Robusna regresija

Lončar, Marko

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Marko Lončar

ROBUSNA REGRESIJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Azra Tafro

Zagreb, rujan, 2021.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Hvala mojoj majci i obitelji na svemu!
Hvala svima koji su mi na bilo koji način bili podrška tijekom studiranja, posebice mojim prijateljima i rodbini.
Hvala mojoj mentorici na ukazanom povjerenju, savjetima i podršci.

Sadržaj

Uvod

Regresija je statistička metoda koja nastoji utvrditi odnos između zavisne i jedne ili više nezavisnih varijabli. Gotovo da nema grane u znanosti u kojoj se neki oblik regresije ne primijenjuje. Dvije osnovne vrste regresije su jednostavna linearna regresija i višestruka linearna regresija. Naravno, postoji i nelinearna regresija koja je uveliko kompliciranija i manje prisutna u znanstvenoj primjeni. Prije više od 200 godina predstavljena je prva metoda za procjenu parametara linearne regresije, a to je metoda najmanjih kvadrata. Zbog svoje duge tradicije i jednostavnosti izračuna, može se reći da je najpopularnija metoda procjene parametara. Općenito u matematici i statistici, razne tehnike i metode efikasne su samo ako su zadovoljeni određeni uvjeti. Tako se vremenom došlo do zaključka da je metoda najmanjih kvadrata jako osjetljiva na podatke koji znatno odudaraju od ostalih podataka, takozvane *outliere*. Razni su uzroci takvih podataka. Primjerice, krivo zabilježen podatak na papir, krivo prenesen podatak preko tipkovnice u računalu, greška u mjernom uređaju, itd. Činjenica je da se *outlieri* vrlo često pojavljuju u stvarnim podacima, količina podataka je prevelika za ručni pregled, a računala ih ne otkrivaju uvijek.

Prema tome, kaže se da je robusna regresija ona koja je otporna (robusna) na *outliere*. Robusnost (otpornost, neosjetljivost) nije samo riječ koja se veže uz regresiju i procjenitelje parametara regresije. Za razne procjene u statistici razvile su se metode koje nisu osjetljive na loše podatke i podatke koji ne zadovoljavaju željene pretpostavke. Začetnikom robusne statistike smatra se Peter J. Huber, švicarski matematičar koji je objavio razne radove i članke na temu robusnosti, a posebice je značajna knjiga [3]. Svijest o opasnosti *outliera* i metode najmanjih kvadrata postajala je sve veća kroz povijest te su predstavljene razne robusne metode za procjenu parametara linearne regresije, a u ovom radu bit će predstavljene dvije.

U prvom poglavlju predstavljamo ukratko opće modele jednostavne linearne regresije i višestruke linearne regresije. Potom prikazujemo detaljan izvod za procjenitelje višestruke linearne regresije metodom najmanjih kvadrata i pokazujemo njihovu nepristranost. Zatim definiramo reziduale i nepristrani procjenitelj varijance. U konačnici navodimo Gauss-Markovljev teorem koji kaže da je procjenitelj metodom najmanjih kvadrata najbolji linearni nepristrani procjenitelj ako vrijede Gauss-Markovljevi uvjeti.

Prije samih metoda, dobro je imati „alate" za mjerenje robusnosti. U drugom poglavlju

uvodimo pojmove robusnosti i pojam *outliera*, tri tipa svojstva ekvivarijantnosti za procjenitelje te dvije mjere robusnosti koje se zovu *breakdown* vrijednost i funkcija utjecaja. Pokazat ćemo koja je maksimalna b*reakdown* vrijednost koju procjenitelj može postići te izračunati *breakdown* vrijednost za metodu najmanjih kvadrata.

U trećem poglavlju obrađujemo dvije robusne metode, metodu najmanjeg medijana kvadrata odstupanja i metodu najmanjih odrezanih kvadrata, i pokazujemo da postižu maksimalnu *breakdown* vrijednost.

Četvrto poglavlje dat će nam uvid u konkretne primjere i usporedbu robusnih metoda sa klasičnom metodom najmanjih kvadrata. Radi se o stvarnim podacima, a analiza je rađena u programskom jeziku R.

Poglavlje 1

Linearna regresija

Pojam *regresija* prvi je uveo poznati britanski biolog Francis Galton 1908. kad se bavio proučavanjem nasljednih svojstava. Jedno od njegovih zapažanja bilo je da su djeca visokih roditelja viša od prosjeka, ali ne tako visoka kao njihovi roditelji. Prema [1], regresija je statistička metoda koja pronalazi odnos između zavisne varijable, koju obično označavamo sa Y, i jedne ili više nezavisnih varijabli. Ako zavisna varijabla ovisi o samo jednoj nezavisnoj varijabli onda to zovemo jednostavna regresija. Ako ovisi o više nezavisnih varijabli, radi se o višestrukoj regresiji. Kada je odnos zavisne i nezavisne varijable linearan, radi se o *linearnoj regresiji*. U sljedeća dva odjeljka bit će predstavljeni opći oblici *jednostavne linearne regresije* i *višestruke linearne regresije* po uzoru na [6] i [5].

1.1 Jednostavna linearna regresija

Jednostavna linearna regresija je linearni regresijski model koji opisuje vezu jedne zavisne i jedne nezavisne varijable. Otuda i naziv „jednostavna". Cilj je pronaći linearnu funkciju koja će predvidjeti vrijednost zavisne varijable kao funkciju nezavisne. Model se tipično navodi u sljedećem obliku:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

gdje je Y zavisna varijabla, X nezavisna varijabla, β_0 slobodni član, β_1 koeficijent smjera, a ε slučajna greška. Za greške se pretpostavlja da im je očekivanje nula, varijanca σ^2 i da su nekorelirane.

Nezavisna varijabla X zove se i varijabla poticaja, prediktorska varijabla, kontrolirana varijabla i ona je neslučajna. Zbog toga je odsada pa nadalje označavamo sa x. Varijabla x se najčešće zadaje, a Y se opaža (mjeri).

Zavisna varijabla Y naziva se i varijabla odgovora, varijabla odziva, kriterijska varijabla i ona je slučajna varijabla. Za njeno očekivanje i varijancu vrijedi:

$$\mathrm{E}[Y \mid x] = \beta_0 + \beta_1 x$$
$$\mathrm{Var}[Y \mid x] = \mathrm{Var}(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2.$$

Primijetimo da je očekivanje od Y linearna funkcija od x, a varijanca od Y ne ovisi o x. β_0 i β_1 zovu se regresijski koeficijenti i oni su nepoznati.

1.2 Višestruka linearna regresija

U mnogim znanstvenim istraživanjima često je potrebno utvrditi odnos između zavisne varijable Y i više od jedne nezavisne varijable, odnosno varijabli $(x_1, \ldots, x_k)$. Opći oblik modela višestruke linearne regresije dan je sa:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

gdje je Y zavisna varijabla, $x_1, \ldots, x_k$ nezavisne varijable, $\beta_0, \beta_1, \ldots, \beta_k$ parametri modela, a ε slučajna greška.

Ako promatramo n opažanja odnosno ako imamo slučajan uzorak $(x_{i1}, x_{i2}, \ldots, x_{ik}, Y_i)$ iz modela višestruke linearne regresije, tada n jednadžbi

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

možemo zapisati u matričnom obliku

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{1.1}$$

gdje su

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

uz pretpostavku da je $k < n$.

Vektor stupac $\mathbf{y}$ sadrži vrijednosti n zavisnih varijabli. Matrica $\mathbf{X}$ je $n \times (k+1)$ matrica nezavisnih varijabli iz uzorka, a zovemo je i matrica dizajna. Vektor β je $(k+1)$-dimenzionalni vektor stupac parametara modela, a ε je vektor stupac slučajnih grešaka koje se nazivaju i „šumovi". Dalje u radu, zbog jednostavnosti, koristit će se umjesto oznaka $\boldsymbol{y}$ i $\boldsymbol{X}$, oznake y i X.

Pretpostavka modela je da greške zadovoljavaju *Gauss-Markovljeve uvjete*:

$$\mathrm{E}[\varepsilon_i] = 0 \qquad \forall i = 1, \ldots, n \tag{1.2}$$

$$\mathrm{Var}[\varepsilon_i] = \sigma^2 \qquad \forall i = 1, \ldots, n \tag{1.3}$$

$$\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = \mathrm{E}[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i \neq j. \tag{1.4}$$

Spomenuto je već da su parametri modela nepoznati. Najčešća metoda korištena za njihovu procjenu je metoda najmanjih kvadrata. Međutim ta metoda je efikasna samo u slučaju kad su u potpunosti zadovoljeni uvjeti (1.2), (1.3) i (1.4), a to u stvarnosti nije baš tako uobičajeno. Više o toj metodi u sljedećem odjeljku.

1.3 Metoda najmanjih kvadrata

Tijekom razdoblja Velikih geografskih otkrića, matematičari su nastojali dati rješenja za probleme i izazove koji su se javljali u astronomiji i geodeziji. Dosljedan opis ponašanja nebeskih tijela bio je ključan za plovidbu i navigaciju brodova na otvorenom moru. Prvi jasan i sažet prikaz metode najmanjih kvadrata objavio je francuski matematičar Adrien-Marie Legendre 1805. Tehnika koju je opisao vrlo brzo je usvojena kao standardni alat u astronomiji i geodeziji. Teoretski dio ovog odjeljka temelji se na sadržajima izvora [6]. Ponekad u radu koristit će se kratica LS za ovu metodu, a dolazi od engleskog izraza *least squares*.

Neka je $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$ uzorak iz modela višestruke linearne regresije. Regresijski model možemo zapisati na sljedeći način:

$$y_i = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Metoda se temelji na minimizaciji sume kvadrata grešaka ε_i. Definiramo funkciju najmanjih kvadrata sa:

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{k} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{i=1}^{k} \beta_j x_{ij} \right)^2. \tag{1.5}$$

Funkciju S minimiziramo obzirom na parametre $\beta_0, \beta_1, \ldots, \beta_k$. Dobiveni procjenitelji metodom najmanjih kvadrata, označimo ih sa $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ moraju zadovoljavati:

$$\frac{\partial S}{\partial \beta_0}|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{i=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

$$\frac{\partial S}{\partial \beta_j}|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{i=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \quad j = 1, \ldots, k.$$

Time dobivamo sustav jednadžbi:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i2k} &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} &= \sum_{i=1}^{n} x_{i1}y_i \\ &\vdots \\ \hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik}y_i. \end{aligned} \tag{1.6}$$

Primijetimo da ima $k+1$ jednadžbi, što je jednako broju nepoznatih regresijskih parametara.

Matrični zapis modela regresije (1.1) svodi funkciju (1.5) na

$$S(\beta) = \sum_{i=1}^{k} \varepsilon_i^2 = (y - X\beta)^T (y - X\beta). \tag{1.7}$$

Primijetimo, $\beta^T X^T y$ je 1×1 matrica odnosno skalar i $(\beta^T X^T y)^T = y^T X\beta$ pa vrijedi:

$$S(\beta) = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta = y^T y - 2\beta^T X^T y + \beta^T X^T X\beta.$$

Funkciju S treba minimizirati, što znači da vektor stupac $\hat{\beta}$ procjenitelja metodom najmanjih kvadrata mora zadovoljavati:

$$\begin{aligned} \frac{\partial S}{\partial \beta}|_{\hat{\beta}} &= -2X^T y + 2X^T X\hat{\beta} = 0 \\ X^T X\hat{\beta} &= X^T y. \end{aligned} \tag{1.8}$$

Uočimo da je to sustav jednadžbi analogan onom raspisanom u (1.6).

Rješenje ćemo dobiti množeći obe strane jednadžbe (1.8) sa $(X^T X)^{-1}$ uz pretpostavku da inverz matrice $X^T X$ postoji[1].

Konačno, procjenitelj metodom najmanjih kvadrata za regresijske koeficijente $\beta_0, \beta_1, \ldots, \beta_k$ je

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Raspišimo detaljno matrični zapis (1.8):

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{bmatrix}.$$

[1] Inverz matrice postoji ako su varijable X_k linearno nezavisne. To znači da nijedan stupac matrice X nije linearna kombinacija ostalih stupaca.

Sada vidimo da je $X^T X$ $(k+1)\times(k+1)$ matrica na čijoj dijagonali se nalaze sume kvadrata elemenata stupaca matrice X. Na ostalim elementima nalaze se sume skalarnih produkata elemenata stupaca matrice X.

Vektor procijenjenih vrijednosti $\hat{y}_i$ u odnosu na opažene vrijednosti y_i glasi

$$\hat{\mathbf{y}} = X\hat{\beta} = X(X^T X)^{-1}X^T \mathbf{y}.$$

Važno svojstvo dobivenog procjenitelja je nepristranost, a dokazujemo ga sljedećim teoremom.

Teorem 1.3.1. *Procjenitelj $\hat{\beta} = (X^T X)^{-1}X^T y$ je nepristrani procjenitelj od β. Nadalje,*

$$\mathrm{Var}(\hat{\beta}) = (X^T X)^{-1}\sigma^2. \tag{1.9}$$

Dokaz. Primijetimo da

$$\mathrm{E}(\hat{\beta}) = \mathrm{E}((X^T X)^{-1}X^T y) = (X^T X)^{-1}X^T \mathrm{E}(y) = (X^T X)^{-1}X^T X\beta = \beta.$$

što dokazuje nepristranost od β.

Preostaje dokazati 1.9, a to činimo računajući direktno:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) = \mathrm{Var}((X^T X)^{-1}X^T y) &= (X^T X)^{-1}X^T \mathrm{Var}(\beta)((X^T X)^{-1}X^T)^T \\ &= (X^T X)^{-1}X^T X(X^T X)^{-1}\sigma^2 \\ &= (X^T X)^{-1}\sigma^2.\end{aligned}$$

□

1.4 Procjena σ^2

Za brojne rezultate vezane uz metodu najmanjih kvadrata i njenog procjenitelja potrebna nam je varijanca σ^2. Ta vrijednost je često nepoznata, stoga radimo njenu procjenu (teorija vezana uz procjenu preuzeta iz [12]). Za procjenu su nam potrebni *reziduali*.

Definirajmo reziduale kao razliku $e_i = y_i - \hat{y}_i$ tj. vektorski

$$e = y - \hat{y}.$$

Sada pogledajmo sumu kvadrata reziduala, koja će nam pomoći u procjeni:

$$\sum_{i=1}^{n} e_i^2 = e^T e = (y - X\hat{\beta})^T(y - X\hat{\beta}) = y^T[I - X(X^T X)^{-1}X^T]y = y^T P y.$$

Definicija 1.4.1. *Matrica $A \in M_{mn}$ je idempotentna ako vrijedi $A^2 = A$.*

Lako je pokazati da je $P = [I - X(X^TX)^{-1}X^T]$ idempotentna matrica:

$$P^2 = [I - X(X^TX)^{-1}X^T][I - X(X^TX)^{-1}X^T] = [I - X(X^TX)^{-1}X^T] = P.$$

Sljedeći teoremi govore o svojstvima idempotentnih matrica, a potrebni su nam za daljne zaključke. Dokazi su izostavljeni, a mogu se pronaći u izvoru [12].

Teorem 1.4.2. *Neka je A idempotentna matrica ranga k. Tada su svojstvene vrijednosti od A 0 ili 1.*

Teorem 1.4.3. *Ako je A idempotentna matrica onda je* tr*(A)*=rang*(A)=k.*

Ukupno je k svojstvenih vrijednosti koji iznose 1, a $n - k$ svojstvenih vrijednosti koji iznose 0. Matrica $X(X^TX)^{-1}X^T$ također je idempotentna, stoga vrijedi:

$$\begin{aligned}\text{rang}(X(X^TX)^{-1}X^T) &= \text{tr}(X(X^TX)^{-1}X^T)\\ &= \text{tr}(X^TX(X^TX)^{-1})\\ &= \text{tr}(I_k) = k.\end{aligned}$$

Zbog $\text{tr}(A - B) = A - B$ slijedi:

$$\begin{aligned}\text{rang}(I - X(X^TX)^{-1}X^T) &= \text{tr}(I - X(X^TX)^{-1}X^T)\\ &= \text{tr}(I_p) - \text{tr}(X^TX(X^TX)^{-1})\\ &= n - k.\end{aligned}$$

Koristeći rezultat matematičkog očekivanja kvadratne forme dobijamo:

$$\begin{aligned}\text{E}[e^Te] &= \text{E}[y^T(I - X(X^TX)^{-1}X^T)y]\\ &= (X\beta)^T(I - X(X^TX)^{-1}X^T)(X\beta) + \sigma^2(n-k)\\ &= (X\beta)^T(X\beta - X(X^TX)^{-1}X^TX\beta) + \sigma^2(n-k)\\ &= \sigma^2(n-k).\end{aligned}$$

Sljedeći teorem daje procjenu varijance:

Teorem 1.4.4. *Nepristrani procjenitelj varijance u višestrukoj linearnoj regresiji je dan sa:*

$$\sigma^2 = \frac{e^Te}{n-k} = \frac{y^T(I - X(X^TX)^{-1}X^T)y}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{1.10}$$

1.5 Gauss-Markovljev teorem

Nepristranost procjenitelja smo već pokazali u Teoremu 1.3.1. Sada ćemo pokazati da procjenitelj dobiven metodom najmanjih kvadrata $\hat{\beta} = (X^T X)^{-1} X^T y$ ima najmanju varijancu među svim linearnim procjeniteljima. Takav procjenitelj zove se najbolji nepristrani linearni procjenitelj. Teorem koji govori o tome zove se Gauss-Markovljev teorem [6] i jedan je od najznačajnijih rezultata u statistici kada je u pitanju regresijska analiza.

Prisjetimo se Gauss-Markovljevih uvjeta:

$$\mathrm{E}[\varepsilon_i] = 0 \qquad \forall i = 1, \ldots, n$$

$$\mathrm{Var}[\varepsilon_i] = \sigma^2 \qquad \forall i = 1, \ldots, n$$

$$\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = \mathrm{E}[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i \neq j.$$

Teorem 1.5.1. *(Gauss-Markov) Neka je $\hat{\beta}$ procjenitelj dobiven metodom najmanjih kvadrata za parametre linearnog regresijskom modela. Ako vrijede Gauss-Markovljevi uvjeti, tada je $\hat{\beta}$ najbolji linearni nepristrani procjenitelj.*

Dokaz. Neka je $\tilde{\beta} = Cy$ neki drugi nepristrani procjenitelj, $C = (X^T X)^{-1} X^T + D$, a D netrivijalna $k \times n$ matrica.

$$\begin{aligned}
\mathrm{E}(\tilde{\beta}) &= \mathrm{E}(Cy) \\
&= \mathrm{E}\Big[((X^T X)^{-1} X^T + D)(X\beta + \varepsilon)\Big] \\
&= \mathrm{E}\big[((X^T X)^{-1} X^T + D)X\beta + ((X^T X)^{-1} X^T + D)\varepsilon\big] \\
&= ((X^T X)^{-1} X^T + D)X\beta + ((X^T X)^{-1} X^T + D)\underbrace{\mathrm{E}[\varepsilon]}_{=0} \\
&= (X^T X)^{-1} X^T X\beta + DX\beta \\
&= I_k\beta + DX\beta \\
&= (I_k + DX)\beta
\end{aligned}$$

Budući da je $\tilde{\beta}$ nepristran, $DX = 0$.

Pokažimo sada potrebnu tvrdnju za varijancu:

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \mathrm{Var}(Cy) \\
&= C\mathrm{Var}(y)C^T \\
&= \sigma^2 CC^T \\
&= \sigma^2((X^TX)^{-1}X^T + D)((X^TX)^{-1}X^T + D)^T \\
&= \sigma^2((X^TX)^{-1}X^T + D)(X(X^TX)^{-1} + D^T) \\
&= \sigma^2((X^TX)^{-1}X^TX(X^TX)^{-1} + (X^TX)^{-1}X^TD^T + \underbrace{DX}_{=0}(X^TX)^{-1} + DD^T) \\
&= \sigma^2(I_k(X^TX)^{-1} + \underbrace{DX}_{=0}(X^TX)^{-1} + DD^T) \\
&= \sigma^2(X^TX)^{-1} + \sigma^2 DD^T \\
&= \mathrm{Var}(\hat{\beta}) + \sigma^2 DD^T.
\end{aligned}$$

Jer je DD^T pozitivno semidefinitna matrica, slijedi da je $\mathrm{Var}(\hat{\beta}) \leq \mathrm{Var}(\tilde{\beta})$ za sve druge linearne nepristrane procjenitelje $\tilde{\beta}$. □

Poglavlje 2

Mjere robusnosti

2.1 Uvod u robusnost

Riječ *robustan* dolazi od latinske riječi *robustus* što znači tvrd, čvrst. Robusnost u statistici, prema Huberu [4], je *otpornost na manje promjene u pretpostavkama*. Problem leži u tome što su mnogi statistički modeli temeljeni na „idealnim" situacijama koje se rijetko javljaju pri radu s podacima iz stvarnog svijeta. Robustan model, metoda ili test bio bi onaj koji daje točne rezultate čak i kada uvjeti nisu u potpunosti ispunjeni. Drugim riječima, izraz robustan ili robusnost u statistici odnosi se na otpornost i snagu statističkog modela, metode, postupka testa itd. Huber kaže da *mala neskladna manjina nikada ne smije nadjačati činjenice koje nam donosi većina opažanja.*

Upoznali smo se s općim oblikom jednostavne i višestruke linearne regresije, te smo predstavili metodu najmanjih kvadrata za procjenu njenih parametara. Općenito, za metodu odnosno procjenu reći ćemo da je robusna ako je otporna na *outliere*. Premda ne postoji precizna definicija *outliera*, možemo reći da su to opažanja koja se značajno razlikuju od ostalih. *Outlieri* mogu uzrokovati značajne probleme u statističkoj analizi, a razni su uzroci njihove pojave. Postoje alati za otkrivanje *outliera*, međutim metode detekcije *outliera* neće biti dio ovog rada.

Robusnih procjenitelja ima puno i nužno je imati određene alate za usporedbu kako bismo znali koji je bolji i pod kojim uvjetima. U ovom poglavlju obradit ćemo dvije mjere robusnosti. Prvi je pojam *breakdown* vrijednosti, koja mjeri koliko se procjena može odupriječiti velikoj promjeni dijela podataka. *Breakdown* vrijednost pripada skupini takozvanih kvantitativnih mjera robusnosti i primjerena je regresijskoj analizi. Druga je *krivulja utjecaja*, koja daje informacije o tome kako samo jedan izdvojeni faktor može utjecati na procjenu. Krivulja utjecaja pripada skupini takozvane infinitezimalne robusnosti.

2.2 Ekvivarijantnost

Prije mjera robusnosti, navest ćemo jedno bitno svojstvo za regresijske procjenitelje, a to je ekvivarijantnost. Točnije, navest ćemo tri tipa ekvivarijantnosti, kao u [9] i [11].

Prisjetimo se općeg oblika modela višestruke linearne regresije:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

gdje je Y zavisna varijabla, $x_1, \ldots, x_k$ nezavisne varijable, $\beta_0, \beta_1, \ldots, \beta_k$ parametri modela, a ε slučajna greška. Promotrimo jedno od n opažanja i označimo taj slučajan uzorak sa $(x_i, y_i) := (x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$.

Procjenitelj T je *regresijski ekvivarijantan* ako

$$T((x_i, y_i + x_i\mathbf{v}) : i = 1, \ldots, n) = T((x_i, y_i) : i = 1, \ldots, n) + \mathbf{v}, \tag{2.1}$$

gdje je $\mathbf{v}$ bilo koji vektor stupac. Ovo svojstvo je ključno svojstvo za regresijske procjenitelje, a često se uzima „a priori".

Procjenitelj T je *ekvivarijantan na skaliranje* ako

$$T((x_i, cy_i) : i = 1, \ldots, n) = cT((x_i, y_i) : i = 1, \ldots, n), \tag{2.2}$$

gdje je c proizvoljna konstanta. Ovo svojstvo kaže da je procjena neovisna o izboru mjerne jedinice za varijablu y. Primjerice, ako modeliramo visinu učenika, promjena mjerne jedinice iz centimetara u metre, ne smije utjecati na procjenu.

Procjenitelj T je *afino ekvivarijantan* ako

$$T((x_i\mathbf{A}, y_i) : i = 1, \ldots, n) = \mathbf{A}^{-1}T((x_i, y_i) : i = 1, \ldots, n), \tag{2.3}$$

za svaku regularnu matricu $\mathbf{A}$. Zapravo, linearna transformacija x_i uzrokuje i transformaciju procjenitelja T jer je $\hat{y}_i = x_i T = (x_i\mathbf{A})(\mathbf{A}^{-1}T)$. Zato ovo svojstvo omogućava korištenje drugih koordinatnih sustava bez utjecaja na procijenjene vrijednosti $\hat{y}_i$.

2.3 *Breakdown* vrijednost

Postoje procjenitelji na koje čak i jedan outlier može jako utjecati, dok s druge strane postoje procjenitelji koji su otporni i na veći broj outliera među podacima. Kako bismo mogli „mjeriti" robusnost procjenitelja uvodimo prvo pojam *breakdown* vrijednosti. *Breakdown* vrijednost je predstavio F. R. Hampel u svojoj doktorskoj disertaciji, a svi teorijski rezultati ovog potpoglavlja temelje se na sadržaju iz [10] i [9].

Pretpostavimo da odaberemo dio podataka. Možemo li izazvati proizvoljno veliku promjenu procjenitelja mijenjanjem odabranih vrijednosti opažanja u odgovarajućoj mjeri?

Neka je T procjenitelj, a Z neki uzorak n opažanja

$$Z = (x_{11}, \ldots, x_{1k}, y_1), \ldots, (x_{n1}, \ldots, x_{nk}, y_n).$$

To znači da djelovanjem T na uzorak Z dobijemo vektor procijenjenih koeficijenata $\hat{\beta}$

$$T(Z) = \hat{\beta}.$$

Zamijenimo bilo kojih m točaka originalnog uzorka s proizvoljnim vrijednostima i promotrimo sve moguće takve Z' uzorke. Definirajmo s $b(m; T, Z)$ maksimalno odstupanje uzrokovano takvom zamjenom

$$b(m; T, Z) = \sup_{Z'} ||T(Z') - T(Z)||.$$

Ako je maksimalno odstupanje beskonačno, to znači da m outliera može imati proizvoljno velik utjecaj na T, pa kažemo da procjenitelj „puca“.

Definicija 2.3.1. *Breakdown vrijednost procjenitelja T na uzorku Z je najmanji dio podataka koji se može promijeniti proizvoljno velikim vrijednostima, a ipak uzrokovati proizvoljno veliku promjenu procjene. Oznaka:*

$$\varepsilon_n^*(T, Z) = min\Big\{\frac{m}{n} : b(m; T, Z) = \infty\Big\}.$$

Uočimo da je ovo *breakdown* vrijednost na konačnom uzorku. Često se promatra asimptotska *breakdown* vrijednost koja se može dobiti kao limes *breakdown* vrijednosti na konačnom uzorku kada veličina uzorka teži u beskonačno.

Najveća *breakdown* vrijednost koju neki procjenitelj može postići je 50%. Naime, kada bi *breakdown* vrijednost bila veća od 50% tada ne bi mogli razlučiti koji podaci su dobri a koji su loši. Tvrdnju formalno pokazujemo sljedećim teoremom i dokazom:

Teorem 2.3.2. *Ako je T procjenitelj koji je regresijski ekvivarijantan, tada je*

$$\varepsilon_n^*(T, Z) \leq \frac{\lfloor (n-k)/2 \rfloor + 1}{n},$$

za sve uzorke Z.

Dokaz. Pretpostavimo suprotno. Neka je *breakdown* vrijednost strogo veća od $(\lfloor (n-k)/2 \rfloor + 1)/n$ i neka je Z neki uzorak. To znači da postoji konstanta b takva da se $T(Z')$ nalazi u kugli[1] $K(T(Z), b)$ oko $T(Z)$ radijusa b za svaki uzorak Z' koji sadrži barem $n - \lfloor (n-k)/2 \rfloor - 1$ točaka iz Z. Definirajmo

$$q := n - \Big\lfloor \frac{n-k}{2} \Big\rfloor - 1.$$

[1] $K(T(Z), b) = \{\beta : ||T(Z) - \beta|| \leq b\}$

Raspisivanjem dobijemo da vrijedi

$$q = \Big\lfloor \frac{n+k+1}{2} \Big\rfloor - 1.$$

Definirajmo sada k-dimenzionalni vektor stupac $\mathbf{v} \neq 0$ tako da vrijedi

$$x_1\mathbf{v} = 0\,, \ldots,\; x_{p-1}\mathbf{v} = 0.$$

Ako je $n + k + 1$ paran tada je $2q - (k-1) = n$, inače $2q - (k-1) = n - 1$. Možemo reći $2q - (k-1) \leq n$. Prema tome, prvih $2q - (k-1)$ točaka u Z možemo zamijeniti s

$$(x_1, y_1), \ldots, (x_{k-1}, y_1),\; (x_k, y_1), \ldots, (x_q, y_1),\; (x_k, y_k + x_k\tau\mathbf{v}), \ldots, (x_q, y_q + x_q\tau\mathbf{v}),$$

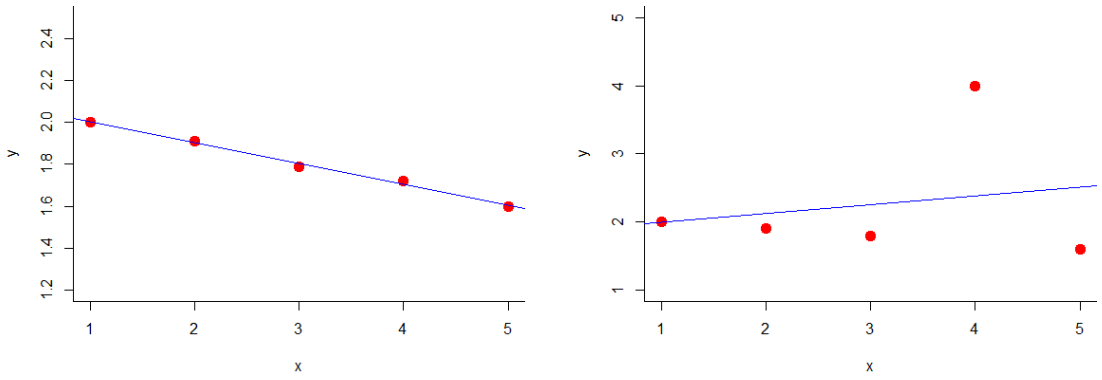
za proizvoljni $\tau > 0$. Za ovako dobiveni uzorak Z', $T(Z') \in K(T(Z), b)$ jer Z' sadrži q točaka iz Z. S druge strane, zbog svojstva regresijske ekvivarijantnosti, $T(Z')$ možemo zapisati i kao $T(Z'') + \tau\mathbf{v}$, pri čemu je $T(Z'') \in K(T(Z), b)$. Dakle, micanjem središta kugle $K(T(Z), b)$, $T(Z')$ je opet u novoj kugli, tj. $T(Z') \in K(T(Z) + \tau\mathbf{v}, b)$.

Time smo došli do kontradikcije jer je presjek između $K(T(Z), b)$ i $K(T(Z) + \tau\mathbf{v}, b)$ prazan za dovoljno veliki τ. □

2.4 *Breakdown* vrijednost metode najmanjih kvadrata

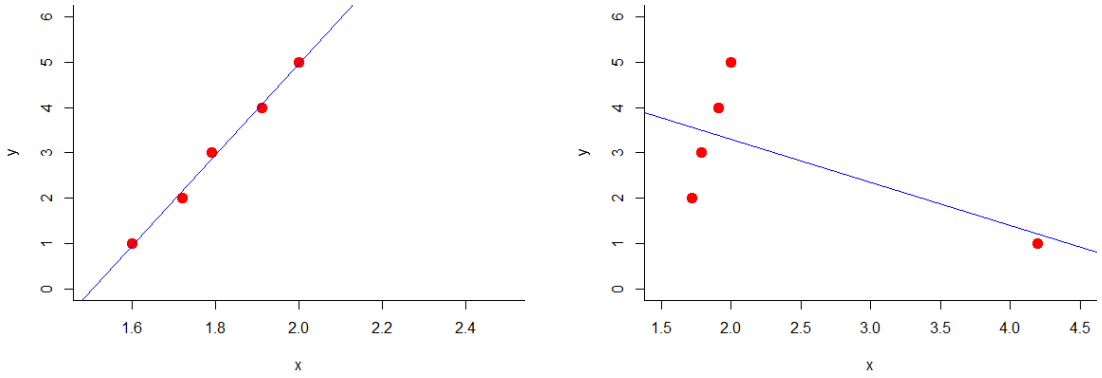
Za bolju intuiciju konstruirajmo model jednostavne linearne regresije. U slučaju jednostavne regresije podatke je moguće prikazati grafički kao parove (x_i, y_i) odnosno točke u ravnini. Uzmimo 5 točaka $(x_1, y_1), \ldots, (x_5, y_5)$ koje leže gotovo na istom pravcu (2.1a). Tada pravac dobiven metodom najmanjih kvadrata vrlo dobro aproksimira podatke što možemo vidjeti na slici (2.1a). Pretpostavimo sada je da je samo jedna vrijednost, neka to bude y_4, krivo izmjerena zbog bilo kakvog razloga. Tada točka (x_4, y_4) značajno odudara od „idealnog" pravca što se očito vidi na slici (2.1b). Novi pravac više ne odgovara podacima jer se četvrta točka „podigla" od originalne pozicije. Takvu točku zovemo *outlier u y-smjeru* i vidimo da već jedna takva točka, tj. outlier, ima veliki utjecaj na procjenu. Ovakva situacija, kada se outlieri pojavljuju u „y-smjeru", česta je u literaturi i praksi jer se, kao što je rečeno, y_i smatraju kao opažanja, a x_i kao fiksne vrijednosti unaprijed zadane.

Međutim, greške su moguće i među nezavisnim varijablama, pogotovo u slučaju modela višestruke regresije kada ima puno nezavisnih varijabli. Da bismo ilustrirali problem, uzmimo opet 5 točaka u ravnini koje leže skoro pa na istom pravcu (slika 2.2a). Pretpostavimo da se dogodila greška u vrijednosti x_1. Sada točka (x_1, y_1) znatno odudara od originalnih podataka i pravca. Takva točka zove se *outlier u x-smjeru* i uzrokuje potpuno zakretanje pravca (slika 2.2b).



(a) pravac bez *outliera* (b) pravac s *outlierom* u y-smjeru

Slika 2.1: Utjecaj *outliera* u y-smjeru na LS pravac



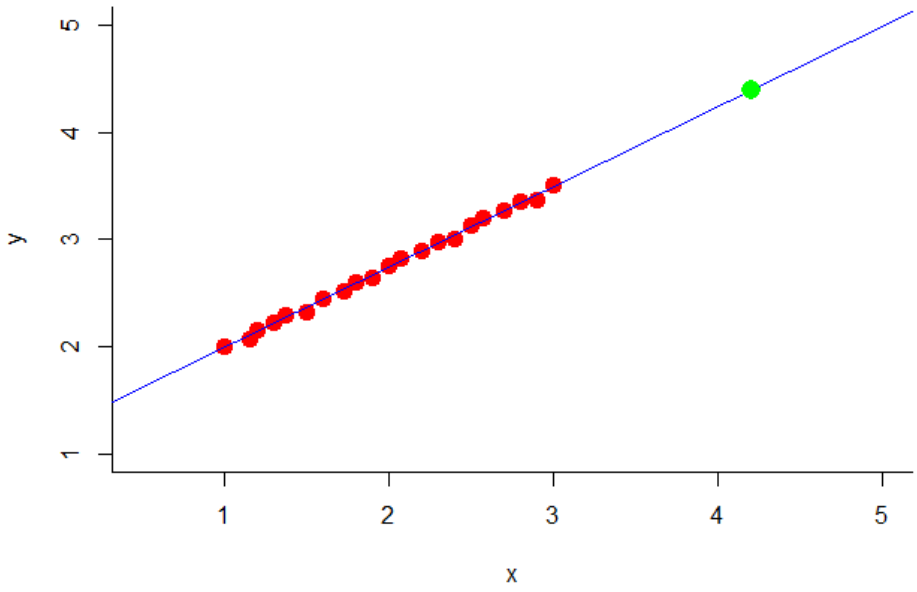
(a) pravac bez *outliera* (b) pravac s *outlierom* u x-smjeru

Slika 2.2: Utjecaj *outliera* u x-smjeru na LS pravac

Iz oba prikazana primjera, očito je da već jedna promjena(greška) podatka utječe značajno na procjenu. Prema definiciji *breakdown* vrijednosti, ona iznosi $\frac{1}{n}$ za metodu najmanjih kvadrata, gdje je n broj podataka. Primijetimo kada n teži u beskonačno, $\frac{1}{n}$ teži u nula. Dakle *breakdown* vrijednost procjenitelja metodom najmanjih kvadrata je 0% i takav procjenitelj nije robustan.

Točka (x_1, y_1) sa slike 2.2 zove se i *točka utjecaja*. Općenito, podatak (x_k, y_k) je točka utjecaja (eng. *leverage point*) ako se x_k nalazi daleko od ostalih x_i. Međutim, nije svaka

točka utjecaja *outlier*. Ako ta točka leži blizu pravca koji je dobiven obzirom na ostale podatke, tada se ta točka zove „dobra" točka utjecaja, a primjer jedne takve možemo vidjeti na slici 2.3. Primjeri u ovom potpoglavlju rađeni su po uzoru na primjere iz [9].



Slika 2.3: Pravac dobiven metodom najmanjih kvadrata i točka utjecaja

2.5 Funkcija utjecaja

Krivulju utjecaja predstavio je prvo Hampel u svom radu [2]. U početku se to zvala krivulja kako bi se naglasili njeni geometrijski aspekti, dok se kasnije češće zove funkcija utjecaja obzirom na generalizaciju na višedimenzionalne prostore.

Funkcija utjecaja opisuje približan utjecaj dodatnog podatka (opažanja) na procjenitelj T, obzirom na uzorak iz distribucije F. Može se ugrubo reći da je to prva derivacija od T obzirom na definiciju koja slijedi.

Pretpostavimo da je F funkcija distribucije. Možemo promatrati malu promjenu F pri vrijednosti $z_0 = (x_0^T, y_0)$ uzimajući u obzir miješanu distribuciju $F_t = (1-t)F + t\delta_{z_0}$ gdje je δ_{z_0} funkcija distribucije od z_0, a $t > 0$ proizvoljno mala vrijednost blizu nula.

Funkcija utjecaja procjenitelja T uz distribuciju F definirana je sa

$$IF(T, F, z_0) = \lim_{t \to 0} \frac{T(F_t) - T(F)}{t}, \tag{2.4}$$

za sve točke z_0 za koje ovaj limes postoji.

Za procjenitelj T kažemo da je robustan ili da ima infinitezimalnu robusnost ako je $IF(T, F, z_0)$ omeđena. Ovime smo dobili jednostavan i snažan alat za mjerenje robusnosti.

Sljedeća dva primjera preuzeta su iz [10].

Primjer 2.5.1. *Neka je X neprekidna slučajna varijabla. Promotrimo $E_F[X]$ kao procjenitelj T za srednju vrijednost u smislu gornjeg izraza 2.4.*

$$T(F) = E_F[X] = \int x dF(x),$$

pa je

$$\begin{aligned} T(F_t) &= \int x dF_t(x) \\ &= (1-t) \int x dF(x) + t \int x d\delta_x 0(x) \\ &= (1-t)T(F) + t x_0, \end{aligned}$$

iz čega slijedi

$$\frac{T(F_t) - T(F)}{t} = x_0 - T(F),$$

odnosno

$$IF(T, F, x_0) = x_0 - T(F).$$

Vidimo da je funkcija utjecaja neograničena u x_0, dakle ovaj procjenitelj nije robustan.

Primjer 2.5.2. *Može se pokazati da funkcija utjecaja za procjenitelja metodom najmanjih kvadrata glasi*

$$IF(T, F, z_0) = E[X^T X]^{-1} x_0 [y_0 - x_0^T T(F)].$$

Funkcija je neograničena i u x_0 i u y_0, što je još jedan pokazatelj da metoda najmanjih kvadrata nije robusna.

Dakle, pokazali smo na dva načina, preko *breakdown* vrijednosti i funkcije utjecaja, da metoda najmanjih kvadrata nije robusna. Ideja je, u nastavku ovog rada, predstaviti metode otpornije na outliere, odnosno robusne metode.

Poglavlje 3

Robusne metode

Pokazali smo u prethodnom poglavlju da metoda najmanjih kvadrata (LS) nije otporna na *outliere* pa ne možemo reći da je robusna. Kroz povijest su se stoga razvijale nove statističke tehnike i metode na koje *outlieri* ne mogu tako lako utjecati. Takve metode daju rezultate koji su relevantni i uz određenu prisutnost *outliera*. Postoje mnoge robusne metode za procjenitelje linearne regresije. U ovom radu bit će predstavljene dvije za koje se pokazalo da postižu *brekdown* vrijednost maksimalnu moguću.

Prije metoda, prisjetit ćemo se potrebnih izraza i jednadžbi otprije. Opći oblik modela višestruke linearne regresije glasi:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

gdje je Y zavisna varijabla, $x_1, \ldots, x_k$ nezavisne varijable, $\beta_0, \beta_1, \ldots, \beta_k$ parametri modela, a ε slučajna greška.

Ako promatramo n opažanja, tada slučajan uzorak označavamo sa $(x_i, y_i) := (x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, \ldots n$.

Reziduale smo definirali kao razliku

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

gdje je $\hat{y}_i$ procijenjena vrijednost od y_i. Dakle reziduali predstavljaju razliku stvarnih i procijenjenih podataka.

3.1 Metoda najmanjeg medijana kvadrata odstupanja

Ako pogledamo još jednom izraze (1.5) i (1.7), možemo reći da bi „dobar" naziv za metodu najmanjih kvadrata bio metoda najmanjih *suma* kvadrata. Međutim, pokazali smo već u prethodnom poglavlju da je *breakdown* vrijednost metode najmanjih kvadrata jednaka $1/n$, odnosno 0%. Postavlja se pitanje - možemo li pristupiti problemu na način da ne

gledamo *sume*. Razmislimo o aritmetičkoj sredini i medijanu kao mjerama srednje vrijednosti. Pokušajmo „izmjeriti" robusnost aritmetičke sredine preko *breakdown* vrijednosti. Uzmimo za primjer uzorak 1,2,2,2,3,4. Aritmetička sredina tog uzorka je 2.33, a medijan 2. Ako zamijenimo podatak 4 sa 100, aritmetička sredina iznimno naraste na vrijednost 18.66, dok se medijan ne mijenja. To nam govori da je medijan otporniji na ekstremne vrijednosti od aritmetičke sredine. Točnije, da već promjena samo jednog podatka značajno utječe na procjenu. Zaista, može se pokazati da je *breakdown* vrijednost aritmetičke sredine $1/n$,a medijana čak 50%, odnosno najviše[1] moguće.

To nas dovodi do procjenitelja metodom najmanjeg medijana kvadrata odstupanja, danog s

$$\hat{\beta} = \arg\min_{\beta} \operatorname*{med}_{i} e_i^2. \tag{3.1}$$

Ideju zamjene sume s medijanom koji je robusniji razvio je Rousseeuw u svom radu [7]. U nastavku ovog rada, koristit će se ponekad kratica LMS za metodu najmanjeg medijana kvadrata odstupanja, a dolazi od engleskog naziva za metodu *least median of squares*.

Sada ćemo promotriti svojstva ovog procjenitelja, njegovu egzistenciju i *breakdown* vrijednost. Zapišimo (x_i, y_i) kao $(x_{i1}, \ldots, x_{ip}, y_i)$. Ovih n podataka pripadaju linearnom prostoru vektor redaka dimenzije $k + 1$. Vektor $\beta = (\beta_1, \ldots, \beta_k)^T$ je vektor stupac nepoznatih parametara dimenzije k. Linearni model sada promatramo kao $y_i = x_i\beta + \varepsilon_i$, gdje ε_i pripada normalnoj distribuciji $\mathcal{N}(\mu, \sigma^2)$. U sljedećim rezultatima pretpostavit ćemo da su sva opažanja $x_i = 0$ zanemarena jer ne daju nikakvu informaciju o β. Nadalje, pretpostavljamo da u $(k + 1)$-dimenzionalnom prostoru (x_i, y_i) ne postoji vertikalna hiperravnina kroz ishodište koja sadrži više od $\lfloor n/2 \rfloor$ točaka. Misli se na hiperravninu koja je k-dimenzionalni potprostor koji sadrži $(0, \ldots, 0)$ i $(0, \ldots, 0, 1)$.

Teorem 3.1.1. *Minimizacijski problem (3.1) uvijek ima rješenje.*

Dokaz ovog teorema nije konstruktivan pa ga izostavljamo, a može se pronaći u [9]. Također, svi daljnji teorijski rezultati ovog potpoglavlja preuzeti su iz [9].

Sljedeća lema govori o ekvivarijantnosti LMS procjenitelja, odnosno dokazuje da zadovoljava svojstva (2.1), (2.2), (2.3).

Lema 3.1.2. *Procjenitelj metodom najmanjeg medijana kvadrata odstupanja (LMS procjenitelj) je regresijski ekvivarijantan, ekvivarijantan na skaliranje i afino ekvivarijantan.*

[1] U Teoremu 2.3.2 smo pokazali da je breakdown vrijednost najviše 50%.

Dokaz. Slijedi iz sljedećih jednakosti:

$$\begin{aligned}\underset{i}{\text{med}}\,((y_i + x_i\mathbf{v}) - x_i(\beta + \mathbf{v}))^2 &= \underset{i}{\text{med}}\,(y_i - x_i\beta)^2\\ \underset{i}{\text{med}}\,(cy_i - x_i(c\beta))^2 &= c^2\underset{i}{\text{med}}\,(y_i - x_i\beta)^2\\ \underset{i}{\text{med}}\,(y_i - (x_i\mathbf{A})(\mathbf{A^{-1}}\beta))^2 &= \underset{i}{\text{med}}\,(y_i - x_i\beta)^2.\end{aligned}$$

□

Za dokazivanje sljedećeg rezultata i nekih drugih koji će se pojaviti kasnije, treba nam pojam *generalne pozicije* uzorka. Kažemo da su opažanja u *generalnoj poziciji* kada svaka kombinacija k točaka podataka jedinstveno određuje β. Ako je $k = 1$, tj. ako imamo jednostavnu regresiju kroz ishodište, to znači da $x_i \neq 0\ \forall i$. Ako je $k = 2$, tada svaki par (x_{i1}, x_{i2}, y_i) i (x_{j1}, x_{j2}, y_j) određuje jedinstvenu ravninu kroz ishodište. Iz toga slijedi da (x_{i1}, x_{i2}, y_i), (x_{j1}, x_{j2}, y_j) i $(0, 0, 0)$ ne smiju ležati na istom pravcu.

Sljedeća ključna stvar koja nam je od interesa, je *breakdown* vrijednost.

Teorem 3.1.3. *Ako je $k > 1$ i opažanja su u generalnoj poziciji, tada je breakdown vrijednost procjenitelja metodom najmanjeg medijana kvadrata odstupanja jednaka*

$$\frac{\lfloor n/2 \rfloor - k + 2}{n}.$$

Dokaz. Prvo pokazujemo da je

$$\varepsilon_n^*(T, Z) \geq \frac{\lfloor n/2 \rfloor - k + 2}{n},$$

za svaki uzorak $Z = \{(x_i, y_i) : i = 1, \ldots, n\}$. Prema Teoremu 3.1.1, Z daje rješenje β. Sada treba pokazati da procjena ostaje ograničena kada $n - (\lfloor n/2 \rfloor - k + 2) + 1$ podataka ostane nepromijenjeno. Stoga konstruiramo uzorak $Z' = \{(x_i', y_i') : i = 1, \ldots, n\}$ na način da $n - \lfloor n/2 \rfloor + k - 1$ podataka iz Z ostavljamo nepromijenjenima i zovemo ih „dobri podaci", a sve ostale podatke zamijenimo s proizvoljnim vrijednostima.

Dovoljno je pokazati da je $\|\beta - \beta'\|^2$ ograničeno, gdje je β' LMS procjenitelj za uzorak Z'. Promatramo $(k + 1)$-dimenzionalni prostor E opažanja (x_i, y_i) i horizontalnu hiperravninu kroz ishodište koju označavamo s $\{y = 0\}$. Neka je

$$\rho = \frac{1}{2}\inf\Big\{\tau > 0 : \text{postoji } (k-1)\text{-dimenzionalni potprostor } V \text{ od } \{y = 0\} \text{ kroz ishodište}$$
$$\text{tako da } V^T \text{ sadrži najmanje k točaka } x_i\Big\},$$

[2] U cijelom radu oznaka $\|\cdot\|$ označava euklidsku normu, formula $\|(x_1, \ldots, x_n)\| = \sqrt{\sum_{i=1}^n |x_i|^2}$

gdje je V^T skup svih $\mathbf{x}$ takvih da je njihova udaljenost do V manja od τ. Znamo da je Z u generalnoj poziciji pa slijedi da je $\rho > 0$.

Definirajmo $M \coloneqq \max_i |e_i|$ gdje su $e_i = y_i - x_i\beta$ reziduali. Sada ćemo pokazati da je

$$||\beta - \beta'|| < 2\left(||\beta|| + \frac{M}{\rho}\right),$$

što će biti dovoljno jer je desna strana konstanta.

Neka su H i H' nevertikalne hiperravnine određene redom jednadžbama $y = \mathbf{x}\beta$ i $y' = \mathbf{x}'\beta'$. Bez smanjenja općenitosti možemo pretpostaviti da je $\beta \neq \beta'$, što znači da je i $H \neq H'$. Znamo iz linearne algebre da presjek hiperravnina $H \cap H'$ ima dimenziju $k-1$. Ako s $pr(H \cap H')$ označimo vertikalnu projekciju od $H \cap H'$ na $\{y = 0\}$, to znači da je $pr(H \cap H')$ $(k-1)$-dimenzionalan potprostor od $\{y = 0\}$. Tada nam definicija od ρ kaže da najviše $k-1$ dobrih x_i leži u $(pr(H \cap H'))^\rho$. Definirajmo A kao skup preostalih dobrih podataka kojih ima najmanje $n - \lfloor n/2 \rfloor + k - 1 - (k-1) = n - \lfloor n/2 \rfloor$. Sada uzmimo neki $(\mathbf{x_a}, \mathbf{y_a}) \in A$ i neka su $e_a = y_a - \mathbf{x_a}\beta$ i $r'_a = y_a - \mathbf{x_a}\hat{\beta}$.

Nadalje, konstruirajmo vertikalnu dvodimenzionalnu ravninu P_a kroz $(\mathbf{x_a}, y_a)$ ortogonalnu na $pr(H \cap H')$. Označimo sa α šiljasti kut koji formiraju H i neki horizontalni pravac iz ravnine P_a. Udaljenost točke $(\mathbf{x_a}, \mathbf{y_a})$ od H je $|e_a|$, a udaljenost projekcije te točke na H od ravnine $\{y = 0\}$ je $|\mathbf{x_a}\beta|$. Sada projekcija točke $(\mathbf{x_a}, \mathbf{y_a})$ na H, projekcija te točke na $\{y = 0\}$ i točka dobivena presjekom $H \cap \{y = 0\}$ čine pravokutan trokut u ravnini P_a, pa vrijedi

$$|e_a| = |\mathbf{x_a}\beta - y_a| \geq \Big|\, |\mathbf{x_a}\beta| - |y_a| \,\Big|.$$

Također,

$$|\tan\alpha| > \frac{|\mathbf{x_a}\beta|}{\rho},$$

pa možemo zaključiti da je kut α u intervalu $\langle -\pi/2, \pi/2 \rangle$. S druge strane, $|\alpha|$ je kut između pravca ortogonalnog na H i vektora $(\mathbf{0}, 1)$, tj. možemo ga promatrati kao kut između $(\beta, 1)$ i $(\mathbf{0}, 1)$. Po formuli za skalarni produkt slijedi

$$|\alpha| = \arccos\left(\frac{|(-\beta, 1), (\mathbf{0}, 1)'|}{||(-\beta, 1)||\ ||(\mathbf{0}, 1)||}\right) = \arccos\left(\frac{1}{\sqrt{1 + ||\beta||^2}}\right).$$

Iz činjenice da je $\cos\alpha = \pm 1/(\sqrt{1 + \tan^2\alpha})$ slijedi $|\tan\alpha| = ||\beta||$. Analogno promotrimo i hiperravninu H' i dobijemo da za odgovarajući kut α vrijedi $\tan\alpha'| = ||\beta'||$. Sada imamo

$$\begin{aligned} |e'_a - e_a| &= |\mathbf{x_a}\beta' - \mathbf{x_a}\beta| \\ &> \rho\,|\tan\alpha' - \tan\alpha| \\ &\geq \rho\,\Big|\,|\tan\alpha'| - |\tan\alpha|\,\Big| \\ &= \rho\,\Big|\,||\beta'|| - ||\beta||\,\Big|. \end{aligned}$$

Znamo da je

$$||\beta - \beta'|| \leq ||\beta|| + ||\beta'|| = 2||\beta|| + (||\beta'|| - ||\beta||) \leq \Big|||\beta'|| - ||\beta||\Big| + 2||\beta||,$$

pa slijedi da je

$$|e'_a - e_a| > \rho\,(||\beta - \beta'|| - 2||\beta||).$$

Sjetimo se, u uzorku Z' imamo $n - \lfloor n/2 \rfloor + k - 1 > \lfloor n/2 \rfloor$ točaka koje su iste kao i u Z. To znači da su reziduali tih točaka obzirom na stari β manji ili jednaki M^2, odnosno medijan kvadrata reziduala je također manji ili jednak M^2. Budući da je β' procjena za Z' dobivena postupkom minimizacije, mora vrijediti

$$\underset{i}{\text{med}}\,(y'_i - x'_i\beta')^2 \leq M^2.$$

Ako pretpostavimo suprotno od onog što želimo pokazati, tj. $||\beta - \beta'|| \geq 2\Big(||\beta|| + M/\rho\Big)$, tada za sve točke indeksa a iz A vrijedi

$$\begin{aligned} |e'_a - e_a| &> \rho\,(||\beta - \beta'|| - 2||\beta||) \\ &> \rho\,(2||\beta|| + 2M\rho - 2||\beta||) = 2M, \end{aligned}$$

pa je

$$|e'_a| \geq |e'_a - e_a| > 2M - M = M,$$

odnosno

$$\underset{i}{\text{med}}\,(y'_i - x'_i\beta')^2 > M^2,$$

jer A ima najmanje $n - \lfloor n/2 \rfloor$ točaka. Time smo dobili kontradikciju, što znači da vrijedi

$$||\beta - \beta'|| < 2\left(||\beta|| + \frac{M}{\rho}\right)$$

za sve Z'. Preostaje pokazati

$$\varepsilon^*_n(T, Z) \leq \frac{\lfloor n/2 \rfloor - k + 2}{n}.$$

Promotrimo uzorke sa $n - \lfloor n/2 \rfloor + k - 2$ dobrih podataka. Uzmimo $k - 1$ takvih i oni će odrediti $(k-1)$-dimenzionalni potprostor L kroz ishodište. Konstruirajmo sada nevertikalnu hiperravninu H' koja sadrži L i određuje neki β' u kontekstu jednadžbe $y = \mathbf{x}\beta'$. Loše podatke proizvoljno biramo, pa ih stavimo u H'. Time smo dobili novi uzorak Z' sa $(\lfloor n/2 \rfloor + k - 2) + (k - 1) = \lfloor n/2 \rfloor + 1$ podataka koji zadovoljavaju $y'_i = x'_i\beta'$. Tada je medijan kvadrata reziduala obzirom na β' jednak nula, što znači da β' ima minimalan medijan reziduala pa je onda to procjena metodom LMS. Odabirući H' sve strmijom i strmijom, dobivat ćemo veći β' odnosno $||\beta - \beta'||$ može se učiniti proizvoljno velikom. Time je tvrdnja teorema dokazana. □

Primijetimo,

$$\lim_{n\to\infty} \frac{\lfloor n/2 \rfloor - k + 2}{n} = \frac{1}{2},$$

odnosno govorit ćemo da je *breakdown* vrijednost LMS metode 50%.

Još jedno zanimljivo svojstvo koje zadovoljava LMS procjena je svojstvo *potpune prilagodbe* (eng. *exact fit property*). Odnosi se na situacije u kojima veliki postotak opažanja točno odgovara nekoj linearnoj jednadžbi. Na primjer, u jednostavnoj regresiji to se događa kad većina podataka leži točno na istom pravcu. U tom slučaju robusna metoda bi trebala „otkriti" tu jednadžbu.

Sljedeći teorem pokazuje da LMS ima svojstvo potpune prilagodbe.

Teorem 3.1.4. *Ako je $k > 1$ i postoji β tako da najmanje $n - \lfloor n/2 \rfloor + k - 1$ opažanja zadovoljava $y_i = x_i\beta$ i opažanja su u generalnoj poziciji, tada je β LMS procjena bez obzira na druga opažanja.*

Dokaz. Postoji neki β tako da najmanje $n - \lfloor n/2 \rfloor + k - 1$ opažanja leži u hiperravnini H danoj jednadžbom $y = x\beta$. Tada β zadovoljava jednadžbu (3.1) jer je $\underset{i}{\text{med}}\, e_i^2(\beta) = 0$. Pretpostavimo da imamo neko drugo rješenje $\beta' \neq \beta$ koje pripada hiperravnini $H' \neq H$ i daje reziduale $e_i(\beta')$. Kao u dokazu teorema 3.1.3, $H \cap H'$ je dimenzije $k - 1$ i stoga sadrži maksimalno $k - 1$ opažanja. Za sva preostala opažanja u H vrijedi $e_i^2(\beta') > 0$, i ima ih minimalno $n - \lfloor n/2 \rfloor$.

Slijedi da je $\underset{i}{\text{med}}\, e_i^2(\beta') > 0$, pa β' ne može biti rješenje. □

Primjer 3.1.5. *Ilustrirajmo ovo svojstvo na jednom primjeru iz jednostavne linearne regresije. Zadano je 9 točaka* $(-4, 0), (-3, 0), (-2, 0), (-1, 0), (0, 0), (1, 0), (2, -5), (3, 5), (12, 1)$. *Od tih 9 točaka, njih 6 leži na pravcu $y = 0$, što vidimo na slici 3.1.*

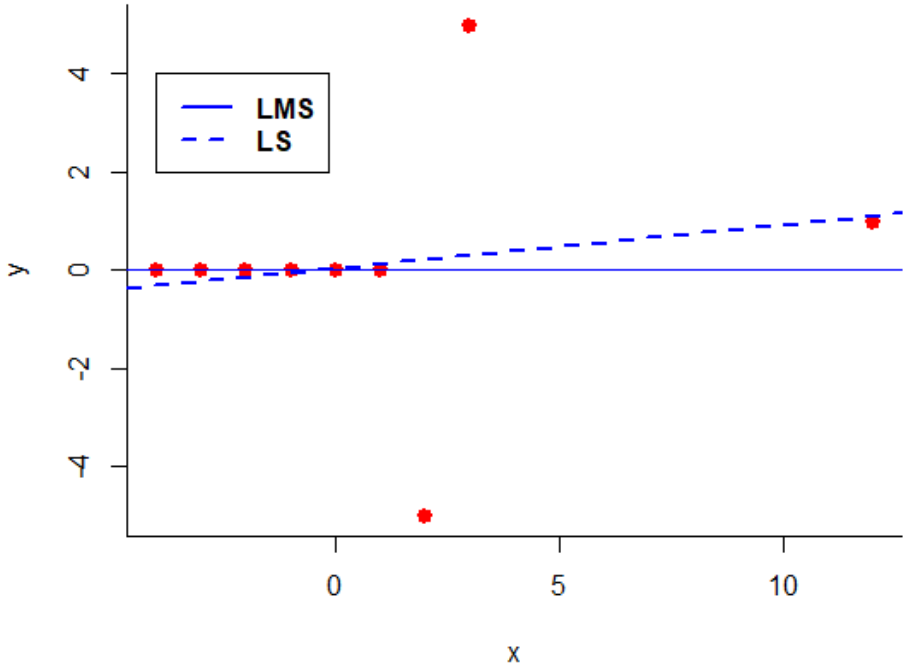
Pravac dobiven metodom najmanjih kvadrata glasi

$$y = 0.09x + 0.032,$$

jer LS nema svojstvo potpune prilagodbe. Budući da 6 od 9 podataka pripada pravcu $y = 0$ to je onda upravo LMS pravac.

3.2 Metoda najmanjih odrezanih kvadrata

Druga robusnija alternativa klasičnoj metodi najmanjih kvadrata koju ćemo predstaviti je metoda najmanjih odrezanih kvadrata, tzv. LTS (eng. *least trimmed squares*). Kao i LMS metodu, i LTS je razvio Rousseeuw [9].



Slika 3.1: Svojstvo potpune prilagodbe

Procjenitelj metodom najmanjih odrezanih kvadrata definiran je s:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{h} e^2_{i:n}, \tag{3.2}$$

gdje $e^2_{i:n}$ predstavlja kvadrate reziduala koji su poredani po veličini od najmanjeg do najvećeg, tj. $e^2_{1:n} \leq e^2_{2:n} \leq \cdots \leq e^2_{n:n}$. Za konstantu h vrijedi $\frac{n}{2} < h \leq n$. Ta konstanta određuje *breakdown* vrijednost, jer ako bolje pogledamo izraz (3.2), on govori da $n - h$ opažanja s najvećim vrijednostima reziduala neće utjecati na procjenitelj. Kasnije ćemo pokazati da se maksimalna *breakdown* vrijednost postiže za $h = \lfloor n/2 \rfloor + \lfloor (k+1)/2 \rfloor$.

Postupak minimizacije (3.2) možemo promatrati na sljedeći način: biramo poduzorak od h opažanja i tražimo β minimizirajući sumu kvadrata reziduala za odabrani poduzorak. Time dobivamo $\binom{n}{h}$ kandidata za LTS procjenitelj pa biramo onog s najmanjom vrijednošću $\sum_{i=1}^{h} e^2_{i:n}$.

Provjerimo zadovoljava li svojstva (2.1), (2.2), (2.3). Dokaz sljedeće leme, kao i svih teorijskih rezultata ovog potpoglavlja mogu se pronaći u [9].

Lema 3.2.1. *Procjenitelj metodom najmanjih odrezanih kvadrata (LTS procjenitelj) je regresijski ekvivarijantan, ekvivarijantan na skaliranje i afino ekvivarijantan.*

Dokaz. Slijedi iz sljedećih jednakosti:

$$\sum_{i=1}^{h}((y_i + x_i\mathbf{v} - x_i\{\mathbf{v}+\beta\})^2)_{i:n} = \sum_{i=1}^{h}((y_i - x_i\beta)^2)_{i:n} = \sum_{i=1}^{h} e^2_{i:n}$$

$$\sum_{i=1}^{h}((cy_i - x_i(c\beta))^2)_{i:n} = c^2\sum_{i=1}^{h}((y_i - x_i\beta)^2)_{i:n}$$

$$\sum_{i=1}^{h}((y_i - x_i\mathbf{A}(\mathbf{A^{-1}}\beta))^2)_{i:n} = \sum_{i=1}^{h}((y_i - x_i\beta)^2)_{i:n}.$$

□

Sljedeći teorem govori o *breakdown* vrijednosti LTS procjenitelja.

Teorem 3.2.2. *Breakdown vrijednost procjenitelja metodom najmanjih odrezanih kvadrata, uz* $h = \lfloor n/2 \rfloor + \lfloor (k+1)/2 \rfloor$*, iznosi*

$$\frac{\lfloor (n-k)/2 \rfloor + 1}{n}. \tag{3.3}$$

Dokaz. Kao i kod Teorema 3.1.3 o *breakdown* vrijednosti LMS procjenitelja, i ovdje pretpostavljamo da su opažanja u generalnoj poziciji.

Prvo pokazujemo:

$$\varepsilon^*_n(T,Z) \geq \frac{\lfloor (n-k)/2 \rfloor + 1}{n}.$$

Budući da se $Z = \{(x_i, y_i) : i = 1, \ldots, n\}$ sastoji od n podataka u generalnoj poziciji, vrijedi da je

$$\rho = \frac{1}{2}\inf\Big\{\tau > 0 : \text{postoji } (k-1)\text{-dimenzionalni potprostor } V \text{ od } \{y = 0\} \text{ kroz ishodište}$$
$$\text{tako da } V^T \text{ sadrži najmanje } k \text{ točaka } x_i \Big\},$$

strogo pozitivna vrijednost.

Neka je β LTS procjenitelj i H pripadajuća hiperravnina određena jednadžbom $y = x\beta$. Nadalje, $M := \max_i |e_i|$ gdje su $e_i = y_i - x_i\beta$ reziduali.

Sada konstruiramo uzorak $Z' = \{(x_i', y_i') : i = 1, \ldots, n\}$ na način da $n - \lfloor (n-k)/2 \rfloor = \lfloor (n+k+1)/2 \rfloor$ podataka iz Z ostavljamo nepromijenjenima, a ostale podatke zamijenimo s proizvoljnim vrijednostima.

Dovoljno je pokazati da je $\|\beta - \beta'\|$ ograničeno, gdje je β' LMS procjenitelj za uzorak Z'. Bez smanjenja općenitosti možemo pretpostaviti da je $\beta \neq \beta'$, što znači da je i $H \neq H'$. Ponavljajući analogno postupak kao u dokazu Teorema 3.1.3, dolazimo do sljedeće nejednakosti:

$$|e_a' - e_a| > \rho\,(\|\beta - \beta'\| - 2\|\beta\|).$$

Sjetimo se, u uzorku Z' imamo najmanje $\lfloor (n+k+1)/2 \rfloor \geq h$ točaka koje su iste kao i u Z. To znači da je suma prvih h kvadrata reziduala tih točaka manja ili jednaka hM^2. Budući da je β' LTS procjena za Z', mora vrijediti

$$\sum_{i=1}^{h} ((y_i' - x_i'\beta')^2)_{i:n} \leq hM^2.$$

Ako pretpostavimo da je $\|\beta - \beta'\| \geq 2\|\beta\| + M(1 + \sqrt{h})/\rho$, tada za sve točke indeksa a iz A vrijedi

$$|e_a' - e_a| > \rho\,(\|\beta - \beta'\| - 2\|\beta\|) \geq M(1 + \sqrt{h}),$$

pa je

$$|e_a'| \geq |e_a' - e_a| - |e_a| > M(1 + \sqrt{h}) - M = M\sqrt{h}.$$

Zbog $n - |A| \leq h - 1$, slijedi

$$\sum_{i=1}^{h} ((y_i' - x_i'\beta')^2)_{i:n} \geq (e_a')^2 > hM^2,$$

a to je kontradikcija. Dakle, $\|\beta - \beta'\| < 2\|\beta\| + M(1 + \sqrt{h})/\rho < \infty$ za svaki Z' definiran kao u dokazu.

Obratna nejednakost

$$\varepsilon_n^*(T, Z) \leq \frac{\lfloor (n-k)/2 \rfloor + 1}{n},$$

slijedi direktno iz Teorema 2.3.2 i Leme 3.2.1. □

Primijetimo,

$$\lim_{n\to\infty} \frac{\lfloor (n-k)/2 \rfloor + 1}{n} = \frac{1}{2},$$

odnosno govorit ćemo da je *breakdown* vrijednost LTS metode 50%.

Na kraju prošlog potpoglavlja spomenuli smo svojstvo potpune prilagodbe koje ima LMS procjena. LTS metoda, kao robusna metoda, također zadovoljava to svojstvo.

Korolar 3.2.3. *Ako postoji β takav da najmanje $\frac{1}{2}(n+k-1)$ opažanja zadovoljava $y_i = x_i\beta$ i opažanja su u generalnoj poziciji, tada je β LTS procjena bez obzira na druga opažanja.*

Primjerice, uzmimo jednostavnu linearnu regresiju i 20 opažanja. Ako 11 opažanja leže na istom pravcu tada je to LTS pravac.

Poglavlje 4

Primjeri

4.1 Primjeri iz ekonomije

Ubrzo nakon predstavljanja metoda LMS i LTS, Rousseeuw je zajedno sa B. Daniels i A. Leroy objavio rad [8] pod nazivom *Applying robust regression to insurance* u kojem analiziraju primjere vezane za osiguranja.

Prvi primjer dolazi nam od velikog belgijskog osiguravajućeg društva. Tablica 4.1 prikazuje mjesečne uplate iz 1979., godine kada je istjecao ugovor o životnom osiguranju. Isplate su evidentirane kao postotak od ukupnog iznosa plaćenog u toj godini. Gledajući ove brojke, primjećujemo blagi opadajući trend tijekom mjeseci. Međutim, u prosincu je isplaćen vrlo velik iznos, što je uglavnom posljedica iznimno visoke dodatne mirovine koja klijentima sjeda. Primjena standardne metode najmanjih kvadrata na ove podatke daje

$$y = 1.327x - 0.294,$$

što odgovara isprekidanoj liniji na slici 4.1. Na taj pravac je značajno utjecala uplata u prosincu i ta točka je *outlier u y-smjeru*.

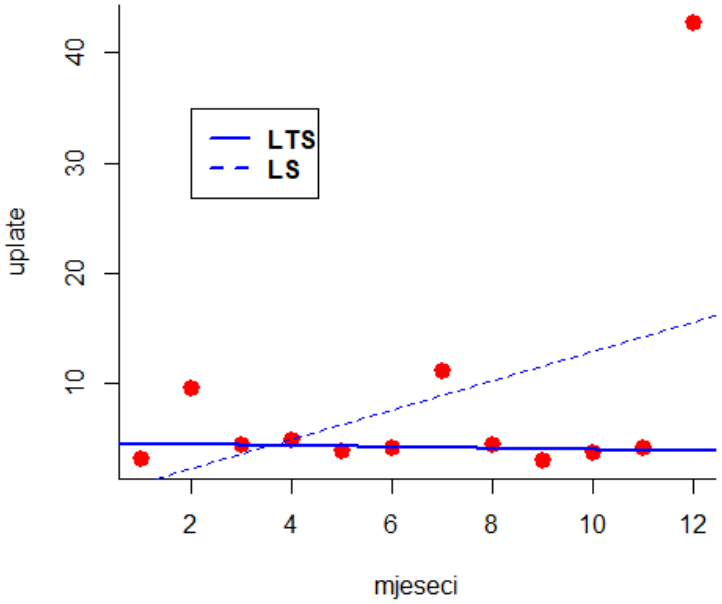
Primjena robusnije metode odrezanih kvadrata (LTS) daje

$$y = -0.052x + 4.661,$$

pravac koji nije pod utjecajem *outliera* odnosno uplate iz prosinca. Na slici je označen punom crtom i jasno je da je on bolje prilagođen većini točaka.

mjeseci (x)	**uplate (y)**
1	3.2
2	9.62
3	4.50
4	4.94
5	4.02
6	4.20
7	11.24
8	4.53
9	3.05
10	3.76
11	4.23
12	42.69

Tablica 4.1: Mjesečne uplate



Slika 4.1: Usporedba pravaca za prvi primjer

Drugi primjer također je vezan uz jedan doista ekstreman *outlier*. Tablica 4.2 prikazuje godišnje stope rasta prosječnih cijena u važnim gradovima Slobodne Kine od 1940. do 1948. Godine 1948. cijene su značajno porasle zbog ogromne državne potrošnje, deficita u proračunu i u konačnici, rata. Sve to dovelo je do hiperinflacije.

Pravac dobiven metodom najmanjih kvadrata glasi

$$y = 24.845x - 1049.468,$$

dok je onaj dobiven robusnom metodom odrezanih kvadrata

$$y = 0.110x - 2.792.$$

Da bismo vidjeli koji od ova dva pravca bolje odgovara podacima, pogledajmo još jednom tablicu 4.2 koja navodi procijenjene vrijednosti prema obje metode. Ispostavlja se da je procjena metodom najmanjih kvadrata loša za svaku točku, dok LTS metoda daje vjerodostojnu aproksimaciju za većinu podataka.

Godina	Rast cijena	Procjena rasta	
(x)	(y)	LTS	LS
40	1.62	1.61	55.67
41	1.63	1.72	-30.82
42	1.90	1.83	-5.98
43	2.64	1.94	18.87
44	2.05	2.05	43.71
45	2.13	2.16	68.56
46	1.94	2.27	93.40
47	15.50	2.38	118.25
48	364.00	2.49	143.09

Tablica 4.2: Usporedba metoda za drugi primjer

4.2 Usporedba LTS i LS u jednostavnoj regresiji

Spominjali smo već kako do *outliera* može doći kada imamo velik broj podataka i neki su jednostavno krivo zapisani, preneseni, izmjereni. Zanimljiv je način na koji je došlo do *outliera* u sljedećem primjeru, preuzetom iz [9].

Naime, radi se o podacima iz jednog statističkog istraživanja u Belgiji koje je objavilo Ministarstvo gospodarstva. Istraživanje sadrži podatke o desecima milijuna internacionalnih odlaznih poziva u Belgiji, u periodu od 1950. do 1973. godine, a možemo ih vidjeti u tablici 4.3. Jasno je da je podaci imaju rastući trend, ali podaci za pozive uspostavljene od 1964. do 1969. djeluju nestvarno veliko. Što se dogodilo?

Ispostavilo se da je u tom periodu korišten drugačiji sustav koji je zapravo prijavljivao ukupan broj minuta ovih poziva umjesto broj poziva kao u drugim godinama (godine 1963. i 1970. također su djelomično zahvaćene jer se prijelazi nisu dogodili baš na Novu godinu, pa je broj poziva od nekih mjeseci dodan broju minuta registriranih u preostalim mjesecima). To je uzrokovalo značajne *outliere u y-smjeru*.

Pravac dobiven metodom najmanjih kvadrata (LS) glasi

$$y = 0.504x - 26.01,$$

i on ne odgovara niti dobrim niti lošim podacima, što se lijepo vidi na slici 4.2. Prikazan je isprekidanom linijom i očito je da su na njega uvelike utjecali loši podaci što je i uzrokovalo tako veliki nagib.

Pravac dobiven robusnom metodom najmanjih odrezanih kvadrata (LTS) glasi:

$$y = 0.116x - 5.616.$$

Godina (x)	Broj poziva (y)
50	0.44
51	0.47
52	0.47
53	0.59
54	0.66
55	0.73
56	0.81
57	0.88
58	1.06
59	1.20
60	1.35
61	1.49
62	1.61
63	2.12
64	11.90
65	12.40
66	14.20
67	15.90
68	18.20
69	21.20
70	4.30
71	2.40
72	2.70
73	2.90

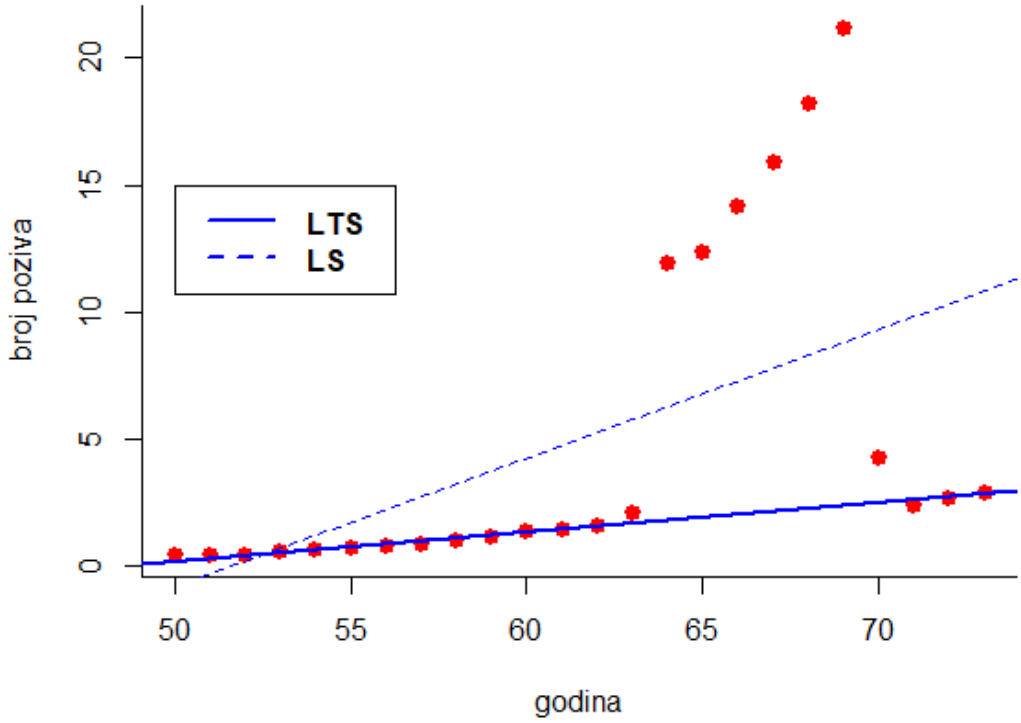
Tablica 4.3: Broj internacionalnih odlaznih poziva iz Belgije

Na slici 4.2 prikazan je punom linijom i vidi se da izbjegava *outliere*, a prati većinu podataka, odnosno prati dobre podatke.

4.3 Usporedba LMS i LS u višestrukoj regresiji

Kao zadnji primjer u ovom radu, napravit ćemo usporedbu klasične LS metode i robusne LMS metode, ali u slučaju višestruke linearne regresije.

Radi se o poznatom primjeru *stackloss* koji je analiziran u brojnim stručnim literaturama od strane mnogih statističara, a preuzet je iz [9]. Podaci opisuju rad postrojenja za



Slika 4.2: Usporedba LS i LTS pravca za internacionalne pozive iz Belgije

oksidaciju amonijaka u dušičnu kiselinu i sastoje se od 21 opažanja koja su navedena u tablici 4.4. Gubitak (y) potrebno je objasniti intezitetom rada (x_1), temperaturom vode koja rashlađuje spremnik (x_2) i koncentracijom kiseline (x_3). Dakle imamo jednu zavisnu i tri nezavisne varijable. Varijabla y zapravo predstavlja deseterostruki postotak ulaznog amonijaka koji se ne iskoristi.

Procjena dobivena metodom najmanjih kvadrata (LS) glasi:

$$y = 0.716x_1 + 1.295x_2 - 0.152x_3 - 39.9.$$

U slučaju jednostavne regresije, *outliere* smo mogli vidjeti „golim okom" preko slike, dok sada ne možemo samo iz jednadžbe pravca zaključiti je li procjena dobra ili ne. U tome nam mogu pomoći *standardizirani reziduali* i Rousseeuw [9]. Oni se računaju kao $e_i/\hat{\sigma}$, pri čemu je σ procjena za standardnu devijaciju. Procjenu za σ^2 imamo u izrazu 1.10, pa

Indeks (i)	Intezitet rada (x_1)	Temperatura (x_2)	Koncentracija kiseline (x_3)	Gubitak (y)
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

Tablica 4.4: Podaci za primjer *stackloss*

je

$$\hat{\sigma} = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4.1}$$

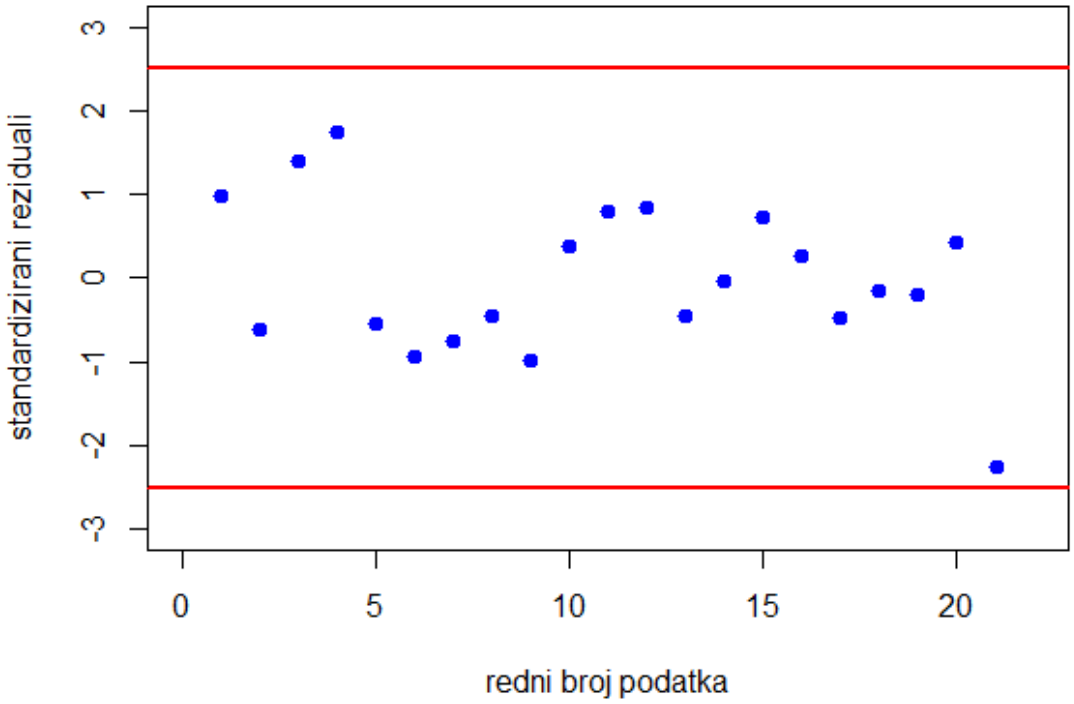
procjena za standardnu devijaciju σ.

Kako bismo bolje interpretirali standardizirane reziduale, Rousseeuw predlaže uvođenje *težina* sa oznakom ω_i.

$$\omega_i = \begin{cases} 1\,, \text{ ako je } |e_i/\hat{\sigma}| \leq 2.5 \\ 0\,, \text{ ako je } |e_i/\hat{\sigma}| > 2.5. \end{cases}$$

Prema ovome bismo rekli da je podatak dobar ako mu je težina 1, a *outlier* ako mu je težina 0. Granica 2.5 je, naravno, samo predložena, ali je sasvim razumna jer ako su greške zaista normalno distribuirane, reziduali ne bi trebali biti veći od $2.5|\hat{\sigma}|$ (više o tome u [9]).

Koristeći formulu 4.1, procjena standardne devijacije za LS procjenu je $\hat{\sigma} = 3.2434$, i možemo računati standardizirane reziduale. Na slici 4.3 nalazi se grafički prikaz standardiziranih reziduala obzirom na redni broj podatka. Crvenom linijom označene su granice -2.5 i 2.5. Prema ovoj slici, svi standardizirani reziduali lijepo upadaju unutar granice i nema *outliera*. Dakle, gledajući jednadžbu LS procjene i graf standardiziranih reziduala ne bismo mogli tvrditi da je procjena loša i da *outlieri* postoje. Radi se o tipičnom *efektu maskiranja*, jer se čini da nema potrebe za robusnom metodom.



Slika 4.3: Standardizirani reziduali LS procjene

Promotrimo sada LMS procjenu. Ona nas dovodi do jednadžbe

$$y = 0.714x_1 + 0.357x_2 - 0x_3 - 34.25.$$

Prema [9], standardna devijacija σ može se također procijeniti na robusniji način. Počinjemo od izračuna s^0 koji se temelji na minimalnom medijanu reziduala i korekcijskom faktoru koji ovisi o n i k:

$$s^0 = 1.4826\Big(1 + \frac{5}{n-k}\Big)\sqrt{\underset{i}{\text{med}}\, e_i^2(\hat{\beta})}.$$

Objašnjenje zašto se koriste faktori 1.4826 i $1 + 5/(n - k)$ može se naći u [9].

Sljedeći korak je uvođenje novih težina obzirom na s^0.

$$\omega_i = \begin{cases} 1\,, \text{ ako je } |e_i/s^0| \leq 2.5 \\ 0\,, \text{ ako je } |e_i/s^0| > 2.5. \end{cases}$$

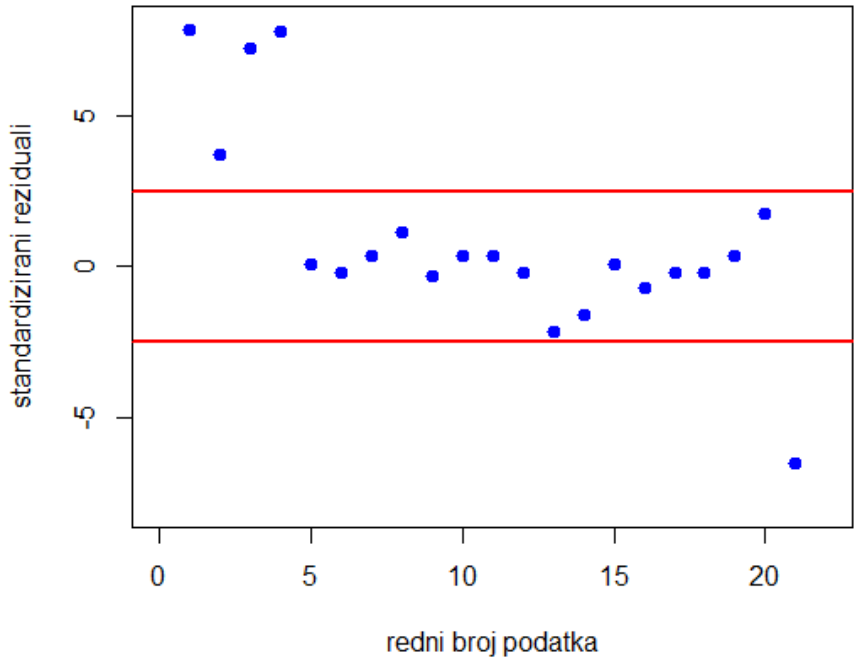
Pomoću navedenih težina, dobije se konačni robusniji procjenitelj standardne devijacije za LMS regresiju

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n} \omega_i e_i^2}{\sum_{i=1}^{n} \omega_i - k}}. \tag{4.2}$$

Na ovakav procjenitelj, *outlieri* nemaju utjecaja. Koristeći formulu 4.2 procjena standardne devijacije za LMS procjenu u ovom primjeru je $\sigma^* = 1.2076$ što je manje od $\hat{\sigma} = 3.2434$ kod LS procjene. Sada možemo izračunati i standardizirane reziduale e_i/σ^2 obzirom na novu procjenu. Na slici 4.4 prikazani su standardizirani reziduali LMS procjene obzirom na redni broj podatka. Sada jasno vidimo da su podaci pod rednim brojevima 1,3,4,21 *outlieri*, a drugi po redu podatak je na granici da bude *outlier*.

Ovaj primjer još jednom ukazuje na opasnost oslanjanja na klasičnu LS procjenu. Potrebno je usporediti standardizirane reziduale metode najmanjih kvadrata i neke robusne metode. Ako se rezultati dva postupka u velikoj mjeri slažu, tada se LS procjeni može vjerovati. Ako se razlikuju, robusna metoda treba se upotrijebiti kao pouzdan alat za procjenu parametara.

Tijekom sedamdesetih i osamdesetih godina prošlog stoljeća statističari su uvelike objavljivali nove radove i knjige na temu robusnih metoda u statistici. Ipak, do danas, ne možemo reći da je neka od robusnijih metoda postala opće prihvaćena kao standardna metoda. Prednosti LMS i LTS metode su svakako njihova dobra svojstva, kao što su ekvivarijantnost, potpuna prilagodba i maksimalna *breakdown* vrijednost. Njihova mana je manjak efikasnosti i velika računska složenost u slučaju velikog broja podataka. Zbog toga nije čudno što su brojni statističari smatrali klasičnu LS metodu dovoljno robusnom. Algoritmi robusnih metoda dugo nisu uopće bili implementirani u statističke pakete. Međutim i to se mijenja, nove modernije knjige govore više o robusnim metodama, a i danas statistički softverski paketi poput R-a imaju alate za robusnu regresiju i metode poput LMS i LTS se više koriste.



Slika 4.4: Standardizirani reziduali LMS procjene

Bibliografija

[1] Y. Dodge, *The Concise Encyclopedia of Statistics*, Springer, 2008.

[2] F. R. Hampel, *The Influence Curve and its Role in Robust Estimation*, Journal of the American Statistical Association **69** (1974), br. 346, 383–393.

[3] P. J. Huber, *Robust Statistics*, John Wiley & Sons, 1981.

[4] P. J. Huber i E. M. Ronchetti, *Robust Statistics, second edition*, John Wiley & Sons, 2009.

[5] M. Huzak, *Linearni regresijski model*, `https://web.math.pmf.unizg.hr/nastava/stat/files/StatRegresija.pdf`, preuzeto 29.7.2021.

[6] D. C. Montgomery, E. A. Peck i G. G. Vining, *Introduction to linear regression analysis, fifth edition*, John Wiley & Sons, 2012.

[7] P. J. Rousseeuw, *Least Median of Squares Regression*, Journal of the American Statistical Association **79** (1984), br. 388, 871–880.

[8] P. J. Rousseeuw, B. Daniels i A. Leroy, *Applying robust regression to insurance*, Insurance: Mathematics and Economics **3** (1984), br. 1, 67–72.

[9] P. J. Rousseeuw i A. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, 1987.

[10] G. A. F. Seber i A. J. Lee, *Linear regression analysis, second edition*, John Wiley & Sons, 2003.

[11] R. R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing, fourth edition*, Academic Press, 2017.

[12] X. Yan i X. Gang Su, *Linear Regression Analysis: Theory and Computing*, World Scientific, 2009.

Dodatak A

R kod korišten u primjerima

Slika 2.1a:

```
data=read.table("primjer1.txt",header = TRUE)
x=data$x
y=data$y
plot(x,y,pch=19,col="red",lwd=6,ylim=c(1.2,2.5),
xlab ="x",ylab ="y",bty="l")
fit=lm(y~x)
abline(fit,col="blue")
```

Slika 2.1b:

```
data=read.table("primjer2.txt",header = TRUE)
x=data$x
y=data$y
plot(x,y,pch=19,col="red",lwd=6,ylim=c(1,5),
xlab ="x",ylab ="y",bty="l")
fit=lm(y~x)
abline(fit,col="blue")
```

Slika 2.2a:

```
data=read.table("primjer3.txt",header = TRUE)
x=data$x
y=data$y
plot(x4,y4,pch=19,col="red",lwd=6,ylim=c(0,6),
xlim=c(1.5,2.5),xlab ="x",ylab ="y",bty="l")
fit=lm(y~x)
abline(fit,col="blue")
```

Slika 2.2b:

```
data=read.table("primjer4.txt",header = TRUE)
x=data$x
y=data$y
plot(x4,y4,pch=19,col="red",lwd=6,ylim=c(0,6),
xlim=c(1.5,4.5),xlab ="x",ylab ="y",bty="l")
fit=lm(y~x)
abline(fit,col="blue")
```

Slika 2.3:

```
data=read.table("leverage.txt",header = TRUE)
x=data$x
y=data$y
plot(x5,y5,pch=19,col="red",lwd=6,ylim=c(1,5),
xlim=c(0.5,5),xlab ="x",ylab ="y",bty="l")
fit=lm(y5~x5)
abline(fit,col="blue")
points(4.2,4.4,pch=19,col="green",lwd=7)
```

Slika 3.1:

```
x_i=c(-4,-3,-2,-1,0,1,2,3,12)
y_i=c(0,0,0,0,0,0,-5,5,1)
plot(x_i,y_i,pch=19,col="red",lwd=3,xlab ="x",
ylab ="y",bty="l")
fit1=lm(y_i~x_i)
fit1
# Call:
#  lm(formula = y_i ~ x_i)
# Coefficients:
# (Intercept)          x_i
#     0.03194      0.08907
abline(fit1,col="blue",lwd=2,lty=2)
library(MASS)
fit2=lmsreg(y_i~x_i)
fit2
# Call:
#   lqs.formula(formula = y_i ~ x_i, method = "lms")
# Coefficients:
# (Intercept)          x_i
#           0            0
abline(fit2,col="blue",lwd=1)
```

```
legend(-4,4, legend=c("LMS", "LS"),
col=c("blue", "blue"),lty = 1:2,text.font = 2 ,lwd=2)
```

Slika 4.1:

```
data=read.table("insurance.txt",header = TRUE)
x=data$mjeseci
y=data$uplate
plot(x,y,pch=19,col="red",lwd=5,xlab ="mjeseci",
ylab ="uplate",bty="l")
fit1=lm(y~x)
fit1
# Call:
#  lm(formula = y ~ x)
# Coefficients:
#   (Intercept)             x
#       -0.2939        1.3273
abline(fit1 ,col="blue",lty=2)
library(MASS)
fit2=ltsreg(y~x)
fit2
# Call:
#   lqs.formula(formula = y ~ x, method = "lts")
# Coefficients:
#   (Intercept)             x
#         4.661        -0.052
abline(fit2 ,col="blue",lwd=2)
legend(2,35, legend=c("LTS", "LS"),
col=c("blue", "blue"),lty = 1:2,text.font = 2 ,lwd=2)
```

Slika 4.2:

```
data=read.table("phonecall.txt",header = TRUE)
x=data$godina
y=data$pozivi
plot(x,y,pch=19,col="red",lwd=3,xlab ="godina",
ylab ="broj poziva",bty="l")
fit1=lm(y~x)
fit1
# Call:
#   lm(formula = y ~ x)
# Coefficients:
```

```
#    (Intercept)              x
#        -26.0059         0.5041
abline(fit1,col="blue",lty=2)
fit2=ltsreg(y~x)
fit2
# Call:
#    lqs.formula(formula = y ~ x, method = "lts")
# Coefficients:
#    (Intercept)              x
#        -5.6162         0.1159
abline(fit2,col="blue",lwd=2)
legend(50,15, legend=c("LTS", "LS"),
col=c("blue", "blue"),lty = 1:2,text.font = 2 ,lwd=2)
```

Slika 4.3:

```
data=read.table("gubitak.txt",header = TRUE)
attach(data)
model=lm(gubitak~intezitet+temperatura+koncentracija)
summary(model)
# Call:
# lm(formula = gubitak ~ intezitet + temperatura
 + koncentracija)
# Residuals:
# Min      1Q  Median      3Q     Max
# -7.2377 -1.7117 -0.4551  2.3614  5.6978
# Coefficients:
# Estimate Std. Error t value Pr(>|t|)
# (Intercept)   -39.9197    11.8960  -3.356  0.00375 **
# intezitet       0.7156     0.1349   5.307  5.8e-05 ***
# temperatura     1.2953     0.3680   3.520  0.00263 **
# koncentracija  -0.1521     0.1563  -0.973  0.34405
# ---
# Residual standard error: 3.243 on 17 degrees of freedom
# Multiple R-squared:  0.9136,  Adjusted R-squared:  0.8983
# F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
y_procjena=0.716*intezitet+1.295*temperatura
-0.152*koncentracija-39.9
sigma=sqrt((1/17)*(sum((y-y_procjena)^2)))
# primijetimo da je sigma jednaka Residual standard error
# kod ispisa summary(model)
```

```
plot(c(1:21),(y-y_procjena)/sigma,pch=19,lwd=2,col="blue",
ylim=c(-3,3),xlim=c(0,22),xlab ="redni broj podatka",
ylab ="standardizirani reziduali")
abline(h=-2.5,col="red",lwd=2)
abline(h=2.5,col="red",lwd=2)
```

Slika 4.4:

```
data=read.table("gubitak.txt",header = TRUE)
attach(data)
model_lms=lmsreg(gubitak~intezitet+temperatura+koncentracija)
model_lms
# Call:
# lqs.formula(formula = gubitak ~ intezitet + temperatura
# + koncentracija,
# method = "lms")
# Coefficients:
# (Intercept)      intezitet     temperatura   koncentracija
# -3.425e+01       7.143e-01      3.571e-01       5.146e-16
y_procjena=0.714*intezitet+0.357*temperatura
-0*koncentracija-34.25
sigma=1.2076
plot(c(1:21),(y-y_procjena)/sigma,pch=19,lwd=2,col="blue",
ylim=c(-8,8),xlim=c(0,22),xlab ="redni broj podatka",
ylab ="standardizirani reziduali")
abline(h=-2.5,col="red",lwd=2)
abline(h=2.5,col="red",lwd=2)
```

Sažetak

Robusne statističke metode osmišljene su kako bi se suzbili neki od problema koji se pojavljuju kod klasične statističke analize, kao što su odstupanja od pretpostavki modela ili neuobičajene vrijednosti (*outlieri*). Posebno, kod procjene parametara modela linearne regresije, standardna metoda najmanjih kvadrata veoma je osjetljiva na opažanja koja znatno odstupaju od ostalih.

Nakon predstavljanja općih modela jednostavne linearne regresije, višestruke linearne regresije i metode najmanjih kvadrata, uvedene su dvije mjere robusnosti - *breakdown* vrijednost i funkcija utjecaja. Zatim su prikazane dvije robusne metode, metoda najmanjeg medijana kvadrata odstupanja i metoda najmanjih odrezanih kvadrata. Za te dvije metode dokazano je da postižu maksimalnu *breakdown* vrijednost. U konačnici, ilustrirani su primjeri na stvarnim podacima radi usporedbe klasične metode najmanjih kvadrata sa robusnim metodama. Primjeri su rađeni u programskom jeziku R.

Summary

Robust statistical methods are designed to combat some of the problems that arise in classical statistical analysis, such as deviations from model assumptions and outliers. In particular, when estimating the parameters of a linear regression model, the standard least squares method is very sensitive to observations that deviate significantly from the others.

After presenting the general forms of the simple linear regression model, multiple linear regression and the least squares method, two measures of robustness were introduced - the breakdown value and the influence function. Furthermore, two robust methods are presented, the least median squares method and the least trimmed squares method. These two methods have been shown to achieve a maximum breakdown value. Finally, examples with some real data and a comparison of the classical least squares method with robust methods are illustrated, using the programming language R.

Životopis

Rođen sam 23. srpnja 1996. godine u Splitu. Odrastao sam u Seget Vranjici, a pohađao sam Osnovnu školu kralja Zvonimira u Segetu Donjem. Srednjoškolsko obrazovanje nastavljam u Trogiru, gdje upisujem Srednju školu Ivana Lucića, smjer opća gimnazija. Završetkom srednje škole, 2014. godine upisujem preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu, kojeg završavam 2019. i stječem titulu sveučilišnog prvostupnika matematike. Te godine na istom fakultetu upisujem diplomski studij Financijska i poslovna matematika kojeg završavam ovim Radom.