“Rešavanje problema heteroskedastičnosti”

Aleksandra Ilić 286/2015, Borisav Damnjanović 399/2014

Uvod

Jedna od stvari koja se najčešće koristi u statitičkom zaključivanju jeste linearna regresija. Da bismo mogli da primenimo linearnu regresiju na podacima potrebno je da određeni uslovi budu zadovoljeni. Jedan od uslova koji treba da bude zadovoljen jeste homoskedastičnost.

U ovom radu prvo ćemo se osvrnuti na pojmove iz linearne regresije i posebno na pojam homoskedastičnosti. Zatim ćemo uvesti i obraditi nove reziduale, koji će bit glavna tema rada. Na kraju ćemo kroz primere prikazati njihovu primenu na realnim podacima i time zaokružiti temu.

Kratak osvrt na linearne statističke modele

Da bi cela priča iz nastavka rada bila što jasnija, prvo ćemo se kratko osvrnuti na osnovne pojmove iz linearne regresije.

Pretpostavimo da imamo p prediktora $X_1, X_2, \ldots, X_p$. Tada linearni model možemo zapisati u obliku

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \;\; i = 1, 2, \ldots, n$$

ili u obliku

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \;\; i = 1, 2, \ldots, n,$$

pri čemu je n obim uzorka koji imamo.

Ovu jednakost možemo zapisati i matrično kao

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

gde su $Y = (Y_1, \ldots, Y_n)^T$ i $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ vektori dimenzije n, $X = (X_{ij})$ $n \times p$ matrica i $\beta = (\beta_1, \ldots, \beta_p)$ vektor dimenzije p.

Pretpostavićemo da za naš model važe sledeće pretpostavke:

1. linearnost - model je linearan po β
2. nezavisnost i heteroskedastičnost grešaka - $Cov(\varepsilon) = diag\{\sigma_1^2, \ldots, \sigma_n^2\} = \Sigma$, $0 < \sigma_i^2 < \infty$, za $i = 1, \ldots, n$
3. maksimalan rang - matrica X je maksimalnog ranga tj. $r(X) = p$
4. normalnost grešaka - $\varepsilon \sim N_n(0, \Sigma)$; $N_n(0, \Sigma)$ je n-dimenziona normalna raspodela sa očekivanjem 0 i kovarijacionom matricom Σ.

Kada bi nam greške bile homoskedastične važilo bi $Cov(\varepsilon) = \sigma^2 I_n$, $\sigma^2 > 0$, gde je I_n matrica dimenzije $n \times n$, sa jedinicama na dijagonali i nulama na svim ostalim mestima.

Pošto smo pretpostavili da je matrica X maksimalnog ranga, matrica $X^T X$ je invertibilna. Ocena za β metodom najmanjih kvadrata je $\hat{\beta} = (X^T X)^{-1} X^T Y$.

S obzirom na pretpostavku normalnosti, $\hat{\beta}$ je ocena i metodom maksimalne verodostojnosti i $\hat{\beta} \sim N_p(\beta, (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1})$. Označimo $P = (X^T X)^{-1} X^T$ i tada je $\hat{\beta} \sim N_p(\beta, P \Sigma P^T)$.

Ocena zavisne promenljive je $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$.

Matrica H, koja se naziva hat matrix, je projektor, tako da $\hat{Y}$ predstavlja ortogonalnu projekciju vektora Y na ravan generisanu sa X. Elemente na njenoj dijagonali (težine) označavamo sa h_i. Reziduale ćemo definisati sa $e \equiv Y - \hat{Y} = Y - HY = (I_n - H)Y$.

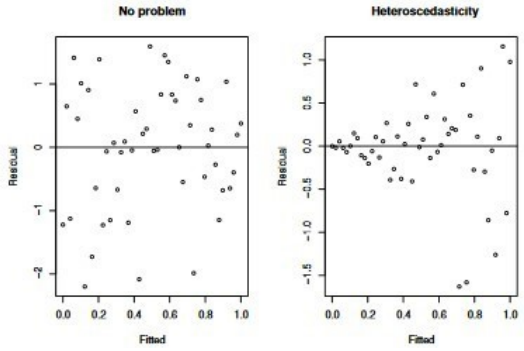
Homoskedastičnost

Nakon konstruisanja linearnog modela treba proveriti da li su ispunjene pretpostavke tog modela koje smo definisali u prethodnom poglavlju. U ovom poglavlju smatraćemo da važi pretpostavka homoskedastičnosti grešaka. Dakle, treba proveriti normalnost, homoskedastičnost i nezavisnost grešaka ε. S obzirom na to da greške nisu obzervabilne, posmatraćemo reziduale e.

Prisetimo se da je $\hat{Y} = X(X^T X)^{-1} X^T Y = HY$, a $e \equiv Y - \hat{Y} = Y - HY = (I_n - H)Y = (I_n - H)X\beta + (I_n - H)\varepsilon$. Pod pretpostavkom da je $Cov(\varepsilon) = \sigma^2 I_n$, važi $Cov(e) = Cov((I_n - H)\varepsilon) = (I_n - H)\sigma^2$. Odatle primećujemo da ukoliko su greške nekorelisane i imaju jednaku disperziju, kod reziduala to ne mora da važi. Međutim, te razlike su uglavnom male tako da proveru pretpostavki grešaka možemo izvršiti na rezidualima i u nastavku rada ćemo smatrati da se sve pretpostavke modela posmatraju na rezidualima.

Jedna od pretpostavki koja je često narušena jeste homoskedastičnost tj. konstantna disperzija grešaka. Da bismo je proverili posmatraćemo grafik reziduala u odnosu na ocenu zavisne promenljive. Ukoliko je homoskedastičnost zadovoljena imaćemo konstantnu raspršenost reziduala po vertikalnoj osi. U slučaju da raspršenost nije konstantna, imamo problem heteroskedastičnosti.

Na sledećoj slici možemo videti kako izgledaju reziduali kod kojih važi (prva slika), a kako oni kod kojih ne važi (druga slika) homoskedastičnost.

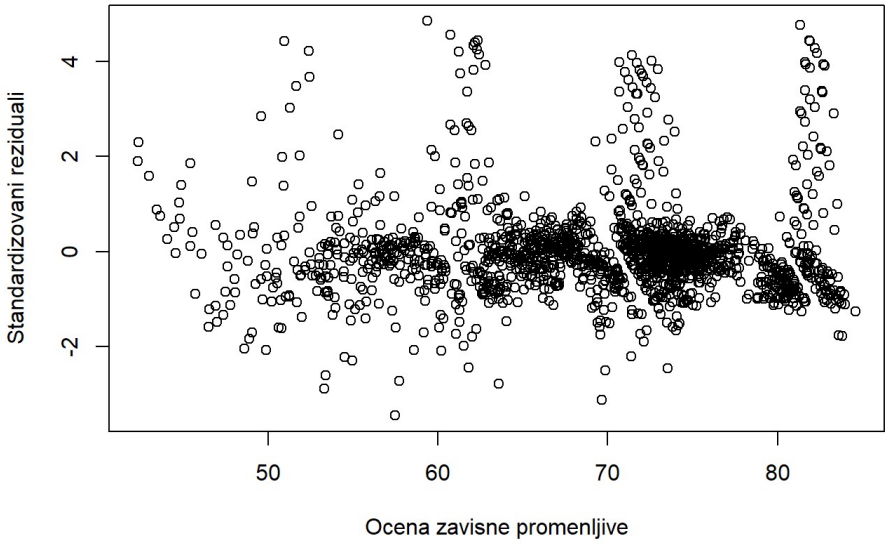


Možemo posmatrati i kvadrate reziduala ili apsolutne vrednosti reziduala u odnosu na ocenu zavisne promenljive i u tim slučajevima ćemo moći bolje da uočimo neku zavisnost, ako ona postoji. Takođe, bolje je posmatrati standardizovane reziduale.

Pokažimo ovo na realnim podacima:

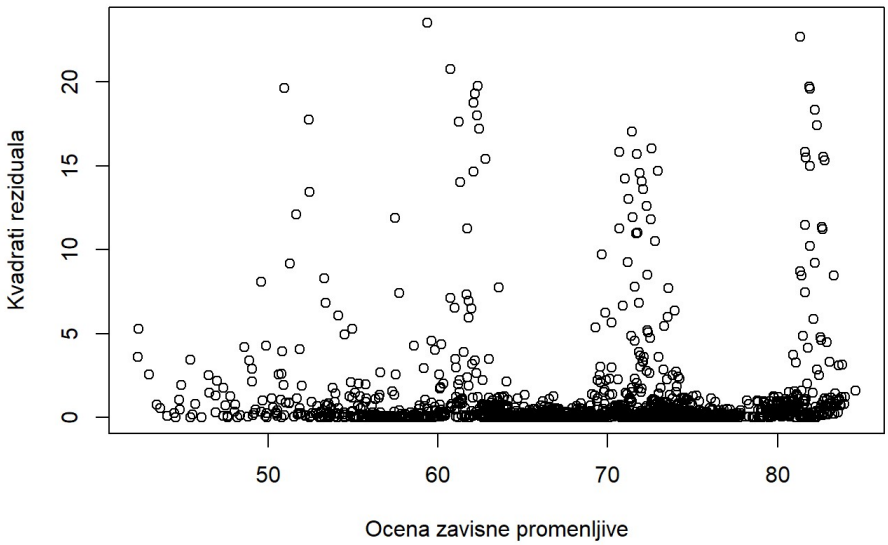
U primeru ćemo koristiti podatke u vezi sa životnim vekom ljudi. Modelujemo prosečan životni vek čoveka u odnosu na sve ostale prediktore iz baze (procenat vakcinisanih za pojedine bolesti, BMI, broj smrtnih slučajeva dece, broj obolelih od određenih bolesti,…) i dobijamo model sa dosta dobrim R^2. Međutim, kada pogledamo grafik standardizovanih reziduala u odnosu na ocenjene vrednosti zavisne promenljive, vidimo da nam uslov homoskedastičnosti nije zadovoljen.

```
model_homosk <- lm(Life.expectancy ~ ., data = podaci)
plot(rstandard(model_homosk) ~ fitted(model_homosk), xlab = "Ocena zavisne promenljive", ylab = "Standardizovani r
eziduali")
```



Podaci se skupljaju i šire i tačno se vidi kako raspršenost nije konstantna.

```
plot(rstandard(model_homosk)^2 ~ fitted(model_homosk), xlab = "Ocena zavisne promenljive", ylab = "Kvadrati rezidu
ala")
```



Kada posmatramo kvadrate reziduala, heteroskedastičnost je još uočljivija.

U ovakvoj situaciji treba da primenimo neku od metoda za rešavanje heteroskedastičnosti.

U daljoj priči transformisaćemo naše početne reziduale i bavićemo se rešavanjem ovog problema pomoću njih.

Novi reziduali i njihova raspodela

Definišimo nove reziduale koje ćemo zvati PCA reziduali (PCA - analiza glavnih komponenti).

Pošto $\hat{\beta}$ prati normalnu raspodelu, primetimo da ce reziduali $e = (I_n - H)Y$ pratiti visedimenzionalnu normalnu raspodelu $N_n(0, (I_n - H)\Sigma)$. Obični reziduali nisu nezavisni, nama je u interesu da formiramo nove reziduale koji će biti.

Pretpostavimo da imamo model za koji važi uslov homoskedastičnosti. Tada će se sopsteveni vektori iz dekompozicije $(I_n - H)\Sigma$ poklapati sa onima iz dekompozicije $(I_n - H)$. Pretpostavimo da je v sopstveni vektor matrice $I_n - H$ kome odgovara sopstvena vrednost 1. Tada je

$$(I_n - H)v = v \Leftrightarrow Hv = 0.$$

Primetimo da v odgovara sopstvenoj vrednosti 1 akko v pripada jezgru od H. Od ranije zamo da je $r(H) = p$, pa iz odnosa $r(H) = n - dim(Ker(H))$ dobijamo $dim(Ker) = n - p$. Dakle, dimenzija prostora sopsvenih vektora koji odgovaraju sopstvenoj vrednosti 1 biće $n - p$.

Razmotrimo slučaj heteroskedastičnosti. $r(I_n - H) = tr(I) - tr(H) = n - p$. Σ je maksimalnog ranga i pošto je $r(I_n - H) = n - p$, važi $r((I_n - H)\Sigma) = n - p$. Odatle je 0 sopstvena vrednost kovarijacione matrice $Cov(\varepsilon) = (I_n - H)\Sigma$ i dimenzija sopstvenog prostora čija je sopstvena vrednost 0 je $dimKer(I_n - H) = p$. Nažalost, pod pretpostavkom heteroskedastičnosti preostale sopstvene vrednosti se mogu razlikovati u parovima (za razliku od slučaja homoskedastičnosti kada su sve preostale sopstvene vrednosti jednake σ^2). Zato imamo sledeću spektralnu dekompoziciju od $(I_n - H)\Sigma$= $(I_n - H)\Sigma = Q\Lambda Q^{-1}$ gde je $\Lambda = diag\{\lambda_1, \ldots, \lambda_{n-p}, 0, \ldots, 0\}$ dimenzije n i Q ortogonalna matrica sopstvenih vektora matrice $(I_n - H)\Sigma$.

Definišimo najzad PCA reziduale sa $R = Qe$, gde je Q prethodno definisana matrica. Pošto reziduali imaju $N_n(0, (I_n - H)\Sigma)$, odnosno $N(0, Q\Lambda Q^{-1})$ raspodelu, tada će iz osobina višedimenzionalne normalne raspodele Qe imati $N_n(0, \lambda)$ raspodelu. Za $R = (R_1, \ldots, R_n)^T$, važi da su PCA reziduali R_i i R_j nezavisni za $i \neq j$ i R_i imaju $N_n(0, \lambda_i)$ raspodelu. Takođe, $R_n, R_{n-1}, \ldots, R_{n-p+1}$ su jednaki 0 jer predstavljaju slučajne veličine sa očekivanjem i disperzijom 0.

Dakle, reziduali treba da budu takvi da je poslednjih p jednako nuli, a preostalih $n - p$ da su nezavisni i normalno raspodeljeni.

Pod pretpostavkom homoskedastičnosti, reziduali $R_1, \ldots, R_{n-p}$ će biti nezavisni, sa $N(0, \sigma^2)$ raspodelom. Preostali će biti jednaki 0. Dakle, reziduali $R_1, \ldots, R_{n-p}$ će biti nezavisni i jednako raspodeljeni i kao takvi, vrlo su pogodni za primenu uobičajenih testova normalnosti, a i preciznost često upotrebljivih QQ plotova biće značajnije poboljšana.

Pri uslovu homoskedastičnosti, pošto nam je σ^2 nepoznato, treba da odredimo ocenu za njega. Iz jakog zakona velikih brojeva zaključujemo da je ocena

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n-p} R_i^2$$

postojana.

Dalje, definišimo standardizovane PCA reziduale pri uslovu homoskedastičnosti. Neka je

$$\hat{\sigma}_i^2 = \frac{1}{n-p-1} \sum_{\substack{j=1 \\ i \neq j}}^{n-p} R_j^2.$$

R_i i $\hat{\sigma}_i^2$ su nezavisne.

Za svako $i = 1, \ldots, n-p$ važi $(n-p-1)\hat{\sigma}_i^2/\sigma^2 \sim \chi^2_{n-p-1}$, dok $R_i/\sigma \sim N(0,1)$ pa su standardizovani PCA reziduali jednaki

$$R_i^* = \frac{R_i}{\hat{\sigma}_i} = \frac{R_i/\sigma}{\hat{\sigma}_i/\sigma} \sim t_{n-p-1},$$

gde je $\hat{\sigma}_i = \sqrt{\hat{\sigma}_i^2}$ i t_{n-p-1} Studentova t raspodela sa $n-p-1$ stepenom slobode.

Dakle, umesto da crtamo QQ plot PCA reziduala u odnosu na teorijske kvantile $N(0, \hat{\sigma}^2)$ raspodele, crtaćemo QQ plot standardizovanih PCA reziduala, R_i^*, u odnosu na teorijske kvantile t_{n-p-1} raspodele.

Na sličan način možemo definisati standardizovane PCA reziduale pri uslovu heteroskedastičnosti. Primetimo da za

$$(I_n - H)\Sigma = Q\Lambda Q^{-1}$$

važi da su reziduali R_i nezavisni i $R_i \sim N(0, \lambda_i)$. Odatle njihov standardizovan oblik možemo definisati sa

$$R_i^* = \frac{R_i}{\sqrt{\lambda_i}}, \;\; i = 1, \ldots, n-p.$$

Pošto su nam λ_i nepoznati moramo ih oceniti.

Nasuprot homeskedastičnom slučaju gde se spektralne dekompozicije matrica $(I_n - H)$ i $(I_n - H)\Sigma$ poklapaju, ovde to nije tako i moramo oceniti matricu kovarijacije koristeći takozvanu “heteroskedastično-konzistentnu ocenu kovariacione matrice”. To je konzistentna ocena $Cov(\hat{\beta})$ pod pretpostavkama modela (heteroskedastičnost).

Definišimo

$$\hat{\Sigma}_i = E_i \hat{\Sigma},$$

za $i = 0, 1, 2, 3, 4$, gde je $\hat{\Sigma} = diag\{e_1^2, \ldots, e_n^2\}$, a E_i predstavljaju:

$$E_0 = I_n, \quad E_1 = \frac{n}{n-p} I_n, \quad E_2 = diag\{1/(1-h_i)\}, \quad E_3 = diag\{1/(1-h_i)^2\}, \quad E_4 = diag\{1/(1-h_i)^{\delta_i}\}, \;\; \delta_i = min\{4, nh_i/p\}, \;\; i = 1, .$$

Naposletku, slično kao i ranije, radimo dekompoziciju $(I_n - H)\Sigma_i$ birajući željeno i iz $\{0, \ldots 4\}$, odnosno :

$$(I_n - H)\hat{\Sigma}_i = Q_i \hat{\Lambda}_i Q_i^T$$

i dobijamo odgovarajuće PCA reziduale:

$$R^{(i)} = Q_i e.$$

Za standardizovanje biće nam potrebne sopstvene vrednosti iz matrice $\hat{\Lambda}_i$.

Ilustrujmo prethodno narednim primerima.

Primena na realnim podacima

Sada ćemo na realnim podacima prikazati primenu PCA reziduala.

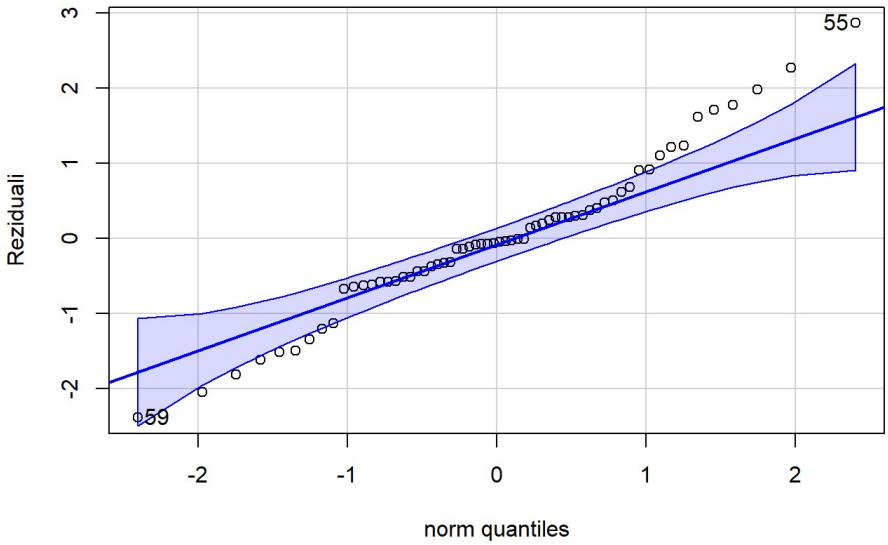
Koristimo bazu cars koja sadrži podatke o dužini kočenja automobila do zaustavljanja i brzini kojom se kretao. Želimo da modeliramo zavisnost dužine kočenja od brzine.

Pravimo model zavisnosti dužine kočenja od brzine i kvadrata brzine.

```
model <- lm(dist ~ speed + I(speed^2) ,data=cars)
```

Prvo što ćemo proveriti jeste normalnost običnih standardizovanih reziduala uz pomoć QQ plota.

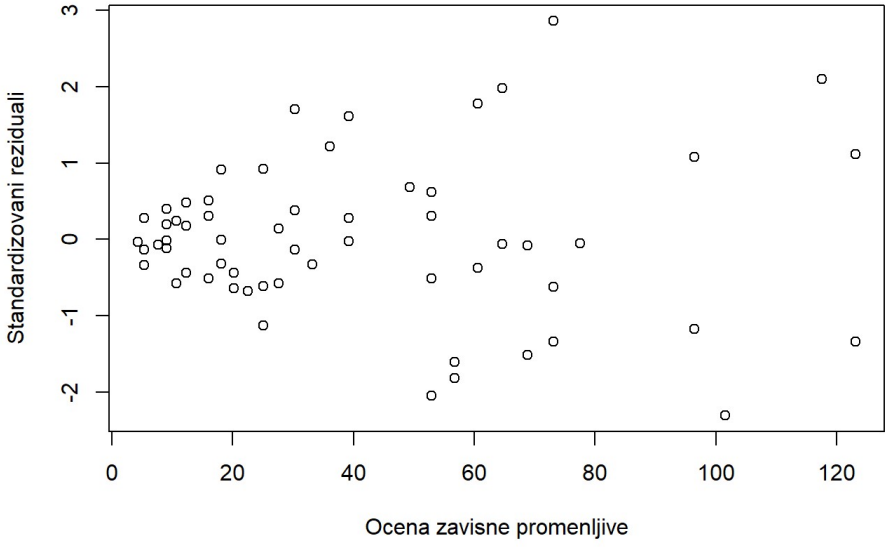
```
qqPlot(rstandard(model), ylab = "Reziduali")
```



Sa grafika primećujemo da oni otprilike prate normalnu raspodelu.

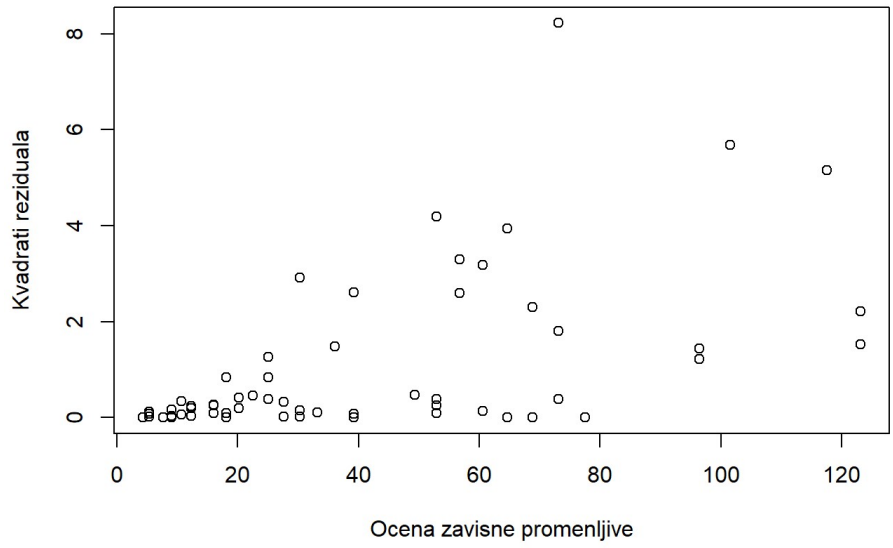
Zatim proveravamo uslov homoskedastičnosti uz pomoć običnih standardizovanih reziduala i sa grafika primećujemo da se njihova raspršenost povećava sa porastom vrednosti ocenjene zavisne promenljive.

```
plot(fitted(model), scale(residuals(model)), ylab = "Standardizovani reziduali", xlab = "Ocena zavisne promenljiv
e")
```



To se još bolje vidi na sledećem grafiku gde posmatramo kvadrate standardizovanih reziduala u odnosu na ocenu zavisne promenljive i ne njemu vidimo još veću razliku u raspršenosti sa porastom vrednosti ocenjene zavisne promenljive.

```
plot(rstandard(model)^2 ~ fitted(model), xlab = "Ocena zavisne promenljive", ylab = "Kvadrati reziduala")
```

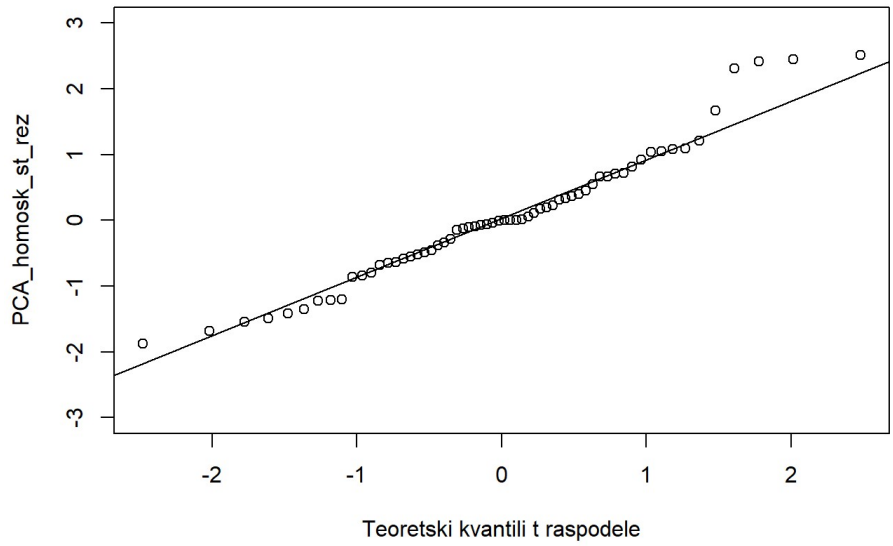


Kako bismo bili sigurniji u zaključke do kojih smo došli, koristićemo PCA reziduale pod pretpostavkom homoskedastičnosti za proveravanje ova dva uslova. PCA reziduale konstruišemo po teoriji i ukoliko proverimo njihove vrednosti možemo videti da su poslednja 3 jednaka nuli, baš kao što smo i očekivali.

```
## [1] -1.332268e-14
## [1] 1.199041e-14
## [1] 8.881784e-16
```

Sledeći korak je, naravno, njihova standardizacija,a zatim proveravamo da li oni prate Studentovu raspodelu sa n-4 stepena slobode uz pomoć QQ plota.

```
qqplot(qt(ppoints(PCA_homosk_st_rez), df = n-(qr(X)$rank)-1), PCA_homosk_st_rez ,
       xlab = "Teoretski kvantili t raspodele", ylim=c(-3,3))
qqline(PCA_homosk_st_rez)
```



Vidimo da prate, a iz toga zaključujemo da je pretpostavka o normalnosti reziduala zadovoljena.

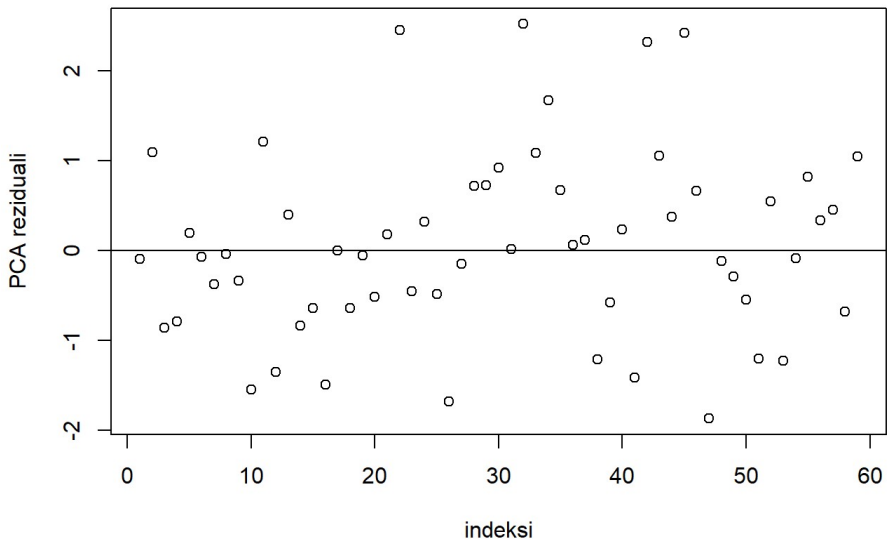
Proverićemo pretpostavku da je očekivana vrednost reziduala 0. To radimo uz pomoć 95% intervala poverenja. Dobijamo da 0 pripada tom intervalu poverenja i samim tim smatramo da je ta pretpostavka zadovoljena.

```
confidence_interval <- function(vector, interval) {
  vec_sd <- sd(vector)

  n <- length(vector)

  vec_mean <- mean(vector)

  error <- qt((interval + 1)/2, df = n-(qr(X)$rank)-1) * vec_sd / sqrt(n)

  result <- c("lower" = vec_mean - error, "upper" = vec_mean + error)
  return(result)
}

confidence_interval(PCA_homosk_st_rez, 0.95)
```

```
##      lower      upper
## -0.2093277  0.2996056
```

Grafičkim prikazom vidimo da PCA reziduali konstruisani pod uslovom homoskedastičnosti izgledaju dosta lepo i odatle ne možemo odmah zaključiti da je uslov homoskedastičnosti ispunjen.

```
plot(c(1:59),PCA_homosk_st_rez[1:59], xlab = "indeksi", ylab = "PCA reziduali")
abline(h=0)
```

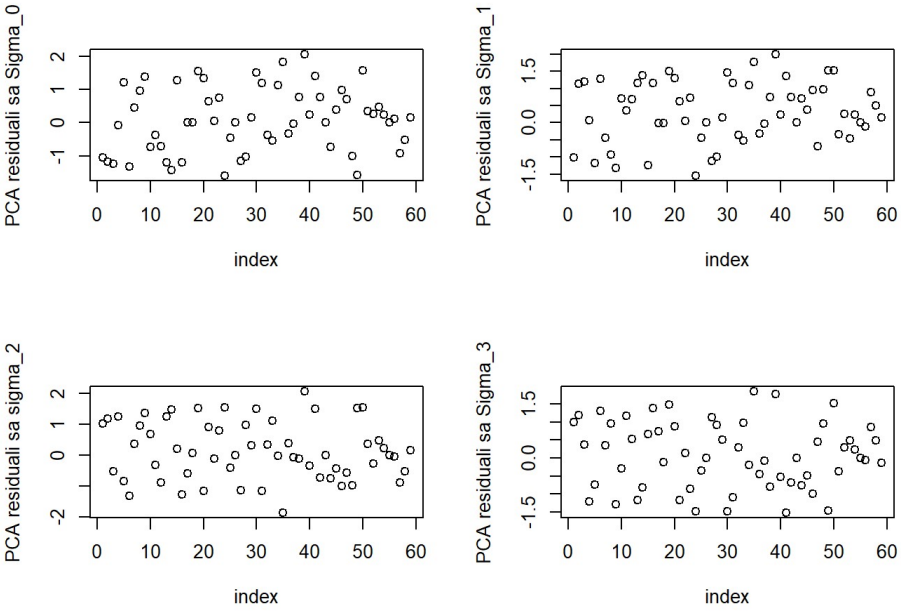


I dalje sumnjamo na to da homoskedastičnost nije zadovoljena i zato prelazimo na konstruisanje PCA reziduala pod uslovom heteroskedastičnosti i vršimo dalju analizu uz pomoć njih.

Nove PCA reziduale konstruišemo opet po definiciji, množeći stare reziduale sa transponovanim matricama sopstvenih vektora matrica Q_i (koje smo konstruisali opet po definiciji uz pomoć matrica E_i, H, Σ). Imamo različite ocene za kovarijacionu matricu i zato konstruišemo različite grupe PCA reziduala. Nove reziduale standardizujemo i spremni smo da ih grafički analiziramo.

Pošto imamo različite mogućnosti za ocenu kovarijacione matrice, crtamo grafike novih reziduala u svim slučajevima.

```
plot(c(1:59), PCA_heter_st_rez$R0[1:59], xlab="index", ylab="PCA residuali sa Sigma_0")
plot(c(1:59), PCA_heter_st_rez$R1[1:59], xlab="index", ylab="PCA residuali sa Sigma_1")
plot(c(1:59), PCA_heter_st_rez$R2[1:59], xlab="index", ylab="PCA residuali sa sigma_2")
plot(c(1:59), PCA_heter_st_rez$R3[1:59], xlab="index", ylab="PCA residuali sa Sigma_3")
```



Vidimo sa sva 4 grafika da se reziduali lepo ponašaju, da je njihova raspršenost ujednačena. S obzirom na to da smo pošli od pretpostavke heteroskedastičnosti i napravili PCA reziduale za taj slučaj, iz ovako ujednačene raspršenosti možemo da zaključimo da naši početni reziduali baš prate tu heteroskedastičnost od koje smo pošli. Ovom analizom smo dodatno potvrdili našu sumnju da su nam originalni reziduali heteroskedastični.

Odradićemo još jedan primer.

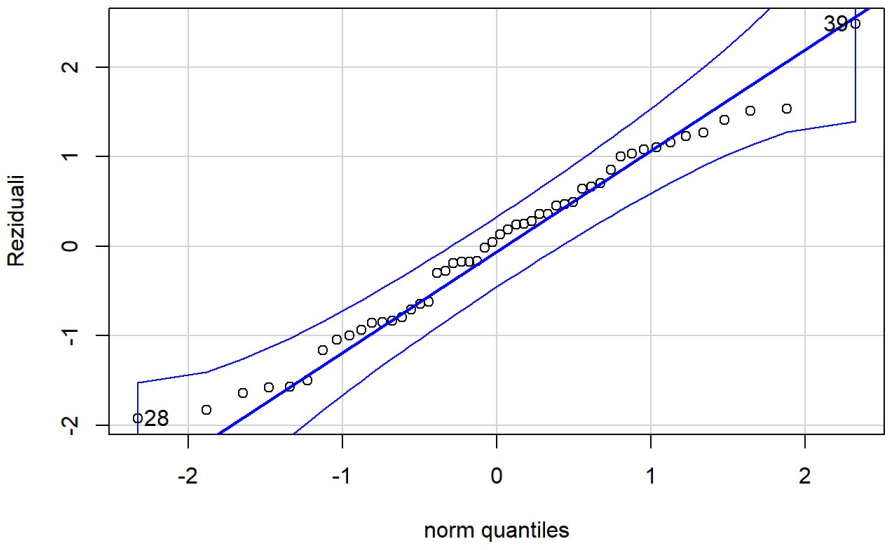
Imamo podatke u kojima posmatramo academic performance index u osnovnim skolama Kalifornije u odnosu na različite prediktore. Modeliramo API u odnosu na broj upisanih đaka, procenat đaka koji dobijaju besplatne obroke, procenat profesora sa potpunom akreditacijom i kvadrat poslednjeg prediktora.

Imamo dobar model, barem posmatrajući R^2. Sada treba proveriti pretpostavke.

```
model <- lm(api00 ~ meals + full + enroll + I(full^2) , data=podaci)
```

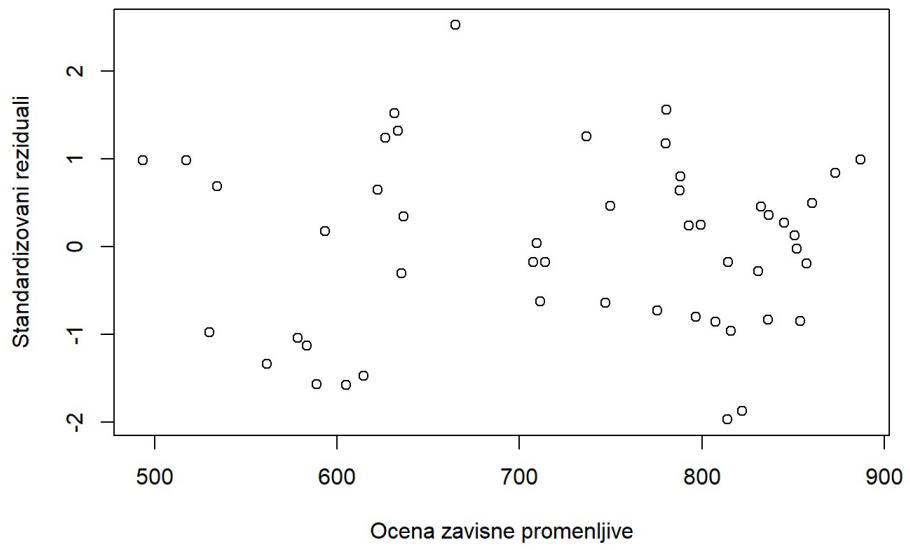
Crtamo QQ plot standardizovanih reziduala i vidimo da oni otprilike prate normalnu raspodelu.

```
qqPlot(rstandard(model), ylab = "Reziduali")
```



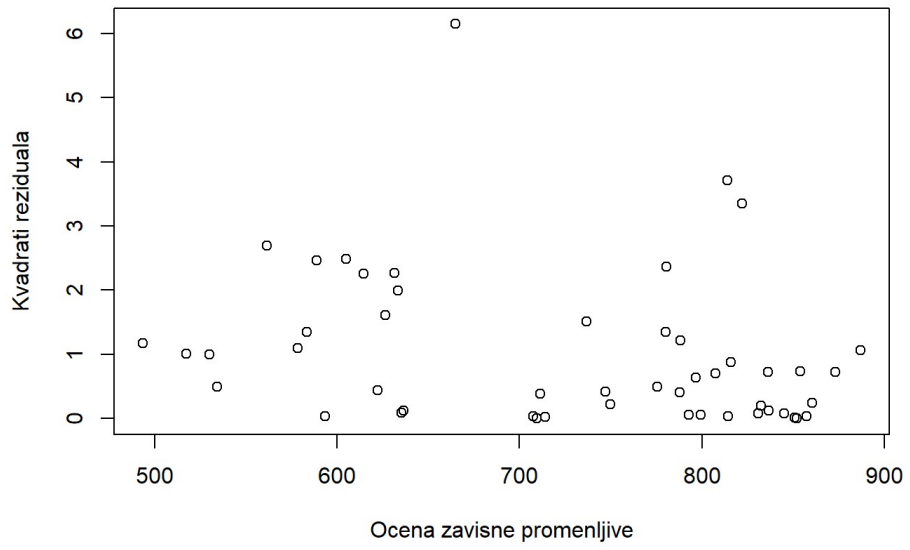
Zatim proveravamo uslov homoskedastičnosti uz pomoć običnih standardizovanih reziduala i primećujemo da taj uslov ne deluje baš ispunjeno.

```
plot(fitted(model), scale(residuals(model)), ylab = "Standardizovani reziduali", xlab = "Ocena zavisne promenljiv
e")
```



Crtajući grafik kvadrata standardizovanih reziduala primećujemo jos očiglednije razliku u raspršenosti.

```
plot(rstandard(model)^2 ~ fitted(model), xlab = "Ocena zavisne promenljive", ylab = "Kvadrati reziduala")
```

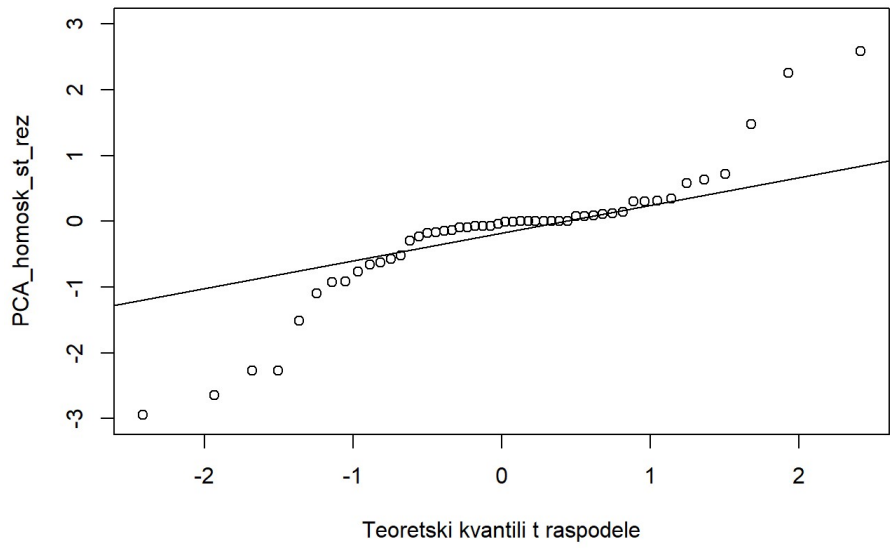


I u ovom primeru kako bismo bili sigurniji u zaključke do kojih smo došli koristićemo PCA reziduale pod pretpostavkom homoskedastičnosti za proveravanje ova dva uslova. PCA reziduale opet konstruišemo po teoriji i ukoliko proverimo njihove vrednosti možemo videti da je poslednjih 5 jednako nuli.

```
## [1] 1.323919e-11
## [1] -8.071321e-13
## [1] -5.912104e-12
## [1] -1.509903e-14
## [1] -5.859202e-14
```

Zatim ih standardizujemo i proveravamo da li oni prate Studentovu raspodelu sa n-6 stepeni slobode uz pomoć QQ plota. Primećujemo da PCA reziduali i ne prate bas studentovu raspodelu, sugerisu vrlo debeo rep.

```
qqplot(qt(ppoints(PCA_homosk_st_rez), df = n-(qr(X)$rank)-1), PCA_homosk_st_rez ,
       xlab = "Teoretski kvantili t raspodele", ylim=c(-3,3))
qqline(PCA_homosk_st_rez)
```

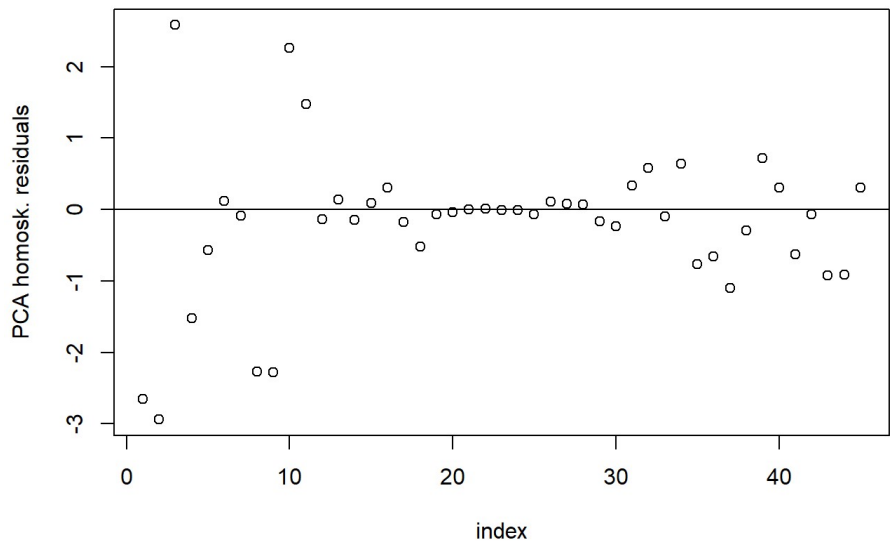


Proveravamo pretpostavku da je očekivana vrednost reziduala 0 uz pomoć 95% intervala poverenja. Dobijamo da 0 pripada tom intervalu poverenja i samim tim smatramo da je ta pretpostavka zadovoljena.

```
##       lower       upper
## -0.46426196  0.09707338
```

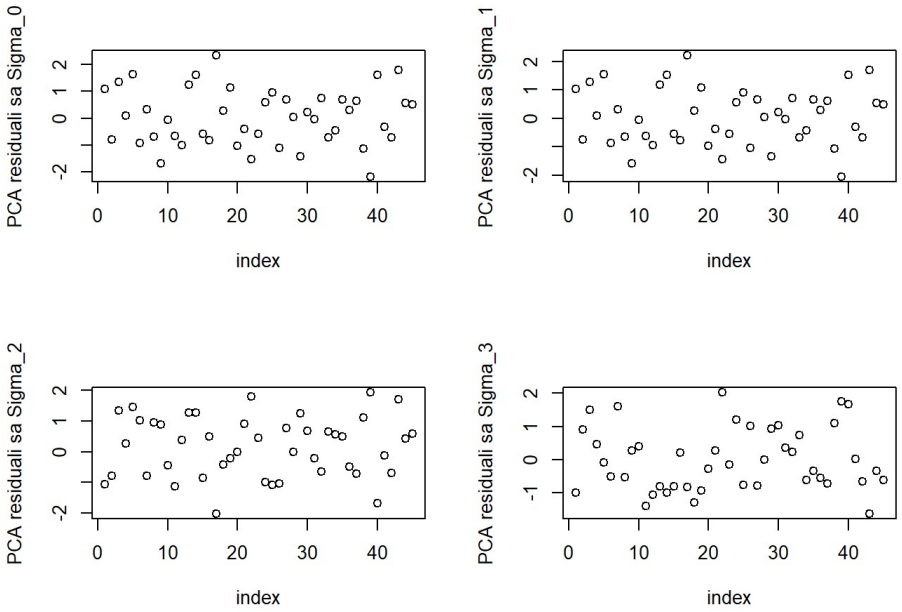
Grafičkim prikazom vidimo da PCA reziduali konstruisani pod uslovom homoskedastičnosti čvrsto potvrđuju da je model heteroskedastičan.

```
plot(c(1:45),PCA_homosk_st_rez[1:45],ylab = "PCA homosk. residuals",xlab = "index")
abline(h=0)
```



Želimo da vidimo i kakve zaključke možemo doneti iz PCA reziduala koji su konstruisani pod uslovom heteroskedastičnosti i kao i u prethodnom primeru konstruišemo ih po definiciji i vršimo analizu homoskedastičnosti za različite ocene kovarijacione matrice.

```
plot(c(1:45), PCA_heter_st_rez$R0[1:45], xlab = "index", ylab = "PCA residuali sa Sigma_0")
plot(c(1:45), PCA_heter_st_rez$R1[1:45], xlab = "index", ylab = "PCA residuali sa Sigma_1")
plot(c(1:45), PCA_heter_st_rez$R2[1:45], xlab = "index", ylab = "PCA residuali sa Sigma_2")
plot(c(1:45), PCA_heter_st_rez$R3[1:45], xlab = "index", ylab = "PCA residuali sa Sigma_3")
```



Vidimo da prva 3 grafika vrlo lepo deluju pod pretpostavkom heteroskedastičnosti, dok je četvrti nije baš toliko ujednačeno raspršen, pa za ocenu treba odabrati neku od prve 3. Tim odabirom, dolazimo do zaključka da su reziduali našeg modela zaista heteroskedastični jer su grafici PCA reziduala pod uslovom heteroskedastičnosti lepo raspršeni.

Zaključak

U ovom radu smo predstavili nove reziduale koji su korisni pri susretu sa heteroskedastičnim modelima. Formiraju se linearnom transformacijom običnih reziduala i spektralnom dekompozicijom ocene kovaracione matrice. Rezultirajući reziduali su nezavisni i normalno raspodeljeni

PCA reziduali imaju značajnu prednost zato što su nezavisni, shodno tome pomoću njih možemo lako proveriti pretpostavke modela. Takođe, lako se računaju i primenjuju u različitim statističkim metodama.Uopšte, sam problem dobijanja nezavisnih reziduala nije nalazio puno uspeha. Mana PCA reziduala je što pri njihovoj formaciji moramo izgubiti “neke” reziduale tj njih p. To je donekle tačno, jer su linearna transformacija običnih reziduala pa nam je njihova informacija i dalje dostupna.

Da bismo prikazali korisnost naših reziduala obradili smo dva primera. U prvom primeru smo prikazali kako uz pomoć PCA reziduala možemo potvrditi naše sumnje u to da nam uslov homoskedastičnosti nije ispunjen, a drugi primer nam je još interesantniji. Pomoću njega smo videli korisnost PCA reziduala u otkrivanju neispunjenosti pretpostavke o normalnosti reziduala iako nam se činilo da je ona zadovoljena pri posmatranju običnih reziduala. Takođe smo tim primerom videli koliko je lakše primetiti heteroskedastičnost uz pomoć novokonstruisanih PCA reziduala i time olakšati sebi put ka pravljenju kvalitetnih modela linearne regresije.

Literatura:

- *Improved residuals for linear regression models under heteroskedasticity of unknown form* (https://arxiv.org/pdf/1607.07926.pdf (https://arxiv.org/pdf/1607.07926.pdf))
- *Linear models with R - Julian J. Faraway*