
بسم الله الرحمن الرحيم

ROBUST REGRESION

رگرسیون نیرومند

ارائه دهندگان:

نیلوفر شاه میوه

مریم داوری

پاییز 97

مدل خطی ساده:

$$Y = X\beta + \varepsilon$$

$$\rightarrow \varepsilon \sim N(0, \delta I)$$
$$\rightarrow Y \sim N(X\beta, \delta I)$$

که در این مدل خطاها دارای توزیع نرمال و مشاهدات Y نیز به صورت نرمال توزیع شده اند بنابر این از روش حداقل مربعات برای براورد β استفاده می کنیم

$$\hat{\beta} = (\acute{X}X)^{-1}\acute{X}Y$$

- اگر مشاهدات 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 معمولا نقاط دور افتاده ایجاد می کند که این نقاط دور افتاده بر براورد حداقل مربعات تاثیر می گذارد. دور افتاده ها برازش حداقل مربعات را تا حد زیادی به سمت خود می کشاند و در نتیجه تعیین و تشخیص این دور افتاده ها مشکل می شود

به روش حداقل مربعات مسئله رگرسیون نرم- L_2 نیز گفته میشود.

- رگرسیون نیرومند برای کاهش اثر مشاهداتی به کار می رود که اگر روش حداقل مربعات به کار گرفته شود اثر بالایی خواهند داشت.

- رگرسیون نیرومند باقی مانده های مرتبط با دور افتاده های بزرگ را کنار می گذارد که بدین وسیله تشخیص دور افتاده ها خیلی ساده تر می شود.

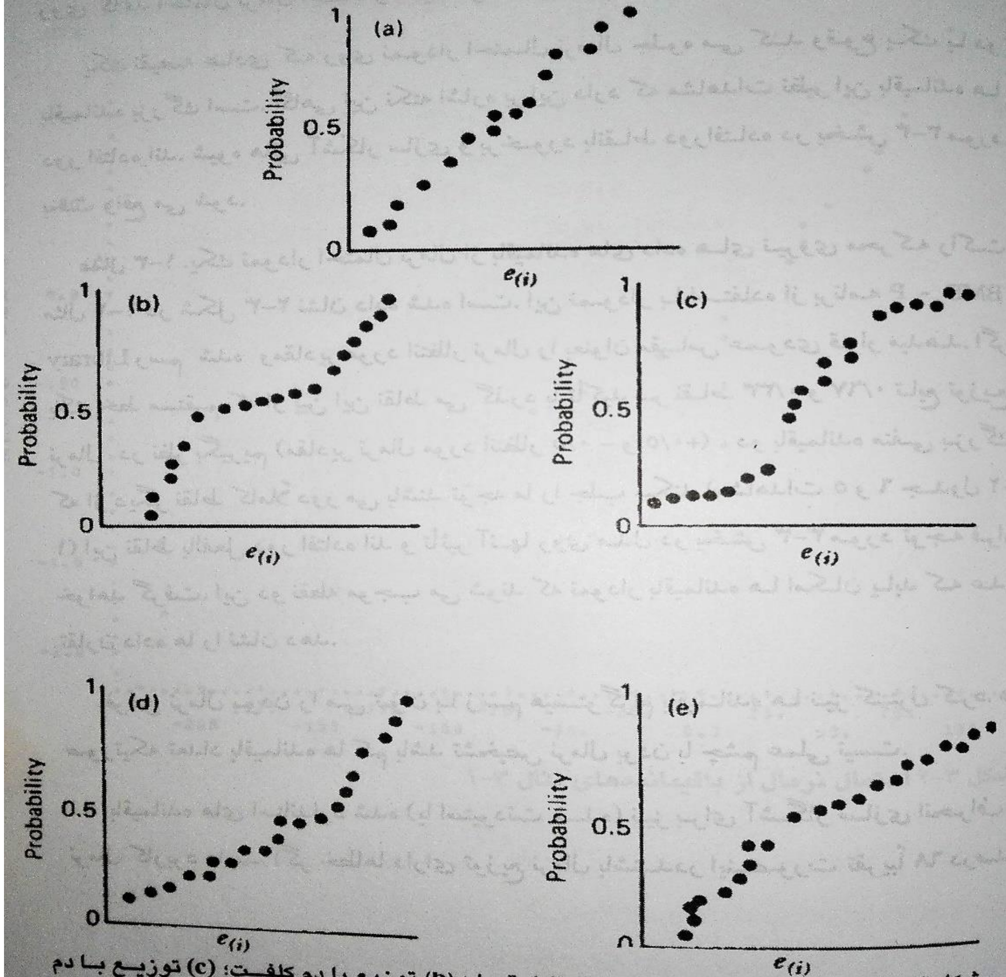
- یک رگرسیون نیرومند علاوه بر حساس نبودن نسبت به دور افتاده ها وقتی مشاهدات توزیع نرمال دارد دارای کارایی ۹۰–۹۵ درصدنسبت به روش براورد حداقل مربعات می باشد.

راه های تشخیص داده پرت:

نمودار پراکنش

مانده های استیودنت شده

آماره کوک



شکل ۳-۱ نمودارهای احتمال نرمال : (a) قابل قبول؛ (b) توزیع با دم کلفت؛ (c) توزیع با دم باریک ؛ (d) عدم تقارن مثبت ؛ (e) عدم تقارن منفی.

آماره کوک را به فرم زیر تعریف می کنیم:

$$D_i = \frac{r_i^2}{2} \times \frac{h_{ii}}{1 - h_{ii}}$$

اگر داشته باشیم

$D_i < 1$

$D_i > \frac{4}{n-2}$

داده پرت است

که درآن:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})}{S_{xx}}$$

$$r_i = \frac{e_i}{\sqrt{MS_E[1 - (h_{ii})]}}$$

که درآن متوسط مجموع مربعات خطا به صورت زیر بدست می آید:

$$MS_E = \frac{SS_E}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

باقی مانده های استیودنت شده را به فرم زیر داریم:

$$r_i = \frac{e_i}{\sqrt{MS_E[1-(h_{ii})]}}$$

- $h_{ii} > \frac{4}{n}$ → اهرمی
  - $|r_i| > 2$ → بد → پرت
  - $|r_i| < 2$ → خوب
- $h_{ii} < \frac{4}{n}$ → معمولی
  - $|r_i| > 2$ → پرت
  - $|r_i| < 2$ → معمولی

علت وجود داده پرت:

اندازه گیری غلط و ثبت یا انتقال نادرست داده ها–از کار افتادن وسیله اندازه گیری و ...

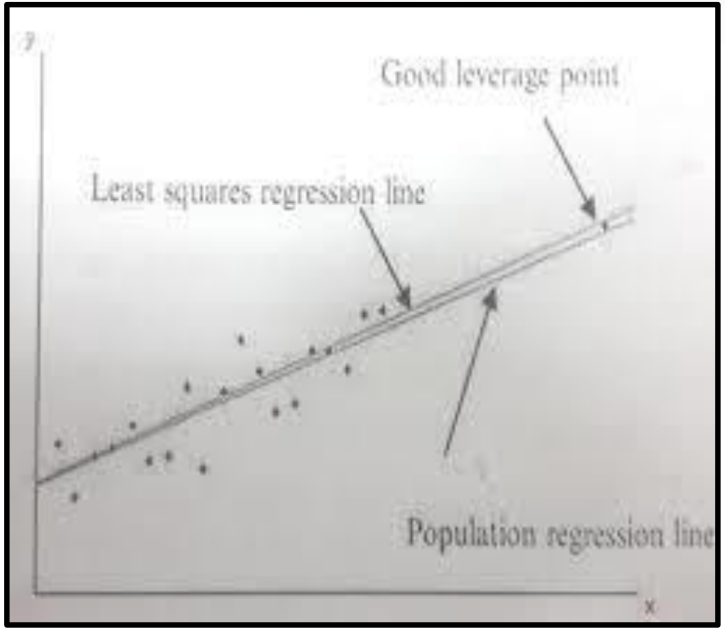
نقاط اهرمی:

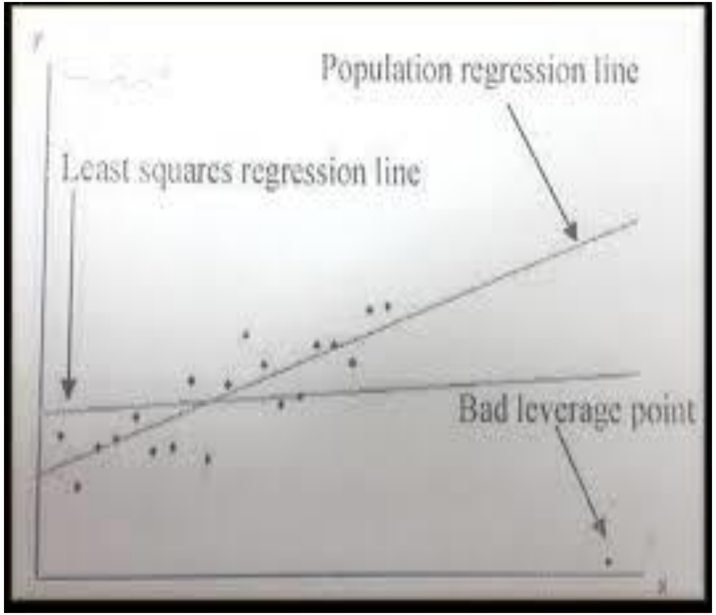
زمانی که داده ها دارای چند مو لفه غیر طبیعی X باشند به این داده ها نقاط اهرمی گفته
می شود.

نقاط اهرمی خوب: از طرح کلی داده ها پیروی می کند یعنی در جهت X پرت است.

نقاط اهرمی بد: از طرح کلی داده ها پیروی نمی کند و در جهت y پرت است.

نقاط اهرمی بد مانند اهرم عمل کرده و خط کمترین مربعات خطا را به شکل قابل توجهی تحت تاثیر خود قرار می دهد.





روش های برآورد یابی با رگرسیون نیرومند

- روش رگرسیون نرم- L_1:
- برآورد GM
- برآورد MM
- برآورد LMs:
- برآورد LTS:
- برآورد مکانی یا برآورد- L :
- برآورد –R :
- روش برآورد M:

روش رگرسیون نرم- L_1:

- حالت خاصی از رگرسین نرم – L_p می باشدکه در آن پارامتر های مدل طوری انتخاب می شوند که $\sum_{i=1}^{n}|\varepsilon_i|^p \;\; (1 \le p \le 2)$ می نیمم شود.

یعنی برای P=1 به جای میینیمم کردن ε_i^2 از قدر مطلق $|\varepsilon_i|$ استفاده می شود.

$$min \sum_{i=1}^{n} |\varepsilon_i|$$

- قدر مطلق $|\varepsilon_i|$در مقابل ε_i^2حساسیت کمتری دارد.

- p=1.5 یک انتخاب قراردادی است که منجر به براورد هایی می شود که وقتی خطا ها نرمال نیستند از براورد حداقل مربعات بهتر است. وقتی هم که خطا ها نرمال باشندبه کار گیری p=1.5 براوردگرهایی را با کارایی به اندازه ٪۹۰ حداقل مربعات می دهد.اما نمی توان روش L_1 را به عنوان جایگزین برای حداقل مربعات در نظر گرفت زیرا به شدت تحت تاثیر مشاهده پرت در جهت X است.

روش برآورد M:

متداول ترین روش برآورد رگرسیون نیرومند روش M است

در حالت کلی می توان دسته ای از برآورد گر های نیرومند را تعریف کرد که یک تابع ρاز

باقی مانده ها را می نیمم می کند .یعنی به جای ε_i^2 از یک تابع $\rho(\varepsilon_i)$استفاده کنیم.

مدل خطی زیر را در نظر بگیرید:

$$y_i = \beta_0 + \beta_2 x_{i2} + \ \ldots\ldots\ldots \beta_k x_{ik} + \varepsilon_i$$

که در آن

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

با توجه به براورد M داریم:

$$min \sum_{i=1}^{n} \rho(e_i) = min \sum_{i}^{n} \rho(y_i - \acute{x}_i \beta)$$

که در آن $\acute{x}_i$ نشان دهنده i امین سطر از ماتریس X می باشد این گونه براورد گر یک برآورد گر M نامیده می شود که در آن M برای درست نمایی ماکزیمم قرار داده می شود بدین معنی که تابع ρ به تابع درستنمایی برای 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) در این صورت:

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

◄ $z = \sum_{i=1}^{n}(c_i + d_i)$ که c_i و d_i انحرافات مثبت و منفی حول خط برازش شده است

$z = \sum_{i=1}^{n}(c_i + d_i)$ که c_i و d_i انحرافات مثبت و منفی حول خط برازش شده

یک ρ قابل قبول باید شرایط زیر را داشته باشد:

۱- همواره نامنفی است $\rho(e) \geq 0$

۲- $\rho(0) = 0$

۳-متقارن است $\rho(e) = \rho(-e)$

۴-یکنوا است $if\ e_i \geq \acute{e}_i\ then\ \rho(e_i) \geq \rho(\acute{e}_i)$

براورد گر M لزوما پایا مقیاس نیست (یعنی اگر باقی مانده های $y_i - \acute{x}_i\beta$ در یک ثابت ضرب شوند جواب جدید $min \sum_{i=1}^{n} \rho(e_i) = min \sum_{i}^{n} \rho(y_i - \acute{x}_i\beta)$ ممکن است مانند جواب قدیم نباشد).

برا ی بدست آوردن شکل جدید پایا مقیاس از براورد گر معمولا معادله زیر را حل می کنیم .

$$min\sum_{i=1}^{n}\rho(\frac{e_i}{s}) = min\sum_{i}^{n}\rho(\frac{y_i - \acute{x}_i\beta}{s})$$

که در آن s یک براورد گر نیرومند مقیاس است و برابر است با:

$$s = \frac{median|e_i - median(e_1, \dots, e_n)|}{0.675}$$

تعریف می کنیم تابع $\psi = \acute{\rho}$ مشتق تابع ρ نسبت به $\beta_j(j = 0,1, \dots ,k)$

برای مینیمم کردن $\sum_i^n \rho(\frac{y_i-\acute{x}_i\beta}{s})$ مشتق ρ نسبت به β را مساوی صفر قرار می دهیم.

$$\sum_i^n x_{ij}\psi(\frac{y_i-\acute{x}_i\beta}{s}) = 0$$

که $\psi = \acute{\rho}$

تعریف می کنیم :

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

بنابراین :

$$\sum_i^n x_{ij}\psi(\frac{y_i - \acute{x}_i\beta}{s})$$

$$= \sum_{i=1}^{n} \frac{x_{ij}\,\psi\{[(y_i - \acute{x}_i\widehat{\beta_0})/s]/(y_i - \acute{x}_i\widehat{\beta_0})/s\}(y_i - \acute{x}_i\widehat{\beta_0})}{s} = 0$$

پس:

$$\sum_{i=1}^{n} x_{ij}w_{i0}(y_i - \acute{x}_i\beta) = 0 \quad , \qquad j = 0,1,\dots,k$$

که با نماد ماتریسی به صورت زیر می شود:

$$\acute{X}W_0X\beta = \acute{X}W_0y$$

که در آن w_0 یک ماتریس قطری $n \times n$ از وزن ها با اعضای قطر $w_{10}, w_{20}, \dots, w_{n0}$ است

برآورد قدم اول عبارتست از :

$$\widehat{\beta_1} = \left(\acute{X} W_0 X\right)^{-1} \acute{X} w_0 y$$

در قدم بعدی وزن ها را بار دیگر محاسبه می کنیم اما $\widehat{\beta_1}$ را به جای $\widehat{\beta_0}$ به کار می بریم و به طور کلی می توان گفت:

$$\widehat{\beta_{j+1}} = \left(\acute{X} W_j X\right)^{-1} \acute{X} W_j y$$

شکل زیر تعدادی از توابع محک نیرومند معروف را ارائه میدهد

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

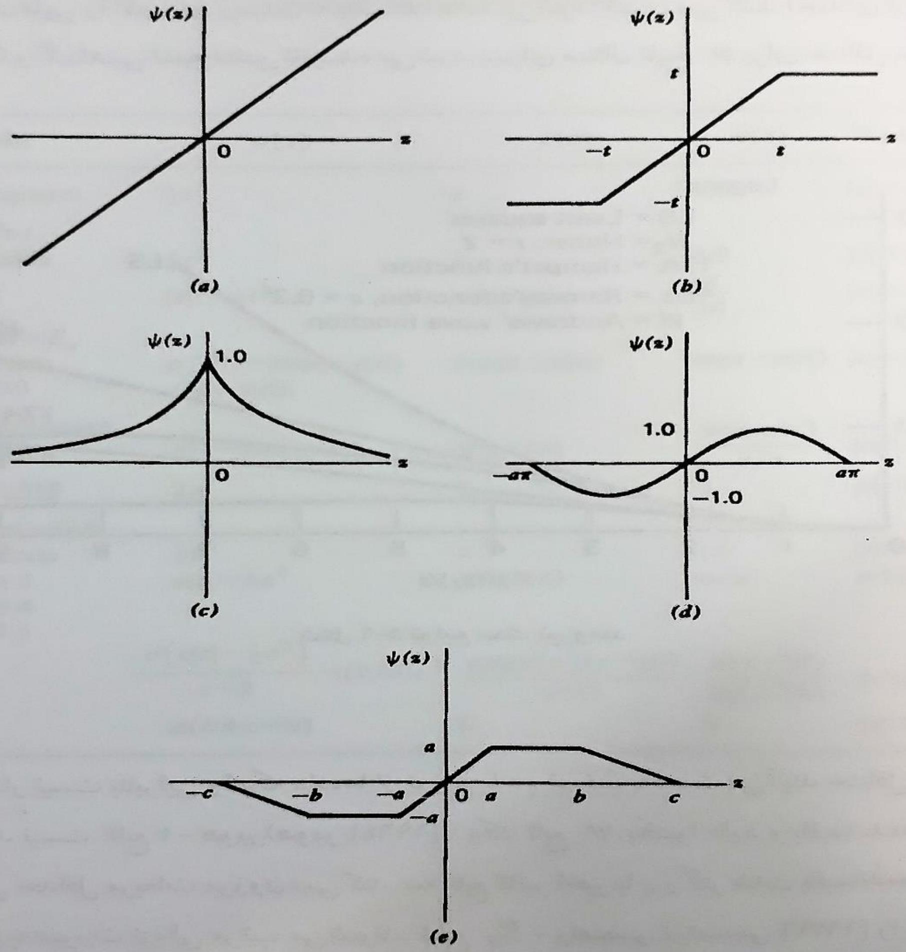
نمودار توابع

محک های

نیرومند نیز در

شکل مقابل

ارائه شده:



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

در جدول های بعدی میتوانید مثال هایی که به روش حداقل مربعات و رگرسیون نیرومند ارزیابی شده اند را مشاهده کنید:

جدول ۹–۱۰ برازش حداقل مربعات به داده‌های زمان تحویل

Observation i	y_i	$\hat{y}_i$	e_i	Weight
1	.166800E+02	.217081E+02	−.502808E+01	.100000E+01
2	.115000E+02	.103536E+02	.114639E+01	.100000E+01
3	.120300E+02	.120798E+02	−.497937E−01	.100000E+01
4	.148800E+02	.995565E+01	.492435E+01	.100000E+01
5	.137500E+02	.141944E+02	−.444398E+00	.100000E+01
6	.181100E+02	.183996E+02	−.289574E+00	.100000E+01
7	.800000E+01	.715538E+01	.844624E+00	.100000E+01
8	.178300E+02	.166734E+02	.115660E+01	.100000E+01
9	.792400E+02	.718203E+02	.741971E+01	.100000E+01
10	.215000E+02	.191236E+02	.237641E+01	.100000E+01
11	.403300E+02	.380925E+02	.223749E+01	.100000E+01
12	.210000E+02	.215930E+02	−.593041E+00	.100000E+01
13	.135000E+02	.124730E+02	.102701E+01	.100000E+01
14	.197500E+02	.186825E+02	.106754E+01	.100000E+01
15	.240000E+02	.233288E+02	.671202E+00	.100000E+01
16	.290000E+02	.296629E+02	−.662928E+00	.100000E+01
17	.153500E+02	.149136E+02	.436360E+00	.100000E+01
18	.190000E+02	.155514E+02	.344862E+01	.100000E+01
19	.950000E+01	.770681E+01	.179319E+01	.100000E+01
20	.351000E+02	.408880E+02	−.578797E+01	.100000E+01
21	.179000E+02	.205142E+02	−.261418E+01	.100000E+01
22	.523200E+02	.560065E+02	−.368653E+01	.100000E+01
23	.187500E+02	.233576E+02	−.460757E+01	.100000E+01
24	.198300E+02	.244029E+02	−.457285E+01	.100000E+01
25	.107500E+02	.109626E+02	−.212584E+00	.100000E+01

$\hat{\beta}_0 = 2.3412$
$\hat{\beta}_1 = 1.6159$
$\hat{\beta}_2 = 0.014385$

جدول ۹-۱۱ برازش نیرومند (t =2 هوبر) داده‌های زمان تحویل

Observation i	y_i	$\hat{y}_i$	e_i	Weight
1	.166800E+02	.217651E+02	−.508511E+01	.639744E+00
2	.115000E+02	.109809E+02	.519115E+00	.100000E+01
3	.120300E+02	.126296E+02	−.599594E+00	.100000E+01
4	.148800E+02	.105856E+02	.429439E+01	.757165E+00
5	.137500E+02	.146038E+02	−.853800E+00	.100000E+01
6	.181100E+02	.186051E+02	−.495085E+00	.100000E+01
7	.800000E+01	.794135E+01	.586521E−01	.100000E+01
8	.178300E+02	.169564E+02	.873625E+00	.100000E+01
9	.792400E+02	.692795E+02	.996050E+01	.327017E+00
10	.215000E+02	.193269E+02	.217307E+01	.100000E+01
11	.403300E+02	.372777E+02	.305228E+01	.100000E+01
12	.210000E+02	.216097E+02	−.609734E+00	.100000E+01
13	.135000E+02	.129900E+02	.510021E+00	.100000E+01
14	.197500E+02	.188904E+02	.859556E+00	.100000E+01
15	.240000E+02	.232828E+02	.717244E+00	.100000E+01
16	.290000E+02	.293174E+02	−.317449E+00	.100000E+01
17	.153500E+02	.152908E+02	.592377E−01	.100000E+01
18	.190000E+02	.158847E+02	.311529E+01	.100000E+01
19	.950000E+01	.845286E+01	.104714E+01	.100000E+01
20	.351000E+02	.399326E+02	−.483256E+01	.672828E+00
21	.179000E+02	.205793E+02	−.267929E+01	.100000E+01
22	.523200E+02	.542361E+02	−.191611E+01	.100000E+01
23	.187500E+02	.233102E+02	−.456023E+01	.713481E+00
24	.198300E+02	.243238E+02	−.449377E+01	.723794E+00
25	.107500E+02	.115474E+02	−.797359E+00	.100000E+01

$\hat{\beta}_0 = 3.3736$

$\hat{\beta}_1 = 1.5282$

$\hat{\beta}_2 = 0.013739$

جدول ۹-۱۲ برازش نیرومند (تابع $E_{0.3}$ رامسی) داده‌های زمان تحویل

Observation i	y_i	$\hat{y}_i$	e_i	Weight
1	$.166800E+02$	$.218009E+02$	$-.512091E+01$	$.388862E+00$
2	$.115000E+02$	$.112454E+02$	$.254584E+00$	$.953571E+00$
3	$.120300E+02$	$.128682E+02$	$-.838215E+00$	$.857082E+00$
4	$.148800E+02$	$.108415E+02$	$.403846E+01$	$.474482E+00$
5	$.137500E+02$	$.147670E+02$	$-.101696E+01$	$.829548E+00$
6	$.181100E+02$	$.186905E+02$	$-.580544E+00$	$.898687E+00$
7	$.800000E+01$	$.826846E+01$	$-.268460E+00$	$.952527E+00$
8	$.178300E+02$	$.170677E+02$	$.762255E+00$	$.868454E+00$
9	$.792400E+02$	$.682279E+02$	$.110121E+02$	$.131649E+00$
10	$.215000E+02$	$.194307E+02$	$.206932E+01$	$.682987E+00$
11	$.403300E+02$	$.369364E+02$	$.339359E+01$	$.535303E+00$
12	$.210000E+02$	$.216035E+02$	$-.603534E+00$	$.895015E+00$
13	$.135000E+02$	$.132081E+02$	$.291877E+00$	$.947139E+00$
14	$.197500E+02$	$.189862E+02$	$.763769E+00$	$.868651E+00$
15	$.240000E+02$	$.232651E+02$	$.734923E+00$	$.873345E+00$
16	$.290000E+02$	$.291901E+02$	$-.190116E+00$	$.964755E+00$
17	$.153500E+02$	$.154431E+02$	$-.931215E-01$	$.983506E+00$
18	$.190000E+02$	$.160129E+02$	$.298707E+01$	$.576113E+00$
19	$.950000E+01$	$.875713E+01$	$.742875E+00$	$.871123E+00$
20	$.351000E+02$	$.395347E+02$	$-.443471E+01$	$.440968E+00$
21	$.179000E+02$	$.205893E+02$	$-.268928E+01$	$.609354E+00$
22	$.523200E+02$	$.534802E+02$	$-.116016E+01$	$.806251E+00$
23	$.187500E+02$	$.232921E+02$	$-.454212E+01$	$.432732E+00$
24	$.198300E+02$	$.243045E+02$	$-.447455E+01$	$.437999E+00$
25	$.107500E+02$	$.117882E+02$	$-.103817E+01$	$.826364E+00$

$\hat{\beta}_0 = 3.8021$
$\hat{\beta}_1 = 1.4894$
$\hat{\beta}_2 = 0.013523$

جدول ۹-۱۳ برازش نیرومند (موج آندروز با $a = 1.48$) داده‌های زمان تحویل

Observation i	y_i	$\hat{y}_i$	e_i	Weight
1	.166800E+02	.216430E+02	−.496300E+01	.427594E+00
2	.115000E+02	.116923E+02	−.192338E+00	.998944E+00
3	.120300E+02	.131457E+02	−.111570E+01	.964551E+00
4	.148800E+02	.114549E+02	.342506E+01	.694894E+00
5	.137500E+02	.152191E+02	−.146914E+01	.939284E+00
6	.181100E+02	.188574E+02	−.747381E+00	.984039E+00
7	.800000E+01	.890189E+01	−.901888E+00	.976864E+00
8	.178300E+02	.174040E+02	.425984E+00	.994747E+00
9	.792400E+02	.660818E+02	.131582E+02	0.
10	.215000E+02	.192716E+02	.222839E+01	.863633E+00
11	.403300E+02	.363170E+02	.401296E+01	.597491E+00
12	.210000E+02	.218392E+02	−.839167E+00	.980003E+00
13	.135000E+02	.135744E+02	−.744338E−01	.999843E+00
14	.197500E+02	.189979E+02	.752115E+00	.983877E+00
15	.240000E+02	.232029E+02	.797080E+00	.981854E+00
16	.290000E+02	.286336E+02	.366350E+00	.996228E+00
17	.153500E+02	.158247E+02	−.474704E+00	.993580E+00
18	.190000E+02	.164593E+02	.254067E+01	.824146E+00
19	.950000E+01	.946384E+01	.361558E−01	.999956E+00
20	.351000E+02	.387684E+02	−.366837E+01	.655336E+00
21	.179000E+02	.209308E+02	−.303081E+01	.756603E+00
22	.523200E+02	.523766E+02	−.566063E−01	.999908E+00
23	.187500E+02	.232271E+02	−.447714E+01	.515506E+00
24	.198300E+02	.240095E+02	−.417955E+01	.567792E+00
25	.107500E+02	.123027E+02	−.155274E+01	.932266E+00

$\hat{\beta}_0 = 4.6532$
$\hat{\beta}_1 = 1.4582$
$\hat{\beta}_2 = 0.012111$

برآورد –R :

روش برآورد–R بر اساس رتبه هایش می باشد.

اگر R_i رتبه $y_i - \acute{x}_i\beta$ با شد در این صورت $\sum_{i=1}^{n}(y_i - \acute{x}_i\beta)R_i$ می نیمم میشود.

رتبه ها را که اعداد صحیح می باشند با $i = 1,2, \dots ,n$ $a(i)$ جانشین می کنیم و بنابراین تابع هدف چنین می شود.

$$min\sum_{i=1}^{n}(y_i - \acute{x}_i\beta)\,a(R_i)$$

اگر تابع امتیاز را با رتبه برابر قرار دهیم یعنی $a(i) = i$ نتیجه نمره های ویلککسون می شود.

امکان دیگر به کار گیری امتیاز میانه است یعنی:

$$a(i) = -1 \quad \text{اگر} \quad i < (n+1)/2$$
$$a(i) = 1 \quad \text{اگر} \quad i > (n+1)/2$$

براوردR از نظر محاسباتی از براوردگر M مشکل تر است اما به طور مجانبی

با براوردگر M معادل است.

برآورد مکانی یا برآورد– L :

برآوردگرهای –L براساس آماره های مرتب پایه گذاری شده اند.برای مثال اگر بخواهیم پارامتر مکانی یک توزیع را با استفاده از نمونه تصادفی $x_1, x_2, \dots, x_n$ برآورد کنیم آماره های مرتب این نمونه عبارتند از:

$$x_{[1]} \le x_{[2]} \le \cdots \le x_{[n-1]} \le x_{[n]}.$$

میانه نمونه یک برآورد گر–L خواهد بود زیرا میانه بر اساس این آماره مرتب یک اندازه مکانی است.

برآورد LMs:

دراین روش میانه توان دوم خطا حداقل می شود.

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

براورد LTS:

$$\hat{\beta} = min \sum_{i=1}^{n} r_i\,(\beta)^2$$

که در آن:

$$r_i(\beta) = y_i - \acute{x}_i\beta$$

$$r_i(\beta)^2 \leq \cdots \leq r_q(\beta)^2$$

$$q = \left({}^{n}/_{2} + 1\right)$$

برآورد GM :

از حل معادله زیر بدست می آید:

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

که در آن:

$$w_i = \sqrt{1 - h_i} \qquad , \qquad \psi(e) = \acute{\rho}(e)$$

نفوذ مشاهده $h_i \rightarrow i$

برآورد MM:

براورد MM از حل معادله زیر به دست می آید.

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

که در آن:

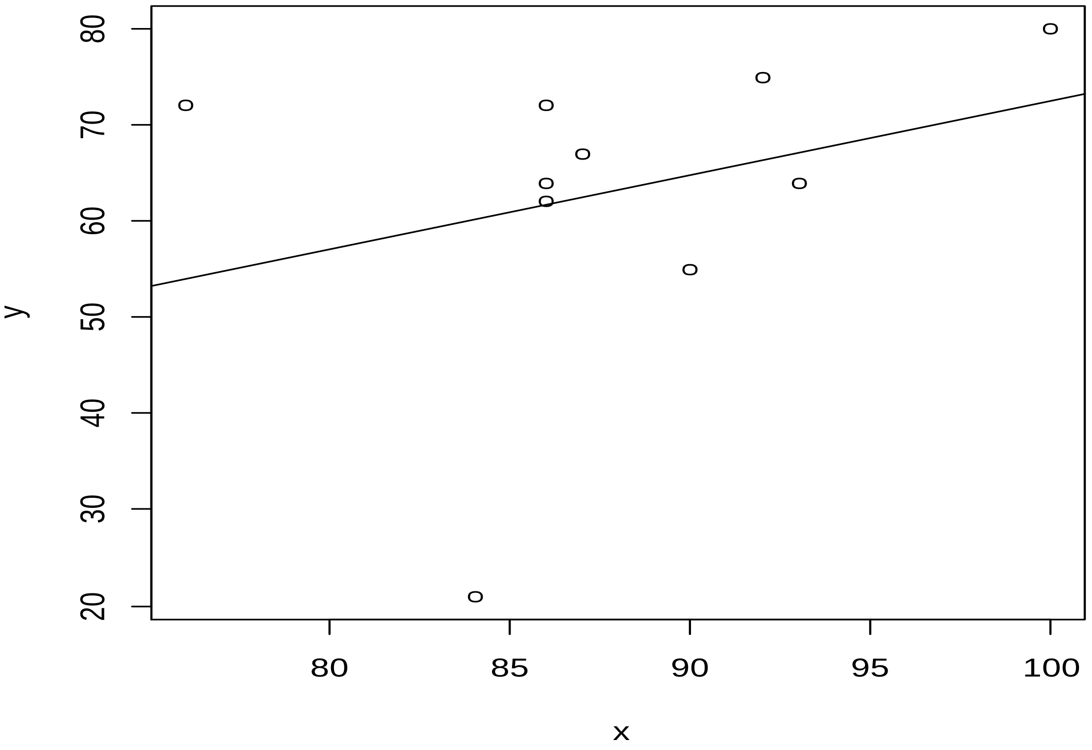
$$r_i(\beta) = y_i - \acute{x}_i\beta$$

در این براورد اگر :

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

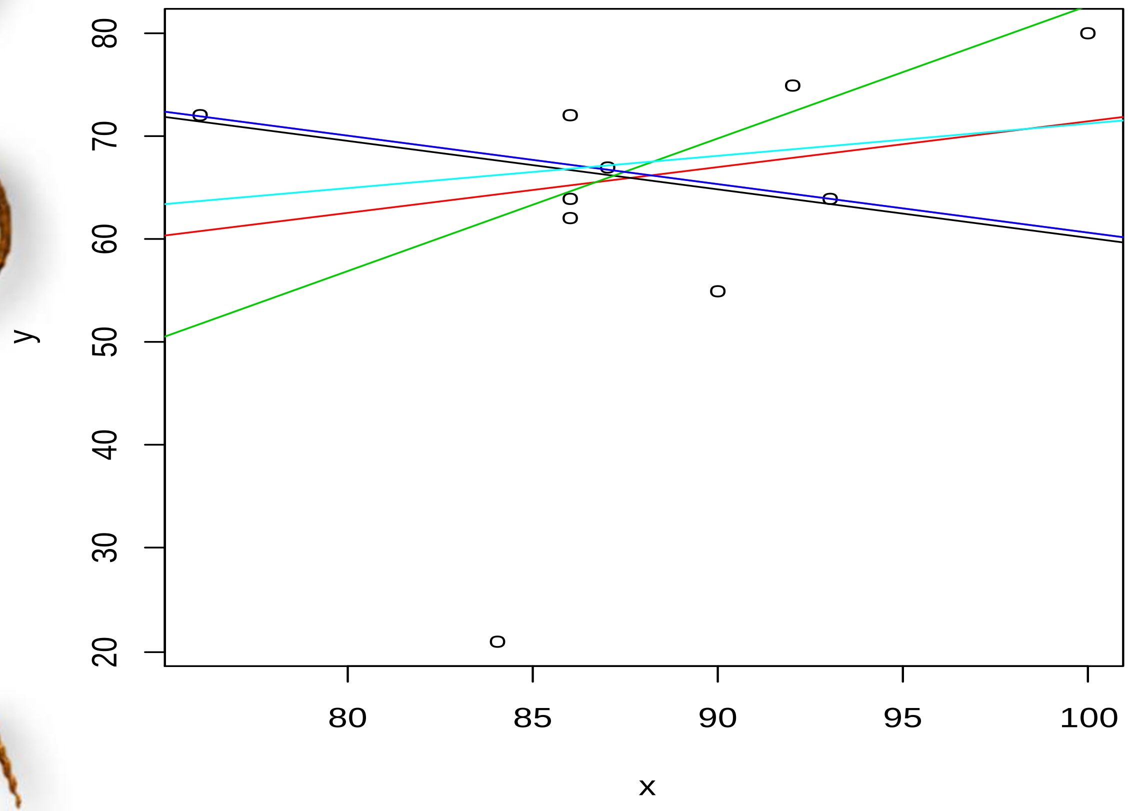
یک مثال درR:

```
> y=incom<-c(62,72,75,55,64,21,64,80,67,72)
> x=education<-c(86,76,92,90,86,84,93,100,87,86)
> model<-lm(y~x)
> plot(y~x)
```



80
70
60
50
40
30
20
y
80
85
90
95
100
x

```
> fitH <- rlm(y~ x, k2 = 1.345)
> fitLMS <- lqs(y ~ x, method = "lms")
> fitLTS <- lqs(y ~ x, method = "lts")
> fitS <- lqs(y ~ x, method = "S")
>fitMM <- rlm(y ~ x, method = "MM")
>plot(y~x)
> abline(fitH, col = 2)
> abline(fitLTS, col = 1)
> abline(fitLMS, col = 3)
> abline(fitS, col = 4)
> abline(fitMM, col = 5)
```



80
70
60
50
40
30
20
y
80
85
90
95
100
x

منابع :

۱-مقدمه ای بر تحلیل رگرسیون خطی: داگلاس مونت کمری

۲-شناسایی نقاط دور افتاده در داده های نرمال بر اساس مقادیر Z اصلاح شده :
محمد ولی احمدی و مجید سرمدی(دانشگاه فردوسی مشهد)

3-Robust Linear Regression: A Review and Comparison
Chun Yu1, Weixin Yao1, and Xue Bai1
1Department of Statistics,
Kansas State University, Manhattan, Kansas, USA 66506-0802

4-Robust Fitting of Parametric Models
Based on M-Estimation
Andreas Ruckstuhl ∗
IDP Institute of Data Analysis and Process Design
ZHAW Zurich University of Applied Sciences in Winterthur
Version 2016†

5-Robust Regression
John Fox & Sanford Weisberg
October 8, 2013

۷-پروژه outlier از خانم مینا خورشیدی و مرضیه روزبهانی سال ۹۵

خوش امد گل وز آن خوشتر نباشد

که در دستت به جز ساغر نباشد

زمان خوش دلی دریاب و دریاب

که دایم در صدف گوهر نباشد

با سپاس فراوان از توجه تان