

Linearna regresija - potencijalni problemi

Kada prilagodimo model linearne regresije određenom skupu podataka, tada se može javiti mnogo problema. Među njima su najčešće:

1. Nelinearnost između prediktora i odgovora (Non-linearity of the response-predictor relationships).
2. Korelacija grešaka (Correlation of error terms).
3. Nekonstantna varijansa grešaka (Non-constant variance of error terms).
4. Outliers.
5. Tačke sa visokom polugom (High-leverage points).
6. Kolinearnost.

U praksi je identifikovanje i prevazilaženje ovih problema umetnost koliko i nauka. O tome je napisano mnogo stranica u bezbroj knjiga. 😊

1. Podaci nisu linearni

Ako je stvarna veza daleko od linearne, tada su sumnjivi praktično svi zaključci koje izvučemo iz modela. Pored toga, tačnost predviđanja modela se može značajno smanjiti.

Grafik reziduala je koristan grafički alat za prepoznavanje nelinearnosti u podacima. Za model *proste linearne regresije*, možemo prikazati reziduale, $e_i = y_i - \hat{y}_i$, naspram prediktora x_i . U slučaju modela višestruke regresije, budući da postoji više prediktora, prikazujemo reziduale u odnosu na predviđene (ili uklopljene) vrednosti $\hat{y}_i$. U idealnom slučaju, grafik reziduala neće pokazivati neki uočljivi obrazac. Prisustvo šablona može ukazivati na problem sa nekim aspektom linearnog modela.

Ako grafik reziduala ukazuje na to da u podacima postoje nelinearne veze, tada je jednostavan pristup upotreba nelinearnih transformacija prediktora, kao što su $\log X$, $\sqrt{X}$ i X^2 , u modelu regresije.

PRIMER: Plot residuals za $mpg \sim horsepower$ i $mpg \sim horsepower^2$

2. Korelacija grešaka

Važna pretpostavka modela linearne regresije je da greške $e_1, e_2, \dots, e_n$, nisu u korelaciji!

Šta ovo znači? Na primer, ako greške nisu korelisane, onda činjenica da je e_i pozitivno daje malo ili nimalo informacija o znaku naredne greške e_{i+1} . Standardne greške koje se izračunavaju za procenjene koeficijente regresije ili predviđene vrednosti zasnivaju se na pretpostavci nekorelisanih grešaka.

Ako postoji korelacija između grešaka, tada će procenjene standardne greške težiti da potcene (eng. *underestimate*) prave standardne greške. Kao rezultat, intervali poverenja (eng. *confidence intervals*) i predviđanja (eng. *prediction intervals*) biće uži nego što to stvarno jesu. Na primer, interval poverenja od 95% može u stvarnosti imati mnogo manju verovatnoću od 0.95 da sadrži pravu vrednost parametra. Pored toga, p -vrednosti koje su povezane sa modelom biće niže nego što bi trebalo da budu; ovo bi moglo dovesti do toga da pogrešno zaključimo da je parametar statistički značajan. *Ukratko, ako su greške u nekoj korelaciji, možda imamo neopravdan osećaj poverenja u naš model.*

Kao krajnji primer, pretpostavimo da smo slučajno udvostručili svoje podatke, što je dovelo do toga da su observacije i izrazi grešaka identični u parovima. Ako bismo ovo zanemarili, naši standardni proračuni grešaka bili bi kao da imamo uzorak veličine $2n$, a u stvari imamo samo n uzoraka. Naši procenjeni parametri bi bili isto za $2n$ uzoraka kao i za n uzoraka, ali intervali pouzdanosti bi bili uži za faktor $\sqrt{2}$!

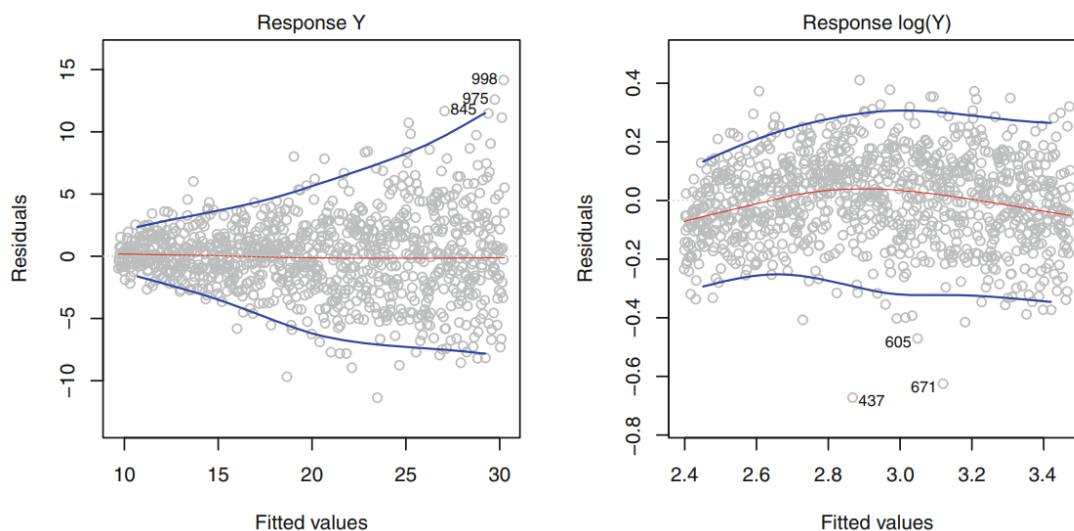
Zašto se mogu pojaviti korelacije među terminima grešaka? Takve korelacije se često javljaju u kontekstu podataka o vremenskim serijama, koje se sastoje od observacija za koja se mere vrednosti u diskretnim tačkama vremena.

3. Nekonstantna varijansa grešaka

Još jedna važna pretpostavka modela linearne regresije je da greške imaju konstantnu varijansu, $Var(e_i) = \sigma^2$. Standardne greške, intervali poverenja i testovi hipoteza koji su odnose na linearni model oslanjaju se na ovu pretpostavku.

Nažalost, često je slučaj da varijanse grešaka nisu stabilne. Na primer, varijanse grešaka mogu se povećati sa vrednošću dobijenog outputa. Nekonstantne varijacije u greškama (eng. *heteroscedasticity*) se mogu detektovati iz postojanja oblika levka (eng. *funnel shape*) na grafiku reziduala.

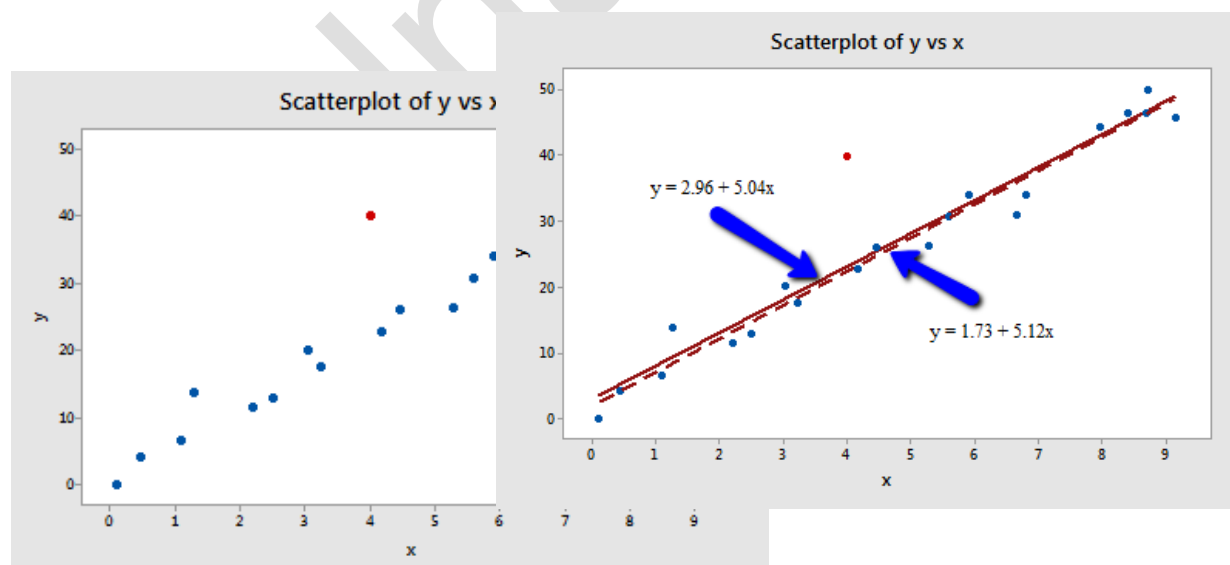
Kada se suočite sa ovim problemom, jedno od mogućih rešenja je transformisanje outputa Y pomoću konkavne funkcije kao što su $\log Y$ ili $\sqrt{Y}$. Takva transformacija rezultira većom količinom skupljanja outputa sa većim vrednostima, što dovodi do smanjenja ove pojave.



Slika 1. Grafik reziduala. **Levo:** levkasti oblik; **Desno:** Prediktor je transformisan pomoću logaritamske funkcije, tako da sada nemamo izgled levka.

4. Anomalije (Outliers)

Izuzetak je tačka za koju je y_i daleko od vrednosti koju predviđa model. Izuzeci mogu nastati iz različitih razloga, kao što je npr. netačno snimanje observacija tokom prikupljanja podataka.

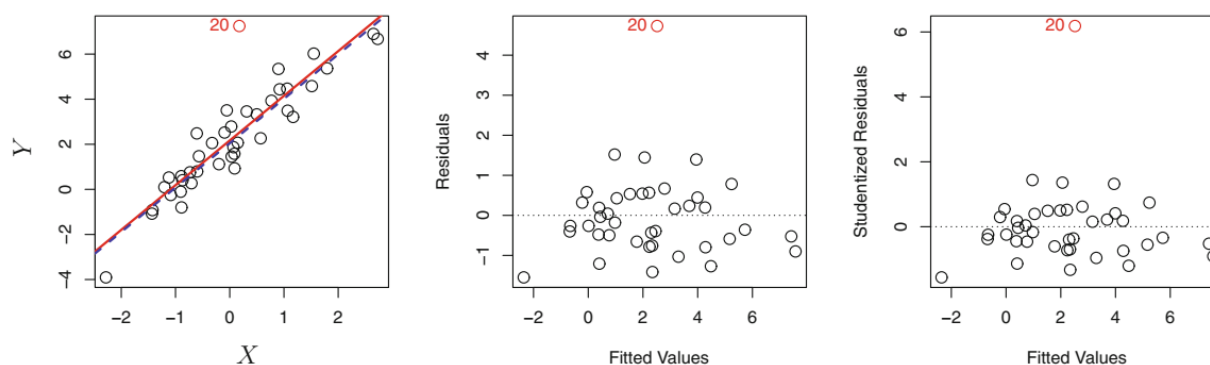


U ovom slučaju, uklanjanje autlajera ima malo efekta na liniju dobijenu pomoću metode najmanjih kvadrata, tj. uklanjanje tačke dovodi vrlo male promene nagiba i malog smanjenja

preseka. Tipično je za autlajer koji nema neobičnu vrednost prediktora da ima mali uticaj na metodu najmanjih kvadrata. Međutim, čak i ako autlajer nema puno uticaja na izgled linije, to može izazvati druge probleme.

Na primer, RSE se može drastično razlikovati kada je outlier uključen u regresiju, i kada se outlier ukloni. Budući da se RSE koristi za izračunavanje svih intervala poverenja i p -vrednosti, tako dramatično povećanje izazvano jednom tačkom podataka može imati implikacije na tumačenje modela.

Grafik reziduala se može koristiti za detektovanje outliers. Ali u praksi može biti teško odlučiti koliko veliki rezidual mora biti pre nego što observaciju proglasimo autlajerom. Sada, umesto da nacrtamo grafik reziduala, možemo da nacrtamo *studentized reziduala*, koji se izračunavaju deljenjem svakog reziduala e_i sa njegovom procenjenom standardnom greškom. **Observacije za koje su *studentized reziduali* po apsolutnoj vrednosti veći od 3 su mogući autlajeri.** *Studentized reziduali* se mogu intepretirati kao broj standardnih grešaka udaljenih od regresione linije.



5. Tačke sa visokom polugom (high leverage)

Upravo smo videli da su odstupanja zapažanja na koja je odgovor y_i neobično s obzirom na prediktor x_i . Suprotno tome, zapažanja sa visokom polugom imaju neobičnu vrednost za x_i . U praksi se može zaključiti da uklanjanje observacije sa velikom polugom ima mnogo značajniji uticaj na liniju najmanjih kvadrata od uklanjanja autlajera. U stvari, zapažanja sa visokom polugom imaju značajan uticaj na procenjenu liniju regresije. Iz tog razloga je važno identifikovati zapažanja sa visokom polugom.

U prosto linearnoj regresiji, observacije sa visokim polugama prilično je lako identifikovati, jer jednostavno možemo tražiti observacije za koja je vrednost prediktora izvan normalnog opsega observacija. Ali u višestrukoj linearnoj regresiji sa više prediktora, moguće je imati observacije koje se nalazi unutar opsega vrednosti svakog pojedinog prediktora, ali da je observacija „neobična“ u slučaju punog skupa prediktora.

Da bismo kvantifikovali polugu observacije, izračunavamo *leverage statistic*. Velika vrednost ove statistike ukazuje na zapažanje sa visokom polugom. Za prostu linearnu regresiju obrazac glasi:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^n (x_{i'} - \bar{x})^2}$$

Postoji jednostavno proširenje za h_i na slučaj više prediktora, ali ovde ne dajemo formulu. *Leverage statistic* h_i je uvek između $1/n$ i 1 , a prosečna poluga za sva posmatranja je uvek jednaka $(p + 1)/n$. Ako je vrednost statistike veća od $2 * (p + 1)/n$, tada se smatra da observacija ima veliku polugu.

```
fitted = predict(hp_model, as.data.frame(Auto["horsepower"]))
```

6. Kolinearnost (Collinearity)

Kolinearnost se odnosi na situaciju u kojoj su dve ili više prediktorskih promenljivih usko povezane jedna sa drugom.

Credit dataset: *limit vs rating* or *limit vs age*.

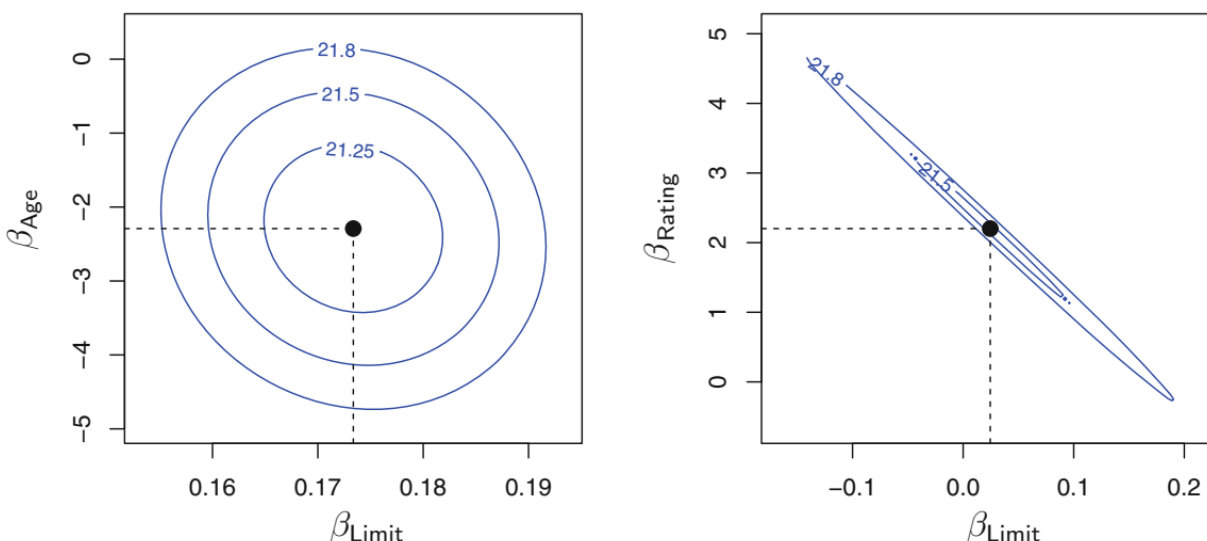


Figure 2. Contour plot za RSS vrednosti u funkciji od parametara β za dve različite regresije koje uključuju podatke o kreditima.

Mala promena podataka može prouzrokovati da se par vrednosti koeficijenata koji daju najmanji RSS pomera bilo gde duž ove doline. To rezultira velikom nesigurnošću u procenama

koeficijenata. **Budući da kolinearnost smanjuje tačnost procena koeficijenata regresije, to dovodi do rasta standardne greške za $\hat{\beta}_j$.** Takođe, kolinearnost rezultira padom t -statistike.

Jednostavan način otkrivanja kolinearnosti je pregled matrice korelacije prediktora. Element ove matrice koji je velik u apsolutnoj vrednosti ukazuje na par visoko koreliranih promenljivih, a samim tim i na problem kolinearnosti u podacima. Na nesreću, svi problemi sa kolinearnošću ne mogu se otkriti pregledom matrice korelacije: *moгуće je da kolinearnost postoji između tri ili više promenljivih, čak i ako nijedan par promenljivih nema posebno visoku korelaciju*. Ovu situaciju nazivamo **multikolinearnost**. Umesto provere matrice korelacije, bolji način za procenu multikolinearnosti je izračunavanje *faktora varijacije inflacije* (VIF, *variance inflation factor*).

U podacima o kreditu, regresija bilansa u odnosu na starost, rejting i limit ukazuje na to da prediktori imaju VIF vrednosti od 1.01, 160.67 i 160.59.

Kada se suočite sa problemom kolinearnosti, postoje dva jednostavna rešenja.

- a) Prvo je brisanje jedne od problematičnih promenljivih iz regresije.
- b) Drugo rešenje je kombinovanje kolinearnih promenljivih u jedan prediktor.

Skup podataka za marketing

1. Da li postoji veza između budžeta za oglašavanje i prodaje?

```
> model = lm(sales ~ ., data = adver)
> summary(model)

Call:
lm(formula = sales ~ ., data = adver)

Residuals:
    Min       1Q   Median       3Q      Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422  <2e-16 ***
TV           0.045765   0.001395  32.809  <2e-16 ***
radio        0.188530   0.008611  21.893  <2e-16 ***
newspaper    -0.001037   0.005871  -0.177    0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

2. Koliko je jak odnos između budžeta za oglašavanje i prodaje?

Za podatke o oglašavanju, RSE je 1.686 jedinica, dok je srednja vrednost outputa 14.022, što ukazuje na procentualnu grešku od oko 12%. Drugo, R^2 statistika beleži procenat varijabilnosti u odgovoru koji se objašnjava prediktorima. Prediktori objašnjavaju gotovo 90% varijanse u prodaji.

3. Koji mediji doprinose prodaji?

Da bismo odgovorili na ovo pitanje, moramo pronaći način da razdvojimo pojedinačne efekte svakog medija kada smo potrošili novac na sva tri medija. U višestrukoj linearnoj regresiji prikazanoj u prethodnoj tabeli, p -vrednosti za TV i radio su niske, ali p -vrednost za novine nije.

4. Koliko tačno možemo proceniti efekat svakog medija na prodaju?

95% intervali poverenja su sledeći: (0,043, 0,049) za TV, (0,172, 0,206) za radio i (-0,013, 0,011) za novine. Interval za novine uključuje nulu, što ukazuje da promenljiva nije statistički značajna u odnosu na vrednosti TV-a i radija.

Kolinearnost može da rezultira vrlo širokim standardnim greškama. Da li bi kolinearnost mogla biti razlog što je interval poverenja vezan za novine tako širok? Rezultati VIF-a su 1.005, 1.145 i 1.145 za TV, radio i novine, što ne ukazuje da postoji kolinearnost.

Dodatno, da bi se procenila povezanost svakog medija pojedinačno na prodaju, možemo napraviti tri odvojene proste linearne regresije.

5. Koliko tačno možemo predvideti buduću prodaju?

Tačnost povezana sa ovom procenom zavisi od toga da li želimo da predvidimo pojedinačni odgovor, $Y = f(X) + e$, ili prosečni odgovor, $f(X)$. Ako je prvo, koristimo interval predikcije, a ako je drugo, koristimo interval poverenja.

Intervali predviđanja uvek će biti širi od intervala poverenja, jer su oni odgovorni za neizvesnost povezanu sa e , ireducibilnom greškom.

6. Da li je odnos linearan?

Ranije smo videli da se grafik rezidula može koristiti za identifikaciju nelinearnosti u podacima. Ako su odnosi linearni, onda grafik reziduala ne bi trebalo da pokaže neki obrazac.

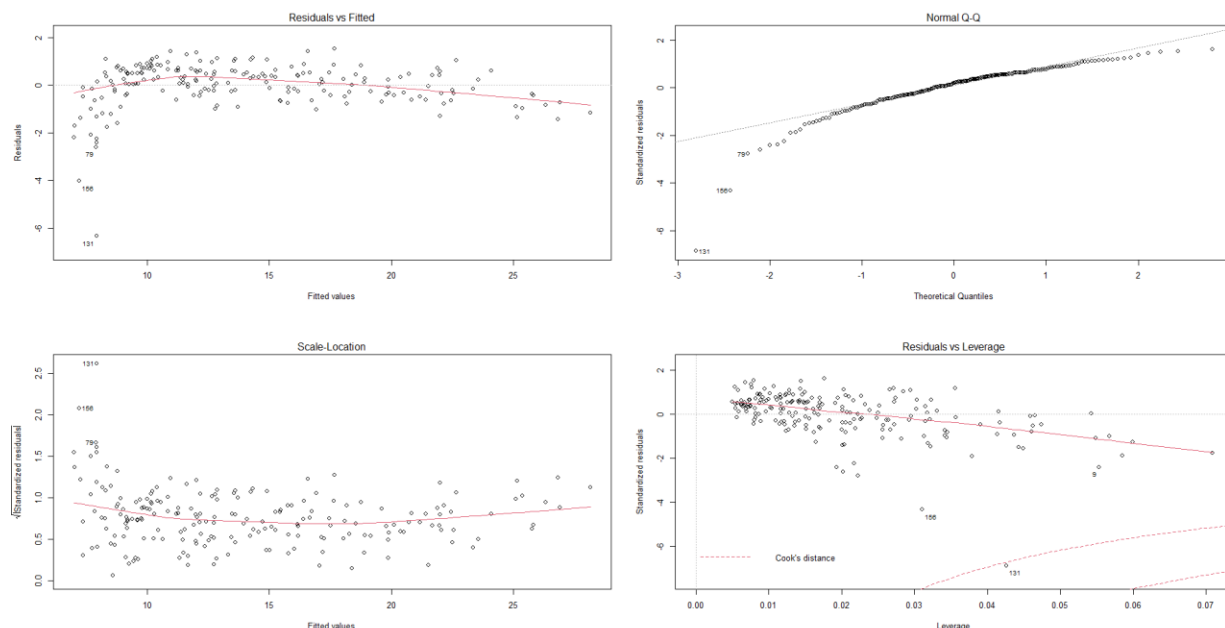
7. Postoji li sinergija među reklamnim medijima?

Možda potrošiti 50.000 USD na televizijsko oglašavanje i 50.000 USD na radio oglašavanje rezultira većom prodajom od izdvajanja 100.000 američkih dolara bilo televizija ili radio pojedinačno. U marketingu je ovo poznato kao *efekat sinergije*, dok se u statistici naziva *efektom interakcije*.

Da li postoji sinergija među prediktorima?

Interpretacija grafika iz R-a

```
par(mfrow = c(2, 2))
plot(model)
```



Residuals vs Fitted. Koristi se za proveru pretpostavki linearnog odnosa. Horizontalna linija, bez jasnih obrazaca, pokazatelj je linearnog odnosa, šta je dobro. (Stavka 1)

Normal Q-Q (Quantile-Quantile plot). Koristi se za ispitivanje da li se reziduali normalno distribuiraju. Dobro je ako reziduali prate ravnu isprekidanu liniju.

Scale-Location (or Spread-Location). Koristi se za proveru homogenosti varijanse reziduala (homoscedasticity). Horizontalna linija sa jednako raširenim tačkama je dobar pokazatelj dobrog svojstva. (Stavka 3)

Residuals vs Leverage. Koristi se za identifikaciju uticajnih slučajeva, odnosno ekstremnih vrednosti koje mogu uticati na rezultate regresije kada se uključe ili izuzmu iz analize. (Stavke 4 i 5)

- Uticajna vrednost (eng. *influential value*) je vrednost koja uključivanjem ili isključivanjem može promeniti rezultate regresione analize. Takva vrednost je povezana sa velikom vrednošću reziduala. Međutim, nisu sva odstupanja (ili ekstremne tačke podataka) uticajna na linearnu regresionu analizu.
- Statističari su razvili metriku koja se naziva **Kukova udaljenost** da bi utvrdili uticaj vrednosti. Ova metrika definiše uticaj kao kombinaciju poluge i rezidualne veličine.

Osnovno pravilo je da observacija ima veliki uticaj ako Kukova udaljenost prelazi vrednost $4/(n - p - 1)$ (P. Bruce and Bruce 2017), gde je n broj observacija, a p broj prediktora.