Heteroskedastičnost u regresijskoj analizi

Matej Petrinović, Lucija Rupčić

1 Uvod

Problem heteroskedastičnosti se javlja kada imamo rasipanje na dijagramu raspršenosti. U regresijskoj analizi govorimo o heteroskedastičnosti u kontekstu reziduala ili pojma pogreške. Naime, heteroskedastičnost je sustavna promjena u širenju reziduala u rasponu izmjerenih vrijednosti. Heteroskedastičnost je problem jer regresija, čiji su koeficijenti dobiveni metodom najmanjih kvadrata, pretpostavlja da su svi reziduali izvučeni iz populacije koja ima konstantnu varijancu (homoskedastičnost).

Da bi se zadovoljile pretpostavke regresije i da bi se moglo pouzdati u rezultate, reziduali bi trebali imati konstantu varijaciju. U ovom seminaru, pokazati ćemo kako prepoznati heteroskedastičnost, objasniti što proizvodi, probleme koje uzrokuje i kroz primjer kako bi pokazali rješenje tog problema. Prije svega definirajmo pojam heteroskedastičnosti.

Definicija 1
Neka je $Y = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon$ linearni regresijski model. Kažemo da je model *heteroskedastičan* ako vrijedi

$$E(\varepsilon^2|\mathbf{X}) = \sigma^2(\mathbf{X})$$

tj. ako je varijanca greške modela fukcija regresora $\mathbf{X}$.

Uobičajeno je provjeriti heteroskedastičnost reziduala nakon što izgradimo model linearne regresije. Razlog tome je što želimo provjeriti je li model tako izrađen može objasniti neki uzorak u varijabli odgovora Y, koji se na kraju pojavljuje u rezidualima. To bi rezultiralo neučinkovitim i nestabilnim regresijskim modelom koji bi kasnije mogao donijeti loša predviđanja. U prisutnosti heteroskedastičnosti modela postoje dvije glavne posljedice za procjenitelje dobivene metodom najmanjih kvadrata

1. Procjena metodom najmanjih kvadrata je još uvijek linearna i nepristrana procjena, ali ona više nije najbolja. To jest, postoji još jedan procjenitelj s manjom varijancom.

2. Standardne pogreške izračunate za procjenu metodom najmanjih kvadrata su netočne. To može utjecati na intervale pouzdanosti i testiranje hipoteza koje koriste te standardne pogreške, što može dovesti do pogrešnih zaključaka

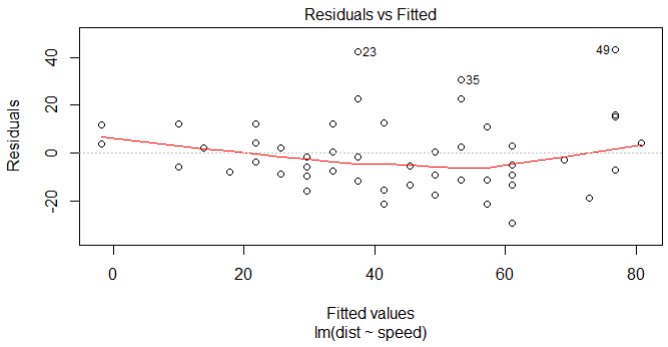
2 Detekcija heteroskedatičnosti

Jedan neformalni način detekcije heteroskedastičnosti je crtanjem dijagrama raspršenja reziduala gdje su oni prikazani u odnosu na teorijske vrijednosti dobivene modelom. Ako postoji očigledan uzorak na slici, tada je prisutna heteroskedastičnost. Ovo ćemo ilustrirati stvarnim regresijskim modelom temeljenim na bazi podataka o automobilima (cars), koja je ugrađena u programski jezik R.

Baza podataka cars sadrži podatke od 50 brzina automobila (mph) i pripadnu udaljenost za zaustavljanje pri toj brzini (ft). Napravit ćemo model jednostavne linearne regresije u kojem ćemo modelirati udaljenost potrebnu za zaustavaljanje (Y) sa brzinom automobila (X). Dobiveni model glasi

$$Y = -17.5791 + 3.9324\,X + \varepsilon$$

Sada kada je model spreman, pogledajmo dijagram raspršenja teorijskih vrijednosti i reziduala.



U našem slučaju, kao što možemo primijetiti na grafičkom prikazu, crvena linija je blago zakrivljena, a reziduali se povećavaju kako se povećava vrijednost Y. Dakle, zaključak ovdje je da postoji heteroskedastičnost. Heteroskedastičnost možemo provesti statističkim testom tako da testiramo hipotezu o (ne)konstantnosi varijance. Najčešće korišteni testovi za testiranje takvih hipoteza su Breush-Paganov test i NCV test. Testiranjem hipoteza za ovaj model dobili smo sljedeće p-vrijednosti

Test	$p-$vrijednost
Breush-Pagan	0.07297
NCV	0.031049

Vidimo kako je NCV test u ovom slučaju odbacio hipotezu o konstantnosti varijance, što potvrđuje naš zaključak sa grafičkog prikaza.

3 Uzrok heteroskedastinosti

Heteroskedastičnost, pojavljuje se često u bazama podataka koji imaju veliki raspon između najvećih i najmanjih promatranih vrijednosti. Iako postoje brojni razlozi zašto heteroskedastičnost može postojati, uobičajeno objašnjenje je da se varijacija pogreške mijenja proporcionalno s regresorima. Ovaj faktor može biti varijabla u modelu.

U nekim slučajevima varijance se povećavaju proporcionalno tom faktoru, ali ostaju konstantne kao postotak. Na primjer, promjena od 10% u broju, kao što je 100, mnogo je manja od 10% promjene u velikom broju, kao što je 100.000. Zato treba biti oprezni kada radimo sa širokim rasponom vrijednosti! Budući da su veliki problemi povezani s ovim problemom, neki tipovi modela su skloniji heteroscedastičnosti.

4 Tipovi heteroskedastičnosti

Heteroskedastičnost općenito možemo kategorizirati u dvije kategorije i to na način

1. *Čista heteroskedastičnost* odnosi se na slučajeve u kojima određujemo ispravan model, a ipak promatramo nekonstantnu varijancu.

2. *Nečista heteroskedastičnost* odnosi se na slučajeve u kojima smo pogrešno specificirali model, a to uzrokuje nekonstantnu varijancu. Kada iz modela ostavimo važnu varijablu, izostavljeni efekt se apsorbira u pojam pogreške. Ako se učinak izostavljene varijable mijenja kroz opaženi raspon podataka, on može proizvesti signalne znakove heteroskedastičnosti.

Kada promatramo heteroskedastičnost u grafičkom prikazu reziduala, važno je odrediti imamo li čistu ili nečistu heteroskedastičnost jer su rješenja različita. Ako imamo nečisti oblik, moramo identificirati važne varijable koje su izostavljene iz modela i prepraviti model s tim varijablama. Nadalje ćemo govoriti o čistom obliku heteroskedastičnosti.

5 Problemi koje heteroskedastičnost uzrokuje

Linearna regresija pretpostavlja da je širenje reziduala konstantno po cijelom dijagramu. Kad god kršimo ovu pretpostavku, postoji mogućnost da ne možemo vjerovati nekim statističkim rezultatima.

Zašto rješavati ovaj problem? Dva su velika razloga zašto želimo homoskedastičnost:

1. Iako heteroskedastičnost ne uzrokuje pristranost u procjenama koeficijenata, ona ih dakako čini manje preciznima. Niža preciznost povećava vjerojatnost da procjene koeficijenata više odstupaju od njihove ispravne vrijednosti

2. Heteroskedastičnost će proizvesti p-vrijednosti koje su manje nego što bi trebale biti. Taj se učinak događa zbog toga što heteroskedastičnost povećava varijancu procjena koeficijenata, ali MNK ne otkriva ovo povećanje. Slijedom toga, MNK izračunava t-vrijednosti i F-vrijednosti koristeći podcijenjenu količinu varijance. Ovaj problem može nas navesti na zaključak da su koeficijenti modela statistički značajni kada zapravo nisu.

6 Rješavanje problema heteroskedastičnosti

Ako možemo otkriti razlog heteroskedastičnosti, možda ćemo ga moći ispraviti i poboljšati svoj model. Dva su najčešća načina za rješavanje problema heteroskedastičnosti.

1. Težinska metoda najmanjih kvadrata
2. Transformacija ovisne varijable

6.1 Težinska metoda najmanjih kvadrata

Kod obične metode najmanjih kvadrata cilj nam je bio minimizirati sumu kvadrata grešaka, odnosno

$$SSE_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \to \min_{\boldsymbol{\beta}}$$

Kao rješenje tog minimizacijskog problema dobili smo LS procjenitelj za $\boldsymbol{\beta}$, tj.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\operatorname{argmin}} \, SSE_n(\boldsymbol{\beta})$$

U ovom slučaju minimizirati ćemo težinsku sumu kvadrata grešaka, tj minimizirati ćemo izraz oblika

$$WSSE_n(\boldsymbol{\beta}, \mathbf{w}) = \sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

Uočimo kako je za $w_i = 1, \forall\, i = 1, \ldots, n$ obična MNK zapravo specijalan slučaj težinske MNK. Iz Gauss-Markovog teorema znamo da je najbolji nepristran linearan procjenitelj za $\boldsymbol{\beta}$, u uvjetima heteroskedastičnosti, dan izrazom

$$\hat{\boldsymbol{\beta}}_D = \left(\mathbf{X}^T D^{-1} \mathbf{X}\right)^{-1} \mathbf{X} D^{-1} \mathbf{Y}$$

pri čemu je $D = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Intuitivan način je pronaći "Cholesky korijen" matricu P tako je $\sigma^2 D^{-1} = P^T P$. To nam daje model oblika $PY = P\mathbf{X}^T\boldsymbol{\beta} + P\varepsilon$. U uvjetima heteroskedastičnosti matrica P je oblika $P = \text{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_n}\right)$. Množeći $Y, \mathbf{X}, \varepsilon$ matricom P svaku observaciju transformiramo djeljnjenom sa σ_i^2. Uočimo sada kako je

$$E\left[\frac{\varepsilon_i^2}{\sigma_i^2} \,\middle|\, \mathbf{X}_i\right] = \frac{1}{\sigma_i^2} E[\varepsilon_i^2 \mid \mathbf{X}_i] = \frac{1}{\sigma_i^2}\sigma_i^2 = 1$$

Uočimo kako smo ovdje dobili konstantu varijancu iznosa 1, te smo time eliminirali heteroskedastičnost.

6.1.1 Procjena σ_i^2

Sljedeći postupak daje nam kako procjeniti težine, odnosno σ_i^2, koje ćemo koristiti za težinsku metodu najmanjih kvadrata.

1. Izračunati reziduale e_i i teorijske vrijednosti dobivene modelom f_i, $i = 1, \ldots, n$, koji je dobiven netežinskom MNK
2. Odrediti $|e_i|$, $i = 1, \ldots, n$, tj. apsolutnu vrijednost reziduala
3. Napraviti regresijski model za apsolutne vrijednosti reziduala na temelju teorijski vrijednosti f_i
4. Za tako dobiven model odrediti njegove teorijske vrijednosti F_i
5. Tako dobive vrijednosti F_i^2 su procjene za σ_i^2

Koristeći ovaj postupak i činjenicu da za težine w_i uzimamo $w_i = \frac{1}{\sigma_i^2}$, umjesto σ_i^2 ćemo za model koristiti F_i^2.

6.2 Transformacija ovisne varijable

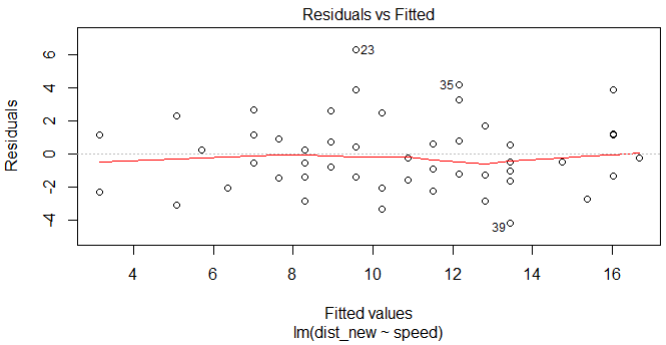
U svrhu rješavanja problema heteroskedastičnosti, ovisnu varijablu Y možemo prikadno transformirati. Najčešće korištena trasformacija ovisne varijabe je Box-Cox transformacija. Box-cox transformacija je matematička transformacija varijable kako bi on bila približna normalnoj distribuciji. Često, transformiranje Y varijable u Box-Cox transformacijom rješava problem. To je transformacija sljedećeg oblika:

$$g_\lambda(y) = \begin{cases} \frac{y^{\lambda-1}}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

Ilustrirajmo to na prethodnom primjeru modela

$$Y = -17.5791 + 3.9324\, X + \varepsilon$$

gdje modeliramo udaljenost potrebnu za zaustavljnje automobila pri određenoj brzini. Procjenjeni λ za ovu transformaciju iznosi $\lambda = 0.5$. Transformirani podaci za naš novi regresijski model su spremni. Kreirajmo model i provjerimo heteroskedastičnost. Ponovno pogledajmo grafički prikaz reziduala i teorijskih vrijednosti.



Vidimo kako je ovdje linija skoro pa pravac, što bi nam moglo sugerirati da smo riješili problem heteroskedastičnosti. Pogledajmo još $p-$vrijednosti statističkih testova.

Test	$p-$vrijednost
Breush-Pagan	0.9157
NCV	0.91258

Vidimo kako su oba testa dali visoke $p-$vrijednosti. Zaista ovom transforamcijom smo eliminirali problem heteroskedastičnosti, te sveli naš model na homoskedastičan.

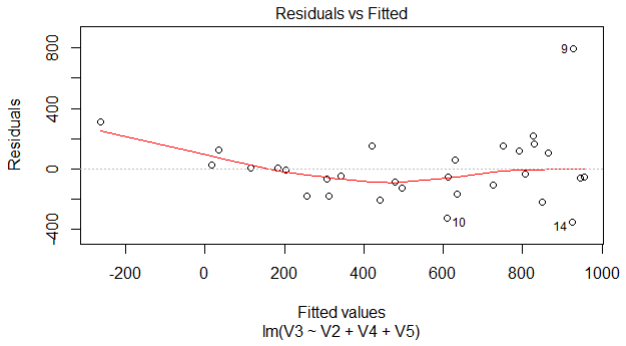
7 Primjene metoda na bazi podataka Birth Rates and Economic Development

Baza podataka Birth Rates and Economic Development sadrži podatke o stopi nataliteta, BDP-u, proporciji stanovnika koji žive na farmama i stopi mortaliteta za 30 zemalja svijeta. Podaci su nastali ranih $50-$tih godina prošlog stoljeća. Za početak napravimo model kojim ćemo modelirati BDP u odnosu na ostale navedene varijable. Dobiveni model je sljedeći:

$$Y = 1087.438 + 6.223\, X_1 - 1113.488\, X_2 - 6.587\, X_3 + \varepsilon$$

pri čemu su: Y - BDP, X_1 - stopa nataliteta, X_2 - proporcija stanovnika na farmama i X_3 - stopa mortaliteta. Ovako dobiven model ima $R^2 = 0.6612$, pa bi mogli reći da ovaj model dobro opisuje podatke. Pogledajmo što je sa homoskedastičnosti modela grafičkim

prikazom reziduala i teorijskih vrijednosti i testirajmo pripadnim testovima.



Uočimo kako ovdje crvena krivulja u početku jako zakrivljena. To bi nam moglo sugerirati da imamo problem sa heteroskedastičnošću. Pogledajmo $p-$ vrijednosti pripadnih testova.

Test	$p-$vrijednost
Breush-Pagan	0.6145
NCV	0.024723

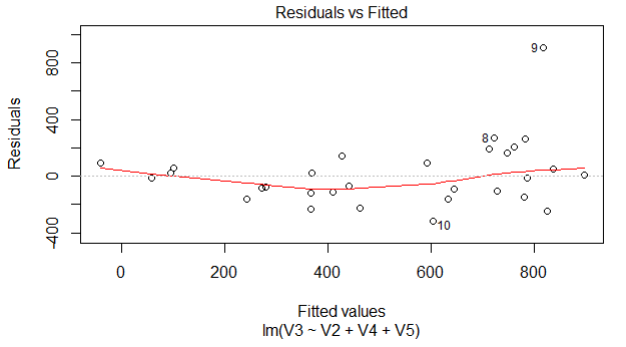
Vidimo kako je zapravo ncvTest odbacio hipotezu o homoskedastičnosti modela. Primjenimo prethodno opisane metode na ovaj model u svrhu rješavanja problema heterskedastičnosti.

7.1 Težinska MNK

Primjenimo ovdje opisani postupak za traženje težina i pogledajmo tako dobiveni model. Model dobiven ovakvim postupkom je

$$Y = 1062.3649 - 0.9355\,X_1 - 938.6496\,X_2 - 4.0121\,X_3 + \varepsilon$$

Ovaj dobiveni model ima $R^2 = 0.7182$, što je veći u odnosu na prethodni model, te ovakav malo bolje opisuje podatke. Za ovaj model pogledajmo također grafički prikaz reziduala i teorijskih vrijednosti.



Uočimo sada kako je crvena krivulja manje za krivljena, pa bi nam to moglo sugerirati da je problem heteroskedastičnosti možda riješen. Pogledajmo što kažu statistički testovi.

Test	$p-$vrijednost
Breush-Pagan	0.6145
NCV	0.12796

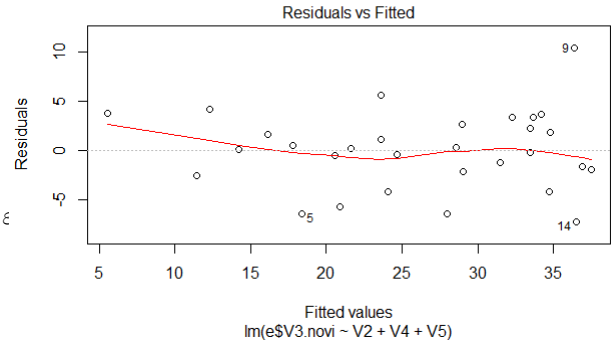
Primjetimo kako je $p-$ vrijednost BP-testa ostala jednaka, dok ncvTest u ovom slučaju nije odbacio hipotezu o homoskedastičnosti modela. Dakle, možemo zaključiti kako je ova metoda riješila problem heteroskedastičnosti.

7.2 Box-Cox transformacija

Kako je naš prvobitni model bio heteroskedastičan, pokušajmo prikadnom transformacijom ovisne varijable Y, dobiti model za koji nemamo razloga sumnjati u homoskedastičnost. Ako iskoristimo Box-Cox transformaciju dobijemo procjenu da je $\lambda = 0.4$. Tada je naš model oblika

$$Z = 41.78855 + 0.11261\,X_1 - 28.00532\,X_2 - 0.17784\,X_3 + \varepsilon$$

gdje je ovdje $Z = 2.5(Y^{\,0.4} - 1)$. Ovdje je $R^2 = 0.813$, pa možemo reći da ovakav model do sada najbolje bi opisivao podatke. Pogledajmo grafički prikaz i pripade $p-$vrijednosti testova kojima testiramo hipoteze o homoskedastičnosti.



Vidimo kako se ovaj grafički prikaz ne razlikuje mnogo od onoga koji smo imali za prvobitni model, ali pogledajmo rezultate testova.

Test	$p-$vrijednost
Breush-Pagan	0.6427
NCV	0.29387

Vidimo da niti jedan test ovdje nije odbacio hipotezu o homoskedastičnosti. Ova metoda kao i težinska MNK je za ovu bazu riješila problem heteroskedastičnosti.

8 Zaključak

Hetoroskedastičnost nastaje kada varijance za sva opažanja nisu jednake. Možemo je detektirati crtanjem dijagrama raspršenosti reziduala i teorijskih vrijednosti dobivenih modelom. Ako postoji prepoznatljiv uzorak, onda može biti prisutna heteroskedastičnost. Formalniji način identificiranja heteroskedastičnosti jest provođenje NCV testa i Breusch-Pagan testa, gdje se procjenjuje funkcija varijance koja ovisi o nezavisnoj varijabli i testira nul hipotezu da heteroskedastičnost nije prisutna u odnosu na alternativu koja je prisutna na heteroskedastičnosti. Postoje dvije glavne posljedice u prisutnosti heteroskedastičnosti. Prvo, procjenitelji dobiveni metodom najmanjih kvadrata su još uvijek linearni i nepristrani, ali više nisu najbolji. Drugo, standardne pogreške mogu biti pogrešne, što može utjecati na procjenu intervala i testiranje hipoteza. Da bismo ispravili ove posljedice, koristimo težinsku metodu najmanjih kvadrata da bismo dobili naše procjene parametara. Također za rješavanje ovoga problema možemo koristiti Box-Cox transformaciju.

Literatura

[1] https://datascienceplus.com/how-to-detect-heteroscedasticity-and-rectify-it/?fbclid=IwAR1MBREJYooLKDHRkfba06tDBwa1rNW3tdxdaCNCO-75ChKsmoZOGIkvwUc

[2] https://www.statisticshowto.datasciencecentral.com/box-cox-transformation/

[3] https://newonlinecourses.science.psu.edu/stat501/node/397/

[4] http://statisticsbyjim.com/regression/heteroscedasticity-regression/?fbclid=IwAR0rk7sIVN8DLW_pWayuUGCUR_aHnbfPZWTeKPI7kvsRZL3KbrIW8Ns1FtM

[5] https://www.stat.cmu.edu/~cshalizi/350/lectures/18/lecture-18.pdf

[6] https://www.reed.edu/economics/parker/s11/312/notes/Notes8.pdf

[7] https://rstudio-pubs-static.s3.amazonaws.com/187387_3ca34c107405427db0e0f01252b3fbdb.html