددصحیح کوچک تر یا مساوی $\frac{n}{2} + 1$ است. اگر خطاها دارای توزیع نرمال باشند این برآوردگر از برآوردگر حداقل مربعات ناکاراتر است.

برآوردگر LMS

دراین روش میانه مربعات باقی مانده ها را بدست می آوریم سپس با مینیمم کردن آن برآورد ضرایب را بدست می آوریم:

$$\hat{\beta} = \min \text{med} \, (y_i - \acute{x}_i \beta)^2$$



براوردگر GM

برآورد GM درسال ۱۹۷۵ توسط مالوس معرفی شد. اساس این روش مبتنی بر به کارگیری توابع وزنی است و برآورد از حل معادله زیر بدست می آید:

$$\sum_{i=1}^{n} w_i \, \psi\left(\frac{r_i(\hat{\beta})}{\hat{\sigma}}\right) x_i = 0$$

اغلب برای سادگی σ را معلوم فرض کرده و یا برآورد آن را در معادله قرار می دهیم.

برآوردگرMM

این برآوردگر از حل معادله زیر به دست می آید:

$$L(\beta)= \sum_{i=1}^{n} \rho_i\left(\frac{r_i(\hat{\beta})}{\widehat{\sigma_n}}\right)$$

که درآن:

$$\rho<\rho_0 \rightarrow L(\hat{\beta}) \leq L(\beta_0)$$

یک مثال واقعی

مقایسه نتایج مدل رگرسیون معمولی و رگرسیون نیرومند در مدل بندی عوامل مرتبط با بیماری پره دیابت:

1. گروه بهداشت عمومی، دانشگاه علوم پزشکی نیشابور
2. گروه آموزشی آمار، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد
3. گروه آمار زیستی، مرکز تحقیقات عوامل اجتماعی موثر بر سلامت، دانشگاه علوم پزشکی مشهد
4. گروه اپیدمیولوژی، مرکز تحقیقات عوامل اجتماعی موثر بر سلامت، دانشگاه علوم پزشکی مشهد
5. گروه جمعیت شناسی، کمیته تحقیقات دانشجویی، دانشگاه آزاد اسلامی واحد تهران

زمینه و هدف: با توجه به اینکه خطر ابتلا به دیابت در افراد پره‌دیابتیک بسیار بالا است، تعیین عوامل موثر بر پره‌دیابت دارای اهمیت می‌باشد. این مطالعه با هدف مقایسه نتایج مدل رگرسیون معمولی و رگرسیون نیرومند در مدل‌بندی عوامل مرتبط با بیماری پره‌دیابت انجام شد.

روش بررسی: این مطالعه که از نوع مقطعی-تحلیلی است روی ۶۴۶۰ نفر از افراد بالای ۳۰ سال، شرکت‌کننده در طرح غربالگری دیابت دانشگاه علوم پزشکی مشهد، از مهر تا آذر ۱۳۸۹ انجام شد. با توجه به میزان قندخون ناشتای افراد، ۵۴۱۴ نفر سالم و ۱۰۴۶ نفر به‌عنوان پره‌دیابتیک شناسایی شدند. سن، جنس، نمایه توده بدن، فشارخون سیستولیک، فشارخون دیاستولیک و نسبت کمر به باسن در مورد هر فرد اندازه‌گیری شد. مدل رگرسیون معمولی روی داده‌ها برازش شد. سپس داده‌های پرت مشخص و سه مدل نیرومندMallow ، WBYو BY برازش شد. آنگاه مدل‌ها با هم مقایسه گردیدند.

یافته‌ها: متغیرهای سن، نمایه توده بدن و فشارخون سیستولیک در همه مدل‌ها از لحاظ آماری معنادار شدند (P<0.01)و متغیر نسبت کمر به باسن معنادار نشد (P>0.1)تعداد ۵۵۲ داده‌ی پرت با خطای بد رده‌بندی در مدل معمولی وجود داشت. مقادیر کای دو پیرسون و سطح زیرمنحنی را که در مدل Mallowبه‌طور تقریبی فرقی با مدل معمولی نداشت. اما در مدل‌های WBY و BY به نسبت بیشتر بود.

نتیجه‌گیری: با توجه به نتایج این پژوهش، سن بالا، نمایه توده بدنی و فشارخون بالا در ابتلا به بیماری پره‌دیابت موثر می‌باشند. همچنین مدل‌های رگرسیون نیرومند WBY و BY برازش بهتر و توان پیشگویی بالاتری نسبت به رگرسیون معمولی در مدل‌بندی عوامل گفته‌شده در ارتباط با پره‌دیابت دارند.

منابع

مقدمه ای بر تحلیل رگرسیون خطی

(داگلاس مونتگمری و الیزابت پک)

Linear Regression Analysis

(Seber, 2003)

رگرسیون استوار(نیرومند)

(سید مهدی امیرجهانشاهی و حسینعلی نیرومد)

Introduction to Robust Estimation Techniques

(Andreas Ruckstuhl,2016)

با تشکر از
همراهی و توجه تان