**UNIVERZITET U NOVOM SADU**
**PRIRODNO-MATEMATIČKI FAKULTET**
**DEPARTMAN ZA MATEMATIKU I INFORMATIKU**

Nina Moldvai

# Narušavanje pretpostavki u linearnoj regresiji

- master rad -

Mentor:

dr Zagorka Lozanov-Crvenković

Novi Sad, 2020.

# Predgovor

Statistika je prema Merriem-Websterovom (Merijem-Vebsterovom) rečniku „grana matematike koja se bavi prikupljanjem, analizom, interpretacijom i prezentacijom mase numeričkih podataka". Veliku primenu nalazi u svakom segmentu jednog društva – privredi, sportu, medicini, politici i tako dalje, a neophodna je za funkcionisanje države, odnosno donošenje odluka.

U statistici se linearna regresija odnosi na svaki pristup modeliranja relacija između jedne ili više promenljivih. Jednostruka linearna regresija je najjednostavniji oblik stohastičkog odnosa između dve promenljive. Jednačina jednostruke linearne regresije se sastoji od zavisne i nezavisne promenljive, slučajne greške i koeficijenata regresije. Ukoliko su koeficijenti nepoznati, za njihovo ocenjivanje se koriste tri metode: metoda najmanjih kvadrata, najbolje linearno nepristrasno ocenjivanje (Best Linear Unbiased Estimation ili skraćeno BLUE) i metoda maksimalne verodostojnosti. Da bi definisanje linearnog regresionog modela bilo kompletno, pored jednačine jednostruke linearne regresije potrebno je još da važe dodatne specifikacije vezane za distribuciju verovatnoće odstupanja.

Tema ovog master rada se odnosi na narušavanje pretpostavki u linearnoj regresiji. Reč je o standardnim pretpostavkama, a to su: normalnost, sredina jednaka nuli, homoskedastičnost, odsustvo autokorelacije i nestohastičnost promenljive.

U prvom delu rada su navedeni osnovni pojmovi jednostruke linearne regresije, tri metode ocenjivanja parametara, kao i njihove osobine. Provereno je da li ocenjivači dobijeni metodom najmanjih kvadrata zadovoljavaju sva poželjna svojstva.

Da li su pretpostavke zadovoljene, proverava se u drugom delu rada. Takođe, biće objašnjene i metode otklanjanja narušenih pretpostavki. Pre primene svake analize, utvrdiće se ispunjenost pretpostavki za njeno sprovođenje. Za proveru normalnosti distribucije podataka, kombinovaće se grafički pokazatelji, testovi Kolmogorov-Smirnov i Shapiro-Wilk (Šapiro-Vilk) za manje uzorke. U slučaju narušene pretpostavke o normalnosti, tretman za otklanjanje će biti nelinearna tranformacija, odnosno logaritamska transformacija. Ako je prisutna heteroskedastičnost, problem će se rešavati korišćenjem logaritamske transformacije. Odsustvo autokorelacije će biti provereno grafičkom metodom odnosno analizom reziduala. Ako se u podacima utvrdi da postoji autokorelacija, tretman za rešavanje ovog problema biće eliminisanje simptoma autokorelacije korišćenjem metode procene modela različite od metode najmanjih kvadrata. Kako je linearna regresija veoma osetljiva na netipične tačke ili outlier-e, detektovaće se na osnovu grafičkog prikaza, a potom proveriti putem *Box plot*-a. Još jedan od načina na koji će se detektovati je i računanje Mahalanobisovih (Mahalanobisovih) ili Cookovih (Kukovih) distanci.

U trećem delu rada, simulacijom podataka kroz tri primera prikazaće se rezultati prethodnih analiza. Detaljno će biti ispitane sve pretpostavke, a tretman će se vršiti na onima koje su narušene.

Cilj ovog rada je uspešna analiza, otkrivanje i otklanjanje narušenih pretpostavki navedenim statističkim metodama.

Za potrebe analize podataka korišćen je softverski paket SPSS 25 for Windows (Statistical Package for Social Sciences), a neophodni statistički podaci su preuzeti sa sajtova: *https://mathstatsresearch.weebly.com/regression.html* *http://staff.bath.ac.uk/pssiw/stats2/page16/page16.html* .

Ovom prilikom želela bih da se zahvalim svojoj profesorici i mentoru, dr Zagorki Lozanov-Crvenković za strpljenje i znanje koje mi je pružila tokom studiranja i pri izradi master rada.

Master rad posvećujem svojoj porodici.

# 1. Jednostruka linearna regresija

U statistici se linearna regresija odnosi na svaki pristup modeliranja relacija između jedne ili više promenljivih. Ona predstavlja prvi tip regresione analize koji se detaljno proučava. Razlog za ovakvo proučavanje je što se upravo modeli koji linearno zavise od svojih nepoznatih parametara lakše modeliraju nego modeli sa nelinearnom zavisnošću.

## 1.1. Opšti pojmovi

Jednostruka linearna regresija je najjednostavniji oblik stohastičkog odnosa između dve promenljive. Takav model se zapisuje u ovom obliku:

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

u kojem je $Y$ zavisna promenljiva, $X$ nezavisna, $\varepsilon$ slučajna greška koja nastaje prilikom merenja, a nepoznate parametre regresije predstavljaju $\alpha$ i $\beta$ (koeficijenti regresije). Za ocenjivanje nepoznatih parametara se koristi uzorak. Zavisna promenljiva $Y$ dobija se na osnovu $n$ vrednosti nezavisne promenljive $X$. Na taj način se dobija $n$ parova $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ koji čine uzorački model jednostruke linearne regresije:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad i = 1,2,3 \dots n$$

Za svaku vrednost promenljive $X$ postoji cela distribucija verovatnoća za vrednosti promenljive $Y$, što znači da se vrednost ove promenljive ne može tačno predvideti. Dakle, za kompletno definisanje regresionog modela, pored jednačine jednostruke linearne regresije, potrebno je još da važe dodatne specifikacije vezane za distribuciju verovatnoće odstupanja. Za slučajnu grešku $\varepsilon_i$, definisane su sledeće standardne pretpostavke:

1. *Normalnost*: $\varepsilon_i$ ima normalnu raspodelu za $i = 1,2,3 \dots n$
2. *Sredina jednaka nuli*: $E(\varepsilon_i) = 0,\ i = 1,2,3 \dots n$
3. *Homoskedastičnost*: $Var(\varepsilon_i) = \sigma^2, i = 1,2,3 \dots n$
4. *Odsustvo autokorelacije*: $Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$
5. *Nestohastičnost promenljive X*: $X$ je nestohastička promenljiva sa fiksnim vrednostima u ponovljenim uzorcima, takva da je za bilo koji uzorak veličine $n$, $\sum_{i=1}^{n}(X_i - \bar{X})^2 \neq 0$ i njena granična vrednost konačna kada je $n \to \infty$ .

Dakle, specifikacija linearnog regresionog modela se sastoji iz regresione jednačine i pet standardnih pretpostavki. To je takozvani „klasični normalni linearni regresioni model". Prve dve pretpostavke govore da je za svaku vrednost promenljive $X$ odstupanje normalno distribuirano oko nule. Stoga je $\varepsilon_i$ neprekidna promenljiva koja uzima vrednosti od $-\infty$ do $+\infty$. Simetrično je distribuirana oko njene sredine i njena je distribucija potpuno određena sredinom i varijansom. Treća pretpostavka se tiče homoskedastičnosti i znači da svako odstupanje ima istu varijansu $\sigma^2$ čija je vrednost nepoznata. Četvrta pretpostavka zahteva da odstupanja budu nekorelirana. Pri toj pretpostavci činjenica je da, na primer, ako je današnja proizvodnja veća od očekivane, to ne bi trebalo da uzrokuje veću ili manju proizvodnju sutra. Dakle, na osnovu prve i četvrte pretpostavke može se zaključiti da su verovatnoće odstupanja nezavisne (za normalne slučajne promenljive nekolinearnost povlači nezavisnost (A.S. Goldberger,

*Econometric Theory* (New York: Wiley, 1964) str 107.-108.). Stoga se $\varepsilon_1, \varepsilon_2 \dots, \varepsilon_n$ mogu posmatrati kao skup $n$ identičnih i nezavisno distribuiranih promenljivih. Može se primetiti da je $var(\varepsilon_i) = E(\varepsilon_i^2)$ i $cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j)$ zbog toga što je $E(\varepsilon_i) = 0$. Poslednja pretpostavka ograničava na razmatranje onih situacija u kojima je vrednost promenljive $X$ ili kontrolisana ili potpuno predvidiva. Važna je implikacija te pretpostavke, a to je da je $E(\varepsilon_i X_j) = X_j E(\varepsilon_i) = 0$ za sve $i, j$. U ovoj pretpostavci je naznačeno da su vrednosti promenljive $X$ fiksne, odnosno da se u ponovljenim uzorcima uzimaju iste vrednosti. Da su sve vrednosti promenljive u uzorku konačne, da ne mogu sve biti jednake istom broju i da ne mogu neograničeno rasti ili opadati, govori sledeća jednačina:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \neq 0.$$

Pretpostavke na kojima se temelji klasični normalni linearni regresioni model se koriste pri izvođenju ocenjivača parametara regresije. Kako se pretpostavlja da odstupanje ima normalnu raspodelu, $N(0, \sigma^2)$ sredina je poznata i jednaka nuli, a varijansa odstupanja $\sigma^2$ je jedina stvar koja je nepoznata u ovoj raspodeli. Stoga ovaj model ima tri nepoznata parametra: parametre regresije $\alpha$ i $\beta$, i varijansu odstupanja $\sigma^2$. Trebalo bi naglasiti da postoji mogućnost da bilo koja ili više standardnih pretpostavki bude narušeno. U ovom radu će biti pokazano šta se dešava sa svojstvima izvedenih ocenjivača kada pretpostavke nisu ispunjene.

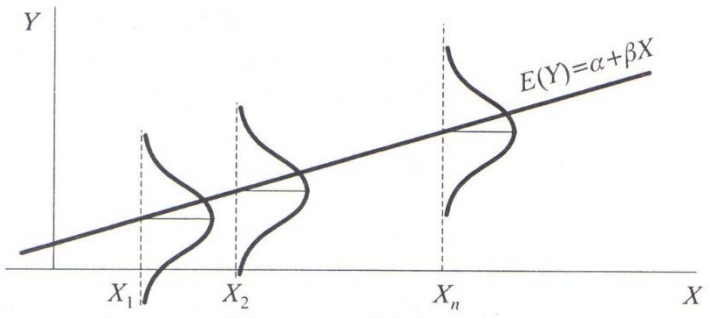
Nakon potpunog definisanja regresionog modela, mogu se bliže prodiskutovati osnovna svojstva. Ako se uzme u razmatranje slučajna promenljiva $Y$, sredina vrednosti $Y_i$ se može izračunati primenom matematičkog očekivanja na obe strane jednačine. Na taj načina dobija se:

$$E(Y_i) = E(\alpha + \beta X_i + \varepsilon_i) = \alpha + \beta X_i \quad (8)$$

Ovo proizilazi iz činjenice da su $\alpha$ i $\beta$ parametri, vrednost $X_i$ determinističke, a u skladu sa drugom pretpostavkom, sredina $E(\varepsilon_i) = 0$. Do varijanse za $Y_i$ se dolazi na sledeći način:

$$Var(Y_i) = E[Y_i - E(Y_i)]^2 = E[(\alpha + \beta X_i + \varepsilon_i) - (\alpha + \beta X_i)]^2 = E(\varepsilon_i{}^2) = \sigma^2$$

Pri ovom izvođenju prvo je iskorišćena opšta definicija varijanse, potom je za $Y_i$ uvrštena jednačina $\alpha + \beta X_i + \varepsilon_i$ kao i za $E(Y_i)$, i konačno je primenjena pretpostavka o homoskedastičnosti iz pretpostavke (3). Kada je reč o raspodeli promenljive $Y_i$, iz jednačine (1) se može videti da je $Y_i$ samo linearna funkcija od $\varepsilon_i$, i budući da $\varepsilon_i$ ima normalnu raspodelu, po Teoremi 14. (*Počela ekonometrije,* Jan Kmenta, str 90.) ako su $X$, $Y$,… $Z$ nezavisne slučajne promenljive sa normalnom raspodelom, i $a$, $b$,…, $c$ konstante, tada i linearna kombinacija $aX + bY + \cdots + cZ$ takođe ima normalnu raspodelu, onda sledi da i $Y_i$ ima normalnu raspodelu. Prema tome, sredina je jednaka ($\alpha + \beta X_i$), a varijansa $\sigma^2$, što se drugačije zapisuje: $Y_i \sim N(\alpha + \beta X_i, \sigma^2)$.




Slika 7-2, *Počela ekonometrije*, Jan Kmenta, str 210.

Iz druge i četvrte pretpostavke, za $i \neq j$ sledi da je:

$$Cov(Y_i, Y_j) = E[Y_i - E(Y_i)][Y_j - E(Y_j)] = E(\varepsilon_i \varepsilon_j) = 0$$

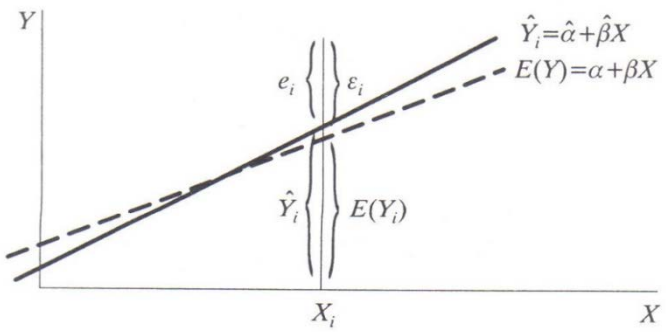
Otuda $Y_1, Y_2, \dots, Y_n$ možemo posmatrati kao skup $n$ nezavisnih promenljivih sa normalnom raspodelom. Međutim, ove promenljive imaju različite sredine i zato nemaju identičnu raspodelu.
Jednačinu (8) koja za svaku vrednost promenljive $X$ daje vrednost promenljivoj $Y$ nazivamo još i regresijska linija populacije. Odsečak te linije, $\alpha$, meri srednju vrednost promenljive $Y$ za $X = 0$. Nagib linije, $\beta$, meri promenu srednje vrednosti promenljive $Y$ koja odgovara jedinici promene vrednosti promenljive $X$. Budući da su vrednosti ovih parametara nepoznate, nepoznata je i regresijska linija populacije. Kada se ocene vrednosti $\alpha$ i $\beta$, dobija se regresijska linija uzorka koja služi kao ocena regresijske linije populacije. Ako su ocene za $\alpha$ i $\beta$, $\hat{\alpha}$ i $\hat{\beta}$, tada je regresijska linija uzorka:

$$\widehat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

gde je sa $\widehat{Y}_i$ obeležena prilagođena vrednost promenljive $Y_i$. Većina opaženih vrednosti promenljive $Y$ neće ležati tačno na regresijskoj liniji i zato će se vrednosti $Y_i$ i $\widehat{Y}_i$ razlikovati. Ta razlika se naziva rezidual i označava se sa $e_i$. Veoma je važno napraviti sledeću razliku:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad \text{(populacija)}$$
$$Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i \qquad \text{(uzorak)}$$

Generalno, $e_i$ se razlikuje od $\varepsilon_i$ zato što se i $\hat{\alpha}$ i $\hat{\beta}$ razlikuju od stvarnih vrednosti za $\alpha$ i $\beta$. Zapravo se ostaci $e_i$, mogu posmatrati kao ocene odstupanja $\varepsilon_i$. To se može jasnije videti na sledećoj slici:




Slika 7-3, *Počela ekonometrije*, Jan Kmenta, str 211.

## 1.2. Ocenjivanje parametara regresije

Problem ocenjivanja parametara regresijskog modela može se posmatrati kao problem ocenjivanja parametara raspodele verovatnoće zavisne promenljive $Y$. Kao što je već pokazano, kada važe pretpostavke modela, $Y_i$ ima normalnu raspodelu, sa sredinom jednakom $E(Y_i) = \alpha + \beta X_i$, i varijansom $Var(Y_i) = \sigma^2$. Problem ocenjivanja parametara regresije $\alpha$ i $\beta$ je prema tome ekvivalentan problemu ocenjivanja sredine promenljive $Y_i$. On se može rešiti pomoću nekoliko različitih načina, a u ovom radu će biti pokazane tri metode: metoda najmanjih kvadrata, najbolje linearno nepristrasno ocenjivanje (Best Linear Unbiased Estimation – BLUE) i metoda maksimalne verodostojnosti. Cilj ovih ocenjivanja je pronaći ocenjivač koji ima što više poželjnih svojstava. Takav ocenjivač će potom biti pogodan za testiranje hipoteza regresijskog modela, kao i za predviđanja.

### 1.2.1. Metoda najmanjih kvadrata

Prvi način ocenjivanja parametara $\alpha$ i $\beta$ je metodom najmanjih kvadrata jer ocene dobijene na ovaj način imaju sva poželjna svojstva ukoliko su ispunjene standardne pretpostavke. Cilj ove metode jeste da se pronađe minimalna suma kvadrata odstupanja registrovanih vrednosti od njihove sredine. Suma koju treba minimizirati obeležena je sa $S$ i izgleda ovako:

$$S = \sum_{i=1}^{n} [Y_i - E(Y_i)]^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$$

Kako bi se izračunale vrednosti $\alpha$ i $\beta$, treba pronaći izvod sume $S$ po $\alpha$ i po $\beta$.

$$\frac{\partial S}{\partial \alpha} = \sum_i \frac{\partial (Y_i - \alpha - \beta X_i)^2}{\partial \alpha} = \sum_i 2(Y_i - \alpha - \beta X_i)(-1)$$

$$= -2 \sum_i (Y_i - \alpha - \beta X_i) \quad (*)$$

$$\frac{\partial S}{\partial \beta} = \sum_i \frac{\partial (Y_i - \alpha - \beta X_i)^2}{\partial \beta} = \sum_i 2(Y_i - \alpha - \beta X_i)(-X_i)$$

$$= -2 \sum_i X_i (Y_i - \alpha \beta X_i) \quad (**)$$

Ocenjene vrednosti $\hat{\alpha}$ i $\hat{\beta}$ se dobijaju izjednačavanjem (*) i (**) sa nulom, odnosno:

$$-2\sum_i \left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) = 0$$

$$-2\sum_i X_i\left(Y_i - \hat{\alpha}\hat{\beta}X_i\right) = 0$$

Ekvivalentno prethodnom zapisu, dobijaju se normalne jednačine metode najmanjih kvadrata:

$$\sum Y_i = n\hat{\alpha} - \hat{\beta}(\sum X_i) \qquad \text{(n1)}$$

$$\sum X_i Y_i = \hat{\alpha}(\sum X_i) - \hat{\beta}\left(\sum {X_i}^2\right) \quad \text{(n2)}$$

Budući da se $Y_i$ može zapisati kao $Y_i = \hat{\alpha} + \hat{\beta}X_i + e_i$, gde su $e_i$ reziduali najmanjih kvadrata, normalne jednačine se mogu zapisati i u ovom obliku:

$$\sum e_i = 0\text{-}$$

$$\sum X_i e_i = 0$$

Jednačine (n1) i (n2) se mogu rešiti po $\hat{\alpha}$ i po $\hat{\beta}$. Rešenje za $\hat{\beta}$ je:

$$\hat{\beta} = \frac{n(\sum X_i Y_i) - (\sum X_i)(\sum Y_i)}{n(\sum {X_i}^2) - (\sum X_i)^2}$$

Može se uočiti da se imenilac i brojilac ovog razlomka mogu napisati i na sledeći način:

$$n\sum (X_i - \bar{X})\,(Y_i - \bar{Y})$$

$$= n\left(\sum X_i Y_i\right) - n\bar{X}\left(\sum Y_i\right) - n\bar{Y}\left(\sum X_i\right) + n^2\bar{X}\bar{Y}$$

$$= -(\sum X_i)n(\sum X_i Y_i)(\sum Y_i) - (\sum X_i)(\sum Y_i) + (\sum X_i)(\sum Y_i)$$

$$= n\left(\sum X_i Y_i\right) - \left(\sum X_i\right)\left(\sum Y_i\right)$$

$$n\sum (X_i - \bar{X})^2$$

$$= n\left(\sum {X_i}^2\right) - 2n\bar{X}\left(\sum X_i\right) + n^2\bar{X}^2$$

$$= n\left(\sum {X_i}^2\right) - 2\left(\sum X_i\right)^2 + \left(\sum X_i\right)^2$$

$$= n\left(\sum {X_i}^2\right) - \left(\sum X_i\right)^2$$

Prema tome, $\hat{\beta}$ se zapisuje:

$$\hat{\beta} = \frac{n\sum(X_i-\bar{X})(Y_i-\bar{Y})}{n\sum(X_i-\bar{X})^2} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} \qquad \text{(T1)}$$

Ako se uvedu oznake $S_{XY} = \sum(X_i - \bar{X})\,(Y_i - \bar{Y})$, $S_{XX} = \sum(X_i - \bar{X})^2$, tada je $\hat{\beta} = \frac{S_{XY}}{S_{XX}}$. Ocena $\hat{\alpha}$ se dobija jednostavno iz jednačine (n1):

$$\hat{\alpha} = \frac{1}{n}(\sum Y_i) - \frac{1}{n}\hat{\beta}(\sum X_i) = \bar{Y} - \hat{\beta}\bar{X} \quad \text{(T2)}$$

Ovo znači da regresijska linija uzorka $\widehat{Y_i} = \hat{\alpha} + \hat{\beta}X_i$ prolazi kroz tačku $(\bar{X}, \bar{Y})$. Vrednost $\hat{\alpha}$ meri odsečak, a $\hat{\beta}$ nagib regresijske linije uzorka.

Dakle, ako su ispunjene standardne pretpostavke, ocene parametara $\alpha$ i $\beta$ dobijene metodom najmanjih kvadrata, BLUE metodom i metodom maksimalne verodostojnosti su ekvivalentne.

### 1.2.2. BLUE metoda

Drugi način za ocenu parametara je metoda najboljeg linearnog nepristrasnog ocenjivanja, BLUE (*best linear unbiased evaluator*) za $\alpha$ i $\beta$.

Nazovimo $\tilde{\beta}$ takvu ocenu za $\beta$. Tada mora da važi:

1. Linearnost $\qquad \tilde{\beta} = \sum_i a_i Y_i$

gde su $a_i, i = 1,2,\dots,n$ konstante koje treba odrediti.

2. Nepristrasnost $\qquad E(\tilde{\beta}) = \beta$

$$E(\tilde{\beta}) = E\left(\sum_i a_i Y_i\right) = \sum_i a_i E(Y_i) = \sum_i a_i E(\alpha + \beta X_i) = \alpha\left(\sum_i a_i\right) + \beta\left(\sum_i a_i X_i\right)$$

Dakle, da bi $\tilde{\beta}$ bio nepristrasan, moraju da važe sledeći uslovi:

$$\sum_i a_i = 0 \qquad \text{i} \qquad \sum_i a_i X_i = 1$$

3. $Var(\tilde{\beta})$ najmanja

$$Var(\tilde{\beta}) = Var\left(\sum a_i Y_i\right) = \sum {a_i}^2 Var(Y_i) = \sum {a_i}^2 \sigma^2 = \sigma^2 \sum {a_i}^2$$

jer su $Y_i$ nezavisni i svako opažanje ima jednaku varijansu $\sigma^2$, dok su $a_i$ nestohastički.
Ostaje još da se nađu $a_i$ takvi da važi $\sum_i a_i = 0$ i $\sum_i a_i X_i = 1$, a da je istovremeno $\sigma^2 \sum {a_i}^2$ minimalno. Ovo je problem minimizacije uz ograničenja koji se može rešiti pomoću metode Lagrangeovih (Lagranžovih) multiplikatora. Izvođenje ove metode se nalazi u knjizi *Počela ekonometrije*, od strane 216. do 218., a u ovom radu će se koristiti već izvedena formula.

Dakle, $a_i$ je jednako:

$$a_i = \frac{-(\sum X_i)+nX_i}{n(\sum {X_i}^2)-(\sum X_i)^2} = \frac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2} \qquad \text{za } i = 1,2,\dots,n$$

To su konstante koje čine $\tilde{\beta}$ nepristrasnim ocenjivačem i minimizuju njegovu varijansu. Kada se $a_i$ uvrsti u početnu formulu, sledi:

$$\tilde{\beta} = \sum_i a_i Y_i = \frac{(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

Tako se dobija isti rezultat kao i metodom najmanjih kvadrata za ocenjivač od $\beta$.

Primenom BLUE načela ne ocenjuju se samo regresioni parametri već se mogu naći i njihove varijanse.

Ubacivanjem rezultata dobijenog za $a_i$ u $Var(\tilde{\beta}) = \sigma^2 \sum {a_i}^2$ dobija se sledeći izraz:

$$\sum a_i^2 = \frac{-(\sum X_i)(\sum a_i) + n(\sum a_i X_i)}{n(\sum {X_i}^2) - (\sum X_i)^2}$$

Poznato je još da je $\sum_i a_i = 0$ i $\sum_i a_i X_i = 1$, s toga prethodni izraz postaje:

$$\sum a_i^2 = \frac{n}{n(\sum {X_i}^2) - (\sum X_i)^2} = \frac{1}{\sum(X_i - \bar{X})^2}$$

$$Var(\tilde{\beta}) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} = \frac{\sigma^2}{S_{XX}}$$

Preostaje još da se pronađu BLUE od $\alpha$, odsečka linearne regresijske linije. Izvođenje je analogno izvođenju za $\tilde{\beta}$, tako da će ovde biti predstavljeno samo krajnje rešenje.

$$\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{X}$$

$$Var(\tilde{\alpha}) = \frac{\sigma^2(\sum {X_i}^2)}{n\sum(X_i - \bar{X})^2} = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right)$$

Može se zapaziti da je i BLUE od $\alpha$, $\tilde{\alpha}$, isti kao ocenjivač koji se dobija za $\alpha$ metodom najmanjih kvadrata,$\hat{\alpha}$. Ovaj rezultat, zajedno sa rezultatom BLUE od $\beta$ čine Gauss-Markovljevu teoremu.

### 1.2.3. Metoda maksimalne verodostojnosti

Poslednji način ocenjivanja parametara koji će se obraditi u ovom radu je metoda maksimalne verodostojnosti (engl. *maximum likelihood estimation* - *MLE*). Suština ovog načina ocenjivanja parametara zasniva se na ideji da se pomoću izabranog uzorka odabere ona vrednost nepoznatog parametra, koja daje najveću verovatnoću da baš taj uzorak bude odabran. Da bi se našli ovi ocenjivači, mora se prvo odrediti a potom maksimizirati funkcija verodostojnosti. U slučaju linearnog regresionog modela, uzorak sadrži $n$ promenljivih $Y_1, Y_2, \dots, Y_n$ koje imaju normalnu raspodelu, redom jednake sredine $(\alpha + \beta X_1), (\alpha + \beta X_2), \dots, (\alpha + \beta X_n)$ i zajedničku varijansu $\sigma^2$. Neka su registrovane vrednosti

ozančene sa $y_1, y_2, \dots, y_n$. Vezu između raspodele verovatnoće promenljive $Y_i$ i raspodele verovatnoće od $\varepsilon_i$ objašnjava sledeća teorema:

Teorema (Zamena promenljivih) Ako slučajna promenljiva $X$ ima gustinu verovatnoće $f(x)$ i ako je promenljiva $Z$ funkcija od $X$ takva da postoji obostrano jednoznačno preslikavanje između $X$ i $Z$, tada je gustina verovatnoće promenljive $Z$ jednaka:

$$f(z) = \left|\frac{dx}{dz}\right| g(x), dx/dz \neq 0$$

Dokaz ove teoreme se nalazi u knjizi *Mathematical Statistics*, John E. Freund i Ronald. Walpole, i u ovom radu se nije dokazivala.

Važnost ove teoreme je zapravo činjenica da se raspodela promenljive može pronaći na osnovu raspodele druge povezane promenljive. U prikazanom regresionom modelu, poznata je raspodela za $\varepsilon_i$, i stoga se primenom ove teoreme može doći do raspodele za $Y_i$.

Dakle, regresioni model izgleda ovako:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Jasno je da postoji obostrano jednoznačno preslikavanje između $Y_i$ i $\varepsilon_i$, i zbog toga se može primeniti prethodno pomenuta teorema.

$$f(y_i) = \left|\frac{d\varepsilon_i}{dY_i}\right| g(\varepsilon_i)$$

Međutim, kako je $\varepsilon_i = Y_i - \alpha - \beta X_i$, onda je $\left|\frac{d\varepsilon_i}{dY_i}\right| = 1$.

I zbog toga je:

$$f(y_i) = g(\varepsilon_i)$$

Promenljive $Y_i$ su nezavisne jer su normalne i nekorelirane. Funkcija verodostojnosti se tada zapisuje:

$$l = f(y_1)f(y_2), \dots, f(y_n)$$

Vrednosti parametara koji maksimizuju $l$ su iste kao i vrednosti koje maksimizuju njen logaritam, definiše se onda $L = logl$, pa se prethodna formula tako može zapisati i na sledeći način:

$$L = \sum_{i=1}^{n} logf(y_i)$$

Ukoliko se sada iskoristi činjenica da $Y_i$ imaju normalnu raspodelu, sa sredinom $(\alpha + \beta X_i)$ i varijansom $\sigma^2$ i to se ubaci u formulu za normalnu raspodelu, dobija se sledeće:

$$log\mathrm{f}(y_i) = -\frac{1}{2}log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{Y_i - \alpha - \beta X_i}{\sigma}\right)^2$$

$$L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (Y_i - \alpha - \beta X_i)^2$$

U formuli $L$ se nalaze tri nepoznata parametra, i stoga će se $L$ diferencirati po svakom od njih:

$$\frac{\partial L}{\partial \alpha} = -\frac{1}{2\sigma^2}\sum_i 2(Y_i - \alpha - \beta X_i)(-1)$$

$$\frac{\partial L}{\partial \beta} = -\frac{1}{2\sigma^2}\sum_i 2(Y_i - \alpha - \beta X_i)(-X_i)$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i 2(Y_i - \alpha - \beta X_i)^2$$

Izjednačavanjem ovih izraza sa nulom, i stavljanjem * na parametre koji se ocenjuju, dobijaju se sledeće jednačine:

$$\frac{1}{2\sigma^{*2}}\sum_i (Y_i - \alpha^* - \beta^* X_i) = 0$$

$$\frac{1}{2\sigma^{*2}}\sum_i X_i(Y_i - \alpha^* - \beta^* X_i) = 0$$

$$-\frac{n}{2\sigma^{*2}} + \frac{1}{2\sigma^{*4}}\sum_i (Y_i - \alpha^* - \beta^* X_i)^2 = 0$$

Iz prve dve jednačine sledi:

$$\sum_i Y_i = -\alpha^* n - \beta^*\left(\sum_i X_i\right)$$

$$\sum_i X_i Y_i = \alpha^*\left(\sum_i X_i\right) + \beta^*\left(\sum_i {X_i}^2\right)$$

Dobijene jednačine su iste kao i normalne jednačine najmanjih kvadrata date sa (n1) i (n2). Odatle se može zaključiti da su maksimalno verodostojni ocenjivači za $\alpha$ i $\beta$, $\alpha^*$ i $\beta^*$, jednaki ocenjivačima metode najmanjih kvadrata.

Treća jednačina daje maksimalnu verodostojnost ocenjvača od $\sigma^2$ i ona glasi:

$$\sigma^{*2} = \frac{1}{n}\sum_i (Y_i - \alpha^* - \beta^* X_i)^2$$

Koristeći prethodno izvedeni zaključak za $\alpha^*$ i $\beta^*$ dobija se sledeća jednakost:

$$\sigma^{*2} = \frac{1}{n}\sum_i {e_i}^2$$

gde su $e_i = Y_i - \widehat{Y}_i$ reziduali, odnosno ocene greške $\varepsilon_i$ . Kako je $\sum e_i = 0$, varijanse odstupanja dobijene metodom maksimalne verodostojnosti jednake su neprilagođenoj varijansi uzorka reziduala koji se dobija metodom najmanjih kvadrata.

Pretpostavka o normalnosti, kao i pretpostavka o neautokoreliranosti, povlače nezavisnost, što znači slučajno uzmanje uzorka. Ove dve pretpostavke ne pružaju nikakvu specifičnost kada je reč o parametrima regresije.

Pretpostavka da je sredina jednaka nuli zajedno sa pretpostavkom o homoskedastičnosti vode do:

$$\frac{1}{n}\sum e_i = 0$$

$$\frac{1}{n}\sum_i {e_i}^2 = \hat{\sigma}^2$$

gde je $\hat{\sigma}^2$ ocenjivač od $\sigma^2$. Pretpostavka o nestohastičnosti nezavisne promenljive povlači da su $\varepsilon$ i $X$ nekorelirani što daje sledeći uslov:

$$\frac{1}{n}\sum e_i X_i = 0$$

Kada se ove tri jednačine reše za ocene od $\alpha$, $\beta$ i $\sigma^2$ dobijaju se isti rezultati kao i za metodu maksimalne verodostojnosti.

## 1.3. Svojstva ocenjivača

Nakon razmatranja svih metoda, dolazi se do zaključka da svaka dovodi do istih ocena parametara regresije. Ukoliko se radi o klasičnom normalnom linearnom regresionom modelu, ocenjivači dobijeni metodom najmanjih kvadrata jednaki su najboljim linearnim nepristrasnim i maksimalno verodostojnim ocenjivačima. Rezultat metode najmanjih kvadrata jesu formule ocenjivača za $\alpha$ i $\beta$, na osnovu BLUE metode dobija se formula za njihove varijanse, a MLE metoda ocenjivanja je kao rezultat dala ocenu za $\sigma^2$. Sada treba razmotriti da li ocenjivači dobijeni metodom najmanjih kvadrata imaju sve poželjne osobine.

Za koeficijente linearne regresije u slučaju malog uzorka (ispod 30 opažanja), poželjno je da imaju sledeće osobine:

1. Nepristrasnost: Ocena je nepristrasna ako je očekivana vrednost ocene $\hat{\beta}$ jednaka pravoj vrednosti $\beta$ tj. $E(\hat{\beta}) = \beta$.

Razlika između ove dve vrednosti predstavlja pristrasnost ocene $\hat{\beta}$ .

2. Efikasnost: Ocena je najefikasnija ako je nepristrasna i ima najmanju disperziju među svim ostalim nepristrasnim ocenama istog parametra.
3. BLUE: Najbolja linearna nepristrasna ocena. Ocena sa ovom osobinom treba da zadovolji uslove da je ocena linearna funkcija opažanja, da je nepristrasna i da ima najmanju disperziju od svih ostalih ocena.

Ukoliko je obim uzorka veliki, poželjne su sledeće asimptotske osobine ocena parametara:

1. Asimptotska nepristrasnost podrazumeva da se povećanjem uzorka dobija što bolja ocena koeficijenta — očekivana vrednost ocene teži pravoj vrednosti koeficijenta kako uzorak raste, odnosno:

$$\lim_{\mathrm{n}\to\infty} \mathrm{E}\left(\hat{\beta}\right) = \beta$$

2. Konzistentnost: Dovoljan uslov za ovu osobinu je

$$\lim_{\mathrm{n}\to\infty} \mathrm{MSE}\left(\hat{\beta}\right) = 0$$

gde je $\mathrm{MSE}\left(\hat{\beta}\right) = E\left(\left(\hat{\beta} - \beta\right)^2\right)$ srednje kvadratna greška ocene $\beta$ .

3. Asimptotska efikasnost podrazumeva da ocena koeficijenta ima osobinu konzistentnosti, najmanju asimptotsku varijansu i asimptotsku distribuciju sa konačnom sredinom i varijansom.

Prvo se može zaključiti da su ovi ocenjivači nepristrasni jer su BLUE. Koristeći Rao-Cramérovu (Rao-Kramerovu) teoremu o najmanjim disperzijama, može se dokazati da su ovi ocenjivači efikasni. Konačno, kako su ovi ocenjivači jednaki MLE ocenjivačima, tada važe i asimptotska nepristrasnost, konzistentnost i asimptotska efikasnost.

Na osnovu svega prethodnog navedenog, dolazi se do važnog zaključka. Ocenjivači dobijeni metodom najmanjih kvadrata imaju sva poželjna svojstva konačnog uzorka i sva poželjna asimptotska svojstva kada važe osnovne pretpostavke. U prethodnom odeljku, izvedene su formule za $Var\left(\tilde{\beta}\right)$ I $Var(\tilde{\alpha})$ koje sadrže nepoznati parametar $\sigma^2$ koji se može oceniti. Koristeći formulu dobijenu ocenjivanjem metodom maksimalne verodostojnosti:

$$\sigma^{*2} = \frac{1}{n}\sum_i (Y_i - \alpha^* - \beta^* X_i)^2$$

Kako je ovo MLE ocenjivač od $\sigma^2$, on dakle ima sva poželjna asimptotska svojstva. Potrebno je još proveriti njegovu nepristrasnost. U prethodnoj jednačini je potrebno zameniti $Y_i$:

$$\sigma^{*2} = \frac{1}{n}\sum_i (\alpha - \beta X_i + \varepsilon_i - \alpha^* - \beta^* X_i)^2 = \frac{1}{n}\sum [-(\alpha^* - \alpha) - (\beta^* - \beta)X_i + \varepsilon_i]^2$$

U knjizi *Počela ekonometrije* je objašnjen ceo postupak računanja kovarijanse od $\tilde{\alpha}$ i $\tilde{\beta}$. U ovom radu će se koristiti izvedene formule za $\tilde{\alpha} - \alpha$ i $\tilde{\beta} - \beta$:

$$\tilde{\alpha} - \alpha = -\left(\tilde{\beta} - \beta\right)\bar{X} + \bar{\varepsilon}$$

$$\tilde{\beta} - \beta = \frac{\sum(X_i - \bar{X})\,\varepsilon_i}{\sum {X_i}^2}$$

Njihovim uvrštavanjem u prethodnu jednačinu za $\sigma^{*2}$ dobija se:

$$\sigma^{*2} = \frac{1}{n}\sum[(\beta^* - \beta)\bar{X} - \bar{\varepsilon} - (\beta^* - \beta)X_i + \varepsilon_i]^2$$

$$= \frac{1}{n}\sum[-(\beta^* - \beta)(X_i - \bar{X}) + (\varepsilon_i - \bar{\varepsilon})]^2$$

$$= \frac{1}{n}\sum[(\beta^* - \beta)^2(X_i - \bar{X})^2 + (\varepsilon_i - \bar{\varepsilon})^2 - 2(\beta^* - \beta)(X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon})]$$

$$= -\frac{1}{n}(\beta^* - \beta)^2\sum(X_i - \bar{X})^2 + \frac{1}{n}\sum(\varepsilon_i - \bar{\varepsilon})^2$$

Ukoliko se primeni matematičko očekivanje na obe strane, dobija se:

$$E(\sigma^{*2}) = -\frac{1}{n}\sum(X_i - \bar{X})^2 E(\beta^* - \beta)^2 + \frac{1}{n}\sum E(\varepsilon_i - \bar{\varepsilon})^2$$

Dalje je:

$$E(\beta^* - \beta)^2 = Var(\beta^*) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

$$E\frac{1}{n}\sum(\varepsilon_i - \bar{\varepsilon})^2 = E\left(\frac{1}{n}\sum {\varepsilon_i}^2 - \bar{\varepsilon}^2\right) = \sigma^2 - \frac{\sigma^2}{n} = \left(\frac{n-1}{n}\right)\sigma^2$$

Dobijeni rezultati se ubacuju u jednačinu $E(\sigma^{*2})$ i odatle sledi da je:

$$E(\sigma^{*2}) = -\frac{1}{n}\sum(X_i - \bar{X})^2\left(\frac{\sigma^2}{\sum(X_i - \bar{X})^2}\right) + \left(\frac{n-1}{n}\right)\sigma^2$$

$$= -\frac{\sigma^2}{n} + \left(\frac{n-1}{n}\right)\sigma^2 = \left(\frac{n-2}{n}\right)\sigma^2$$

Iz ovoga se može zaključiti da je $\sigma^{*2}$ pristrasan ocenjivač od $\sigma^2$.

Da bi se dobio nepristrasan ocenjivač od $\sigma^2$, potrebno je pomnožiti obe strane sa $\left(\frac{n}{n-2}\right)$:

$$\left(\frac{n}{n-2}\right)E(\sigma^{*2}) = \sigma^2$$

$$E\left(\frac{n}{n-2}\right)\frac{1}{n}\sum_i (Y_i - \alpha^* - \beta^* X_i)^2 = \sigma^2$$

$$E\left(\frac{1}{n-2}\right)\sum_i (Y_i - \alpha^* - \beta^* X_i)^2 = \sigma^2$$

Odatle se može izraziti $s^2$:

$$s^2 = \frac{1}{n-2}\sum_i (Y_i - \alpha^* - \beta^* X_i)^2 = \frac{1}{n-2}\sum_i {e_i}^2$$

Ukoliko se koristi $s^2$ kao nepristrasan ocenjivač od $\sigma^2$, mogu se takođe izvesti formule za nepristrasne varijanse od $\alpha^*$ i $\beta^*$, obeležene sa $s^2_{\alpha^*}$ i $s^2_{\beta^*}$:

$$s^2_{\alpha^*} = s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right)$$

$$s^2_{\beta^*} = \frac{s^2}{S_{XX}}$$

# 2. Narušavanje osnovnih pretpostavki

U ovom radu je do sada prikazan linearni regresioni model i kako se isti može iskoristiti za ocenjivanje. Pri izvođenju ovih rezultata polazilo se od pet osnovnih pretpostavki. Prve četiri su vezane za odstupanje $\varepsilon_i$, odnosno da $\varepsilon_i$ imaju normalnu raspodelu, sredinu jednaku nuli, homoskedastični su i neautokorelirani. Peta pretpostavka se odnosi na $X_i$, govori da je ovo nezavisna deterministička promenljiva i da je varijansa iz uzorka bilo koje veličine konačan broj. Takođe je pokazano da ocenjivači dobijeni metodom najmanjih kvadrata imaju sva poželjna svojstva kada su ove pretpostavke ispunjene. U nastavku rada će se utvrditi kako narušavanje osnovnih pretpostavki može da utiče na svojstva ocenjivača dobijenih prethodno pomenutom metodom.

## 2.1. Nenormalnost

Prva pretpostavka o normalnosti je osnovna pretpostavka modela linearne regresije

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

gde je $\varepsilon_i$ stohastičko odstupanje sa normalnom raspodelom. Kao što je pokazano u prethodnom poglavlju, $\varepsilon_i$ predstavlja sumu ili prosek velikog broja nezavisnih učinaka koji imaju istu verovatnoću da budu pozitivni ili negativni i ne postoji nijedan koji ima dominantnu varijabilnost. Pretpostavka o normalnosti je bila neophodna kod dokaza efikasnosti ocenjivača parametara regresije dobijenih metodom najmanjih kvadrata. Međutim, dokle god su ispunjene ostale osnovne pretpostavke i dokle god je $\varepsilon_i$ konačno, ocenjivači dobijeni metodom najmanjih kvadrata su najbolji linearni nepristrasni ocenjivači (BLUE) bez obzira na oblik distribucije stohastičke promenljive $\varepsilon_i$. Što se tiče drugih poželjnih svojstava, LSE od $\alpha\, i\, \beta$ su asimptotski nepristrasni, jer su nepristrasni za bilo koju veličinu uzorka. Takođe se može lako izvesti i dokaz njihove konzistentnosti. Prema tome, kad odstupanje nije normalno distribuirano (ali ima konačnu varijansu $\sigma^2$), ocenjivači metode najmanjih kvadrata su još i BLUE i konzistentni, ali ne nužno i efikasni ili asimptotski efikasni.

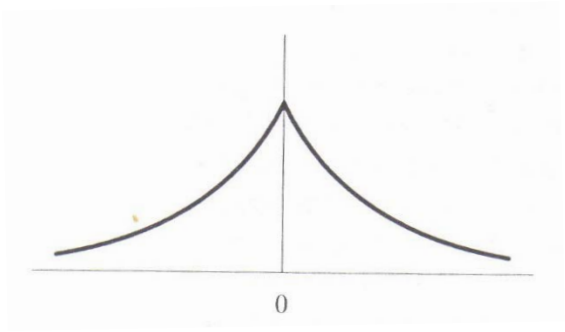
### 2.1.1. Ocenjivanje u uslovima nenormalnosti

Na osnovu prethodno navedenih činjenica, narušavanje pretpostavke o normalnosti nema ozbiljne posledice za ocenjivanje metodom najmanjih kvadrata. Prema navodima nekih kritičara, ekstremna odstupanja, koja se dešavaju veoma retko, imaju neopravdano jak uticaj na ovaj način ocenjivanja. Takva odstupanja mogu uzrokovati izrazito nenormalni vremenski uslovi, grube greške, politički događaji, glasine i slični događaji. Tvrdi se da se odstupanja mogu bolje modelirati nekom distribucijom koja ima manje osetljive krajeve i veću varijansu (beskonačnu). Tada ocenjivači regresionih koeficijenata dobijeni metodom najmanjih kvadrata mogu imati znatno veću varijansu od ocenjivača metode maksimalne verodostojnosti, jer distribucija odstupanja ima beskonačnu varijansu, pa je tada i varijansa ocenjivača metode najmanjih kvadrata beskonačna.

Za ilustrovanje neefikasnosti ocenjivača metode najmanjih kvadrata koristi se Laplaceova (Laplasova) distribucija od $\varepsilon_i$ definisana sa:

$$f(\varepsilon_i) = \frac{1}{2\phi} e^{-|\varepsilon_i/\phi|}\,, (\phi > 0)$$

Osnovna razlika između ove dve distribucije je ta što Laplaceova sadrži apsolutne vrednosti $\varepsilon_i$, a normalna kvadrate od $\varepsilon_i$. Laplaceova distribucija od $\varepsilon_i$ je simetrična, ima sredinu jednaku nuli i varijansu $2\phi^2$, a njeni krajevi su i manje osetljivi od krajeva pomenute normalne distribucije (A.C. Harvey *The Econometric Analysis of Time Series* (New York: Wiley, 1981) str 114.-115.).




Slika 8-1, *Počela ekonometrije*, Jan Kmenta, str 263.

Ocenjivanje regresionih koeficijenata metodom maksimalne verodostojnosti ekvivalentno je minimiziranju sume apsolutnih vrednosti reziduala:

$$\sum_i |Y_i - \alpha - \beta X_i|$$

Takva minimizacija je problem linearnog programiranja, rešenje sadrži izbor dva između $n$ mogućih opažanja iz uzoraka kroz koja prolazi ocenjena regresiona linija (L.D. Taylor "*Estimation by Minimizing the Sum of Absolute Errors*" i P. Zarembka *Frontiers in Econometrics* (New York: Academic Press, 1974)). Dobijeni ocenjivač od β, je asimptotski efikasan i ima asimptotsku varijansu:

$$Asimpt.Var(\dot{\beta}) = \frac{\phi^2}{\sum(X_i - \bar{X})^2}$$

Asimptotska varijansa ocenjivača metode najmanjih kvadrata od β je:

$$Asimpt.Var(\hat{\beta}) = \frac{2\phi^2}{\sum(X_i - \bar{X})^2}$$

Dakle, dvostruko je veća od $Asimpt.Var(\dot{\beta})$ (A.C. Hervey op. cit. str. 114 – 115 ocenjivač maksimalne verodostojnosti od φ je $\sum_i |Y_i - \dot{\alpha} - \dot{\beta} X_i|/n$).

Ocenjivač regresionih koeficijenata dobijen minimiziranjem sume apsolutnih odstupanja se obeležava sa MAD. Budući da na taj ocenjivač manje utiču ekstremna odstupanja nego na ocenjivač metode najmanjih kvadrata, preporuka je da se koristi u svim slučajevima kada se na krajevima pojavljuje, na primer, zadebljana distribucija odstupanja. Pokazalo se da je ocenjivač MAD regresionog koeficijenta asimptotski nepristrasan i normalno distribuiran, kao da je i njegova asimptotska varijansa manja od varijanse ocenjivača metode najmanjih kvadrata za veliki broj na krajevima zadebljanih distribucija (G. Bassett, Jr., i R. Koenker, *Asymptotic Theory of Least Aboslute Error Regression*, Journal of the American

Statistical Association, 73, str 618.-622.; formula za asimptotsku varijansu se takođe nalazi u ovom članku). Kada je reč o normalnoj distribuciji $\varepsilon_i$ ocenjivač MAD je neefikasan jer je njegova varijansa 57% veća od varijanse ocenjivača dobijenih metodom najmanjih kvadrata.

Na osnovu prethodnog navedenog, može se zaključiti da se ocenjivači metode najmanjih kvadrata ponašaju vrlo dobro kada je odstupanje normalno distribuirano i mogu se ponašati veoma slabo kada odstupanje nije normalno distribuirano. U pokušaju konstruisanja ocenjivača koji bi se ponašali mnogo bolje od ocenjivača metode najmanjih kvadrata kada je distribucija odstupanja na krajevima zadebljana i koji bi se ponašali gotovo kao ocenjivači LSE kada je distribucija normalna, došlo se do takozvanih *jakih ocenjivača*. Najpoznatiji su M ocenjivači (za ocenjivače maksimalno verodostojnog tipa) dok se ocenjivači najmanjih kvadrata dobijaju minimiziranjem $\sum_i (Y_i - \alpha - \beta X_i)^2$, tačnije minimiziranjem:

$$\sum_i f(Y_i - \alpha - \beta X_i)$$

gde je $f(Y_i - \alpha - \beta X_i)$ neka funkcija od $(Y_i - \alpha - \beta X_i)$ koja bi dala ocenjivače sa prethodno navedenim osobinama jakosti, zamisao na kojoj se temelje ocenjivači M. Peter J. Huber (Huber) je preporučio jednu takvu funkciju, ona se često navodi u statističkoj literaturi i uključuje upotrebu ocenjivanja metodom najmanjih kvadrata za sva odstupanja čija je apsolutna vrednost manja ili jednaka od prethodno određene vrednosti. On savetuje upotrebu ocenjivanja MAD za sva odstupanja veća od te vrednosti (P.J. Huber, *Robust Statistics*). Kada je ovaj broj beskonačan, onda se metoda svodi na ocenjivanje metodom najmanjih kvadrata. Za ostale vrednosti, neophodne su obe metode. Preporuka nekih autora koji su se bavili ovom temom je i da može jednostavno da se isključe sva opažanja koja odgovaraju velikim odstupanjima, a da se na ostale primeni metoda najmanjih kvadrata. Ovaj postupak je poznatiji kao „uređeno" ocenjivanje. Poteškoća kod ovog ocenjivanja je kako proceniti koja opažanja treba izostaviti, dok je u slučaju jakih ocenjivača, koje preporučuje Huber, izbor konstante. Takođe, baratanje ekstremnim odstupanjima na način da se zamagli njihov uticaj manje je povoljno od objašnjavanja njihove prisutnosti. Kad neobični događaji, kao što su na primer nenormalni vremenski uslovi, uzrokuju ekstremna odstupanja, tada je poželjnije primeniti postupak koji bi te događaje uključio kao kvalitetne promenljive u jednačinu regresije. Upravo je sve ovo gore navedeno razlog zašto, iako se može primeniti, nije naročito zastupljeno u statistici.

### 2.1.2. Testiranje normalnosti

Kada je normalnost odstupanja sumnjiva, najpraktičnije je izračunati ocene regresionih koeficijenata metodom najmanjih kvadrata i minimiziranjem sume apsolutnih odstupanja. Ukoliko se ove dve ocene značajno razlikuju, onda je sledeći korak identifikacija onog opažanja koji odgovara ekstremnim odstupanjima i ispitivanje njegovog uzroka. Ovo se detaljnije obrađuje u sledećem primeru.

*Primer 2.1*

| Cena: RSD/kg | Količina: kg |
|---|---|
| **100** | 55 |
| **90** | 70 |
| **80** | 90 |
| **70** | 100 |
| **70** | 90 |
| **70** | 105 |
| **70** | 80 |
| **65** | 110 |
| **60** | 125 |
| **60** | 115 |
| **55** | 130 |
| **50** | 130 |

U datoj tabeli su podaci o ceni i količini prodatih narandži u jednoj prodavnici za 12 uzastopnih dana. Neka $X_i$ bude cena, a $Y_i$ prodata količina $i$-tog dana. Funkcija potražnje je dakle sledećeg oblika:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Neka je funkcija takva da zadovoljava osnovne pretpostavke. Ocenu za $\alpha$ i $\beta$ dobijamo metodom najmanjih kvadrata:

$$\bar{X} = 70\ , \sum(X_i - \bar{X})\,(Y_i - \bar{Y}) = -3550$$

$$\bar{Y} = 100\ , \sum(X_i - \bar{X})^2 = 2250, \sum(Y_i - \bar{Y})^2 = 6300$$

$$\hat{\beta} = \frac{-3550}{2250} = -1.578$$

$$\hat{\alpha} = 100 - (-1.578) \times 70 = 210.460$$

i odatle je ocenjena linija regresije $\hat{Y}_i = 210.460 - 1.578X_i$

Model sada izgleda ovako:

$$Y_i = 210.460 - 1.578X_i + \varepsilon_i$$

Kada se izračunaju ocene minimiziranjem sume apsolutnih odstupanja, dobija se:

$$Y_i = 205 - 1.5X_i + \tilde{\varepsilon}_i$$

Može se primetiti da su ocene slične i u tom slučaju se može zaključiti da nema problema sa normalnošću.

Normalnost distribucije podataka je uslov za sprovođenje mnogih statističkih tehnika. Da li podaci imaju normalnu raspodelu, može da se proveri na više načina. Ono što je uobičajeno je da se kombinuju grafički pokazatelji i testovi. Od grafičkih pokazatelja to su: histogram, Box plot i normalni i detrendovani Q-Q plot. Od testova koriste se Kolmogorov - Smirnov za uzorke veće od 50 ili Shapiro -Wilk test za manje uzorke. Osim navedenih, koristi se i inspekcija deskriptivnih pokazatelja distribucije podataka (AS, SD, skewness i kurtosis). Normalnost podataka u regresionoj analizi se proverava preko raspodele reziduala. U sklopu procedure linearne regresije, većina softvera proizvodi dijagram rasturanja reziduala. Ukoliko su reziduali normalno raspodeljeni oko predviđenih vrednosti zavisne promenljive, smatra se da je pretpostavka o normalnosti ispunjena.

### 2.1.3. Tretmani otklanjanja narušene normalnosti

Kada podaci nemaju normalnu distribuciju, ovaj problem može da se reši pomoću neke od nelinearnih transformacija, a obično se koristi Log-transformacija. Drugi način je korenovanjem podataka. U ovom radu koristiće se Log-transformacija kako bi se ispravila narušena pretpostavka o normalnosti. Nakon toga, može se primeniti regresiona analiza. Treća mogućnost je korišćenjem neparametarskih testova, ali se time ovaj rad neće baviti.

## 2.2. Sredina različita od nule

Druga pretpostavka klasičnog regresionog modela jeste da je sredina odstupanja jednaka nuli. Ako je regresiona linija populacije jednaka $E(Y_i) = \alpha + \beta X_i$, a sredina odstupanja nije jednaka nuli nego na primer $\mu_i$, narušavanje ove pretpostavke se tada zapisuje:

$$E(Y_i) = \alpha + \beta X_i + \mu_i$$

Neophodno je razlikovati slučaj kada $\mu_i$ ima istu vrednost za sva opažanja i slučaja kada su ove vrednosti različite.

a) Kada se obeleži ista vrednost za sva opažanja sa $\mu_i = \mu$ , regresiona linija populacije izgleda:

$$E(Y_i) = \alpha + \mu + \beta X_i$$

$$E(Y_i) = \alpha^* + \beta X_i$$

$$Y_i = \alpha^* + \beta X_i + \varepsilon_i{}^*$$

gde je $\alpha^* = \alpha + \mu$, $\varepsilon_i{}^* = \varepsilon_i - \mu$, i sredina od $\varepsilon_i{}^*$ je nula.

Iako ovo ne utiče na ocenjivač metode najmanjih kvadrata od $\beta$, jasno je da tada formula metode najmanjih kvadrata za ocenjivanje odsečka daje ocenu od $\alpha^*$, a ne od $\alpha$. Sve dok je $\varepsilon_i$ normalno distribuirana promenljiva ne postoji način na koji bi se odvojeno ocenili $\alpha$ i $\mu$ kako bi se dobile konzistentne ocene.

*Primer 2.2.* (*Granična proizvodna funkcija*)

Na ovom primeru biće pokazana situacija u kojoj sredina od $\varepsilon_i$ nije jednaka nuli. Najjednostavnija verzija granične proizvodne funkcije može se zapisati:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \text{ gde je } \varepsilon_i \leq 0$$

Y predstavlja logaritam količine proizvodnje, X logaritam količine utroška faktora gde se $i$ odnosi na $i$-to preduzeće. Ukoliko se proizvodna funkcija odnosi na maksimalnu količinu proizvodnje za datu količinu utroška, postojanje odstupanja ima učinak na zavisnu promenljivu. Može biti ili negativan ili jednak nuli, ali nikako pozitivan. Posledica ovoga je da je sredina od $\varepsilon_i$ negativna. Ako se pretpostavi da je ova sredina jednaka za svako preduzeće, ona se može interpretirati kao mera prosečne efikasnosti industrije. Kako je odstupanje ograničeno odozgo, ono ne može biti normalno distribuirano. Međutim, ukoliko su zadovoljene tri preostale osnovne pretpostavke, ocenjivač najmanjih kvadrata za $\beta$ će imati sva poželjna

asimptotska svojstva. Konzistentna ocena od $\alpha$ i sredina od $\varepsilon_i$ mogu se izvesti uz određene pretpostavke o distribuciji $\varepsilon_i$ (F.R. Forsund, C.A.K. Lovell i P. Schmidt, *A Survey of Frontier Production Functions and of Their Relationship to Efficiency Measurement* Journal of Econometrics, 13, str 5.-25.).

b) Neka su različite vrednosti opažanja obeležene sa $(\alpha + \mu_i)$. Ovo implicira da se srednja vrednost zavisne promenljive $E(Y_i)$ menja ne samo zbog promene nezavisne promenljive $X_i$ već i zbog promene vrednosti $\alpha + \mu_i$.

U ovom slučaju veza između $X_i$ i $Y_i$ je jasno precizirana. Ovo se može dogoditi kada na $E(Y_i)$ ne utiče samo $X_i$ već i neka druga nestohastička promenljiva $Z_i$. To znači da model zapravo izgleda ovako:

$$Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i{}^*$$

gde $\varepsilon_i{}^*$ zadovoljava sve osnovne pretpostavke dok se ocenjuje sledeće:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \,, \quad (\mu_i = \gamma Z_i)$$

### 2.2.2. Testiranje pretpostavke i tretman

Za dijagnostiku ove pretpostavke nema jednostavnog, a efikasnog načina provere. Jedan od pristupa je da se izračunaju reziduali za sve ispitanike, a zatim ispita korelacija između reziduala i ispitanika. Još jedan od pristupa je ispitivanje *intra-class* koeficijenta korelacije, ali nema mnogo statističkih paketa koji će ga izračunati bez posebnog programiranja. Prilikom sprovođenja regresione analize mogu se ispitati dijagrami distribucije reziduala radi provere da li se reziduali distribuiraju nezavisno i da li imaju normalnu raspodelu.

## 2.3. Heteroskedastičnost

Treća pretpostavka klasičnog modela govori o tome da svako odstupanje ima istu varijansu, odnosno:

$$Var(\varepsilon_i) = \sigma^2, \text{ za svako } i$$

Ova pretpostavka se naziva homoskedastičnost. Kako se pretpostavlja da je sredina od $\varepsilon_i$ jednaka nuli, prethodna jednačina se može zapisati i na sledeći način:

$$E(\varepsilon_i{}^2) = \sigma^2$$

Odstupanje od ove osobine nazivamo heteroskedastičnost. To je pojava koja se najčešće javlja u podacima preseka. Ovi podaci predstavljaju određene vrednosti prikupljene u jednom trenutku vremena za različite jedinice posmatranja (pojedinci, domaćinstva, preduzeća, industrije, geografske oblasti i slično). Elementi u uzorcima ovog tipa mogu biti različitih veličina kao što su male, srednje ili velike firme ili nizak, srednji ili visoki prihod.

Važno je napomenuti, preliminarno, da kada je reč o prostornim uzorcima (posebno sa regionalnim podacima), heteroskedastičnost je uobičajena pojava zbog prirode prikupljanja podataka. Očigledni izvori heteroskedastičnosti su povezani sa različitim dimenzijama za različite regione u istraživanom području, neravnopravne koncentracije stanovništva i privredne aktivnosti u ruralnim i urbanim područjima.

Dakle, kada je odstupanje od regresione linije heteroskedastično, odnosno kada se varijansa odstupanja razlikuje u opažanjima, tada je:

$$E(\varepsilon_i{}^2) = \sigma_i{}^2$$

Potrebno je ispitati kako ovo ponašanje varijanse utiče na svojstva ocenjivača koeficijenata regresije koji se dobijaju metodom najmanjih kvadrata.

### 2.3.1. Ispitivanje osobina ocenjivača u uslovima heteroskedastičnosti

Ocenjivač metode najmanjih kvadrata za $\beta$ je izveden u prethodnom poglavlju, on je jednak:

$$\hat{\beta} = \frac{n\sum(X_i - \bar{X})\,(Y_i - \bar{Y})}{n\sum(X_i - \bar{X})^2} = \beta + \frac{\sum(X_i - \bar{X})\,\varepsilon_i}{\sum(X_i - \bar{X})^2}$$

a očekivana vrednost od $\hat{\beta}$ je:

$$E(\hat{\beta}) = \beta + E\left(\frac{\sum(X_i - \bar{X})\,\varepsilon_i}{\sum(X_i - \bar{X})^2}\right) = \beta$$

Takođe je već izvedena formula za ocenjivač od $\alpha$, a to je:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = (\alpha + \beta\bar{X} + \bar{\varepsilon}) - \hat{\beta}\bar{X}$$

i očekivana vrednost od $\hat{\alpha}$ je:

$$E(\hat{\alpha}) = \alpha + \beta\bar{X} + E(\bar{\varepsilon}) - E(\hat{\beta})\bar{X} = \alpha$$

Dolazi se do zaključka da su ocenjivači za $\alpha$ i $\beta$ nepristrasni i u uslovima heteroskedastičnosti.

Sada bi trebalo proveriti da li su dobijeni ocenjivači i najbolji linearni nepristrasni ocenjivači. Kada se izvedu formule BLUE u slučaju heteroskedastičnosti i ukoliko postoji razlika između njih i ocenjivača dobijenih metodom najmanjih kvadrata, tada oni nisu BLUE.

Prvo se mora transformisati regresiona jednačina tako da odstupanje bude homoskedastično. Ovo se postiže deljenjem obe strane početne jednačine sa $\sigma_i$:

$$\frac{Y_i}{\sigma_i} = \alpha\left(\frac{1}{\sigma_i}\right) + \beta\left(\frac{X_i}{\sigma_i}\right) + \frac{\varepsilon_i}{\sigma_i} \quad \text{(T3)}$$

odnosno

$$Y_i{}^* = \alpha W_i^* + \beta X_i^* + \varepsilon_i{}^* \qquad \text{(T3*)}$$

gde je $Y_i{}^* = \frac{Y_i}{\sigma_i}$, $W_i^* = \frac{1}{\sigma_i}$, $X_i^* = \frac{X_i}{\sigma_i}$ i $\varepsilon_i{}^* = \frac{\varepsilon_i}{\sigma_i}$. Trebalo bi uočiti da sada tranformisana jednačina ima dve determinističke promenljive $W^*$ i $X^*$, i nema odsečak. Kada bi se računale njihove vrednosti, neophodna bi bila vrednost $\sigma_i$. Dalje se računa:

$$E(\varepsilon_i{}^*) = \frac{E(\varepsilon_i)}{\sigma_i} = 0$$

$$Cov(\varepsilon_i{}^*\varepsilon_j{}^*) = \frac{E(\varepsilon_i\varepsilon_j)}{\sigma_i\sigma_j} = 0$$

$$Var(\varepsilon_i{}^*) = Var\left(\frac{\varepsilon_i}{\sigma_i}\right) = \frac{Var(\varepsilon_i)}{\sigma_i{}^2} = 1$$

U jednačini (T3*) su nezavisne nestohastičke promenljive, odstupanja jednaka nuli, nisu u korelaciji i imaju konstantnu varijansu, ispunjeni su svi preduslovi za jednakost ocenjivača dobijanih metodom najmanjih kvadrata (LSE) i najboljih linearnih nepristrasnih ocenjivača (BLUE) od $\alpha$ i $\beta$.

Dalje se primenjuju načela metode najmanjih kvadrata, n1 i n2, i dobijaju se sledeće normalne jednačine:

$$\sum W_i{}^*Y_i{}^* = \tilde{\alpha}\sum W_i{}^{*2} - \tilde{\beta}\sum W_i{}^*X_i{}^*$$

$$\sum X_i{}^*Y_i{}^* = \tilde{\alpha}\sum W_i{}^*X_i{}^* - \tilde{\beta}\sum X_i{}^{*2}$$

Ako se označe u skladu sa (T3):

$$\sum \frac{Y_i}{\sigma_i{}^2} = \tilde{\alpha}\sum \frac{1}{\sigma_i{}^2} + \tilde{\beta}\sum \frac{X_i}{\sigma_i{}^2}$$

$$\sum \frac{X_iY_i}{\sigma_i{}^2} = \tilde{\alpha}\sum \frac{X_i}{\sigma_i{}^2} + \tilde{\beta}\sum \frac{X_i{}^2}{\sigma_i{}^2}$$

Obeležavanjem $\frac{1}{\sigma_i{}^2} = w_i$, dobija se prostiji oblik prethodnih jednačina:

$$\sum w_iY_i = \tilde{\alpha}\sum w_i + \tilde{\beta}\sum w_iX_i$$

$$\sum w_iX_iY_i = \tilde{\alpha}\sum w_iX_i + \tilde{\beta}\sum w_iX_i{}^2$$

Rešavanjem ovih jednačina dobija se:

$$\tilde{\beta} = \frac{(\sum w_i)(\sum w_iX_iY_i) - (\sum w_iX_i)(\sum w_iY_i)}{(\sum w_i)(\sum w_iX_i{}^2) - (\sum w_iX_i)^2}$$

$$= \frac{\sum w_i(X_i - \tilde{X})(Y_i - \tilde{Y})}{\sum W_i\,(X_i - \tilde{X})^2}$$

$$\tilde{\alpha} = \tilde{Y} - \tilde{\beta}\tilde{X}$$

gde je $\tilde{X} = (\sum w_i X_i)/(\sum w_i)$, a $\tilde{Y} = (\sum w_i Y_i)/(\sum w_i)$.

Dobijene formule za najbolje linearne nepristrasne ocenjivače od $\alpha$ i $\beta$ su različite od onih dobijenih metodom najmanjih kvadrata. Dakle, kada je pretpostavka o homoskedastičnosti narušena, ocenjivači LSE nisu BLUE. Odavde kao posledica proizilazi činjenica da dobijeni ocenjivači nemaju najmanju varijansu od svih nepristrasnih ocenjivača i stoga nisu efikasni.

Kada je reč o asimptotskim svojstvima, proverava se konzistentnost u uslovima heteroskedastičnosti, odnosno ispituje se granična vrednost sredine kvadratne greške ocenjivača regresionih koeficijenata dobijenih metodom najmanjih kvadrata kada veličina uzorka teži beskonačnosti. Ukoliko je jednaka nuli, konzistentni su.

Provera konzistentnosti za $\hat{\beta}$, ocenjivača od $\beta$ dobijenog metodom najmanjih kvadrata, polazi od nepristrasnosti, odnosno sledi da je $MSE(\hat{\beta}) = Var(\hat{\beta})$. Iz formule varijanse se dobija:

$$Var(\hat{\beta}) = E(\hat{\beta} - \beta)^2 = E\left(\frac{\sum(X_i - \bar{X})\,\varepsilon_i}{\sum(X_i - \bar{X})^2}\right)^2$$

ali zbog $E({\varepsilon_i}^2) = {\sigma_i}^2$, i $E(\varepsilon_i \varepsilon_j) = 0$, $Var(\hat{\beta})$ je:

$$Var(\hat{\beta}) = \frac{\sum(X_i - \bar{X})^2\,{\sigma_i}^2}{(\sum(X_i - \bar{X})^2)^2}$$

Ukoliko je ${\sigma_i}^2 = \sigma^2$, za svako $i$, ova formula se svodi na klasičnu formulu.

Uvodeći oznake $\theta_i' = {\sigma_i}^2 - \bar{\sigma}^2$, gde je $\bar{\sigma}^2 = \sum {\sigma_i}^2/n$, pa je $\sum \theta_i' = 0$. Tada je:

$$Var(\hat{\beta}) = \frac{\sum(X_i - \bar{X})^2(\bar{\sigma}^2 + \theta_i')}{(\sum(X_i - \bar{X})^2)^2} = \frac{\bar{\sigma}^2}{\sum(X_i - \bar{X})^2} + \frac{\sum(X_i - \bar{X})^2\theta_i'}{(\sum(X_i - \bar{X})^2)^2}$$

$$= \frac{\bar{\sigma}^2/n}{\sum(X_i - \bar{X})^2/n} + \frac{(\sum(X_i - \bar{X})^2\theta_i'/n)(1/n)}{(\sum(X_i - \bar{X})^2/n)^2}$$

gde je $\sum(X_i - \bar{X})^2\theta_i'/n$ kovarijansa iz uzorka između $(X_i - \bar{X})^2$ i ${\sigma_i}^2$. Za $\sum(X_i - \bar{X})^2\theta_i'/n$ sa sigurnošću možemo reći da je konačan kada $n \to \infty$, dokle god je ${\sigma_i}^2 < \infty$ za svako $i$. Kako je $\lim_{n\to\infty}(\bar{\sigma}^2/n) = 0$ odatle se dobija:

$$\lim_{n\to\infty} Var(\hat{\beta}) = 0$$

Dakle, $\hat{\beta}$ je konzistentan i u uslovima heteroskedastičnosti. Analogno se dobijaju identični rezultati za $\hat{\alpha}$.

Da bi se utvrdila asimptotska efikasnost, neophodno je pronaći odgovarajuće maksimalo verodostojne ocenjivače, a potom uporediti varijanse sa onima dobijenim metodom najmanjih kvadrata. Ukoliko nisu jednake, ocenjivači metode najmanjih kvadrata nisu asimptotski efikasni.

Funkcija verodostojnosti u uslovima heteroskedastičnosti stoga izgleda ovako:

$$L = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n} log{\sigma_i}^2 - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{Y_i - \alpha - \beta X_i}{\sigma_i}\right)^2$$

Diferenciranjem $L$ u odnosu na $\alpha$ i $\beta$ dobija se:

$$\frac{\partial L}{\partial \alpha} = \sum_i \left(\frac{Y_i - \alpha - \beta X_i}{\sigma_i}\right)$$

$$\frac{\partial L}{\partial \beta} = \sum_i \left[\frac{X_i(Y_i - \alpha - \beta X_i)}{{\sigma_i}^2}\right]$$

Daljim izjednačavanjem sa nulom i rešavanjem za ocenjivače od $\alpha$ i $\beta$, dolazi se do rešenja koje je identično formuli najboljih linearnih nepristrasnih ocenjivača. Odatle se zaključuje da i njihove varijanse moraju biti jednake. Formule za ove varijanse se izvode na sledeći način. Ako je:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

i

$$\tilde{Y} = \alpha + \beta\tilde{X} + \tilde{\varepsilon}$$

gde je $\tilde{\varepsilon} = \sum w_i\varepsilon_i / \sum w_i$, dobija se sledeće:

$$\tilde{\beta} = \frac{\sum w_i(X_i - \tilde{X})(Y_i - \tilde{Y})}{\sum w_i(X_i - \tilde{X})^2} = \frac{\sum w_i(X_i - \tilde{X})[\beta(X_i - \tilde{X}) + (\varepsilon_i - \tilde{\varepsilon})]}{\sum w_i(X_i - \tilde{X})^2}$$

$$= \beta + \frac{\sum w_i(X_i - \tilde{X})(\varepsilon_i - \tilde{\varepsilon})}{\sum w_i(X_i - \tilde{X})^2} = \beta + \frac{\sum w_i(X_i - \tilde{X})\varepsilon_i}{\sum w_i(X_i - \tilde{X})^2}$$

Dakle, $Var(\tilde{\beta})$ je tada:

$$Var(\tilde{\beta}) = E(\tilde{\beta} - \beta)^2 = E\left(\frac{\sum w_i(X_i - \tilde{X})\varepsilon_i}{\sum w_i(X_i - \tilde{X})^2}\right)^2 = \frac{\sum {w_i}^2(X_i - \tilde{X})^2{\sigma_i}^2}{\left[\sum w_i(X_i - \tilde{X})^2\right]^2}$$

$$= \frac{\sum w_i(X_i - \tilde{X})^2}{\left[\sum w_i(X_i - \tilde{X})^2\right]^2} = \frac{\sum w_i}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2} = \frac{1}{\sum w_i(X_i - \tilde{X})^2}$$

Na sličan način se dolazi i do formule za varijansu BLUE od $\alpha$ u slučaju heteroskedastičnosti:

$$Var(\tilde{\alpha}) = \frac{\sum w_i X_i^2}{\sum w_i \sum w_i X_i^2 - (\sum w_i X_i)^2} = \frac{1}{\sum w_i} + \frac{X_i^2}{\sum w_i(X_i - \tilde{X})^2}$$

Može se uočiti da su izvedene formule za $Var(\tilde{\alpha})$ i $Var(\tilde{\beta})$ iste kao i kod klasičnog modela kada je $w_i = w = 1/\sigma^2$ za svako $i$.

Potrebno je još ispitati asimptotsku efikasnost ocenjivača od $\alpha$ i $\beta$ dobijenih metodom najmanjih kvadrata u slučaju prisustva heteroskedastičnosti. Ovo se rešava upoređivanjem asimptotskih varijansi sa varijansama maksimalno verodostojnih ocenjivača. Varijansa od $\beta$ dobijena metodom najmanjih kvadrata je:

$$Var(\hat{\beta}) = \frac{\sum(X_i - \bar{X})^2 \sigma_i{}^2}{(\sum(X_i - \bar{X})^2)^2}$$

I različita je od varijanse maksimalno verodostojnih ocenjivača:

$$Var(\tilde{\beta}) = \frac{1}{\sum w_i (X_i - \tilde{X})^2}$$

bez obzira na veličinu uzorka. Do sličnog zaključka se dolazi i upoređivanjem ocenjivača za $\alpha$. Prema tome, kako se ove vrednosti varijansi razlikuju, može se zaključiti da ocenjivači dobijeni metodom najmanjih kvadrata nisu asimptotski efikasni u uslovima heteroskedastičnosti.

Zaključak: Ocene parametara $\alpha$ i $\beta$ dobijene metodom najmanjih kvadrata u uslovima heteroskedastičnosti su nepristrasne i konzistentne, ali nemaju osobinu efikasnosti i asimptotske efikasnosti, a ocena varijanse je pristrasna.

### 2.3.2. Oblici heteroskedastičnosti

Najčešći oblici heteroskedastičnosti na koje se nailazi u literaturi, a i u praksi, jesu multiplikativna i aditivna heteroskedastičnost.

#### *2.3.2.1. Multiplikativna heteroskedastičnost*

Multiplikativna heteroskedastičnost se primenjuje češće, a njen opšti oblik zapisujemo na sledeći način:

$$log\sigma_i^2 = log\sigma^2 + \delta_1 logZ_{i1} + \delta_2 logZ_{i2} + \cdots + \delta_p logZ_{ip}$$

gde je prikazana zavisnost $\sigma_i^2$ od $p$ promenljivih. Ovakav zapis je pogodan kod višestruke regresije, gde su $Z_s$, $s = 1, \ldots, p$ nezavisne promenljive u regresijskom modelu. Kada je reč o jednostrukoj regresiji, tada je:

$$log\sigma_i^2 = log\sigma^2 + \delta logZ_i$$

$$\sigma_i^2 = \sigma^2 Z_i^{\delta}$$

gde su $\sigma^2$ i $Z_i^{\delta}$ parametri. Jačina heteroskedastičnosti zavisi najviše od δ. Što je vrednost δ manja, manje su razlike pojedinačnih varijansi. Za $\delta = 0$, model je homoskedastičan.

Kada su oba pomenuta parametra nepoznata, potrebno ih je oceniti zajedno sa regresijskim koeficijentima $\alpha$ i $\beta$. Postoji i mogućnost da se vrednost jednog parametra fiksira, najčešće se postavi da je $\delta = 2$, što dovodi do toga da je standardna devijacija odstupanja proporcionalna promenljivoj $Z_i$.

Regresijski model u uslovima heteroskedastičnosti izgleda ovako:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma_i^2),$$

$$\sigma_i^2 = \sigma^2 Z_i^{\delta}, (\delta > 0; Z_i > 0)$$

uz pretpostavke da su $\varepsilon_i$ nekorelisane, a $X_i$ i $Z_i$ nestohastičke promenljive. Dalje je neophodno primeniti funkciju maksimalne verodostojnosti za $\sigma_i^2 = \sigma^2 Z_i^{\delta}$, kako bi se dobili ocenjivači od $\alpha$, $\beta$, $\sigma^2$ i $\delta$.

$$L = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(log\sigma^2 + \delta log Z_i) - \frac{1}{2}\sum_{i=1}^{n}\left[\frac{Y_i - \alpha - \beta X_i}{\sigma Z_i^{\delta/2}}\right]^2$$

Diferenciranjem po $\alpha$, $\beta$, $\sigma^2$ i $\delta$ dobija se:

$$\frac{\partial L}{\partial \alpha} = \frac{1}{\sigma^2}\sum_i\left[\frac{Y_i - \alpha - \beta X_i}{Z_i^{\delta}}\right]$$

$$\frac{\partial L}{\partial \beta} = \frac{1}{\sigma^2}\sum_i\left[\frac{(Y_i - \alpha - \beta X_i)X_i}{Z_i^{\delta}}\right]$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i\left[\frac{Y_i - \alpha - \beta X_i}{Z_i^{\delta/2}}\right]^2$$

$$\frac{\partial L}{\partial \delta} = -\frac{1}{2}\sum_i logZ_i + \frac{1}{2\sigma^2}\sum_i\left[\frac{(Y_i - \alpha - \beta X_i)^2 logZ_i}{Z_i^{\delta}}\right]$$

Izjednačavanjem sa nulom dalje se dobijaju četiri jednačine, sa četiri nepoznate, koje su nelinearne zbog $\delta$. Jednostavnije rešenje se dobija kada se izjednače prve tri jednačine sa nulom, a potom za određene vrednosti $\alpha$, $\beta$ i $\sigma^2$ izračuna $L$:

$$\sum_i\left(\frac{Y_i}{Z_i^{\delta}}\right) = \bar{\bar{\alpha}}\sum_i\left(\frac{1}{Z_i^{\delta}}\right) + \bar{\bar{\beta}}\sum_i\left(\frac{X_i}{Z_i^{\delta}}\right)$$

$$\sum_i\left(\frac{X_i Y_i}{Z_i^{\delta}}\right) = \bar{\bar{\alpha}}\sum_i\left(\frac{X_i}{Z_i^{\delta}}\right) + \bar{\bar{\beta}}\sum_i\left(\frac{{X_i}^2}{Z_i^{\delta}}\right)$$

$$\bar{\bar{\sigma}}^2 = \frac{1}{2}\left(\sum_i\frac{X_i}{Z_i^{\delta}} - \bar{\bar{\alpha}}\sum_i\frac{1}{Z_i^{\delta}} - \bar{\bar{\beta}}\sum_i\frac{X_i}{Z_i^{\delta}}\right)^2$$

Za većinu ekonomskih modela, $\delta$ uzima vrednosti između 0 i 3 ili 4, pa se potom traže ocene za $\alpha$, $\beta$ i $\sigma^2$ kada je $\delta = 0$, zatim $\delta = 0.1\ldots$ i tako do 3 ili 4. Zatim se računa $L$ za sve ove vrednosti i traži se onaj slučaj kada je $L$ maksimalno. Ovaj način traženja maksimalne funkcije verodostojnosti se zove još i

metoda traženja, koji je najjednostavniji i sprečava mogućnost dobijanja lokalnog maksimuma umesto globalnog (A. C. Harvey, *Estimating Regression Models with Multipilactive Heteroscedasticity*, Econometrica, 44, str 461.-465.). Maksimalno verodostojni ocenjivači imaju sva poželjna asimptotska svojstva.

Najveći problem u modelu multiplikativne heteroskedastičnosti je izbor promenljive $Z$. Mala je verovatnoća da je poznata promenljiva povezana sa varijansom odstupanja koja nije uključena u regresijsku jednačinu. Stoga se najčešće uzima nezavisna promenljiva $X$ kao izbor za $Z$. Specijalni slučaj multiplikativne heteroskedastičnosti je tada za $Z_i = X_i$ i $\delta = 2$ jednak:

$$\sigma_i^2 = \sigma^2 X_i^2$$

uz pretpostavku je da je standardna devijacija regresijskog odstupanja proporcionalna vrednostima nezavisne promenljive.

#### *2.3.2.2. Aditivna heteroskedastičnost*

Aditivna heteroskedastičnost je drugi oblik heteroskedastičnosti. Opšti oblik se zapisuje:

$$\sigma_i^2 = a_0 + a_1 Z_{i1} + a_2 Z_{i2} + \cdots + a_p Z_{ip}$$

a za $p = 2$, $Z_{i1} = X_i$ i $Z_{i2} = X_i^2$ se dobija:

$$\sigma_i^2 = a + bX_i + cX_i^2$$

gde su $a$, $b$ i $c$ konstante čije se vrednosti pretpostavljaju ili ocenjuju. U slučaju kada su $b$ i $c$ jednake nuli, model je homoskedastičan, a kada su $a$ i $b$ jednaki nuli dobija se specijalan slučaj multiplikativne heteroskedastičnosti. Dakle, ovaj oblik ima manja ograničenja u odnosu na prethodni jer pruža mogućnost homoskedastičnosti, kao i dva modela heteroskedastičnosti.

Heteroskedastičnost zavisne promenljive je specijalni slučaj aditivne heteroskedastičnosti. Pretpostavka je da je varijansa odstupanja proporcionalna kvadratnoj sredini od $Y_i$, $\sigma_i^2 = \sigma^2[E(Y_i)]^2 = \sigma^2(\alpha + \beta X_i)^2$.

U jednoj empirijskoj analizi potrošnje čaja u zavisnosti od dohotka i veličine domaćinstva, utvrđen je ovaj oblik varijanse (Z. Mladenović, *Uvod u ekonometriju*). Sva domaćinstva su bila grupisana po veličini i dohotku, pa se variranje potrošnje unutar grupe pripisuje dejstvu ostalih faktora. Rezultat ove analize je da standardna devijacija potrošnje čaja unutar grupe $i$ odgovara $\sigma_i$. One su približno proporcionalne aritmetičkoj sredini potrošnje čaja u svakoj grupi, a kako one odgovaraju očekivanim vrednostima zavisne promenljive, došlo se do pretpostavke o proporcionalnosti standardne devijacije i očekivane vrednosti zavisne promenljive.

### 2.3.3. Testiranje heteroskedastičnosti

U ovom radu je diskutovano o različitim posledicama heteroskedastičnosti u modelu linearne regresije. Ispitivale su se osobine ocenjivača dobijenih metodom najmanjih kvadrata, kao i osobine alternativnih

ocenjivača heteroskedastičnih modela. Međutim, ukoliko je nepoznato da li je model heteroskedastičan ili ne, može se izvesti testiranje hipoteze:

$$H_0 = \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2 \qquad (m \leq n)$$

gde je $m$ broj različitih vrednosti varijanse, naspram alternativne da $H_0$ nije istinita.

Heteroskedastičnost smanjuje preciznost koeficijenata i teži da „proizvede“ manju p-vrednost i time navodi na pogrešan zaključak da je model statistički značajan kada zapravo ne postoji statistička značajnost. Jedan od najjednostavnijih metoda za ispitivanje postojanja homoskedastičnosti odnosno pojave heteroskedastičnosti, sastoji se u vizuelnom pregledu reziduala ocenjenog modela. Uobičajeno je da se formiraju scatterplot dijagrami rasturanja tačaka reziduala ili njihove apsolutne vrednosti i nezavisne promenljive. Na osnovu dijagrama donosi se zaključak o tome da li heteroskedastičnost postoji odnosno kako se generiše varijansa slučajne greške. Za otkrivanje prisustva heteroskedastičnosti koriste se i testovi od kojih su najpoznatiji Goldfeld-Quandtov (Goldfled Kvantov) test, Breusch-Pagan-Godfrey (Brojš-Pagan-Godfri) test, Whiteov (Vajtov) test.

#### *2.3.3.1. Goldfeld-Quandtov test*

Goldfeld-Quandtov test je jedan od najstarijih testova i veoma se često koristi zbog svoje jednostavnosti. Test se temelji na ideji da je varijansa odstupanja jednog dela opažanja iz uzorka generisana u uslovima homoskedastičnosti, odnosno kada je $H_0$ istinita. Odgovarajuće varijanse uzorka će se tada razlikovati zbog kolebanja u uzorku. Zbog toga se test homoskedastičnosti svodi na test jednakosti dveju varijansi. Kako svaka varijansa ima hi-kvadrat raspodelu kada je $H_0$ istinita, njihov odnos ima F statistiku pod uslovom da su dve varijanse uzorka nezavisne. Dakle, potrebno je oceniti dve regresijske jednačine, po jednu za svaki deo opažanja uzorka. Goldfeld-Quandtova test veličina je tada:

$$\frac{s_1^2}{s_2^2} \sim F_{n_2-2, n_1-2}$$

gde su:

$$s_1^2 = \frac{\sum(Y_i - \hat{\alpha}_1 - \hat{\beta}_1 X_i)^2}{n_1 - 2} \qquad (i = 1,2, \dots n_1)$$

$$s_2^2 = \frac{\sum(Y_i - \hat{\alpha}_2 - \hat{\beta}_2 X_i)^2}{n_2 - 2} \qquad (i = n_1 + p + 1, n_1 + p + 2, \dots n_1 + p + n_2)$$

$n_1$ je broj opažanja iz prvog dela uzorka, $n_2$ iz drugog, a $p$ je broj opažanja iz sredine koja nisu uključena ni u jedan deo. Ovaj test je egzaktan ali ne jako moćan, odnosno ima veoma veliku verovatnoću prihvatanja $H_0$ kada je pogrešna, kad su odstupanja heteroskedastična i kada se prosečna varijansa iz prvog dela uzorka ne razlikuje mnogo od varijanse drugog dela. Iz ovog razloga, Goldfeld-Quandtov test se uglavnom preporučuje u okolnostima gde se opažanja mogu poređati u rastući niz prema varijansi. To je jednostavno kada je varijansa povezana sa vrednostima nezavisne promenljive $X$. Dalje, ako se opažanja poređaju i podele na dva jednaka ili približno jednaka dela, varijanse nekoliko poslednjih odstupanja prve polovine verovatno bi bile slične prvih nekoliko odstupanja u drugoj polovini čak i u

uslovima heteroskedastičnosti. Zato je poželjno izostaviti $p$ srednjih opažanja. Eksperimentalni rezultati pokazuju da je razumno ostaviti jednu šestinu srednjih opažanja. Ovo je nezadovoljavajući aspekt testa, jer se izborom $p$ otvara mogućnost prilagođavanja rezultata nečijim željama.

#### *2.3.3.2. Breusch-Pagan-Godfrey test*

Drugi test homoskedastičnosti je Breusch-Pagan-Godfrey test (BPG test). On se temelji na činjenici da ocene regresijskih koeficijenata dobijene metodom najmanjih kvadrata ne bi trebalo značajno da se razlikuju od ocena dobijenih metodom maksimalne verodostojnosti ako je hipoteza o homoskedastičnosti istinita (T.S. Breuch i A.R. Pagan, *A Simple Test for Heteroskedasticity and Random Coefficient Variation*). Konkretno, kad je $L$ funkcija verodostojnosti, koja uzima u obzir heteroskedastičnost, prvi izvodi funkcije $L$ su jednaki nuli, kada se nepoznati parametri zamene njihovim maksimalno verodostojnim ocenama. Ako se, umesto toga, nepoznati parametri zamene ocenama dobijenim metodom najmanjih kvadrata i ako su odstupanja homoskedastična, tada se prvi izvodi funkcije $L$ ne bi trebalo značajno razlikovati od nule. Dakle, testira se hipoteza $H_0$ nasuprot alternativnoj hipotezi $H_A$:

$$H_A\colon \sigma_i^2 = g\left(\gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_p Z_{ip}\right) \qquad (i = 1,2,\dots,n)$$

gde je $g$ neprekidna funkcija koja ima neprekidne prve izvode. Promenljive $Z$ su neke poznate nestohastičke promenljive. One su obično jednake nezavisnim promenljivama regresijske jednačine ili neke poznate funkcije nezavisnih promenljivih. Funkcija $g$ je dovoljno opšta da obuhvata i multiplikativnu i aditivnu heteroskedastičnost kao specijalne slučajeve. Testiranje se vrši u nekoliko koraka:

1. Model linearne regresije se ocenjuje metodom najmanjih kvadrata kako bi se dobili reziduali
2. Ocenjena vrednosti varijanse se potom računa na sledeći način

   $$\hat{\sigma}^2 = \sum \frac{e_i^2}{n}$$

   a to je ocena varijanse dobijena metodom maksimalne verodostojnosti.

3. Zatim se kvadriraju reziduali i podele ocenjenom varijansom

   $$p_i = \frac{e_i^2}{n}$$

4. Metoda najmanjih kvadrata se sada primenjuje na novu jednačinu linearne regresije:

   $$p_i = \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_p Z_{ip} + v_i$$

   gde su $Z_{ij}$ već pomenute promenljive.

5. Na kraju još treba formirati statistiku testa kao polovinu regresijske sume kvadrata modela dobijenog u četvrtom koraku, označenu sa $SSR_{BP}$, koja pod uslovima normalne distribucije grešaka ima $\chi^2$ raspodelu sa brojem stepeni slobode jednakim broju ocenjenih parametara.

$$\frac{\mathrm{SSR}_{\mathrm{BP}}}{2} \sim \chi_p^2$$

Ukoliko su vrednosti test statistike veće od tabličnih, nulta hipoteza se odbacuje.

#### *2.3.3.3. Whiteov test*

Breusch-Pagan-Godfrey test je kritikovan zbog svoje osetljivosti na minimalna narušavanja pretpostavke o normalnosti regresijskog odstupanja. Malom promenom test veličine, ta zavisnost o normalnosti se može otkloniti. Nakon ovih izmena, BPG test postaje ekvivalentan testu koji je preporučio White.

Whiteov test se bazira na poređenju varijanse ocenjivača dobijenih metodom najmanjih kvadrata u uslovima homoskedastičnosti i heteroskedastičnosti. Ako je nulta hipoteza tačna, dve bi se ocenjene varijanse razlikovale samo zbog kolebanja u uzorku. Nulta hipoteza o homoskedastičnosti slučajne greške testira se protiv široko postavljene alternativne hipoteze da je varijansa slučajne greške zavisna od objašnjenih promenljivih, njihovih kvadrata i međuproizvoda tj. ispituje se variranje reziduala pod kombinovanim dejstvom regresora. Test se sastoji od sledećih koraka:

a) Model linearne regresije se oceni kako bi se dobili reziduali $e_i$, odnosno njihove kvadrirane vrednosti
b) Metoda najmanjih kvadrata se primenjuje na sledeću jednačinu:

$$e_i{}^2 = \delta_0 + \delta_1 Z_{i1} + \delta_2 Z_{i2} + \cdots + \delta_p Z_{ip} + u_i\,, \quad i = 1,2,\ldots n$$

gde je za prostu regresiju $p = 2$, $Z_{i1} = X_i$ i $Z_{i2} = X_i^2$, pa se test zapravo zasniva na analizi sledeće jednačine:

$$e_i{}^2 = \delta_0 + \delta_1 X_i + \delta_2 X_i^2 + u_i\,, \quad i = 1,2,\ldots n$$

Kao ni u prethodna dva testa, nije neophodna specifikacija oblika heteroskedastičnosti.

Za razliku od BPG testa, Whiteov test nije osetljiv na odstupanje grešaka od normalnosti i jednostavniji je, pa se i češće upotrebljava u testiranju heteroskedastičnosti. U slučaju da postoji više regresora, uvođenje kvadrata i svih međuproizvoda u pomoćnu regresiju može značiti veliki gubitak u broju stepeni slobode. Zato se često Whiteov test izvodi bez međuproizvoda. Ukoliko rezultat testa pokaže značajno visoku statistiku, razlog ne mora biti heteroskedastičnost, već greška u specifikaciji modela (izostavljeni regresor). Dakle, Whiteov test se može koristiti u otkrivanju heteroskedastičnosti ili greške specifikacije.

Potrebno je napomenuti i to da je Whiteov test često neuspešan za male obime uzoraka. Takođe, ako rezultat testa govori da bi trebalo odbaciti nultu hipotezu, odatle se ne može dobiti informacija o tome šta je prouzrokovalo heteroskedastičnost. To može dovesti do problema pri pokušaju konstruisanja generalizovane metode najmanjih kvadrata za ocenjivanje parametara, za koji je potrebno znati oblik greške.

### 2.3.4. Tretman za otklanjanje heteroskedastičnosti

Heteroskedastičnost se može otkloniti na više načina u zavisnosti od forme $\sigma_i^2$. Ukoliko je $\sigma_i^2$ poznato, potrebno je samo polaznu liniju regresije podeliti sa $\sigma_i^2$ i na dobijeni model primeniti metodu najmanjih

kvadrata. Međutim, $\sigma_i^2$ u većini slučajeva nije poznato, već se mogu pretpostaviti specifikacije varijanse. Razlikuju se sledeće specifikacije:

1. Multiplikativna heteroskedastičnost gde je $\sigma_i^2 = \sigma^2 X_i^2$,
2. Varijansa proporcionalna sa $X_i$, $\sigma_i^2 = \sigma^2 X_i$,
3. Aditivna heteroskedastičnost, $\sigma_i^2 = a + bX_i + cX_i^2$,
4. Heteroskedastičnost zavisne promenljive, $\sigma_i^2 = \sigma^2[E(Y_i)]^2$,
5. Logaritamska transformacija, $lnY_i = \alpha + \beta lnX_i + \varepsilon_i$.

Kod multiplikativne heteroskedastičnosti, lako se dolazi do homoskedastičnog modela, deljenjem obe strane regresijske jednačine sa $X_i$,

$$\frac{Y_i}{X_i} = \alpha\left(\frac{1}{X_i}\right) + \beta + \frac{\varepsilon_i}{X_i}$$

odnosno:

$$Y_i^* = \beta + \alpha X_i^* + \varepsilon_i^*$$

gde je $Y_i^* = \frac{Y_i}{X_i}$, $X_i^* = \frac{1}{X_i}$ i $\varepsilon_i^*$ zadovoljava osnovne pretpostavke klasičnog modela sa $Var(\varepsilon_i^*) = \sigma^2$. Sa ovako dobijenom jednačinom regresije može se baratati kao i sa klasičnom, osim što su $\alpha$ i $\beta$ zamenili mesta. Njihovi ocenjivači dobijeni metodom najmanjih kvadrata imaju sva poželjna svojstva. Prednost ovog oblika heteroskedastičnosti je njena jednostavnost, a mana je što ova specifikacija isključuje mogućnost da odstupanje bude homoskedastično ili da heteroskedastičnost bude različita od stepena koji odgovara $\delta = 2$.

Na sličan način se jednačina regresije transformiše i u slučaju varijanse proporcionalne sa $X_i$:

$$\frac{Y_i}{\sqrt{X_i}} = \alpha\left(\frac{1}{\sqrt{X_i}}\right) + \beta\sqrt{X}_i + \frac{\varepsilon_i}{\sqrt{X_i}},\ X_i > 0$$

i tada se primenjuje metoda najmanjih kvadrata.

U slučaju aditivne heteroskedastičnosti, potrebno je postupiti na sledeći način:

a) Primeniti prvo metodu najmanjih kvadrata na

$$e_i^2 = a + bX_i + cX_i^2 + u_i$$

gde su $e_i$ reziduali regresije $Y$ na $X$ dobijeni metodom najmanjih kvadrata i gde je $u_i = e_i^2 - \sigma_i^2$. Ovako dobijeni ocenjivači za $a$, $b$ i $c$ se nazivaju ocenjivači „prve runde" i oni su označeni sa $\hat{a}$, $\hat{b}$ i $\hat{c}$. Ocenjivač prve runde za $\sigma_i^2$ je tada:

$$\hat{\sigma}_i^2 = \hat{a} + \hat{b}X_i + \hat{c}X_i^2$$

b) Dobijeni ocenjivači prve runde od $a$, $b$ i $c$ nisu asimptotski efikasni jer je $u_i$ heteroskedastična. Zatim se primenjuje metoda najmanjih kvadrata na:

$$\frac{e_i^2}{\hat{\sigma}_i^2} = a\frac{1}{\hat{\sigma}_i^2} + b\frac{X_i}{\hat{\sigma}_i^2} + c\frac{X_i^2}{\hat{\sigma}_i^2} + {u_i}^*$$

i dobijaju se ocenjivači „druge runde" od $a$, $b$ i $c$, u oznaci $\tilde{a}$, $\tilde{b}$ i $\tilde{c}$ koji su asimptotski efikasni. Ocenjivač druge runde za $\sigma_i^2$ je tada jednak:

$$\tilde{\sigma}_i^2 = \tilde{a} + \tilde{b}X_i + \tilde{c}X_i^2$$

c) Ponovo se primenjuje metoda najmanjih kvadrata, ali sada na:

$$\frac{Y_i}{\tilde{\sigma}_i^2} = \alpha\frac{1}{\tilde{\sigma}_i^2} + \beta\frac{X_i}{\tilde{\sigma}_i^2} + {\varepsilon_i}^*$$

Može se dokazati da dobijeni ocenjivači od $\alpha$ i $\beta$ imaju ista asimptotska svojstva kao i maksimalno verodostojni ocenjivači. Konzistentni ocenjivači njihovih varijansi mogu se izračunati uvrštavanjem $\frac{1}{\tilde{\sigma}_i^2}$ umesto $w_i$ u formulu za $Var(\tilde{\alpha})$ i $Var(\tilde{\beta})$.

U slučaju heteroskedastične zavisne promenljive, računamo funkciju maksimalne verodostojnosti koja sadrži tri nepoznata parametra, $\alpha$, $\beta$ i $\sigma^2$:

$$L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n} log(\alpha + \beta X_i) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\frac{Y_i - \alpha - \beta X_i}{(\alpha + \beta X_i)}\right]^2$$

Izračunavanje vrednosti ovih parametara nije jednostavno. Izvođenje se može pojednostaviti ako se primeni metoda najmanjih kvadrata na $\frac{Y_i}{\hat{Y}_i} = \alpha\frac{1}{\hat{Y}_i} + \beta\frac{X_i}{\hat{Y}_i} + {\varepsilon_i}^*$, gde su $\hat{Y}_i$ predviđene vrednosti od $Y$ iz regresije $Y$ na $X$ dobijene metodom najmanjih kvadrata. Ceo postupak proizilazi iz činjenice da je $\hat{Y}_i$ nepristrasan i konzistentan ocenjivač od $E(Y_i)$.

Logaritamska transformacija često redukuje heteroskedastičnost u odnosu na klasičan model regresije. Ova transformacija je popularna zbog bolje ekonomske interpretacije. Koeficijentom pravca $\beta$ može se izmeriti elastičnost $Y$ u odnosu na $X$. U slučaju da se ne može pretpostaviti oblik heteroskedastičnosti, predlaže se korišćenje Whiteove korekcije u izračunavanju standardnih grešaka ocena parametara.

## 2.4. Autokorelirana odstupanja

Prema pretpostavci klasičnog modela linearne regresije poznato je sledeće:

$$Cov(\varepsilon_i, \varepsilon_j) = E[\varepsilon_i - E(\varepsilon_i)]\left[\varepsilon_j - E(\varepsilon_j)\right] = 0, i \neq j$$

Kako se pretpostavlja da su sredine $\varepsilon_i$ i $\varepsilon_j$ jednake nuli, dobija se da je $E(\varepsilon_i\varepsilon_j) = 0$.

Uzimajući u obzir pretpostavku normalnosti, pretpostavka da je kovarijansa između $\varepsilon_i$ i $\varepsilon_j$ jednaka nuli podrazumeva da su i nezavisni. Ova osobina regresijskih odstupanja poznata je kao neautokoreliranost. To zapravo znači da odstupanje koje se javlja u jednoj tački opažanja nije korelirano ni sa jednim drugim odstupanjem.

Potrebno je razmotriti verovatnoću pretpostavke o neautokorelaciji, ispitati posledice njenog narušavanja za svojstva ocenjivača dobijenih metodom najmanjih kvadrata i pronaći alternativne metode ocenjivanja ukoliko je potrebno. Uobičajena je tvrdnja da je pretpostavka o neautokorelaciji najčešće narušena kada se ocenjuju veze iz vremenskih nizova podataka nego u slučaju veza ocenjenih iz podataka vremenskog preseka. Ova se tvrdnja zapravo oslanja na tumačenje da je odstupanje sažetak velikog broja slučajnih i nezavisnih, ali i nemerljivih uzoraka koji se tiču proučavane veze. Zato postoji sumnja da bi se učinak faktora koji deluju u jednom razdoblju mogao delimično preneti na sledeća razdoblja.

Autokorelacija se najbolje može uporediti sa zvučnim učinkom udarene žice na instrumentu – iako je zvuk najjači u vreme udarca, on ne prestaje odmah, nego postepeno slabi jedno vreme dok se konačno ne ugasi. To može biti i karakteristika odstupanja, gde njegov učinak može trajati jedno vreme nakon što se pojavi. Međutim, tokom trajanja učinka jednog odstupanja, javljaju se druga odstupanja, kao kada bi se žica na intrumentu ponovo udarala promenljivom jačinom. Što je vreme između udaraca kraće, veća je verovatnoća da se prethodni zvuk još uvek može čuti. Slično ovoj priči, što su kraća razdoblja između registrovanja zvuka, veća je verovatnoća pojavljivanja autokoreliranih odstupanja.

Zbog verovanja da veze ocenjene iz opažanja tokom vremena sadrže autokorelirana odstupanja u literaturi se umesto $i$ piše $t$, i prema tome kada su odstupanja autokorelirana sledi da je:

$$E(\varepsilon_t \varepsilon_{t-s}) \neq 0, \qquad t > s$$

Ovaj izraz zapravo predstavlja da je odstupanje koje se desilo u trenutku $t$ povezano sa odstupanjem koje se dogodilo u trenutku $(t-s)$. Posledice autokorelacije za ocenjivanje se mogu najbolje odrediti ukoliko se prvo precizira narav autokorelacije. Većina radova je izvedena na pretpostavci da regresijsko odstupanje sledi autoregresivnu šemu prvog reda. Ona se skraćeno obeležava sa AR (1).

### 2.4.1. Autoregresivna odstupanja prvog reda

Kod autoregresivnih odstupanja prvog reda, odstupanja se generišu prema sledećoj šemi:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad \text{za svako } t$$

gde je $\rho$ parameter čija je apsolutna vrednost manja od jedan, a $u_t$ je normalna nezavisna promenljiva sa sredinom jednakom nuli i varijansom ${\sigma_u}^2$. Takođe, promenljiva $u_t$ ne zavisi od $\varepsilon_{t-1}$. Sve ovo se zapisuje:

$$u_t \sim N(0, {\sigma_u}^2), \quad \text{za svako } t$$

$$E(u_t, u_s) = 0, \text{ za svako } t \neq s$$

$$E(u_t\varepsilon_{t-1}) = 0, \text{ za svako } t$$

### 2.4.2. Testiranje odsustva autokorelacije

Pretpostavka o autokorelaciji se uobičajeno proverava na dva načina. Prvi je grafička metoda, odnosno analiza reziduala koji se dobijaju standardnom procedurom. Drugi način je korišćenje nekog od testova, uobičajeno se koristi Durbin-Watson (Derbin-Votson) test. Dakle, testira se hipoteza o neautoregresiji:

$$H_0: \quad \rho = 0,$$

alternativna hipoteza u tom slučaju je najčešće ona da postoji pozitivna autoregresija:

$$H_A: \quad \rho > 0.$$

Da bi se koristio Durbin-Watsonov test, neophodno je izračunati vrednost *d*:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} {e_t}^2}$$

gde su $e_t$ reziduali metode najmanjih kvadrata.

Pravila odlučivanja u slučaju ove alternativne hipoteze su sledeća:

1. Odbaciti $H_0$ ukoliko je $d < d_L$
2. Ne odbaciti $H_0$ ukoliko je $d > d_U$
3. Test je neodređen ako je $d_L \leq d \leq d_U$

gde su $d_U$ i $d_L$ gornja i donja granica čije su vrednosti konstruisali Durbin i Watson i mogu se pronaći u tablici. Ako je alternativna hipoteza dvostrana, pravila odlučivanja za Durbin-Watsonov test su:

1. Odbaciti $H_0$ ukoliko je $d < d_L$ i ako je $d > 4 - d_L$
2. Ne odbaciti $H_0$ ukoliko je $d_U < d < 4 - d_U$
3. Test je neodređen ako je $d_L \leq d \leq d_U$ ili ako je $4 - d_U \leq d \leq 4 - d_L$

Koristeći postupak Cochrane-Orcuttovog (Kokran-Orketovog) ocenjivanja (*Počela Ekonometrije*, Jan Kmenta, str 314.), dobija se ocena $\hat{\rho}$:

$$\hat{\rho} = \frac{\sum e_t e_{t-1}}{\sum {e_{t-1}}^2}, \text{ za } t\text{=2,3...}n$$

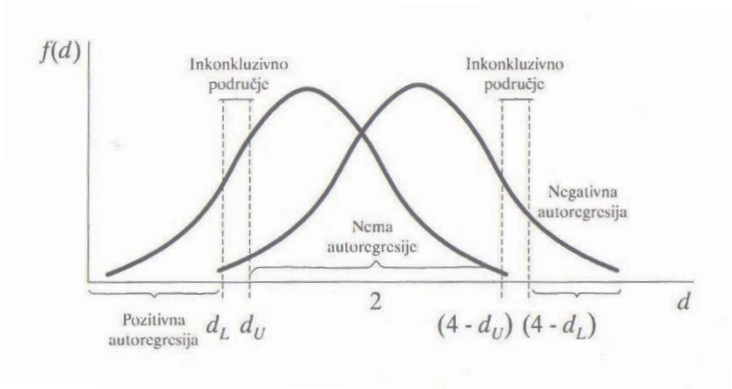
Kako *d* može da se zapiše kao

$$d = \frac{\sum_{t=2}^{n} {e_t}^2}{\sum_{t=1}^{n} {e_t}^2} + \frac{\sum_{t=2}^{n} {e_{t-1}}^2}{\sum_{t=1}^{n} (e_t)^2} - 2\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} {e_t}^2}$$

i odatle je *d*:

$$plim d = 2(1 - \rho)$$

Dakle, ako kritična vrednost Durbin-Watson statistike pripada ovom intervalu $1.5 < d < 2.5$, prihvata se $H_0$, odnosno može se reći da nema autokorelacije.




Slika 8-1, *Počela ekonometrije,* Jan Kmenta, str 330.

### 2.4.3. Tretman u slučaju autokorelacije

Ako se u podacima utvrdi da postoji autokorelacija, uobičajen tretman za rešavanje ovog problema je eliminisanje simptoma autokorelacije korišćenjem metode procene modela različite od metode najmanjih kvadrata. Ali mnogi autori ipak preporučuju da se na prvom mestu spreči da do autokorelacije dođe, nego da se model oslobađa posledica autokorelacije.

## 2.5. Stohastičnost promenljive

Poslednja ili peta pretpostavka klasičnog modela govori o nestohastičnosti promenljive *X*, da su njene vrednosti fiksne i da je $\sum_{i=1}^{n}(X_i - \bar{X})^2 \neq 0$ konačan kada $n \to \infty$. Uslov o fiksnim vrednostima primenljive *X* se postavlja kako bi se izbegle poteškoće koje bi se pojavile kada bi se vrednost za *X* menjala od uzorka do uzorka. Drugi uslov se nameće kako bi se osigurala asimptotska svojstva ocenjivača dobijenih metodom najmanjih kvadrata. Kod ispitivanja stohastičnosti promenljive *X*, potrebno je razlikovati slučajeve zavisnosti od $\varepsilon$:

a) X i $\varepsilon$ nezavisne
b) X i $\varepsilon$ su istovremeno nekorelirani
c) X i $\varepsilon$ su nezavisne i istovremeno nekorelirani.

Nakon uvođenja i pretpostavke da je *X* konačan broj različit od nule, proveriće se svojstva ocenjivača parametara regresije u sva tri slučaja.

a) Kada su X i $\varepsilon$ nezavisne, ocenjivači $\tilde{\alpha}$ i $\tilde{\beta}$ zadržavaju svojstvo nepristrasnosti (*Počela Ekonometrije*, Jan Kmenta, str 336.), ali se više ne mogu posmatrati kao najbolji linearni nepristrasni ocenjivači jer ne predstavljaju više linearne funkcije od $Y_1, Y_2, \dots, Y_n$. Međutim ovi ocenjivači zadržavaju osobinu efikasnosti. Što se tiče asimptotskih svojstava, može se dokazati da oni zadržavaju sva poželjna svojstva kada *X* ne zavisi od $\varepsilon$. Ukoliko se još razmotre varijanse

ovih ocenjivača, dobijaju se iste varijanse od $\tilde{\alpha}$ I $\tilde{\beta}$ kao kada je *X* nestohastička, osim što su članovi koji sadrže *X* zamenjeni njihovim matematičkim očekivanjima.
Dakle, dolazi se do zaključka da zamena pretpostavke o nestohastičnosti da su X i $\varepsilon$ nezavisne ne menja poželjna svojstva ocenjivača dobijenih metodom najmanjih kvadrata.

b) U drugom slučaju kada su X i $\varepsilon$ su istovremeno nekorelirani pretpostavlja se da su međusobne kovarijanse jednake nuli, odnosno $Cov(X_1, \varepsilon_1) = Cov(X_2, \varepsilon_2) = \cdots Cov(X_n, \varepsilon_n) = 0$. U ovom slučaju ocenjivači dobijeni metodom najmanjih kvadrata gube svoja poželjna svojstva, zadržavaju samo asimptotska.

c) Kada X i $\varepsilon$ nisu u korelaciji, niti su nezavisne, tu ocenjivači dobijeni metodom najmanjih kvadrata gube konzistentnost (*Počela Ekonometrije*, Jan Kmenta, str 340.) Srećom, razvijaju se alternativne metode ocenjivanja koje će dati konzistentne ocene za $\tilde{\alpha}$ i $\tilde{\beta}$.

### 2.5.1. Testiranje i tretman u slučaju stohastičnosti promenljive

Nažalost, ne postoji efikasan način provere pretpostavke o stohastičnosti promenljive. Slično kao i prilikom narušenja pretpostavke da je sredina jednaka nuli, jedan od pristupa može biti računanje reziduala za sve ispitanike, a potom ispitivanje korelacije između reziduala i ispitanika.

## 2.6. Višestruka regresija

U prethodnim poglavljima predstavljene su osnovne pretpostavke jednostruke linearne regresije, ocenjivači dobijeni metodom najmanjih kvadrata i objašnjena su njihova poželjna svojstva. Takođe je pokazano šta se dešava sa njima u slučaju narušenosti neke pretpostavke, kako se vrši testiranje i koji su tretmani za njihovo otklanjanje. Zaključci do kojih se došlo, mogu se bez mnogo poteškoća proširiti i na višestruku regresiju.

Jednačina višestruke linearne regresije se zapisuje na sledeći način:

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \qquad i = 1,2,3 \ldots n$$

gde je $Y_i$ zavisna, $X_{i1}, \ldots X_{ik}$ nezavisne promenljive, $\alpha$, $\beta_1, \ldots, \beta_k$ koeficijenti i $\varepsilon_i$ odstupanje. Pored već navedenih pet pretpostavki dodaju se još dve:

6. Obim uzorka je veći ili jednak od broja koeficijenata koje treba oceniti
7. Ne postoji linearna zavisnost između nezavisnih promenljivih.

Pretpostavka da broj koeficijenata ne sme biti veći od obima uzorka osigurava dovoljan broj stepena slobode u ocenjivanju. Sledeća pretpostavka govori da nijedna nezavisna promenljiva ne sme biti u korelaciji sa drugom nezavisnom promenljivom niti sa njenom linearnom kombinacijom, što je takođe neophodno za ocenjivanje. Narušavanje ove pretpostavke naziva se multikolinearnost.

### 2.6.1. Multikolinearnost

Ukoliko između nezavisnih promenljivih, odnosno prediktora postoji jaka povezanost, tačnije visoka korelacija (viša od *r=0.9*), onda se smatra da postoji multikolinearnost. Multikolinearnost se dijagnostikuje i pomoću vrednosti Tolerance i VIF (*Variance inflation factor*). Tolerance je pokazatelj koji pokazuje koji deo varijanse kriterijuma nije objašnjen varijansama prediktora koje su u modelu, a računa se za svaku promenljivu prema sledećoj formuli: $1 - R^2$, gde je *R* koeficijent determinacije. Vrednosti Tolerance manji od 0.10 ukazuju da postoji visoka korelacija promenljive sa ostalim prediktorima u modelu. VIF je pokazatelj koji predstavlja recipročnu vrednost pokazatelja Tolerance, a vrednosti veće od 10 ukazuju na postojanje multikolinearnosti.

### 2.6.2. Tretman u slučaju multikolinearnosti

Multikolinearnost umanjuje performanse regresionog modela, stoga se predlaže da se iz modela uklone promenljive čija je linearna kombinacija veća od *r=0.7*. Osim uklanjanja jedne ili obe promenljive, tretman može da bude i pravljenje jedne zajedničke promenljive od dve promenljive koje su u korelaciji. Pošto je multikolinearnost svojstvo uzorka, a ne populacije, na pojavu multikolinearnosti posebno su osetljivi mali uzorci. Jedan od tretmana za otklanjanje multikolienarnosti je povećanje uzorka na odgovarajuči način. Koristi se i takozvana ORR metoda (engl. Ordinary Ridge Regression), razmena male pristrasnosti za veliko smanjenje varijanse.

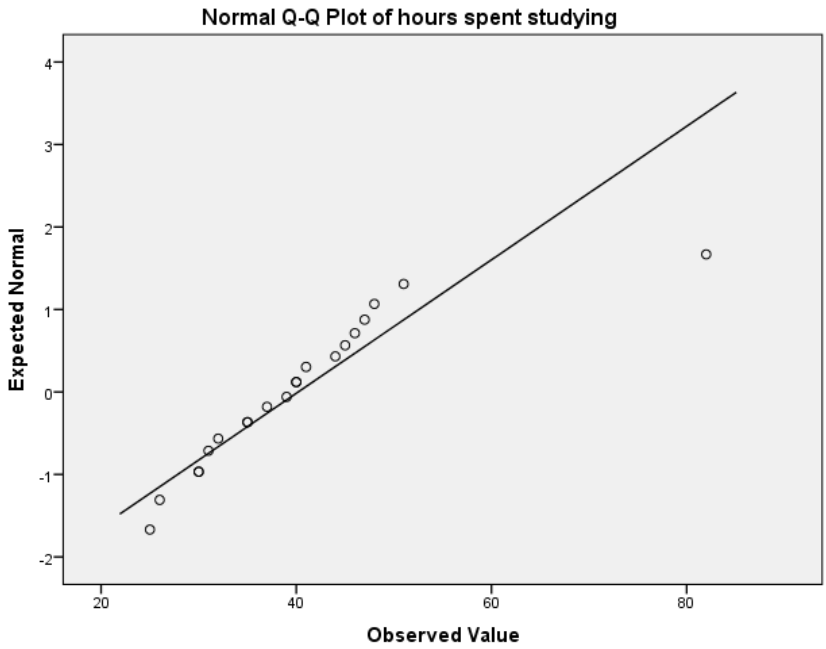
# 3. Simulacije primera u softveru SPSS

U svrhu ilustracije narušenih pretpostavki za korišćenje linearne regresije, koristiće se podaci preuzeti sa sledećih sajtova: *https://mathstatsresearch.weebly.com/regression.html* i *http://staff.bath.ac.uk/pssiw/stats2/page16/page16.html*. Pre primene svake analize proveriće se ispunjenost pretpostavki za njeno sprovođenje. U prethodnom delu ovog master rada detaljno je objašnjeno na koji način se obično proverava ispunjenost pretpostavki koje je moguće ispitati upotrebom klasičnih statističkih softvera. U ovom delu će se kroz tri primera prikazati kako izgleda kada se detektuje da pretpostavke nisu ispunjene i kakav je tretman u tom slučaju. Za potrebe analize podataka korišćen je softverski paket SPSS 25 for Windows *(Statstical Package for Social Sciences).* U primerima će se prikazati samo one pretpostavke koje su narušene. Uobičajeni načini za provere pretpostavki su grafikoni i testovi. Nekoliko grafičkih metoda provere su veoma praktični, a mnogi statistički programi već imaju integrisane ove metode, pa je vrlo lako doći do njih. Međutim, poteškoće se mogu javiti prilikom tumačenja istih. Testovi daju jednoznačne rezultate, tako da ponekad zahtevaju dodatne informacije koje nije moguće uvek prikupiti. Stoga se, ukoliko je moguće, radi kombinacija testova i grafičkih metoda.

Jedan od uslova za sporovođenje regresione analize je da kriterijumska varijabla bude kontinuirana, a prediktori najmanje na intervalnom nivou, kao i da broj parametara bude manji ili jednak broju merenja koji su na raspolaganju. Ove pretpostavke su ispunjene u svim primerima i neće se proveravati.

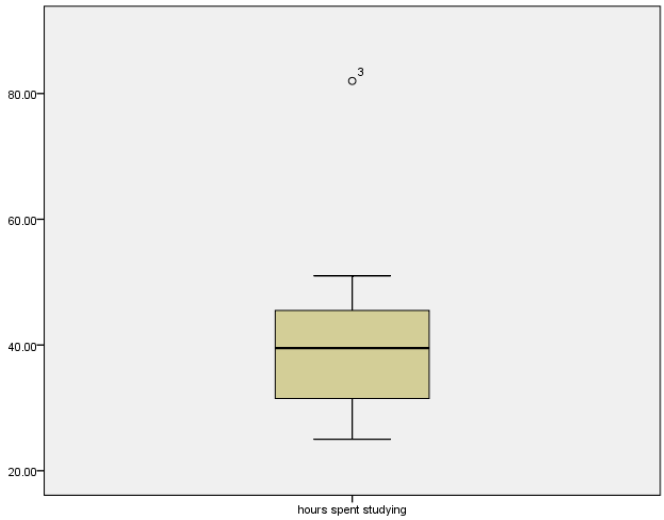
## 3.1. Prvi primer

U prvom primeru korišćeni su podaci o broju sati provedenih u učenju i ocene na ispitu. Opisan je najprostiji slučaj regresione analize, odnosno jednostruka regresiona analiza, kada imamo jednu prediktorsku i jednu kriterijumsku varijablu. Kriterijumsku varijablu predstavlja ocena na ispitu, a prediktor je broj sati provedenih u učenju. Ispitana je varijabla broj sati provedenih u učenju.




*Grafikon 1*

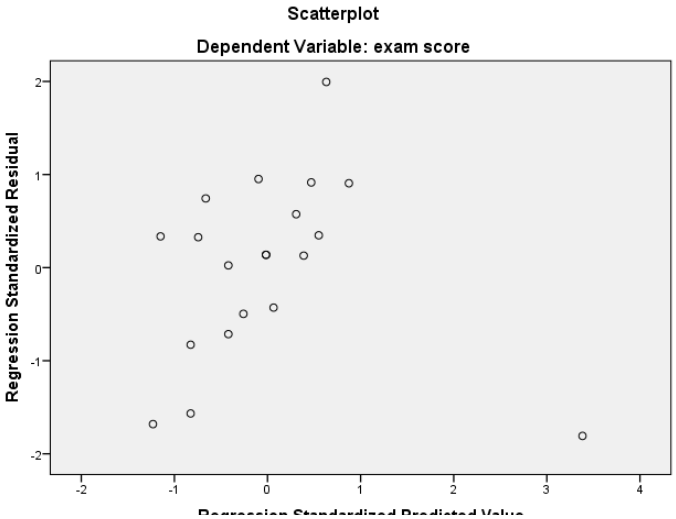
U Q-Q grafikonu se primećuje da postoji jedna vrednost koja ne prati liniju i ona se može detektovati kao outlier. Isto se može primeti i u *Box Plot* grafikonu koji izoluje slučaj koji je prepoznat kao netipična vrednost.




*Grafikon 2*

Outlier može da se javi zbog različitih razloga: pogrešan upis vrednosti prilikom unosa ili merenja, pogrešan podatak ili stvarna neuobičajena vrednost. Na primer, visina žena se obično kreće u rasponu od 160 do 180 cm, žene koje su više ili niže od ovih vrednosti spadaju u outliere. Iako je regresija „osetljiva“ na netipične tačke, mora se dobro razmisliti na koji način će se tretirati outlier, odnosno da li ga zadržati, izbaciti ili transformisati. Ukoliko je došlo do greške u unosu ili prilikom merenja, ovaj podatak se može izbaciti ili izmeniti pravom vrednošću ukoliko je ona poznata. Ako nije moguće otkriti

uzrok nastanka outlier-a, tada se prave modeli sa i bez spomenutog elementa. Ukoliko su modeli jednaki, elemenat ne utiče bitno. Međutim ako se modeli bitno razlikuju potrebno je dalje ispitivanje i pronalazak odgovarajućeg tretmana. Formiran je model i prikazan je *scatterplot*, odnosno grafikon rasturanja standardizovanih reziduala u kom se jasno uočava da postoji outlier.




*Grafikon 3*

Program SPSS, prilikom sprovođenja regresione analize, nudi opciju ispisa informacija o netipičnim tačkama, gde se kao atipični slučajevi smatraju oni sa standardizovanim vrednostima većim od 3.0 ili manjim od -3.0. U tabeli se može primetiti da postoje slučajevi čije su standardizovane vrednosti izvan ovog opsega.

**Residuals Statistics[a]**

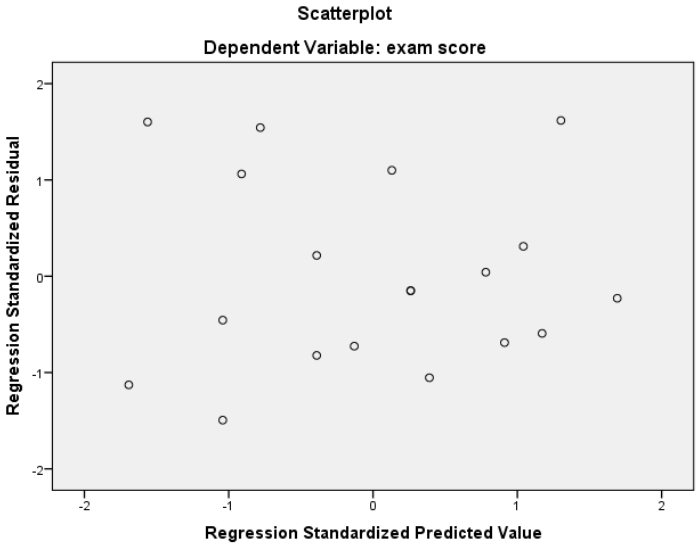
| | Minimum | Maximum | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Predicted Value | 51.6585 | 86.6892 | 61.0000 | 7.59426 | 20 |
| Residual | -14.68916 | 16.20633 | .00000 | 7.90808 | 20 |
| Std. Predicted Value | -1.230 | 3.383 | .000 | 1.000 | 20 |
| Std. Residual | -1.808 | 1.995 | .000 | .973 | 20 |

a. Dependent Variable: exam score

Takođe, program daje i Mahalanobisove i Cookove udaljenosti reziduala. One nisu prikazane u ispisu rezultata nego su sačuvane u datoteci sa podacima kao dodatna promenljiva. Kritična vrednost za Mahalanobisove distance se određuje na osnovu broja promenljivih i zahteva više proračuna, tako da se u primeru koristila Cookova distanca. Kod Cookovih distanci smatra se da je potrebno tretirati one slučajeve gde je distanca veća od 1.

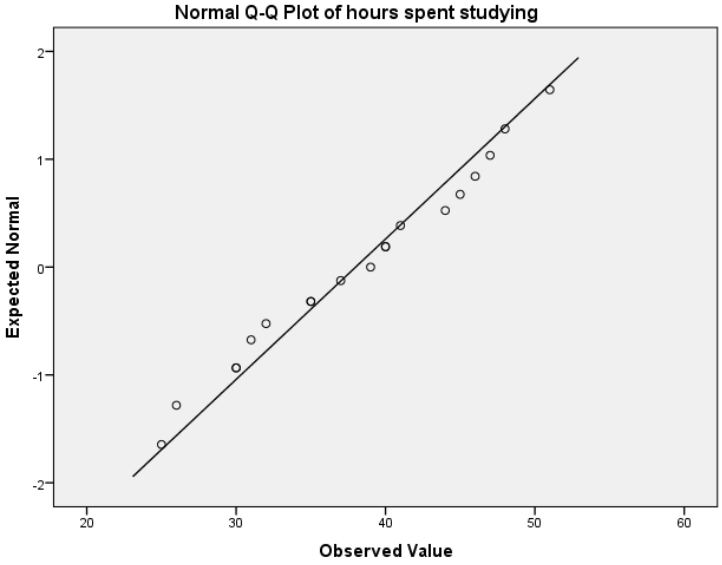
U ovom primeru, pokazalo se da postoji slučaj sa standardizovanom vrednošću reziduala koja je veća od 3, stoga su se posmatrale Cookove distance na osnovu kojih bi se takve vrednosti eliminisale. U datoteci

je utvrđeno da postoji jedna vrednost kod koje Cookova distanca pokazuje udaljenost veću od 1 (8.814). Nakon eliminisanja ove vrednosti, ponovila se regresiona analiza. U grafikonu se primećuje kako se dijagram rasturanja promenio nakon uklanjanja netipične tačke.




*Grafikon 4*

Nakon izbacivanja outlier-a ponovljena je i deskriptivna analiza, odnosno ponovo je nacrtan Q-Q grafikon. Sada se u ovom grafikonu primećuje da ne postoje vrednosti koje mnogo odstupaju od prave, i stoga se može zaključiti da nema netipičnih tačaka.




*Grafikon 5*

## 3.2. Drugi primer

Drugi primer ilustruje takođe jednostruku linearnu regresiju. Korišćeni su podaci o broju saobraćajnih nesreća i broju pređenih kilometara. Kriterijumsku varijablu predstavlja broj saobraćajih nesreća, a prediktor broj pređenih kilometara u milijardama kilometara. Pre sprovođenja analize, proverene su pretpostavke. Za početak je ispitana distribucija prediktorske i kriterijumske promenljive. Utvrđeno je da je narušena pretpostavka o normalnosti distribucije kod varijable broja pređenih kilometara.

Kao što je navedeno, za proveru pretpostavke o normalnoj distribuciji koristi se kombinacija testovnih i grafičkih pokazatelja. Pregledom tabele deskriptivnih pokazatelja mogu se dobiti neke informacije o distribuciji. Jedan od indikatora je da Vrednosti Mean, 5% Trimmed Mean i Median budu slične. Druga vrednost koju posmatramo je vrednosti skewness-a i kurtosisa. Vrednosti od -1 do 1 ukazuju na normalnu raspodelu, a vrednosti veće od 3 i manje od -3 ukazuju na odstupanje od normalne raspodele. U ovom slučaju sve navedene vrednosti zadovoljavaju kriterijume.

**Descriptives**

| | | | Statistic | Std. Error |
|---|---|---|---|---|
| Billions km driven | Mean | | 272,24 | 18,503 |
| | 95% Confidence Interval for Mean | Lower Bound | 235,14 | |
| | | Upper Bound | 309,33 | |
| | 5% Trimmed Mean | | 270,55 | |
| | Median | | 252,99 | |
| | Variance | | 18829,822 | |
| | Std. Deviation | | 137,222 | |
| | Minimum | | 73 | |
| | Maximum | | 502 | |
| | Range | | 429 | |
| | Interquartile Range | | 272 | |
| | Skewness | | ,194 | ,322 |
| | Kurtosis | | -1,303 | ,634 |

Rečeno je da se za statističko testiranje normalnosti koristi Kolmogorov-Smirnov test (ili Shapiro-Wilk test za manje uzorke). Ukoliko je *p* manje od 0.05 u ovim testovima, empirijska raspodela značajno

odstupa od normalne raspodele. Ovo je slučaj u prikazanom primeru. Vrednost Kolmogorov-Smirnov testa je $p = 0.012$, što je manje od 0.05 i upućuje na to da podaci nemaju normalnu raspodelu.

**Tests of Normality**

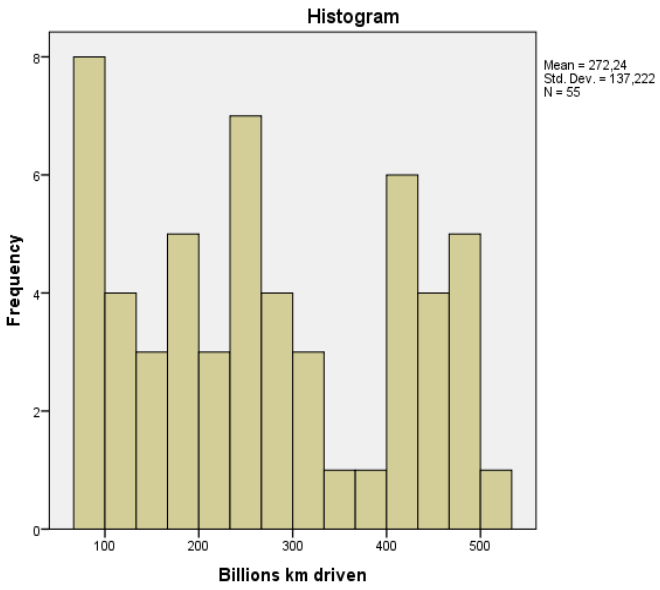
| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Billions km driven | ,137 | 55 | ,012 | ,927 | 55 | ,002 |

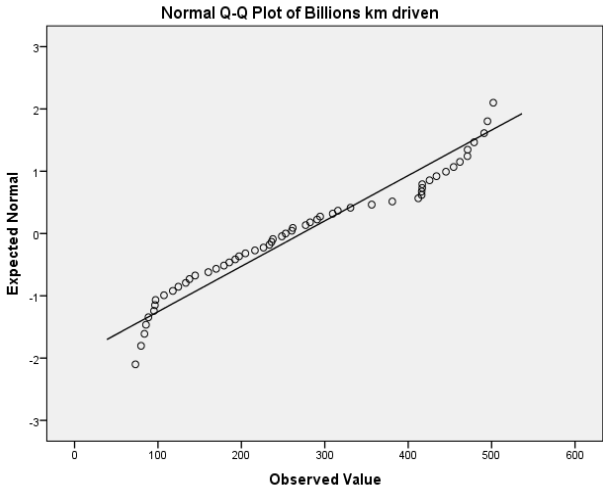
a. Lilliefors Significance Correction

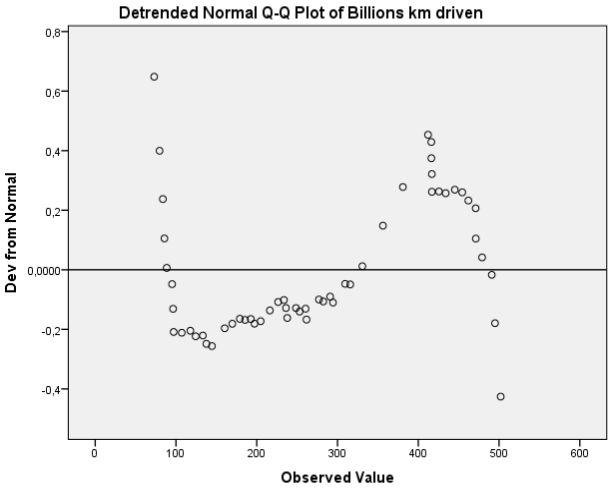
Ovo je česta situacija u praksi, da različiti načini provere pretpostavki daju različite rezultate. Upravo iz tog razloga se koristi kombinacija više tehnika pre donošenja konačnog zaključka.

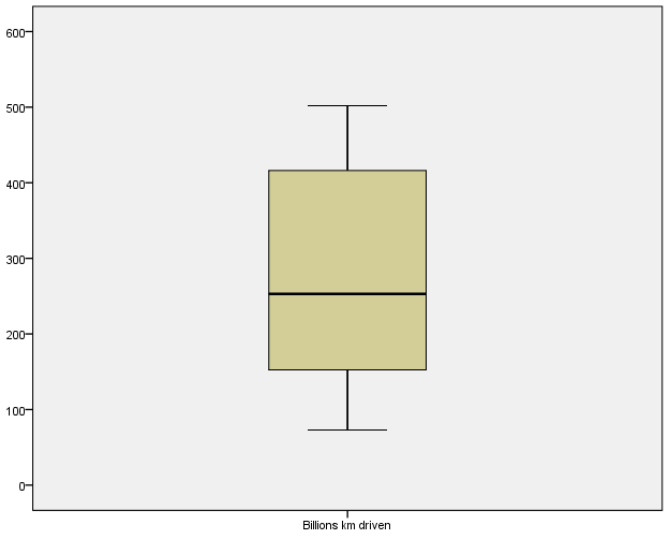
Što se tiče grafičkih metoda rečeno je da se uobičajeno koriste sledeći grafikoni: histogram – vizuelna procena da li je empirijska raspodela slična zvonastoj simetričnoj raspodeli, zatim normalni Q–Q grafikon, kod kog, ukoliko je raspodela normalna, tačke budu na pravoj liniji. Odstupanje tačaka od prave linije ukazuje na odstupanje raspodele od normalne. Koristi se i detrendovan normalni Q–Q grafikon. Ako je raspodela normalna, tačke će biti ravnomerno raspoređene iznad i ispod horizontalne linije. Ako raspodela nije normalna, raspored tačaka će imati neki oblik. Još se koristi i *Box Plot*, ako kod ovog grafikona postoji nekoliko ekstremnih vrednosti ili neobičnih vrednosti na bilo kom kraju raspodele, to ukazuje na odstupanje od normalne raspodele. Ako medijana nije u centru grafikona kutije već je znatno bliža jednom od krajeva kutije, to ukazuje na odstupanje od normalne raspodele.

Pregledom grafikona u ovom primeru primećuje se da histogram nema klasičan simetričan oblik zvona što ukazuje da podaci nisu normalno distribuirani. Dok se kod normalnog Q-Q grafikona primećuje da u velikoj meri tačke prate liniju sa nekim odstupanjima, kod detrendovanog Q-Q grafikona primećuje se da se tačke raspoređuju u obliku polegnutog slova S, što takođe ukazuje da raspodela nije normalna. Kod *Box Plot* grafikona se primećuje da nema karakterističnih netipičnih tačaka ili ekstremnih vrednosti, a medijana blago odstupa od centra.

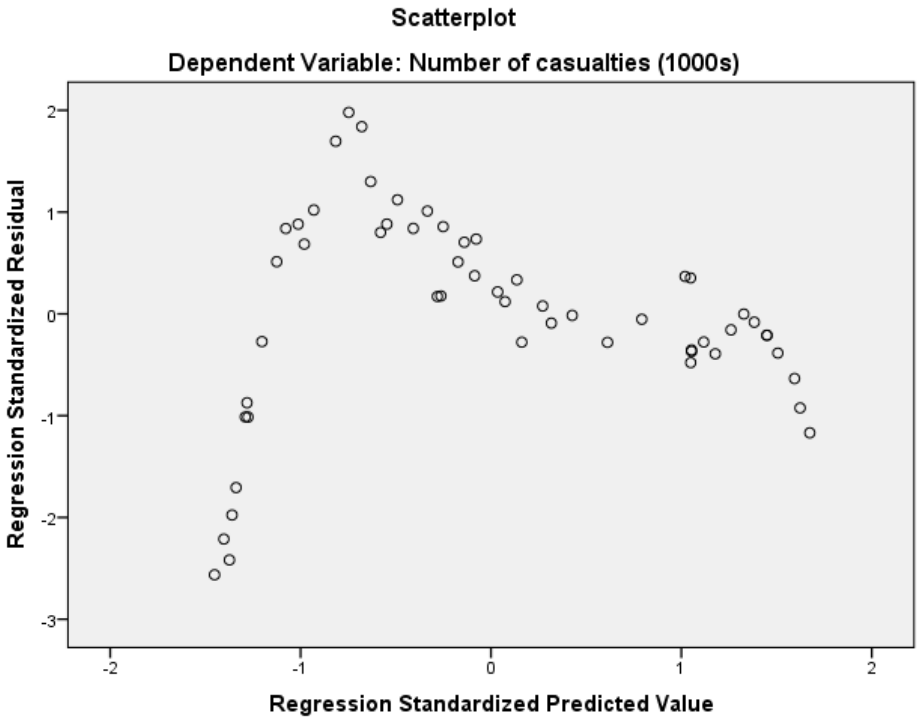

Histogram
Mean = 272,24
Std. Dev. = 137,222
N = 55
Frequency
Billions km driven


Normal Q-Q Plot of Billions km driven
Expected Normal
Observed Value


Detrended Normal Q-Q Plot of Billions km driven
Dev from Normal
Observed Value


Billions km driven

Rečeno je i da se u regresionoj analizi normalnost podataka proverava preko dijagrama rasturanja reziduala, tako da je formiran model i urađena regresiona analiza. Iz dijagrama rasturanja standardizovanih reziduala primećuje se da postoji karakterističan obrazac. Oblik reziduala formira obrnuto U, što nagoveštava da podaci nemaju normalnu raspodelu.

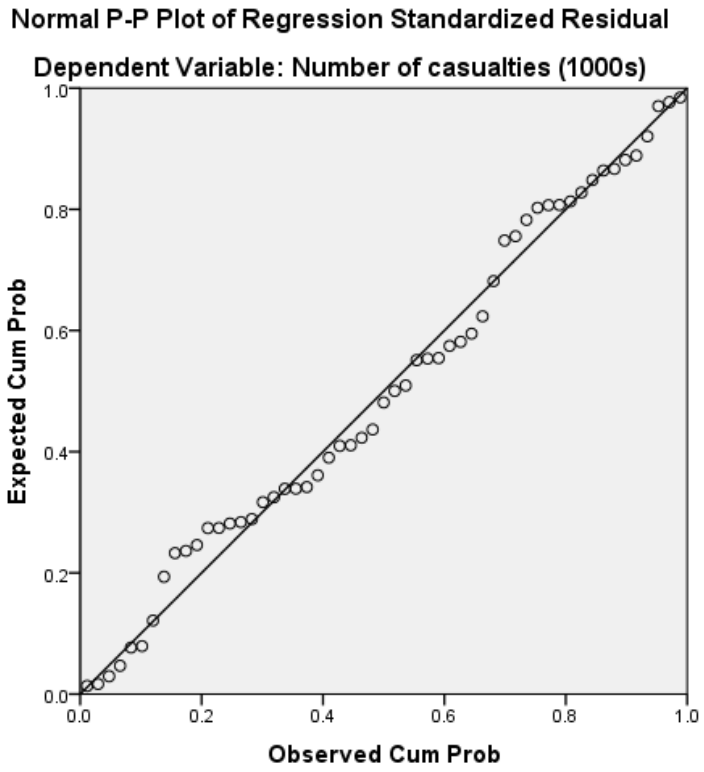



*Grafikon 6*

Iz svega navedenog, sa posebnim osvrtom na dijagram rasturanja reziduala može da se izvede zajključak da je narušena pretpostavka o normalnosti i može se pristupiti nekom od tretmana za rešenje ovog problema. Kao tretman korišćena je logaritamska transformacija podataka. Međutim, nakon transformacije grafički pokazatelji sugerišu da su podaci normalno distribuirani, ali dijagram rasturanja reziduala i dalje pokazuje da je pretpostavka o normalnosti narušena. Neki autori smatraju da kršenje ove pretpostavke nije ključno za regresionu analizu, odnosno da normalnost podataka ne garantuje validan i upotrebljiv model. Bitno je napomenuti i veličinu uzorka kod sprovođenja analize i uticaj na normalnost podataka. Neke smernice za veličinu uzorka su da se određuje na osnovu broja nezavisnih varijabli u modelu, i to da je minimalan broj posmatranja po varijabli 10 (takozvano pravilo palca – *"rule of thumb"*). Tabachnick (Tabašnik) i Fidell (Fidel) predlažu pravilo prema sledećoj formuli "50 + 8*m*", gde *m* predstavlja broj prediktora. S obzirom na to da je u ovom primeru ilustrovana jednostavna linearna regresija, sa jednim prediktorom, može da se kaže da je veličina uzorka od $n = 55$ odgovarajuća i da pretpostavka o normalnosti može da se zanemari.

Ispis regresione analize daje i Q-Q plot, gde se uzorački reziduali porede sa teorijskim rezidualima iz normalne raspodele. S obzirom na to da ovaj grafikon ne prikazuje da postoje velika odstupanja od

normalne distribucije, i gotovo je identičan i pre i nakon tretmana normalnosti, može da se zaključi kako nema potrebe da se tretira ova pretpostavka.




*Grafikon 7*

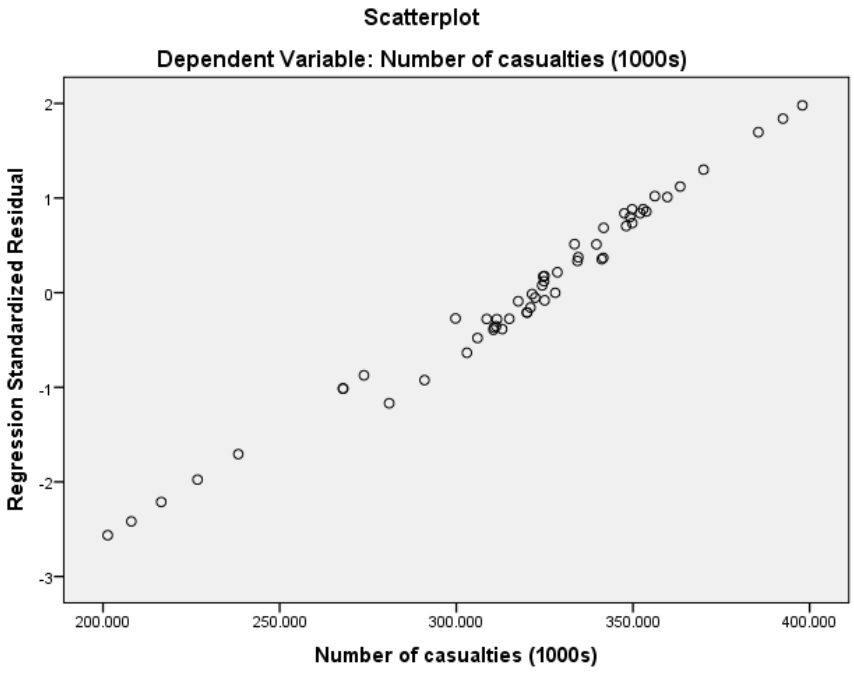
Prilikom sprovođenja regresione analize utvrđeno je da je narušena i pretpostavka o odsutnosti autokorelacije. Kao što je rečeno, za detekciju autokorelacije koristi se kombinacija testovnih i grafičkih metoda. Korišćen je Durbin-Watson test, a vrednost statistike je 2.952 što prelazi kritičnu granicu i ukazuje na postojanje autokorelacije. Iz grafikona se takođe može primetiti da postoji autokorelacija.

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | ,699[a] | ,488 | ,473 | 11657,24823 | 2,952 |

a. Predictors: (Constant), Causalities

b. Dependent Variable: Bilions km driven




*Grafikon 8*

Kao što je napomenuto, tretman autokorelacije predstavlja eliminisanje simptoma autokorelacije, ali se preporučujue da se na prvom mestu spreči da do autokorelacije dođe, nego da se model oslobađa posledica autokorelacije.

Na istoj bazi podataka predstaviće se i primer narušene pretpostavke o multikolinearnosti. Ovaj put koristiće se višestruka linearna regresija. Kriterijumsku varijablu i dalje predstavlja broj saobraćajnih nesreća, a prediktori su broj pređenih kilometara u milijardama kilometara i vreme izraženo u godinama (od 1950. godine do 2004. godine). Iz tabele korelacije može da se primeti kako su godina i broj pređenih kilometara u veoma visokoj korelaciji ($r = .993$), što ukazuje na multikolinearnost.

**Correlations**

| | | Year | Number of casualties (1000s) | Billions km driven |
|---|---|---|---|---|
| Year | Pearson Correlation | 1 | .207 | .993** |
| | Sig. (2-tailed) | | .029 | .000 |
| | N | 55 | 55 | 55 |

| | | | | |
|---|---|---|---|---|
| Number of casualties (1000s) | Pearson Correlation | .207 | 1 | .157 |
| | Sig. (2-tailed) | .029 | | .053 |
| | N | 55 | 55 | 55 |
| Billions km driven | Pearson Correlation | .993** | .157 | 1 |
| | Sig. (2-tailed) | .000 | .053 | |
| | N | 55 | 55 | 55 |

**. Correlation is significant at the 0.01 level (2-tailed).

U analizi linearne regresije, kao deo procedure može da se sprovede „dijagnostika kolinearnosti" promenljivih pri čemu se dobijaju vrednosti Tolerance i VIF. Iz tabele se može primetiti da su vrednosti Tolerance na granici preporučljive (0.10), a da su vrednosti VIF mnogo više od preporučene granice koja iznosi 10, tako da može da se donese zaključak da postoji multikolinearnost.

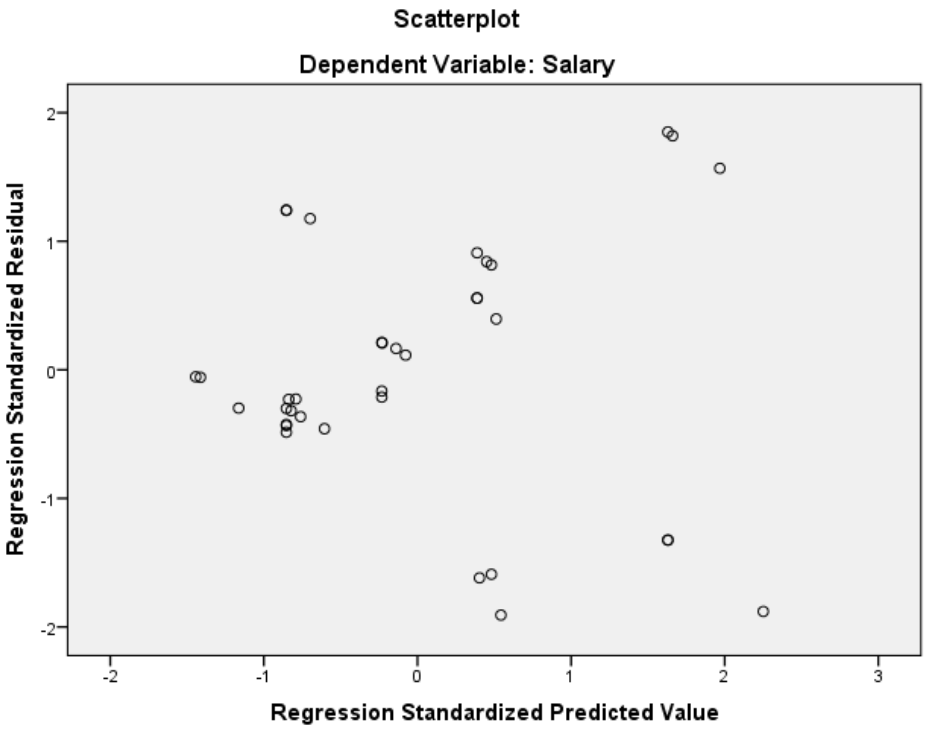
**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | -19167.702 | 5385.877 | | -3.559 | .001 | | |
| | Billions km driven | -1.112 | .323 | -3.602 | -3.441 | .001 | .014 | 73.097 |
| | Year | 10.010 | 2.768 | 3.785 | 3.616 | .001 | .014 | 73.097 |

Kada se u podacima uoči multikolienarnost, uobičajeno je da se iz modela izbaci jedna od promenljivih koje jako koreliraju. Takav tretman je primenjen i ovde. Napravljen je model sa jednom promenljivom, a odabrana je nezavisna promenljiva vreme, odnosno godina. Ova promenljiva ima nešto višu korelaciju sa nezavisnom promenljivom $(r = .207)$. Sada imamo jednostavan model linearne regresije, sa jednom promenljivom i otuda ne postoji multikolinearnost.

## 3.3. Treći primer

U trećem primeru korišćeni su podaci o visini plate i dužini radnog staža. U ovom primeru kriterijumsku varijablu predstavlja visina plate, dok prediktor predstavlja dužina radnog staža, a detektovano je da je narušena pretpostavka o homoskedastičnosti podataka. Ova pretpostavka se najlakše ispituje grafičkom metodom. Iako ima manji broj slučajeva, iz grafikona može da se primeti da varijansa greške nije konstantna, odnosno da nije jednaka za sve vrednosti nezavisne promenljive, na šta nam ukazuje blagi oblik levka, tako da bi se moglo reći da postoji heteroskedastičnost.




*Grafikon 9*

Grafička metoda nije dovoljna da se donese ovakav zaključak. Od testovnih metoda, kao što je navedeno najčešće se koriste Goldfled-Quandtov test, Breusch-Pagan-Godfrey test (u daljem tekstu BPG test) i Whiteov test. Novije verzije statističkog softvera SPSS nude opciju testiranja heteroskedastičnosti. Kao primer odabran je BP test. Testira se nulta hipoteza da postoji homoskedastičnost, odnosno da su greške varijansi jednake, nasuprot alternativne hipoteze da se greške razlikuju. Ukoliko se dobije statistički značajan test, onda može da se odbaci nulta hipoteza o homoskedastičnosti, odnosno donosi se zaključak da postoji heteroskedastičnost.

Kao što se može primetiti iz tabele, test nije dostigao statističku značajnost što bi značilo da je pretpostavka o homoskedastičnosti ispunjena.

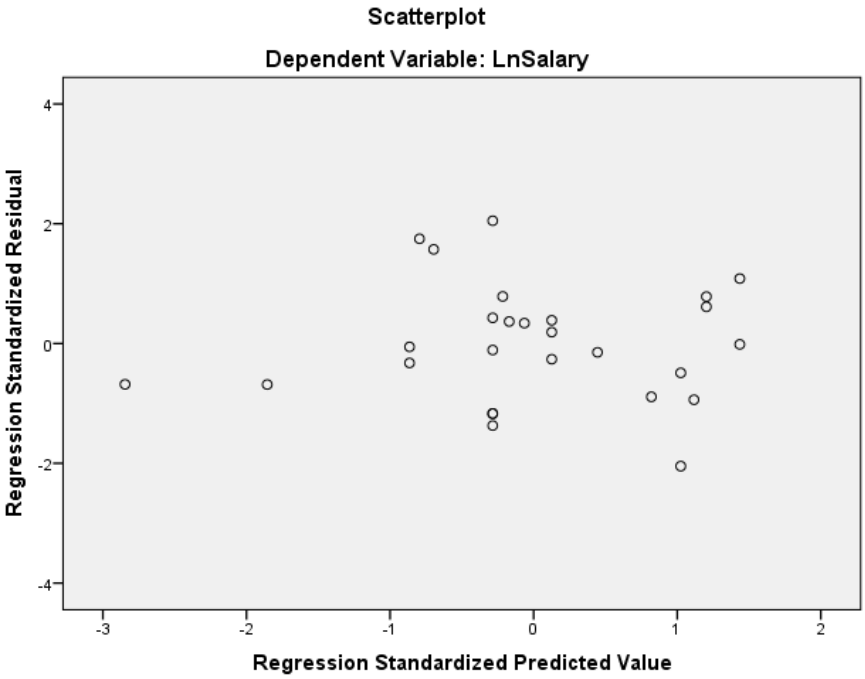
**Breusch-Pagan Test for Heteroskedasticity[a,b,c]**

| Chi-Square | Df | Sig. |
|---|---|---|
| .625 | 1 | .429 |

a. Dependent variable: Salary
b. Tests the null hypothesis that the variance of the errors does not depend on the values of the independent variables.
c. Predicted values from design: Intercept + YOS

Primer pokazuje da test i grafikon pokazuju suprotne rezultate. Zbog toga što grafikon pokazuje postojanje heteroskedastičnosti, a i radi ilustracije tretmana, zavisna promenljiva je logaritmovana.

Nakon logaritmovanja podataka, urađena je još jedna regresiona analiza. Sada je kao zavisna promenljiva odabrana transformisana promenljiva visina plate. Iz dijagrama reziduala može se primetiti da vrednosti nemaju neki karakterističan obrazac, kao i da su jednako raspoređene oko nule, tako da se može zaključiti da je otklonjena heteroskedastičnost.




*Grafikon 10*

# 4. Zaključak

Ovaj rad se bavio narušavanjem pet osnovnih pretpostavki u modelu linearne regresije: *normalnost*, *sredina jednaka nuli, homoskedastičnost, odsustvo autokorelacije* i *nestohastičnost promenljive X*. Ocenjeni su parametri regresije i ispitana su svojstva ovih ocenjivača dobijenih korišćenjem metode najmanjih kvadrata, BLUE metodom i metodom maksimalne verodostojnosti. Ustanovljeno je da ocenjivači dobijeni metodom najmanjih kvadrata imaju sva poželjna svojstva. Pokazalo se da je ocenjivač MAD regresionog koeficijenta asimptotski nepristrasan i normalno distribuiran u uslovima nenormalnosti, kao i da je njegova asimptotska varijansa manja od varijanse ocenjivača metode najmanjih kvadrata za veliki broj na krajevima zadebljanih distribucija. Ocene parametara $\alpha$ i $\beta$ dobijene metodom najmanjih kvadrata u uslovima heteroskedastičnosti su nepristrasne i konzistentne, ali nemaju osobinu efikasnosti i asimptotske efikasnosti, a ocena varijanse je pristrasna.

Cilj rada je bio sprovođenje analize, detekcija narušenih pretpostavki i njihovo uspešno otklanjanje. U prvom primeru se pojavila netipična vrednost, outlier, sa standardizovanom vrednošću reziduala koja je veća od 3, stoga su morale da se posmatraju Cookove distance na osnovu kojih bi se takve vrednosti eliminisale. U datoteci je utvrđeno da postoji jedna vrednost kod koje Cookova distanca pokazuje udaljenost veću od 1 (8.814). Nakon eliminacije ove vrednosti i ponovljene regresione analize, primećeno je da se dijagram rasturanja promenio, a nakon ponovnog crtanja Q-Q plota moglo se zaključiti da ne postoje više vrednosti koje mnogo odstupaju od prave. Drugi primer je ilustrovao zavisnost saobraćajnih nesreća od broja pređenih kilometara. Prilikom provere pretpostavki bilo je utvrđeno da je narušena pretpostavka o normalnosti, stoga se pristupilo analizi kombinovanjem testova i grafičkih pokazatelja. U ovom slučaju su vrednosti deskriptivne statistike dale vrednosti koje zadovoljavaju kriterijume, a Kolmogorov-Smirov test je dao vrednost *p* manju od 0.05, što je uputilo na to da podaci nemaju normalnu raspodelu. Daljim pregledom grafikona primećeno je da histogram nema klasični simetrični oblik zvona i to je ukazalo da podaci nisu normalno distribuirani. Iz dijagrama rasturanja standardizovanih reziduala primećeno je da postoji karakterističan obrazac. Oblik reziduala je formirao obrnuto U, što je još jedan nagoveštaj da podaci nemaju normalnu raspodelu. Na osnovu svega navedenog, konstatovano je da je pretpostavka o normalnosti narušena i da se može pristupiti nekom od tretmana za rešavanje ovog problema. Nakon logaritamske transformacije podataka, grafički pokazatelji su sugerisali da su podaci normalno distribuirani, ali je dijagram rasturanja reziduala i dalje pokazivao da je pretpostavka o normalnosti narušena. Kako je veličina uzorka bila odgovarajuća, pretpostavka o normalnosti je mogla da se zanemari. U ovom primeru je takođe primećeno narušenost pretpostavke o odsutnosti autokorelacije. Korišćeni su grafikoni i Durbin-Watson test čija je vrednost prešla kritičnu granicu. Na istoj bazi je proverena i multikolinearnost, uočena je visoka korelacija između godina i pređenih kilometara. Tretman koji je ovde primenjen je izbacivanje jedne promenljive koja je jako korelirala, i multikolinearnost je eliminisana. U poslednjem primeru je bila narušena pretpostavka o homoskedastičnosti. Blagi oblik levka na grafikonu je naveo na sumnju da varijansa greške nije konstantna, ali to nije bilo dovoljno da se donese konačan zaključak. Breusch-Pagan-Godfrey test nije dostigao statistički značajnu vrednost, odnosno po tom testu ne postoji heteroskedastičnost. Zbog dobijenih suprotnih rezultata, ponovljena je analiza nakon logaritmovanja zavisne promenljive. Kako se iz dijagrama reziduala moglo primetiti da vrednosti nemaju karakterističan obazac, jednako su raspoređene oko nule, zaključeno je da je otklonjena heteroskedastičnost.

Čak i kad se detektuje narušenost neke pretpostavke, nije lako korigovati problem. Mogućnosti statističkog softvera SPSS su velike, ali i dalje nisu svi testovi pokriveni. Ponekad on daje različit rezultat u odnosu na grafičke pokazatelje i tada je neophodno dodatno prodiskutovati rezultate, kako bi se došlo do zaključka.

# 5. Literatura


[1] Jan Kmenta, *Počela Ekonometrije*, Zagreb: Mate, 1997.

[2] Z. Mladenović, *Uvod u ekonometriju*, Centar za izdavačku delatnost Ekonomskog fakulteta, Beograd, 2011.

[3] A.S. Goldberger, *Econometric Theory*, New York: Wiley, 1964

[4] John E. Freund and Ronald. Walpole, *Mathematical Statistics*, 1986

[5] Z. Lozanov-Crvenković, *Statistika*, Novi Sad, 2012

[6] D. Rajter-Ćirić, *Verovatnoća*, Novi Sad, 2009

[7] M. Jovičić, *Ekonometrijski metodi i modeli*, Centar za izdavačku delatnost Ekonomskog fakulteta, Beograd, 2011

[8] G.S. Maddala, *Introduction to Econometrics*, third edition, John Wiley & sons, LTD, 2001

[9] William H. Greene, *Econometric Analysis*, fifth edition, New York, 2003

[10] A.C. Harvey, *The Econometric Analysis of Time Series*, New York: Wiley, 1981

[11] P.J. Huber, *Robust Statistics*, New York: Wiley, 1981

[12] *https://mathstatsresearch.weebly.com/regression.html*

[13] *http://staff.bath.ac.uk/pssiw/stats2/page16/page16.html*

# 6. Biografija

Nina Moldvai je rođena 21. decembra 1991. godine u Novom Sadu. Završila je osnovnu školu „Jožef Atila“, kao i osnovnu muzičku školu „Josip Slavenski“ smer klavir. Potom je pohađala specijalno-matematičko odeljenje gimnazije „Jovan Jovanović Zmaj“ u Novom Sadu gde je maturirala 2010. godine sa odličnim uspehom.

Iste godine se upisuje na Prirodno-matematički fakultet Univerziteta u Novom Sadu, smer primenjena matematika—matematika finansija. Osnovne akademske studije završava 2014. godine kada upisuje master studije primenjene matematike. Iste godine odlazi u Milano, na Erasmus razmenu studenata i tamo polaže ispite na prvoj godini. Sve ispite predviđene planom i programom polaže do 2016. godine čime ostvaruje pravo na odbranu master rada.

Zapošljava se u ProCredit banci na poziciji Savetnika za klijente, gde radi do decembra 2017. godine kada dobija posao u firmi „Synechron“ gde i danas radi kao „Murex test consultant“.

Nina Moldvai,

Novi Sad, 2020. godina

UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
**RBR**

Identifikacioni broj:
**IBR**

Tip dokumentacije: Monografska dokumentacija
**TD**

Tip zapisa: Tekstualni štampani materijal
**TZ**

Vrsta rada: Master rad
**VR**

Autor: Nina Moldvai
**AU**

Mentor: dr Zagorka Lozanov-Crvenković
**MN**

Naslov rada: Narušavanje pretpostavki u linearnoj regresiji
**NR**

Jezik publikacije: srpski (latinica)
**JP**

Jezik izvoda: srpski/engleski
**JI**

Zemlja publikovanja: Srbija
**ZP**

Uže geografsko područje: Vojvodina
**UGP**

Godina: 2020.
**GO**

Izdavač: Autorski reprint
**IZ**

Mesto i adresa: Novi Sad, Departman za matematiku i informatiku, Prirodno-matematički fakultet, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 4

**MA**

Fizički opis rada: (3/61/13/0/0/0)

**FO**

Naučna oblast: Matematika

**NO**

Naučna disciplina: Statistika

**ND**

Ključne reči: Linearna regresija, osnovne pretpostavke, normalnost, homoskedastičnost, heteroskedastičnost, autokorelacija, Durbin-Watsonov test, Breusch-Pagan-Godfreyjev test, Whiteov test, multikolinearnost, SPSS

**PO,UKR**

Čuva se: Biblioteka Departmana za matematiku i informatiku Prirodno-matematičkog fakulteta Univerziteta u Novom Sadu

**ČU**

Važna napomena: Nema

**VN**

Izvod: U prvom delu master rada dat je model jednostruke linearne regresije i definisane su osnovne pretpostavke, pored toga ocenjeni su parametri regresije na tri načina i ispitana njihova svojstva. U drugom delu rada su posmatrane osobine ovih ocenjivača u slučaju narušenih pretpostavki. Navedeni su testovi koji se koriste, ali je objašnjen i tretman u slučaju narušenosti jedne ili više pretpostavki. Treći deo ovog master rada se bavio simulacijom primera u softveru SPSS.

**IZ**

Datum prihvatanja teme

od strane NN veća: 25.09.2020.

**DP**

Datum odbrane: 2020.

**DO**

Članovi komisije:

**KO**

Predsednik: dr Ljiljana Gajić, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Mentor: dr Zagorka Lozanov-Crvenković, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu
Član: dr Ivana Štajner-Papuga, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

UNIVERSITY OF NOVI SAD
FACULTY OF NATURAL SCIENCE AND MATHEMATICS
KEY WORDS DOCUMENTATION

Accession number:
**ANO**

Identification number:
**INO**

Document type: Monograph documentation
**DT**

Type of record: Textual printed material
**TR**

Contents code: Master thesis
**CC**

Author: Nina Moldvai

**AU**

Mentor: Zagorka Lozanov-Crvenković, PhD
**MN**
Title: Violation of assumptions in linear regression
**TI**

Language of text: Serbian (latin)
**LT**

Language of abstract: Serbian/English
**LA**

Country of publication: Serbia
**CP**

Locality of publication: Vojvodina

**LP**

Publication year: 2020.
**PY**

Publisher: Author's reprint
**PU**

Publ. place: Novi Sad, Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Dositeja Obradovića Sq.4

**PP**
Physical description: (3/61/13/0/0/0/0)
**PD**

Scientific field: Mathematics
**SF**

Scientific discipline: Statistics
**SD**
Subject Key words: Linear regression, assumptions, normality, homoscedasticity, heteroscedasticity, autocorrelation, Durbin-Watson, Breusch-Pagan-Godfrey, White, multicollinearity, SPSS

**SKW, UC**

Holding data: The Library of the Department of Mathematics and Informatics, Faculty of Science, University of Novi Sad
**HD**

Note: None
**N**
Abstract: The first part of this master thesis presents basics of linear regression model and its assumptions. Parameters of regression and its properties are analysed in three different ways. Second part is about violation of assumptions. It is described how to detect and treat one or more violations. In the end, there are simulations of examples done in SPSS software.

**AB**
Accepted on the Scientific board on: 25.09.2020.
**AS**

Defended: 2020.

**DE**

Thesis Defend board:

**D**

President: Ljiljana Gajić, PhD, full professor, Faculty of Science, University of Novi Sad

Mentor: Zagorka Lozanov-Crvenković, PhD, full professor, Faculty of Science, University of Novi Sad

Member: Ivana Štajner-Papuga, PhD, full professor, Faculty of Science, University of Novi Sad