Univerzitet u Beogradu
Matematički fakultet

Master Rad

Problemi sa preduslovima u višestrukom linearnom modelu i njihovo prevazilaženje

Mentor:
Dr Vesna Jevremović

Studentkinja:
Maida Hodžić

Beograd, 2015.

Sadržaj

I Linearni modeli **3**

1 Linearna regresija **3**

2 Jednostruki linearni regresioni modeli **5**

3 Višestruki linearni regresioni modeli **8**

4 Pretpostavke višestrukog linearnog modela **10**
4.1 Pretpostavke o parametrima linearnog modela 10
4.2 Pretpostavke o slučajnom odstupanju ε 12

5 Vektorska reprezentacija modela **13**

II Problemi u pretpostavkama modela **16**

6 Pretpostavke o grešci **16**
6.1 Konstantna disperzija 17
6.1.1 Metoda odmerenih najmanjih kvadrata (Weighted Least Squares) 21
6.2 Normalna raspodela grešaka 25
6.3 Korelisane greške 29
6.3.1 Metoda uopštenih najmanjih kvadrata (Generalized Least Squares) 33

7 Pretpostavke o nezavisnim promenljivim X_i **39**
7.1 Greške u promenljivim X_i 39
7.2 Kolinearnost promenljivih X_i 44

8 Linearne transformacije **50**
8.1 Goodness of fit 54

III Primena opisanih rešenja **60**

9 Provera pretpostavki o greškama ε_i **60**
9.1 Problem nekonstantne disperzije grešaka 60
9.2 Provera normalne raspodele grešaka 63

10 Provera pretpostavki o promenljivim X_i **63**
10.1 Problem korelisanosti promenljivih 63

11 Transformacija promenljivih u modelu **66**

Uvod

Regresioni modeli se koriste u mnogim naučnim oblastima. Koriste se za predstavljanje podataka dobijenih u istraživanjima i donošenje zaključaka na osnovu tih modela. Zato je jako bitno da modeli predstavljaju podatke na najbolji mogući način. Postoji nekoliko načina procene tačnosti modela. Ovaj rad će predstaviti probleme koji se mogu javiti u pretpostavkama i ponuditi načine na koji ti problemi mogu da se reše. Rad takođe sadrži i puno praktičnih primera.

U prvom dijelu rada su definisani linearni modeli. Navedene su glavne osobine jednostrukih i višestrukih linearnih regresionih modela, kao i pretpostavke o parametrima i slučajnom odstupanju u modelu.

Drugi deo opisuje probleme koji se mogu javiti u pretpostavkama navedenim u prvom dijelu kao i moguća rešenja za te probleme.
Sva rešenja su objašnjena na primerima sa graficima. Većina primera je rađena u programskom jeziku R, zbog podataka i implementiranih funkcija.

Treći deo pokazuje kako se opisana rešenja mogu koristiti u realnim primerima.

Poglavlje I
Linearni modeli

1 Linearna regresija

Regresiona analiza je tehnika koja se koristi da bi se predstavila veza između dve promenljive koja opusuje vrijednost zavisne promenljive Y na osnovu izabrane vrednosti nezavisne promenljive X. Regresija predstavlja statističku metodu kojom se proverava i opisuje povezanost između različitih pojava.
Linearna regresija je data modelom:

$$Y = aX + b + \varepsilon \tag{1.1}$$

Relacija (1.1) definiše vezu između dve promenljive koje su linearno zavisne.
U ovom modelu:

- a je procenjena vrijednost zavisne promenljive Y gde regresiona linija seče y-osu kada je X nula.
- b je nagib linije, ili prosečna promena Y za svaku jediničnu promenu (povećanje ili smanjenje) nezavisne promenljive X.

Pojava na osnovu kojih se dobija predviđanje može biti više od jedne. Neka su to: $X_1, X_2, ..., X_k$, su nezavisne (determinističke) promenljive ili faktori, a pojava koja zavisi od ovih promenljivih, Y zove se zavisna (stohastička) promenljiva. Zavisnost spomenutih pojava se opisuje regresionim modelom:

$$Y = \varphi(X_1, X_2, ..., X_k) + \varepsilon \tag{1.2}$$

gde je ε slučajna greška, a φ linearna funkcija.
Radi mogućnosti predviđanja potrebno je pronaći koeficijente u linearnoj funkciji φ kojom bi se definisala međusobna zavisnost promenljivih.

Linearna regresija je bila prvi tip zavisnosti dvije promenljive koja je detaljno proučavana i koja se često koristila u praksi. Razlog za ovo je taj što se modeli koji linerano zavise od svojih nepoznatih parametara lakše predstavljaju nego modeli sa nelinearnom zavisnošću od parametara i promenljive X. Takođe, statistička svojstva rezultirajućih promenljivih se lakše određuju.

Linearna regresija ima mnogo praktičnih primjena. Većina primera primene linearne regresije spada u jednu od sledeće dvije kategorije:

- Ako je cilj predviđanje ili prognoza, linearna regresija se može koristiti za podešavanje prediktivnog modela prema posmatranom skupu podataka vrijednosti Y i X. Nakon razvoja ovakvog modela, ako je data vrijednost za promenljivu X: X^* bez pripadajuće vrijednosti promenljive Y: Y^*, dobijeni model se može koristiti za predviđanje vrijednosti Y^*.
- Ako imamo promenljivu Y i veći broj promenljivih $X_1, ..., X_p$ koje mogu biti povezane sa Y, možemo koristiti lineranu regresionu analizu za određivanje jačine relacije između Y i X_j, za procenu koji je X_j uopšte vezan za Y, i da bismo odredili koji podskupovi od X_j sadrže najbitnije informacije o Y, tako da, kad je jedan od tih podskupova poznat, ostali više ne daju dovoljno korisne informacije. [1] [1]

[1] "Višestruka regresija", Paul D. Allison, Poglavlje 1, Str 1

2 Jednostruki linearni regresioni modeli

To je najjednostavniji regresioni model, a opštiji model višestruke linearne regresije je u mnogo čemu samo njegovo logično uopštenje. Modelom jednostruke regresije izražena je veza između zavisne promenljive Y i nezavisne promenljive X, koja se može formalno opisati izrazom:

$$Y = f(X) + \varepsilon \tag{2.1}$$

Dva su osnovna cilja koja se žele postići stvaranjem regresionog modela:

1. odrediti tip realne funkcije f (linearna, kvadratna, eksponencijalna, logaritamska,) koja najbolje opisuje vezu između posmatranih promenljivih;
2. procijeniti parametre te funkcije tako da zbir kvadrata svih rezidualnih odstupanja ε bude što manji.

Ako je model dat sa:

$$Y = aX + b + \varepsilon \tag{2.2}$$

treba odrediti parametre a i b tako da disperzija svih rezidualnih odstupanja ε bude minimalna.
Svakim statističkim istraživanjem dobijamo ukupno n različitih uređenih parova vrijednosti promenljive X i vrijednosti promenljive Y, tj. dobijamo uređene parove $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
Budući da je grafik svake linearne funkcije prava, to praktično znači da tražimo jednačinu prave koja će najbolje aproksimirati skup tačaka predstavljen dijagramom rasipanja.

Za svaki od tih parova vrijedi jednakost (2.2):

$$y_i = a \cdot x_i + b + \varepsilon_i, \tag{2.3}$$

pri čemu je $i = 1, 2, \ldots, n$, a ε_i rezidualno odstupanje, odnosno "promašaj" koji činimo ako zamijenimo vrijednost y_i računski dobijenom vrijednošću $a \cdot x_i + b$.

Time, dobijamo sistem od n linearnih jednačina sa $n + 2$ nepoznate promenljive: $a, b, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ čije se rešenje dobija metodom

najmanjih kvadrata. Rezultat se traži po formuli:

$$\min \sum (y_i - ax_i - b)^2 \tag{2.4}$$

Ako sa y_i obeležimo sve stvarne vrijednosti promenljive Y, sa $\hat{y}_i$ procenjena (očekivana) vijednost promenljive Y, osnovna ideja metode najmanjih kvadrata je postići da zbir kvadrata odstupanja empirijskih vrijednosti y_i od očekivanih vrijednosti $\hat{y}_i$ te promenljive bude minimalan. Ovom metodom dobijamo ocene parametara modela a i b i obeležavamo ih sa: $\hat{a}$ i $\hat{b}$:

$$\hat{a} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{2.5}$$

pri čemu su $\bar{x}$ i $\bar{y}$ redom aritmetičke sredine vrijednosti promenljive X, odnosno Y.[2] [2]
Tako dobijamo model jednostruke linearne regresije:

$$\hat{Y} = \hat{a} \cdot X + \hat{b} \tag{2.6}$$

Vrijednost b naziva se naziva konstantni član. Njegova teorijska interpretacija je sledeća:
b je očekivana vrijednost zavisne promenljive Y kada je vrijednost nezavisne promenljive X jednaka nuli. To se lako vidi uvrštavanjem vrijednosti $X = 0$ u jednačinu (2.5). Ovaj parametar nema veliko praktično značenje jer često nema razloga da se koristi (npr. u modelima koji mere zavisnost novca izdvojenog za prehranu od ukupnih mesečnih primanja jer je u takvim slučajevima praktično besmisleno procenjivati iznos za prehranu osobe bez ikakvih ukupnih mesečnih primanja).

Vrijednost a naziva se regresioni koeficijent i to je ujedno najvažniji parametar dobijenog modela. Njegova teorijska interepretacija je sledeća: vrijednost a je očekivana prosečna promena zavisne promenljive Y kada se nezavisna promenljiva X poveća za jednu jedinicu mere.

Pozitivna vrijednost regresionog koeficijenta upućuje na to da porast vrijednosti nezavisne promenljive X uzrokuje porast vrijednosti

[2] "Metoda najmanjih kvadrata",F. Bruckler, Str 2

zavisne promenljive Y, a negativna vrijednost regresionog koeficijenta upućuje na to da porast vrijednosti nezavisne promenljive X uzrokuje smanjenje vrijednosti zavisne promenljive Y.

Primer 1 *Vremena procesuiranja ulaznih i izlaznih podataka 7 programa su data na sledeći način:[3]*

$$(14, 2), (16, 5), (27, 7), (42, 9), (39, 10), (50, 13), (83, 20)$$ [3]

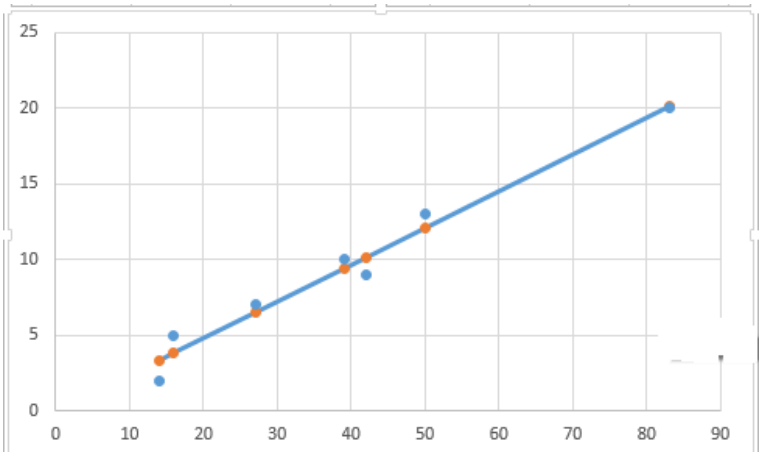
Da bismo odredili model koji će tačno predstavljati ove podatke, računamo:
$n = 7, \Sigma xy = 3375, \Sigma x = 271, \Sigma x^2 = 13855, \sum y = 66, \sum y^2 = 828, \bar{x} = 38.71, \bar{y} = 9.43.$

Iz ovog dobijamo koeficijente modela:

$$b_1 = 0.2438, b_2 = -0.0083.$$

Traženi linearni model ima oblik:

$$\hat{Y} = 0.2438 \cdot X - 0.0083$$



Slika 1: Dobijeni model u odnosu na podatke iz uzorka

[3]"Ostali regresioni modeli", R.Jain, Poglavlje 1, Str 5

3 Višestruki linearni regresioni modeli

Višestruki linearni model predstavlja vezu zavisne promenljive Y i nezavisnih, kontrolisanih, promenljivih $X_1, ..., X_k$. Opšti oblik višestrukog linearnog modela je:

$$Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \tag{3.1}$$

gde su $X_1, ..., X_k$ poznati vektori konstanti, $\beta_1, ..., \beta_k$ nepoznati parametri i ε slučajna greška.

Cilj ispitivanja višestrukom linearnom regresijom je definisati uzoračku hiperravan regresije sa najmanjim mogućim rezidualima. U tu svrhu potrebno je oceniti nepoznate koeficijente regresije, $\beta_1, ..., \beta_k$, na neki od sledećih načina:

1. **Tačkasta ocena** Metodom tačkastog ocenjivanja dobija se ocena, statistika: $\hat{\beta}\_i$, i=0,...,k, za svaki nepoznati koeficijent, β_i, i=0,...,k. Na osnovu realizovanog uzorka dobijaju se realizovane vrijednosti ovih statistika. Ukoliko je ocenjena vrijednost blizu stvarne vrijednosti koeficijenta, navedena metoda predstavlja dobar način ocene. U daljem radu će biti objašnjeno nekoliko metoda za tačkasto ocenjivanje.
2. **Intervalna ocena**- Intervalnom ocenom dobija se interval poverenja koji sa verovatnoćom $1-\alpha$ sadrži željeni parametar, koeficijent. Realizovani interval na osnovu uzorka sadrži ili ne sadrži koeficijent, stoga se za slučajni interval kaže samo da sa $1-\alpha$ verovatnoćom sadrži koeficijent. Postoje dvostrani i jednostrani intervali poverenja. Dvostrani su dati u obliku

   $$P[L < \theta < U] = 1 - \alpha, \tag{3.2}$$

   gde je L donja granica, a U gornja granica intervala poverenja i obično su u obliku:

   $$\hat{\theta} - T * SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + T * SE(\hat{\theta}) \tag{3.3}$$

   $\hat{\theta}$ - ocijenjeni parametar
   T - tablična vrijednost
   $SE(\hat{\theta})$ standardna greška parametra.
3. **Testiranje hipoteza pomoću testova** - Ukoliko želimo da odredimo da li je parametar tj. koeficijent veći ili manji od

određene vrijednosti q, možemo da koristimo testiranje hipoteza i testove.
Postavljaju se dvije hipoteze za parametar θ, nulta hipoteza H_0 i alternativna hipoteza H_1.
Cilj testiranja je utvrditi da li ima dokaza za odbacivanje hipoteze H_1. U tu svrhu fiksira se veličina testa α (obično je 0.05 ili 0.01) i p-vrednost: veličina kritične oblasti ako joj je granica registrovana vrednost test statistike.
U slučaju desne jednostruke kritične oblasti, ako je $p \leq \alpha$ odbacuje se hipoteza H_0 , a ako je $p > \alpha$ hipoteza H_0 se ne odbacuje.[8] [4]

[4]"Modeliranje višestrukom linearnom regresijom", A.Vaš, Poglavlje 1, Str 3

4 Pretpostavke višestrukog linearnog modela

4.1 Pretpostavke o parametrima linearnog modela

Za ocene koeficijenata $\beta_i, i = 0, ..., k$ u višestrukom linearnom modelu:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon \tag{4.1}$$

poželjno je da imaju sledeće osobine:

1. **Nepristrasnost** odnosno da je očekivana vrijednost ocene $\hat{\beta}$ jednaka pravoj vrijednosti parametra β:

   $$E(\hat{\beta}) = \beta \tag{4.2}$$

   Ukoliko postoji razlika između ove dvije vrijednosti, kažemo da je ocena parametra $\hat{\beta}$ pristrasna, a $E(\hat{\beta}) - \beta$ predstavlja pristrasnost ocene.

2. **Efikasnost** - odnosno da je ocena nepristrasna i ima manju disperziju od svih ostalih nepristrasnih ocena tog parametra. U zavisnosti na osnovu čega se posmatra, efikasnost može biti apsolutna i relativna. Ukoliko postoje dvije ocene uzorka tada se za ocenu koja ima manju disperziju kaže da je relativno efikasnija. Za pronalaženje najefikasnije ocene koristi se teorema **Rao-Cramera** koja sledi.

   Neka je $X_1, ..., X_n$ prost slučajan uzorak za obeležje X u čijoj raspodeli figuriše nepoznati parametar θ. Neka je gustina raspodele obeležja X: $g(x; \theta), x \epsilon R, \theta \epsilon \Theta$, gde je Θ parametarski prostor.
   Funkcija raspodele u ovom slučaju je:

   $$L(X; \theta) = \prod_{j=1}^{n} g(X_j; \theta) \tag{4.3}$$

   Pod uslovima **regularnosti**:

   (a) skup $X : L(X; \theta) > 0$ ne zavisi od θ

   (b) L je dvaput diferencijabilna po θ

važi nejednakost Rao-Cramera:

$$D(\hat{\theta}) \geq -\frac{1}{E(\frac{\partial^2 \ln L}{\partial \theta^2})} = \frac{1}{E(\frac{1}{\frac{\partial \ln L}{\partial \theta}})^2} \tag{4.4}$$

gde je $\hat{\theta}$ nepristrasna ocena za parametar θ.

3. **BLUE** (best linear unbiased estimator) najbolja linearna nepristrasna ocena. Ocena sa ovom osobinom treba da zadovolji sledeće uslove:
   - Ocena je linearna funkcija opažanja iz uzorka: x_i i y_i.
   - Ocena je nepristrasna.
   - Ocena ima manju disperziju od svih ostalih ocena za taj parametar.

Navedene osobine važe samo u slučaju kada je uzorak mali. Mogu se primenjivati u slu caju da je obim uzorka manji od 30. Ukoliko je obim uzorka veliki, poželjne su sledeće asimptotske osobine ocena parametara:

1. Asimptotska nepristasnost - podrazumeva da se povećanjem veličine uzorka dobija što bolja ocena koeficijenta, tj. očekivana vrijednost ocene teži stvarnoj vrijednosti parametra kako veličina uzorka raste:
$$\lim_{n\to\infty} E(\hat{\beta}) = \beta \tag{4.5}$$
2. Konzistentnost - ocena $\hat{\beta}$ teži u verovtnoči ka koeficijentu β i za svako $\varepsilon > 0$ važi:
$$\lim_{n\to\infty} P(|\hat{\beta} - \beta| < \varepsilon) = 1. \tag{4.6}$$
3. Asimptotska efikasnost - podrazumeva da ocena parametra ima sledeće osobine:
   - Asimptotsku raspodelu koja teži raspodeli parametra sa konačnom sredinom i disperzijom.
   - Konzistentnost

- Manju disperziju od disperzija ostalih ocena tog parametra. [6] [5]

4.2 Pretpostavke o slučajnom odstupanju ε

Metode za ocenu parametara se zasnivaju na sledećim pretpostavkama o slučajnoj greški ε:

1. Matematičko očekivanje ili srednja vrednost odstupanja ε_i je jednaka 0:
$$E(\varepsilon_i) = 0, i = 0, 1..., k \tag{4.7}$$
2. Disperzija odstupanja ε_i je konstantna, tj. homoskedastična:
$$D(\varepsilon_i) = S^2, i = 1, ..., k \tag{4.8}$$
Ako disperzija nije konstantna tada imamo heteroskedastičnost.
3. Kovarijansa odstupanja ε_i i ε_j je jednaka nuli, odnosno greške nisu korelisane:
$$E(\varepsilon_i\varepsilon_j) = 0, i = 1, ..., k; j = 1, ..., k; i \neq j \tag{4.9}$$
4. Odstupanje: ε_i nije korelisano sa promenljivom X_j:
$$E(\varepsilon_i X_j) = 0, i = 1, .., k; j = 1, ..., s \tag{4.10}$$
5. Slučajna greška ε ima normalnu raspodelu $N(0, S^2)$. [6] [6]

Ovi uslovi ne važe uvijek. U slučaju kada nisu ispunjeni, primenjuju se razne metode da bi se napravio odgovarajući linearni model. Ovaj rad će opisati rešenja mogućih problema u modelu kada osnovne pretpostavke o promenljivim i slučajnim greškama ne važe.

[5] "Višestruka linearna regresija: procena i osobine", E.Uriel, Poglavlje 3, Str 11
[6] "Višestruka linearna regresija: procena i osobine", E.Uriel, Poglavlje 3, Str 12

5 Vektorska reprezentacija modela

Neka je Y n-dimenzioni vektor koji se može predstaviti kao višestruki linerni u obliku:

$$Y = \beta_1 X_1 + ... + \beta_k X_k + \varepsilon \tag{5.1}$$

gde su $X_1, ..., X_k$ poznati vektori konstanti i vektor grešaka $\varepsilon \sim N_n(0, \sigma^2 I_n)$ (greške ε_i imaju normalnu raspodelu $N(0, \sigma^2)$), a $\beta_1, ..., \beta_k$ nepoznati parametri.

Ovaj model se može predstaviti i na drugačiji način:

$$Y = \theta + \varepsilon \tag{5.2}$$

gde je:

$$\theta \epsilon V = L(X_1, ..., X_k) i \varepsilon \sim N(0, \sigma^2 I_n) \tag{5.3}$$

Odnosno, Y se može predstaviti preko elemenata linerarnog prostora promenljivih X_i.[4] [7]

Neka se matrica X sastoji od svih vektora $X_i, i = 1, ..., k$: $X = (X_1, ..., X_k)$. Tada je $\theta = X\beta$, i ako su $X_1, ..., X_k$ linearno nezavisni,važi:

$$(X'X)^{-1}X'\theta = \beta \tag{5.4}$$

Metoda najmanjih kvadrata dovodi do procene θ preko $\hat{\theta}$, gde $\theta = \hat{\theta}$ minimizira izraz:

$$\|Y - \theta\|^2 \equiv Q(\theta) \tag{5.5}$$

gde je $\theta \epsilon V$.

Odnosno, metodom najmanjih kvadrata dolazimo do ocene $\hat{\theta} = P(Y|V) = \hat{Y}$.

U slučaju kada je X punog ranga, tj. X_i su linearno nezavisne, važi:

$$\hat{\theta} = X(X'X)^{-1}X'Y i \hat{\beta} = (X'X)^{-1}X'Y \tag{5.6}$$

[7] "Linearni statistički modeli", J. Stapleton, Poglavlje 3, Str 88

Tada važi:

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon \tag{5.7}$$

Ako su vektori u kolonama X ortogonalni, tada je:

$$\hat{Y} = \sum P(Y|X_j) = \sum \hat{\beta}_j X_j \tag{5.8}$$

gde je:

$$\hat{\beta}_j = (Y, X_j)/\|X_j\|^2 = \beta_j + (\varepsilon, X_j)/\|X_j\|^2 \tag{5.9}$$

Odnosno, svaka komponenta $\hat{\beta}$ je jednaka zbiru odgovarajuće komponente β i linearne kombinacije komponenti ε_i. Ako definišemo $M = X'X$, iz prethodnih jednakosti dobijamo:

- $E(\hat{\beta}) = \beta + M^{-1}X'E(\varepsilon)$, tako da ako je $E(\varepsilon) = 0$, tada je $E(\hat{\beta}) = \beta$.
- $Cov(\hat{\beta}) = Cov(M^{-1}X'\varepsilon) = M^{-1}X'Cov\varepsilon XM^{-1}$, tako da ako je $Cov(\varepsilon) = \sigma^2 I_n$, tada je $Cov(\hat{\beta}) = M^{-1}\sigma^2$.
- Ako ε ima višedimenzionu normalnu raspodelu, tada i $\hat{\beta}$ višedimenzionu normalnu raspodelu.

U praksi je teško da se pretpostavi da ε ima više dimenzionu normalnu raspodelu. Ipak, centralna granična teorema daje da za veliko n, k različitih linearnih kombinacija nezavisnih komponenti ε ima približno višedimenzionu normalnu raspodelu čak i kad samo ε nema. [4]

[8] U nastavku rada višestruki linearni modeli će uglavnom biti predstavljeni na sledeći način:

$$Y = \beta \cdot X + \varepsilon \tag{5.10}$$

gde Y predstavlja zavisnu promenljivu, X matricu koja se sastoji od n promenljivih X_i, β vektor koeficijenata koji stoje uz svaku promenljivu, i ε vektor grešaka ε_i.

[8] "Linearni statistički modeli", J. Stapleton, Poglavlje 3.1, Str 89

Literatura u ovom poglavlju:

U ovom poglavlju su korišćene sledeće jedinice iz literature:

- "Višestruka regresija", P. Allison [1],
- "Metoda najmanjih kvadrata",F. Bruckler [2],
- "Modeliranje višestrukom linearnom regresijom", A.Vaš [8],
- "Višestruka linearna regresija: procena i osobine", E.Uriel [6],
- "Linearni statistički modeli", J. Stapleton [4].

Poglavlje II

Problemi u pretpostavkama modela

Procena i rezultati dobijeni pomoću regresionog modela zavise od nekoliko pretpostavki. Ove pretpostavke moraju da važe da bi model predstavljao podatke na pravi način. Moguće probleme u višestrukim linearnim regresionim modelima možemo da podelimo u tri grupe:

- Greška - Pretpostavljamo da za slučajne greške u modelu važi: ε_i ima $N(0, \sigma^2 I)$ raspodelu, odnosno da su greške nezavisne međusobno, imaju konstantnu dispeziju i normalno su raspodeljene.
- Model - Pretpostavljamo da je struktura modela $E(Y) = X\beta$ ispravna.
- Neuobičajena opažanja - Nekada se samo nekoliko elemenata uzorka ne uklapa u model. Ova opažanja mogu da utiču na izbor i valjanost modela.

U slučaju problema u pretpostavkama, model treba da se popravi, što znači da pravljenje linearnog modela predstavlja iterativni i interaktivni proces.

6 Pretpostavke o grešci

Greške se ne mogu posmatrati, pa zato koristimo reziduale $\widehat{\varepsilon}$ da bismo proverili da li odstupanja imaju normalnu raspodelu i da li su međusobno nezavisna. Reziduali nisu isto što i greške, pošto nemaju sve iste osobine.[5] [9] Ako posmatramo višestruki regresioni model: $Y = X\beta + \varepsilon$, ocena za β dobijena metodom najmanjih kvadrata: $\hat{\beta}$ je ona ocena za koju je suma kvadrata grešaka najmanja. Dakle, ocena dobijena metodom najmanjih kvadrata minimizira sledeću vrijednost:

[9] "Linearni modeli u R-u", J. Faraway, Poglavlje 4, Str 53

$$\sum \varepsilon_i^2 = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta) \tag{6.1}$$

Kada ovaj izraz diferenciramo po β i izjednačimo sa nulom, dobijamo sledeći izraz za $\hat{\beta}$:

$$X^T X\hat{\beta} = X^T Y \tag{6.2}$$

Ove jednačine se nazivaju normalne jednačine. Pod pretpostavkom da je matrica $X^T X$ inverzibilna, dobijamo izraz:

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = H * Y \tag{6.3}$$

gde je H hat-matrica. Hat matrica je ortogonalna projekcija Y na prostor definisan vektorima X.
Koristeći matricu H i vrijednosti iz uzorka, definišemo reziduale na sledeći način:

$$\hat{\varepsilon} = y - \hat{y} = (I - H)y = (I - H)X\beta + (I - H)\varepsilon = (I - H)\varepsilon \tag{6.4}$$

Disperziju reziduala $\hat{\varepsilon}_i$ takođe možemo izraziti koristeći prethodno dobijenu jednakost:

$$D(\hat{\varepsilon}) = D((I - H)\varepsilon) = (I - H)D(\varepsilon) = (I - H)\sigma^2 \tag{6.5}$$

ako pretpostavimo da je $D(\varepsilon) = \sigma^2 I$.
Dakle, može da se desi da zbog razlike reziduala i stvarnih grešaka u modelu, greške imaju konstantnu disperziju i nisu korelisane, a da za reziduale važi obratno. Srećom ta razlika je uglavnom suviše mala i nema uticaja na testiranje osobina greške.[5] [10]

6.1 Konstantna disperzija

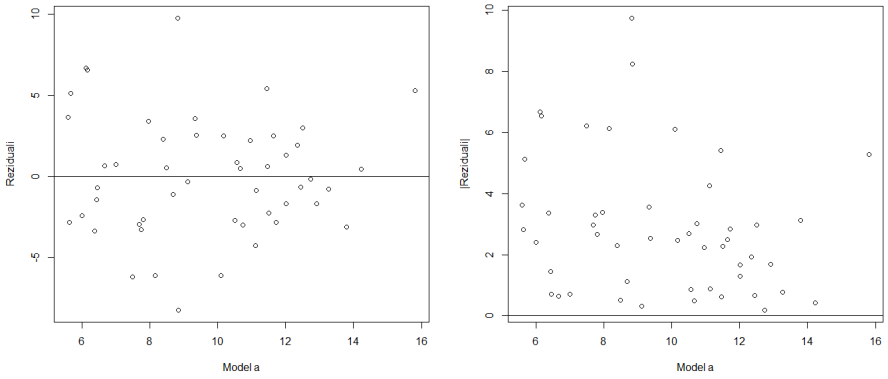
Nije moguće proveriti pretpostavku o konstantnoj disperziji posmatrajući samo reziduale jer će neki biti mali a neki veliki, ali to neće ništa dokazati, već mora da se proveri da li disperzija reziduala zavisi od vrednosti promenljivih u modelu.
Treba proveriti kakva je veza između reziduala $\hat{\varepsilon}$ i $\hat{y}$ i $\hat{x}$. Ovo se može videti na grafiku.

[10] "Linearni modeli u R-u", J. Faraway, Poglavlje 4.1, Str 53

Ako je sve u redu, na grafiku se vidi konstantna disperzija u vertikalnom pravcu ($\hat{\varepsilon}$) i tačke treba da budu simetrično raspoređene oko nule. Ono na šta treba obratiti pažnju su nekonstantna disperzija i nelinearnost - jer se u tom slučaju mora transformisati model.

Primer 2 *Posmatramo podatke iz seta "savings", paketa faraway u R-u. Podaci predstavljaju prosečnu uštedevinu u odnosu na prihode kod populacije ispod 15 i iznad 75 godina, tokom perioda 1960-1970 godine.*

```
data(savings)
a<-lm(sr~pop15+pop75+dpi+ddpi,savings)
par(mfrow=c(1,2))
plot(fitted(a),residuals(a),xlab="Model a",
ylab="Reziduali")
abline(h=0)
plot(fitted(a), abs(residuals(a),xlab="Model a",
ylab="|Reziduali|")
abline(h=0)
```



Slika 2: Reziduali

Drugi grafik na slici 1. koristimo da bismo proverili konstantnost disperzije, ali i prvi grafik je neophodan da bismo proverili nelinearnost. Na datim graficima se ne vidi ništa što bi ukazivalo na to da disperzija reziduala nije konstantna.[5] [11]

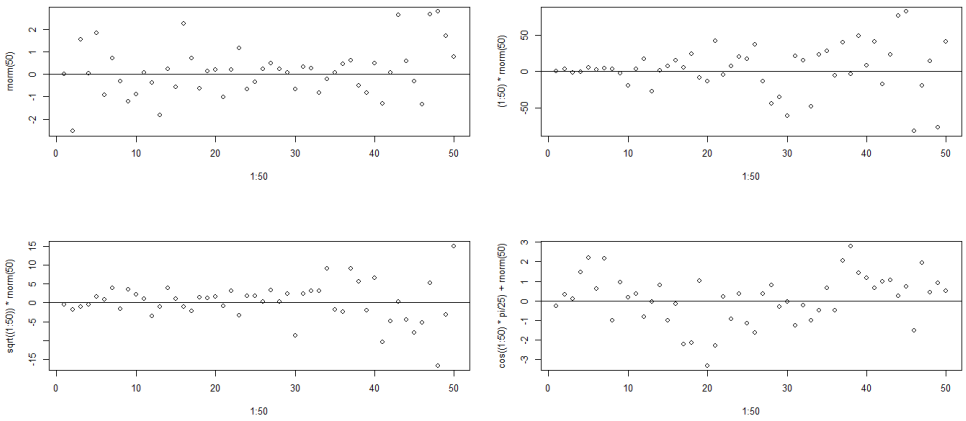
[11]"Linearni modeli u R-u", J. Faraway, Poglavlje 4, Str 55

U sledećem primeru na slici 2. se vidi kako izgledaju grafici za svaki od četiri slučaja redom:

1. Konstantna disperzija
2. Jako nekonstantna disperzija
3. Srednje nekonstantna disperzija
4. Nelinearnost

Primer 3 `par(mfrow=c(2,2))`

```
plot(1:50,rnorm(50), abline(h=0))
plot(1:50,(1:50)*rnorm(50), abline(h=0))
plot(1:50, sqrt((1:50))*rnorm(50), abline(h=0))
plot(1:50, cos((1:50)*pi/25)+rnorm(50), abline(h=0))
```

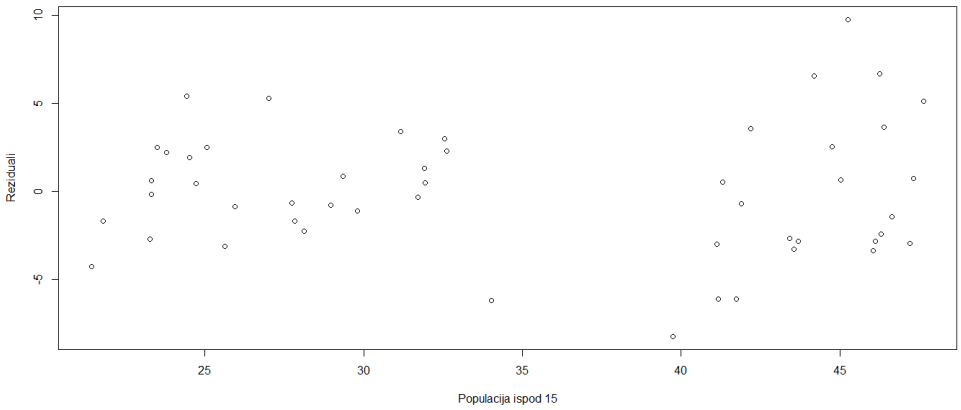


Slika 3: Primeri reziduala sa različitim disperzijama

Konstantnost disperzije se može proveriti i korišćenjem dva nezavisna slučajna uzorka iz normalne raspodele. Koristićemo F test da uporedimo disperzije iz dva uzorka.

Primer 4 *Uporedićemo disperziju iz dva uzorka iz podataka "savings" iz paketa "faraway".*

```
> data(savings)
> plot(savings$pop15,residuals(a),xlab="Populacija
 ispod 15",ylab="Reziduali")
```



Slika 4: Reziduali iz dva uzorka

Na grafiku se vidi da postoje dvije grupe tačaka. Zato upoređujemo disperzije reziduala iz 2 uzorka:

```
> var.test(residuals(a)[savings$pop15>35],
  residuals(a)[savings$pop15<35])

        F test to compare two variances

data:  residuals(a)[savings$pop15 > 35] and
residuals(a)[savings$pop15 < 35]
F=2.7851, num df=22, denom df=26, p-value=0.01358
alternative hypothesis: true ratio of variances
is not equal to 1

95 percent confidence interval:
 1.240967 6.430238
sample estimates:
ratio of variances
          2.785067
```

Vidimo da postoji velika razlika u disperzijama, na osnovu čega možemo da zaključimo da disperzija nije konstantna.

Postoje 2 pristupa za rešavanje problema nekonstantne disperzije:

1. Metoda odmerenih najmanjih kvadrata (weighted least squares - WLS)
2. Transformacija parametara modela

Transformacija promenljivih će biti detaljno opisana u poglavlju "Linearne transformacije", dok će ovde biti naveden postupak za rešavanje problema konstantnosti disperzije reziduala korišćenjem metode najmanjih kvadrata.

6.1.1 Metoda odmerenih najmanjih kvadrata (Weighted Least Squares)

U slučaju da su slučajne greške nekorelisane, a njihova disperzija iako nije konstantna ima neku poznatu formu, koristimo ovu metodu za određivanje najboljeg modela.
Pretpostavimo da je disperzija grešaka:

$$D(\varepsilon) = \sigma^2 \Sigma \tag{6.6}$$

gde je σ^2 nepoznato ali je matrica Σ poznata. Kada je Σ dijagonalna matrica, greške su nekorelisane ali ne moraju da imaju istu disperziju.
U tom slučaju možemo da napišemo:

$$\Sigma = diag(\frac{1}{w_1}, ..., \frac{1}{w_n}) \tag{6.7}$$

gde su w_i takvi da je:

$$S = diag(\sqrt{\frac{1}{w_1}, ..., \frac{1}{w_n}}. \tag{6.8}$$

Možemo da napravimo regresiju $\sqrt{w_i}y_i$ po $\sqrt{w_i}x_i$ (u tom slučaju kolona jedinica u matrici X mora da se zamijeni sa $\sqrt{w_i}$. Slučajevi sa malom disperzijom dobijaju veliku težinu, a sa velikom disperzijom malu težinu. Neki primeri:

1. Greške proporcionalne vrednostima nezavisne promenljive: $w_i = x_i^{-1}$.
2. Kada su Y_i prosečne vrijednosti iz n_i posmatranja, tada je $D(y_i) = D(\varepsilon_i) = \sigma^2/n_i$, što sugeriše da je $w_i = n_i$. Vrednosti koje su prosečne se lako uklapaju, ali treba voditi računa

da je dispezija rezultujuće promenljive stvarno proporcionalna veličini grupe.

Primer 5 *Posmatramo očekivanu dužinu života u različitim državama.*
Na prvi pogled, može da se zaključi da mere w_i treba da budu jednake populaciji tih država, ali primećujemo da postoje mnogi drugi uzroci varijacije u očekivanoj dužini života pored veličine populacije. Postavljanje uslova $w_i = n_i$ će verovatno odgovarati za malo n_i.

Kada se koriste w_i, reziduali moraju da se modifikuju. U tom slučaju koristimo $w_i\hat{\varepsilon}_i$ za dijagnozu modela.[5] [12]

U slučaju kada oblik disperzije ε nije potpuno poznat, možemo da modeliramo Σ koristeći mali broj parametara. Na primer:

$$D(\varepsilon_i) = \gamma_0 + \gamma_1 x_1 \tag{6.9}$$

se može iskoristiti u datoj situaciji. Algoritam iterativno premerene metode najmanjih kvadrata (IRWLS) je:

1. Početi sa $w_i = 1$.
2. Koristiti metodu najmanjih kvadrata za procenu β.
3. Koristiti reziduale za procenu γ, na primer regresija $\hat{\varepsilon}^2$ po x.
4. Preračunati mere w_i i vratiti se na korak 2.

Proces se ponavlja do konvergencije rezultata.
Postoje neki problemi na koje treba obratiti pažnju kad se koristi ova metoda. Procena za γ nije u potpunosti tačna, što dalje utiče na ocenu parametra β.

Još jedan pristup rešavanju problema grešaka koje nemaju konstantu disperziju jeste da se modelira disperzija i da se zajedno procijene regresiona funkcija i parametri koristeći metod baziran na verovatnoći.

Primer 6 *Posmatramo odnos određenih čestica pri sudaru sa protonima. Cilj eksperimenta je bio da se ispitaju teorije o jakoj vezi*

[12] "Linearni modeli u R-u", J. Faraway, Poglavlje 6.2, Str 94

određenih čestica u prirodi. Za crossx promenljivu se veruje da je linearno povezana sa inverznom vijednošću promenljive energije (u podacima prikazanim dole energy kolona već ima inverzne vrijednosti energije). Kolona sd sadrži standardna odstupanja rezultujuće promenljive.
Učitavamo podatke iz ovog eksperimenta iz baze strongx. *Prikazano je prvih deset elemenata iz baze.*

```
> data(strongx)
> strongx
   momentum energy crossx sd
1         4  0.345    367 17
2         6  0.287    311  9
3         8  0.251    295  9
4        10  0.225    268  7
5        12  0.207    253  7
6        15  0.186    239  6
7        20  0.161    220  6
8        30  0.132    213  6
9        75  0.084    193  5
10      150  0.060    192  5
```

Definišemo model i mere w_i:

```
> g <- lm(crossx  energy, strongx, weights=sd-2)
> summary(g)5.2. WEIGHTED LEAST SQUARES 63
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 148.47 8.08 18.4 7.9e-08
energy 530.84 47.55 11.2 3.7e-06
Residual standard error:1.66 on 8 degrees of freedom
Multiple R-Squared: 0.94, Adjusted R-squared: 0.932
F-statistic: 125 on 1 and 8 degrees of freedom,
p-value: 3.71e-06
```

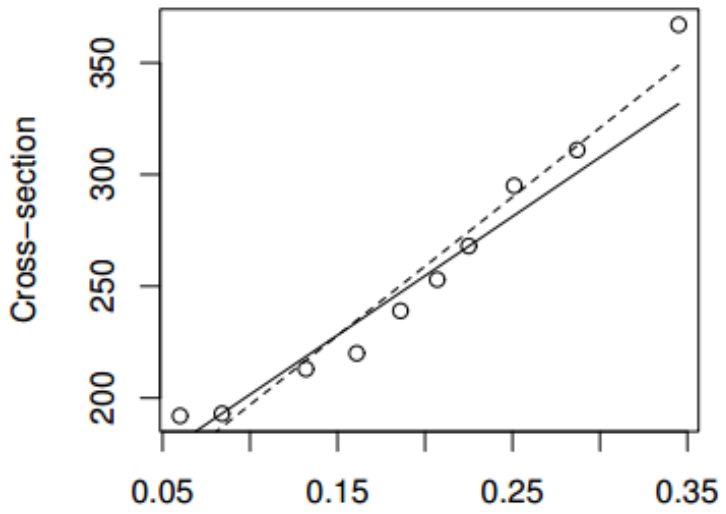
Kada uporedimo ovu procenu modela sa modelom dobijenim bez računanja mera:

```
> gu <- lm(crossx  energy, strongx)
```

```
> summary(gu)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 135.0 10.1 13.4 9.2e-07
energy 619.7 47.7 13.0 1.2e-06

Residual standard error: 12.7 on 8 degrees of freedom
Multiple R-Squared: 0.955, Adjusted R-squared: 0.949
F-statistic: 169 on 1 and 8 degrees of freedom,
p-value: 1.16e-06
```

i to predstavimo na grafiku (Slika 5):

```
> plot(crossx  energy, data=strongx)
> abline(g)
> abline(gu,lty=2)
```



Slika 5: Upoređivanje modela dobijenih na 2 različita načina

Model koji nije dobijen metodom odmerenih kvadrata na grafiku izgleda kao da bolje predstavlja podatke (isprekidana linija), ali za manje vrijednosti energije, disperzija je manja, a model koji je dobijen metodom odmerenih kvadrata bolje predstavlja ove vrijednosti.[11] [13]

[13]"Metoda uopštenih najmanjih kvadrata", I. Ruczinski, Poglavlje 5.1, Str 59

Ovo se može implementirati u R-u koristeći funkciju *gls()* iz biblioteke *nlme*. Ova funkcija konstruiše linearni model koristeći metodu uopštenih najmanjih kvadrata (Generalised least squares) koja će biti opisana u nastavku. Ona se koristi za rešavanje problema konstantnosti disperzije i korelisanosti reziduala.

6.2 Normalna raspodela grešaka

Svi testovi i intervali poverenja koje koristimo se zasnivaju na pretpostavci da reziduali imaju normalnu raspodelu. Normalnost reziduala se može proceniti koristeći Q-Q grafik. On poredi reziduale sa opažanjima iz normalne raspodele. Crtamo grafik za reziduale koristeći funkciju $\Phi^{-1}(\frac{i}{n+1})$ za $i = 1, 2, ..., n$.

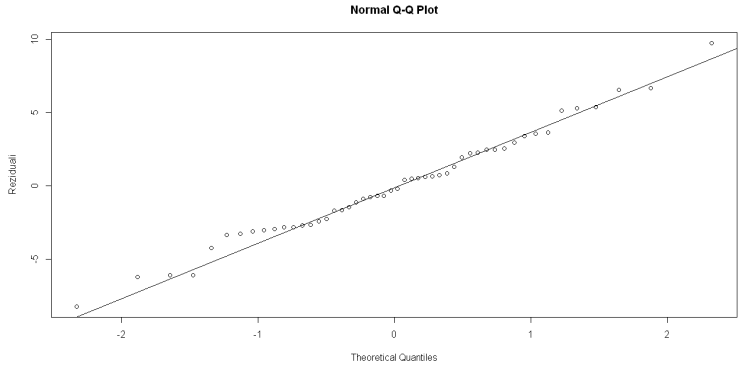
Pretpostavimo da posmatramo slučajan uzorak $W_1, ..., W_n$ za neku neprekidnu raspodelu; tj. $F(x) = \Phi(\frac{x-\mu}{\sigma})$ za neko μ i σ. Neka su $W_{(1)} < W_{(2)} < ... < W_{(n)}$ odgovarajuće statistike poretka. Pošto $F(W_j)$ imaju uniformnu raspodelu na $[0, 1]$, može se pokazati da je $E(W_j)$ je približno $F^{-1}(u_j)$, za $u_j = (j - 1/2)/n$.

Pošto, u normalnoj raspodeli $F^{-1}(u_j) = \mu + \sigma\Phi^{-1}(u_j)$ aproksimirano očekivanje će biti: $E(W_j) = \mu + \sigma\Phi^{-1}(u_j)$, tako da parovi $(Z_j, W_{(j)})$ predstavljaju promenljive koje su vezane jednostrukom linearnom regresijom, gde je $Z_j = \Phi^{-1}(u_j)$. Tako da ako predstavimo ove parove na grafiku, oni bi trebalo da formiraju približno pravu liniju, sa odstupanjem σ i srednjom vrednošću μ.

Primer 7 *Proveravamo normalnost grešaka u podacima iz baze* savings.

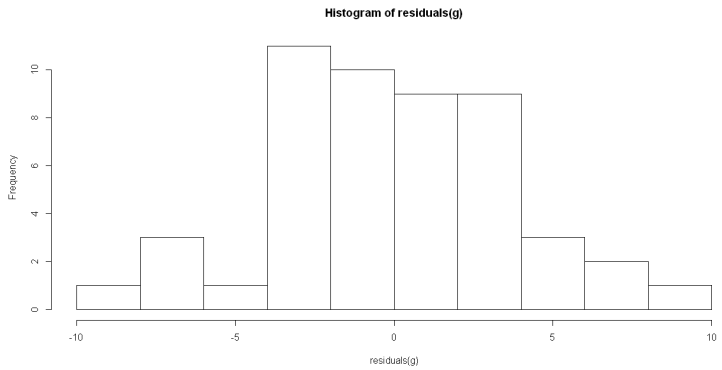
```
data(savings)
g<-lm(sr~pop15+pop75+dpi+ddpi, savings)
qqnorm(residuals(g), ylab="Reziduali")
qqline(residuals(g))
```

Normalni reziduali bi trebalo da prate liniju na grafiku (Slika 6). U ovom primeru reziduali se ponašaju kao da imaju normalnu raspodelu. Histogrami se mogu koristiti za prikaz reziduala, ali nisu prikladni,



Slika 6: Q-Q grafik

kao što se može videti na ovom primeru (Slika 7).

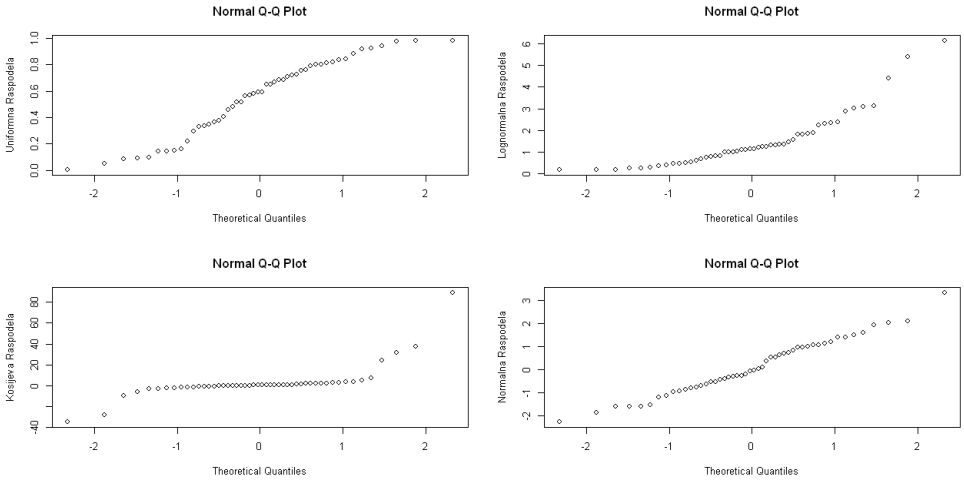


Slika 7: Histogram

Nije uvijek lako zaključiti da postoji problem u Q-Q graficima. Nekada su ekstremni slučajevi znak raspodele sa velikim repom kao u Košijevoj raspodeli ili su prosto autlajeri - tj. vrijednosti koje se puno razlikuju od većine ostalih iz uzorka. Ovo se proverava tako što otklonimo ekstremne slučajeve koji se ne nalaze na pravoj liniji iz grafika. U slučaju da kada otklonimo samo te tačke iz grafika, druge tačke postaju izraženije, onda je problem verovatno u raspodeli sa dugim repom.

Primer 8 *Primer* Q-Q *grafika za neke od raspodela.*

1. *Normalna raspodela*
2. *Log-normalna raspodela - primer raspodele sa srednjim repom*
3. *Košijeva raspodela - primer raspodele sa dugim repom*
4. *Uniformna raspodela - primer raspodele sa kratkim repom*



Slika 8: Q-Q grafici za različite raspodele

Kada greške nisu normalne, procena metodom najmanjih kvadrata nije optimalna. Takođe, testovi i intervali poverenja neće biti tačni. Ipak, samo raspodele sa dugim repom izazivaju velika odstupanja u tačnosti. Mala odstupanja od normalne raspodele se lako mogu ignorisati i što je veći obim uzorka, to je manji problem ako reziduali zaista nemaju normalnu raspodelu.

Kada se otkrije problem sa normalnom raspodelom reziduala, samo rešenje zavisi od tipa problema koji se pojavi.

- Za raspodele sa malim repom, posledice nenormalosti grešaka nisu ozbiljne i mogu se ignorisati (primer: uniformna raspodela).
- Za greške sa dužim repom na jednoj strani, transformacija rezultujuće promenljive može da reši problem (primer: log-normalna raspodela).

- Za greške sa dugim repom na obe strane, možemo da prihvatimo nenormalnost i da baziramo zaključke na pretpostavci druge raspodele ili da koristimo metode biranja uzoraka kao što su *bootstrap* ili testovi permutacije (primer: Košijeva raspodela). [5] [14]

  Bootstrap je metoda kojom se na osnovu raspoloživih podataka iz nekog uzorka kreira veliki broj novih uzoraka, istog obima kao i početni uzorak, slučajnim biranjem sa vraćanjem iz skupa raspoloživih podataka. Ovo znači da svaki element ima jednaku verovatnoću da uđe u uzorak i da neki element može da se pojavi više puta a neki nijednom.

  Alternativno, mogu se koristiti robusne metode, u kojima autlajeri ne utiču na rezultat.

U drugim načinima dijagnostikovanja problema u raspodeli grešaka se takođe predlaže promena modela. U novom, promenjenom modelu je moguće da se neće pojaviti problem sa normalnom raspodelom grešaka.

Formalan test (sa test statistikom i p- vrijednošću) za normalnost raspodele je **Shapiro-Wilk-ov** test. Nulta hipoteza u ovom testu je da promenljiva koja se ispituje ima normalnu raspodelu. Test statistika je:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.10}$$

gde je $x_{(i)}$ i-ta statistike poretka, $\bar{x}$ srednja vrijednost podataka u uzorku, a konstante a_i su definisane na sledeći način:

$$(a_1, ..., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}} \tag{6.11}$$

gde je $m = (m_1, ..., m_n)^T$, a $m_1, ...m_n$ su srednje vrijednosti statistika poretka nezavisnih i jednako raspodeljenih podataka izvučenih iz populacije sa standarnom normalnom raspodelom. V je matrica kovarijacije tih statistika poretka.

[14] "Linearni modeli u R-u", J. Faraway, Poglavlje 4, Str 59

Primer 9 *Koristeći Shapiro-Wilk test proveravamo normalnu raspodelu reziduala u modelu na podacima iz baze* savings *iz paketa* faraway.

```
data(savings)
g<-lm(sr~pop15+pop75+dpi+ddpi, savings)

  shapiro.test(residuals(g))

        Shapiro-Wilk normality test

data:  residuals(g)
W = 0.987, p-value = 0.8524
```

Nulta hipoteza je da su reziduali sa normalnom raspodelom. Vidimo da je p *vrijednost velika, pa prihvatamo ovu hipotezu.*

Ovaj test se može koristiti sa Q-Q grafikom, pošto sama p vrijednost ne pokazuje šta treba da se promijeni.
Za velike uzorke, čak i mala odstupanja od normalne raspodele će se primetiti, ali sam efekat grešaka koje nemaju normalnu raspodelu je zanemarljiv u tom slučaju.
Za male uzorke, svi formalni testovi se teže primenjuju.

6.3 Korelisane greške

Kada konstruišemo linearni model pretpostavljamo da su greške nekorelisane, ali za povezane podatke to ne mora uvijek da bude tačno, i zato treba da se proveri.

Postoje dva načina da se proveri korelisanost grešaka:

- Grafičke provere uključuju grafike na kojima se predstavlja $\widehat{\varepsilon}$ u odnosu na vrijeme i $\widehat{\varepsilon_i}$ u odnosu na $\widehat{\varepsilon_{i-1}}$.
- Durbin-Watson-ov test koristi statistiku:

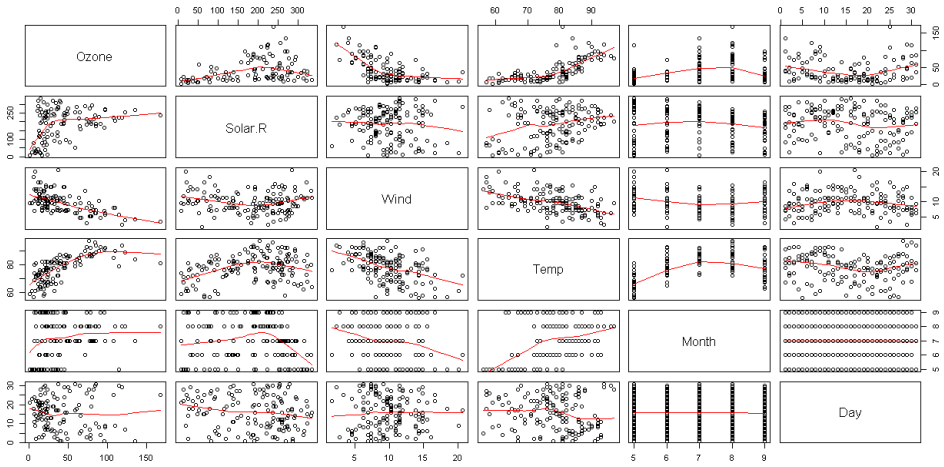
$$DW = \frac{\sum_{i=2}^{n}(\widehat{\varepsilon_i} - \widehat{\varepsilon_{i-1}})^2}{\sum_{i=1}^{n}\widehat{\varepsilon_i}^2} \tag{6.12}$$

gde su $\hat{\varepsilon}_i$ reziduali u datom linearnom modelu.

Primer 10 *Posmatramo podatke dobijene istraživanjem životne sredine (prikazano je prvih 5 podataka iz baze), gde su praćene 4*

promenljive: ozon, radijacija, temperatura i brzina vetra, tokom 153 dana u Njujorku (Slika 8).

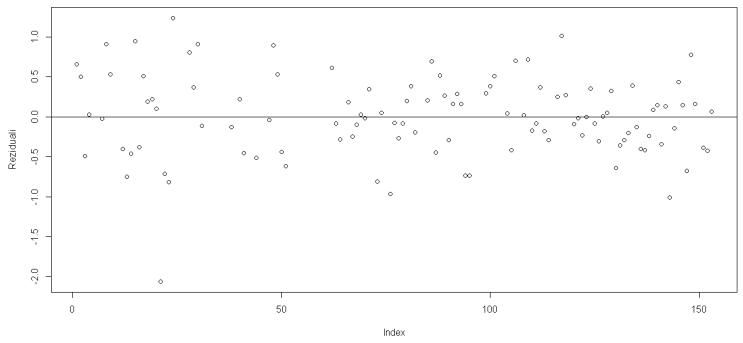
```
data(airquality)
airquality
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149  12.6  74     5   3
4    18     313  11.5  62     5   4
5    NA      NA  14.3  56     5   5

pairs(airquality, panel=panel.smooth)
```



Slika 9: Odnos promenljivih u modelu

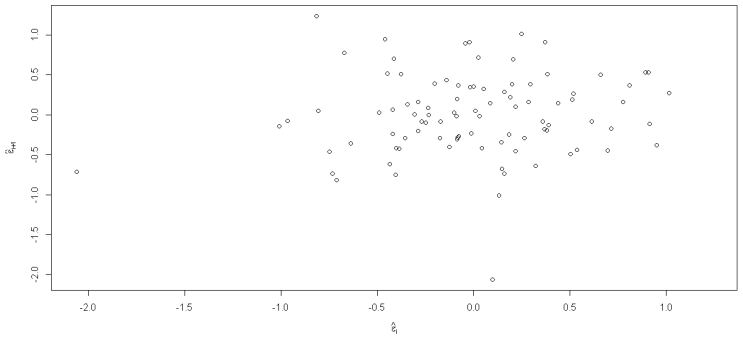
Pravimo model i predstavljamo reziduale na grafiku da bismo proverili da li postoji korelisanost grešaka. Sami podaci u ovoj bazi nisu potpuno i neke vrijednosti nedostaju (predstavljene su kao NA *vrijednosti), pa ih isključujemo.*

```
g<-lm(log(Ozone)~Solar.R + Wind + Temp, airquality,
na.action=na.exclude)
plot(residuals(g),ylab="Reziduali")
abline(h=0)
```



Slika 10: Reziduali

Da postoji korelacija među rezidualima, videli bismo ili duže nizove reziduala iznad ili ispod linije na slici 9. Ako ovi efekti nisu jaki, to će se teško primetiti na grafiku. Na grafiku vidimo da su tačke raspodeljene ravnomerno oko 0, pa možemo da pretpostavimo da je ovde sve u redu.
Često je bolje da se predstave samo reziduali na grafiku (Slika 11):



Slika 11: $\hat{\varepsilon_i}$ u odnosu na $\hat{\varepsilon_{i+1}}$

Vidimo da nema problema sa korelacijom na osnovu ovih grafika.

To možemo da proverimo i računanjem Durbin-Watson-ove statistike:

```
dwtest(Ozone~Solar.R+Wind+Temp, data=na.omit(airquality))
```

```
        Durbin-Watson test

data:  Ozone ~ Solar.R + Wind + Temp
DW = 1.9355, p-value = 0.3347
alternative hypothesis: true autocorrelation is greater
 than 0
```

Na osnovu p *vrednosti zaključujemo da nema dokaza za korelisanost. Ipak, rezultate treba posmatrati sa rezervom, pošto smo isključili sve* NA *vrijednosti iz baze.*

U slučaju kada postoji korelacija među greškama, može se koristiti metoda uopštenih najmanjih kvadrata (GLS) da bi se rešio taj problem.

6.3.1 Metoda uopštenih najmanjih kvadrata (Generalized Least Squares)

U modelu:

$$Y = \theta + \varepsilon = \sum_{1}^{k} \beta_j x_j \tag{6.13}$$

pretpostavke $E(\varepsilon) = 0$ i $Cov(\varepsilon) = \sigma^2 I_n$ daju, po teoremi Gauss-Markov-a , optimalno rešenje metodom najmanjih kvadrata.

Ipak, postoje mnogi slučajevi kada uslov $Cov(\varepsilon) = \sigma^2 I_n$ nije realan. Ako komponente Y odgovaraju opažanjima u određenim tačkama u vremenu, što je često slučaj sa ekonomskim podacima, često se dešava da postoji korelacija među opažanjima. Veće vrijednosti za komponente θ dovode do većih odgovarajućih vrijednosti za $Cov(\varepsilon)$.

Nepoznavanje tačne vrijednosti za $\sum = Cov(\varepsilon)$ može da izazove ozbiljne probleme. Postoje slučajevi kada je $\sum$ poznato ili se zna do na konstantu. U tom slučaju možemo da sredimo model tako da možemo da primenimo teoremu Gauss-Markov-a.
Neka je:

$$Y = \theta + \varepsilon = \sum_{1}^{k} \beta_j x_j = X\beta + \varepsilon, Cov(\varepsilon) = \Sigma = \sigma^2 A \tag{6.14}$$

gde je A poznata $n \times n$ nesingularna matrica, $I\sigma^2$ nepoznata konstanta (ovo je disperzija u slučaju da su dijagonalni elementi A svi jednaki 1).
Neka je $BB^T = A$, gde je B $n \times n$ matrica. Pošto A mora biti pozitivno definitna i $(BF)(BF)^T = A$, za bilo koju ortogonalnu matricu F, beskonačno takvih matrica B može biti izabrano. Tako da ga možemo izabrati i da ima neke posebne osobine.
Za takvo B, gde važi $BB^T = A$ definišemo:

$$Z = B^{-1}Y = B^{-1}\theta + B^{-1}\varepsilon = \sum_{1}^{k} \beta_j (B^{-1}x_j) + B^{-1}\varepsilon = \sum_{1}^{k} \beta_j w_j + \eta \tag{6.15}$$

gde je $w_j = B^{-1}x_j$ i $\eta = B^{-1}\varepsilon$.

Tada je:

$$E(\eta) = B^{-1}E(\varepsilon) = 0 \tag{6.16}$$

, a kovarijacija:

$$Cov(\eta) = B^{-1}(\sigma^2 A)B^{T-1} = \sigma^2 B^{-1}AB^{T-1} = \sigma^2 I_n \tag{6.17}$$

tako da Z zadovoljava standardnu linearnu hipotezu koju smo naveli na početku.
U modelu važi:

$$E(Z) = B^{-1}\theta\epsilon L(w_1, w_2, ..., w_k). \tag{6.18}$$

Procena za β dobijena metodom najmanjih kvadrata je onda:

$$(W^T W)^{-1}W^T Z = \hat{\beta} \tag{6.19}$$

gde je $W = (w_1, w_2, ..., w_k) = B^{-1}X$. Sledi da je:

$$\hat{\beta} = (X^T B^{T-1}B^{-1}X)^{-1}X^T B^{T-1}B^{-1}Y\hat{\beta} = (X^T A^{-1}X)^{-1}X^T A^{-1}Y \tag{6.20}$$

Pošto je $\hat{\beta}$ funkcija od A (kao i X i Y), ne zavisi od samog rastavljanja A: $BB^T = A$. I pošto je ovo procena za β metodom najmanjih kvadrata, ima optimalne uslove koji slede iz teoreme Gauss-Markova. Svaki test sa hipotezama se može izvršiti koristeći Z. Naravno, svaka funkcija se može preraditi tako da se umesto Z koristi Y, uvođenjem smene: $Z = B^{-1}Y$. [4] [15]
Glavni problem u primeni metode uopštenih kvadrata jeste što Σ nije uvijek poznato. U tom slučaju moramo prvo da ga procijenimo. Ceo proces korišćenja ove metode će biti prikazan u sledećem primeru.

Primer 11 *Posmatramo Longley regresione podatke u R - u. Posmatramo to kako se broj zaposlenih menjao od 1947. do 1962. u zavisnosti od doma ćeg proizvoda i populacije preko 14 godina. Podaci su se originalno pojavili u Longley-u 1967. godine.*
Učitavamo podatke i pravimo standardni model:

```
data(longley)
g<-lm(Employed~GNP + Population,longley)
> summary(g)
```

[15] "Linearni statistički modeli", J. Stapleton, Poglavlje 4.3, Str 177

```
Call:
lm(formula = Employed ~ GNP + Population, data = longley)
Residuals:
     Min       1Q   Median       3Q      Max
-0.80899 -0.33282 -0.02329  0.25895  1.08800
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.93880   13.78503   6.452 2.16e-05 ***
GNP          0.06317    0.01065   5.933 4.96e-05 ***
Population  -0.40974    0.15214  -2.693   0.0184 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5459 on 13 degrees of freedom
Multiple R-squared:  0.9791, Adjusted R-squared:  0.9758
F-statistic: 303.9 on 2 and 13 DF,  p-value: 1.221e-11
```

Proveravamo korelaciju među promenljivim u modelu.

```
>cor(longley$GNP,longley$Population)
[1] 0.9910901
```

Vidimo da je korelacija među promenljivim jako pozitivna.

U ovakvim primerima, gde su podaci skupljani tokom dužeg vremena, korelacija među greškama se može proveriti pravljenjem sledećeg modela:

$$\varepsilon_{i+1} = \rho\varepsilon_i + \delta_i \tag{6.21}$$

gde je $\delta_i \sim N(0, \tau^2)$. Σ je u tom slučaju: $\Sigma_{ij} = \rho^{|i-j|}$.[5] [16]

Nastavak primera: *Proveravamo korelaciju među greškama u datom modelu.*

```
cor(residuals(g)[-1], residuals(g)[-16])
[1] 0.3104092
```

Dakle, ρ će imati vrijednost 0.3104092, u ovom primeru.
Sledeći korak je da konstruišemo Σ matricu i primenimo metod

[16]"Linearni modeli u R-u", J. Faraway, Poglavlje 6.1, Str 90

uopštenih kvadrata da bismo procijenili β i greške.
Primenjujemo postupak opisan u teorijskom dijelu i nalazimo sve matrice potrebne za računanje ocena.

```
> Ro<-cor(residuals(g)[-1], residuals(g)[-16])
> X<-model.matrix(g)
> Sigma<- diag(16)
> Sigma<-Ro^abs(row(Sigma)- col(Sigma))
> Sigi<-solve(Sigma)
> Xt<-t(X)
> XtXi<-solve(Xt%*%Sigi%*%X)
> Y<-longley$Employed

> beta<-XtXi%*%(Xt%*%Sigi%*%Y)
> beta
                   [,1]
(Intercept) 94.89887712
GNP          0.06738948
Population  -0.47427390
```

Računamo i greške za ove koeficijente:

```
> res<-longley$Employed - X%*%beta
> sig<- sqrt((t(res)%*%Sigi%*%res)/g$df)
> sig
         [,1]
[1,] 0.542443
> sqrt(diag(XtXi))*sig
(Intercept)         GNP  Population
```

Ovaj metod je implementiran u R-u u paketu *nlme*, koji su napravili Pinheiro and Bites 2000.
Ovaj paket se koristi za konstruisanje linearnih i nelinearnih modela i proveravanje njihovih osobina.

Primenjujemo funkciju iz nlme paketa da bi ocijenili koeficijente u datom modelu.

```
> library(nlme)
```

```
> g1<-gls(Employed ~ GNP+Population,
  correlation = corAR1(form= ~ Year), data=longley)
> summary(g1)
Generalized least squares fit by REML
  Model: Employed ~ GNP + Population
  Data: longley
       AIC      BIC    logLik
  44.66377 47.48852 -17.33188

Correlation Structure: AR(1)
 Formula: ~Year
 Parameter estimate(s):
      Phi
0.6441692

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 101.85813 14.198932  7.173647  0.0000
GNP           0.07207  0.010606  6.795485  0.0000
Population   -0.54851  0.154130 -3.558778  0.0035

 Correlation:
           (Intr) GNP
GNP         0.943
Population -0.997 -0.966

Standardized residuals:
       Min         Q1        Med        Q3       Max
-1.5924564 -0.5447822 -0.1055401 0.3639202 1.3281898

Residual standard error: 0.689207
Degrees of freedom: 16 total; 13 residual

> intervals(g1)
Approximate 95% confidence intervals

 Coefficients:
                 lower         est.       upper
(Intercept) 71.18320461 101.85813306 132.5330615
```

```
GNP          0.04915865   0.07207088   0.0949831
Population  -0.88149053  -0.54851350  -0.2155365
attr(,"label")
[1] "Coefficients:"

 Correlation structure:
       lower      est.     upper
Phi -0.44239 0.6441692 0.9644304
attr(,"label")
[1] "Correlation structure:"

 Residual standard error:
    lower      est.     upper
0.2479415 0.6892070 1.9157998
```

Vidimo da je $\rho = 0.64417$*, ali pošto je interval za njega* $(-0.44, 0.96)$*, da nije značajno veći od 0, pa stoga u možemo da zaključimo da u novom modelu nema korelacije među greškama.* [5] [17]

[17]"Linearni modeli u R-u", J. Faraway, Poglavlje 6.1, Str 92

7 Pretpostavke o nezavisnim promenljivim X_i

Za nezavisne promenljive X_i u modelu se pretpostavlja:

- da se računaju bez greške ili sa jako malom greškom
- da su međusobno nezavisne

7.1 Greške u promenljivim X_i

Pri definisanju modela pretpostavljamo da su vrednosti za promenljive X_i izračunate bez greške. Ali ovo često nije slučaj u praksi. Ne treba mešati greške u vrednostima koja koristimo za modeliranje Y sa pretpostavkom da je slučajna promenljiva. Za posmatrane podatke, X promenljive se mogu posmatrati kao slučajne promenljive, ali regresiona veza je uslovljena fiksiranom vrednošcu za X. U ovom slučaju imamo pretpostavku da je Y generisano u odnosu na fiksiranu vrednost za promenljive X_i.

Posmatramo parove (x_i^0, y_i^0) za $i = 1, 2, ..., n$ koji su povezani sa pravim vrijednostima (x_i^A, y_i^A):

$$y_i^0 = y_i^A + \varepsilon_i x_i^0 = x_i^A + \delta_i \tag{7.1}$$

gde su greške ε i δ nezavisne. Veza između promenljivih X i Y se može predstaviti na sledeći način:

$$y_i^A = \beta_0 + \beta_1 x_i^A \tag{7.2}$$

ali mi vidimo samo (x_i^0, y_i^0). Kada predstavimo sve to zajedno, dobijamo:

$$y_i^0 = \beta_0 + \beta_1 x_i^0 + (\varepsilon_i - \beta_1 \delta_i) \tag{7.3}$$

Pretpostavimo da koristimo metod najmanjih kvadrata da procenimo koeficijente β_0 i β_1. Takođe pretpostavimo da važi:
$E(\varepsilon_i) = E(\delta_i) = 0$ i da je $D(\varepsilon_i) = \sigma_\varepsilon^2$, $D(\delta_i) = \sigma_\delta^2$. Neka je:

$$\sigma_x^2 = \Sigma(x_i^A - \bar{x}^A)^2/2\sigma_{x\delta} = cov(x^A, \delta) \tag{7.4}$$

Za posmatrane podatke, σ_x^2 je uzoračka disperzija za X^A dok za kontrolisani eksperiment možemo da ga posmatramo kao meru rasipanja. Sličan zaključak se može donijeti i za $cov(x^A, \delta)$, mada ćemo

u većini slučajeva da pretpostavimo da je ova vrijednost jednaka 0. Sada je:

$$\hat{\beta}_1 = \Sigma(x_i - \bar{x})y_i / \sum (x_i - \bar{x})^2 \tag{7.5}$$

i nakon računanja nalazimo da je:

$$E(\hat{\beta}_1) = \beta_1 \frac{(\sigma_x^2 + \sigma_{x\delta})}{(\sigma_x^2 + \sigma_\delta^2 + 2\sigma_{x\delta})} \tag{7.6}$$

Postoje dva specijalna slučaja koje posmatramo:

- Ako ne postoji veza između X^A i δ, $\sigma_{x\delta} = 0$, ovo se pojednostavljuje u:

  $$E(\hat{\beta}_1) = \beta_1 \frac{1}{1 + \sigma_\delta^2/\sigma_x^2} \tag{7.7}$$

  Tako da je $\hat{\beta}_1$ blisko 0 bez obzira na veličinu uzorka. Ako je σ_δ^2 malo u odnosu na σ_x^2, onda se problem može ignorisati. Drugim riječima, ako je disperzija grešaka opažanja za X relativno mala u odnosu na rang X, onda nema potrebe da brinemo.
  Rang matrice X dimenzija $m \times n$ je broj r jednak redu najveće kvadratne regularne ($\det X \neq 0$) podmatrice matrice X i važi: $r \leq \min(m, n)$.
  Za više promenljivih, uobičajeni efekat merenja grešaka je da $\hat{\beta}$ teži ka 0.

- U kontrolisanim eksperimentima, treba da odredimo dva načina na koja može da se pojavi greška u X. U prvom slučaju, merimo x tako da iako je prava vrijednost x^A, posmatramo x^0. Ako bismo ponovili merenje, imali bismo isto x^A, ali različito x^0. U drugom slučaju, popravimo x^0 - na primer ako se istraživanj odnosi na traženje prave koncentracije supstanci u nekom hemijskom sastavu, smislimo hemijsko rešenje sa određenom koncentracijom x^0. Prava koncentracija bi bila x^A. Ako bismo ovo ponovili, dobili bismo isto x^0, ali bi x^A bilo različito. U ovom slučaju imamo:

  $$\sigma_{x\delta} = cov(X^0 - \delta, \delta) = -\sigma_\delta^2 \tag{7.8}$$

  i tada bi važilo da je $E(\hat{\beta}_1) = \beta_1$. Tako da bi naša procena za β bila neograničena. Ovo izgleda paradoksalno dok ne primetimo da drugi slučaj zamenjuje uloge x^A i x^0 i ako posmatramo

pravo X, dobijamo neograničenu ocenu za β.
Ako se model koristi za prognoziranje, možemo koristiti iste argumente kao u drugom slučaju. U ponovljenim eksperimentima, vrijednost x u kojoj se vrši pognoza će biti fiksirana, iako može predstavljati različite prave vrijednosti za X.

U slučaju kada greška u X ne može biti ignorisana, razmatramo alternativne načine za procenu koeficijenata, umesto metode najmanjih kvadrata. Jednačina linearne regresije može da se zapiše kao:

$$\frac{y-\bar{y}}{SD_y} = r\frac{(x-\bar{x})}{SD_x} \tag{7.9}$$

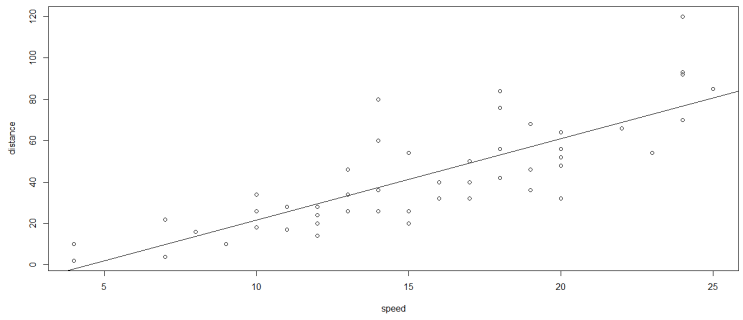
tako da je $\hat{\beta}_1 = rSD_y/SD_x$. Ako zamenimo uloge x i y ovde, ne dobijamo istu jednačinu regresije. Pošto u našem problemu imamo greške i u promenljivim X i Y, možemo da zaključimo da jednačina treba da bude ista u oba slučaja.
Jedan način da se ovo postigne jeste da postavimo da je $\hat{\beta}_1 = SD_y/SD_x$. Ovo je poznato kao veza funkcionalne geometrijske sredine.
Drugi način jeste da koristimo SIMEX metod Cook-a i Stefanskog (1994) koji je ilustrovan u primeru.

Primer 12 *Posmatramo podatke o brzini vozila i dužini zaustavljanja 1920. godine.*

```
data(cars)
g<-lm(dist~speed,cars)
plot(dist~speed,cars,ylab="distance")
abline(g)
```

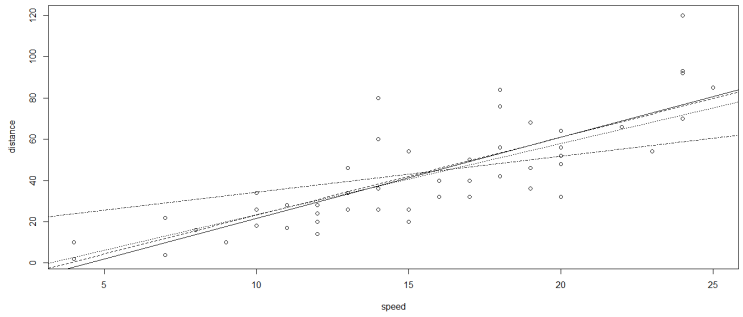
Posmatramo šta se dešava kad dodamo grešku pri merenju promenljive X (speed):

```
> g1<-lm(dist~I(speed+rnorm(50)),cars)
> coef(g1)
         (Intercept) I(speed + rnorm(50))
          -14.566616             3.773571
> abline(g1,lty=2)
```



Slika 12: Dužina zaustavljanja u odnosu na brzinu kretanja auta

```
> g2<-lm(dist~I(speed+2*rnorm(50)),cars)
> coef(g2)
             (Intercept) I(speed + 2 * rnorm(50))
              -10.904692                 3.439259
> abline(g2,lty=3)
> g5<-lm(dist~I(speed+5*rnorm(50)),cars)
> coef(g5)
             (Intercept) I(speed + 5 * rnorm(50))
               17.051929                 1.733042
> abline(g5,lty=4)
```



Slika 13: Modeli sa greškom u promenljivoj speed

Vidimo da se koeficijent pravca smanjuje sa povećavanjem broja tačaka.
Pretpostavimo da znamo da se promenljiva Speed u originalnim po-

dacima meri sa poznatom disperzijom greške - 0.5. Imajući u vidu simulirane modele sa greškom, možemo da procenimo krivu modela kada su merenja bez greške. Odnosno, konstruisanjem modela gde promenljive imaju grešku dobijamo koeficijente kojima teže koeficijenti iz tih modela i tako konstruišemo tačan model za podatke bez greške. Ovo je ideja SIMEX metode.
U ovom primeru simuliramo efekat dodavanja greške sa normalnom raspodelom sa disperzijom od 0.1 do 0.5, ponavljajući eksperiment 1000 puta za svaki od slučajeva.

```
vv<-rep(1:5/10,each = 1000)
slopes<-numeric(1000)
for(i in 1:5000)slopes[i]<-lm(dist~I(speed+sqrt(vv[i])
    *rnorm(50)), cars)$coef[2]
```

Sada predstavljamo na grafiku srednje vrednosti slopes za svaku od disperzija. Pretpostavimo da podaci imaju disperziju 0.5 tako da se dodatna disperzija dodaje na to:

```
betas<-c(coef(g)[2],colMeans(matrix(slopes,nrow=1000)))
disp<-c(0,1:5/10)+0.5
plot(disp,betas,xlim=c(0,1),ylim=c(3.86,4))
```

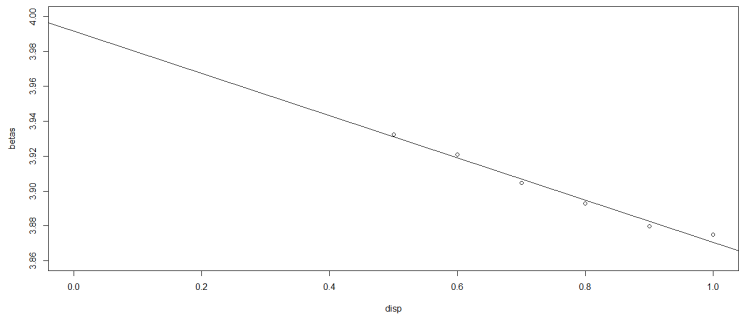
Pravimo linearni model za nove podatke i predstavljamo ga na grafiku:

```
gv<-lm(betas~disp)
> coef(gv)
(Intercept)        disp
  3.9915481  -0.1208646
> abline(gv)
```

Predviđena vrijednost za $\hat{\beta}$ pri disperziji 0, bez greške u merenju, je jednaka 4.0. Na ovaj način smo definisali model za date promenljive bez grešaka u merenju. [5] [18]

Postoje i drugi modeli za ekstrapolaciju koji se mogu koristiti u ovakvim slučajevima.

[18]"Linearni modeli u R-u", J. Faraway, Poglavlje 5, Str 80



Slika 14: Novi model

7.2 Kolinearnost promenljivih X_i

Kolinearnost je veza između vektora $X_1, X_2, ..., X_k$ u kojoj jedan ili više njih linearna kombinacija ostalih. U takvim slučajevima u modelu, kada je jedna promenljiva komnibacija ostalih , matrica $X^T X$ je singularna. Ne postoji jedinstvena procena metodom najmanjih kvadrata za β. Ovo izaziva ozbiljne probleme u proceni β i povezanih vrijednosti, kao i u interpretaciji modela.
Kolinearnost se može otkriti na neki od sledećih načina:

1. Proučavanje matrice korelacije promenljivih X_i može otkriti veliku korelaciju među parovima promenljivih.
2. Regresija X_i po svim ostalim promenljivama daje R_i^2. Ovo se računa za svaku od promenljivih. U slučaju da je R_i^2 blizu 1 ukazuje na problem. Linearna kombinacija među promenljivim se može otkriti proučavanjem regresionih koeficijenata.
3. Proučavanje sopstvenih vrijednosti matrice $X^T X$, gde je λ_1 najveća sopstvena vrijednost sa ostalim u opadajućem redu. Relativno male sopstvene vrijednosti ukazuju na problem. Uslovni broj se definiše kao:

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}} \tag{7.10}$$

gde se $\kappa \geq 30$ smatra velikim. κ se naziva uslovni broj. Ostali uslovni brojevi, $\sqrt{\lambda_1/\lambda_p}$ su isto bitni jer oni ukazuju na to da li je više od jedne linearne kombinacije među promenljivim problem u modelu. Alternativan način je standardizovanje

promenljivih.

Kolinearnost dovodi do toga da je neke parametre teže procijeniti. Definišemo:

$$S_{x_j x_j} = \sum_i (x_{ij} - \bar{x_j})^2, \tag{7.11}$$

$$D(\hat{\beta}_j) = \sigma^2 (\frac{1}{1 - R_j^2}) \frac{1}{S_{x_j x_j}}. \tag{7.12}$$

Vidimo da ako x_j ne varira puno, onda će disperzija $\hat{\beta}_j$ biti velika. Iz ove jednačine takođe vidimo i šta možemo da iskoristimo da smanjimo disperziju koeficijenata ako možemo da izaberemo X. Ortogonalnost znači da je $R_j^2 = 0$ što smanjuje disperziju. Takođe možemo da uvećamo $S_{x_j x_j}$ tako što ćemo da raširimo vrijednosti za X što je više moguće. Maksimum se postiže tako što stavimo pola tačaka u najmanju tačku a pola u najveću. Nažalost ovaj dizajn pretpostavlja linearnost efekta i ne omogućava ni na koji način da se proveri zakrivljenost. U praksi, većina bi stavila neke tačke u sredinu intervala da bi se proverilo koliko model odgovara podacima.
Ako je R_j^2 blizu 1, onda će faktor inflacije disperzije koji se definiše formulom: $\frac{1}{1-R_j^2}$ biti veliki.
Kolinearnost dovodi do neprecizne procene za β. Znaci koeficijenata mogu biti suprotni od onoga što nam intuicija govori o uticaju promenljivih u modelu. Ovo utiče i na standardne greške tako da t-testovi ne prepoznaju značajne faktore. Model postaje osetljiv na greške pri merenju, tako da male promene u y mogu dovesti do velikih promena u $\hat{\beta}$.[4] [19]

Primer 13 *Istraživači u HumoSim laboratoriji na Univerzitetu u Mičigenu su sakupili podatke o 38 vozača. Merili su starost u godinama, težinu u funtama, visinu sa cipelama i bez cipela u cm, visinu kad sede, dužinu ruku, dužinu butina, dužinu potkolenica i rastojanje između sredine kuka i fiksirane tačke u autu u mm.*
Pravimo model od tih podataka, uzimajući u obzir sve promenljive:

```
> data(seatpos)
```

[19]"Linearni statistički modeli", J. Stapleton, Poglavlje 4.7, Str 199

```
> m<-lm(hipcenter~.,seatpos)
> summary(m)

Call:
lm(formula = hipcenter ~ ., data = seatpos)

Residuals:
    Min      1Q  Median      3Q     Max
-73.827 -22.833  -3.678  25.017  62.337

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 436.43213  166.57162   2.620   0.0138 *
Age           0.77572    0.57033   1.360   0.1843
Weight        0.02631    0.33097   0.080   0.9372
HtShoes      -2.69241    9.75304  -0.276   0.7845
Ht            0.60134   10.12987   0.059   0.9531
Seated        0.53375    3.76189   0.142   0.8882
Arm          -1.32807    3.90020  -0.341   0.7359
Thigh        -1.14312    2.66002  -0.430   0.6706
Leg          -6.43905    4.71386  -1.366   0.1824
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 37.72 on 29 degrees of freedom
Multiple R-squared:  0.6866, Adjusted R-squared:  0.6001
F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

Ovaj model već pokazuje neke znake kolinearnosti. p vrijednost za F statistiku je jako mala, a R^2 je značajno veliko, ali nijedna od individualnih promenljivih nije značajna. Proveravamo korelaciju među parovima promenljivih:

```
> round(cor(seatpos),3)
             Age Weight HtShoes     Ht Seated    Arm
Age        1.000  0.081  -0.079 -0.090 -0.170  0.360
Weight     0.081  1.000   0.828  0.829  0.776  0.698
HtShoes   -0.079  0.828   1.000  0.998  0.930  0.752
```

```
Ht        -0.090  0.829   0.998  1.000  0.928  0.752
Seated    -0.170  0.776   0.930  0.928  1.000  0.625
Arm        0.360  0.698   0.752  0.752  0.625  1.000
Thigh      0.091  0.573   0.725  0.735  0.607  0.671
Leg       -0.042  0.784   0.908  0.910  0.812  0.754
hipcenter  0.205 -0.640  -0.797 -0.799 -0.731 -0.585

           Thigh    Leg hipcenter
           0.091 -0.042     0.205
           0.573  0.784    -0.640
           0.725  0.908    -0.797
           0.735  0.910    -0.799
           0.607  0.812    -0.731
           0.671  0.754    -0.585
           1.000  0.650    -0.591
           0.650  1.000    -0.787
          -0.591 -0.787     1.000
```

Vidimo da postoji više velikih korelacija među promenljivim koje koristimo za predviđanje i između ovih promenljivih i rezultujuće Hipcenter promenljive. Zato proveravamo sopstvene vrijednosti:

```
> x<- model.matrix(m)[,-1]
> e<-eigen(t(x)%*%x)
> e$val
[1] 3.653671e+06 2.147948e+04 9.043225e+03 2.989526e+02
    1.483948e+02
[6] 8.117397e+01 5.336194e+01 7.298209e+00
> sqrt(e$val[1]/e$val)
[1]   1.00000  13.04226  20.10032 110.55123 156.91171
      212.15650 261.66698
[8] 707.54911
```

Vidimo da su sopstvene vrijednosti u velikom intervalu i da je nekoliko uslovnih brojeva dosta veliko. To znači da je probleme u modelu izazvalo više od jedne linearne kombinacije među promenljivim.
Sada proveravamo faktor promene disperzije. Za prvu promenljivu imamo:

```
 > summary(lm(x[,1]~x[,-1]))$r.squared
[1] 0.4994823
> 1/(1-0.49948)
[1] 1.997922
```

Računamo automatski sve ostale promene disperzije koristeći funkciju iz paketa faraway u R-u:

```
> vif(x)
       Age      Weight    HtShoes          Ht      Seated
  1.997931    3.647030 307.429378 333.137832    8.951054
       Arm       Thigh     Leg
    4.496368    2.762886 6.694291
```

Postoji velika promena disperzije, dakle imamo problem korelisanosti promenljivih.
Merenje hipcentra (razdaljine izmedju kukova i kola) je teško da se izvede tačno i možemo očekivati neke varijacije u ovim vrijednostima.[5]
[20]

Jedan od načina da se riješi kolinearnost jeste da iz modela izbacimo promenljivu koja pravi probleme. Ako imamo dosta promenljivih koje koristimo da bismo objasnili neku pojavu, to nije takav problem. Ako nekoliko promenljivih, koje su jako korelisane, su sve povezane sa rezultujućom promenljivom u modelu, moramo da pazimo da ne zaključimo da promenljive koje sklonimo iz modela nemaju nikakve veze sa rezultatom.

Nastavak primera: *Posmatramo matricu korelacije promenljivih, samo za promenljive koje se odnose na dužine.*

```
> round(cor(x[,3:8]),2)
        HtShoes   Ht Seated  Arm Thigh  Leg
HtShoes    1.00 1.00   0.93 0.75  0.72 0.91
```

[20]"Linearni modeli u R-u", J. Faraway, Poglavlje 5.2, Str 85

```
Ht          1.00 1.00   0.93 0.75  0.73 0.91
Seated      0.93 0.93   1.00 0.63  0.61 0.81
Arm         0.75 0.75   0.63 1.00  0.67 0.75
Thigh       0.72 0.73   0.61 0.67  1.00 0.65
Leg         0.91 0.91   0.81 0.75  0.65 1.00
```

Ovih 6 promenljivih su jako korelisane jedna sa drugom - bilo koja bi bila dobra za prikazivanje ostalih. Izaberemo visinu kao najjednostavniju za merenje. Ne tvrdimo da ostale promenljive ne utiču na rezultujuću promenljivu, ali nam ne trebaju da bismo predvideli vrijednosti.
Pravimo novi model bez ostalih promenljivih:

```
> m2<-lm(hipcenter~Age+Weight+Ht,seatpos)
> summary(m2)

Call:
lm(formula = hipcenter ~ Age + Weight + Ht,
   data = seatpos)

Residuals:
    Min      1Q  Median      3Q     Max
-91.526 -23.005   2.164  24.950  53.982

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 528.297729 135.312947   3.904 0.000426 ***
Age           0.519504   0.408039   1.273 0.211593
Weight        0.004271   0.311720   0.014 0.989149
Ht           -4.211905   0.999056  -4.216 0.000174 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 36.49 on 34 degrees of freedom
Multiple R-squared:  0.6562, Adjusted R-squared:  0.6258
F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08
```

Kada uporedimo ovaj model sa prvobitnim, vidimo da je sličan u smislu R^2, dakle dobro predstavlja date podatke, ali da je mnogo manje promenljivih korišćeno u modelu.

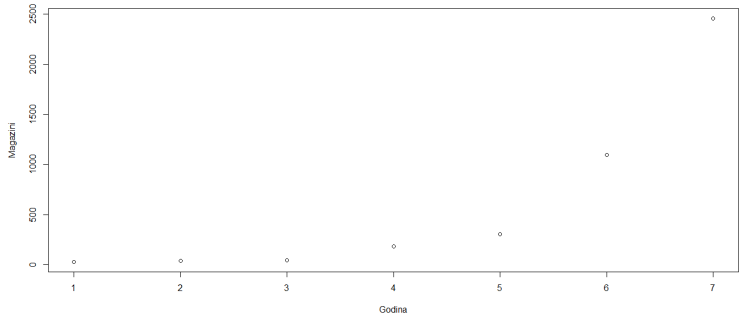
U slučaju da moramo da zadžimo sve promenljive u modelu, koristimo alternativne metode za rešavanje problema korelacije.

8 Linearne transformacije

Posmatramo problem nalaženja regresione funkcije $g(x) = E(Y|x)$ za promenljivu x. Postoji nekoliko načina za procenu $g(x)$ u slučaju kada je $g(x)$ linearna kombinacija $h(x) = \sum_{j=1}^{k} \beta_j f_j(x)$, gde su f_j poznate funkcije od x; odnosno $g(x)$ se može predstaviti kao linearna funkcija nepoznatih parametara β_j. Ako je g glatka funkcija i x nema preveliki domen, tada $g(x)$ može biti aproskimirana funkcijom u obliku $h(x)$ na traženom intervalu za x.
Problem se javlja kada ovo nije slučaj, kao što je prikazano u sledećem primeru.

Primer 14 *Podaci predstavljaju broj akademskih magazina objavljenih na Internetu tokom perioda 1991-1997.*

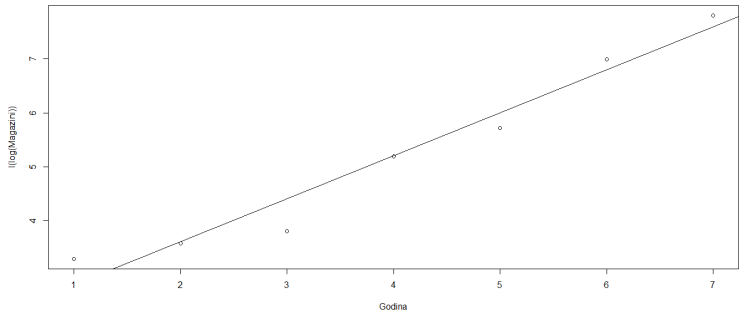
```
Godina<-seq(1,7)
Magazini<-c(27,36,45,181,306,1093,2459)
Internet<-data.frame(Godina,Magazini)
plot(Godina,Magazini)
```



Slika 15: Broj magazina objavljenih tokom datog perioda

Grafik pokazuje da veza izmedju broja magazina i godine nije linearna. Zbog očiglednog eksponencijalnog rasta broja magazina, možemo da pretpostavimo da bi pomoglo da uzmemo logaritam broja magazina prije nego što uklopimo vrednosti u model.
Ovo možemo da uradimo direktno u pozivu regresione funkcije lm() u R-u.

```
attach(Internet)
rezultat<-lm(I(log(Magazini))~Godina)
plot(Godina,I(log(Magazini)))
abline(rezultat)
```



Slika 16: Model nakon transformacije

Kada se javi ovakav problem da regresiona funkcija nije linearna može se izvršiti tranformacija kao u primeru. Transformacije možemo vršiti na zavisnim promenljivama ili na nezavisnoj promenljivoj.

Ako posmatramo aproksimaciju:

$$Y \approx h(x, \gamma_0, \gamma_1) \approx \gamma_0 e^{\gamma_1 x \varepsilon} \tag{8.1}$$

ili drugačije zapisano:

$$logY \approx log\gamma_0 + \gamma_1 x_1 + \varepsilon. \tag{8.2}$$

Koristeći ovaj model dobijamo linearnu funkciju za β_0, β_1, a samim tim i ocene za ove parametre koristeći jednostavni linearni model.

U praksi, ne možemo tako lako da odredimo kako se greška pojavljuje u modelu (da li se dodaje ili množi). Standardni postupak je da probamo nekoliko transformacija pa da na osnovu reziduala odredimo da li zadovoljavaju uslove neophodne za linearnu regresiju.
Iako transformišemo nezavisnu promenljivu, zavisne promenljive se obično predstavljaju u originalnom obliku. Regresioni koeficijenti će morati da se predstave u odnosu na transformisani model.
Kada se koristi logaritamska transformacija rezultata, koeficijenti se predstavljaju na sledeći način:

$$log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p \hat{y} = e^{\hat{\beta}_0} e^{\hat{\beta}_1 x_1} ... e^{\hat{\beta}_p x_p} \tag{8.3}$$

Na ovaj način se regresioni koeficijenti predstavljaju na multiplikativni umesto na aditivni način.

U prethodnom primeru je bilo očigledno da može da se iskoristi logaritamska transformacija. Ali to neće uvijek biti tako jednostavno.

Box - Cox metoda je popularni način za određivanje transformacije rezultujuće promenljive. Dizajnirana je isključivo za pozitivne rezultate i pomaže da se nađe transformacija koja najbolje odgovara podacima.
Metoda transformiše rezultujuću promenljivu y u $g_\lambda(y)$, koja ima sledeći oblik:

$$g_\lambda(y) = \begin{array}{ll} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ logy & \lambda = 0 \end{array} \tag{8.4}$$

Za fiksirano $y > 0$, $g_\lambda(y)$ je neprekidno po λ. Biramo λ sa najvećom verovatnoćom. Funkcija verodostojnosti, pod pretpostavkom da su greške normalno raspodeljene je:

$$L(\lambda) = -\frac{n}{2} log(\frac{RSS_\lambda}{n}) + (\lambda - 1) \sum log(y_i) \tag{8.5}$$

gde je RSS_λ rezidualna suma kvadrata kada je $g_\lambda(y)$ rezultat.[12] [21]

Transformacija rezultujuće promenljive može da učini model još težim za interpretaciju, pa ne želimo to da uradimo osim ako je

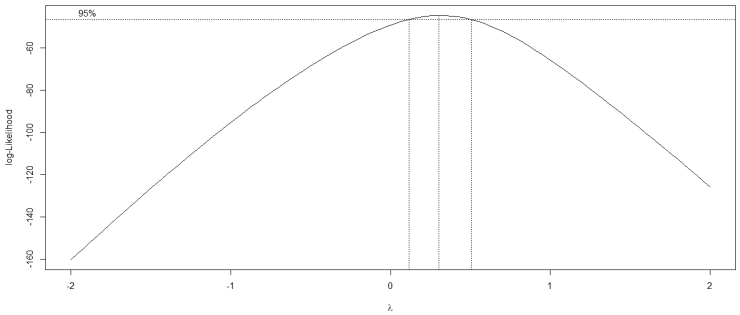
[21] "Normalizacija podataka koristeći Box-Cox transformaciju", A. Burthman

neophodno. Jedan od načina da se to proveri je da se napravi interval poverenja za λ.
$100(1-\lambda)\%$ interval poverenja za λ je:

$$\lambda : L(\lambda) > L(\hat{\lambda}) - \frac{1}{2}\chi_1^{2(1-\alpha)}. \tag{8.6}$$

Primer 15 *Posmatramo podatke o vrstama ptica na Galapagosu. Primenjujemo Box-Cox metodu na model koji predstavlja ove podatke:*

```
data(gala)
m<-lm(Species~Area+Elevation+Nearest+Scruz+Adjacent,
gala)
boxcox(m,plotit=T)
boxcox(m,lambda=seq(0.0,1.0,by=0.05),plotit=T)
```

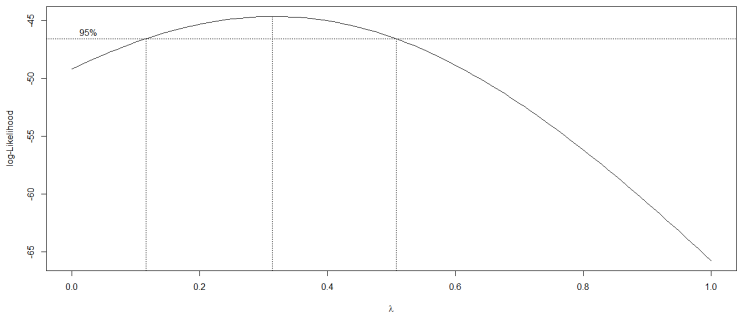


Slika 17: Box-Cox grafik za sve vrijednosti

Sa grafika možemo da vidimo je verovatno najbolja transformacija promenljive Species, kubni koren. Kvadratni koren je takođe moguća transformacija, pošto to λ pripada intervalu poverenja.

Osobine Box-Cox metode na koje treba obratiti posebnu pažnju:

1. Autlajeri utiču na Box-Cox metodu.
2. Ako su neki $y_i < 0$, možemo da dodamo konstantu na sve y. Ovo je dobro rešenje u slučaju da je konstanta mala.
3. Ako je $\max_i y_i / \min_i y_i$ malo, tada Box-Cox neće imati pravog efekta zato što početna funkcija može biti dobro aproksimirana linearnom transformacijom na kratkim intervalima.



Slika 18: Box-Cox grafik za određene vrijednosti λ

4. Postoji sumnja da li ocena λ treba da se uračuna u stepene slobode. Ovo je teško pitanje pošto λ nije linearni parametar i njegova procena nije deo procene metodom najmanjih kvadrata.

Box-Cox metoda nije jedini način za transformaciju rezultujuće promenljive. Za promenljivu Y koja je predstavljena razlomcima, koristi se logaritamska transformacija oblika $\log(y(1-y))$, dok se za promenljive koje predstavljaju korelacije koristi Fišerova z transformacija: $y = 0.5\log(\frac{(1+y)}{(1-y)})$.

Box-Cox metod se može primeniti za svaku od pomenljivih X. Ipak, to može da potraje neko vrijeme. Zato je možda jednostavnije koristiti grafičke metode za određivanje prave transformacije. Ove metode su dizajnirane tako da zamene X u modelu sa $f(X)$ za neko izabrano f.[5] [22]

8.1 Goodness of fit

Da bismo uporedili dva ili više pokušaja uklapanja podataka u model, treba nam mera saglasnosti modela, tj. mera toga koliko model odgovara stvarnim podacima (goodness of the fit). Ako je aproksimacija y_i data modelom $\widehat{Y_i}$ onda se ovakva mera računa na sledeći način:

[22]"Linearni modeli u R-u", J. Faraway, Poglavlje 7.1, Str 110

$$R^2(\hat{Y}, y) \equiv 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}. \tag{8.7}$$

Ovo je višestruki koeficijent korelacije samo u slučaju da se $\hat{y}$ dobijaju iz linearnog modela. R^2 moze biti manje od 0. [4] [23]

Test koji se koristi za merenje saglasnosti modela sa podacima (goodness of fit test) ima početnu pretpostavku (H_0) da model odgovara podacima.

Pretpostavimo da izvršena transformacija promenljive y: $z = g(y)$. Neka je $\hat{z}$ pretpostavljena vrijednost za z dobijena iz linearnog modela za z preko x. Neka je $\hat{Y}$ pretpostavljena vrijednost za y iz linearne regresije y po x, i neka je $\hat{Y}_z = g^{-1}(\hat{z})$. Scott i Wild su 1991. godine dali primer koji se sastoji iz 6 parova (x, y):

$$(x, y) : (0, 0.1), (3, 0.4), (8, 2), (13, 10), (16, 15), (20, 16).$$

Za date vrijednosti se dobija da je $R^2(\hat{Y}, y) = 0.88$, a $R^2(\hat{Y}_z, y) = -0.316$. U smislu najmanjih kvadrata, $\bar{y}$ je bolja ocena za y od $\hat{Y}_z$. Na z- skali (posmatramo R^2 u odnosu na z promenljivu) $R^2(z, \hat{z}) = 0.94$. Scott i Wild su upozorili, pokazujući to i na primerima, da vrijednosti R^2 zasnovane na različitim skalama ne mogu da se porede.

Merenje vrednosti transfomacije treba da se bazira na nekoj skali da bi se koristilo u donošenju odluke. Oni koji koriste ovu metodologiju moraju da izaberu skalu na kojoj će meriti valjanost aproksimacije.
U sledećem primeru predstavljeni su različiti modeli , sa i bez transformacija i upoređena njihova saglasnost sa realnim podacima.

Primer 16 *Date su vrijednosti* (x_i, y_i)*, i njihove tranformacije* $w_i = log(x_i)$*,* $z_i = log(y_i)$*,* $u_i = \frac{1}{x_i}$*. Odrediti najbolji model za date podatke.*

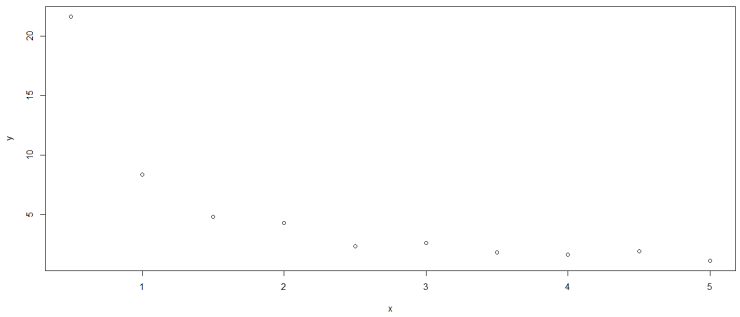
```
> x<-c(0.5,1,1.5,2,2.5,3,3.5,4,4.5,5)
```

[23] "Linearni statistički modeli", J. Stapleton, Poglavlje 4.1, Str 166

```
> y<-c(21.64,8.343,4.833,4.3,2.343,2.623,1.818,1.628,
1.909,1.120)
> w<-log(x)
> w
[1]-0.6931472 0.0000000 0.4054651 0.6931472 0.9162907
[7] 1.0986123 1.2527630 1.3862944 1.5040774 1.6094379
> z<-log(y)
> z
 [1] 3.0745435 2.1214229 1.5754674 1.4586150 0.8514322
 [8] 0.9643187 0.5977370 0.4873523 0.6465795 0.1133287


> summary(m1)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-4.055 -2.811 -1.302  1.765  9.687

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.4860     2.9927   4.506  0.00199 **
x            -3.0656     0.9646  -3.178  0.01304 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.381 on 8 degrees of freedom
Multiple R-squared:  0.558, Adjusted R-squared:  0.5027
F-statistic:  10.1 on 1 and 8 DF,  p-value: 0.01304

> plot(x,y)
```

Prvi model (m_1) predstavlja podatke x *i* y *bez transformacije.* R^2 *je u ovom modelu jednako 0.5027, a i koeficijenti imaju veliku grešku. Konstruišemo drugi model ali ovoga puta transformišemo podatke.*

```
> m2<-lm(y~w)
> summary (m2)
```

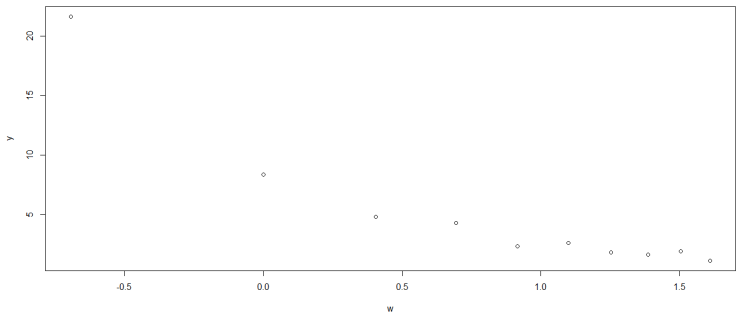


Slika 19: Model koji predstavlja podatke x i y bez transformacije

```
Call:
lm(formula = y ~ w)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3858 -1.8916 -0.0825  1.8318  4.9832

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.333      1.359   8.336 3.24e-05 ***
w             -7.681      1.267  -6.063 0.000302 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.786 on 8 degrees of freedom
Multiple R-squared:  0.8213, Adjusted R-squared:  0.7989
F-statistic: 36.76 on 1 and 8 DF,  p-value: 0.0003016
```

Drugi model (m_2) predstavlja podatke y *koristeći transformisane podatke za* x *(logaritamtska transformacija). Vidimo da je u ovom modelu R^2 veće i sada iznosi 0.7989. Greška iznosi 2.786, a p vrijednost 0.0003016.*

```
> m3<-lm(z~w)
> summary(m3)
```

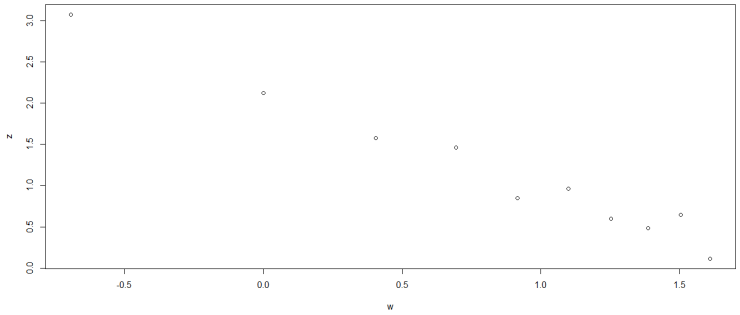


Slika 20: Model koji predstavlja podatke y i log(x) bez transformacije

```
Call:
lm(formula = z ~ w)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21931 -0.09712 -0.03309  0.10362  0.27845

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.16604    0.07607   28.47 2.50e-09 ***
w           -1.19536    0.07089  -16.86 1.55e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1559 on 8 degrees of freedom
Multiple R-squared:  0.9726, Adjusted R-squared:  0.9692
F-statistic: 284.3 on 1 and 8 DF,  p-value: 1.551e-07
```

Treći model (m_3) predstavlja transformisane podatke za y ($\log(y)$) i transformisane podatke za x ($\log(x)$). Vidimo da je R^2 u ovom modelu 0.9692, tj. možemo da zaključimo da model dobro predstavlja podatke. Greška je manja nego u prethodnim modelima i sada iznosi 0.1559.

A i ako uporedimo grafike, vidimo da treći model najbolje odgo-



Slika 21: Model koji predstavlja transformisane podatke i za x i za y

vara datim podacima. Zato zaključujemo da je ovo najbolji model za predstavljanje datih podataka.

Literatura u ovom poglavlju:
U ovom poglavlju su korišćene sledeće jedinice iz literature:

- "Linearni modeli u R-u", J. Faraway [5],
- "Metoda uopštenih najmanjih kvadrata", I. Ruczinski [11],
- "Linearni statistički modeli", J. Stapleton [4],
- "Normalizacija podataka koristeći Box-Cox transformaciju", A. Burthman [12].

Poglavlje III
Primena opisanih rešenja

9 Provera pretpostavki o greškama ε_i

9.1 Problem nekonstantne disperzije grešaka

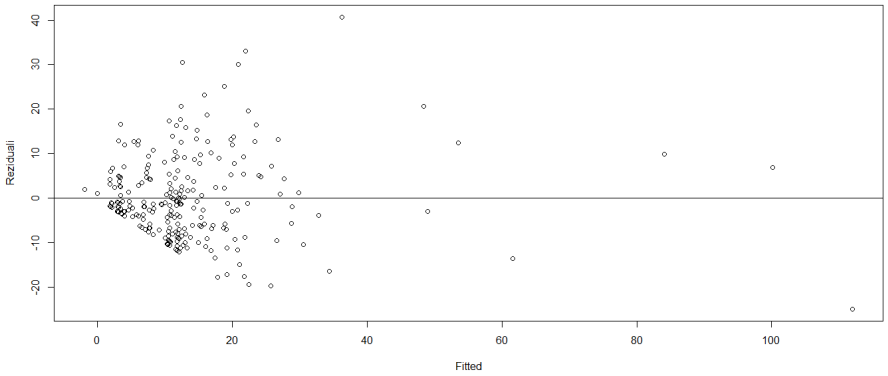
Posmatramo podatke *Ornstein* iz paketa *car*. Podaci predstavljaju informacije o 248 najvećih firmi u Kanadi sredinom 1970-tih. Za date podatke konstruišemo model:

```
> library(car)
> data(Ornstein)
> m<-lm(interlocks~assets+sector+nation, data=Ornstein)
> summary(m)

Call:
lm(formula = interlocks ~ assets + sector + nation,
data = Ornstein)

Residuals:
    Min      1Q  Median      3Q     Max
-25.001  -6.602  -1.629   4.780  40.728

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.027e+01  1.561e+00   6.575 3.14e-10 ***
assets       8.096e-04  6.119e-05  13.231  < 2e-16 ***
sectorBNK   -1.781e+01  5.906e+00  -3.016  0.00284 **
sectorCON   -4.709e+00  4.728e+00  -0.996  0.32034
sectorFIN    5.153e+00  2.646e+00   1.948  0.05266 .
sectorHLD    8.777e-01  4.004e+00   0.219  0.82669
sectorMAN    1.149e+00  2.065e+00   0.556  0.57849
sectorMER    1.491e+00  2.636e+00   0.566  0.57206
sectorMIN    4.880e+00  2.067e+00   2.361  0.01905 *
sectorTRN    6.171e+00  2.760e+00   2.236  0.02629 *
sectorWOD    8.228e+00  2.679e+00   3.072  0.00238 **
nationOTH   -1.241e+00  2.695e+00  -0.461  0.64555
nationUK    -5.775e+00  2.674e+00  -2.159  0.03184 *
```

```
nationUS    -8.618e+00  1.496e+00  -5.760 2.64e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.827 on 234 degrees of freedom
Multiple R-squared:  0.6463, Adjusted R-squared:  0.6267
F-statistic: 32.89 on 13 and 234 DF,  p-value: < 2.2e-16
```

Posmatramo reziduale u ovom modelu i proveravamo na grafiku da li imaju neku specijalnu formu ili je disperzija konstantna.

```
> plot(fitted(m),residuals(m),xlab="Fitted",
ylab="Reziduali")
> abline(h=0)
```



Slika 22: Reziduali iz datog modela m

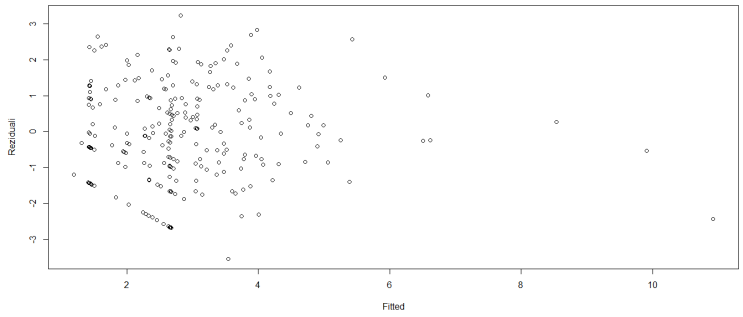
Na grafiku vidimo da se reziduali grupišu, pa stoga zaključujemo da nemaju konstantnu disperziju.
Probaćemo da rešimo taj problem transformisanjem promenljivih u modelu. U ovom slučaju ne može da se koristi obična box-plot metoda pošto nisu sve vrijednosti u bazi pozitivne.
Zato koristimo spread.level funkciju iz paketa cars, koja vraća najbolju transformaciju za date podatke.

```
> spread.level.plot(m)
```

```
Suggested power transformation:  0.4788627
Warning messages:
1: 'spread.level.plot' is deprecated.
Use 'spreadLevelPlot' instead.
See help("Deprecated") and help("car-deprecated").
2: In spreadLevelPlot.lm(...) : 2 negative fitted
values removed
```

Na osnovu ove procene, transformacija koja bi rešila problem nekonstantne disperzije reziduala je da se transformiše za $\lambda = 0.4788627$. Konstruišemo model sa ovom transformacijom i proveravamo da li reziduali u tom modelu imaju konstantnu disperziju:

```
>m3<-lm(interlocks^0.4788~assets+sector+nation,
data=Ornstein)
>plot(fitted(m3),residuals(m3),xlab="Fitted",
ylab="Reziduali")
```



Slika 23: Reziduali iz transformisanog modela m2

Na grafiku vidimo da je problem sa ne-konstantnom disperzijom ispravljen transformisanjem podataka.

Drugi način za rešavanje ovog problema bi bilo korišcenje metode odmerenih kvadrata (weighted least squares) koja je opisana ranije u radu:

```
> m.wls<-lm(interlocks~assets+nation+sector,
weights=1/assets, data=Ornstein)
```

U ovom primeru za mere (weights) koristimo recipročne vrijednosti jedne od nezavisnih promenljivih (assets).

9.2 Provera normalne raspodele grešaka

Imamo date podatke za x i y i model koji je konstruisan za njih:

```
  > x = c(21,34,6,47,10,49,23,32,12,16,29,49,28,8,57,
  9,31,10,21,26,31,52,21,8,18,5,18,26,27,26,32,2,59,
  58,19,14,16,9,23,28,34,70,69,54,39,9,21,54,26)
> y = c(47,76,33,78,62,78,33,64,83,67,61,85,46,53,
  55,71,59,41,82,56,39,89,31,43,29,55,
  81,82,82,85,59,74,80,88,29,58,71,60,86,91,72,89,
  80,84,54,71,75,84,79)
> model=lm(y~x)
```

Hoćemo da proverimo normalnu raspodelu reziduala u ovom modelu. Konstruišemo Q-Q grafike da bismo proverili da li je raspodela grešaka normalna:

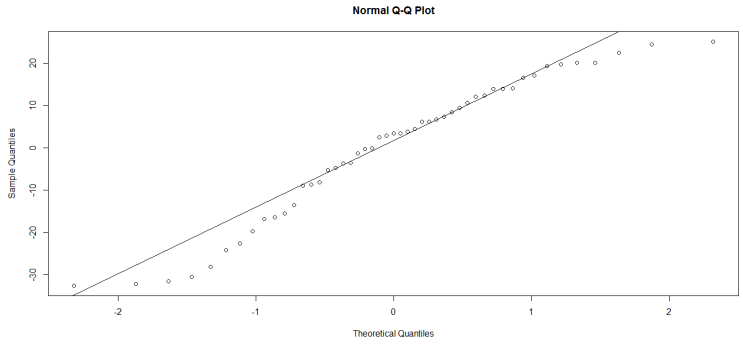
```
> qqnorm(residuals(model))
> qqline(residuals(model))
```

Sa grafika vidimo da reziduali nemaju potpuno normalnu raspodelu (sve tačke na grafiku se ne nalaze na pravoj). Ali pošto ne odstupaju puno od normalne prave, možemo to da zanemarimo (raspodela sa kratkim repom).

10 Provera pretpostavki o promenljivim X_i

10.1 Problem korelisanosti promenljivih

Posmatramo podatke iz sledeće baze (dobijene iz dva paketa u R-u : *car* i *latice* i konstruisane tako da se dobiju korelisani podaci).



Slika 24: QQ grafik za dati model za podatke x i y

Ovde je prikazano samo prvih 10 elemenata iz promenljivih:

```
> rdata
            resp       pred1       pred2       pred3
1   -0.596777525 -0.17505561  0.07416697 -0.11684624
2    1.212922693  0.97651258  1.05554932  1.18139180
3   -0.615503509 -0.98215647 -0.64043086 -1.16973291
4    0.072565198  0.42519539  0.36234438 -1.86437066
5    0.629743594  1.64818502  1.62997132 -1.09955165
6    1.559369772  2.24120863  2.01536690  0.77667011
7    1.579147111  2.14110586  1.59886856 -0.90108276
8    1.389210834  1.68797428  1.21365934  2.03412244
9   -1.477952435 -0.04964882  0.37505072 -0.56296768
10  -0.004776042  0.16038799  0.47016039  0.97589715
. . .
```

Konstruišemo model za date podatke.

```
> m3 = lm(resp ~ pred1 + pred2 + pred3, rdata)
> summary(m3)

Call:
lm(formula = resp ~ pred1 + pred2 + pred3,
data = rdata)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1.93887 -0.41198 -0.03508  0.40897  1.44148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02015    0.06402  -0.315 0.753631
pred1        0.74344    0.21040   3.533 0.000633 ***
pred2       -0.06396    0.21493  -0.298 0.766668
pred3        0.27115    0.06076   4.463  2.2e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.629 on 96 degrees of freedom
Multiple R-squared:  0.6678,Adjusted R-squared:  0.6574
F-statistic: 64.33 on 3 and 96 DF,  p-value: < 2.2e-16
```

Vidimo da postoji velika greška u modelu, iako po R^2 možemo da zaključimo da model odgovara podacima.
Proveravamo kolinearnost među promenljivim u modelu i promenu disperzije:

```
> cor(rdata)
           resp     pred1     pred2     pred3
resp  1.0000000 0.7738799 0.7424648 0.5275810
pred1 0.7738799 1.0000000 0.9584316 0.3668424
pred2 0.7424648 0.9584316 1.0000000 0.3719551
pred3 0.5275810 0.3668424 0.3719551 1.0000000

> vif(m3)
   pred1    pred2    pred3
12.30247 12.35640  1.16234
```

Vidimo da postoji velika korelacija među promenljivim *pred1* i *pred2*. Zato konstruišemo model bez promenljive *pred1*:

```
> m2 = lm(resp ~ pred2 + pred3, rdata)
> summary(m2)

Call:
lm(formula = resp ~ pred2 + pred3, data = rdata)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.77844 -0.42705 -0.04775  0.41461  1.33927

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03991    0.06744  -0.592    0.555
pred2        0.65893    0.06966   9.459 1.97e-15 ***
pred3        0.27953    0.06420   4.354 3.32e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6652 on 97 degrees of freedom
Multiple R-squared:  0.6246, Adjusted R-squared:  0.6169
F-statistic:  80.7 on 2 and 97 DF,  p-value: < 2.2e-16
```

Vidimo da ovaj model bolje predstavlja podatke, a gore smo videli da je korelacija među promenljivama *pred2* i *pred3* mala, tako da to rešava problem korelisanosti promenljivih u modelu.

11 Transformacija promenljivih u modelu

Posmatramo podatke za X_1, X_2, X_3 i Y i tražimo model koji im najviše odgovara:

```
Y<-c(11.2,13.4,40.7,5.3,24.8,12.7,
20.9,35.7,8.7,9.6,14.5,26.9,15.7,
36.2,18.1,28.9,14.9,25.8,21.7,25.7)
X1<-c(587000,643000,635000,692000,
1248000,643000,1964000,1531000,713000,
749000,7895000,762000,2793000,741000,
625000,854000,716000,921000,595000,
3353000)
X2<-c(16.5,20.5,26.3,16.5,19.2,16.5,
20.2,21.3,17.2,14.3,18.1,23.1,19.1,
24.7,18.6,24.9,17.9,22.4,20.2,16.9)
X3<-c(6.2,6.4,9.3,5.3,7.3,5.9,6.4,
7.6,4.9,6.4,6.0,7.4,5.8,8.6,6.5,8.3,
6.7,8.6,8.4,6.7)
```

Konstruišemo osnovni model u kom učestvuju sve promenljive:

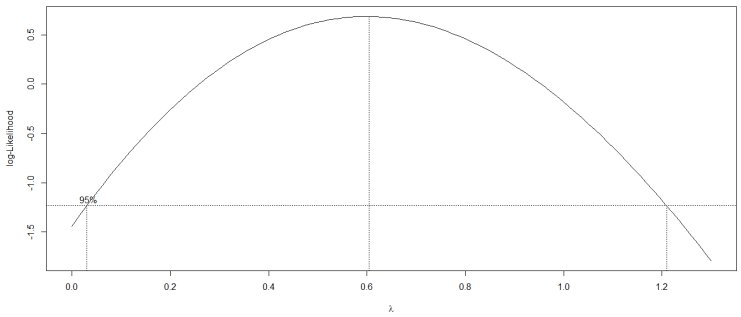
```
> model<-lm(Y~X1+X2+X3)
> summary(model)

Call:
lm(formula = Y ~ X1 + X2 + X3)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7174 -3.3233  0.4031  1.7684 10.0329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.676e+01  7.011e+00  -5.244 8.03e-05 ***
X1           7.629e-07  6.363e-07   1.199  0.24798
X2           1.192e+00  5.617e-01   2.123  0.04974 *
X3           4.720e+00  1.530e+00   3.084  0.00712 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.59 on 16 degrees of freedom
Multiple R-squared:  0.8183, Adjusted R-squared:  0.7843
F-statistic: 24.02 on 3 and 16 DF,  p-value: 3.629e-06
```

Vidimo da ovaj model odgovara podacima (R^2 je veliko), ali poveravamo da li bi neka transformacija predstavila podatke na bolji način.
Koristimo Box-Cox metodu da bi našli najbolju transformaciju.

```
> boxcox(model,plotit=T,lambda=seq(0,1.3,by=0.1))
```

Vidimo da je 95% interval za λ otprilike od 0 do 1.2, a da je $\lambda = 0.6$. Dakle, transformacija bi bila $y^{0.6}$.
Pravimo takav model:

```
> tran.model<-lm(Y^0.6 ~X1+X2+X3)
> summary(tran.model)
```



Slika 25: Nalaženje λ za dati model

```
Call:
lm(formula = Y^0.6 ~ X1 + X2 + X3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.01635 -0.64463  0.05712  0.36234  1.64599

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.424e+00  1.262e+00  -3.505  0.00293 **
X1           1.673e-07  1.146e-07   1.461  0.16348
X2           1.994e-01  1.011e-01   1.972  0.06610 .
X3           8.969e-01  2.755e-01   3.255  0.00497 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8263 on 16 degrees of freedom
Multiple R-squared:  0.8194, Adjusted R-squared:  0.7855
F-statistic:  24.2 on 3 and 16 DF,  p-value: 3.462e-06
```

Vidimo da je ovaj model minimalno povećao procenat objašnjene disperzije, promenio stepen značajnosti, kao i parametre modela.

Zaključak

U radu smo prošli kroz osobine modela, njihove pretpostavke i neke probleme koji se mogu javiti pri konstruisanju modela. Problemi se mogu javiti u osobinama grešaka i tada se rešavaju transformacijom promenljivih ili metodom uopstenih ili odmerenih najmanjih kvadrata.

Najčešći problemi koji se mogu javiti u pretpostavkama o promenljivim jesu ti da se vrijednosti promenljivih mere sa greškom ili da su promenljive međusobno kolinearne. Ovi problemi se rešavaju eliminacijom određene promenljive iz modela ili nekom drugom metodom.

U radu nisu detaljnije opisani načini za upoređivanje modela koristeći disperzionu analizu (ANOVA), već su modeli upoređivani samo na osnovu toga koliko odgovaraju podacima, odnosno po koeficijentu determinacije (R^2).

Postoje i neki drugi problemi koji se mogu javiti pri konstruisanju višestrukih linearnih modela kao posledica malog obima uzorka ili autlajera, ali pošto ovi problemi ne zavise od pretpostavki modela, nisu detaljnije objašnjeni u radu.

Višestruki linearni modeli se koriste u mnogim naučnim oblastima i kao sto smo videli u ovom radu, greške u osnovnim pretpostavkama imaju veliki uticaj na samu korektnost modela i na zaključke koji se mogu izvesti iz njega.
Većina problema u pretpostavkama se može rešiti primenjivanjem metoda opisanih u drugom dijelu ovog rada, a tako će se i model kojim se predstavljaju dati podaci učiniti tačnijim i korisnijim za dalju analizu podataka.

Literatura

[1] Paul Allison: Multiple Regression, New Delhi, 1998

[2] F. Bruckler: Metoda najmanjih kvadrata, 2000

[3] Raj Jain: Other Regression Models, St.Louis, 2008

[4] James H. Stapleton: Linear Statistical Models, Canada, 1995

[5] Julian J. Faraway: Linear models with R, Broken Sound Parkway NW, 2004

[6] Ezequiel Uriel: Multiple linear regression: estimation and properties, Valencia, 2013

[7] Ovidiu Calin: An Introduction to Stochastic Calculus with Applications to Finance, Ann Arbor,Michigan,

[8] Anita Vaš: Modeliranje višestrukom linearnom regresijom, Novi Sad

[9] David Gamarnik: Advanced Stochastic Processes, Massachusetts Ave,Cambridge, 2013

[10] Perla Sousi: Advanced Probability, Cambridge, 2013

[11] Ingo Ruczinski: Generalized Least Squares, Baltimore

[12] Arne Burthman: Making Data Normal Using Box-Cox Power Transformation