# Problem multikolinearnosti u višestrukoj regresiji: detekcija i moguća rješenja

**Nakić, Jelena**

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

JELENA NAKIĆ

# PROBLEM MULTIKOLINEARNOSTI U VIŠESTRUKOJ REGRESIJI: DETEKCIJA I MOGUĆA RJEŠENJA

DIPLOMSKI RAD

Split, prosinac 2023.

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

ODJEL ZA MATEMATIKU

# PROBLEM MULTIKOLINEARNOSTI U VIŠESTRUKOJ REGRESIJI: DETEKCIJA I MOGUĆA RJEŠENJA

DIPLOMSKI RAD

Student(ica):
Jelena Nakić

Neposredna voditeljica:
dr. sc. Ana Perišić
Mentor:
izv. prof. dr. sc. Jurica Perić

Split, prosinac 2023.

TEMELJNA DOKUMENTACIJSKA KARTICA

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU
ODJEL ZA MATEMATIKU

DIPLOMSKI RAD

# PROBLEM MULTIKOLINEARNOSTI U VIŠESTRUKOJ REGRESIJI: DETEKCIJA I MOGUĆA RJEŠENJA

Jelena Nakić

**Sažetak**:

*Linearna regresija je metoda modeliranja veze između varijable odgovora i jedne ili više eksplanatornih varijabli. Jedan od problema koji može nastupiti je slučaj kada su barem dvije eksplanatorne varijable linearno zavisne ili približno linearno zavisne. Kažemo da tada imamo problem multikolinearnosti. U ovom radu navodimo metode za detekciju i rješavanje ovoga problema, a zatim prikazujemo kako ih koristiti na stvarnom skupu podataka. Gradimo različite regresijske modele na kojima provjeravamo prisutnost problema multikolinearnosti te ukoliko je prisutan rješavamo ga izostavljanjem varijabli i analizom glavnih komponenti. Pomoću obje metode možemo doći do modela bez problema multikolinearnosti, ali jasnija je interpretacija modela dobivenog izostavljanjem varijabli.*

**Ključne riječi**:

*linearna regresija, eksplanatorne varijable, multikolinearnost*

**Podatci o radu**:

*48 stranica, 1 slika, 5 tablica, 11 literaturnih navoda, izvornik na hrvatskom*

TEMELJNA DOKUMENTACIJSKA KARTICA

*jeziku)*

**Mentor**: *izv. prof. dr. sc. Jurica Perić*

**Neposredna voditeljica**: *dr. sc. Ana Perišić*

**Član povjerenstva**: *mag. math. Marcela Mandarić*

Povjerenstvo za diplomski rad je prihvatilo ovaj rad *12. prosinca 2023.*

BASIC DOCUMENTATION CARD

FACULTY OF SCIENCE, UNIVERSITY OF SPLIT

DEPARTMENT OF MATHEMATICS

MASTER'S THESIS

# MULTICOLINEARITY IN MULTIPLE REGRESSION: DETECTION METHODS AND POSSIBLE SOLUTIONS

Jelena Nakić

**Abstract**:

*Linear regression is a method of modeling the relationship between a response variable and one or more explanatory variables. One of the problems that can occur is the case when at least two explanatory variables are linearly dependent or approximately linearly dependent. In this case, we say that the problem of multicollinearity is present. In this paper, we present several methods for detecting and solving this problem, and then present the application on a real data set. We build different regression models where we check the presence of multicollinearity. We tackle the presence of multicollinearity by omitting variables or applying principal component analysis. In each case it is possible to build a model having no serious problem of multicollinearity. However, the model built on the basis of principal components offers limited interpretation.*

**Key words**:

*linear regression, explanatory variables, multicollinearity*

**Specifications**:

*48 pages, 1 image, 5 tables, 11 references, original in Croatian*

**Mentor**: *associate professor Jurica Perić*

**Immediate mentor**: *Ana Perišić, PhD*

**Committee**: *Marcela Mandarić, mag.math.*

This thesis was approved by a Thesis commettee on *December 12th, 2023.*

*Hvala Josipu, mojoj obitelji i prijateljima koji su bili uz mene tijekom studiranja.*

# Uvod

Francis Galton je u 19.st. proučavao vezu između visine djece i njihovih roditelja. Primjetio je da, iako visoki roditelji imaju visoku djecu te niski roditelji imaju nisku djecu, postoji tendencija prema prosjeku. To bi značilo da visoki roditelji imaju djecu nešto nižu od njih samih, a niski roditelji imaju djecu nešto višu od njih samih. Tendenciju prema prosjeku nazvao je regresijom prema prosjeku. Pomoću regresije mogao je predvidjeti visinu djeteta ukoliko je znao visinu njegovog roditelja. Dakle, regresijska analiza se bavi proučavanjem veze između varijable odgovora i jedne ili više eksplanatornih varijabli. Problem koji može nastupiti je slučaj kada su eksplanatorne varijable međusobno zavisne. Intuitivno je jasno da ćemo u tom slučaju teško razlikovati pojedinačni utjecaj svake eksplanatorne varijable na zavisnu varijablu. Ovaj problem nazivamo problem multikolinearnosti, a cilj ovoga rada je pokazati ocjene prisutnosti multikolinearnosti te metode za uklanjanje multikolinearnosti ukoliko ona postoji.

U prvom poglavlju uvodimo pojam jednostavne linearne regresije, to jest regresije u kojoj postoji samo jedna regresorska varijabla. Obrađena je Metoda najmanjih kvadrata kojom procjenjujemo parametre jednostavne linearne regresije. Također, prikazan je način na koji možemo ispitati značajnost regresijskog parametra. Zatim povećavamo broj regresorskih varijabli i tako uvodimo pojam višestruke linearne regresije. Na isti način pristupamo analizi

kao i kod jednostruke linearne regresije. Dodatno, u ovom dijelu navodimo Gauss-Markovljev teorem koji daje najbolje linearne nepristrane procijenitelje parametara.

U drugom poglavlju uvodimo pojam multikolinearnosti. Radi se o problemu koji se pojavljuje kada su barem dvije regresorske varijable linearno zavisne ili približno linearno zavisne. Uvedeno je nekoliko pokazatelja kojima možemo provjeriti prisutnost multikolinearnosti: faktor inflacije varijance, generalizirani faktor inflacije varijance i kondicioni broj. Također, prikazane su dvije procedure koje uključuju statističke testove Farrar Glauberov test i Mtest koji služe kao pomoć pri otkrivanju multikolinearnosti.

U trećem poglavlju pokušavamo otkloniti problem multikolinearnosti. Navedene su dvije metode: analiza glavnih komponenti i izostavljanje varijabli.

U četvrtom poglavlju je dan sveobuhvatni primjer u kojem, na stvarnom skupu podataka, prikazujemo kako se otkriva i rješava problem multikolinearnosti. U zadnjem poglavlju dana su zaključna razmatranja.

# Sadržaj

# Poglavlje 1

# Linearna regresija

Regresijska analiza bavi se proučavanjem odnosa između odabrane varijable (koju često nazivamo zavisna varijabla) i jedne ili više nezavisnih varijabli temeljem kojih želimo predvidjeti vrijednosti zavisne varijable. Također, često je cilj regresijske analize opisati odnos između zavisne i nezavisnih varijabli. Zavisnu varijablu ćemo označavati s y, dok ćemo nezavisne varijable označavati s $x_1, x_2, ..., x_k$.

## 1.1 Jednostavna linearna regresija

U slučaju kada je k=1, tj. kada postoji samo jedna nezavisna varijabla, govorimo o jednostavnoj linearnoj regresiji.

Relacija između x i y je određena s y = f(x).
Možemo razlikovati dvije vrste relacija:

1. Deterministička ili matematička relacija
2. Statistička relacija koja za dani x ne daje jedinstvenu vrijednost y, ali ju opisuje u terminima vjerojatnosti.

**Primjer 1.1** *y = cijena vožnje*

*x = broj prijeđenih kilometara*

*U ovom primjeru prikazujemo vezu između cijene vožnje i prijeđenih kilometara u nekoj taxi službi.*

*Pretpostavimo da je relacija između broja prijeđenih kilometara x i cijene vožnje y dana s y = 2 + x - 0.01$x^2$. Ovo je primjer determinističke relacije. Za svaki x možemo izračunati jedinstveni y. Primjerice, ako je x = 10, pripadna y vrijednost jednaka je 11. Dakle, osoba će za prijeđeni put od 10km taxistu platiti 11 eura.*

*S druge strane, ako je relacija između y i x dana s x = 2 + x - 0.01$x^2$ + u, gdje je*

$$u = \begin{cases} +0.5\,, \text{ s vjerojatnošću } \frac{1}{2} \\ -0.5\,, \text{ s vjerojatnošću } \frac{1}{2} \end{cases}$$

*tada vrijednost y za dani x nije jedinstvena, već je dana s određenom vjerojatnošću. Primjerice, za x = 10, y = 11.5 s vjerojatnošću $\frac{1}{2}$ ili y = 10.5 s vjerojatnošću $\frac{1}{2}$.*

U nastavku ćemo pretpostavljati da je funkcija f(x) linearna, tj.

$$\mathrm{f(x)} = \alpha + \beta\ \mathrm{x}$$

i pretpostavit ćemo da je relacija stohastička, tj.

$$\mathrm{y} = \alpha + \beta\mathrm{x} + \mathrm{u}$$

pri čemu $u$ ima poznatu vjerojatnosnu distribuciju. Varijablu $u$ nazivamo pogreška ili smetnja.

U gornjoj jednadžbi $\alpha + \beta$x je deterministička komponenta od y, dok je $u$ stohastička ili slučajna komponenta. $\alpha$ i $\beta$ se nazivaju koeficijenti ili parametri regresije, a procijenjujemo ih na temelju opaženih vrijednosti varijabli.

### 1.1. Jednostavna linearna regresija

Postoje tri glavna razloga zašto dodajemo pogrešku $u$:

1. Nepredvidivi element slučajnosti. Na primjer, ako varijabla y predstavlja vrijeme vožnje, a varijabla x predstavlja broj prijeđenih kilometara, postoji nepredvidiv element slučajnosti u vremenu vožnje. To se, na primjer, može odnositi na neočekivani zastoj u prometu zbog pometne nesreće.

2. Učinak velikog broja izostavljenih varijabli. Ponovno, u našem primjeru x nije jedina varijabla koja utječe na varijablu y. Primjerice, brzina vožnje također utječe na varijablu y.

3. Pogreška mjerenja u y. U našem primjeru to bi bila pogreška mjerenja u vremenu vožnje.

Ako imamo n opažanja za y i x, možemo napisati

$$y_i = \alpha + \beta x_i + u_i, \text{ i = 1,2, ... ,n}$$

Naš cilj je procijeniti nepoznate parametre $\alpha$ i $\beta$. Da bismo ih procijenili, potrebno je uvesti sljedeće pretpostavke o pogrešci $u_i$:

1. $\mathrm{E}(u_i) = 0$, $\forall_i \in \{1, ..., n\}$.

2. (Zajednička varijanca) $\mathrm{Var}(u_i) = \sigma^2$, $\forall_i \in \{1, ..., n\}$.

3. (Nezavisnost) $u_i$ i $u_j$ su nezavisne $\forall_{i,j} \in \{1, ..., n\}$, i$\neq$ j.

4. (Nezavisnost s $x_j$) $u_i$ i $x_j$ su nezavisne $\forall_{i,j} \in \{1, ..., n\}$.

5. (Normalnost) $u_i$ su normalno distribuirane $\forall_i \in \{1, ..., n\}$.

Uvjeti 1.-3. nazivaju se Gauss Markovljevi uvjeti.
Kako smo pretpostavili da je $\mathrm{E}(u_i) = 0$, možemo pisati

$$\mathrm{E}(y_i) = \alpha + \beta x_i$$

Kada u ovu jednadžbu uvrstimo procjenitelje parametara $\alpha$ i $\beta$, dobivamo regresijsku funkciju uzorka.

Parametre $\alpha$ i $\beta$ možemo procijeniti Metodom najmanjih kvadrata.

### 1.1.1 Metoda najmanjih kvadrata

Pomoću ove metode određujemo procijenitelje $\hat{\alpha}$ i $\hat{\beta}$ parametara $\alpha$ i $\beta$ redom minimizirajući vrijednost

$$Q = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

Pri tome je Q suma kvadrata pogreški predviđanja pri procjenjivanju vrijednosti $y_i$ uz dane $x_i$.
Intuitivno, ideja metode najmanjih kvadrata je odrediti pravac dan jednadžbom regresije koji prolazi što bliže svim točkama $(x_i, y_i)$, $i = 1, ..., n$. Minimiziranjem vrijednosti Q zapravo minimiziramo sumu kvadrata udaljenosti točaka od pravca.
Da bismo minimizirali Q s obzirom na $\hat{\alpha}$ i $\hat{\beta}$, izjednačavamo s nulom njene prve derivacije po $\hat{\alpha}$ i $\hat{\beta}$. Tako dobivamo sljedeće jednadžbe:

$$\sum_{i=1}^{n} y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} x_i, \text{ tj. } \bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$$
$$\sum_{i=1}^{n} y_i x_i = \hat{\alpha}\sum_{i=1}^{n} x_i + \hat{\beta}\sum_{i=1}^{n} x_i^2$$

Ove dvije jednadžbe nazivamo normalne jednadžbe.
Supstitucijom $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ u jednadžbu $\sum_{i=1}^{n} y_i x_i = \hat{\alpha}\sum_{i=1}^{n} x_i + \hat{\beta}\sum_{i=1}^{n} x_i^2$ dobivamo

$$\sum_{i=1}^{n} y_i x_i = n\bar{x}(\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}\sum_{i=1}^{n} x_i^2$$

Definiramo li

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$
$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

možemo pisati

$$\hat{\beta} S_{xx} = S_{xy}$$

Sada su procijenitelji za $\alpha$ i $\beta$ dani sa

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \tag{1.1}$$

i

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{1.2}$$

Procijenjeni reziduali su

$$\hat{u_i} = y_i - \hat{\alpha} - \hat{\beta} x_i$$

Suma kvadrata reziduala (oznaka RSS) dana je sa:

$RSS = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$
$= \sum_{i=1}^{n}[y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})]^2$
$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + \hat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\hat{\beta} \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$
$= S_{yy} + \hat{\beta}^2 S_{xx} - 2\hat{\beta} S_{xy}$

Kako je $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ dobivamo

$$\text{RSS} = S_{yy} - \hat{\beta} S_{xy}$$

$S_{yy}$ obično označavamo s TSS (ukupna suma kvadrata), dok $\hat{\beta} S_{xy}$ označavamo s ESS (objašnjena suma kvadrata). Iz ovoga slijedi:

$$\text{TSS} = \text{ESS} + \text{RSS}$$

Omjer objašnjene sume kvadrata i ukupne sume kvadrata označavamo s $r_{xy}^2$, a u slučaju jednostavne regresije $r_{xy}$ je jednak koeficijentu korelacije. Dakle, vrijedi $r_{xy}^2 = \frac{ESS}{TSS}$ i $1 - r_{xy}^2 = \frac{RSS}{TSS}$. Vrijednost $r_{xy}^2$ naziva se koeficijent determinacije. Koeficijent determinacije poprima vrijednosti iz segmenta [0,1] te temeljem vrijednosti ovog koeficijenta moguće je ocijeniti reprezentativnost modela. Ako je vrijednost $r_{xy}^2$ blizu nule, varijabla x objašnjava jako mali dio varijabilnosti od y. Nasuprot tome, što je koeficijent determinacije bliži 1, to je model reprezentativniji, odnosno varijabla x objašnjava veći dio varijabilnosti zavisne varijable y.

Koeficijent determinacije dan je sa

$$r_{xy}^2 = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS} = \frac{\hat{\beta}S_{xy}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

iz čega slijedi da je koeficijent korelacije dan sa

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \hat{\beta}\sqrt{\frac{S_{xx}}{S_{yy}}}$$

Procijenitelji $\hat{\alpha}$ i $\hat{\beta}$ najmanjih kvadrata daju procijenjeni pravac koji ima manju vrijednost RSS od bilo kojeg drugog pravca.

Uvrstimo li u jednadžbu $\hat{y} = \hat{\alpha} + \hat{\beta}x$ izraze (1.1) i (1.2) sređivanjem izraza dobit ćemo

$$\frac{\hat{y}-\bar{y}}{\sqrt{S_{yy}}} = r_{xy}\frac{x-\bar{x}}{\sqrt{S_{xx}}}$$

Interpretiramo:

Pretpostavimo li da je $r_{xy} > 0$ i da je x vrijednost varijable koja je veća od prosjeka za jednu standardnu devijaciju, tada će i vrijednost varijable y biti veća od prosjeka $r_{xy}$ standardnih devijacija, pri čemu je $r_{xy} \leq 1$.

Stoga predviđena vrijednost odstupa od svog prosjeka za manje standardnih devijacija nego prediktor.

**Primjer 1.2** *Ponovno ćemo promatrati podatke iz prethodnog primjera.*

| *Opažanje* | $x$ | $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|---|
| *1* | *5* | *6.75* | *25* | *45.56* | *33.75* |
| *2* | *4* | *5.85* | *16* | *34.22* | *23.4* |
| *3* | *10* | *11* | *100* | *121* | *110* |
| *4* | *3* | *4.90* | *9* | *24.01* | *14.7* |
| *5* | *8* | *9.35* | *64* | *87.42* | *74.8* |
| *6* | *15* | *14.75* | *225* | *217.56* | *221.25* |
| *Ukupno* | *45* | *52.6* | *439* | *529.77* | *477.9* |

*Računanjem dobivamo sljedeće podatke:*

$$\bar{x} = \frac{45}{6} = 7.5$$
$$\bar{y} = \frac{52.6}{6} = 8.77$$
$$S_{xx} = 439 - 6 * 7.5^2 = 101.5$$
$$S_{xy} = 477.9 - 6 * 7.5 * 8.77 = 83.25$$
$$S_{yy} = 529.77 - 6 * 8.77^2 = 68.29$$
$$r_{xy} = \frac{83.25}{\sqrt{101.5 * 68.29}} = 0.9999 \qquad ili \qquad r_{xy}^2 = 0.9999$$
$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{83.25}{101.5} = 0.82$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 8.77 - 0.82 * 7.5 = 2.62$$

*Stoga je jednadžba regresije dana s* $y = 2.62 + 0.82x$

### 1.1.2 Značajnost parametara

Prilikom donošenja zaključaka o značajnosti parametara u statističkom modelu važna je pretpostavka o distribuciji zavisne varijable, odnosno pogrešaka. Zbog pretpostavke normalnosti (i nezavisnosti) grešaka slijedi da su procjenitelji $\hat{\alpha}$ i $\hat{\beta}$ normalno distribuirani te vrijedi:

$$E(\hat{\alpha}) = \alpha, \qquad Var(\hat{\alpha}) = \sigma^2(\tfrac{1}{n} + \tfrac{\bar{x}^2}{S_{xx}})$$

Nadalje vrijedi

$$E(\hat{\beta}) = \beta, \qquad Var(\hat{\beta}) = \tfrac{\sigma^2}{S_{xx}}$$
$$\text{i } Cov(\hat{\alpha}, \hat{\beta}) = \sigma^2(\tfrac{-\bar{x}}{S_{xx}})$$

Istaknimo da uvjet normalnosti i nezavisnosti pogrešaka nije nužan da bi procjenitelji $\hat{\alpha}$ i $\hat{\beta}$ bili nepristrani. Ovi rezultati će biti korisni ukoliko je poznata varijanca pogreške $\sigma^2$. Međutim, u praksi, $\sigma^2$ nije poznata i treba biti procijenjena.

Nepristrani procjenitelj zajedničke varijance slučajnih pogrešaka $\sigma^2$ je statistika

$$\hat{\sigma}^2 = \tfrac{RSS}{n-2}$$

Vrijedi, $\frac{RSS}{\sigma^2}$ ima $\chi^2$-distribuciju sa n-2 stupnja slobode.

Da bismo donijeli zaključke o $\alpha$ i $\beta$ koristit ćemo t-distribuciju.

Budući da su dane dvije varijable $X_1 \sim N(0,1)$ i $X_2 \sim \chi^2(k)$ pri čemu su $X_1$ i $X_2$ nezavisne, onda

$$X = \frac{X_1}{\sqrt{\frac{X_2}{k}}}$$

ima t(k) distribuciju.

U ovom slučaju vrijedi $\frac{\hat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$.

Također, $\frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$ te su slučajne varijable $\frac{\hat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}}$ i $\frac{RSS}{\sigma^2}$ nezavisne. Tada vrijedi

$$\frac{\frac{\hat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}}}{\sqrt{\frac{RSS}{(n-2)\sigma^2}}} \sim t(n-2)$$

Ukoliko u gornjoj jednadžbi zamijenimo $\sigma^2$ sa njenim procijeniteljem, kao rezultat dobivamo da $\frac{\hat{\beta}-\beta}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$ ima t(n-2) distribuciju. Vrijednost $\frac{\hat{\sigma}^2}{S_{xx}}$ procijenjena na temelju uzorka je procijenjena varijanca od $\hat{\beta}$, a njezin korijen se

naziva standardna pogreška i označava se sa $SE(\hat{\beta})$. Na isti način dobivamo da i $\frac{\hat{\alpha}-\alpha}{SE(\hat{\alpha})}$ ima t(n-2) distribuciju.
Ove distribucije možemo koristiti da bismo dobili pouzdane intervale za $\alpha$ i $\beta$ te za testiranje hipoteza o $\alpha$ i $\beta$.

**Primjer 1.3** *Ponovno ćemo iskoristiti prethodni primjer kako bi pokazali kako se računa standardna pogreška procjenitelja. Prethodno smo procijenili regresijsku jednadžbu*

$$y = 2.62 + 0.82x$$

*Imamo*

$$Var(\hat{\alpha}) = \sigma^2(\tfrac{1}{n} + \tfrac{\bar{x}^2}{S_{xx}}) = \sigma^2(\tfrac{1}{6} + \tfrac{56.25}{101.5}) = 0.72\sigma^2$$
$$Var(\hat{\beta}) = \tfrac{\sigma^2}{S_{xx}} = \tfrac{\sigma^2}{101.5} = 0.0099\sigma^2$$
$$\hat{\sigma}^2 = \tfrac{1}{n-2}(S_{yy} - \tfrac{S_{xy}^2}{S_{xx}}) = \tfrac{1}{4}[68.29 - \tfrac{83.25^2}{101.5}] = 0.00215$$
$$SE(\hat{\alpha}) = \sqrt{0.72 * 0.00215} = 0.04$$
$$SE(\hat{\beta}) = \sqrt{0.0099 * 0.00215} = 0.0046$$

*Koristeći tablicu t-distribucija za n-2=4 stupnja slobode dobivamo*

$$P[-2.78 < \tfrac{\hat{\alpha}-\alpha}{SE(\hat{\alpha})} < 2.78] = 0.95$$
$$i\ P[-2.78 < \tfrac{\hat{\beta}-\beta}{SE(\hat{\beta})} < 2.78] = 0.95$$

*Uvrštavanjem vrijednosti $\hat{\alpha}$, $\hat{\beta}$, $SE(\hat{\alpha})$ i $SE(\hat{\beta})$ dobivamo da su 95%-pouzdani intervali (2.51,2.73) i (0.81,0.83) za $\alpha$ i $\beta$ redom.*
*Možemo primjetiti kako su pouzdani intervali za $\alpha$ i $\beta$ simetrični oko $\hat{\alpha}$ i $\hat{\beta}$ redom.*

Sada pretpostavimo da želimo testirati hipotezu o značajnosti modela, $H_0$: $\beta = 0$.
Znamo da $T = \frac{\hat{\beta}-\beta}{SE(\hat{\beta})}$ ima t(n-2) distribuciju.

Ako je alternativna hipoteza $\beta \neq 0$ onda je $|T|$ testna statistika. Vrijednost testne statistike je tada

$$T = \frac{0.82-0}{0.0046} = 178.26.$$

Kako je vrijednost testne statistike veća od kritične vrijednosti t-distribucije, odbacujemo nultu hipotezu. Isto smo mogli zaključiti proučavanjem pouzdanog intervala za $\beta$: budući da procijenjeni interval ne sadrži 0 možemo zaključiti kako odbacujemo nul-hipotezu.

## 1.2 Višestruka linearna regresija

U višestrukoj linearnoj regresiji proučavamo vezu između zavisne varijable y i nezavisnih varijabli $x_1, x_2, ..., x_k$ , $k \geq 2$.
Pretpostavljeni model je

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i, i = 1, 2, ..., n$$

Pogreške $u_i$ ponovno dodajemo zbog mogućih odstupanja u mjerenju varijable y te zbog specifičnosti odnosa varijable y i varijabli $x_1, x_2, ..., x_k$.
Kao i za pogreške $u_i$ u linearnoj regresiji, tako ćemo pretpostaviti i za $u_i$ kod višestruke regresije:

1. $\mathrm{E}(u_i) = 0, \forall_i \in \{1, ..., n\}$.
2. $\mathrm{Var}(u_i) = \sigma^2$, $\forall_i \in \{1, ..., n\}$.
3. $u_i$ i $u_j$ su nezavisni $\forall_i \in \{1, ..., n\}, i \neq j$.
4. $u_i$ i $x_j$ su nezavisni $\forall_{i,j} \in \{1, ..., n\}$.
5. $u_i$ su normalno distribuirani $\forall_i \in \{1, ..., n\}$.

Dodatno, pretpostavit ćemo da $x_1, x_2, ..., x_k$ nisu kolinearni, tj. ne postoji deterministička linearna veza između njih.
U sljedećem primjeru prikazat ćemo slučaj kada su nezavisne varijable kolinearne.

**Primjer 1.4** *Pretpostavimo da je dana jednadžba regresije*

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$$

*pri čemu su* $x_1$ *i* $x_2$ *povezani relacijom*

$$x_1 - 3x_2 = -2$$

*Tada je* $x_1 = 3x_2 - 2$ *pa jednadžbu regresije možemo zapisati na sljedeći način:*

$$y = \alpha + \beta_1(3x_2 - 2) + \beta_2 x_2 + u$$
$$y = (\alpha - 2\beta_1) + (3\beta_1 + 2\beta_2)x_2 + u$$

*Stoga možemo procijeniti* $(\alpha - 2\beta_1)$ *i* $(3\beta_1 + 2\beta_2)$*, ali ne možemo procijeniti* $\alpha, \beta_1, \beta_2$ *zasebno.*

Slučaj kada postoji egzaktna linearna veza između nezavisnih varijabli $x_1, x_2, ..., x_k$ poznat je kao egzaktna ili savršena kolinearnost. U slučaju kada imamo višestruku linearnu regresiju s dvije nezavisne varijable, egzaktna veza povlači da je koeficijent korelacije između varijabli $x_1$ i $x_2$ jednak +1 ili -1. Zasad ćemo pretpostaviti da ne postoji savršena kolinearnost.

### 1.2.1 Metoda najmanjih kvadrata

U ovom odjeljku prikazat ćemo jednu metodu procjene nepoznatih parametara u višestrukoj regresiji. Jednako kao u slučaju jednostavne linearne

regresije, prikazana je metoda najmanjih kvadrata, a jednostavnosti radi, promatramo slučaj s dva prediktora, odnosno promatramo model s dvije nezavisne varijable.

Pretpostavit ćemo model s dvije nezavisne varijable

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, i = 1, 2, ..., n$$

Pomoću metode najmanjih kvadrata dobivamo procijenitelje $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$ od $\alpha, \beta_1, \beta_2$ redom minimizirajući vrijednost

$$Q = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2$$

Derivirajući Q po $\hat{\alpha}$, $\hat{\beta}_1$ i $\hat{\beta}_2$ te izjednačavajući s 0 dobivamo

$$\begin{aligned}
\frac{\partial Q}{\partial \hat{\alpha}} = 0 &\longrightarrow \sum_{i=1}^{n} 2(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})(-1) = 0 \\
\frac{\partial Q}{\partial \hat{\beta}_1} = 0 &\longrightarrow \sum_{i=1}^{n} 2(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})(-x_{1i}) = 0 \\
\frac{\partial Q}{\partial \hat{\beta}_2} = 0 &\longrightarrow \sum_{i=1}^{n} 2(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})(-x_{2i}) = 0
\end{aligned}$$

Ove tri jednadžbe nazivamo normalne jednadžbe. Možemo ih pisati redom

$$\bar{y} = \hat{\alpha} + \hat{\beta}_1 \bar{x_1} + \hat{\beta}_2 \bar{x_2} \tag{1.3}$$

$$\sum_{i=1}^{n} x_{1i} y_i = \hat{\alpha} \sum_{i=1}^{n} x_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{2i} \tag{1.4}$$

$$\sum_{i=1}^{n} x_{2i} y_i = \hat{\alpha} \sum_{i=1}^{n} x_{2i} + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i}^2 \tag{1.5}$$

Uvrštavajući vrijednost $\bar{y}$ u jednadžbu (1.4) dobivamo

$$\sum_{i=1}^{n} x_{1i} y_i = n\bar{x_1}(\bar{y} - \hat{\beta}_1 \bar{x_1} - \hat{\beta}_2 \bar{x_2}) + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{2i}$$

Definirajmo

$$\begin{aligned}
S_{11} &= \sum_{i=1}^{n} x_{1i}^2 - n\bar{x_1}^2, & S_{1y} &= \sum_{i=1}^{n} x_{1i} y_i - n\bar{x_1}\bar{y} \\
S_{12} &= \sum_{i=1}^{n} x_{1i} x_{2i} - n\bar{x_1}\bar{x_2}, & S_{2y} &= \sum_{i=1}^{n} x_{2i} y_i - n\bar{x_2}\bar{y} \\
S_{22} &= \sum_{i=1}^{n} x_{2i}^2 - n\bar{x_2}^2, & S_{yy} &= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2
\end{aligned}$$

Sada uvrštavanjem u normalne jednadžbe dobivamo

$$S_{1y} = \hat{\beta}_1 S_{11} + \hat{\beta}_2 S_{12}$$
$$S_{2y} = \hat{\beta}_1 S_{12} + \hat{\beta}_2 S_{22}$$

Rješavanjem ovih dviju jednadžbi dobivamo $\hat{\beta}_1$ i $\hat{\beta}_2$. Dobivamo

$$\hat{\beta}_1 = \frac{S_{22}S_{1y} - S_{12}S_{2y}}{S_{11}S_{22} - S_{12}^2}$$
$$\hat{\beta}_2 = \frac{S_{11}S_{2y} - S_{12}S_{1y}}{S_{11}S_{22} - S_{12}^2}$$

Kada izračunamo $\hat{\beta}_1$ i $\hat{\beta}_2$ možemo izračunati

$$\hat{\alpha} = \bar{y} - \hat{\beta}_1 \bar{x_1} - \hat{\beta}_2 \bar{x_2}$$

Također, vrijede sljedeći izrazi

$$RSS = S_{yy} - \hat{\beta}_1 S_{1y} - \hat{\beta}_2 S_{2y}$$
$$ESS = \hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y}$$
$$R^2_{y.12} = \frac{\hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y}}{S_{yy}}$$

$R^2_{y.12}$ se naziva koeficijent višestruke determinacije, a njegov korijen se naziva koeficijent višestruke korelacije. Prvi podindeks odnosi se na zavisnu varijablu, a podindeksi navedeni nakon točke odnose se na nezavisne varijable. Procedura je analogna za više od dvije nezavisne varijable.

**Primjer 1.5** *U tablici su dani podaci o masi 10 srednjoškolaca i njihovih roditelja.*

### 1.2. Višestruka linearna regresija

| *masa djeteta (kg)* | *masa majke (kg)* | *masa oca (kg)* |
|---|---|---|
| *58* | *65* | *87* |
| *66* | *74* | *95* |
| *51* | *67* | *80* |
| *56* | *50* | *70* |
| *42* | *66* | *88* |
| *54* | *63* | *95* |
| *52* | *59* | *100* |
| *46* | *65* | *77* |
| *85* | *90* | *105* |
| *70* | *81* | *93* |

*Vrijedi*

$$\bar{Y} = 58$$
$$\bar{X}_1 = 68$$
$$\bar{X}_2 = 89$$

*U sljedećoj tablici su izračuni potrebni za daljnje računanje:*

### 1.2. Višestruka linearna regresija

| $i$ | $X_{1i}^2$ | $X_{2i}^2$ | $X_{1i}X_{2i}$ | $X_{1i}Y_i$ | $X_{2i}Y_i$ | $Y_i^2$ |
|---|---|---|---|---|---|---|
| *1* | *4225* | *7569* | *5655* | *3770* | *5046* | *3364* |
| *2* | *5476* | *9025* | *7030* | *4884* | *6270* | *4356* |
| *3* | *4489* | *6400* | *5360* | *3417* | *4080* | *2601* |
| *4* | *2500* | *4900* | *3500* | *2800* | *3920* | *3136* |
| *5* | *4356* | *7744* | *5808* | *2772* | *3696* | *1764* |
| *6* | *3969* | *9025* | *5985* | *3402* | *5130* | *2916* |
| *7* | *3481* | *10000* | *5900* | *3068* | *5200* | *2704* |
| *8* | *4225* | *5929* | *5005* | *2990* | *3542* | *2116* |
| *9* | *8100* | *11025* | *9450* | *7650* | *8925* | *7225* |
| *10* | *6561* | *8649* | *7533* | *5670* | *6510* | *4900* |
| *Ukupno* | *47382* | *80266* | *61226* | *40423* | *52319* | *35082* |

*Slijedi*

$$S_{11} = 1142 \qquad S_{1y} = 983$$
$$S_{12} = 706 \qquad S_{2y} = 699$$
$$S_{22} = 1056 \qquad S_{yy} = 1442$$

*Normalne jednadžbe su*

$$1142\hat{\beta}_1 + 706\hat{\beta}_2 = 983$$
$$706\hat{\beta}_1 + 1056\hat{\beta}_2 = 699$$

*Rješavanjem sustava dobivamo*

$$\hat{\beta}_1 = 0.77$$
$$\hat{\beta}_2 = 0.15$$

*Također vrijedi*

$$\hat{\alpha} = \bar{y} - \hat{\beta}_1\bar{x_1} - \hat{\beta}_2\bar{x_2} = -7.71 \; R^2 = \frac{\hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y}}{S_{yy}} = 0.5976$$

*Jednadžba regresije je*

$$\hat{Y} = -7.71 + 0.77X_1 + 0.15X_2$$

### Matrični pristup

Jednostavniji pristup metodi najmanjih kvadrata je matrični.
Neka je dan model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1}$$

kojemu odgovaraju podaci

$$y_i, x_{i1}, x_{i2}, ..., x_{i,p-1}, \qquad i = 1, 2, ..., n$$

Vrijednosti $y_i, i = 1, 2, ..., n$, su prikazane u matrici Y. Nepoznanice $\beta_0, \beta_1, ..., \beta_{p-1}$ su prikazane pomoću vektora $\beta$. Neka je $X_{n\times p}$ matirca

$$\begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1,p-1} \\ 1 & x_{21} & x_{22} & ... & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{n,p-1} \end{bmatrix}$$

Za dani $\beta$, vektor predviđenih vrijednosti, $\hat{Y}$, možemo pisati u obliku

$$\underbrace{\hat{Y}}_{n\times 1} = \underbrace{X}_{n\times p} \underbrace{\beta}_{p\times 1}$$

Tada problem najmanjih kvadrata prelazi u problem pronalaženja vrijednosti $\beta$ tako da vrijednost $S(\beta)$ bude minimalna, pri čemu je
$S(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_{p-1} x_{i,p-1})^2$
$= ||Y - X\beta||^2$
$= ||Y - \hat{Y}||^2$

Deriviramo li S po svim varijablama $\beta_k, k = 0, ..., p-1$, i derivacije izjednačimo s nulom, vrijednosti parametara za koje se postiže minimum $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p-1}$ će zadovoljavati sljedećih p jednadžbi

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + ... + \hat{\beta}_{p-1} \sum_{i=1}^n x_{i,p-1} = \sum_{i=1}^n y_i$$
$$\hat{\beta}_0 \sum_{i=1}^n x_{ik} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}x_{ik} + ... + \hat{\beta}_{p-1} \sum_{i=1}^n x_{ik}x_{i,p-1} = \sum_{i=1}^n y_i x_{ik},$$
$$k = 1, ..., p-1$$

Ovih p jednadžbi možemo zapisati u matričnom obliku

$$X^T X\hat{\beta} = X^T Y,$$

a nazivaju se normalne jednadžbe.
Ako je $X^T X$ nesingularna matrica, onda je rješenje jednadžbe

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Sljedeća lema daje kriterij za postojanje i jedinstvenost rješenja normalnih jednadžbi.

**Lema 1.6** *$X^T X$ je nesingularna ako i samo ako je rang od X jednak p.*

**Dokaz.** Prvo pretpostavimo da je $X^T X$ singularna. Tada postoji ne-nul vektor $u$ takav da je $X^T Xu = 0$. Množenjem ove jednadžbe slijeva s $u^T$ dobivamo

$$\begin{aligned} 0 &= u^T X^T Xu \\ &= (Xu)^T (Xu) \end{aligned}$$

pa je $Xu = 0$, stupci od X su linearno nezavisni i rang od X je manji od $p$. Obratno, pretpostavimo da je rang od X manji od $p$. Sada postoji ne-nul vektor $u$ takav da je $Xu = 0$. Slijedi da je $X^T Xu = 0$ pa je $X^T X$ singularna.

■

Lako se pokaže da vrijedi:

1. Ako pogreške imaju srednju vrijednost jednaku nuli, procjenitelji najmanjih kvadrata su nepristrani.

2. Ako su pogreške nekorelirane, imaju srednju vrijednost jednaku nuli i konstantnu varijancu $\sigma^2$, matrica kovarijanci procjenitelja najmanjih kvadrata $\hat{\beta}$ je $\sum_{\hat{\beta}\hat{\beta}} = \sigma^2(X^TX)^{-1}$.

3. Ako pogreške nisu korelirane i imaju konstantnu varijancu $\sigma^2$, nepristrani procjenitelj od $\sigma^2$ je $s^2 = \frac{||Y-\hat{Y}||^2}{n-p}$.

### 1.2.2 Značajnost parametara

Kako su komponente od $\hat{\beta}$ linearne kombinacije nezavisnih normalno distribuiranih slučajnih varijabli, one su također normalno distribuirane. Preciznije, svaka komponenta $\hat{\beta}_i$ od $\hat{\beta}$ je normalno distribuirana s očekivanom vrijednosti $\beta_i$ i varijancom $\sigma^2 c_{ii}$, gdje je $C = (X^TX)^{-1}$. Stoga standardna pogreška od $\hat{\beta}_i$ može biti procijenjena s

$$s_{\hat{\beta}_i} = \sigma\sqrt{c_{ii}}.$$

Ovaj rezultat ćemo koristiti da bismo konstruirali intervale pouzdanosti i testne hipoteze.
Pod pretpostavkom normalnosti vrijedi

$$\frac{\hat{\beta}_i-\beta_i}{s_{\hat{\beta}_i}} \sim t_{n-p}.$$

Slijedi da je $100(1-\alpha)\%$ pouzdani interval za $\beta_i$ dan s

$$\hat{\beta}_i \pm t_{n-p}(\alpha/2)s_{\hat{\beta}_i}.$$

Da bismo testirali nul hipotezu $H_0 : \beta_i = \beta_{i0}$ možemo koristiti testnu statistiku

$$t = \frac{\beta_i-\beta_{i0}}{s_{\hat{\beta}_i}}.$$

Pod pretpostavkom da vrijedi $H_0$, ova statistika slijedi t-distribuciju s n-p stupnjeva slobode. Najčešće testirana nul hipoteza je $H_0 : \beta_i = 0$, koja tvrdi da $x_i$ nije značajna varijabla u modelu.

U sljedećem teoremu pokazat ćemo da procjenitelj dobiven metodom najmanjih kvadrata ima najmanju varijancu od svih linearnih nepristranih procjenitelja.

**Teorem 1.7 (Gauss-Markovljev teorem)** *Ako je* $E(Y) = X\beta$ *i* $Cov(Y) = \sigma^2 I$, *procjenitelji dobiveni metodom najmanjih kvadrata* $\hat{\beta}_j, j = 0, ..., k$, *imaju minimalnu varijancu među svim linearnim nepristranim procjeniteljima.*

**Dokaz.** Pretpostavimo da je AY linearni procjenitelj za $\beta$ i tražimo A takav da je AY nepristrani procjenitelj minimalne varijance za $\beta$. Da bi AY bio nepristrani procjenitelj za $\beta$ mora vrijediti $E(AY) = \beta$. Koristeći pretpostavku linearnosti $E(Y) = X\beta$ dobivamo

$$E(AY) = AE(Y) = AX\beta = \beta$$

iz čega slijedi uvjet AX = I.
Matrica varijanci i kovarijanci procjenitelja AY je dana sa

$$Cov(AY) = A(\sigma^2 I)A^T = \sigma^2 AA^T.$$

Varijance od $\hat{\beta}_j$ nalaze se na dijagonali matrice $\sigma^2 AA^T$, te je stoga matrica A (uz uvjet AX = I) takva da su dijagonalni elementi od $AA^T$ minimalni. Nadalje, vrijedi

$$AA^T = [A-(X^TX)^{-1}X^T+(X^TX)^{-1}X^T][A-(X^TX)^{-1}X^T+(X^TX)^{-1}X^T]^T$$

Sređivanjem izraza dobivamo

$$AA^T = [A - (X^TX)^{-1}X^T][A - (X^TX)^{-1}X^T]^T + (X^TX)^{-1}$$

Matrica $[A - (X^TX)^{-1}X^T][A - (X^TX)^{-1}X^T]^T$ je pozitivno semidefinitna i dijagonalni elementi su joj nenegativni. Lako se pokaže da su dijagonalni elementi jednaki nula za $A = (X^TX)^{-1}X^T$. (Ovakva matrica A zadovoljava uvjet AX = I). Dobiveni procjenitelj minimalne varijance za $\beta$ je

$$AY = (X^TX)^{-1}X^TY,$$

a on je jednak procjenitelju $\hat{\beta}$. ■

Gauss-Markovljev teorem može se izreći na sljedeći način:
Ako je $E(Y) = X\beta$ i $Cov(Y) = \sigma^2 I$ procjenitelji najmanjih kvadrata $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ su najbolji linearni nepristrani procjenitelji (BLUE).
U ovom izrazu najbolji označava minimalnu varijancu, a linearni označava da su procjenitelji linearne funkcije.

Postoje i druge metode procjene parametara regresijskog modela. Na primjer, metoda maksimalne vjerodostojnosti (MLE). U slučaju kada su zadovoljeni uvjeti nezavisnosti i normalnosti grešaka, te homoskedastičnosti, MLE procjenitelj linearnog regresijskog modela jednak je procjenitelju dobivenom metodom najmanjih kvadrata.

# Poglavlje 2

# Problem multikolinearnosti i odabrane metode detekcije

Gauss-Markovljev teorem navodi kako među svim linearnim nepristranim procjeniteljima, procjenitelj dobiven metodom najmanjih kvadrata ima najmanju varijancu. Ipak, samim teoremom ne možemo zaključiti ništa o veličini same varijance procjenitelja. Procjenitelj može imati i neograničenu varijancu u slučaju postojanja linearno zavisnih ili približno linearno zavisnih varijabli u modelu. U tom slučaju govorimo o prisutnosti problema multikolinearnosti. Problem multikolinearnosti je prisutan ako su barem dvije regresorske varijable linearno zavisne ili približno linearno zavisne. Razlikujemo dva tipa multikolinearnosti: savršena multikolinearnost i približna multikolinearnost.

Savršena multikolinearnost prisutna je ako su dvije ili više regresorskih varijabli linearno zavisne. U tom slučaju je determinanta matrice $X^TX$ jednaka nuli te je matrica $X^TX$ singularna. S obzirom da je vektor procijenjenih

parametara, određen metodom najmanjih kvadrata

$$\hat{\beta} = (X^T X)^{-1} X^T y = \frac{adj(X^T X)}{det(X^T X)} X^T y, \tag{2.1}$$

to ne postoji jedinstveno rješenje jednadžbe (2.1), tj. procjene parametara nisu jednoznačno određene.

Savršena multikolinearnost se rijetko javlja pri primjeni stvarnih podataka, no može se pojaviti ukoliko se rabe eksplanatorne varijable, a ne vodi se računa o tome da u regresiji koja sadrži konstantni član, broj dummy varijabli za svaku kvalitativnu varijablu mora biti za jedan manji nego li je broj njezinih modaliteta.

Mnogo je češći i ozbiljniji problem približne multikolinearnosti ili približne linearne zavisnosti regresorskih varijabli, a prisutan je ako su dvije ili više regresorskih varijabli jako korelirane. U tom slučaju je determinanta matrice $X^T X$ približno jednaka nuli pa je matrica $X^T X$ neprikladna za invertiranje. Zbog jednadžbe (2.1), vrijednosti vektora procijenjenih parametara bit će brojčano nepouzdane.

Nadalje, kako je

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}, \tag{2.2}$$

$$Var(\hat{\beta}) = \sigma^2 \frac{adj(X^T X)}{det(X^T X)}, \tag{2.3}$$

to će zbog $det(X^T X) \approx 0$, varijance pa i standardne pogreške parametara biti velike. Velike vrijednosti postizat će i elementi van glavne dijagonale, odnosno kovarijance $Cov(\hat{\beta}_i, \hat{\beta}_j)$, te je vrijednosti procijenjenih parametara moguće izračunati s pogrešnim predznakom. Zbog velikih standardnih pogrešaka

$$SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{c_{jj}}, \qquad c_{jj} = (X^T X)^{-1}_{jj} \tag{2.4}$$

empirijski t omjeri ($t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$) će biti nerealno mali, što će navoditi zaključak da pojedine regresorske varijable nisu značajne, te da ih treba isključiti iz modela. Također, zbog velikih standardnih pogrešaka intervalne procjene parametara bit će vrlo neprecizne.

Problem približne linearne zavisnosti često se javlja u vremenskim regresijskim modelima koji kao varijable sadrže vremenske nizove uključenih pojava. Kako se pojave tijekom vremena često razvijaju na sličan način, vremenski nizovi (stupci u matrici X) mogu biti jako korelirani.

**Primjer 2.1** *Za početak ćemo promotriti primjer u kojem postoji skoro savršena kolinearnost između nezavisnih varijabli.*

*Neka je dana normalno distribuirana varijabla* $X_1 \sim N(5, 9)$*, te definirajmo varijablu*

$$X_2 = 6 + 2X_1 + \epsilon_1, \text{ pri čemu je } \epsilon_1 \sim N(0, 1) \text{ greška.}$$

*Korelacija između ovih dviju varijabli jednaka je 0.99837, dakle skoro pa savršena.*

*Sada ćemo definirati zavisnu varijablu*

$$Y = 2 + 3X_1 + 7X_2 + \epsilon, \text{ gdje je } \epsilon \sim N(0, 1) \text{ greška.}$$

*Promatramo li univarijantni model linearne regresije koji uključuje samo nezavisnu varijablu* $X_2$*, dobivamo da su koeficijenti linearne regresije*

$$\beta_0 = -6.77806,$$
$$\beta_1 = 8.48328$$

*S druge strane, ako promatramo multivarijantni model koji uključuje i varijablu $X_1$ i varijablu $X_2$ dobivamo da su koeficijenti linearne regresije*

$$\beta_0 = 2.8186,$$
$$\beta_1 = 3.2268,$$
$$\beta_2 = 6.8887$$

*Možemo primjetiti kako postoje velike razlike u koeficijentima linearne regresije u univarijantnom i multivarijantnom modelu. To ukazuje na probleme s multikolinearnosti.*

Može se pokazati da je varijanca procjenitelja $\hat{\beta}_l$ dana s

$$Var(\hat{\beta}_l) = \hat{\sigma}^2 (X^T X)^{-1}_{ll} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_{il} - \overline{x_l})^2 (1 - R_l^2)} = \frac{\hat{\sigma}^2}{S_{ll}(1 - R_l^2)},$$

gdje $R_l^2$ predstavlja kvadrat koeficijenta višestruke korelacije između $x_l$ i ostalih nezavisnih varijabli, odnosno koeficijent determinacije. Lako je vidjeti da će $Var(\hat{\beta}_l)$ imati veliku vrijednost ukoliko

1. $\sigma^2$ ima visoku vrijednost, odnosno, što je prilagodba modela bolja, manja je varijanca procjenitelja.

2. $S_{ll}$ ima nisku vrijednost, odnosno, što je disperzija prediktora manja, veća je varijanca procjenitelja.

3. $R_l^2$ ima visoku vrijednost, odnosno, što je prediktor jače koreliran s ostalim prediktorima, varijanca procjenitelja će biti veća.

## 2.1 Faktor inflacije varijance (VIF)

Jedan od najčešće korištenih pokazatelja multikolinearnosti je faktor inflacije varijance (VIF). Za procjenitelja $\hat{\beta}_l$ faktor inflacije varijance definiramo kao

$$VIF(\hat{\beta}_l) = \frac{1}{1-R_l^2},$$

gdje je $R_l^2$ kvadrat koeficijenta višestruke korelacije između $x_l$ i ostalih nezavisnih varijabli, odnosno koeficijent determinacije. Promatrajući formulu $Var(\hat{\beta}_l) = \frac{\sigma^2}{S_{ll}(1-R_l^2)}$, $VIF(\hat{\beta}_l)$ možemo interpretirati kao omjer stvarne varijance od $\hat{\beta}_l$ i varijance od $\hat{\beta}_l$ koju bismo dobili kada $x_l$ ne bi bio koreliran s $x_1, x_2, ..., x_{l-1}, x_{l+1}, ..., x_n$.

Idealna situacija bi bila kada bi svi $x_i, i = 1, ..., n$ bili nekorelirani. Dakle, $VIF_l$ uspoređuje stvarno stanje s idealnim stanjem.

Ne postoji konsenzus oko granice vrijednosti ovog pokazatelja koja upućuje na postojanje multikolinearnosti. Na primjer, može se smatrati da je ozbiljan problem multikolinearnosti prisutan ako je $R_l^2 > 0.8$, odnosno $VIF_l > 5$. Ponekad se uzima i stroga granica gdje vrijednost $VIF_l > 10$, što odgovara vrijednosti $R_l^2 > 0.9$, upućuje na postojanje problema multikolinearnosti.

**Primjer 2.2** *U Primjeru 2.1. je spomenuto kako je koeficijent korelacije između varijabli $x_1$ i $x_2$ jednak 0.99837. Odavde slijedi da je*

$$VIF_1 = VIF_2 = \frac{1}{1-0.99837^2} = 307.82,$$

*što je mnogo veće od 5 pa možemo zaključiti kako postoji ozbiljan problem multikolinearnosti.*

## 2.2 Generalizirani faktor inflacije varijance (GVIF)

Faktor inflacije varijance generaliziran je na slučaj kada u linearnom modelu $y = X\beta + \epsilon$ promatramo podskupove parametara. U tom slučaju predložen je generalizirani faktor inflacije varijance (GVIF) kao mjera multikolinearnosti. Neka je dan model

$$y = X_0\beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$$

pri čemu je $n \times p$ matrica dizajna X podijeljena na matrice: $X_0$ tipa $n \times p_0$, $X_1$ tipa $n \times p_1$, $X_2$ tipa $n \times p_3$. Matrica $X_0$ sadrži varijable čije su vrijednosti fiksne i ne mogu biti odabrane na drugačiji način (na primjer konstantni član u regresiji), $X_1$ predstavlja skup varijabli čiji utjecaj želimo promatrati simultano (na primjer skup dummy varijabli vezanih uz jednog kategorijalnog prediktora), te $X_2$ skup ostalih varijabli. Promotrimo slučaj kada $X_0$ sadrži samo konstantni član, a ostale varijable su standardizirane. Generalizirani faktor inflacije varijance možemo računati kao

$$GVIF_l = \frac{detR_{11} \times detR_{22}}{detR}$$

gdje je R korelacijska matrica varijabli svih prediktora uključenih u $X_1$ i $X_2$, dok je $R_{ii}$ korelacijska matrica prediktora uključenih u skup varijabli predstavljenih matricom dizajna $X_i$, $i = 1, 2$. Nadalje, s ciljem usporedivosti vrijednosti pokazatelja GVIF s obzirom na različite podskupove varijabli predlaže se korištenje modificirane, osnosno skalirane verzije $GVIF^{(1/2p_1)}$. Napomenimo kako u slučaju $p_1 = 1$ vrijedi GVIF=VIF.

Više o ovom pokazatelju moguće je pronaći u (Fox  Monette, 1992). Tipičan primjer korištenja generaliziranog faktora inflacije varijance vezan je za slučaj kada su prediktori kategorijalne varijable. Naime, za jednu kategorijalnu varijablu s $k$ kategorija u regresijski model je uključeno $(k-1)$ dummy varijabli. Stoga je uz jednu kategorijalnu varijablu povezano $(k-1)$ parametara pa postojanje multikolinearnosti nije moguće ispitati primjenom faktora inflacije varijance. Dakle, kada je u modelu prisutna kategorijalna varijabla, problem multikolinearnosti moguće je detektirati primjenom generaliziranog faktora inflacije varijance i modificirane, odnosno skalirane verzije ovog parametra.

## 2.3 Kondicioni broj

Da bismo razumjeli pojam kondicionog broja, prvo ćemo definirati karakteristične korijene.
Neka je A simetrična kvadratna matrica reda n. Pretpostavimo da želimo minimizirati izraz $X^TAX$ uz uvjet $X^TX = 1$. Uvođenjem Lagrangeovog multiplikatora $\lambda$, minimiziramo izraz

$$X^TAX - \lambda(X^TX - 1).$$

Deriviramo li ovaj izraz po X i izjednačimo ga s nula, dobivamo

$$2AX - 2\lambda X = 0 \qquad ili \qquad (A - \lambda I)X = 0.$$

Da bi ove jednadžbe imale ne-nul rješenje, mora vrijediti

$$Rank(A - \lambda I) < n \qquad ili \qquad |A - \lambda I| = 0.$$

Korijeni ove determinističke jednadžbe, koju nazivamo karakteristična jednadžba, nazivaju se karakteristični korijeni matrice A (alternativni nazivi su latentni korijeni ili svojstvene vrijednosti).
Karakteristična jednadžba $|A-\lambda I| = 0$ je jednadžba n-tog stupnja u varijabli $\lambda$ te ima n korijena. Za svako rješenje $\lambda_l$ postoji odgovarajući vektor $x_l$ koji je rješenje jednadžbe $(A - \lambda_l I)X = 0$. Ti vektori se nazivaju karakteristični vektori (ili svojstveni vektori).

Kondicioni broj mjeri osjetljivost procjenitelja regresije na male promjene u podacima. Definira se kao kvadratni korijen omjera najveće i najmanje svojstvene vrijednosti matrice $X^TX$, tj.

$$CN = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

Što je kondicioni broj bliže 1, to u većoj mjeri možemo zaključiti kako multikolinearnost nije prisutna. Niti u slučaju primjene kondicionog broja kao pokazatelja postojanja multikolinearnosti ne postoji konsenzus oko granice vrijednosti pokazatelja koja upućuje na postojanje multikolinearnosti. Ponekad se kao granična vrijednost koja ukazuje na postojanje multikolinearnosti uzima $CN > 20$, dok neki autori sugeriraju da postoji umjerena do jaka multikolinearnost ako je CN između 100 i 1000, a ako je CN iznad 1000 postoji ozbiljan problem multikolinearnosti.
Napomenimo da je primjenom kondicionog broja moguće utvrditi postojanje multikolinearnosti u modelu, dok je primjenom faktora inflacije varijance moguće utvditi postojanje multikolinearnosti vezano uz individualni prediktor.

## 2.4 Kleinovi kriteriji

Korisno je poznavati Kleinove kriterije koji nam mogu pomoći pri utvrđivanju problema postojanja multikolinearnosti.
Prvi Kleinov kriterij kaže da postoji ozbiljan problem s multikolinearnosti ukoliko je barem jedan od koeficijenata korelacije nultog reda između regresorskih varijabli po apsolutnoj vrijednosti veći od koeficijenta višestruke linearne korelacije R.
Koeficijenti korelacije nultog reda su elementi korelacijske matrice

$$C = \begin{bmatrix} 1 & & & & \\ r_{1y} & 1 & & & \\ r_{2y} & r_{21} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ r_{ky} & r_{k1} & r_{k2} & \dots & 1 \end{bmatrix}, \tag{2.5}$$

$$r_{ij} = \frac{Cov(x_i,x_j)}{\sqrt{Var(x_i)}\sqrt{Var(x_j)}}$$

Ako postoji barem jedan $r_{ij}$ takav da je $r_{ij} > R$ tada postoji problem multikolinearnosti. Zapravo, najjednostavniji način provjere postojanja multikolinearnosti je upravo promatranje matrice C, odnosno koeficijenata korelacije među regresorskim varijablama. Koeficijenti korelacije koji su po apsolutnoj vrijednosti veći od 0.8 ukazuju na mogućnost postojanja problema multikolinearnosti.

Drugi Kleinov kriterij kaže da postoji ozbiljan problem multikolinearnosti ako je koeficijent determinacije dovoljno velik ($0.7 < R^2 < 0.9$), a istovremeno su empirijski t-omjeri mali. Naime, empirijski F-omjer za skupni test može se izraziti kao funkcija koeficijenata determinacije:

$$F = \frac{SP/k}{SR/n-(k+1)} = \frac{\frac{SP}{ST}}{\frac{SR}{ST}/[n-(k+1)]} = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}, \tag{2.6}$$

te bi zbog velike vrijednosti koeficijenata determinacije empirijski F-omjer bio značajan, što bi značilo da je barem jedna od regresorskih varijabli značajna u modelu. S druge strane, male vrijednosti empirijskih t-omjera upućuju na zaključak o neznačajnosti regresorskih varijabli, što je kontradiktorno prethodnom.

## 2.5 Statistički testovi

### 2.5.1 Farrar Glauberov test

Koristan test u ispitivanju postojanja multikolinearnosti je Farrar-Glauberov test. Test se provodi u tri stadija:

1. Ispitivanje postojanja multikolinearnosti primjenom $\chi^2$-testa (primjenjuje se Bartlettov test).

2. Ukoliko prvi korak ukaže na postojanje multikolinearnosti primjenom F-testa detektira se skup varijabli koje su korelirane.

3. Primjenom t-testa detektiraju se varijable koje uzrokuju multikolinearnost.

Pojasnimo pobliže prvi korak. U prvom koraku provjeravamo je li korelacijska matrica značajno različita od jedinične matrice, odnosno ispitujemo postojanje multikolinearnosti. Postavljamo hipotezu:
$H_0$: Nije prisutan problem multikolinearnosti.
$H_1$: Problem multikolinearnosti je prisutan.
Nultom se hipotezom pretpostavlja da ne postoji multikolinearnost , tj. da su regresorske varijable međusobno nezavisni vektori, odnosno da je korelacijska matrica jedinična matrica.
Empirijska test veličina pripada $\chi^2$-distribuciji s $k(k-1)/2$ stupnjeva slobode i glasi:

$$\chi^2 = -[n - 1 - \tfrac{1}{6}(2k + 5)]ln(detC),$$

pri čemu je $n$ broj opažanja, a $k$ broj regresorskih varijabli.
U prethodnom izrazu C je matrica koja sadrži koeficijente linearne korelacije parova regresorskih varijabli:

$$\mathrm{C} = \begin{bmatrix} 1 & & & & \\ r_{21} & 1 & & & \\ r_{31} & r_{32} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ r_{k1} & r_{k2} & r_{k3} & \dots & 1 \end{bmatrix}.$$

Matrica C dobiva se izostavljanjem prvog stupca i prvog retka korelacijske matrice C u izrazu (2.5).

Odluka se donosi na uobičajen način , tj. usporedbom vrijednosti test veličine i teorijske vrijednosti $\chi^2$-distribucije za zadanu razinu signifikantnosti.

U drugom koraku promatraju se koeficijenti višestruke korelacije i testiramo hipotezu $H_0 : R_i^2 = 0$. Primjenjuje se F-test (2.5). Odbacivanjem ove hipoteze možemo zaključiti da su varijable kolinearne. U trećem koraku promatraju se koeficijenti parcijalne korelacije među parovima varijabli i testira se hipoteza o nepostojanju korelacije između dva prediktora. Primjenjuje se t-test i odbacivanjem hipoteze detektiraju se prediktori odgovorni za postojanje multikolinearnosti.

### 2.5.2 Mtest

Mtest je neparametarski test koji daje statističku potporu dvjema najpoznatijima metodama za otkrivanje multikolinearnosti u primjeni: Kleinovo pravilo i faktor inflacije varijance (VIF).

Definirajmo regresijski model

$$\mathrm{y}_i = \beta_0 x_{0i} + \beta_1 x_{1i} + ... + \beta_p x_{pi} + u_i, \qquad i = 1, ..., n, \ (2.7)$$

$x_{ji}$ su prediktori za $j = 1, ..., p$, $x_0 = 1$ za svaki $i \in \{1, ..., n\}$ i $u_i$ je normalno distribuirana greška.

Primjer sporednog regresijskog modela pridruženog ovom modelu je

$$x_{2i} = \gamma_1 x_{1i} + \gamma_3 x_{3i} + ... + \gamma_p x_{pi} + u_i.$$

Općenito, postoji $p$ sporednih regresijskih modela gdje je jedan prediktor uzet kao zavisna varijabla, dok su ostali prediktori nezavisne varijable. S $R_g^2$ označimo koeficijent determinacije od (2.7), a s $R_j^2$ j-ti koeficijent determinacije za svaku regresiju.

Kleinovo pravilo uspoređuje $R_g^2$ s $R_j^2$. Pravilo sugerira da ako je $R_j^2 > R_g^2$, onda varijabla $X_j$ uzrokuje multikolinearnost.

S druge strane, VIF metoda računa $VIF_j = \frac{1}{1-R_j^2}$ i sugerira da je multikolinearnost generirana s $X_j$ ako je $VIF_j > 10$ (napomenimo ponovo da su predložene i druge granice).

Moguće je povezati obje metode izrazom

$$10 = \frac{1}{1-R_j^2}$$
$$R_j^2 = 0.9,$$

što znači da prema VIF metodi varijabla $X_j$ uzrokuje multikolinearnost ako je $R_j^2 \geq 0.9$. Ovo znači da ako je, na primjer, $R_g^2 = 0.85$ i $R_j^2 = 0.88$, za neki $j \in \{1, ..., n\}$, Kleinovo pravilo će detektirati multikolinearnost, no VIF neće. Važno je napomenuti da vrijednost 0.9 nije fiksna u svim primjenama. Ovdje prikazujemo slučaj kada je granična vrijednost 0.9, ali ta vrijednost može biti promijenjena.

Uz dani regresijski model Mtest se temelji na računanju procjena $R_g^2$ i $R_j^2$ iz $n$ bootstrap uzoraka dobivenih iz skupa podataka, $R_{gboot}^2$ i $R_{jboot}^2$ redom.

Stoga se detekcija multikolinearnosti temeljem pokazatelja VIF prevodi u statističku hipotezu

$$H_0 : \mu_{R^2_{j_{boot}}} \geq 0.9 \qquad i$$
$$H_1 : \mu_{R^2_{j_{boot}}} < 0.9$$

Tražimo postignutu razinu značajnosti (ASL)

$$ASL = Prob_{H_0}\{\mu_{R^2_{j_{boot}}} \geq 0.9\}$$

procijenjenu s

$$\hat{ASL}_{n_{boot}} = card\{\mu_{R^2_{j_{boot}}} \geq 0.9\}/n_{boot}.$$

Na sličan način, Kleinovo pravilo se prevodi u

$$H_0 : \mu_{R^2_{j_{boot}}} \geq \mu_{R^2_{g_{boot}}} \qquad i$$
$$H_1 : \mu_{R^2_{j_{boot}}} < \mu_{R^2_{g_{boot}}}.$$

Tražimo postignutu razinu značajnosti

$$ASL = Prob_{H_0}\{\mu_{R^2_{j_{boot}}} \geq \mu_{R^2_{g_{boot}}}\}$$

procijenjenu s

$$\hat{ASL}_{n_{boot}} = \{\mu_{R^2_{j_{boot}}} \geq \mu_{R^2_{g_{boot}}}\}/n_{boot}.$$

Treba napomenuti da nam ovaj raspis omogućuje formuliranje VIF-a i Kleinova pravila u smislu testiranja statističkih hipoteza.

# Poglavlje 3

# Neke metode uklanjanja multikolinearnosti

## 3.1 Analiza glavnih komponenti

Glavni cilj analize glavnih komponenti (eng. Principal component analysis (PCA)) je odrediti nekoliko komponenti koje objašnjavaju najveći mogući dio varijance mjerenih varijabli. Pri tome su komponente linearne kombinacije nezavisnih varijabli. PCA se koristi kako bi se smanjila dimenzionalnost podataka uz najmanji mogući gubitak informacija.
Pretpostavimo da su $x_1, x_2, ..., x_k$ eksploratorne varijable s matricom kovarijanci V. Sada možemo razmotriti linearne kombinacije nezavisnih varijabli:

$$z_1 = a_1x_1 + a_2x_2 + ... + a_kx_k$$
$$z_2 = b_1x_1 + b_2x_2 + ... + b_kx_k, \text{ itd.}$$

Želimo pronaći linearnu kombinaciju $\alpha^T X$ koja ima najveću varijancu uz uvjet $\alpha^T\alpha = 1$. Ovaj uvjet se naziva uvjet normalizacije. Bez ovog uvjeta varijancu od $z_1$ bismo mogli neograničeno povećavati. Ovaj problem se svodi

na rješavanje $|V - \lambda I| = 0$. Najveći karakteristični korijen od V je tražena najveća varijanca, a odgovarajući karakteristični vektor je traženi $\alpha$.
Proces maksimiziranja varijance linearne kombinacije $z$, uz uvjet da je suma kvadrata koeficijenata jednaka 1, daje $k$ rješenja. U skladu s njima konstruiramo $k$ linearnih kombinacija $z_1, z_2, ..., z_k$. One se nazivaju glavne komponente od $x_i, i = 1, ..., k$. Poredamo li ih na sljedeći način:

$$Var(z_1) > Var(z_2) > ... > Var(z_k).$$

Komponenta $z_1$, koja ima najveću varijancu, naziva se prva glavna komponenta, $z_2$ koja ima sljedeću najveću varijancu naziva se druga glavna komponenta itd.
Ove glavne komponente imaju sljedeća svojstva:

1. $Var(z_1) + Var(z_2) + ... + Var(z_k) = \lambda_1 + \lambda_2 + ... + \lambda_k = Tr(V) = Var(x_1) + Var(x_2) + ... + Var(x_k)$.

2. Za razliku od $x_1, ..., x_k$ koji su korelirani, $z_1, ..., z_k$ su ortogonalni ili nekorelirani. Stoga ne postoji multikolinearnost među varijablama $z_1, ..., z_k$.

Stoga je važnost komponenti označena svojstvenim vrijednostima, iz kojih se može izračunati postotak objašnjene varijabilnosti. Obično se dobivene glavne komponente ponovno skaliraju.
Razvijene su različite smjernice o tome koliko linearnih kombinacija k treba zadržati. Na primjer: kriterij latentnog korijena (zadržavaju se samo komponente koje imaju svojstvenu vrijednost veću od 1), apriorni kriterij, scree test i kriterij udjela varijance (Hair et al.,1995).

### 3.1. Analiza glavnih komponenti

Metoda glavnih komponenata može biti primijenjena i u cilju uklanjanja problema multikolinearnosti. Osnovna ideja je zamijeniti skup linearno zavisnih varijabli njihovim linearnim kombinacijama.

Ponekad se preporučuje da umjesto regresije y na $x_1, x_2, ..., x_k$ promatramo regresiju y na $z_1, z_2, ..., z_k$. Međutim, ovo nije dobro rješenje za problem multikolinearnosti. Činjenica da $z_1, z_2, ..., z_k$ nisu korelirani ne znači da ćemo dobiti bolje procjenitelje koeficijenata u početnoj jednadžbi regresije. Stoga ima smisla koristiti glavne komponente samo ako promatramo regresiju y na podskupu skupa $\{z_1, z_2, ..., z_k\}$. Ali i ovdje postoje neki problemi:

1. Prva glavna komponenta $z_1$, iako ima najveću varijancu, ne mora biti najviše korelirana s y. Dakle, ne mora nužno postojati povezanost između redoslijeda glavnih komponenti i stupnja korelacije sa zavisnom varijablom y.

2. Može se pomisliti kako je najbolje izabrati samo glavne komponente koje imaju visoku korelaciju s y, a ostale odbaciti, ali isti postupak se može koristiti s originalnim skupom varijabli $x_1, x_2, ..., x_k$ izabiranjem varijable s najvećom korelacijom s y, zatim varijable s najvećom parcijalnom korelacijom itd.

3. Linearna kombinacija varijabli $z_1, z_2, ..., z_k$ često nema interpretabilno značenje.

4. Mijenjanje mjernih jedinica od $x_1, x_2, ..., x_k$ će promijeniti glavne komponente. Ovaj problem se može izbjeći standardizirajući sve varijable.

## 3.2 Izostavljanje varijabli

U nekim slučajevima nas ne zanimaju svi parametri. U takvim slučajevima možemo dobiti procjenitelje za parametre koji nas zanimaju, a koji imaju manje srednje kvadratne pogreške od procjenitelja najmanjih kvadrata (OLS procjenitelja). To postižemo odbacivanjem pojedinih varijabli.
Pretpostavimo model

$$y = \beta_1 x_1 + \beta_2 x_2 + u \tag{3.1}$$

u kojem je problem visoka korelacija između $x_1$ i $x_2$. Pretpostavimo da je naš glavni interes $\beta_1$. Zatim izostavimo $x_2$ i procijenimo jednadžbu

$$y = \beta_1 x_1 + v. \tag{3.2}$$

Neka je procjenitelj od $\beta_1$ u modelu (3.1) označen s $\hat{\beta}_1$, a procjenitelj od $\beta_1$ u modelu (3.2) označen s $\beta_1^*$. $\hat{\beta}_1$ je OLS procjenitelj, a $\beta_1^*$ je procjenitelj parametra $\beta_1$ s izostavljenom varijablom $x_2$. Za OLS procjenitelj znamo da vrijedi

$$E(\hat{\beta}_1) = \beta_1 \qquad i \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{11}(1-r_{12}^2)}$$

Za procjenitelja parametra $\beta_1$ u modelu s izostavljenom varijablom (OV (omitted variable) procjenitelj) trebamo izračunati $E(\beta_1^*)$ i $Var(\beta_1^*)$. Sada supstitucijom početnog modela

$$\beta_1^* = \frac{\sum_{i=1}^n x_{i1} y_i}{\sum_{i=1}^n x_{i1}^2}$$

dobivamo

### 3.2. Izostavljanje varijabli

$$\beta_1^* = \frac{\sum_{i=1}^n x_{i1}(\beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum_{i=1}^n x_{i1}^2} = \beta_1 + \frac{S_{12}}{S_{11}}\beta_2 + \frac{\sum_{i=1}^n x_{i1} u_i}{S_{11}}$$

Stoga je

$$E(\beta_1^*) = \beta_1 + \beta_2 \frac{S_{12}}{S_{11}} \qquad i$$
$$Var(\beta_1^*) = Var(\frac{\sum_{i=1}^n x_{i1} u_i}{S_{11}}) = \frac{\sigma^2 S_{11}}{S_{11}^2} = \frac{\sigma^2}{S_{11}}$$

Stoga je $\beta_1^*$ pristran, ali ima manju varijancu od $\hat{\beta}_1$. Vrijedi

$$\frac{Var(\beta_1^*)}{Var(\hat{\beta}_1)} = 1 - r_{12}^2$$

i ako je $r_{12}$ vrlo visoka, onda će $Var(\beta_1^*)$ biti znatno manji od $Var(\hat{\beta}_1)$. Sada je

$$\frac{(\text{pristranost u } \beta_1^*)^2}{Var(\hat{\beta}_1)} = (\frac{\beta_2 S_{12}}{S_{11}})^2 \frac{S_{11}(1-r_{12}^2)}{\sigma^2} = \frac{S_{12}^2}{S_{11} S_{22}} \beta_2^2 \frac{S_{22}(1-r_{12}^2)}{\sigma^2}$$

Ovaj izraz možemo zapisati kao $r_{12}^2 t_2^2$, gdje je

$$t_2^2 = \frac{\beta_2^2}{Var(\hat{\beta}_2)}$$

$t_2$ predstavlja t-omjer za $x_2$ u početnom modelu (naglasimo kako je to stvarni, a ne procijenjeni t-omjer).
Kako je $MSE = pristranost^2 + varijanca$ i kako je za $\hat{\beta}_1$ MSE $= Var(\beta_1)$ imamo

$$\begin{aligned}\frac{MSE(\beta_1^*)}{MSE(\hat{\beta}_1)} &= \frac{(\text{pristranost u } \beta_1^*)^2}{Var(\hat{\beta}_1)} + \frac{Var(\beta_1^*)}{Var(\hat{\beta}_1)} \\ &= r_{12}^2 t_2^2 + (1 - r_{12}^2) \\ &= 1 + r_{12}^2(t_2^2 - 1)\end{aligned}$$

Stoga, ako je $|t_2| < 1$, onda je $MSE(\beta_1^*) < MSE(\hat{\beta}_1)$. Ako je $t_2$ nepoznat koristi se procijenjena t-vrijednost $\hat{t}_2$ iz početnog modela. Za procjenitelja od $\beta_1$ koristimo procjenitelj uvjetno izostavljene varijable (COV, conditional omitted variable), definiran na sljedeći način

$$\tilde{\beta}_1 = \begin{cases} \hat{\beta}_1 & \text{,ako je } |\hat{t_2}| \geq 1 \\ \beta_1^* & \text{,ako je } |\hat{t_2}| < 1 \end{cases}$$

Također, umjesto korištenja $\hat{\beta}_1$ ili $\beta_1^*$, ovisno o $\hat{t_2}$ možemo koristiti linearnu kombinaciju

$$\lambda\hat{\beta}_1 + (1-\lambda)\beta_1^*$$

Ovo se naziva ponderirani procjenitelj i ima minimalnu srednju kvadratnu pogrešku za $\lambda = \frac{t_2^2}{1+t_2^2}$. Ponovno, $t_2$ je nepoznat i moramo koristiti njegovu procijenjenu vrijednost $\hat{t_2}$.

# Poglavlje 4

# Primjer

U ovom poglavlju primijenit ćemo teorijsko znanje na stvarnom skupu podataka. Provest ćemo regresijsku analizu u programu R, gdje ćemo pomoću ugrađenih metoda provjeriti postoji li multikolinearnost u raznim modelima linearne regresije. Ukoliko multikolinearnost postoji, pokušat ćemo riješiti problem metodama navedenim u prethodnom poglavlju. Na taj način ćemo izgraditi model koji nam najviše odgovara.
Analizu provodimo na skupu podataka *Boston* koji se može pronaći u R paketu *MASS*. Ovaj skup podataka daje informacije o vrijednosti stanova u predgrađu Bostona. Varijabla koju želimo procijeniti je *medv*, odnosno srednja vrijednost naseljenih stanova (u 1000 USD). Koristimo sljedećih 11 kontinuiranih prediktora:

1. *crim* - stopa kriminala po glavi stanovnika
2. *zn* - udio stambenog zemljišta za parcele veće od 25000 kvadratnih stopa
3. *indus* - udio nemaloprodajnih poslovnih prostora u gradu
4. *nox* - koncentracija dušikovih oksida (na 10 milijuna)

5. *rm* - prosječni broj soba po stanu

6. *age* - udio naseljenih jedinica koje su izgrađene prije 1940.

7. *dis* - ponderirana srednja vrijednost udaljenosti do pet centara za zapošljavanje u Bostonu

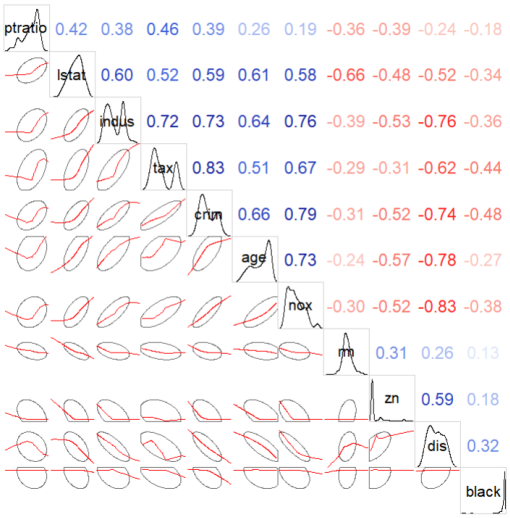
8. *tax* - puna vrijednost stope poreza na imovinu na 10000 USD

9. *ptratio* - omjer učenika i nastavnika po gradu

10. *black* - $1000(B_k - 0.63)^2$, gdje je $B_k$ udio crnaca po gradu

11. *lstat* - postotak stanovništva koji pripada nižem statusu

Cilj ovoga dijela nije pronalaženje najboljeg modela već primjer izgradnje linearnog modela uz prisustvo multikolinearnosti. Stoga nisu promatrani nelinearni modeli, nego su provedene transformacije varijabli koje nisu u linearnoj vezi s varijablom *medv*. Tako smo logaritmirali varijable *dis*, *lstat* i *crim*.




Slika 4.1: Prediktori grupirani prema korelaciji.

506 observacija smo podijelili na podatke za treniranje i podatke za testiranje u omjeru 70%:30%.

Možemo vidjeti na Slici 4.1 da među prediktorima postoji jako pozitivno povezana grupa, a u njoj se nalaze varijable *age*, *nox*, *indus* i *tax*. S druge strane, možemo uočiti da je prediktor *dis* u jakoj negativnoj korelaciji s prediktorima *age*, *nox* i *indus*.

Za početak smo procijenili univarijantne modele linearne regresije. Rezultate možemo vidjeti u Tablici 4.1.

| Zavisna varijabla | Procijenjene vrijednosti parametara | $R^2$ | Standardna greška | t-statistika | p-vrijednost |
|---|---|---|---|---|---|
| $log(dis)$ | 4.44 | 0.066 | 0.8612 | 5.156 | 4.07e-07 |
| $nox$ | -31.34 | 0.16 | 3.693 | -8.488 | 4.81e-16 |
| $indus$ | -0.6 | 0.21 | 0.05993 | -10.08 | <2.2e-16 |
| $tax$ | -0.02 | 0.2 | 0.002484 | -9.826 | <2.2e-16 |
| $age$ | -0.12 | 0.12 | 0.01623 | -7.257 | 2.261e-12 |
| $log(lstat)$ | -12.36 | 0.65 | 0.4708 | -26.25 | < 2.2e-16 |
| $zn$ | 0.12 | 0.1 | 0.01904 | 6.499 | 2.549e-10 |
| $rm$ | 8.91 | 0.45 | 0.5087 | 17.53 | < 2.2e-16 |
| $ptratio$ | -2.05 | 0.23 | 0.192 | -10.68 | < 2.2e-16 |
| $log(crim)$ | -1.73 | 0.17 | 0.1977 | -8.767 | 2.2e-16 |
| $black$ | 0.03 | 0.11 | 0.004662 | 7.060 | 7.98e-12 |

Tablica 4.1: Tablica procijenjenih vrijednosti u univarijantnim modelima.

| Prediktor | $\hat{\beta}$ | VIF | Standardna greška | t-statistika | p-vrijednost |
|---|---|---|---|---|---|
| $log(crim)$ | 0.53 | 5.518 | 0.247986 | 2.131 | 0.033717 |
| $zn$ | 0.003 | 2.031 | 0.013929 | 0.196 | 0.844598 |
| $indus$ | -0.05 | 3.539 | 0.061824 | -0.743 | 0.457707 |
| $nox$ | -20.23 | 4.6 | 4.207456 | -4.809 | 2.21e-06 |
| $rm$ | 2.56 | 1.943 | 0.464678 | 5.508 | 6.82e-08 |
| $age$ | 0.02 | 3.504 | 0.015791 | 1.122 | 0.262394 |
| $log(dis)$ | -5.93 | 4.78 | 0.948488 | -6.253 | 1.11e-09 |
| $tax$ | -0.004 | 4.455 | 0.002860 | -1.388 | 0.165826 |
| $ptratio$ | -0.83 | 1.719 | 0.139843 | -5.904 | 8.06e-09 |
| $black$ | 0.01 | 1.355 | 0.002811 | 3.628 | 0.000326 |
| $log(lstat)$ | -9.85 | 3.123 | 0.680116 | -14.486 | $<$ 2e-16 |

Tablica 4.2: Tablica procijenjenih vrijednosti u multivarijantnom modelu.

Zatim smo promotrili multivarijantni model linearne regresije sa svih navedenih 11 prediktora:

$medv = 60.99 + 0.53log(crim) + 0.003zn - 0.05indus - 20.23nox + 2.56rm +$
$0.02rm + 0.02age - 5.93log(dis) - 0.004tax - 0.83ptratio + 0.01black - 9.85log(lstat)$

Rezultati su sistematizirani u tablici 4.2.

Vrijednost korigiranog koeficijenta determinacije jednaka je 0.7622, što znači da je modelom objašnjeno 76.22% varijabilnosti zavisne varijable. Vrijednost F testne statistike iznosi 111.7, a pripadna p-vrijednost je manja od 2.2e-16, pa možemo zaključiti kako je model značajan. Model je primijenjen na podatke iz test seta te RMSE iznosi 3.817.

U Tablici 4.1. i 4.2. možemo primjetiti da prediktori *crim*, *age* i *dis* imaju koeficijente različitih predznaka u univarijantnom i multivarijantnom modelu. To bi moglo ukazivati na postojanje multikolinearnosti. Nadalje, primijetimo

kako se značajnost prediktora $zn$, $indus$ i $tax$ promijenila u multivarijantnom modelu.

Iz tablice možemo iščitati da prediktor $log(crim)$ ima najveću VIF vrijednost (5.518). Stoga ispitujemo postojanje multikolinearnosti. Razvijeno je više funkcija, odnosno paketa, koji mogu biti korisni za ispitivanje postojanja multikolinearnosti. U ovome radu služili smo se paketima: $Mtest$, $car$ i $mctest$. Kondicioni broj izračunat za slučaj modela sa uključenih 11 prediktora iznosi 89.117, što ukazuje na postojanje problema multikolinearnosti. Nadalje, proveden je Farrar Glauberov test te rezultati ($\chi^2 = 3071.93$) također sugeriraju kako je multikolinearnost prisutna.

Sada dolaze na red dvije mogućnosti korekcije: isključivanje varijabli iz modela i analiza glavnih komponenti.

1. Isključivanje varijabli iz modela.

   U ovome slučaju iz modela s uključenim kolinearnim varijablama kolinearnost otklanjamo na način da iz modela isključujemo prediktore koji uzrokuju kolinearnost. Odluku o tome koji prediktor isključiti iz modela istraživač može donijeti na temelju vlastite prosudbe (na primjer, teorijske smjernice, pouzdanost prediktora) ili na temelju statističkih performansi (na primjer, isključujemo prediktor koji nije bio značajan u univarijantnim modelima ili je imao najmanji postotak objašnjene varijabilnosti u univarijantnom modelu).

   Iz multivarijantnog modela s 11 prediktora isključujemo varijable $log(crim)$, $nox$, $indus$, $tax$ i $rm$ temeljem VIF kriterija. Pri tom isključujemo varijable koje su manje značajne. Također, iz modela isključujemo varijable $age$ i $zn$ zbog velike p-vrijednosti koja nam govori da navedene varijable nisu značajne u modelu.

Procijenjeni model je

$$medv = 68.66 - 4.13log(dis) - 0.82ptratio + 0.01black - 12.54log(lstat).$$

Dakle, promjenom varijable $log(lstat)$ za jednu jedinicu, promjena u varijabli $medv$ će biti $-12.54$ ako ostale varijable držimo fiksnima.
Isključivanjem navedenih varijabli nije došlo do velikog smanjenja korigiranog koeficijenta determinacije - ovim modelom objašnjeno je 71.9% varijabilnosti zavisne varijable. Model je primijenjen na podatke iz test seta te RMSE iznosi 4.34.
Najveća VIF vrijednost u ovom modelu je 1.61 što je zadovoljavajuće. Primjena različitih testova dostupnih u korištenim paketima sugerira kako više nije prisutan problem multikolinearnosti.

2. Analiza glavnih komponenti.
   Drugi način uklanjanja multikolinearnosti koji primjenjujemo je metoda glavnih komponenti. Primjenom metode glavnih komponenti formirali smo tri linearne kombinacije koje objašnjavaju 76% ukupne varijabilnosti prediktora. Te tri linearne kombinacije (u oznaci PC1, PC2 i PC3) dalje koristimo u izgradnji regresijskog modela. Glavne komponente su procijenjene primjenom paketa *psych*. Nakon izgradnje tri linearne kombinacije originalnih varijabli, iste koristimo za izgradnju regresijskog modela sa zavisnom varijablom $medv$. Procijenjeni model dan je s:

$$medv = 22.49 - 2.27PC1 - 3.71PC2 + 1.61PC3.$$

Koeficijent determinacije jednak je 0.666. Vrijednost F-statistike jednaka je 254 i $p$-vrijednost je manja od 2e-16. Dakle model je značajan.

Model je primijenjen na podatke iz test seta te RMSE iznosi 6.3. Problema multikolinearnosti nema, ali model više nema jasnu interpretaciju.

# Zaključak

U ovom radu smo se bavili problemom multikolinearnosti u višestrukoj linearnoj regresiji. Dana je teorijska podloga vezana uz linearnu regresiju te detekciju i rješavanje multikolinearnosti. Multikolinearnost uzrokuje velike standardne pogreške parametara te je moguće izračunati vrijednosti procijenjenih parametara s pogrešnim predznakom. Prikazali smo kriterije za detekciju problema: faktor inflacije varijance, generalizirani faktor inflacije varijance i kondicioni broj. Također, dali smo prikaz dvaju statističkih testova: Farrar Glauberov test i Mtest. Zatim smo prikazali metode kojima se uklanja problem, a to su analiza glavnih komponenti i izostavljanje varijabli. Napomenimo da je ovo tek dio metoda detekcije, te kako su razvijene različite metode uklanjanja (npr. ridge regresija). Na kraju je dan sveobuhvatni primjer kojim prikazujemo primjenu prikazane teorije na stvarnom skupu podataka. U programu R je provedena analiza skupa podataka *Boston* koji sadrži 506 opažanja. Ovaj skup podataka je dovoljno velik te nam omogućava podjelu na podatke za treniranje i podatke za testiranje modela. Tako su na podacima za treniranje izgrađeni različiti univarijantni i multivarijantni modeli linearne regresije, a na podacima za testiranje su uspoređene njihove performanse. Pomoću više testova utvrđeno je postojanje multikolinearnosti u modelu sa svim varijablama, a zatim smo pristupili problemu na dva načina: isključivanjem problematičnih varijabli i analizom glavnih kom-

ponenata. Isključivanje problematičnih varijabli provedeno je na način da se iz modela isključuje jedna po jedna varijabla s najvećom vrijednosti faktora inflacije varijance. Tako u nekoliko koraka dolazimo do modela koji nam daje zadovoljavajuće rezultate. Analiza glavnih komponenti nam daje nove varijable linearne regresije koje su linearna kombinacija originalnih varijabli i model bez problema multikolinearnosti, ali model ne daje jasnu interpretaciju.

U slučaju da nam je cilj dobiti model najveće prediktivne moći tada multikolinearnost nije problem, ali ako nas zanima odnos između zavisne i nezavisnih varijabli tada je nužno otkloniti multikolinearnost. Primjenom dodatnih informacija, u kombinaciji sa selektivnosti, možemo savladati problem multikolinearnosti i doći do željenog modela.

# Literatura


[1] G.S.Maddala, *Introduction to econometrics - second edition*, University of Florida and Ohio State University, 1992.

[2] John A. Rice, *Mathematical Statistics and Data Analysis Third Edition*, University of California, Berkeley, 2007.

[3] Alvin C. Rencher, G. Bruce Schaalje, *Linear models in statistics Second Edition*, Department of Statistics, Brigham Young University, Provo, Utah, 2008.

[4] Arthur S.Goldberger, *A Course in Econometrics*, Harvard University Press Cambridge, Massachusetts, London, England, 1991.

[5] John Fox, Georges Monette, *Generalized Collinearity Diagnostics*, Journal of the American Statistical Association , Mar., 1992

[6] https://cran.r-project.org/web/packages/MTest/MTest.pdf

[7] Morales-Oñate, Víctor, Morales-Oñate, Bolívar, *MTest: a bootstrap test for multicollinearity*, Banco Solidario, Universidad Técnica de Ambato, Universidad de las Américas, Escuela Superior Politécnica de Chimborazo, 2021.

**Literatura**

[8] Prased Perera, *The Boston Housing Dataset*, 2028., https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset

[9] Greene, W. H. (2003), Econometric Analysis , Pearson Education

[10] Morales Oñate, V., Morales-Oñate, B. (2023). MTest: una Prueba bootstrap para Multicolinealidad. Revista Politécnica, 51(2), 53–62. https://doi.org/10.33333/rp.vol51n2.05

[11] Bahovec, V.; Erjavec, N. (2009) Uvod u ekonometrijsku analizu. Zagreb: Element, 2009