Određivanje glavnih prediktora razine adiponektina, homocisteina, Cistacina C i ekskrecije albumina u urinu kod dijabetesa tipa 2 koristeći regresijsku analizu

Šeketa, Valentina

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Valentina Šeketa

ODREĐIVANJE GLAVNIH PREDIKTORA RAZINE ADIPONEKTINA, HOMOCISTEINA, CISTACINA C I EKSKRECIJE ALBUMINA U URINU KOD DIJABETESA TIPA 2 KORISTEĆI REGRESIJSKU ANALIZU

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, rujan, 2016

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Mojim roditeljima – mami Jadranki i tati Josipu

Sadržaj

Uvod

Regresijska analiza je jedna od raširenijih statističkih metoda. Pitamo se: Zašto je to tako? Naime, koristeći upravo rezultate regresijske analize možemo pomoću danog skupa podataka predvidjeti ponašanje varijable koja nas zanima. Dobiveni regresijski model može se tada primijeniti na novi ulazni skup vrijednosti te time dobiti procjena varijable koju promatramo. Uočavamo da takav pristup zaista ima široki raspon primjena. U ovom radu bazirat ćemo se na područje medicine, tj. nastojat ćemo predložiti pogodan model za četiri čimbenika koji se prate kod pacijenata koji boluju od dijabetesa tipa 2.

Činjenica da je veći dio regresijske analize ustaljen postupak, tj. postoji „kostur“ koji je osnova njezine provedbe, dodatno ohrabruje kao i pogled na same pretpostavke takvog modeliranja koje nisu toliko rigorozne. Nadalje, točnosti i brzini rezultata će dodatno pridonijeti računarska podloga. U ovom radu koristit ćemo statistički program SAS (University Edition) koji ima implemetiranu većinu potrebnih procedura, iako je zadatak moguće riješiti i pomoću drugih statističkih programa.

Da ne bi sve ostalo „visiti“ u zraku, također ćemo objasniti matematičku pozadinu potrebnih procedura prije nego što ih primijenimo na zadani problem. Pritom ćemo upozoriti na neke moguće ishode prilikom modeliranja te naći prikladna rješenja.

Iako će SAS znatno ubrzati cijeli postupak, modeliranje tu ne staje. Prilikom interpretacije dobivenih rezultata i odabira pogodnog modela od presudnog će značaja biti upravo statističko razmišljanje i savjetovanje sa strukom kako bismo odabrali model koji je „najbolji“. U svrhu lakše interpretacije rezultata ispisanih u SAS-u, navodimo prijevod korištenog nazivlja.

Nazivlje

Analysis of Variance (ANOVA) – analiza varijance
Corrected Total – ukupno ispravljeno
Degrees of Freedom (DF) – stupnjevi slobode
Distribution – distribucija; **Normal Distribution** – normalna distribucija
Error – greška; **Standard Error** – standardna greška
Frequency (Freq) – Frekvencija; **Cumulative Frequency** – kumulativna frekvencija
Goodness-of-Fit Tests – Ocjena prilagodbe modela
Intercept – slobodni član
Kurtosis – koeficijent spljoštenosti
Label – oznaka **Mean** – očekivanje
Mean square – Varijanca
Median – medijan
Minimum – minimum; **Maximum** – maksimum
Number of Observation Read – broj učitanih observacija
Number of Observation Used – broj korištenih observacija
Number of Observation with Missing Values – broj observacija sa vrijednostima koje nedostaju
Parameter Estimate – procjena parametara
Pearson Correlation Coefficients – Pearsonov koeficijent korelacije
Percent – postotak; **Cumulative Percent** – kumulativni postotak
Probability (Pr) – vjerojatnost
Proportion Less – proporcija manjih vrijednosti
Quantile – kvantil
R-square – R_2 ; **Partial R-square** – parcijalni R_2; **Adjusted R-square** – prilagođeni R_2
Regression (REG) – regresija
Residual – rezidual
Skewness – koeficijent asimetrije
Source – izvor
Standard Deviation (Std Dev) – standardna devijacija
Statistic – statistika
Step – korak
Sum of Squares – suma kvadrata; **Type II SS** – Suma kvadrata tipa 2
Summary of Stepwise Selection – sažetak stepwise odabira
Univariate – jednostruka
Variable (Var) – varijabla; **Variable Entered** – varijabla koja ulazi; **Variable Removed** – varijabla koja izlazi; **Dependent Variable** – zavisna varijabla
Value – vrijednost; **Predicted Value** ... predviđena vrijednost

Poglavlje 1

Linearna regresija

Regresija kao znanstvena metoda pojavila se prvi put oko 1885. godine, iako je metoda najmanjih kvadrata otkrivena 80-ak godina prije. Ubrzo regresija postaje jedan od najraširenijih metoda u statistici što je rezultiralo njenim daljnjim proučavanjem i razvijanjem. Tako postoje linearna i nelinearna regresija, te parametrijska i neparametrijska regresija. Ovisno o broju varijabli koje ulaze i izlaze iz regresije u ovom radu ćemo koristiti jednostruku (jedan ulazna i jedna izlazna varijabla) i višestruku (više ulaznih i jedna izlazna varijabla) linearnu regresiju.

Koristeći regresiju možemo predvidjeti ili objasniti ponašanje neprekidne varijable pomoću nekoliko drugih varijabli, istovremeno ih modelirajući u odnosu na zavisnu varijablu. Prilikom tog postupka važno je odabrati podskup nezavisnih varijabli iz velikog skupa potencijalnih nezavisnih varijabli kako bismo pronašli prihvatljiv model. U tome će nam pomoći statističko zaključivanje i savjetovanje sa strukom. Na kraju preostaje procijeniti dobiveni model.

U ovom radu proučavat ćemo kako odabrani skup varijabli utječe na razinu adiponektina, homocisteina, cistacina C i ekskrecije albumina u urinu kod pacijenata koji boluju od dijabetesa tipa 2. Prilikom istraživanja opažene su vrijednosti više varijabli za svakog pacijenta te nas zanima istovremeni utjecaj podskupa tih varijabli na varijablu koju proučavamo. Budući da su potrebne procjene parametara vrlo često zahtjevne, koristit ćemo statistički program SAS kako bismo ubrzali cijeli postupak.

Poglavlje 2

Jednostruka linearna regresija

Ispitivanje i analiza ovisnosti jedne (zavisne) varijable o jednoj (nezavisnoj) varijabli naziva se jednostruka regresija. Ako je veza između zavisne i nezavisne varijable linearna, tada je riječ o jednostrukoj linearnoj regresiji. Rezultat svake regresijske analize je regresijski model.

2.1 Jednostruki linearni regresijski model

U slučaju jednostavne linearne regresije model je dan jednadžbom pravca koji opisuje vezu između zavisne i nezavisne varijable. Dakle, jednostavni regresijski model je oblika

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

pri čemu su ε slučajne greške za koje vrijedi da je $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ i da su nekorelirane, tj. da vrijednost jedne greške ne ovisi o vrijednosti drugih grešaka. Slobodni član β_0 i koeficijent smjera β_1 zovu se regresijski koeficijenti i oni su nepoznati.

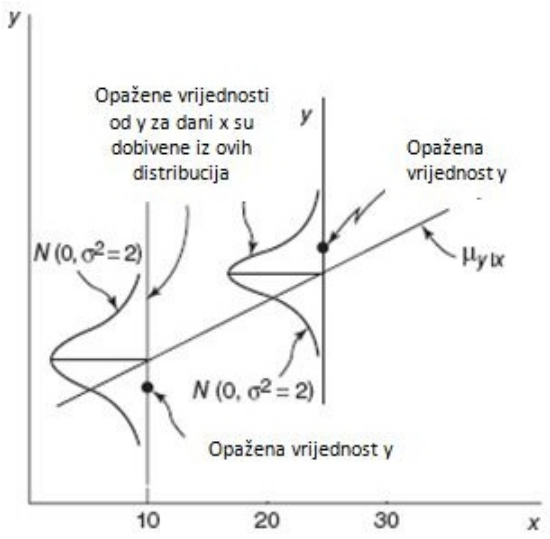
Prilikom prikupljanja podataka nezavisna varijabla x je kontrolirana varijabla. Vrijednost u toj varijabli mjeri se sa zanemarivom slučajnom greškom ε. Tada je y zavisna slučajna varijabla, tj. za svaki x postoji vjerojatnosna distribucija od y. Njezino očekivanje i varijanca su jednaki

$$E(y|x) = \beta_0 + \beta_1 x \tag{2.2a}$$

$$Var(y|x) = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2 \tag{2.2b}$$

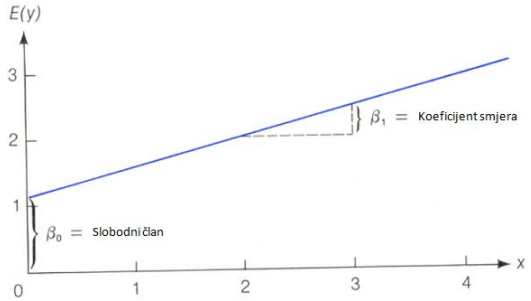
Uočimo da je očekivanje od y linearna funkcija od x. Dakle, vrijednost pravca za neki x predstavlja očekivanje od y za taj x. Također, varijanca od y ne ovisi o x i određena je greškom modela σ^2. Budući da su greške nekorelirane, vrijednosti zavisne varijable su također nekorelirane.

Distribucija podataka u jednostrukoj linearnoj regresiji prikazana je na slici 2.1. Budući da u primjeni gotovo nikad neće biti moguće povući pravac kroz sve točke, ε će objašnjavati odudaranje točaka od pravca.



Slika 2.1: Observacije u linearnoj regresiji (preuzeto iz [1])

Regresijski koeficijenti β_0 i β_1 imaju jednostavnu i vrlo korisnu interpretaciju, kao što je prikazano na slici 2.2. Ako se vrijednost od x poveća za jednu jediničnu veličinu, tada je koeficijent smjera β_1 promjena očekivanja distribucije od y. Ako je $x = 0$ unutar raspona podataka, tada je za $x = 0$ slobodni član β_0 očekivanje distribucije zavisne varijable y.



Slika 2.2: Interpretacija regresijskih koeficijenata (preuzeto iz [3])

2.2 Procjena parametara metodom najmanjih kvadrata

Parametri β_0 i β_1 su nepoznati, ali ih možemo procijeniti koristeći dobivene podatke. Pretpostavimo da imamo n parova podataka, recimo$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Ti podaci mogu biti rezultat dizajniranog kontroliranog eksperimenta, opažačke studije ili postojećih povijesnih bilješki (retrospektivna studija).

2.2.1 Procjena β_0 i β_1

Da bismo procijenili β_0 i β_1 koristimo metodu najmanjih kvadrata. Dakle, procjenjujemo β_0 i β_1 tako da suma kvadrata razlika između observacija y i željenog pravca bude minimalna. Koristeći jednakost (2.1) možemo zapisati

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, ..., n. \tag{2.3}$$

Relaciju (2.1) zovemo populacijski regresijski model, dok relaciju (2.3) zovemo uzorački regresijski model za dani skup podataka (x_i, y_i) $(i = 1, \ldots, n)$. Funkcija najmanjih kvadrata definirana je sa

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.4}$$

Procjenitelji β_0 i β_1 dobiveni metodom najmanjih kvadrata, recimo $\hat{\beta}_0$ i $\hat{\beta}_1$, moraju zadovoljavati sljedeće uvjete:

$$\frac{\partial S}{\partial \beta_0} \mid_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0$$

$$\frac{\partial S}{\partial \beta_1} \mid_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Pojednostavimo li ove dvije relacije dobivamo sustav normalnih jednadžbi:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned} \tag{2.5}$$

Rješavanjem dobivenog sustava dobivamo procjenitelje $\hat{\beta}_0$ i $\hat{\beta}_1$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.6}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \tag{2.7}$$

gdje su $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ i $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ redom aritmetičke sredine od y_i i x_i.

Fitani jednostavni regresijski model je sada oblika

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.8}$$

te za određeni x možemo izračunati procjenu očekivanja od y.

Nazivnik u jednakosti (2.7) je korigirana suma kvadrata od x_i, a brojnik je korigirana suma skalarnih produkata od x_i i y_i pa ih možemo zapisati na kompaktniji način kao

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2.9}$$

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x}) \tag{2.10}$$

Dakle, relaciju (2.7) možemo zapisati u obliku

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{2.11}$$

Razlika između opaženih vrijednosti y_i i odgovarajućih fitanih vrijednosti $\hat{y}_i$ je i-ti rezidual. Definiramo ga na sljedeći način

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right), \quad i = 1, ..., n \tag{2.12}$$

Reziduali igraju važnu ulogu u ispitivanju pretpostavki i adekvatnosti modela.

2.2.2 Svojstva procjenitelja dobivenih metodom najmanjih kvadrata

Procjenitelji $\hat{\beta}_0$ i $\hat{\beta}_1$ dobiveni metodom najmanjih kvadrata imaju nekoliko važnih svojstava. U relacijama (2.6) i (2.7) vidimo da su $\hat{\beta}_0$ i $\hat{\beta}_1$ linearne kombinacije observacija y_i, tj.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i$$

gdje je $c_i = (x_i - \bar{x}) / S_{xx}$ za $i = 1, ..., n$.

Procjenitelji $\hat{\beta}_0$ i $\hat{\beta}_1$ dobiveni metodom najmanjih kvadrata su nepristrani procjenitelji parametara $\hat{\beta}_0$ i $\hat{\beta}_1$. Za $\hat{\beta}_1$ vrijedi:

$$E\left(\hat{\beta}_1\right) = E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E\left(y_i\right) = \sum_{i=1}^{n} c_i \left(\beta_0 + \beta_1 x_i\right) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

jer je pretpostavci $E(\varepsilon_i) = 0$. Direktno se vidi da je $\sum_{i=1}^{n} c_i = 0$ i $\sum_{i=1}^{n} c_i x_i = 1$ pa je

$$E\left(\hat{\beta}_1\right) = \beta_1.$$

Analogno se pokaže da je $\hat{\beta}_0$ nepristrani procjenitelj za β_0, tj. da vrijedi

$$E\left(\hat{\beta}_0\right) = \beta_0.$$

Varijanca od $\hat{\beta}_1$ jednaka je

$$Var\left(\hat{\beta}_1\right) = Var\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 Var\left(y_i\right) \tag{2.13}$$

jer su observacije y_i nekorelirane pa je varijanca sume zapravo jednaka sumi varijanci. Iskoristimo sada pretpostavku da je $Var\left(y_i\right) = \sigma^2$:

$$Var\left(\hat{\beta}_1\right) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2 \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}} \tag{2.14}$$

Varijanca od $\hat{\beta}_0$ jednaka je

$$Var\left(\hat{\beta}_0\right) = Var\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) = Var\left(\bar{y}\right) + \bar{x}^2 Var\left(\hat{\beta}_1\right) - 2\bar{x} Cov\left(\bar{y}, \hat{\beta}_1\right)$$

Sada je varijanca od $\bar{y}$ jednaka $Var\left(\bar{y}\right) = \sigma^2 / n$, a može se pokazati da je kovarijanca između $\bar{y}$ i $\hat{\beta}_1$ jednaka 0. Dakle,

$$Var\left(\hat{\beta}_0\right) = Var\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) = Var\left(\bar{y}\right) + \bar{x}^2 Var\left(\hat{\beta}_1\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_x x}\right) \tag{2.15}$$

Još jedan važan rezultat vezan za procjenitelje $\hat{\beta}_0$ i $\hat{\beta}_1$ dobivenih metodom najmanjih kvadrata je Gauss-Markovljev teorem koji kaže da su za regresijski model (2.1) sa pretpostavkama $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ i nekoreliranim greškama, procjenitelji dobiveni metodom najmanjih kvadrata nepristrani procjenitelji i imaju minimalnu varijancu u usporedbi sa svim ostalim nepristranim procjeniteljima koji su linearna kombinacija od y_i. Često kažemo da su procjenitelji dobiveni metodom najmanjih kvadrata najbolji linearni nepristrani procjenitelji, gdje se „najbolji“ odnosi na minimalnu varijancu.

Navedimo još nekoliko korisnih svojstava procjene modela metodom najmanjih kvadrata:

1. Suma reziduala u bilo kojem regresijskom modelu koji sadrže slobodni član β_0 je uvijek 0, tj.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0.$$

Ovo svojstvo direktno slijedi iz prve normalne jednakosti (2.5). Zaokruživanje grešaka može utjecati na rezultat.

2. Suma opaženih vrijednosti y_i je jednaka sumi procjenjenih vrijednosti $\hat{y}_i$, tj.

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

3. Regresijski pravac dobiven metodom najmanjih kvadrata uvijek prolazi kroz centroid [točka $(\bar{x}, \bar{y})$] podataka.

4. Suma reziduala pondiranih sa odgovarajućim vrijednostima nezavisnih varijabli je uvijek jednaka nula, tj.

$$\sum_{i=1}^{n} x_i e_i = 0$$

5. Suma reziduala pondiranih sa odgovarajućim procjenjenim vrijednostima je uvijek jednaka 0, tj.

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

2.3 Testiranje hipoteza za slobodni član i koeficijent smjera

Za testiranje hipoteza i konstruiranje pouzdanih inervala potrebna je dodatna pretpostavka da su greške ε_i u modelu normalno distribuirane. Dakle, sve zajedno mora vrijediti da su greške normalno nezavisno distribuirane sa očekivanjem 0 i varijancom σ^2 ili kraće $\varepsilon_i \overset{nez}{\sim} N(0, \sigma^2)$. Kasnije ćemo vidjeti kako možemo provjeriti ove pretpostavke pomoću analize reziduala.

2.3.1 Uporaba t testa

Pretpostavimo da želimo testirati hipotezu da je koeficijent smjera jednak konstanti β_{10}. Testiramo sljedeće hipoteze:

$$\begin{aligned} H_0 &: \beta_1 = \beta_{10} \\ H_1 &: \beta_1 \neq \beta_{10} \end{aligned} \tag{2.16}$$

gdje je alternativna hipoteza dvostrana. Budući da su greške $\varepsilon_i \overset{nez}{\sim} N(0, \sigma^2)$, za observacije vrijedi $y_i \overset{nez}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$. Parametar $\hat{\beta}_1$ je linearna kombinacija observacija pa je $\hat{\beta}_1$ normalno distribuirana sa očekivanjem $\hat{\beta}_1$ i varijancom σ^2/S_{xx} (odjeljak 2.2.2). Ako vrijedi nulta hipoteza $H_0 : \beta_1 = \beta_{10}$, onda statistika

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2/S_{xx}}}$$

slijedi $N(0, 1)$ distribuciju. Nažalost, σ^2 je obično nepoznata. Može se pokazati da je očekivana vrijednost od SS_{Res} jednaka $E(SS_{Res}) = (n-2)\sigma^2$. Dakle, nepristrani procjenitelj od σ^2 je

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res} \tag{2.17}$$

Mjeru MS_{Res} zovemo srednje kvadratno odstupanje. Dakle, ako vrijedi nulta hipoteza $H_0 : \beta_1 = \beta_{10}$, onda statistika

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} \tag{2.18}$$

slijedi t_{n-2} distribuciju. Odbacujemo nultu hipotezu ako za gornji $\alpha/2$ percentil t_{n-2} distribucije $t_{\alpha/2,n-2}$ vrijedi

$$|t_0| > t_{\alpha/2,n-2} \tag{2.19}$$

Nazivnik testne stastike t_0 u relaciji (2.18) se često zove standardna greška koeficijenta smjera, tj.

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{MS_{Res}}{S_{xx}}} \tag{2.20}$$

Slična procedura se može koristiti za testiranje hipoteze za slobodni član. Neka je β_{00} pretpostavljena konstanta. Da bismo testirali

$$\begin{aligned} H_0 &: \beta_0 = \beta_{00} \\ H_1 &: \beta_0 \neq \beta_{00} \end{aligned} \tag{2.21}$$

koristimo test statistiku

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0 - \beta_{00}}{se\left(\hat{\beta}_0\right)} \tag{2.22}$$

gdje je

$$se\left(\hat{\beta}_1\right) = MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \tag{2.23}$$

standardna greška slobodnog člana. Odbacujemo nultu hipotezu $H_0 : \beta_0 = \beta_{00}$ ako je $|t_0| > t_{\frac{\alpha}{2},n-2}$.

2.3.2 Testiranje značajnosti regresije

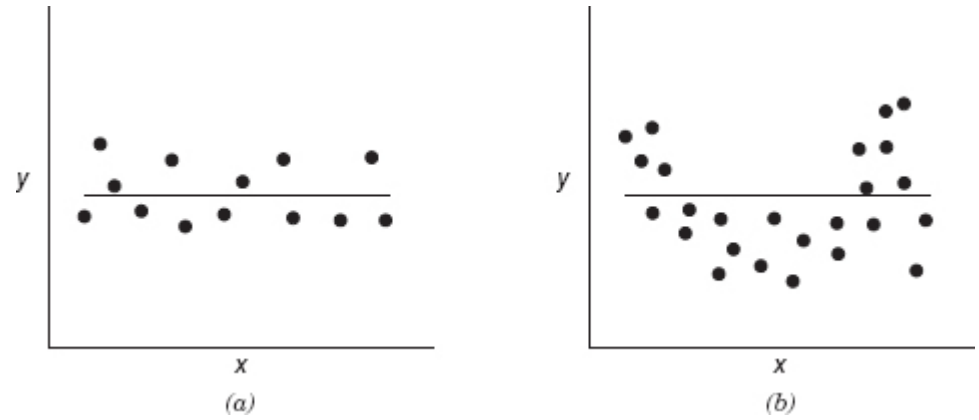
Vrlo važan poseban slučaj hipoteza u relaciji (2.16) je testiranje sljedećih hipoteza:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned} \tag{2.24}$$

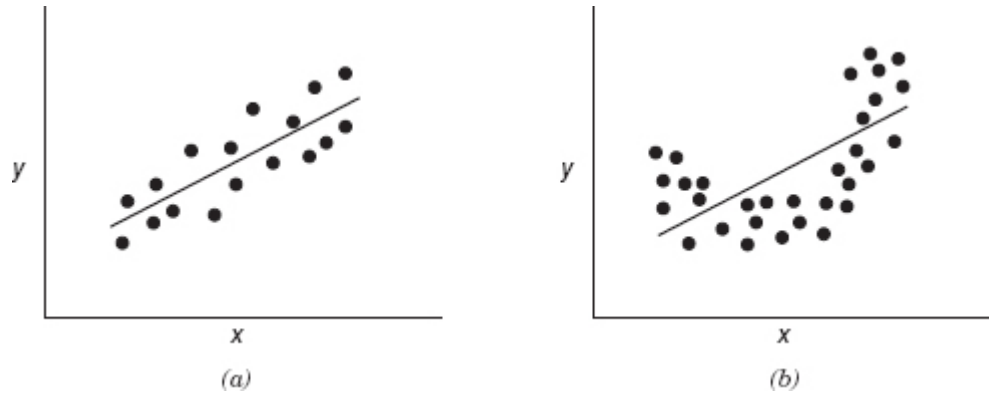
Ove hipoteze se odnose na značajnost regresije. Ako ne možemo odbaciti $H_0 : \beta_1 = 0$, to povlači da ne postoji linearna veza između x i y. Ovo je prikazano na slici 2.3. Uočimo da ovo može značiti da x ne objašnjava značajno varijaciju u y i da je najbolji procjenitelj od y za bilo koji x zapravo $\hat{y} = \bar{y}$ (slika 2.3*a*) ili da prava veza između x i y nije linearna (slika 2.3*b*). Dakle, ako ne možemo odbaciti $H_0 : \beta_1 = 0$, to je isto kao da kažemo da ne postoji linearna veza između y i x.
Alternativno, ako možemo odbaciti $H_0 : \beta_1 = 0$, to povlači da x značajno objašnjava varijabilnost od y. To je prikazano na slici 2.4. Ipak, odbacivanje $H_0 : \beta_1 = 0$ može značiti ili da je model opisan pravcem adekvatan (slika 2.4*a*) ili da iako postoji linearni efekt u x, bolji rezultati bi se mogli dobiti dodavanjem polinomijalnog člana višeg reda u x (slika 2.4*b*). Testna statistika za testiranje $H_0 : \beta_1 = 0$ može se razviti iz dva pristupa. Prvi pristup je jednostavno korištenje t statistike u relaciji (2.17) za $\beta_{10} = 0$, tj.

$$t_0 = \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} \tag{2.25}$$

Odbacujemo nultu hipotezu o značajnosti regresije ako je $|t_0| > t_{\frac{\alpha}{2},n-2}$.



Slika 2.3: Slučajevi kada ne možemo odbaciti hipotezu $H_0 : \beta_1 = 0$ (preuzeto iz [1])



Slika 2.4: Slučajevi kada možemo odbaciti hipotezu $H_0 : \beta_1 = 0$ (preuzeto iz [1])

2.3.3 Analiza varijance (ANOVA)

Drugi pristup testiranja značajnosti regresije je analiza varijance. Osnova analize varijance je particija ukupne varijabilnosti zavisne varijable y. Da bismo dobili tu particiju, počnimo s identitetom

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \tag{2.26}$$

Kvadriramo li obje strane jednakosti (2.26) i sumiramo po svih n observacija dobivamo

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

Uočimo da treći član na desnoj strani ovog izraza možemo zapisati kao

$$2 \sum_{i=1}^{n} (\hat{y}_i - \bar{y}) (y_i - \hat{y}_i) = 2 \sum_{i=1}^{n} \hat{y}_i (y_i - \hat{y}_i) - 2\bar{y} \sum_{i=1}^{n} (y_i - \hat{y}_i) = 2 \sum_{i=1}^{n} \hat{y}_i e_i - 2\bar{y} \sum_{i=1}^{n} e_i = 0$$

budući da je suma reziduala uvijek nula i da je suma reziduala pondiranih odgovarajućim fitanim vrijednostima $\hat{y}_i$ također nula. Dakle,

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.27}$$

Lijevu stranu jednakosti (2.27) označimo sa SS_T. To je korigirana suma kvadrata observacija SS_T koja predstavlja ukupnu varijabilnost observacija. Sumu kvadrata regresije $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ s desne strane označimo sa SS_R. Ona mjeri mjere količinu varijabilnosti observacija y_i objašnjenih regresijskim pravcem. Neka je $SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ suma kvadrata grešaka. Ona mjeri varijancu reziduala koja je ostala neobjašnjena dobivenim pravcem. S uvedenim oznakama dobivamo sljedeći osnovni identitet analize varijance regresijskog modela:

$$SS_T = SS_R + SS_{Res} \tag{2.28}$$

Ukupna suma kvadrata SS_T ima $df_T = n - 1$ stupnjeva slobode jer je jedan stupanj slobode izgubljen kao rezultat ograničenja $\sum_{i=1}^{n} (y_i - \bar{y})$. Suma kvadrata regresije SS_R ima $df_R = 1$ stupanj slobode jer je SS_R potpuno određena jednim parametrom, $\hat{\beta}_1$. Konačno, prethodno smo označili da SS_{Res} ima $df_{Res} = n - 2$ stupnjeva slobode jer su uvedena dva ograničenja na devijacije $y_i - \hat{y}_i$ kao rezultat procjene $\hat{\beta}_0$ i $\hat{\beta}_1$. Uočimo da stupnjevi slobode imaju svojstvo aditivnosti:

$$\begin{aligned} df_T &= df_R + df_{Res} \\ n - 1 &= 1 + (n - 2) \end{aligned} \tag{2.29}$$

Za testiranje hipoteze $H_0 : \beta_1 = 0$ možemo koristiti uobičajen F test analize varijance. Statistika

$$F_0 = \frac{SS_R/df_R}{SS_{Res}/df_{Res}} = \frac{SS_R/1}{SS_{Res}/(n-2)} = \frac{MS_R}{MS_{Res}} \tag{2.30}$$

slijedi $F_{1,n-2}$ distribuciju. Da bismo testirali hipotezu $H_0 : \beta_1 = 0$, računamo testnu statistiku F_0 i odbacujemo H_0 ako

$$F_0 > F_{\alpha,1,n-2}$$

Procedura testa je sažeta u ANOVA tablici:

Tablica 2.1: Analiza varijance za testiranje značajnosti regresije

Izvor varijabilnosti	Suma kvadrata	Stupnjevi slobode	Varijanca	F
Regresija	SS_R	1	MS_R	MS_R/MS_{Res}
Rezidual	SS_{Res}	$n-2$	MS_{Res}	
Ukupno	SS_T	$n-1$		

2.3.4 Više o t testu

Spomenuli smo da t statistiku

$$t_0 = \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1}{\sqrt{MS_{Res}/S_{xx}}} \tag{2.31}$$

možemo koristiti za testiranje značajnosti regresije. Ako kvadriramo obje strane, dobivamo

$$t_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_{Res}} = \frac{\hat{\beta}_1 S_{xy}}{MS_{Res}} = \frac{MS_R}{MS_{Res}} \tag{2.32}$$

Dakle, t_0^2 iz relacije (2.32) je identičnaF_0 za analizu varijance iz relacije (2.30). Općenito, kvadrat slučajne varijable t sa n stupnjeva slobode je slučajna varijabla F sa 1 stupnjem slobode u brojniku i n stupnjeva slobode u nazivniku. Iako je t test za $H_0 : \beta_1 = 0$ ekvivalentan F testu u slučaju jednostruke linearne regresije, t test je prilagodljivi jer ga možemo koristiti i za testiranje jednostranih alternativnih hipoteza ($H_1 : \beta_1 < 0$ ili $H_1 : \beta_1 > 0$). Korisnost analize varijance doći će do izražaja u višestrukim regresijskim modelima.

2.4 Pouzdani intervali

Pretpostavka o normalnosti iz poglavlja 2.3 i dalje vrijedi. Osim procjene parametara β_0, β_1 možemo također dobiti i procjenu njihovih pouzdanih intervala. Širina pouzdanih intervala govori o kvaliteti dobivenog regresijskog pravca.

2.4.1 Pouzdani intervali za β_0, β_1

Ako su greške normalno i nezavisno distribuirane, tada $\left(\hat{\beta}_1 - \beta_1\right) / se\left(\hat{\beta}_1\right)$ i $\left(\hat{\beta}_0 - \beta_0\right) / se\left(\hat{\beta}_0\right)$ slijede t distribuciju s $n-2$ stupnjeva slobode. Prema tome, $100(1-\alpha)$ postotni pouzdani interval (p.i.) koeficijenta smjera β_1 je dan sa

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2} se\left(\hat{\beta}_1\right) \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2} se\left(\hat{\beta}_1\right)$$

a $100(1-\alpha)$ postotni pouzdani interval slobodnog člana β_0 je dan sa

$$\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2} se\left(\hat{\beta}_0\right) \leq \beta_0 \leq \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2} se\left(\hat{\beta}_0\right)$$

Ovi intervali pouzdanosti imaju uobičajenu frekvencijsku interpretaciju. Ako uzmemo ponovljene uzorke iste veličine za iste vrijednosti od x i konstruiramo, npr. 95 posto pouzdani interval koeficijenta smjera β_1 za svaki uzorak, tada će 95 posto tih intervala sadržavati pravu vrijednost β_1.

2.4.2 Pouzdani intervali za očekivanje zavisne varijable

Glavna primjena regresijskog modela je procjena očekivanja zavisne varijable $E(y)$ za određenu vrijednost nezavisne varijable x. Neka je x_0 bilo koja vrijednost nezavisne varijable koja se nalazi unutar raspona originalnih podataka za koji želimo procijeniti očekivanje zavisne varijable, recimo $E(y \mid x_0)$. Nepristrani procjenitelj od $E(y \mid x_0)$ fitanog modela je

$$\widehat{E(y \mid x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Uočimo da je $\hat{\mu}_{y|x_0}$ normalno distribuirana slučajna varijabla jer je linearna kombinacija observacija y_i. Varijanca od $\hat{\mu}_{y|x_0}$ jednaka je

$$Var\left(\hat{\mu}_{y|x_0}\right) = Var\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) = Var\left[\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})\right] = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

budući da je $Cov(\bar{y}, \hat{\beta}_1)$=0.

Prema tome, uzoračka distribucija od

$$\frac{\hat{\mu}_{y|x_0} - E(y \mid x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}}$$

je t distribucija sa $n-2$ stupnjeva slobode. $100(1-\alpha)$ postotni interval pouzdanosti za očekivanje zavisne varijable u točki $x = x_0$ je

$$\hat{\mu}_{y|x_0} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \leq E(y \mid x_0) \leq \hat{\mu}_{y|x_0} + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Širina intervala pouzdanosti za $E(y \mid x_0)$ je funkcija od x_0. Minimalna je za $x_0 = \bar{x}$ i povećava se kako $|x_0 - \bar{x}|$ raste. Ovo je logično jer očekujemo da preciznost procjene opada kako se pomičemo prema rubovima prostora x.

2.5 Predviđanje novih observacija

Važna primjena regresijskog modela je predikcija novih observacija y u ovisnosti o x. Ako je x_0 vrijednost nezavisne varijable koja nas zanima, tada je procjena nove vrijednosti zavisne varijable y_0:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{2.33}$$

Pouzdani intervali očekivanja zavisne varijable za $x = x_0$ su nisu prikladni za procjenu intervala buduće observacije y_0 jer je to procjena intervala očekivanja od y (parametra), a ne vjerojatnost budućih observacija iz te distribucije. Uočimo da je slučajna varijabla

$$\Psi = y_0 - \hat{y}_0 \tag{2.34}$$

normalno distribuirana sa očekivanjem nula i varijancom

$$Var(\Psi) = Var(y_0 - \hat{y}_0) = \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right] \tag{2.35}$$

jer su buduće observacije y_0 nezavisne od $\hat{y}_0$. Ako koristimo $\hat{y}_0$ za predviđanje y_0, tada je prikladna statistika na kojoj temeljimo konstrukciju predikcijskog intervala standardna greška $\Psi = y_0 - \hat{y}_0$. Prema tome, $100(1-\alpha)$ postotni interval pouzdanosti za buduće observacije za x_0 je

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \leq y_0 \leq \hat{y}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \tag{2.36}$$

Predikcijski interval (2.36) je najuži za $x_0 = \bar{x}$ i širi se kako se $|x_0 - \bar{x}|$ povećava. Uočavamo da je predikcijski interval u x_0 (2.36) uvijek širi nego pouzdani interval u x_0 (2.43) jer predikcijski interval ovisi i o greškama fitanog modela i o greškama budućih observacija.

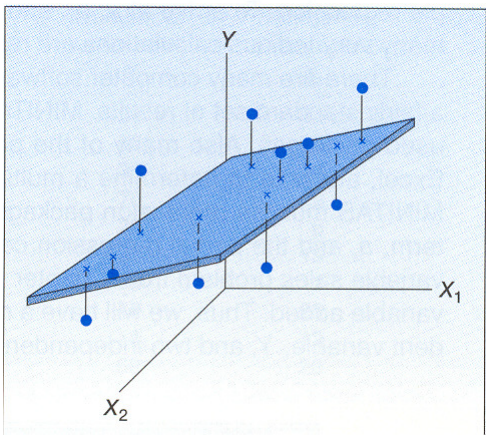
Poglavlje 3

Višestruka linearna regresija

Linearni regresijski model koji se sastoji od jedne zavisne varijable i više od jedne nezavisnih varijabli zove se višestruki regresijski model. Rezultati ovog poglavlja su poopćenja rezultata iz jednostruke linearne regresije iz prethodnog poglavlja.

3.1 Višestruki regresijski model

Općenito, zavisna varijabla y može biti povezana sa k nezavisnih varijabli. Njihova veza opisana je jednadžbom hiperravnine u k-dimenzionalnom prostoru nezavisnih varijabli. Na slici 3.1 zavisna varijabla ovisi o dvjema nezavisnim varijablama.



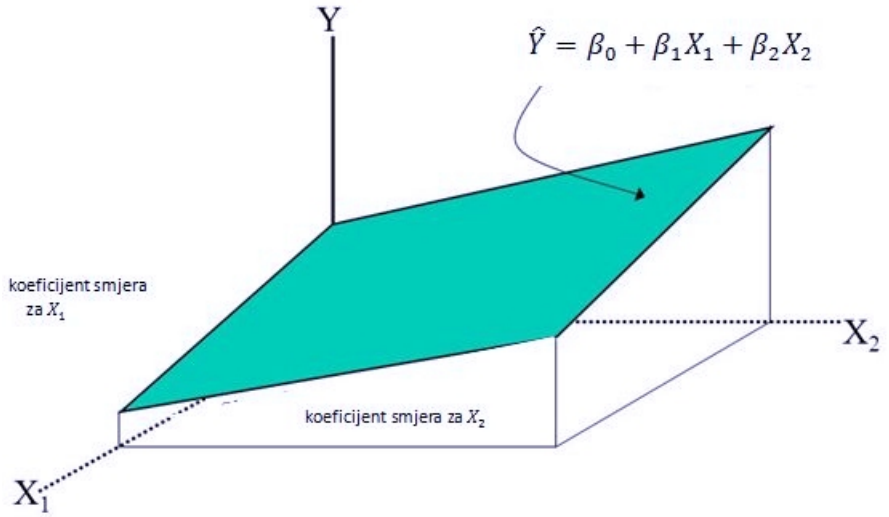
Slika 3.1: Višestruka linearna regresija za dvije nezavisne varijable (preuzeto iz [3])

Dakle, višestruki linearni regresijski model sa k nezavisnih varijabli je oblika

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon. \tag{3.1}$$

Parametri β_j, $j = 0, 1, \ldots k$ se zovu regresijski koeficijenti i oni su nepoznati.

Interpretacija regresijskih koeficijenata slična je interpretaciji u jednostrukoj linearnoj regresiji i prikazana je na slici 3.2. Ako raspon vrijednosti podataka uključuje i slučaj kada su svi $x_j = 0$, tada je slobodni član β_0 očekivanje y. Inače je β_0 vrijednost gdje ravnina presijeca y. Svaki β_j,$j = 1, \ldots, k$ predstavlja nagib u odnosu na x_j, tj. vrijednost promjene očekivanja od y kada je x_j veći za jednu jediničnu veličinu, a sve ostale nezavisne varijable x_i ($i \neq j$) su konstantne. Zboga toga se β_j još zovu i parcijalni regresijski koeficijenti.



Slika 3.2: Interpretacija regresijskih koeficijenata (preuzeto sa web stranice `http://slideplayer.com/slide/5317437/`)

Pomoću višestruke linearne regresije mogu se također analizirati i modeli kompleksnije strukture (npr. polinomi višeg stupnja i interakcije). Općenito, bilo koji regresijski model koji ima linearne parametre β je linearni regresijski model. Višestruki linearni modeli se često koriste kao empirijski modeli ili za aproksimaciju funkcija. Dakle, prava veza između y i $x_1, x_2, \ldots, x_k$ je nepoznata, ali za određen raspon vrijednosti nezavisnih varijabli dobiveni model će aproksimirati pravu nepoznatu funkciju.

3.2 Procjena parametara metodom najmanjih kvadrata

U primjeni vrijednosti parametara (regresijskih koeficijenata β_i) su nepoznati, ali ih možemo procijeniti iz podataka. Budući da su podaci u našem slučaju prikupljeni prilikom dizajniranog eksperimentalnog (kliničkog) istraživanja, pretpostavljamo da su regresori fiksne (matematičke, neslučajne) varijable bez grešaka pri mjerenju. Pretpostavimo da je dano $n > k$ podataka. Neka y_i predstavlja $i-$tu opaženu vrijednost zavisne varijable te x_{ij} $i-$tu opaženu vrijednost nezavisne varijable x_j. Podaci su prikazani u tablici 3.2.

Tablica 3.1: Podaci za višestruku linearnu regresiju (preuzeto iz [1])

		Nezavisne varijable			
Observacija, i	Zavisna varijabla, y	x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Pretpostavimo da za slučajne greške ε vrijedi $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ i da su one nekorelirane. Za testiranje hipoteza i konstrukciju pouzdanih intervala je potrebna dodatna pretpostavka da je uvjetna distribucija y uz dane x_1,x_2,...,x_k normalna sa očekivanjem $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ i varijancom σ^2.

3.2.1 Procjena regresijskih koeficijenata metodom najmanjih kvadrata

Regresijske koeficijente β_0,β_1,...,β_k modela (3.1) možemo procijeniti metodom najmanjih kvadrata. Ideja je pronaći aproksimacijsku funkciju minimizirajući reziduale, tj. razliku između opaženih i procijenjenih vrijednosti. Koristeći model (3.1) regresijski model možemo zapisati na sljedeći način:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

$$y_i = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{3.2}$$

Definiramo funkciju najmanjih kvadrata

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{k} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{i=1}^{k} \beta_j x_{ij} \right)^2 \tag{3.3}$$

Minimizirajući funkciju S po svakoj komponenti dobit ćemo procjenitelje $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ za $\beta_0, \beta_1, \ldots, \beta_k$.

$$\frac{\partial S}{\partial \beta_0} |_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \tag{3.4a}$$

$$\frac{\partial S}{\partial \beta_j} |_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \quad j = 1, 2, \ldots, k \tag{3.4b}$$

Dobivamo sljedeći sustav normalnih jednadžbi:

$$\begin{gathered} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1} y_i \\ \vdots \\ \hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik} y_i \end{gathered} \tag{3.5}$$

Uočimo da sustav ima $p = k + 1$ normalnih jednadžbi što je jednako broju regresijskih parametara.

Kompaktniji matrični zapis dobivenog sustava glasi

$$y = x\beta + \varepsilon,$$

gdje su

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Problem minimizacije je ekvivalentan sljedećem problemu:

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon = (y - X\beta)' (y - X\beta)$$

Budući da je $(\beta' X' y)' = y' X\beta$ jer je $\beta' X' y$ skalar, tj. 1×1 matrica dobivamo

$$S(\beta) = y'y - \beta' X' y - y' X\beta + \beta' X' X\beta = y'y - 2\beta' X' y + \beta' X' X\beta$$

Minimiziramo funkciju S kako bismo dobili procjenitelje metode najmanjih kvadrata.

$$\frac{\partial S}{\partial \beta} \mid_{\hat{\beta}} = -2X' y + 2X' X\hat{\beta} = 0$$

Pojednostavimo li gornji izraz dobivamo normalne jednadžbe analogne jednadžbama (3.5)

$$X' X\hat{\beta} = X' y \tag{3.6}$$

Pomnožimo obje strane sa $(X'X)^{-1}$. Inverz matrice $X'X$ postoji ako su nezavisne varijable linearno nezavisne, tj. nijedan stupac matrice X nije linearna kombinacija ostalih stupaca. Dobivamo sljedeći procjenitelj za regresijske koeficijente $\beta_0, \beta_1, \ldots, \beta_k$:

$$\hat{\beta} = (X'X)^{-1} X' y \tag{3.7}$$

Uočavamo da je matrični zapis normalnih jednadžbi (3.6) identičan skalarnom zapisu (3.5). Raspisujući matrični izraz dobivamo

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_k \end{bmatrix}$$

te njegovim množenjem dolazimo do skalarnog oblika. Uočimo da je $X'X$ $p \times p$ simetrična matrica. Dijagonalni elementi predstavljaju sumu kvadrata stupca matrice X, a elementi izvan dijagonale predstavljaju sumu skalarnog produkta elemenata stupaca matrice X. $X'y$ je $p \times 1$ vektor stupac koji se sastoji od suma skalarnih produkata stupaca matrice X i observacija y_i.

Procijenjeni regresijski model za nezavisne varijable $x' = [1, x_1, x_2, \ldots, x_k]$ glasi

$$\hat{y} = x'\hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j$$

Odgovarajuće procijenjene vrijednosti $\hat{y}_i$ u odnosu na opažene vrijednosti y_i iznose

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1} X' y = Hy \tag{3.8}$$

Razlika između opaženih vrijednosti y_i i odgovarajućih procijenjenih vrijednosti $\hat{y}_i$ zove se rezidual $e_i = y_i - \hat{y}_i$. Dobivenih n reziduala možemo matrično zapisati kao

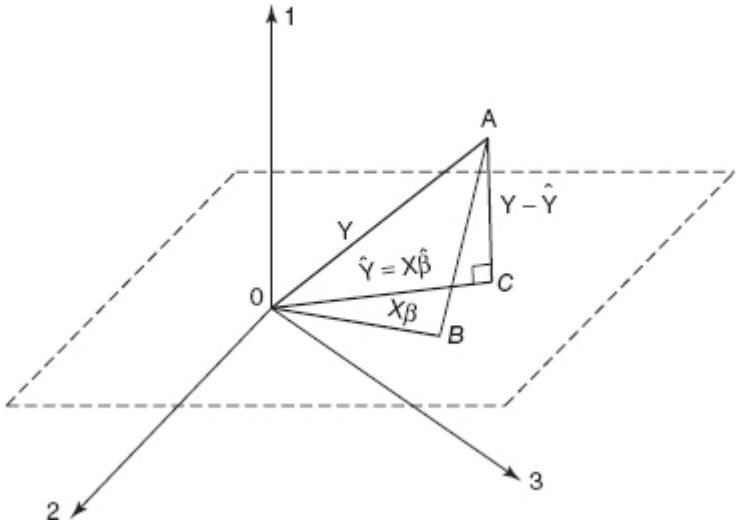
$$e = y - \hat{y}. \tag{3.9}$$

Možemo ih prikazati i kao

$$e = y - X\hat{\beta} = y - Hy = (I - H)y. \tag{3.10}$$

3.2.2 Geometrijska interpretacija metode najmanjih kvadrata

Svaku točku možemo interpretirati kao vektor tako da ju spojimo sa ishodištem prostora u kojem se nalazi. Stoga predočimo vektor observacija $y = [y_1, y_2, \ldots, y_n]$ kao vektor koji spaja ishodište n dimenzionalnog koordinatnog sustava sa točkom A. Na slici 3.3 je primjer 3-dimenzionalnog prostora.



Slika 3.3: Grafički prikaz višestruke linearne regresije (preuzeto iz [1])

Matrica X se sastoji od p $(n \times 1)$ vektora stupaca, tj. 1 (vektor stupac jedinica) te $x_1, x_2, \ldots, x_k$. Svaki od ovih stupaca definira vektor iz ishodišta. Ovih p vektora formiraju p-dimenzionalni potprostor koji zovemo prostor procjene (na slici 3.3 je $p = 2$). Bilo koju točku u ovom potprostoru možemo prikazati kao linearnu kombinaciju vektora $1, x_1, x_2, \ldots, x_k$. Dakle, bilo koja točka u ovom prostoru procjene je oblika $X\beta$. Neka je točka B na slici 3.3 određena sa $X\beta$. Kvadrirana udaljenost od točke B do točke A jest

$$S(\beta) = (y - X\beta)'(y - X\beta)$$

Dakle, da bismo minimizirali kvadratnu udaljenost vektora observacija y označenog točkom A do prostora procjene moramo naći točku u prostoru procjene koja je najbliža točki A. Kvadratna udaljenost će biti minimalna ako je točka u prostoru procjene nožište linije iz A normalne (okomite) na prostor procjene. To je točka C na slici 3.3. Ova točka je definirana vektorom $\hat{y} = X\hat{\beta}$. Budući da je $y - \hat{y} = y - X\hat{\beta}$ okomita na prostor procjene, možemo zapisati

$$X'\left(y - X\hat{\beta}\right) = 0 \quad ili \quad X'X = \hat{\beta}X'y$$

koje prepoznajemo kao normalne jednadžbe metode najmanjih kvadrata.

3.2.3 Svojstva procjenitelja dobivenih metodom najmanjih kvadrata

Statistička svojstva procjenitelja metode najmanjih kvadrata $\hat{\beta}$ mogu se lagano dokazati. Provjerimo prvo pristranost:

$$\begin{aligned} E(\hat{\beta}) &= E[(X'X)^{-1}X'y] \\ &= E[(X'X)^{-1}X'(X\beta + \varepsilon)] \\ &= E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon] \\ &= \beta \end{aligned}$$

jer je $E(\varepsilon) = 0$ i $(X'X)^{-1}X'X = I$. Dakle, $\hat{\beta}$ je nepristrani procjenitelj za β.

Svojstva varijance $\hat{\beta}$ su izražena preko kovarijacijske matrice

$$Cov\left(\hat{\beta}\right) = E\left\{\left[\hat{\beta} - E\left(\hat{\beta}\right)\right]\left[\hat{\beta} - E\left(\hat{\beta}\right)\right]'\right\}$$

što je $p \times p$ simetrična matrica čiji je j-ti dijagonalni element varijanca od $\hat{\beta}_j$ i čiji je ij-ti element izvan dijagonale kovarijanca između $\hat{\beta}_i$ i $\hat{\beta}_j$. Kovarijacijska matrica $\hat{\beta}$ se dobije primjenom operatora varijance na $\hat{\beta}$:

$$Cov\left(\hat{\beta}\right) = Var\left(\hat{\beta}\right) = Var\left[\left(X'X\right)^{-1}X'y\right].$$

Sada je $(X'X)^{-1}X'$ matrica konstanti i varijanca od y je $\sigma^2 I$ pa je

$$\begin{aligned} Var\left(\hat{\beta}\right) &= Var\left[\left(X'X\right)^{-1}X'y\right] \\ &= \left(X'X\right)^{-1}X'Var\left(y\right)\left[\left(X'X\right)^{-1}X'\right]' \\ &= \sigma^2\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1} \\ &= \sigma^2\left(X'X\right)^{-1} \end{aligned}$$

Ako stavimo $C = (X'X)^{-1}$, varijanca $\hat{\beta}_j$ je $\sigma^2 C_{jj}$ i kovarijanca između $\hat{\beta}_i$ i $\hat{\beta}_j$ je $\sigma^2 C_{ij}$.

Gauss Markovljev teorem tvrdi da je procjenitelj $\hat{\beta}$ metode najmanjih kvadrata najbolji linearni nepristrani procjenitelj za β. Ako dodatno pretpostavimo da su pogreške ε_i normalno distribuirane, tada je $\hat{\beta}$ također procjenitelj maksimalne vjerodostojnosti za β. Procjenitelj dobiven metodom maksimalne vjerodostojnosti je nepristrani procjenitelj za β minimalne varijance.

3.3 Testiranje hipoteza višestruke linearne regresije

Nakon što smo procijenili parametre modela, zanima nas kolika je adekvatnost modela i koje su nezavisne varijable važne. Test koji ćemo koristiti kao odgovor na ova pitanja je generalizacija analize varijance koju smo koristili u jednostavnoj linearnoj regresiji. I dalje vrijedi prepostavka da su slučajne greške nezavisne i normalno distribuirane sa očekivanjem $E(\varepsilon_i) = 0$ i varijancom $Var(\varepsilon_i) = \sigma^2$.

3.3.1 Testiranje značajnosti višestruke linearne regresije

Test značajnosti regresije je globalni test kojim testiramo postoji li linearna veza između zavisne varijable y i bilo koje nezavisne varijable $x_1, x_2, \ldots, x_k$. Odgovarajuće hipoteze su

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0 \; za \; bar \; jedan \; j$$

Odbacivanje nulte hipoteze povlači da barem jedna od nezavisnih varijabli $x_1, x_2, \ldots, x_k$ značajno doprinosi modelu. Ukupna suma kvadrata SS_T je particionirana u sumu kvadrata regresije SS_R i sumu kvadrata reziduala SS_{Res}. Dakle, $SS_T = SS_R + SS_{Res}$ Ako je nulta hipoteza istinita, tada statistika

$$F_0 = \frac{\frac{SS_R}{k}}{\frac{SS_{Res}}{n-k-1}} = \frac{MS_R}{MS_{Res}}$$

slijedi $F_{k,n-k-1}$ distribuciju. Dakle, da bismo testirali hipotezu $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, računamo testnu statistiku F_0 i odbacujemo H_0 ako je

$$F_0 > F_{\alpha,k,n-k-1}$$

Procedura testa obično je sažeta u ANOVA tablici:

Tablica 3.2: Analiza varijance za testiranje značajnosti regresije

Izvor varijabilnosti	Suma kvadrata	Stupnjevi slobode	Varijanca	F
Regresija	SS_R	k	MS_R	MS_R/MS_{Res}
Rezidual	SS_{Res}	$n-k-1$	MS_{Res}	
Ukupno	SS_T	$n-1$		

3.4 Pouzdani intervali u višestrukoj regresiji

3.4.1 Intervali pouzdanosti za koeficijente regresije

Da bismo konstruirali procjenu pouzdanih intervala regresijskih koeficijenata β_j i dalje ćemo pretpostavljati da su greške ε_i nezavisno i normalno distribuirane sa očekivanjem 0 i varijancom σ^2. Dakle, observacije y_i su nezavisne i normalno distribuirane s očekivanjem $\beta_0 + \sum_{j=1}^{k} \beta_j x_i j$ i varijancom σ^2. Budući da je procjenitelj metodom najmanjih kvadrata $\hat{\beta}$ linearna kombinacija observacija, slijedi da je $\hat{\beta}$ normalno distribuiran s vektorom očekivanja β i kovarijacijskom matricom $\sigma^2(X'X)^{-1}$. To povlači da je marginalna distribucija bilo kojeg regresijskog koeficijenta $\hat{\beta}_j$ normalna s očekivanjem β_j i varijancom $\sigma^2 C_{jj}$, gdje je C_{jj} j-ti dijagonalni element matrice $(X'X)^{-1}$. Posljedica toga je da svaka od statistika

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, \quad j = 0, 1, \ldots, k \tag{3.11}$$

slijedi t distribuciju sa $n - p$ stupnjeva slobode, gdje je $\hat{\sigma}^2 = MS_{Res}$. Pomoću relacije (3.11), možemo definirati $100(1 - \alpha)$ postotni interval pouzdanosti za regresijske koeficijente β_j, $j = 0, 1, \ldots, k$ kao

$$\hat{\beta}_j - t_{\frac{\alpha}{2}, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\frac{\alpha}{2}, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \tag{3.12}$$

Mjera

$$se\left(\hat{\beta}_j\right) = \sqrt{\hat{\sigma}^2 C_{jj}} \tag{3.13}$$

se zove standardna greška regresijskog koeficijenta $\hat{\beta}_j$.

3.4.2 Procjena pouzdanih intervala za očekivanje zavisne varijable

Možemo konstruirati pouzdani interval za očekivanje zavisne varijable u određenoj točki, kao što su $x_{01}, x_{02}, \ldots, x_{0k}$. Definiramo vektor x_0 kao

$$x_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}.$$

Fitana vrijednost u ovoj točki je

$$\hat{y}_0 = x_0' \hat{\beta} \tag{3.14}$$

Ovo je nepristrani procjenitelj od $E(y \mid x_0)$ jer je $E(\hat{y}_0) = x_0' \hat{\beta} = E(y \mid x_0)$, i varijanca od $\hat{y}_0$ je

$$Var(\hat{y}_0) = \sigma^2 x_0' (X'X)^{-1} x_0 \tag{3.15}$$

Prema tome, $100(1-\alpha)$ postotni interval pouzdanosti za očekivanje zavisne varijable u točki $x_{01}, x_{02}, \ldots, x_{0k}$ je

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-p}\sqrt{\sigma^2 x_0' (X'X)^{-1} x_0} \leq E(y \mid x_0) \leq \hat{y}_0 + t_{\frac{\alpha}{2},n-p}\sqrt{\sigma^2 x_0' (X'X)^{-1} x_0} \tag{3.16}$$

3.4.3 Istovremeni pouzdani intervali za regresijske koeficijente

Raspravljali smo o procedurama za konstrukciju pouzdanih i predikcijskih intervala, ali to su bili odvojeni intervali, tj. uobičajen tip pouzdanih ili predikcijskih intervala gdje koeficijent pouzdanosti $1-\alpha$ predstavlja proporciju točnih tvrdnji koje dobijemo kada uzmemo ponovljeni slučajni uzorak i konstruiramo odgovarajuću procjenu intervala koristeći isti uzorak podataka. Sada nas zanimaju pouzdani ili predikcijski intervali koji su istovremeno istiniti sa vjerojatnošću $1-\alpha$.

Promotrimo primjer modela jednostavne linearne regresije. Pretpostavimo da analitičar želi izvesti zaključke o slobodnom članu β_0 i koeficijentu smjera β_1. Jedna mogućnost bila bi (recimo) konstruirati 95 postotne pouzdane intervale za svaki parametar. Međutim, ako su ove procjene intervala nezavisne, vjerojatnost da su obje izjave točne je $(0.95)^2 = 0.9025$. Prema tome, nismo dobili 95 postotni interval pouzdanosti istovremeno za oba zahtjeva. Nadalje, budući da konstruirani intervali koriste isti skup uzoračkih podataka, oni nisu nezavisni. Ovo uvodi novu komplikaciju u određivanju nivoa pouzdanosti za istovremeni skup zahtjeva.

Udruženo pouzdano područje za parametre β modela višestruke regresije je relativno lako definirati. Možemo pokazati da je

$$\frac{\left(\hat{\beta}-\beta\right)' X'X \left(\hat{\beta}-\beta\right)}{pMS_{Res}} \sim F_{p,n-p}$$

i to povlači

$$P\left\{\frac{\left(\hat{\beta}-\beta\right)' X'X \left(\hat{\beta}-\beta\right)}{pMS_{Res}} \leq F_{\alpha,n-p}\right\}$$

Prema tome, $(1-\alpha)$ postotni udruženi interval pouzdanosti za sve parametre u vektoru β je

$$\frac{\left(\hat{\beta}-\beta\right)' X'X \left(\hat{\beta}-\beta\right)}{pMS_{Res}} \leq F_{\alpha,p,n-p} \tag{3.17}$$

Ova nejednakost definira područje u obliku elipse. Konstrukcija ovog udruženog područja je relativno izravna za jednostavnu linearnu regresiju (p=2).

3.5 Predviđanje novih observacija

Regresijski model možemo koristiti za predviđanje novih observacija y za određenu vrijednost regresora, npr. $x_{01}, x_{02}, \ldots, x_{0k}$. Ako je $x_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$, tada je procjena nove observacije y_0 u točki $x_{01}, x_{02}, \ldots, x_{0k}$ jednaka

$$\hat{y}_0 = x_0' \hat{\beta} \tag{3.18}$$

Prema tome, $100(1-\alpha)$ postotni predikcijski interval nove observacije je

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-p} \sqrt{\sigma^2 \left(1 + x_0' \left(X'X\right)^{-1} x_0\right)} \leq E\left(y \mid x_0\right) \leq \hat{y}_0 + t_{\frac{\alpha}{2},n-p} \sqrt{\sigma^2 \left(1 + x_0' \left(X'X\right)^{-1} x_0\right)} \tag{3.19}$$

Ovo je poopćenje predikcijskog intervala novih observacija u jednostrukoj linearnoj regresiji (2.45).

Poglavlje 4

Testiranje normalnosti

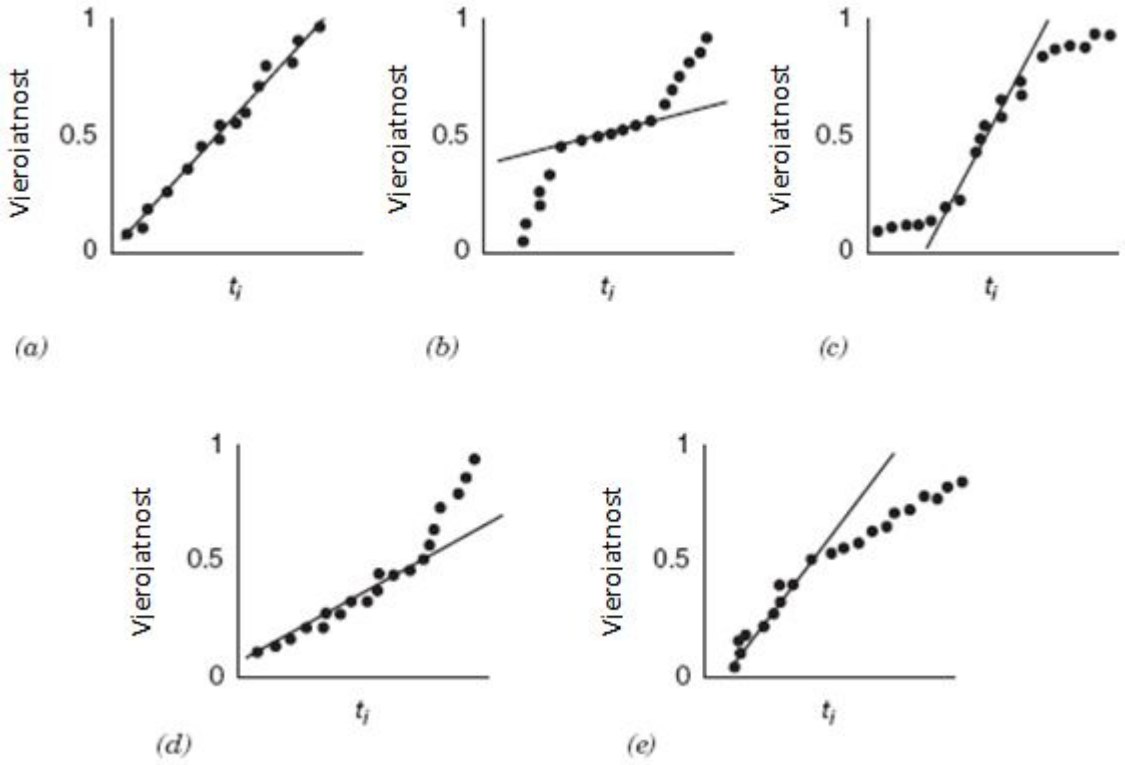
Testiramo nultu hipotezu o pripadnosti normalnoj distribuciji. Pri tom ćemo navesti i definirati grafičke testove i ocjene prilagodbe modela koje koristimo. Grafički testovi nam pomažu da intuitivno provjerimo ima li smisla pretpostaviti da uzorak dolazi iz normalne distribucije. Ocjena prilagodbe modela daje konačnu odluku. Pretpostavka normalnosti je razumna jer statistike t i F te pouzdani i predikcijski intervali ovise o njoj. Ako greške dolaze iz distribucija s lakšim ili težim repom od normalne, fit metodom najmanjih kvadrata bi mogao biti osjetljiv na male skupove podataka. Distribucije sa težim repom često generiraju outliere koji „vuku“ fit najmanjih kvadrata prema sebi.

4.1 Grafički test

U tu kategoriju spadaju histogram i normalni vjerojatnosni graf. Ovdje ćemo pobliže objasniti normalni vjerojatnosni graf. Pravcem je prikazana kumulativna normalnu distribuciju. Neka su $t_{[1]} < t_{[2]} < \ldots < t_{[n]}$ R-student reziduali rangirani u rastućem poretku. Ako nacrtamo $t_{[i]}$ nasuprot kumulativnim vjerojatnostima, $P_i = (i–1/2)/n, i = 1, 2, \ldots, n$, na normalnom vjerojatnosnom grafu, dobivene točke bi trebale približno ležati na pravcu. Znatna odstupanja od pravca ukazuju na to da distribucija nije normalna.

Normalni vjerojatnosni graf izgleda u redu čak i ako greške ε_i nisu normalno distribuirane. Razlog tomu je što reziduali slučajni uzorak; oni su ostaci procesa procjene parametara. Reziduali su zapravo linearne kombinacije grešaka modela (ε_i). Prema tome, fitanjem parametara uništavamo dokaz o nenormalnosti reziduala te se ne možemo oslanjati na normalni vjerojatnosni graf za otkrivanje odstupanja od normalnosti. Također, česti se na na normalnom vjerojatnosnom grafu mogu pojaviti jedan ili dva velika reziduala. To ponekad ukazuje da su odgovarajuće vrijednosti outlieri.

Na slici 4.1 su prikazani neki od uzoraka normalnih vjerojatnosnih grafova. Na slici *a*) je prikazan idealan vjerojatnosni graf koji vodi do zaključka da bi uzorak zaista mogao biti normalno distribuiran. Na slici *b*) prikazana je distribucija uzorka sa lakšim repom, a na slici *c*) sa težim repom od normalne distribucije. Na slici *d*) prisutna je distribucija sa pozitivnom asimetrija, a na slici *e*) distribucija sa negativnom asimetrijom.



Slika 4.1: Normalni vjerojanosni grafovi (preuzeto iz [1])

4.2 Ocjena prilagodbe modela (*engl. Goodness-of-Fit)*

Empirijska funkcija distribucije (EDF) za n nezavisnih observacija $X_1, \ldots, X_n$ je definirana pomoću uobičajene funkcije distribucije $F(x)$. Neka su sa $X_{(1)}, \ldots, X_{(n)}$ označene observacije sortirane od najmanje do najveće. Empirijska funkcija distribucije $F_n(x)$ je definirana sa

$$\begin{aligned} F_n(x) &= 0, \quad x < X_{(1)} \\ F_n(x) &= \frac{i}{n}, \quad X_{(i)} \leq x \leq X_{(i+1)}, \quad i = 1, \ldots, n-1 \\ F_n(x) &= 1, \quad X_{(n)} \leq x \end{aligned}$$

Uočimo da je $F_n(x)$ step funkcija visine $\frac{1}{n}$ za svaku observaciju. Ova funkcija daje procjenu funkcije distribucije $F(x)$. Za bilo koji x, $F_n(x)$ je proporcija observacija manjih ili jednakim od x, dok je $F(x)$ vjerojatnost da je observacija manje ili jednake od x. EDF statistike mjere razliku između $F_n(x)$ i $F(x)$. Formule za računanje EDF statistika koriste transformaciju vjerojatnosnog integrala $U = F(x)$. Ako je $F(x)$ funkcija distribucije od X, tada je slučajna varijabla U uniformno distribuirana između 0 i 1. U nastavku teksta dat ćemo formalnu definiciju triju EDF statistika:

- Kolmogorov-Smirnov
- Anderson-Darling
- Cramér-von Mises

4.2.1 Kolmogorov statistika

Kolmogorov-Smirnov statististika D je definirana sa

$$D = sup_x \left| F_n(x) - F(x) \right|.$$

Temelji se na najvećoj okomitoj razlici između $F(x)$ i $F_n(x)$. Kolmogorov-Smirnov statististika se računa kao maksimum D^+ i D^-, gdje je D^+ najveća okomita udaljenost između EDF i funkcije distribucije kada je EDF veća od funkcija distribucije, i D^- je najveća okomita udaljenost kada je EDF manja od funkcije distribucije.

$$D^+ = max_i \left(\frac{i}{n} - U_{(i)} \right) D^- = max_i \left(U_{(i)} - \frac{i-1}{n} \right) D = max\left(D^+, D^-\right)$$

SAS koristi modificiranu Kolmogorov statistiku D za testiranje podataka o pripadnosti normalnoj distribuciji sa očekivanjem i varijancom jednakima očekivanju i varijanci uzorka.

4.2.2 Anderson-Darling statistika

Anderson-Darling statistika i Cramér-von Mises statistike se temelje na kvadriranoj udaljenosti razlike $(F_n(x) - F(x))^2$. Kvadratne statistike imaju sljedeću općenitu formu:

$$Q = n \int_{-\infty}^{-\infty} (F_n(x) - F(x))^2 \Psi(x) dF(x)$$

Funkcija $\Psi(x)$ je težinska funkcija kvadrirane razlike $(F_n(x) - F(x))^2$. Anderson-Darling statistika A^2 je definirana sa

$$Q = n \int_{-\infty}^{-\infty} (F_n(x) - F(x))^2 [F(x)(1 - F(x))]^{-1} dF(x)$$

Ovdje je težinska funkcija $\Psi(x) = [F(x)(1 - F(x))]^{-1}$. Anderson-Darling statistika se računa kao

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} [(2i - 1) log U_{(i)} + (2n + 1 - 2i) log(1 - U_{(i)})]$$

4.2.3 Cramér-von Mises statistika

Cramér-von Mises statistika W^2 je definirana sa

$$Q = n \int_{-\infty}^{-\infty} (F_n(x) - F(x))^2 dF(x)$$

Ovdje je težinska funkcija $\Psi(x) = 1$ Cramér-von Mises statistika se računa kao

$$W^2 = \sum_{i=1}^{n} \left(U_{(i)} - \frac{2i - 1}{2n}) \right)^2 + \frac{1}{12n}$$

Kada izračuna EDF statistike, SAS izračuna i odgovarajuće p-vrijednosti. Ispitivanjem p vrijednosti odlučujemo da li možemo odbaciti nultu hipotezu ili ne. Kada je p vrijednost manja od kritične vrijednosti α, odbacujemo nultu hipotezu u korist alternative i zaključujemo da dani podaci ne dolaze iz normalne distribucije.

Poglavlje 5

Transformacije varijabli

U ovom poglavlju ćemo se usredotočiti na metode i procedure modeliranja regresijskog modela kada su narušene neke od dosadašnjih pretpostavki. Naglasak će biti na transformacijama podataka. Nije neuobičajeno da kada su zavisna varijabla i/ili nezavisna varijabla izraženi na ispravnoj mjernoj skali, određena kršenja prepostavki, kao što je nejednakost varijanci, više nisu prisutna. U odabiru prikladne transformacije ponekad nas može voditi iskustvo ili teoretska podloga modela. Idealno bi bilo da metriku izabere inženjer ili znanstvenik sa potrebnim znanjem područja, ali ima puno situacija kada ova informacija nije dostupna. U tim slučajevima, transformaciju podataka možemo izabrati pomoću neke analitičke procedure.

5.1 Transformacije za stabilizaciju varijance

Pretpostavka konstantne varijance je osnovni zahtjev regresijske analize. Najčešći razlog kršenja ove pretpostavke je ako zavisna varijabla y slijedi vjerojatnosnu distribuciju u kojoj varijanca funkcijski ovisi o očekivanju. Npr. ako je y Poissonova slučajna varijabla u jednostrukom linearnom regresijskom modelu, tada je varijanca od y jednaka očekivanju. Budući da je očekivanje od y povezano sa nezavisnom varijablom x, varijanca od y će biti proporcionalna x.

Važno je ispitati konstantnost varijance. Prva naznaka o nekonstatnosti varijance može se uočiti iz grafova ili analize reziduala. Ako ne stabiliziramo varijancu, procjenitelji dobiveni metodom najmanjih kvadrata će i dalje biti nepristrani, ali više neće imati minimalnu varijancu. To znači da će koeficijenti regresora imati veću standardnu grešku nego što je potrebno. Zadatak transformacija je preciznije procjeniti parametre modela i povećati osjetljivost statističkih testova.

Nekoliko najčešćih transformacija kojima stabiliziramo varijancu su prikazane u tablici 5.1. Jačina transformacija ovisi o jačini krivulje. Transformacije u tablici se kreću od relativno blagog drugog korijena do relativno snažne recipročne vrijednosti. Općenito, blage transformacije nad relativno uskim rasponom vrijednosti (npr. $y_{max}/y_{min} < 2,3$) imaju mali utjecaj, dok će snažne transformacije nad širokim rasponom vrijednosti imat drastičan utjecaj na analizu.

Tablica 5.1: Korisne transformacije za stabilizaciju varijance (preuzeto iz [1])

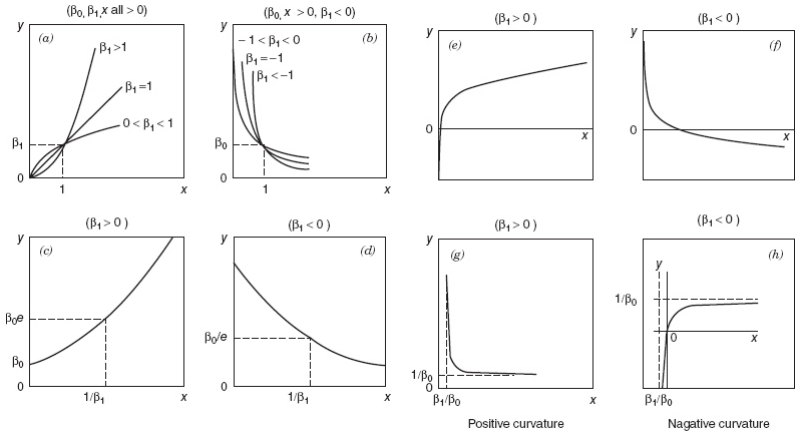
Veza između σ^2 i $E(y)$	***Transformacija***
$\sigma^2 \propto$ konstanta	$y' = y$ (nema transformacije)
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$ (drugi korijen; Poissonova distribucija)
$\sigma^2 \propto E(y)[1 - E(y)]$	$y' = \sin^{-1}(\sqrt{y})$ (arcsin; binomne proporcije $0 \leq y_i \leq 1$)
$\sigma^2 \propto [E(y)]^2$	$y' = \ln(y)$ (prirodni logaritam)
$\sigma^2 \propto [E(y)]_3$	$y' = y^{-1/2}$ (recipročna vrijednost drugog korijena)
$\sigma^2 \propto [E(y)]_4$	$y' = y^{-1}$ (recipročna vrijednost)

Nakon što transformiramo varijablu, predviđanja su unutar iste transformirane skale. Obično želimo vratiti predviđene vrijednosti na originalnu skalu. Međutim, direktna primjena inverzne transformacije na predviđene vrijednosti daje procjenu medijana distribucije zavisne varijable umjesto njezinog očekivanja. Pouzdani ili predikcijski intervali se mogu direktno pretvarati iz jedne metrike u drugu jer su te procjene intervala percentili distribucije i kao takvi nisu pod utjecajem transformacija. Međutim, ne postoji garancija da su to najuži mogući dobiveni intervali u originalnim jedinicama.

5.2 Transformacije za linearizaciju modela

Jedna od pretpostavki regresijske analize jest da je veza između zavisne varijable y i nezavisnih varijabli linearna. Nažalost, to ponekad nije točno. Nelinearnost se može vidjeti pomoću (matrice) grafova kombinacija varijabli ili (parcijalnih) grafova reziduala. U nekim slučajevima primjenom odgovarajuće transformacije možemo linearizirati nelinearnu funkciju.

Na slici 5.1 prikazano je nekoliko funkcija koje možemo linearizirati. Nelinearne funkcije, njihove transformacije i rezultirajuće linearne forme prikazane su u tablici 5.2 Ako uočimo krivulju na grafu y naspram x ukazuje na krivulju, možemo spojiti opaženo ponašanje grafa sa krivuljama na slici5.1 te iskoristiti linearizirani oblik funkcije. Ove transformacije zahtijevaju da su transformirane greške nezavisno normalno distribuirane sa očekivanjem nula i konstantnom varijancom.



Slika 5.1: Funkcije koje možemo linearizirati (preuzeto iz [1])

Tablica 5.2: Funkcije koje možemo linearizirati i njihov linearni oblik (preuzeto iz [1])

Slika	Nelinearna funkcija	Transformacija	Linearni oblik
5.4a, b	$y = \beta_0 x^{\beta_1}$	$y' = \log y, x' = \log x$	$y' = \log \beta_0 + \beta_1 x'$
5.4c, d	$y = \beta_0 e^{\beta_1 x}$	$y' = \ln y$	$y' = \ln \beta_0 + \beta_1 x$
5.4e, f	$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y' = \beta_0 + \beta_1 x'$
5.4g, h	$y = \frac{x}{\beta_0 x - \beta_1}$	$y' = \frac{1}{y}, x' = \frac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

Poglavlje 6

Adekvatnost modela

Proučavajući linearnu regresijsku analizu do sada smo pretpostavili sljedeće:
1. Veza između zavisne varijable y i nezavisnih varijabli je (barem približno) linearna.
2. Greške ε imaju očekivanje 0.
3. Greške ε imaju konstantnu varijancu σ^2.
4. Greške su nekorelirane.
5. Greške su normalno distribuirane.

Pretpostavke 4 i 5 zajedno povlače da su greške nezavisne slučajne varijable. Pretpostavka 5 je potrebna za testiranje hipoteza i procjenu intervala.

Uvijek moramo provjeriti navedene pretpostavke te ispitati adekvatnost dobivenog modela. U ovom ćemo poglavlju predstaviti nekoliko korisnih metoda provjere adekvatnosti modela koje se uglavnom temelje na proučavanju reziduala modela.

6.1 Definicija reziduala

Reziduale smo definirali kao

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n \tag{6.1}$$

gdje je y_i observacija i $\hat{y}_i$ odgovarajuća procijenjena vrijednost. Budući da reziduale možemo smatrati devijacijom između podataka i fita, oni su također i mjera varijabilnosti zavisne varijable neobjašnjene regresijskim modelom.

Reziduali imaju nekoliko važnih svojstava: očekivanje im je nula, a njihova procijenjena prosječna varijanca je

$$\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_{Res}}{n-p} = MS_{Res}$$

6.2 Skalirani reziduali

U ovom poglavlju predstavit ćemo četiri metode skaliranja reziduala. Reziduali su korisni u otkrivanje outliera ili ekstremnih vrijednosti, tj. observacija koje odudaraju od trenda ostatka dijela podataka.

Standardizirani reziduali. Budući da je pomoću MS_{Res} procijenjen približni prosjek varijance reziduala, logično skaliranje reziduala bili bi standardizirani reziduali

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}, \quad i = 1, 2, \ldots, n \tag{6.2}$$

Standardizirani reziduali imaju očekivanje nula i približno jediničnu varijancu. Potencijalni outlieri će imati velike standardizirane reziduale (npr. $d_i > 3$).

Studentizirani reziduali. Ako za varijancu i-tog reziduala koristimo MS_{Res}, za e_i dobivamo samo aproksimaciju. Možemo poboljšati skaliranje reziduala dijeljenjem e_i sa točnom standardnom devijacijom i-tog reziduala. Prisjetimo se ((3.10)) da vektor reziduala možemo zapisati kao

$$e = (I - H)y \tag{6.3}$$

gdje je $H = X(X'X)^{-1}X'$. Matrica H ima nekoliko korisnih svojstava: simetrična je ($H' = H$) i idempotentna ($HH = H$). Matrica $I - H$ je također simetrična i idempotentna. Ako uvedemo supstituciju $y = X\beta + \varepsilon$ u jednakost (6.3), dobivamo

$$e = (I-H)(X\beta+\varepsilon) = X\beta-HX\beta+(I-H)\varepsilon = X\beta-X(X'X)^{-1}X'X\beta+(I-H)\varepsilon = (I-H)\varepsilon \tag{6.4}$$

Dakle, reziduali su ista linearna transformacija observacija y kao i grešaka ε.

Kovarijacijska matrica reziduala je

$$Var(e) = Var\left[(I - H)\varepsilon\right] = (I - H)Var(\varepsilon)(I - H)' = \sigma^2(I - H) \tag{6.5}$$

budući da je $Var(e) = \sigma^2 I$ i $I - H$ je simetrična i idempotentna. Matrica $I - H$ općenito nije dijagonalna pa reziduali imaju različite varijance i korelirani su.

Varijanca i-tog reziduala je

$$Var(e_i) = \sigma^2(1 - h_{ii}) \tag{6.6}$$

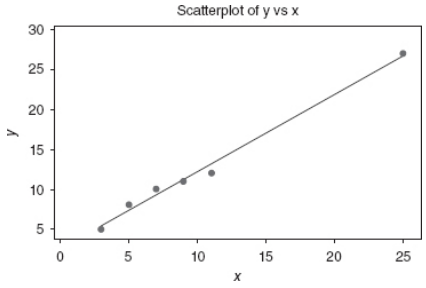
gdje je h_{ii} i-ti dijagonalni element matrice H. Kovarijanca između reziduala e_i i e_j jednaka je

$$Cov(e_i, e_j) = -\sigma^2 h_{ij} \tag{6.7}$$

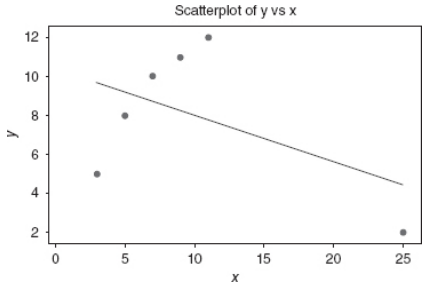
gdje je h_{ij} ij-ti element matrice H. Budući da je $0 \leq h_{ii} \leq 1$, koristeći MS_{Res} za procjenu varijance reziduala zapravo precjenjujemo $Var(e_i)$.

Nadalje, h_{ii} je mjera lokacije i-te točke na osi x pa varijanca od e_i ovisi o položaju točke x_i. Općenito, točke imaju veću varijancu oko središta osi x (lošiji fit metodom najmanjih kvadrata) nego reziduali na udaljenijim mjestima. Kršenje pretpostavki modela češće se događa u udaljenijim točkama i ono se teže može otkriti iz pregleda običnih residuala e_i (ili standardiziranih reziduala d_i) jer će obično njihovi reziduali biti manji.

Kako se udaljavamo od središta x osi, reziduali točaka su manji i teže prema 0. Na sljedećim slikama su prikazane dvije različite situacije. Na 1. slici za $x = 25$ vrijedi $y = 25$ i to je primjer udaljene točke koja prati trend podataka. Takva točka je udaljena u smislu uobičajenih vrijednosti nezavisnih varijabli, ali je opažena vrijednost zavisne varijable konzistentna sa predikcijom temeljenom na ostalim vrijednostima podataka. Na 2. slici za $x = 25$ vrijedi $y = 2$ i to je primjer udaljene točke koja ne prati trend podataka. Takav podatak nije samo udaljen u smislu uobičajenih vrijednosti nezavisnih varijabli, već i opažena vrijednost od y nije konzistentna s vrijednostima sa vrijednostima koje bismo predvidjeli na temelju ostalih podataka. Na obje slike povučen je pravac dobiven metodom najmanjih kvadrata nad cijelim skupom podataka. Uočavamo da influentual točka vuče pravac prema sebi.



Slika 6.1: Primjer udaljene točke koja prati trend podataka (preuzeto iz [1])



Slika 6.2: Primjer udaljene točke koja ne prati trend podataka (preuzeto iz [1])

Dakle, logično je ispitati studentizirane reziduale

$$r_i = \frac{e_i}{\sqrt{MS_{Res}\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}}, \quad i = 1, 2, \ldots, n \tag{6.8}$$

umjesto e_i (ili d_i). Ako je forma modela ispravna, studentizirani reziduali imaju konstantnu varijancu $Var(r_i) = 1$ bez obzira gdje se nalazi x_i. U mnogim situacijama varijanca reziduala se stabilizira, naročito za velike skupove podataka. Tada je razlika između standardiziranih i studentiziranih reziduala mala te oni zapravo daju ekvivalentnu informaciju. Međutim, kako svaka točka sa velikim rezidualom i velikim h_{ii} jako utječe na fit metodom najmanjih kvadrata, općenita preporuka je ispitivanje studentiziranih reziduala.

PRESS reziduali. Zanima nas y_i–$\hat{y}_{(i)}$ gdje je $\hat{y}_{(i)}$ fitana vrijednost i-te zavisne varijable temeljena na svim observacijama osim na i-toj. Ideja je ako je i-ta observacija y_i zaista neobična, regresijski model temeljen na svim observacijama bi mogao biti previše pod utjecajem te observacije. To bi moglo stvoriti fitanu vrijednost $\hat{y}_i$ veoma sličnu opaženoj vrijednosti y_i i kao posljedicu će imati male reziduale e_i. Prema tome, bit će teško otkriti outlier. Međutim, ako je i-ta observacija izbrisana, tada $\hat{y}_{(i)}$ ne mogu biti pod utjecajem te observacije pa bi rezltirani rezidual trebao ukazati na prisutnost outliera.

Ako izbrišemo i-tu observaciju, fitamo regresijski model na preostalih $n-1$ observacija i izračunamo odgovarajuću predikcijsku vrijednost od y_i za izbrisanu observaciju, odgovarajuća predikcijska greška je

$$e_{(i)} = y_i - \hat{y}_{(i)} \tag{6.9}$$

Ovo računanje predikcijske greške se ponavlja za svaku observaciju $i = 1, 2, \ldots, n$. Ove predikcijske greške obično zovemo PRESS reziduali. U početku se čini da je za računanje PRESS reziduala potrebno fitanje n različitih regresija. Međutim, moguće je izračunati PRESS reziduale od rezultata jednog fitanja najmanjih kvadrata za svih n observacija. Ispada da je i-ti PRESS rezidual

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}, \quad i = 1, 2, \ldots, n \tag{6.10}$$

Iz jednakosti (6.10) lako se vidi da je PRESS rezidual zapravo rezidual ponderiran dijagonalnim elemenatima h_{ii} matrice H. Reziduali točaka za koje imaju veliki h_{ii} imati će i velike PRESS reziduale. To su u većini slučajeva jako utjecajne točke. Velika razlika između običnih reziduala i PRESS reziduala će uglavnom ukazivati na točku gdje model dobro fita podatke, ali model bez te točke slabo predviđa.

Naposljetku, varijanca i-tog PRESS reziduala je

$$Var[e_{(i)}] = Var\left[\frac{e_i}{1-h_{ii}}\right] = \frac{1}{(1-h_{ii})^2}\left[\sigma^2(1-h_{ii})\right] = \frac{\sigma^2}{(1-h_{ii})}$$

pa je standardizirani PRESS rezidual jednak

$$\frac{e_{(i)}}{\sqrt{Var\left[e_{(i)}\right]}} = \frac{\frac{e_i}{1-h_{ii}}}{\sqrt{\frac{\sigma_i^2}{1-h_{ii}}}} = \frac{e_i}{\sqrt{\sigma_i^2(1-h_{ii})}}$$

koji je, ako uzmemo MS_{Res} za procjenu σ^2, zapravo studentizirani rezidual.

R-student. Prilikom računanja studentiziranih reziduala obično se koristi MS_{Res} kao procjena σ^2. MS_{Res} je unutarnje generirana procjena σ^2 dobivena fitanjem modela na svih n observacija. Drugi pristup bi bio uzeti procjenu σ^2 temeljenu na skupu podataka gdje je izbačena i-ta observacija. Označimo tako dobivenu procjenu σ^2 sa $S^2_{(i)}$. Tada vrijedi

$$S^2_{(i)} = \frac{(n-p)MS_{Res} - e_i^2/(1-h_{ii})}{n-p-1} \tag{6.11}$$

Da bismo dobili vanjske studentizirani rezidual, obično zvan R-student, dan sa

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}, \quad i = 1, 2, \ldots, n \tag{6.12}$$

u jednakosti (6.11) se koristi procjena σ^2 umjesto MS_{Res}. U mnogim situacijama t_i će se malo razlikovati od studentiziranih reziduala r_i. Međutim, ako je i-ta observacija utjecajna, tada se $S^2_{(i)}$ može značajno razlikovati od MS_{Res} stoga će statistika R-student biti osjetljivija u toj točki.

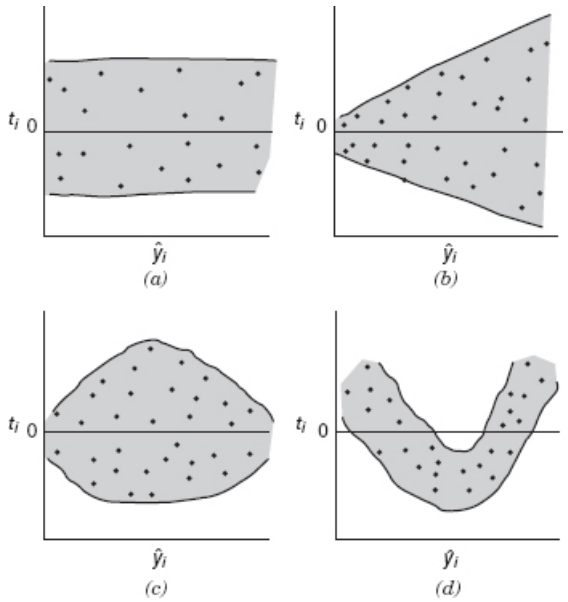
6.3 Grafovi reziduala

Grafička analiza reziduala je vrlo efikasan način za provjeru osnovnih pretpostavki i adekvatnosti regresijskog modela. Grafove reziduala koje ćemo objasniti u ovom poglavlju dostupni su u većini statističkih softverskih paketa. Često crtamo R–student reziduale jer imaju konstantu varijancu.

Graf reziduala naspram fitanih vrijednosti $\hat{y}_i$. Graf reziduala (po mogućnosti vanjskih studentiziranih reziduala, t_i) naspram odgovarajućih fitanih vrijednosti y_i su korisni za otkrivanje nekoliko čestih tipova neadekvatnosti modela. Ako je dobiveni graf sličan slici 6.3*a* gdje se reziduali nalaze unutar horizontalne "pruge", tada ne postoje očiti defekti u modelu. Grafovi koji su slični nekoj od slika $b-d$ upozoravaju na neke nedostatke modela.

Na slikama *b* i *c* varijanca nije konstantna. Uzorak na slici b povlači da je varijanca rastuća funkcija od *y* (analogno može biti i obratno). Uzorak na slici c se najčešće pojavljuje kada je *y* proporcija između 0 i 1. Naime, varijanca binomne proporcije s parametrom 0.5 je veća nego ona sa parametrom 1. Da bismo se riješili nekonstantnosti varijance u ovim slučajevima najčešće primjenjujemo pogodnu transformaciju bilo nezavisnih varijabli ili zavisne varijable. U praksi se to svodi na transformaciju zavisne varijable.

Slika *d* ukazuje na nelinearnost. Ovo može značiti da su potrebni drugi članovi u modelu, npr. kvadratni član. Transformacije nezavisnih varijabli i/ili zavisne varijable također mogu biti potrebne u ovim slučajevima.



Slika 6.3: Uzorci grafova reziduala: a) zadovoljavajući; b)lijevak; c) dvostruki luk; d) nelinearan (preuzeto iz [1])

Graf reziduala naspram $\hat{y}_i$ mogu također otkriti jednu ili više neobično velikih reziduala, tj. potencijalnih outliera. Veliki reziduali pri ekstremnim vrijednostima $\hat{y}_i$ mogu biti pokazatelj nekonstantnosti varijance ili da prava veza između *y* i *x* nije linearna.

Graf reziduala naspram nezavisnih varijabli. Na ovim grafovima često su slični uzorci kao na slici 6.3, osim što se na vodoravnoj osi umjesto $\hat{y}_i$ nalaze x_{ij} *j*-t nezavisne varijable. I dalje težimo tome da su reziduali sadržani unutar vodoravne pruge. Slike b i c pokazuju problem sa nekonstantnom varijancom. Slika d ili bilo koji drugi nelinearan uzorak povlače da je pretpostavljena veza između *y* i nezavisne varijable x_j netočna. Prema tome, ili su potrebni članovi višeg reda u x_j (npr. x_j^2) ili je potrebna transformacija.

Poglavlje 7

Odabir varijabli u modelu

Prilikom dosadašnjeg modeliranja regresijskog modela pretpostavili smo da znamo koje su nezavisne varijable važne i samim time čine konačan model. U primjenama uglavnom na raspolaganju imamo veći skup nezavisnih varijabli od kojih je najčešće samo manji dio njih zaista važan.

U ovoj cjelini bavit ćemo se upravo sljedećom problematikom: Kako odabrati „najbolji“ podskup danog skupa nezavisnih varijabli? Suočit ćemo se sa dva suprotna zahtjeva:
1. Želimo da model sadrži što je moguće više nezavisnih varijabli koji će opisivati ponašanje zavisne varijable.
2. Želimo da model sadrži što je moguće manje nezavisnih varijabli jer se sa povećanjem broja nezavisnih varijabli povećava i varijanca predviđanja te je skuplje sakupljanje podataka i održavanje modela.
Proces traženja modela koji je kompromis između ova dva zahtjeva zove se odabir „najbolje“ regresijske jednadžbe.

Što je manje varijabli u modelu, varijanca preostalih procjena parametara se smanjuje. Pritom se može pojaviti pristranost u procjenama, ali ako je efekt varijabli koje smo izbacili mali, tada je pristranost manja od redukcije varijance. Izbacivanje varijabli najčešći je način rješavanja problema multikolinearnosti unutar modela. Također, razni algoritmi odabira varijabli u modelu implementirani su u statističkim softverskim paketima pa je traženje konačnog/ih modela uvelike olakšano.

Nažalost, nijedna od procedura koje ćemo opisati u ovom poglavlju nije garancija da smo pronašli „najbolju“ regresijsku jednadžbu. Naprotiv, različite procedure će često predložiti nekoliko jednako dobrih regresijskih modela. Prilikom modeliranja oslanjat ćemo se na statističko razmišljanje, iskustvo i subjektivne odluke sa završnim savjetovanjem sa strukom. Također smo uvidjeli problematiku izbacivanja nezavisnih varijabli i observacija iz baze. Naime, tim postupkom mijenja se veličina baze ili prostora te se dobivaju novi rezultati.

7.1 Kriteriji za procjenu modela

Od sada pa nadalje, podmodelom smatramo model koji sadrži neki podskup nezavisnih varijabli punog modela sa K nezavisnih varijabli. U ovom poglavlju uvodimo kriterije za procjenu modela pomoću kojih ćemo lakše usporediti dobivene modele.

7.1.1 Koeficijent determinacije R^2

Neka R_p^2 označava koeficijent determinacije regresijskog p-članog podmodela, tj. modela koji sadrži $p-1$ nezavisnih varijabli i slobodni član β_0. Tada je

$$R_p^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{Res}(p)}{SS_T} \tag{7.1}$$

gdje je $SS_R(p)$ suma kvadrata regresije, a $SS_{Res}(p)$ suma kvadrata reziduala p-članog podmodela.

Uočimo da je R_p^2 rastuća fukcija od p te da se za $p = K + 1$ postiže maksimum. Dakle, dodavanje nezavisnih varijabli u model je opravdano ukoliko je promjena u R^2 značajna. R_p^2 je bolji kriterij za usporedbu modela jednakih veličina. Za svaki skup p-članih modela, R_p^2 može poprimiti $(K \mid (p-1))$ vrijednosti. Bolji modeli su oni sa većim R_p^2.

7.1.2 Prilagođeni koeficijent determinacije R_{Adj}^2

Za p-člani model prilagođeni koeficijent determinacije R_{Adj}^2 definiramo na sljedeći način:

$$R_{Adj,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2) \tag{7.2}$$

Uočavamo da dodavanjem dodatnih nezavisnih varijabli u model statistika $R_{Adj,p}^2$ ne mora nužno rasti. Ako dodamo s nezavisnih varijabli u model, $R_{Adj,p+s}^2$ će biti veći od $R_(Adj, p)^2$ ako i samo ako je parcijalna F statistika za testiranje značajnosti s dodatnih nezavisnih varijabli veća od 1. Bolji modeli su oni sa maksimalnim $R_{Adj,p}^2$.

7.1.3 Prosječni kvadratni rezidual MS_{Res}

Prosječni kvadratni rezidual definiran je sa

$$MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p} \tag{7.3}$$

Budući da se $SS_{Res}(p)$ smanjuje kako p raste, $MS_{Res}(p)$ se prvo smanjuje, zatim se stabilizira da bi se na kraju možda povećao.

Do povećanja $MS_{Res}(p)$ dolazi ako redukcija $SS_{Res}(p)$ od dodavanja nezavisnih varijabli u model nije dovoljna da bi nadoknadila gubitak jednog stupnja slobode u nazivniku. Podmodel koji minimizira $MS_{Res}(p)$ također maksimizira $R^2_{Adj,p}$:

$$R^2_{Adj,p} = 1 - \frac{n-1}{n-p}(1-R^2_p) = 1 - \frac{n-1}{n-p}\frac{SS_{Res}(p)}{SS_T} = 1 - \frac{MS_{Res}(p)}{\frac{SS_T}{n-1}} \tag{7.4}$$

Dakle, bolji modeli su oni sa minimalnim MS_{Res}.

7.1.4 Mallowa C_p statistika

Ova statistika je definirana na sljedeći način:

$$E[\hat{y}_i - E(y_i)]^2 = [E(y_i) - E(\hat{y}_i)]^2 + Var(\hat{y}_i) \tag{7.5}$$

pri čemu je $E(y_i)$ očekivanje zavisne varijable punog modela i $E(\hat{y}_i)$ je očekivanje zavisne varijabla p-članog podmodela.

Mallowa C_p statistika predlaže manje modele nego prilagođeni R^2. Kod regresijskih jednadžbi gdje ima malo pristranosti vrijednosti od C_p će se nalaziti blizu pravca $C_p = p$, dok će kod jednadžbi sa značajnom pristranošću biti iznad ovog pravca. Bolji modeli su oni sa malim vrijednostima statistike C_p.

7.1.5 Akaike informacijski kriterij (AIC)

AIC se temelji na maksimiziranju očekivane entropije modela, tj. mjere očekivane informacije. Neka je L vjerodostojnost za model s p parametara. AIC definiramo sa

$$AIC = -2ln(L) + 2p$$

U slučaju najmanjih kvadrata regresije

$$AIC = n\ln\left(\frac{SS_{Res}}{n}\right) + 2p$$

Zapravo je AIC penalizirana log vjerodostojnost. Ideja je slična ideji kod R^2_{Adj} i Mallowe C_p statistike. Dodavanjem nezavisnih varijabli modelu, SS_{Res} ne može rasti. Pitamo se da li smanjenje SS_{Res} opravdava uključivanje dodatnog člana. Bolji modeli su oni sa minimalnim AIC-om.

7.2 Metode odabira varijabli

7.2.1 Sve moguće regresije

Da bismo pronašli podskup varijabli koje ćemo koristiti u završnoj jednadžbi, prirodno je razmotriti modeliranje zavisne varijable pomoću raznih kombinacija kandidata nezavisnih varijabli.

U ovoj proceduri traže se sve moguće kombinacije potencijalnih nezavisnih varijabli u modelu. Najprije se modelira model koji uključuje jednu nezavisnu varijablu, zatim dvije nezavisne varijable, itd. Svaki od danih modela se procjenjuje nekim od navedenih kriterija te se izabire „najbolji“ regresijski model.

Ako pretpostavimo da je slobodni član β_0 prisutan u svakoj jednadžbi i imamo K potencijalnih nezavisnih varijabli, tada je ukupno 2^K mogućih jednadžbi koje treba procijeniti i ispitati. Uočavamo da broj dobivenih jednadžbi brzo raste kako raste broj potencijalnih nezavisnih varijabli.

Postoji nekoliko algoritama korisnih za generiranje svih mogućih regresija. Temeljna ideja svih ovih algoritama je izvedba 2^K mogućih podmodela na način da se svaki sljedeći podmodel u nizu razlikuje za samo jednu varijablu. Ova procedura provediva je ukoliko na raspolaganju imamo 30 ili manje potencijalnih nezavisnih varijabli. Pomoću ove metode dobivamo preporuku nekoliko potencijalnih modela što daje struci prostora da pomogne pri daljnjem modeliranju.

7.2.2 Stepwise regresijske metode

Budući da procjena svih mogućih regresija može biti mukotrpna, razvile su se različite metode procjene samo malog podskupa regresijskih modela tako da ili dodajemo ili brišemo jednu po jednu nezavisnu varijablu. Te metode obično se odnose na stepwise procedure. Možemo ih svrstati u 3 kategorije:

1. forward selection
2. backward elimination
3. stepwise regression, koji je kombinacija procedura 1 i 2.

Forward selection

Ova procedura kreće od modela koji se sastoji samo od slobodnog člana. Zatim traži optimalni podskup tako da ubacuje nezavisne varijable u model jedan po jedan. Prva nezavisna varijabla koja ulazi u jednadžbu je ona koja ima najveću korelaciju sa zavisnom varijablom y. Pretpostavimo da je to nezavisna varijabla x_1. To je također nezavisna varijabla koja će imati najveću vrijednost F statistike za testiranje značajnosti regresije. Ova nezavisna varijabla ulazi ako F statistika prelazi prvotno odabranu vrijednost F, recimo F_{IN}. Druga nezavisna varijabla koja ulazi je ona koja sad ima najveću korelaciju sa y nakon prilagodbe efekta prve nezavisne varijable koja je ušla (x_1) za y. Ovu korelaciju nazivamo parcijalnom korelacijom. To je jednostruka korelacija između reziduala regresijske jednadžbe $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ i reziduala regresijske jednadžbe ostalih potencijalnih nezavisnih varijabli na x_1, recimo $\hat{x}_j = \hat{\alpha}_0 \hat{\alpha}_{1j} x_1$, $j = 2, 3, \ldots, K$.
Pretpostavimo da je u 2. koraku x_2 nezavisna varijabla sa najvišom parcijalnom korelacijom sa y. To povlači da je najveća parcijalna F statistika

$$F = \frac{SS_R(x_2 \mid x_1)}{MS_{Res}(x_1, x_2)}$$

Ako ova F vrijednost prelazi F_{IN}, tada x_2 dodajemo u model. Općenito, u svakom koraku nezavisna varijabla koja ima najvišu parcijalnu korelaciju sa y (ili ekvivalentno najveću parcijalnu F statistiku dobivenu kada su ostale nezavisne varijable već u modelu) je dodana u model ako njezina parcijalna F statistika koraku prelazi F_{IN}. Procedura završava kada parcijalna F statistika u nekom koraku ne prelazi F_{IN} ili kada je u model dodana zadnja potencijalna nezavisna varijabla. Neki kompjuterski paketi ispisuju t statistike za ulazak ili izlazak varijabli. To je u skladu sa opisanom procedurom jer vrijedi $t_{\frac{\alpha}{2},v} = F_{\alpha,1,v}$.

Backward elimination

Forward selection počinje bez ijedne nezavisne varijable u modelu i dodaje varijable dok ne dobijemo prikladan model. Backward elimination funkcionira obrnuto, tj. počinjemo s modelom koji uključuje svih K kandidata nezavisnih varijabli. Tada se računa F statistika (ili ekvivalentno t statistika) za svaku nezavisnu varijablu kao da je zadnja varijabla koja je ušla u model. Najmanja od ovih parcijalnih F (ili t) statistika se uspoređuje sa prvotno izabranom vrijednošću, F_{OUT} (ili t_{OUT}) i ako je najmanja parcijalna F vrijednost manja od F_{OUT} tu nezavisnu varijablu izbacujemo iz modela. Sada modeliramo regresijski model sa $K - 1$ nezavisnih varijabli. Računamo parcijalnu F statistiku za taj novi model, i procedura se ponavlja. Backward elimination algoritam završava kada najmanja parcijalna F statistika nije manja od prvotno definirane vrijednosti F_{OUT} (ili t_{OUT}).

Stepwise regression

Stepwise regresija je modifikacija forward selectiona u kojem u svakom koraku parcijalnom F (ili t) statistikom ispitujemo sve nezavisne varijable koje su prethodno ušle u model. Može se dogoditi da je nezavisna varijabla koja je dodana u ranijem koraku sada postala suvišna zbog veze između nje i nezavisnih varijabli koje su sada u jednadžbi. Ako je parcijalna F (ili t) statistika varijable manja od F_{OUT} (ili t_{OUT}), tada izbacujemo varijablu iz modela.

Stepwise regression zahtijeva dvije unaprijed definirane vrijednosti: jednu za ulazak varijabli i jednu za njihovo izbacivanje. Neki analitičari preferiraju izabrati F_{IN} (ili t_{IN})= F_{OUT} (ili t_{OUT}), iako to nije potrebno. Često biramo F_{IN} (ili t_{IN}) > F_{OUT} (ili t_{OUT}), tako da je teže dodati nezavisnu varijablu nego ju izbrisati.

Osvrt na Stepwise procedure

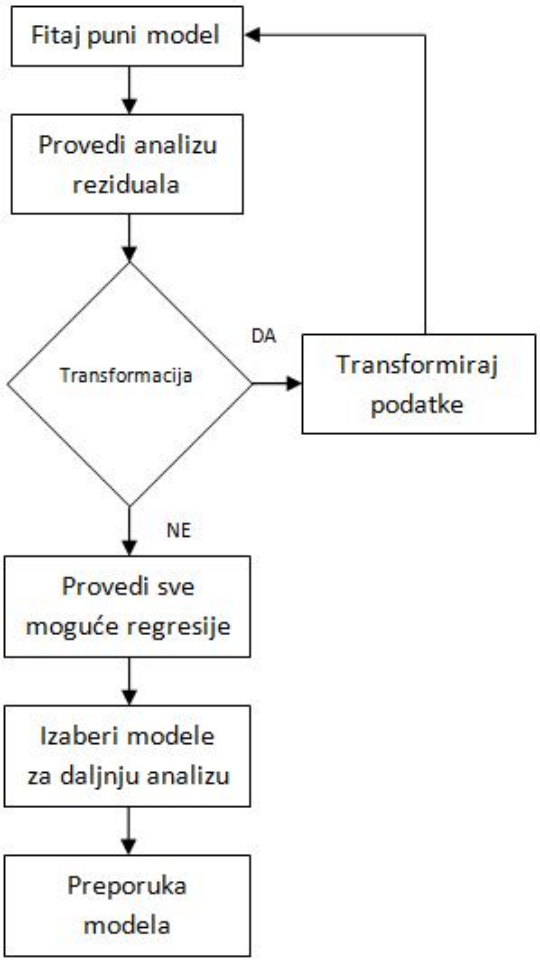
Najčešća kritika stepwise procedura je da ne garantiraju pronalazak najboljeg podmodela bilo koje veličine. Jedan od razloga je što najbolji podmodel niti ne postoji. Uočimo da forward selection, backward elimination i stepwise regression ne vode obavezno do istog modela.

Također, treba imati na umu da red kojim nezavisne varijable ulaze ili izlaze iz modela ne povlače obavezno redoslijed važnosti nezavisnih varijabli. Naprotiv, ponekad upravo nezavisna varijabla koja je ušla ranije u model može izaći iz modela u sljedećem koraku. Kod forward selection procedure problem je što jednom kad je dodana nezavisna varijabla više ju ne možemo izbaciti u sljedećem koraku. Backward elimination algoritam je manje pod negativnim utjecajem korelacijske strukture nezavisnih varijabli od forward selectiona.

Najbolja preporuka je primijeniti sve procedure u nadi da ćemo ili vidjeti nego suglasje ili naučiti nešto o strukturi podataka koju smo mogli previdjeti koristeći samo jednu proceduru odabira. Forward selection algoritam se slaže sa metodom svih mogućih regresija za male podmodele, ali ne i za velike, dok se backward elimination slaže sa svim mogućim regresijama za velike podmodele, ali ne i za male.

7.3 Strategije odabira varijabli i modeliranje

Sljedeća slika sažima osnovni pristup odabiru varijabli i modeliranju modela.



Slika 7.1: Proces modeliranja (preuzeto iz [1])

Način na koji smo mi radili:

1. Normalizirali zavisnu varijablu
2. Koristili stepwise metodu da bismo istaknuli potencijalne nezavisne varijable.
3. Proveli regresiju koristeći dobivene nezavisne varijable.
4. Izbacili varijable koje nisu bile statistički značajne na razini značajnosti od 5%.
    - pritom pazili da radimo na istoj bazi
    - ukoliko je neka od nezavisnih varijabli diskretna varijabla, modelirali smo model za svaku vrijednost te diskretne varijable

Poglavlje 8

Primjena regresijske analize

Podaci koje koristimo u ovom radu prikupljeni su u sveučilišnoj klinici za dijabetes, endokrinologiju i bolesti metabolizma Vuk Vrhovac. Naš zadatak je odrediti glavne prediktore razine adiponektina, homocisteina, cistacina C i ekskrecije albumina u urinu kod dijabetesa tipa 2. Pritom se služimo do sada navedenim rezultatima regresijske analize.

8.1 Opis podataka

Obzirom na ubrzani način života, dijabetes ili šećerna bolest je goruć problem današnjice. Svakodnevni stres je jedan od glavnih okidača ove bolesti čije ignoriranje simptoma može dovesti do ozbiljnih zdravstvenih problema. Kod dijabetičara gušterača proizvodi premalo inzulina koji bi omogućio da šećer unešen hranom iz krvi prijeđe u mišiće i druge stanice koje proizvode energiju. Ako šećer ne može ući u stanice, on se nakuplja u krvi. Zato je osnovna značajka dijabetesa povišen šećer u krvi.

Postoje dva osnovna tipa dijabetesa: dijabetes tipa 2 i dijabetes tipa 1. Kod dijabetesa tipa 2 gušterača ili dalje izlučuje inzulin, ali u manjim količinama ili proizvedeni inzulin ne djeluje pravilno (neosjetljivost mišića na inzulin). Ova dijagnoza se može regulirati boljom prehranom, određivanjem vremena obroka, vježbanjem ili smanjenjem težine. U krajnjem slučaju je potrebno uzimati lijekove koji povećavaju proizvodnju inzulina ili osjetljivost na njega. Dijabetes tipa 1 je teži oblik dijabetesa uzrokovan potpunim nedostatkom inzulina. To je stalno stanje i za njegovo kontroliranje potrebno je primanje inzulina putem injekcija svaki dan.

U ovom radu koristimo podatke pacijenata koji boluju od dijabetesa tipa 2. Kod svakog od njih mjerile su se sljedeće varijable:

DISKRETNA VARIJABLA:

1. SPOL (oznaka *sex*)

NEPREKIDNE VARIJABLE:

a) ZAVISNE VARIJABLE:

1. ADIPONEKTIN (oznaka *ApN*)
2. HOMOCISTEIN (oznaka *HCY*)
3. CYSTATIN C (oznaka *Cys_C*)
4. EKSKRECIJA ALBUMINA (oznaka *AER*)

b) POTENCIJALNE NEZAVISNE VARIJABLE:

1. DOB (oznaka *age*)
2. TRAJANJE (oznaka *duration*)
3. OPSEG STRUKA (oznaka *waste*)
4. OMJER OPSEGA STRUKA I BOKOVA (oznaka *WHR*)
5. GLIKEMIJA NATAŠTE (oznaka *fBG*)
6. RAZINA GLUKOZE U KRVI NAKON JELA (oznaka *ppBG*)
7. HEMOGLOBIN A1c (oznaka *HbA1c*)
8. C-REAKTIVNI PROTEIN (oznaka *CRP*)
9. FIBRINOGEN (oznaka *FIB*)
10. INTERLEUKIN-6 (oznaka *IL_6*)
11. SIALIČNA KISELINA (oznaka *Sijal_ac_*)
12. KLIRENS KREATININA (oznaka *klirens*)
13. SISTOLIČKI TLAK (oznaka *SBP*)

14. DIJASTOLIČKI TLAK (oznaka *DBP*)
15. BIJELE KRVNE STANICE (oznaka *WBC*)
16. LIPOPROTEIN VISOKE GUSTOĆE (oznaka *HDL*)
17. LIPOPROTEIN NISKE GUSTOĆE (oznaka *LDL*)
18. TRIGLICERIDI (oznaka *TG*)
19. ASPARTATE AMINOTRANSFERASE (oznaka *AST*)
20. ALANINE AMINOTRANSAMINASE (oznaka *ALT*)
21. GAMMA-GLUTAMYL TRANSPEPTIDASE (oznaka *GGT*)
22. MOKRAĆNA KISELINA (oznaka *MK*)
23. C-PEPTID natašte (fasting) (oznaka *C1*)
24. INDEX MASNE JETRE (fatty liver index) (oznaka *FLI*)
25. MJERA POHRANE GLUKOZE (oznaka *GDR*)

Kao što smo već naveli, u ovom poglavlju ćemo nastojati odrediti glavne prediktore razine adiponektina, homocisteina, cistacina C i ekskrecije albumina u urinu. Svaka od ovih varijabli pomoći će nam u predviđanju daljnjeg tijeka bolesti.

Prilikom obrade podataka koristit ćemo statistički program SAS (University Edition). Podatke ćemo pohraniti u `bazu`. Sintaksa za učitavanje podataka je sljedeća:

```
data baza;
   input sex age duration waste WHR fBG ppBG HbA1c ApN CRP HCY FIB
   IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST ALT
   GGT MK C1 FLI GDR;
   datalines;
podaci;
run;
```

8.2 Deskriptivna statistika

Spol je jedina diskretna varijabla u bazi pa ćemo prvo nju analizirati pomoću `proc freq`. Ona je kodirana pomoću dvije vrijednosti: 1 označava žene, a 2 muškarce.

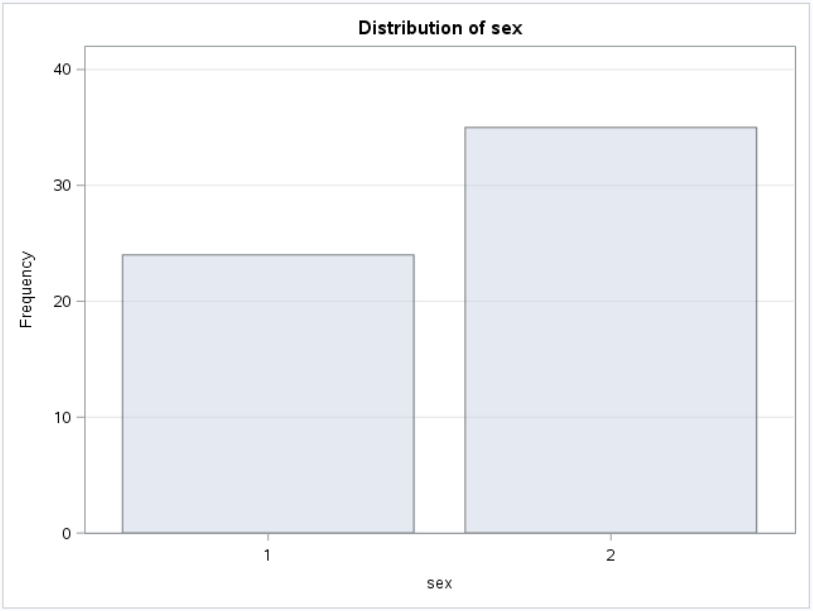
Kod u SAS-u:

```
1    proc freq data=baza;
2        table sex/ chisq binomial(level='1') plots=freqplot;
3    run;
```

The FREQ Procedure

sex

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	24	40.68	24	40.68
2	35	59.32	59	100.00



Chi-Square Test for Equal Proportions	
Chi-Square	2.0508
DF	1
Pr > ChiSq	0.1521

Slika 8.1: Deskriptivna statistika za varijablu sex (spol)(ispis iz SAS-a)

Testom χ^2 za testiranje jednakosti proporcija ne možemo odbaciti nultu hipotezu o jednakosti proporcija žena i muškaraca.

Ostale varijable u bazi su neprekidne i njihova deskriptivna statistika dana je u tablici 8.1.

Kod u SAS-u:

```
proc means data=baza N Min Median Max Mean Std Skew Kurt nolabels
maxdec=3;
    var age duration waste WHR fBG ppBG HbA1c ApN CRP HCY FIB
    IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST
    ALT GGT MK C1 FLI GDR;
run;
```

Tablica 8.1: Deskriptivna statistika neprekidnih varijabli (ispis iz SAS-a)

The MEANS Procedure

Variable	N	Minimum	Median	Maximum	Mean	Std Dev	Skewness	Kurtosis
age	59	34.000	60.000	79.000	59.814	10.768	-0.325	-0.660
duration	59	0.500	10.000	27.000	11.017	7.118	0.488	-0.632
waste	53	73.000	103.000	144.000	106.849	12.907	0.748	1.243
WHR	56	0.800	0.960	1.850	0.971	0.140	4.688	28.994
fBG	59	4.700	8.700	19.700	9.705	3.502	1.155	0.926
ppBG	58	3.700	10.050	23.500	11.428	4.240	0.719	-0.092
HbA1c	59	5.200	7.000	12.700	7.212	1.540	1.078	1.465
ApN	58	2.280	7.290	14.580	7.705	2.853	0.718	0.054
CRP	59	0.100	2.000	27.200	4.093	6.042	2.845	7.773
HCY	59	4.900	12.200	42.600	14.290	7.502	2.010	4.328
FIB	59	2.100	3.800	5.400	3.773	0.794	-0.071	-0.604
IL_6	55	0.000	2.200	19.100	3.962	4.394	1.387	1.743
Cys_C	59	0.500	0.850	2.590	1.001	0.458	2.183	4.406
Sijal_ac_	59	1.340	2.070	2.810	2.042	0.305	0.055	0.404
klirens	57	0.300	1.800	3.440	1.796	0.748	0.194	-0.496
AER	58	2.690	14.580	3622.000	225.546	654.629	4.145	17.611
SBP	59	80.000	135.000	220.000	135.678	22.234	0.724	3.497
DBP	59	60.000	80.000	100.000	82.881	10.180	-0.463	0.181
WBC	59	4.400	7.500	14.400	7.732	1.838	0.906	1.597
HDL	59	0.530	1.330	2.500	1.317	0.369	0.278	0.664
LDL	59	1.500	2.940	6.320	3.048	0.952	1.096	1.930
TG	59	0.670	2.090	16.170	2.593	2.477	3.797	17.469
AST	59	14.000	23.000	56.000	25.644	9.240	1.192	1.310
ALT	59	10.000	23.000	70.000	27.356	13.821	0.999	0.454
GGT	59	9.000	27.000	195.000	38.458	34.392	2.688	8.503
MK	59	133.000	298.000	562.000	300.102	79.558	0.734	1.537
C1	57	0.100	0.680	2.130	0.713	0.445	1.274	1.912
FLI	53	3.941	85.114	99.855	74.299	24.797	-0.970	0.000
GDR	56	0.828	5.649	11.041	6.006	2.181	0.290	0.017

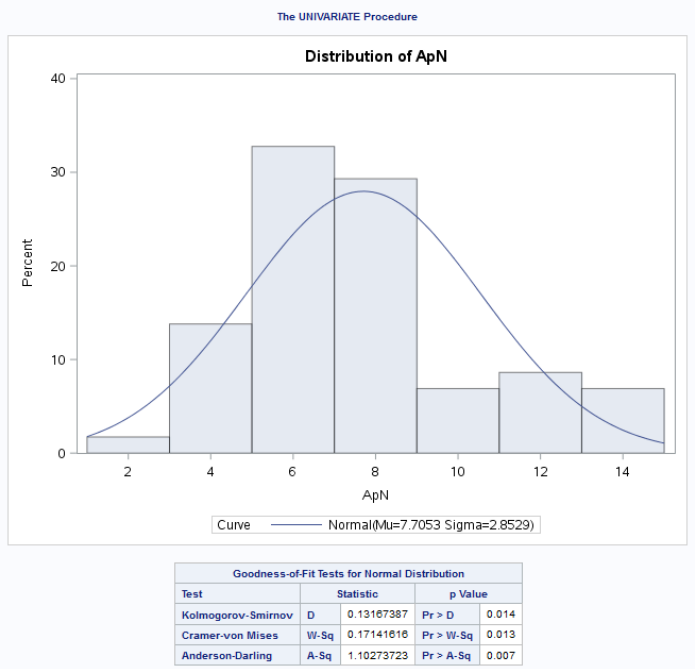
8.3 Transformacije zavisnih varijabli

Da bi pretpostavke potrebne za provedbu linearne regresije bile zadovoljene, potrebno je ispitati normalnost podataka. Ako podaci nisu normalno distribuirani, potrebno je primijeniti odgovarajuće transformacije za njihovu normalizaciju. Prit tome će od velike koristi biti opcija `ASSIST` implementirana u SAS-u. U ovom radu ćemo se ograničiti na zahtjev normalnosti zavisnih varijabli, tj. na normalnost adiponektina, homocisteina, cistacina C te ekskrecije albumina.

8.3.1 Transformacija adiponektina, ApN

Kod u SAS-u:

```
1  proc univariate data=baza;
2      var ApN;
3      histogram / normal;
4  run;
```



Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.13167387	Pr > D	0.014
Cramer-von Mises	W-Sq	0.17141616	Pr > W-Sq	0.013
Anderson-Darling	A-Sq	1.10273723	Pr > A-Sq	0.007

Slika 8.2: Histogram i testiranje normalnosti zavisne varijable ApN (ispis iz SAS-a)

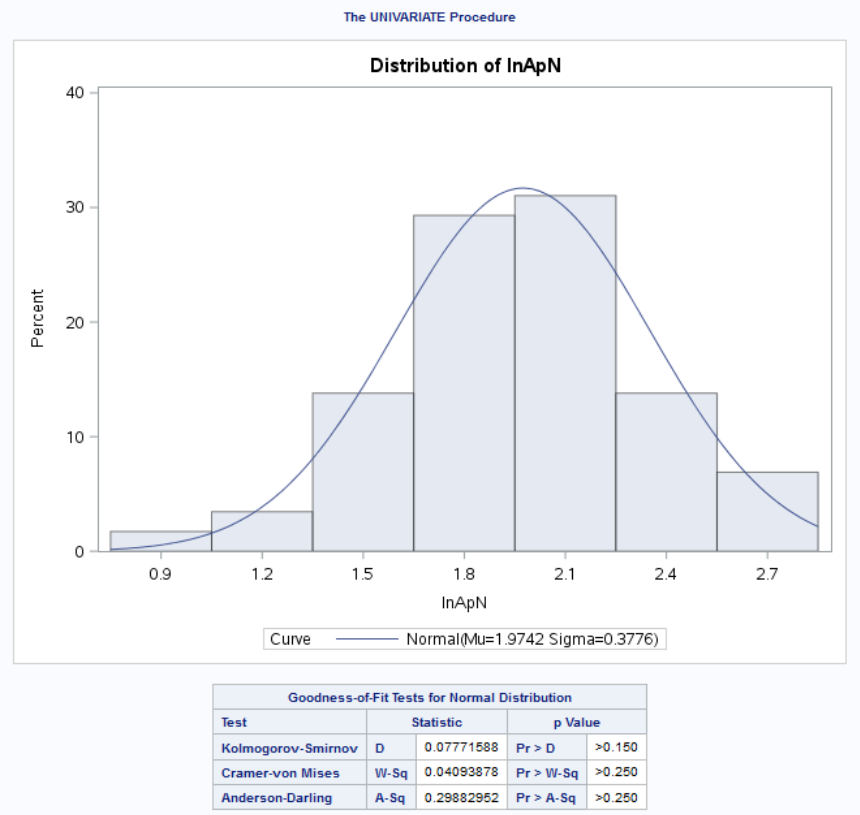
Na razini značajnosti od 5% možemo odbaciti nultu hipotezu o normalnosti adiponektina. Dakle, potrebno je transformirati adiponektin.

Primjenjena je sljedeća transformacija:

$$lnApN = ln(ApN)$$

Kod u SAS-u:

```
data novabaza;
    set baza;
    lnApN=log(ApN);
run;
proc univariate data=novabaza;
    var lnApN;
    histogram/normal;
run;
```



Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.07771588	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.04093878	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.29882952	Pr > A-Sq	>0.250

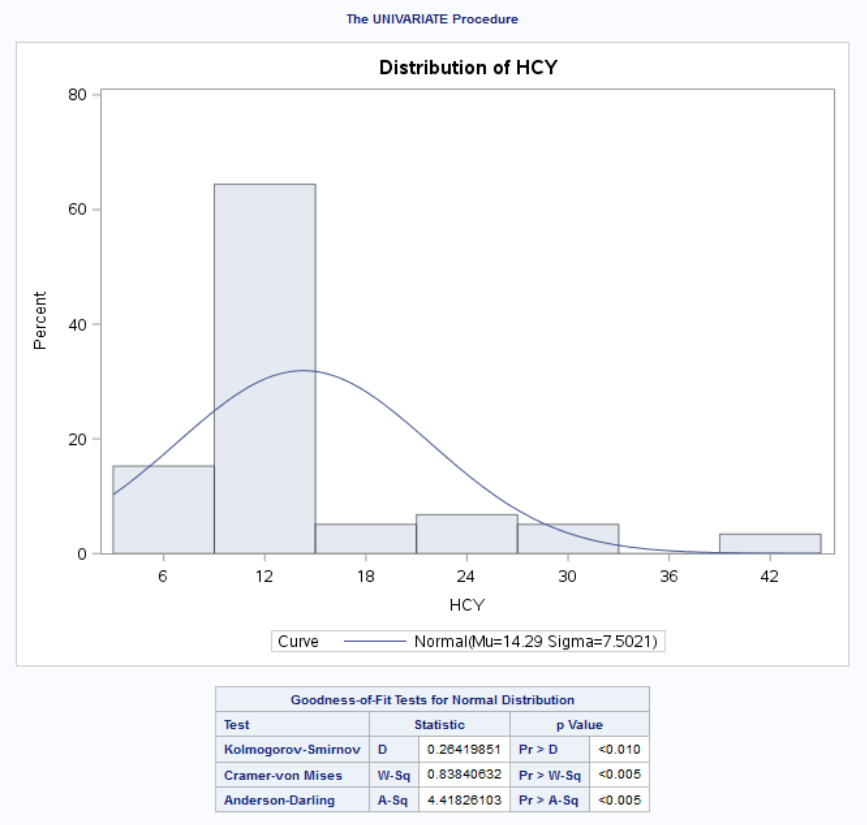
Slika 8.3: Histogram i testiranje normalnosti transformacije zavisne varijable ln(ApN) (ispis iz SAS-a)

Sada ne možemo odbaciti nultu hipotezu o normalnosti podataka transformiranog ApN.

8.3.2 Transformacija homocisteina, *HCY*

Kod u SAS-u:

```
proc univariate data=baza;
    var HCY;
    histogram/normal;
run;
```



Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.26419851	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.83840632	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	4.41826103	Pr > A-Sq	<0.005

Slika 8.4: Histogram i testiranje normalnosti zavisne varijable HCY (ispis iz SAS-a)

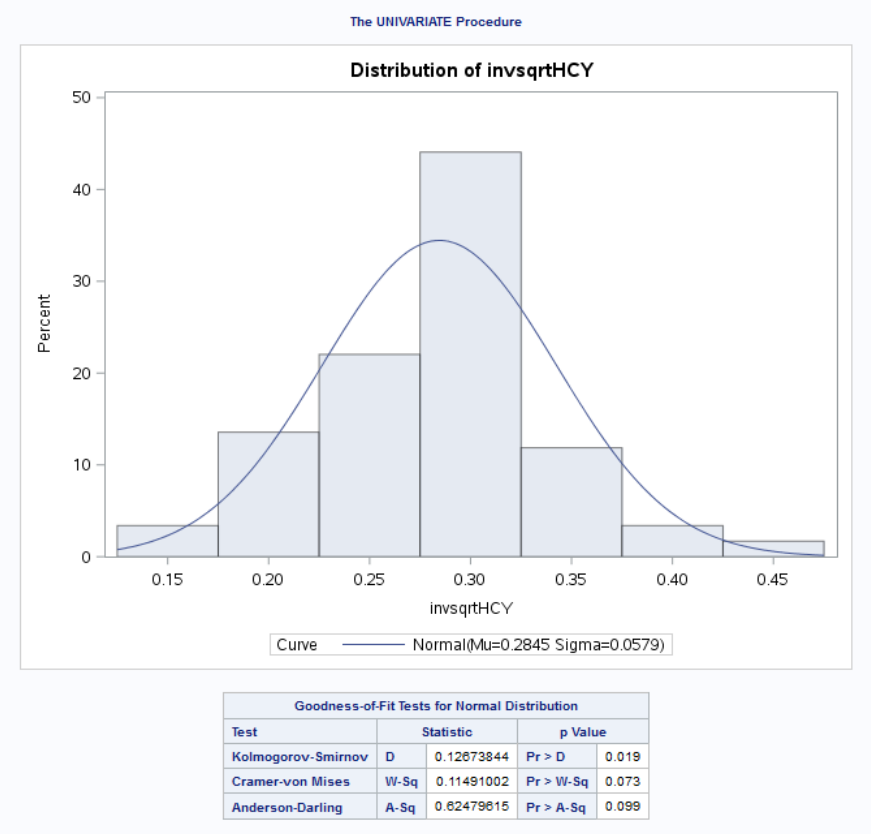
Na razini značajnosti od 5% možemo odbaciti nultu hipotezu o normalnosti homocisteina. Dakle, potrebno je transformirati homocistein. Primjenjena je sljedeća transformacija:

$$invsqrtHCY = \frac{1}{\sqrt{HCY}}$$

Kod u SAS-u:

```
data novabaza;
    set baza;
    lnApN=log(ApN);
    invsqrtHCY=1/sqrt(HCY);
run;

proc univariate data=novabaza;
    var invsqrtHCY;
    histogram/normal;
run;
```



Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.12673844	Pr > D	0.019
Cramer-von Mises	W-Sq	0.11491002	Pr > W-Sq	0.073
Anderson-Darling	A-Sq	0.62479615	Pr > A-Sq	0.099

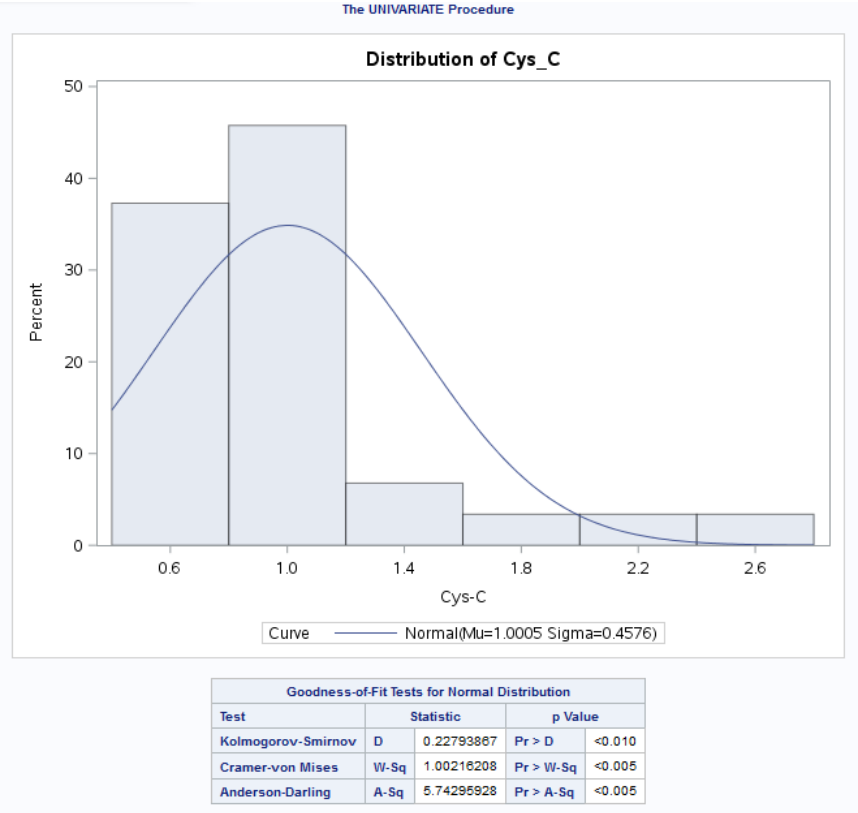
Slika 8.5: Histogram i testiranje normalnosti transformacije zavisne varijable $1/\sqrt{HCY}$ (ispis iz SAS-a)

Sada ne možemo odbaciti nultu hipotezu o normalnosti podataka transformiranog HCY.

8.3.3 Transformacija Cistacina C, *Cys_C*

Kod u SAS-u:

```
1  proc univariate data=baza;
2      var Cys_C;
3      histogram/normal;
4  run;
```



Slika 8.6: Histogram i testiranje normalnosti zavisne varijable *Cys_C* (ispis iz SAS-a)

Možemo odbaciti nultu hipotezu o normalnosti cistacina C. Dakle, potrebno je transformirati cistacin C. Primjenjena je sljedeća transformacija:

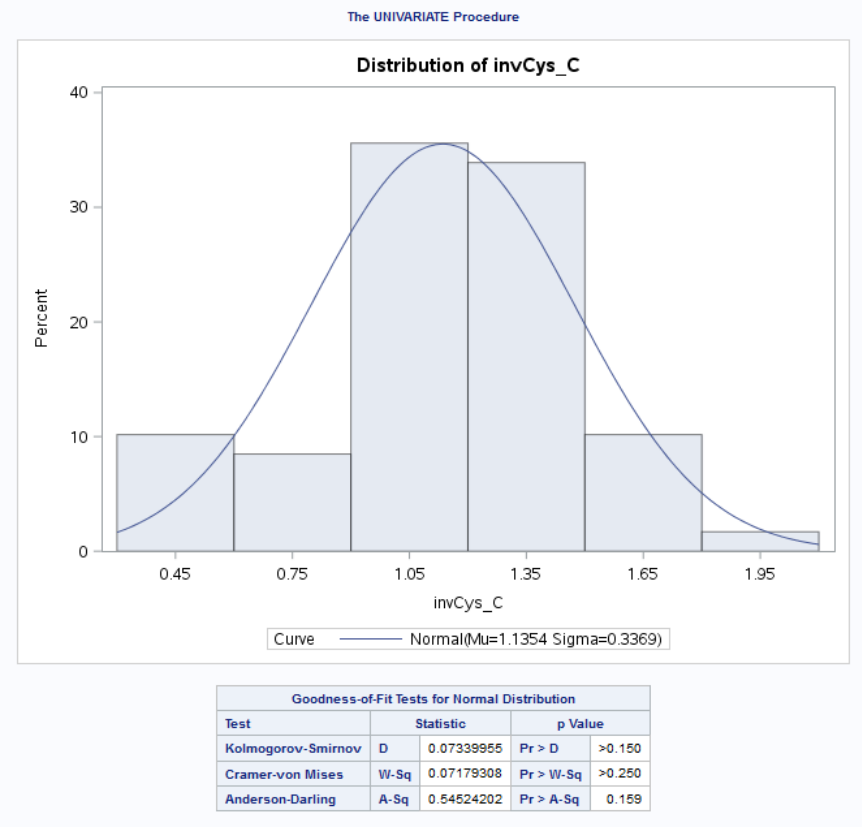
Primjenjena je sljedeća transformacija:

$$invCys\_C = \frac{1}{Cys\_C}$$

Kod u SAS-u:

```
data novabaza;
    set baza;
    lnApN=log(ApN);
    invsqrtHCY=1/sqrt(HCY);
    invCys_C=1/Cys_C;
run;

proc univariate data=novabaza;
    var invCys_C;
    histogram/normal;
run;
```



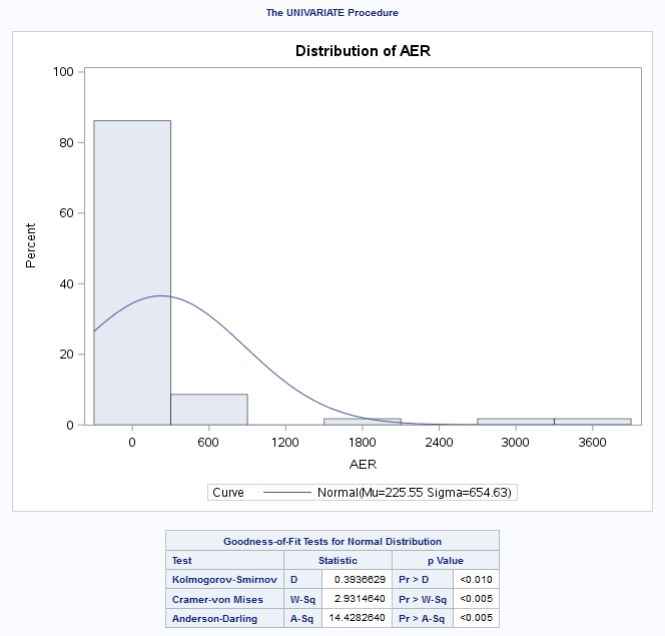
Slika 8.7: Histogram i testiranje normalnosti transformacije zavisne varijable$1/Cys_C$ (ispis iz SAS-a)

Sada ne možemo odbaciti nultu hipotezu o normalnosti podataka transformiranog Cys_C.

8.3.4 Transformacija ekskrecije albumina, *AER*

Kod u SAS-u:

```
1  proc univariate data=baza;
2      var AER;
3      histogram/normal;
4  run;
```



Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.3936629	Pr > D	<0.010
Cramer-von Mises	W-Sq	2.9314640	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	14.4282640	Pr > A-Sq	<0.005

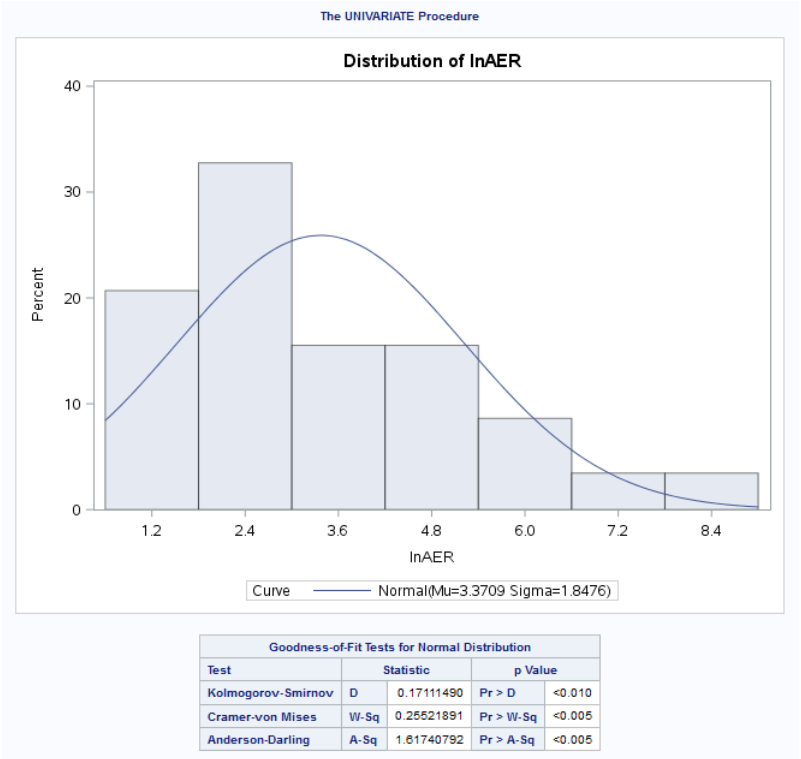
Slika 8.8: Histogram i testiranje normalnosti zavisne varijable *AER* (ispis iz SAS-a)

Možemo odbaciti nultu hipotezu o normalnosti ekskrecije albumina. Dakle, potrebno je transformirati ekskreciju albumina. Primjenjena je sljedeća transformacija:

$$lnAER = ln(AER)$$

Kod u SAS-u:

```
data novabaza;
    set baza;
    lnApN=log(ApN);
    invsqrtHCY=1/sqrt(HCY);
    invCys_C=1/Cys_C;
    lnAER=log(AER);
run;
proc univariate data=novabaza;
    var lnAER;
    histogram/normal;
run;
```



Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.17111490	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.25521891	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	1.61740792	Pr > A-Sq	<0.005

Slika 8.9: Histogram i testiranje normalnosti zavisne varijable $ln(AER)$ (ispis iz SAS-a)

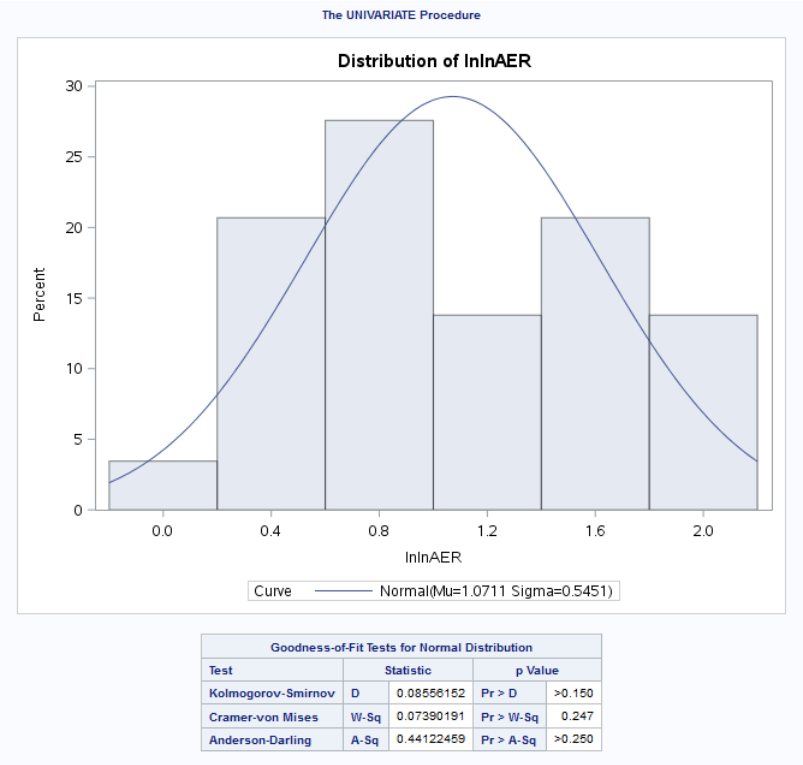
Nakon logaritimiranja ekskrecije albumina i dalje možemo odbaciti nultu hipotezu o normalnosti podataka. Primijenjena je sljedeća transformacija na logaritmirane podatke:

$$lnlnAER = ln(ln(AER))$$

Kod u SAS-u:

```
data novabaza;
    set baza;
    lnApN=log(ApN);
    invsqrtHCY=1/sqrt(HCY);
    invCys_C=1/Cys_C;
    lnlnAER=log(log(AER));
run;

proc univariate data=novabaza;
    var lnlnAER;
    histogram/normal;
run;
```



Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.08556152	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.07390191	Pr > W-Sq	0.247
Anderson-Darling	A-Sq	0.44122459	Pr > A-Sq	>0.250

Slika 8.10: Histogram i testiranje normalnosti transformacije zavisne varijable $ln(ln(AER))$ (ispis iz SAS-a)

Sada ne možemo odbaciti nultu hipotezu o normalnosti podataka transformiranog *AER*.

U sljedećoj tablici prikazane su vrijednosti deskriptivnih statistika zavisnih varijabli nakon provedenih transformacija.
Kod u SAS-u:

```
1  proc means data=novabaza N Min Median Max Mean Std Skew Kurt nolabels
     maxdec=3;
2        var lnApN invsqrtHCY invCys_C lnlnAER;
3    run;
```

Tablica 8.2: Deskriptivna statistika transformiranih zavisnih varijabli (ispis iz SAS-a)

The MEANS Procedure

Variable	N	Minimum	Median	Maximum	Mean	Std Dev	Skewness	Kurtosis
lnApN	58	0.824	1.986	2.680	1.974	0.378	-0.294	0.423
invsqrtHCY	59	0.153	0.286	0.452	0.285	0.058	0.013	0.685
invCys_C	59	0.386	1.176	2.000	1.135	0.337	-0.289	0.303
lnlnAER	58	-0.011	0.986	2.103	1.071	0.545	0.055	-0.926

8.3.5 Korelacije među varijablama

Pearsonov koeficijent korelacije ispituje koliko je jaka i kojeg je predznaka linearna veza među varijablama.

Pomoću njega možemo ispitati linearnu vezu zavisne varijable sa nekom nezavisnom varijablom u jednostavnoj linearnoj regresiji. Na taj način možemo naći nezavisnu varijablu koja najbolje opisuje ponašanje zavisne varijable.

Ovi zaključci se mogu proširiti i na donošenje odluka u višestrukoj linearnoj regresiji, ali samo ako su sve nezavisne varijable zaista nezavisne. Ako su nezavisne varijable blizu linearne zavisnosti, to može dovesti do problema multikolinearnosti u modelu. Naime, multikolinearnost unutar modela može imati velik utjecaj na procjenu regresijskih koeficijenata (npr. suprotan predznak, veća varijanca). Dobra strana je da se multikolinearnost može lako ispraviti, uglavnom izbacivanjem iz modele neke/ih od nezavisne/ih varijabli koje su korelirane. Sljedećim kodom u SAS-u smo ispitali Pearsonov koeficijent korelacije podataka u bazi:

Kod u SAS-u:

```
1  proc corr data=baza Pearson vardef=df;
2     var age duration waste WHR fBG ppBG HbA1c ApN CRP HCY FIB IL_6
3     Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST ALT GGT
4     MK C1 FLI GDR;
5  run;
```

Tablica 8.3: Pearsonov koeficijent korelacije i njegova značajnost (ispis iz SAS-a; 1/8)

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	age	duration	waste	WHR	fBG	ppBG	HbA1c
age age	1.00000 59	0.40833 0.0013 59	-0.08208 0.5590 53	0.03220 0.8138 56	-0.25047 0.0557 59	-0.35012 0.0071 58	-0.23078 0.0787 59
duration duration	0.40833 0.0013 59	1.00000 59	-0.04645 0.7412 53	0.06063 0.6571 56	0.08194 0.5373 59	0.06187 0.6445 58	0.18951 0.1506 59
waste waste	-0.08208 0.5590 53	-0.04645 0.7412 53	1.00000 53	0.18272 0.1903 53	0.30030 0.0289 53	0.38347 0.0050 52	0.46520 0.0004 53
WHR WHR	0.03220 0.8138 56	0.06063 0.6571 56	0.18272 0.1903 53	1.00000 56	-0.11127 0.4143 56	0.11535 0.4017 55	-0.00006 0.9997 56
fBG fBG	-0.25047 0.0557 59	0.08194 0.5373 59	0.30030 0.0289 53	-0.11127 0.4143 56	1.00000 59	0.83404 <.0001 58	0.74949 <.0001 59
ppBG ppBG	-0.35012 0.0071 58	0.06187 0.6445 58	0.38347 0.0050 52	0.11535 0.4017 55	0.83404 <.0001 58	1.00000 58	0.74143 <.0001 58
HbA1c HbA1c	-0.23078 0.0787 59	0.18951 0.1506 59	0.46520 0.0004 53	-0.00006 0.9997 56	0.74949 <.0001 59	0.74143 <.0001 58	1.00000 59
ApN ApN	0.43801 0.0006 58	0.02021 0.8803 58	-0.08112 0.5675 52	-0.19066 0.1632 55	-0.21954 0.0977 58	-0.30961 0.0191 57	-0.31893 0.0147 58
CRP CRP	-0.11952 0.3673 59	0.00860 0.9484 59	0.38098 0.0049 53	0.01648 0.9040 56	0.35368 0.0060 59	0.44591 0.0005 58	0.43623 0.0006 59
HCY HCY	0.39570 0.0019 59	0.08027 0.5456 59	0.04388 0.7550 53	0.24302 0.0711 56	-0.11772 0.3746 59	-0.14751 0.2691 58	-0.17520 0.1844 59
FIB FIB	0.04217 0.7512 59	0.22333 0.0891 59	0.16358 0.2419 53	0.19937 0.1407 56	0.25416 0.0521 59	0.28885 0.0279 58	0.28929 0.0263 59
IL_6 IL-6	0.09256 0.5015 55	0.08347 0.5446 55	0.13126 0.3687 49	0.20823 0.1385 52	0.11390 0.4077 55	0.23857 0.0823 54	0.29206 0.0305 55
Cys_C Cys-C	0.37448 0.0035 59	0.16852 0.2020 59	0.23124 0.0957 53	0.32475 0.0146 56	-0.07729 0.5607 59	0.00902 0.9464 58	-0.05841 0.6603 59
Sijal_ac_ Sijal-ac.	0.08873 0.5040 59	0.14707 0.2663 59	0.36494 0.0072 53	0.15531 0.2531 56	0.14422 0.2758 59	0.20722 0.1186 58	0.30913 0.0172 59
klirens klirens	-0.53995 <.0001 57	-0.10284 0.4465 57	0.23929 0.0908 51	-0.10596 0.4457 54	0.18624 0.1654 57	0.25509 0.0578 56	0.22775 0.0884 57

Tablica 8.4: Pearsonov koeficijent korelacije i njegova značajnost (ispis iz SAS-a; 2/8)

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	ApN	CRP	HCY	FIB	IL_6	Cys_C	Sijal_ac_	klirens
age age	0.43801 0.0006 58	-0.11952 0.3673 59	0.39570 0.0019 59	0.04217 0.7512 59	0.09256 0.5015 55	0.37448 0.0035 59	0.08873 0.5040 59	-0.53995 <.0001 57
duration duration	0.02021 0.8803 58	0.00860 0.9484 59	0.08027 0.5456 59	0.22333 0.0891 59	0.08347 0.5446 55	0.16852 0.2020 59	0.14707 0.2663 59	-0.10284 0.4465 57
waste waste	-0.08112 0.5675 52	0.38098 0.0049 53	0.04388 0.7550 53	0.16358 0.2419 53	0.13126 0.3687 49	0.23124 0.0957 53	0.36494 0.0072 53	0.23929 0.0908 51
WHR WHR	-0.19066 0.1632 55	0.01648 0.9040 56	0.24302 0.0711 56	0.19937 0.1407 56	0.20823 0.1385 52	0.32475 0.0146 56	0.15531 0.2531 56	-0.10596 0.4457 54
fBG fBG	-0.21954 0.0977 58	0.35368 0.0060 59	-0.11772 0.3746 59	0.25416 0.0521 59	0.11390 0.4077 55	-0.07729 0.5607 59	0.14422 0.2758 59	0.18624 0.1654 57
ppBG ppBG	-0.30961 0.0191 57	0.44591 0.0005 58	-0.14751 0.2691 58	0.28885 0.0279 58	0.23857 0.0823 54	0.00902 0.9464 58	0.20722 0.1186 58	0.25509 0.0578 56
HbA1c HbA1c	-0.31893 0.0147 58	0.43623 0.0006 59	-0.17520 0.1844 59	0.28929 0.0263 59	0.29206 0.0305 55	-0.05841 0.6603 59	0.30913 0.0172 59	0.22775 0.0884 57
ApN ApN	1.00000 58	-0.17927 0.1781 58	0.21167 0.1107 58	-0.05726 0.6694 58	-0.08562 0.5343 55	0.11846 0.3758 58	-0.02327 0.8624 58	-0.33519 0.0116 56
CRP CRP	-0.17927 0.1781 58	1.00000 59	0.09603 0.4694 59	0.51657 <.0001 59	0.62416 <.0001 55	0.36529 0.0044 59	0.46581 0.0002 59	-0.11571 0.3914 57
HCY HCY	0.21167 0.1107 58	0.09603 0.4694 59	1.00000 59	0.14630 0.2689 59	0.23011 0.0910 55	0.59526 <.0001 59	0.17378 0.1881 59	-0.56610 <.0001 57
FIB FIB	-0.05726 0.6694 58	0.51657 <.0001 59	0.14630 0.2689 59	1.00000 59	0.45936 0.0004 55	0.34223 0.0080 59	0.57969 <.0001 59	-0.15008 0.2652 57
IL_6 IL-6	-0.08562 0.5343 55	0.62416 <.0001 55	0.23011 0.0910 55	0.45936 0.0004 55	1.00000 55	0.41863 0.0015 55	0.38260 0.0039 55	-0.24043 0.0799 54
Cys_C Cys-C	0.11846 0.3758 58	0.36529 0.0044 59	0.59526 <.0001 59	0.34223 0.0080 59	0.41863 0.0015 55	1.00000 59	0.22943 0.0805 59	-0.56643 <.0001 57
Sijal_ac_ Sijal-ac.	-0.02327 0.8624 58	0.46581 0.0002 59	0.17378 0.1881 59	0.57969 <.0001 59	0.38260 0.0039 55	0.22943 0.0805 59	1.00000 59	-0.14276 0.2894 57
klirens klirens	-0.33519 0.0116 56	-0.11571 0.3914 57	-0.56610 <.0001 57	-0.15008 0.2652 57	-0.24043 0.0799 54	-0.56643 <.0001 57	-0.14276 0.2894 57	1.00000 57

Tablica 8.5: Pearsonov koeficijent korelacije i njegova značajnost (ispis iz SAS-a; 3/8)

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	AER	SBP	DBP	WBC	HDL	LDL	TG
age age	0.06313 0.6378 58	0.26123 0.0457 59	-0.14366 0.2777 59	-0.05876 0.6585 59	0.20001 0.1288 59	-0.19913 0.1305 59	-0.19584 0.1371 59
duration duration	0.25487 0.0535 58	0.29026 0.0257 59	-0.14048 0.2886 59	-0.04498 0.7351 59	0.25200 0.0542 59	-0.26719 0.0408 59	-0.08945 0.5005 59
waste waste	0.20820 0.1386 52	0.06281 0.6550 53	0.21404 0.1238 53	0.06287 0.6547 53	-0.29315 0.0331 53	0.04923 0.7263 53	0.07202 0.6083 53
WHR WHR	0.12859 0.3495 55	0.11376 0.4038 56	-0.18783 0.1657 56	0.09143 0.5028 56	-0.29011 0.0301 56	0.07531 0.5812 56	0.14515 0.2858 56
fBG fBG	0.07333 0.5844 58	0.12320 0.3526 59	0.22402 0.0881 59	-0.06823 0.6076 59	-0.18007 0.1723 59	-0.25884 0.0478 59	0.00616 0.9631 59
ppBG ppBG	0.16289 0.2260 57	0.08859 0.5084 58	0.20137 0.1296 58	-0.01307 0.9224 58	-0.22740 0.0860 58	-0.16053 0.2287 58	0.16873 0.2055 58
HbA1c HbA1c	0.12614 0.3454 58	0.12438 0.3479 59	0.18200 0.1677 59	0.09507 0.4738 59	-0.29057 0.0256 59	-0.12580 0.3424 59	0.07445 0.5752 59
ApN ApN	-0.09084 0.5016 57	-0.13247 0.3216 58	-0.20871 0.1159 58	-0.14725 0.2700 58	0.46454 0.0002 58	-0.08120 0.5445 58	-0.30129 0.0215 58
CRP CRP	0.17719 0.1833 58	0.03469 0.7942 59	0.11652 0.3795 59	0.19415 0.1406 59	-0.27088 0.0380 59	-0.15347 0.2458 59	0.20496 0.1194 59
HCY HCY	0.23935 0.0704 58	0.09023 0.4968 59	-0.03776 0.7764 59	0.10601 0.4242 59	-0.13366 0.3128 59	-0.03453 0.7952 59	0.09086 0.4937 59
FIB FIB	0.28175 0.0321 58	0.34057 0.0083 59	0.16349 0.2160 59	0.29053 0.0256 59	-0.22093 0.0927 59	-0.15040 0.2555 59	0.08713 0.5117 59
IL_6 IL-6	0.18735 0.1749 54	-0.00017 0.9990 55	-0.02507 0.8558 55	0.36923 0.0055 55	-0.33615 0.0121 55	-0.20488 0.1335 55	0.15072 0.2720 55
Cys_C Cys-C	0.54040 <.0001 58	-0.03698 0.7810 59	-0.20094 0.1270 59	0.19660 0.1356 59	-0.26940 0.0391 59	-0.00514 0.9692 59	0.24402 0.0625 59
Sijal_ac_ Sijal-ac.	0.33287 0.0107 58	0.33130 0.0104 59	0.31448 0.0153 59	0.29533 0.0232 59	-0.17130 0.1945 59	-0.04020 0.7624 59	0.32872 0.0110 59
klirens klirens	-0.09950 0.4615 57	0.02848 0.8334 57	0.16403 0.2227 57	-0.05339 0.6933 57	0.02092 0.8772 57	0.31325 0.0177 57	-0.00452 0.9734 57

Tablica 8.6: Pearsonov koeficijent korelacije i njegova značajnost (ispis iz SAS-a; 4/8)

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	AST	ALT	GGT	MK	C1	FLI	GDR
age age	-0.30689 0.0181 59	-0.56072 <.0001 59	-0.29237 0.0246 59	0.12674 0.3388 59	-0.37180 0.0044 57	-0.21732 0.1180 53	0.03636 0.7902 56
duration duration	-0.10857 0.4130 59	-0.18505 0.1606 59	-0.07557 0.5695 59	-0.08053 0.5443 59	-0.34974 0.0077 57	-0.12171 0.3853 53	-0.18000 0.1844 56
waste waste	0.25568 0.0646 53	0.29193 0.0339 53	0.09950 0.4784 53	0.14497 0.3003 53	0.33838 0.0141 52	0.73048 <.0001 53	-0.39380 0.0035 53
WHR WHR	-0.04641 0.7341 56	0.03489 0.7985 56	0.04807 0.7250 56	0.23033 0.0877 56	0.41470 0.0018 54	0.24617 0.0756 53	-0.22976 0.0885 56
fBG fBG	0.21024 0.1100 59	0.30873 0.0174 59	0.13739 0.2994 59	-0.12341 0.3517 59	0.29377 0.0266 57	0.35345 0.0094 53	-0.42422 0.0011 56
ppBG ppBG	0.18703 0.1598 58	0.35031 0.0070 58	0.22751 0.0859 58	-0.01388 0.9176 58	0.42149 0.0012 56	0.44094 0.0011 52	-0.42292 0.0013 55
HbA1c HbA1c	0.24021 0.0669 59	0.39030 0.0022 59	0.15328 0.2464 59	-0.11264 0.3957 59	0.26103 0.0499 57	0.49886 0.0001 53	-0.63287 <.0001 56
ApN ApN	-0.28689 0.0290 58	-0.44979 0.0004 58	-0.31552 0.0158 58	-0.01549 0.9082 58	-0.28014 0.0365 56	-0.34472 0.0123 52	0.32960 0.0140 55
CRP CRP	0.26974 0.0388 59	0.19957 0.1297 59	0.08750 0.5099 59	0.30959 0.0170 59	0.24403 0.0673 57	0.34386 0.0117 53	-0.23903 0.0760 56
HCY HCY	-0.06256 0.6378 59	-0.24859 0.0576 59	-0.07654 0.5645 59	0.40268 0.0016 59	0.12360 0.3597 57	0.07383 0.5993 53	0.05836 0.6692 56
FIB FIB	0.09835 0.4587 59	0.07147 0.5907 59	-0.01369 0.9181 59	0.36871 0.0041 59	0.12007 0.3736 57	0.21616 0.1201 53	-0.31552 0.0178 56
IL_6 IL-6	-0.00897 0.9482 55	-0.01710 0.9014 55	-0.03171 0.8182 55	0.26376 0.0517 55	0.32354 0.0181 53	0.22552 0.1192 49	-0.10363 0.4647 52
Cys_C Cys-C	-0.15818 0.2315 59	-0.35359 0.0060 59	-0.16752 0.2047 59	0.60032 <.0001 59	0.14759 0.2733 57	0.17732 0.2040 53	0.06371 0.6409 56
Sijal_ac_ Sijal-ac.	-0.09936 0.4540 59	-0.00628 0.9624 59	-0.18772 0.1545 59	0.28861 0.0266 59	0.08095 0.5494 57	0.35181 0.0098 53	-0.32128 0.0158 56
klirens klirens	0.13573 0.3141 57	0.33770 0.0102 57	0.28468 0.0319 57	-0.31833 0.0158 57	0.01446 0.9166 55	0.22578 0.1112 51	-0.16039 0.2466 54

Tablica 8.7: Pearsonov koeficijent korelacije i njegova značajnost (ispis iz SAS-a; 5/8)

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations							
	age	duration	waste	WHR	fBG	ppBG	HbA1c
AER AER	0.06313 0.6378 58	0.25487 0.0535 58	0.20820 0.1386 52	0.12859 0.3495 55	0.07333 0.5844 58	0.16289 0.2260 57	0.12614 0.3454 58
SBP SBP	0.26123 0.0457 59	0.29026 0.0257 59	0.06281 0.6550 53	0.11376 0.4038 56	0.12320 0.3526 59	0.08859 0.5084 58	0.12438 0.3479 59
DBP DBP	-0.14366 0.2777 59	-0.14048 0.2886 59	0.21404 0.1238 53	-0.18783 0.1657 56	0.22402 0.0881 59	0.20137 0.1296 58	0.18200 0.1677 59
WBC WBC	-0.05876 0.6585 59	-0.04498 0.7351 59	0.06287 0.6547 53	0.09143 0.5028 56	-0.06823 0.6076 59	-0.01307 0.9224 58	0.09507 0.4738 59
HDL HDL	0.20001 0.1288 59	0.25200 0.0542 59	-0.29315 0.0331 53	-0.29011 0.0301 56	-0.18007 0.1723 59	-0.22740 0.0860 58	-0.29057 0.0256 59
LDL LDL	-0.19913 0.1305 59	-0.26719 0.0408 59	0.04923 0.7263 53	0.07531 0.5812 56	-0.25884 0.0478 59	-0.16053 0.2287 58	-0.12580 0.3424 59
TG TG	-0.19584 0.1371 59	-0.08945 0.5005 59	0.07202 0.6083 53	0.14515 0.2858 56	0.00616 0.9631 59	0.16873 0.2055 58	0.07445 0.5752 59
AST AST	-0.30689 0.0181 59	-0.10857 0.4130 59	0.25568 0.0646 53	-0.04641 0.7341 56	0.21024 0.1100 59	0.18703 0.1598 58	0.24021 0.0669 59
ALT ALT	-0.56072 <.0001 59	-0.18505 0.1606 59	0.29193 0.0339 53	0.03489 0.7985 56	0.30873 0.0174 59	0.35031 0.0070 58	0.39030 0.0022 59
GGT GGT	-0.29237 0.0246 59	-0.07557 0.5695 59	0.09950 0.4784 53	0.04807 0.7250 56	0.13739 0.2994 59	0.22751 0.0859 58	0.15328 0.2464 59
MK MK	0.12674 0.3388 59	-0.08053 0.5443 59	0.14497 0.3003 53	0.23033 0.0877 56	-0.12341 0.3517 59	-0.01388 0.9176 58	-0.11264 0.3957 59
C1 C1	-0.37180 0.0044 57	-0.34974 0.0077 57	0.33838 0.0141 52	0.41470 0.0018 54	0.29377 0.0266 57	0.42149 0.0012 56	0.26103 0.0499 57
FLI FLI	-0.21732 0.1180 53	-0.12171 0.3853 53	0.73048 <.0001 53	0.24617 0.0756 53	0.35345 0.0094 53	0.44094 0.0011 52	0.49886 0.0001 53
GDR GDR	0.03636 0.7902 56	-0.18000 0.1844 56	-0.39380 0.0035 53	-0.22976 0.0885 56	-0.42422 0.0011 56	-0.42292 0.0013 55	-0.63287 <.0001 56

Tablica 8.8: Pearsonov koeficijent korelacije i njegova značajnost (ispis iz SAS-a; 6/8)

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	ApN	CRP	HCY	FIB	IL_6	Cys_C	Sijal_ac_	klirens
AER AER	-0.09084 0.5016 57	0.17719 0.1833 58	0.23935 0.0704 58	0.28175 0.0321 58	0.18735 0.1749 54	0.54040 <.0001 58	0.33287 0.0107 58	-0.09950 0.4615 57
SBP SBP	-0.13247 0.3216 58	0.03469 0.7942 59	0.09023 0.4968 59	0.34057 0.0083 59	-0.00017 0.9990 55	-0.03698 0.7810 59	0.33130 0.0104 59	0.02848 0.8334 57
DBP DBP	-0.20871 0.1159 58	0.11652 0.3795 59	-0.03776 0.7764 59	0.16349 0.2160 59	-0.02507 0.8558 55	-0.20094 0.1270 59	0.31448 0.0153 59	0.16403 0.2227 57
WBC WBC	-0.14725 0.2700 58	0.19415 0.1406 59	0.10601 0.4242 59	0.29053 0.0256 59	0.36923 0.0055 55	0.19660 0.1356 59	0.29533 0.0232 59	-0.05339 0.6933 57
HDL HDL	0.46454 0.0002 58	-0.27088 0.0380 59	-0.13366 0.3128 59	-0.22093 0.0927 59	-0.33615 0.0121 55	-0.26940 0.0391 59	-0.17130 0.1945 59	0.02092 0.8772 57
LDL LDL	-0.08120 0.5445 58	-0.15347 0.2458 59	-0.03453 0.7952 59	-0.15040 0.2555 59	-0.20488 0.1335 55	-0.00514 0.9692 59	-0.04020 0.7624 59	0.31325 0.0177 57
TG TG	-0.30129 0.0215 58	0.20496 0.1194 59	0.09086 0.4937 59	0.08713 0.5117 59	0.15072 0.2720 55	0.24402 0.0625 59	0.32872 0.0110 59	-0.00452 0.9734 57
AST AST	-0.28689 0.0290 58	0.26974 0.0388 59	-0.06256 0.6378 59	0.09835 0.4587 59	-0.00897 0.9482 55	-0.15818 0.2315 59	-0.09936 0.4540 59	0.13573 0.3141 57
ALT ALT	-0.44979 0.0004 58	0.19957 0.1297 59	-0.24859 0.0576 59	0.07147 0.5907 59	-0.01710 0.9014 55	-0.35359 0.0060 59	-0.00628 0.9624 59	0.33770 0.0102 57
GGT GGT	-0.31552 0.0158 58	0.08750 0.5099 59	-0.07654 0.5645 59	-0.01369 0.9181 59	-0.03171 0.8182 55	-0.16752 0.2047 59	-0.18772 0.1545 59	0.28468 0.0319 57
MK MK	-0.01549 0.9082 58	0.30959 0.0170 59	0.40268 0.0016 59	0.36871 0.0041 59	0.26376 0.0517 55	0.60032 <.0001 59	0.28861 0.0266 59	-0.31833 0.0158 57
C1 C1	-0.28014 0.0365 56	0.24403 0.0673 57	0.12360 0.3597 57	0.12007 0.3736 57	0.32354 0.0181 53	0.14759 0.2733 57	0.08095 0.5494 57	0.01446 0.9166 55
FLI FLI	-0.34472 0.0123 52	0.34386 0.0117 53	0.07383 0.5993 53	0.21616 0.1201 53	0.22552 0.1192 49	0.17732 0.2040 53	0.35181 0.0098 53	0.22578 0.1112 51
GDR GDR	0.32960 0.0140 55	-0.23903 0.0760 56	0.05836 0.6692 56	-0.31552 0.0178 56	-0.10363 0.4647 52	0.06371 0.6409 56	-0.32128 0.0158 56	-0.16039 0.2466 54

Tablica 8.9: Pearsonov koeficijent korelacije i njegova značajnost (ispis iz SAS-a; 7/8)

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	AER	SBP	DBP	WBC	HDL	LDL	TG
AER	1.00000	0.11377	0.07296	0.19799	-0.14529	0.07951	0.39273
AER		0.3951	0.5863	0.1363	0.2765	0.5530	0.0023
	58	58	58	58	58	58	58
SBP	0.11377	1.00000	0.56635	-0.07860	0.07313	0.16197	-0.00085
SBP	0.3951		<.0001	0.5540	0.5820	0.2204	0.9949
	58	59	59	59	59	59	59
DBP	0.07296	0.56635	1.00000	-0.03499	-0.06926	0.07384	0.14667
DBP	0.5863	<.0001		0.7925	0.6022	0.5783	0.2676
	58	59	59	59	59	59	59
WBC	0.19799	-0.07860	-0.03499	1.00000	-0.26954	0.08251	0.15402
WBC	0.1363	0.5540	0.7925		0.0390	0.5344	0.2441
	58	59	59	59	59	59	59
HDL	-0.14529	0.07313	-0.06926	-0.26954	1.00000	-0.00098	-0.47074
HDL	0.2765	0.5820	0.6022	0.0390		0.9941	0.0002
	58	59	59	59	59	59	59
LDL	0.07951	0.16197	0.07384	0.08251	-0.00098	1.00000	0.28854
LDL	0.5530	0.2204	0.5783	0.5344	0.9941		0.0267
	58	59	59	59	59	59	59
TG	0.39273	-0.00085	0.14667	0.15402	-0.47074	0.28854	1.00000
TG	0.0023	0.9949	0.2676	0.2441	0.0002	0.0267	
	58	59	59	59	59	59	59
AST	-0.21768	0.03351	0.29981	0.08506	-0.25367	-0.05414	0.15827
AST	0.1007	0.8011	0.0211	0.5218	0.0525	0.6838	0.2312
	58	59	59	59	59	59	59
ALT	-0.19220	-0.02100	0.30385	0.09375	-0.31052	-0.00869	0.22931
ALT	0.1483	0.8746	0.0193	0.4800	0.0167	0.9479	0.0806
	58	59	59	59	59	59	59
GGT	-0.03275	-0.07324	0.16361	0.00096	-0.09685	0.01135	0.19259
GGT	0.8072	0.5815	0.2157	0.9942	0.4655	0.9320	0.1439
	58	59	59	59	59	59	59
MK	0.25175	0.04869	-0.06168	0.23261	-0.21343	0.23845	0.25197
MK	0.0566	0.7142	0.6426	0.0762	0.1046	0.0690	0.0542
	58	59	59	59	59	59	59
C1	-0.02667	-0.00124	0.10049	0.03260	-0.55592	0.09486	0.34650
C1	0.8453	0.9927	0.4570	0.8098	<.0001	0.4827	0.0083
	56	57	57	57	57	57	57
FLI	0.17781	0.14163	0.34782	0.14385	-0.51339	0.18873	0.43170
FLI	0.2073	0.3117	0.0107	0.3041	<.0001	0.1759	0.0012
	52	53	53	53	53	53	53
GDR	-0.16401	-0.47381	-0.42301	-0.13880	0.21953	-0.04124	-0.14977
GDR	0.2315	0.0002	0.0012	0.3076	0.1040	0.7628	0.2706
	55	56	56	56	56	56	56

Tablica 8.10: Pearsonov koeficijent korelacije i njegova značajnost (ispis iz SAS-a; 8/8)

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	AST	ALT	GGT	MK	C1	FLI	GDR
AER AER	-0.21768 0.1007 58	-0.19220 0.1483 58	-0.03275 0.8072 58	0.25175 0.0566 58	-0.02667 0.8453 56	0.17781 0.2073 52	-0.16401 0.2315 55
SBP SBP	0.03351 0.8011 59	-0.02100 0.8746 59	-0.07324 0.5815 59	0.04869 0.7142 59	-0.00124 0.9927 57	0.14163 0.3117 53	-0.47381 0.0002 56
DBP DBP	0.29981 0.0211 59	0.30385 0.0193 59	0.16361 0.2157 59	-0.06168 0.6426 59	0.10049 0.4570 57	0.34782 0.0107 53	-0.42301 0.0012 56
WBC WBC	0.08506 0.5218 59	0.09375 0.4800 59	0.00096 0.9942 59	0.23261 0.0762 59	0.03260 0.8098 57	0.14385 0.3041 53	-0.13880 0.3076 56
HDL HDL	-0.25367 0.0525 59	-0.31052 0.0167 59	-0.09685 0.4655 59	-0.21343 0.1046 59	-0.55592 <.0001 57	-0.51339 <.0001 53	0.21953 0.1040 56
LDL LDL	-0.05414 0.6838 59	-0.00869 0.9479 59	0.01135 0.9320 59	0.23845 0.0690 59	0.09486 0.4827 57	0.18873 0.1759 53	-0.04124 0.7628 56
TG TG	0.15827 0.2312 59	0.22931 0.0806 59	0.19259 0.1439 59	0.25197 0.0542 59	0.34650 0.0083 57	0.43170 0.0012 53	-0.14977 0.2706 56
AST AST	1.00000 59	0.84051 <.0001 59	0.53685 <.0001 59	-0.02950 0.8245 59	0.37627 0.0039 57	0.36822 0.0067 53	-0.37603 0.0043 56
ALT ALT	0.84051 <.0001 59	1.00000 59	0.51235 <.0001 59	-0.10488 0.4292 59	0.44459 0.0005 57	0.44053 0.0010 53	-0.46125 0.0003 56
GGT GGT	0.53685 <.0001 59	0.51235 <.0001 59	1.00000 59	-0.09282 0.4844 59	0.05746 0.6711 57	0.39353 0.0036 53	-0.25389 0.0590 56
MK MK	-0.02950 0.8245 59	-0.10488 0.4292 59	-0.09282 0.4844 59	1.00000 59	0.16256 0.2270 57	0.13650 0.3298 53	0.04826 0.7239 56
C1 C1	0.37627 0.0039 57	0.44459 0.0005 57	0.05746 0.6711 57	0.16256 0.2270 57	1.00000 57	0.43614 0.0012 52	-0.27424 0.0448 54
FLI FLI	0.36822 0.0067 53	0.44053 0.0010 53	0.39353 0.0036 53	0.13650 0.3298 53	0.43614 0.0012 52	1.00000 53	-0.48723 0.0002 53
GDR GDR	-0.37603 0.0043 56	-0.46125 0.0003 56	-0.25389 0.0590 56	0.04826 0.7239 56	-0.27424 0.0448 54	-0.48723 0.0002 53	1.00000 56

8.4 Jednostruka linearna regresija

Jednostrukom linearnom regresijom modeliramo vezu između jedne zavisne varijable i jedne nezavisne varijable. Na taj način možemo vidjeti koje su varijable značajne u objašnjavanju zavisne varijable. Općeniti kod za modeliranje jednostrukog linearnog regresijskog modela je vrlo jednostavan i glasi:

Kod u SAS-u:

```
1  proc reg data=novabaza;
2      model zavisna_varijabla=nezavisna_varijabla;
3   run;
```

U našem slučaju nezavisne varijable su transformirani adiponektin ($ln(ApN)$), homocistein ($1/\sqrt{HCY}$), cistacin C ($1/Cys\_C$) i eksrecija albumina ($ln(ln(AER))$). Jednom kada fiksiramo zavisnu varijablu, gledamo njezinu vezu sa ostalim varijablama. Rezultati su dani u sljedećim tablicama.

8.4.1 Jednostruka linearna regresija za transformirani adiponektin

Tablica 8.11: Jednostruka linearna regresija za zavisnu varijablu $ln(ApN)$ (rezultati iz SAS-a)

				ANOVA		PROCJENA PARAMETARA				
naziv	**N**	R^2	R^2_{adj}	**F**	**p**	**Sl. član**	**nagib**	**SE**	**t**	**p**
sex	**58**	**0,1107**	**0,0949**	**7,0**	**0,0107**	**2,38**	**-0,253**	**0,096**	**-2,6**	**0,0107**
age	**58**	**0,2122**	**0,1981**	**15,1**	**0,0003**	**1,02**	**0,016**	**0,004**	**3,9**	**0,0003**
duration	58	0,0010	-0,0169	0,1	0,8160	1,96	0,002	0,007	0,2	0,8160
waste	52	0,0034	-0,0165	0,2	0,6803	2,15	-0,002	0,004	-0,4	0,6803
WHR	55	0,0187	0,0002	1,0	0,3197	2,32	-0,369	0,368	-1,0	0,3197
fBg	58	0,0560	0,0392	3,3	0,0737	2,22	-0,026	0,014	-1,8	0,0737
ppBG	**57**	**0,0947**	**0,0782**	**5,8**	**0,0199**	**2,30**	**-0,028**	**0,012**	**-2,4**	**0,0199**
HbA1c	**58**	**0,1222**	**0,1066**	**7,8**	**0,0071**	**2,60**	**-0,086**	**0,031**	**-2,8**	**0,0071**
CRP	58	0,0350	0,0177	2,0	0,1599	2,02	-0,012	0,008	-1,4	0,1599
HCY	58	0,0352	0,0179	2,0	0,1586	1,84	0,009	0,007	1,4	0,1586
FIB	58	0,0123	-0,0054	0,7	0,4076	2,17	-0,053	0,063	-0,8	0,4076
IL_6	55	0,0143	-0,0043	0,8	0,3845	2,03	-0,010	0,012	-0,9	0,3845
Cys_C	58	0,0162	-0,0014	0,9	0,3410	1,87	0,104	0,108	1,0	0,3410
Sijal_ac_	58	0,0017	-0,0162	0,1	0,7615	2,08	-0,050	0,164	-0,3	0,7615
klirens	**56**	**0,1178**	**0,1015**	**7,2**	**0,0096**	**2,28**	**-0,174**	**0,065**	**-2,7**	**0,0096**
AER	57	0,0088	-0,0092	0,5	0,4870	1,98	0,000	0,000	-0,7	0,4870
SBP	58	0,0195	0,0020	1,1	0,2963	2,29	-0,002	0,002	-1,1	0,2963
DBP	58	0,0505	0,0336	3,0	0,0898	2,66	-0,008	0,005	-1,7	0,0898
WBC	58	0,0233	0,0059	1,3	0,2524	2,24	-0,035	0,031	-1,2	0,2524
HDL	**58**	**0,2397**	**0,2262**	**17,7**	< .0001	**1,32**	**0,496**	**0,118**	**4,2**	< .0001
LDL	58	0,0083	-0,0094	0,5	0,4957	2,08	-0,036	0,053	-0,7	0,4957
TG	**58**	**0,0787**	**0,0622**	**4,8**	**0,0330**	**2,09**	**-0,043**	**0,019**	**-2,2**	**0,0330**
AST	**58**	**0,0761**	**0,0596**	**4,6**	**0,0360**	**2,26**	**-0,011**	**0,005**	**-2,2**	**0,0360**
ALT	**58**	**0,1904**	**0,1759**	**13,2**	**0,0006**	**2,30**	**-0,012**	**0,003**	**-3,6**	**0,0006**
GGT	**58**	**0,0990**	**0,0829**	**6,2**	**0,0162**	**2,11**	**-0,003**	**0,001**	**-2,5**	**0,0162**
MK	58	0,0030	-0,0148	0,2	0,6835	2,05	0,000	0,001	-0,4	0,6835
C1	**56**	**0,0714**	**0,0542**	**4,2**	**0,0466**	**2,14**	**-0,231**	**0,113**	**-2,0**	**0,0466**
FLI	**52**	**0,1320**	**0,1147**	**7,6**	**0,0081**	**2,39**	**-0,006**	**0,002**	**-2,8**	**0,0081**
GDR	**55**	**0,0889**	**0,0717**	**5,2**	**0,0270**	**1,65**	**0,052**	**0,023**	**2,3**	**0,0270**

Uočavamo da su na razini značajnosti od 5% značajni sex (spol), age (dob), ppBG, HbA1c, klirens, HDL, TG, AST, ALT, GGT, C1, FLI i GDR.

8.4.2 Jednostruka linearna regresija za transformirani homocistein

Tablica 8.12: Jednostruka linearna regresija za zavisnu varijablu $1/\sqrt{HCY}$ (rezultati iz SAS-a)

				ANOVA		PROCJENA PARAMETARA				
naziv	**N**	R^2	R^2_{adj}	**F**	**p**	**Sl. član**	**nagib**	**SE**	**t**	**p**
sex	59	0,0000	-0,0175	0,0	0,9973	0,28	0,000	0,015	0,0	0,9973
age	**59**	**0,1987**	**0,1847**	**14,1**	**0,0004**	**0,43**	**-0,002**	**0,001**	**-3,8**	**0,0004**
duration	59	0,0067	-0,0108	0,4	0,5384	0,29	-0,001	0,001	-0,6	0,5384
waste	53	0,0218	0,0026	1,1	0,2914	0,35	-0,001	0,001	-1,1	0,2914
WHR	56	0,0450	0,0273	2,6	0,1164	0,37	-0,089	0,056	-1,6	0,1164
fBg	59	0,0209	0,0037	1,2	0,2749	0,26	0,002	0,002	1,1	0,2749
ppBG	58	0,0280	0,0106	1,6	0,2094	0,26	0,002	0,002	1,3	0,2094
HbA1c	59	0,0406	0,0238	2,4	0,1257	0,23	0,008	0,005	1,6	0,1257
ApN	58	0,0225	0,0051	1,3	0,2608	0,31	-0,003	0,003	-1,1	0,2608
CRP	59	0,0066	-0,0108	0,4	0,5400	0,29	-0,001	0,001	-0,6	0,5400
FIB	59	0,0184	0,0012	1,1	0,3056	0,32	-0,010	0,010	-1,0	0,3056
IL_6	55	0,0308	0,0125	1,7	0,1999	0,30	-0,002	0,002	-1,3	0,1999
Cys_C	**59**	**0,3519**	**0,3405**	**31,0**	< .0001	**0,36**	**-0,075**	**0,013**	**-5,6**	< .0001
Sijal_ac_	59	0,0280	0,0109	1,6	0,2053	0,35	-0,032	0,025	-1,3	0,2053
klirens	**57**	**0,2829**	**0,2698**	**21,7**	< .0001	**0,21**	**0,041**	**0,009**	**4,7**	< .0001
AER	58	0,0592	0,0424	3,5	0,0657	0,29	0,000	0,000	-1,9	0,0657
SBP	59	0,0188	0,0016	1,1	0,3006	0,33	0,000	0,000	-1,0	0,3006
DBP	59	0,0000	-0,0175	0,0	0,9906	0,29	0,000	0,001	0,0	0,9906
WBC	59	0,0188	0,0016	1,1	0,3000	0,32	-0,004	0,004	-1,1	0,3000
HDL	59	0,0192	0,0020	1,1	0,2957	0,26	0,022	0,021	1,1	0,2957
LDL	59	0,0000	-0,0175	0,0	0,9714	0,29	0,000	0,008	0,0	0,9714
TG	59	0,0095	-0,0079	0,6	0,4631	0,29	-0,002	0,003	-0,7	0,4631
AST	59	0,0007	-0,0169	0,0	0,8444	0,28	0,000	0,001	0,2	0,8444
ALT	59	0,0508	0,0341	3,1	0,0861	0,26	0,001	0,001	1,8	0,0861
GGT	59	0,0058	-0,0116	0,3	0,5667	0,28	0,000	0,000	0,6	0,5667
MK	**59**	**0,2193**	**0,2057**	**16,0**	**0,0002**	**0,39**	**0,000**	**0,000**	**-4,0**	**0,0002**
C1	57	0,0074	-0,0106	0,4	0,5245	0,29	-0,011	0,017	-0,6	0,5245
FLI	53	0,0271	0,0080	1,4	0,2389	0,31	0,000	0,000	-1,2	0,2389
GDR	56	0,0017	-0,0168	0,1	0,7616	0,29	-0,001	0,004	-0,3	0,7616

Uočavamo da su na razini značajnosti od 5% značajni age (dob), IL_6, klirens i MK.

8.4.3 Jednostruka linearna regresija za transformirani cistacin C

Tablica 8.13: Jednostruka linearna regresija za zavisnu varijablu $1/Cys\_C$ (rezultati iz SAS-a)

				ANOVA		PROCJENA PARAMETARA				
naziv	**N**	R^2	R^2_{adj}	**F**	**p**	**Sl. član**	**nagib**	**SE**	**t**	**p**
sex	59	0,0060	-0,0114	0,3	0,5596	1,05	0,053	0,090	0,6	0,5596
age	**59**	**0,2209**	**0,2072**	**16,2**	**0,0002**	**2,02**	**-0,015**	**0,004**	**-4,0**	**0,0002**
duration	59	0,0107	-0,0066	0,6	0,4348	1,19	-0,005	0,006	-0,8	0,4348
waste	53	0,0716	0,0533	3,9	0,0528	1,85	-0,007	0,003	-2,0	0,0528
WHR	56	0,0676	0,0503	3,9	0,0530	1,74	-0,625	0,316	-2,0	0,0530
fBg	59	0,0117	-0,0056	0,7	0,4148	1,03	0,010	0,013	0,8	0,4148
ppBG	58	0,0046	-0,0132	0,3	0,6129	1,07	0,005	0,011	0,5	0,6129
HbA1c	59	0,0020	-0,0155	0,1	0,7366	1,06	0,010	0,029	0,3	0,7366
ApN	58	0,0214	0,0039	1,2	0,2737	1,27	-0,017	0,016	-1,1	0,2737
CRP	**59**	**0,1034**	**0,0877**	**6,6**	**0,0130**	**1,21**	**-0,018**	**0,007**	**-2,6**	**0,0130**
HCY	**59**	**0,3499**	**0,3385**	**30,7**	< .0001	**1,52**	**-0,027**	**0,005**	**-5,5**	< .0001
FIB	**59**	**0,0790**	**0,0629**	**4,9**	**0,0310**	**1,59**	**-0,119**	**0,054**	**-2,2**	**0,0310**
IL_6	**55**	**0,1577**	**0,1419**	**9,9**	**0,0027**	**1,26**	**-0,031**	**0,010**	**-3,2**	**0,0027**
Sijal_ac_	59	0,0553	0,0387	3,3	0,0731	1,67	-0,260	0,142	-1,8	0,0731
klirens	**57**	**0,2614**	**0,2479**	**19,5**	< .0001	**0,71**	**0,225**	**0,051**	**4,4**	< .0001
AER	**58**	**0,1709**	**0,1561**	**11,5**	**0,0013**	**1,17**	**0,000**	**0,000**	**-3,4**	**0,0013**
SBP	59	0,0055	-0,0119	0,3	0,5755	1,29	-0,001	0,002	-0,6	0,5755
DBP	59	0,0080	-0,0094	0,5	0,5010	0,89	0,003	0,004	0,7	0,5010
WBC	59	0,0511	0,0345	3,1	0,0851	1,46	-0,041	0,024	-1,8	0,0851
HDL	**59**	**0,0733**	**0,0571**	**4,5**	**0,0381**	**0,81**	**0,247**	**0,116**	**2,1**	**0,0381**
LDL	59	0,0000	-0,0175	0,0	0,9943	1,13	0,000	0,047	0,0	0,9943
TG	59	0,0223	0,0051	1,3	0,2592	1,19	-0,020	0,018	-1,1	0,2592
AST	59	0,0159	-0,0013	0,9	0,3409	1,02	0,005	0,005	1,0	0,3409
ALT	**59**	**0,1299**	**0,1146**	**8,5**	**0,0051**	**0,90**	**0,009**	**0,003**	**2,9**	**0,0051**
GGT	59	0,0616	0,0451	3,7	0,0581	1,04	0,002	0,001	1,9	0,0581
MK	**59**	**0,3085**	**0,2964**	**25,4**	< .0001	**1,84**	**-0,002**	**0,000**	**-5,0**	< .0001
C1	57	0,0164	-0,0014	0,9	0,3419	1,20	-0,098	0,103	-1,0	0,3419
FLI	53	0,0313	0,0123	1,7	0,2052	1,31	-0,002	0,002	-1,3	0,2052
GDR	56	0,0019	-0,0166	0,1	0,7522	1,10	0,007	0,021	0,3	0,7522

Uočavamo da su na razini značajnosti od 5% značajni age (dob), CRP, HCY, FIB, IL_6, klirens, AER, HDL, ALT i MK.

8.4.4 Jednostruka linearna regresija za transformiranu ekskreciju albumina

Tablica 8.14: Jednostruka linearna regresija za $ln(ln(AER))$ (rezultati iz SAS-a)

				ANOVA		PROCJENA PARAMETARA				
naziv	**N**	R^2	R^2_{adj}	**F**	**p**	**Sl. član**	**nagib**	**SE**	**t**	**p**
sex	58	0,0249	0,0075	1,4	0,2366	0,80	0,173	0,145	1,2	0,2366
age	58	0,0118	-0,0059	0,7	0,4175	1,41	-0,006	0,007	-0,8	0,4175
duration	58	0,0602	0,0434	3,6	0,0634	0,86	0,019	0,010	1,9	0,0634
waste	**52**	**0,0826**	**0,0642**	**4,5**	**0,0388**	**-0,20**	**0,012**	**0,006**	**2,1**	**0,0388**
WHR	55	0,0543	0,0365	3,0	0,0868	0,20	0,888	0,509	1,7	0,0868
fBg	58	0,0445	0,0274	2,6	0,1121	0,75	0,033	0,020	1,6	0,1121
ppBG	**57**	**0,1239**	**0,1080**	**7,8**	**0,0072**	**0,55**	**0,045**	**0,016**	**2,8**	**0,0072**
HbA1c	**58**	**0,1270**	**0,1114**	**8,1**	**0,0060**	**0,16**	**0,126**	**0,044**	**2,9**	**0,0060**
ApN	57	0,0497	0,0324	2,9	0,0955	1,40	-0,043	0,025	-1,7	0,0955
CRP	**58**	**0,0798**	**0,0633**	**4,9**	**0,0317**	**0,97**	**0,025**	**0,012**	**2,2**	**0,0317**
HCY	58	0,0002	-0,0176	0,0	0,9130	1,06	0,001	0,010	0,1	0,9130
FIB	**58**	**0,0904**	**0,0742**	**5,6**	**0,0218**	**0,28**	**0,209**	**0,088**	**2,4**	**0,0218**
IL_6	**54**	**0,0772**	**0,0595**	**4,4**	**0,0419**	**0,97**	**0,034**	**0,016**	**2,1**	**0,0419**
Cys_C	**58**	**0,1667**	**0,1518**	**11,2**	**0,0015**	**0,58**	**0,485**	**0,145**	**3,4**	**0,0015**
Sijal_ac_	58	0,0369	0,0197	2,2	0,1484	0,36	0,348	0,238	1,5	0,1484
klirens	57	0,0054	-0,0127	0,3	0,5871	0,98	0,054	0,098	0,6	0,5871
SBP	58	0,0055	-0,0122	0,3	0,5792	0,82	0,002	0,003	0,6	0,5792
DBP	58	0,0062	-0,0115	0,4	0,5554	1,42	-0,004	0,007	-0,6	0,5554
WBC	58	0,0184	0,0009	1,1	0,3094	0,76	0,040	0,039	1,0	0,3094
HDL	**58**	**0,0997**	**0,0836**	**6,2**	**0,0158**	**1,68**	**-0,462**	**0,186**	**-2,5**	**0,0158**
LDL	58	0,0046	-0,0132	0,3	0,6126	0,95	0,039	0,076	0,5	0,6126
TG	**58**	**0,1198**	**0,1041**	**7,6**	**0,0078**	**0,87**	**0,076**	**0,027**	**2,8**	**0,0078**
AST	58	0,0000	-0,0178	0,0	0,9793	1,08	0,000	0,008	0,0	0,9793
ALT	58	0,0036	-0,0142	0,2	0,6545	1,01	0,002	0,005	0,5	0,6545
GGT	58	0,0050	-0,0128	0,3	0,5976	1,03	0,001	0,002	0,5	0,5976
MK	58	0,0551	0,0383	3,3	0,0760	0,58	0,002	0,001	1,8	0,0760
C1	56	0,0423	0,0246	2,4	0,1282	0,89	0,253	0,164	1,5	0,1282
FLI	**52**	**0,1273**	**0,1099**	**7,3**	**0,0094**	**0,47**	**0,008**	**0,003**	**2,7**	**0,0094**
GDR	**55**	**0,0708**	**0,0533**	**4,0**	**0,0495**	**1,46**	**-0,066**	**0,033**	**-2,0**	**0,0495**

Uočavamo da su na razini značajnosti od 5% značajni waste, ppBG, HbA1c, CRP, FIB, IL_6, Cys_C, HDL, TG, FLI i GDR.

8.5 Odabir modela

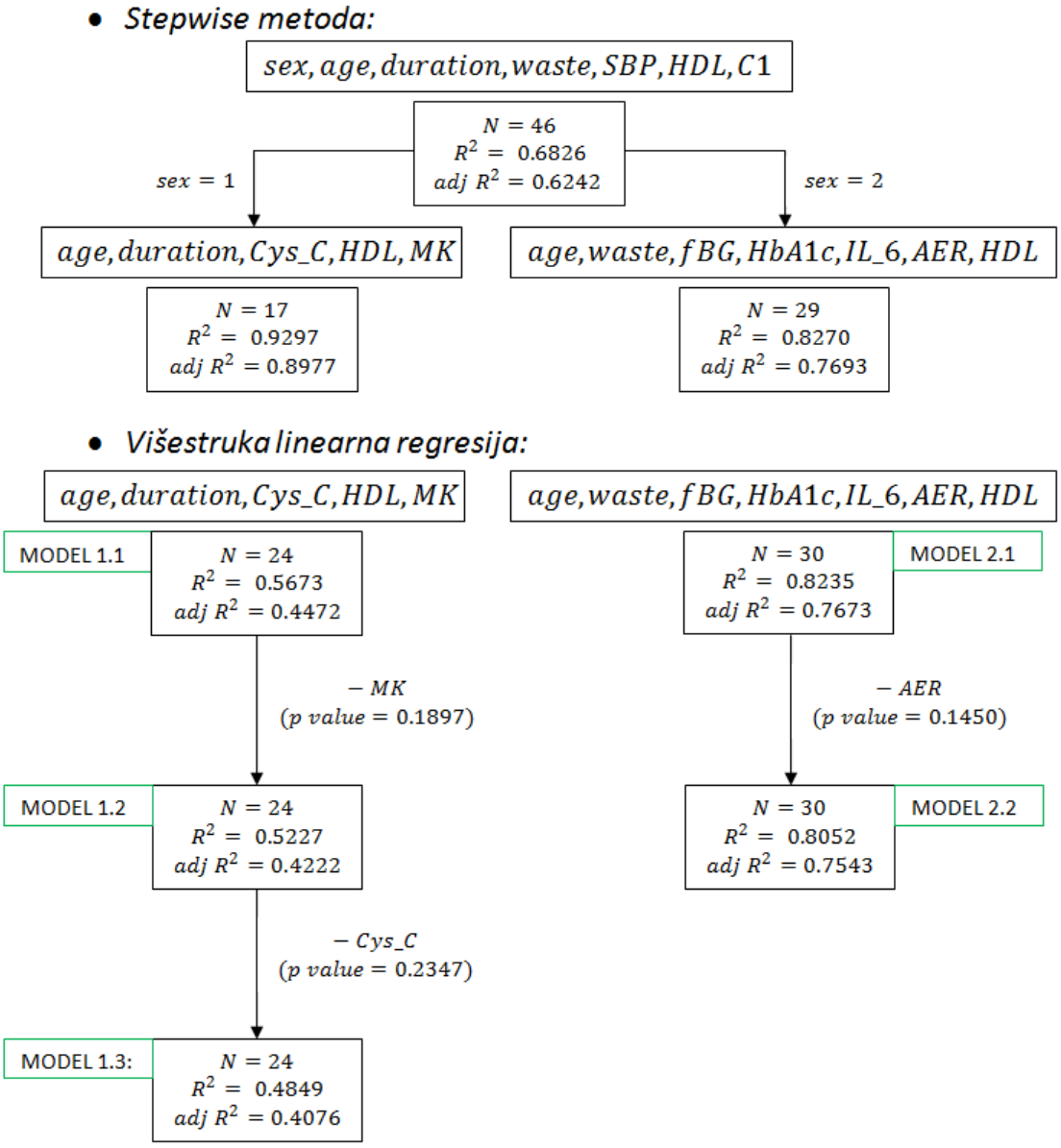
Prilikom modeliranja koristit ćemo stepwise metodu koju smo objasnili u prošlom poglavlju. SAS u provedbi stepwise metode uzima razinu značajnosti od 15 %. Budući da smo se mi ograničili na razinu značajnosti od 15 %, sve varijable koje prelaze tu razinu značajnosti izbacujemo jednu po jednu iz danog modela. Time se mijenja i veličina baze pa dobivamo nove rezultate.

Budući da SAS zanemaruje retke u kojima nedostaje barem jedan podatak, provest ćemo višestruku linearnu regresiju (`PROC REG`) uključujući samo varijable koje su stepwise metodom istaknute kao potencijalne nezavisne varijable višestrukog linearnog regresijskog modela. Na taj način dobivamo više podataka koji ulaze u analizu početnog modela.

Također, procijenit ćemo kvalitetu modela pomoću (prilagođenog) koeficijenta determinacije, Mallowe C_p statistike te AKAIKE kriterija. Dobivene vrijednosti usporedit ćemo sa procjenama modela dobivenih izbacivanjem varijabli koje nisu značajne na razini značajnosti od 5%. Svaku drastičniju promjenu u vrijednostima navedenih pokazatelja kvalitete modela nastojat ćemo dodatno pojasniti.

Dakle, naš zadatak je modelirati modele za transformacije adiponektina, homocisteina, cistacina C i ekskrecije albumina.

8.5.1 Odabir linearnog regresijskog modela za logaritmirani adiponektin, $ln(ApN)$



Slika 8.11: Proces modeliranja linearnog regresijskog modela za logaritmirani adiponektin

Odabir linearnog regresijskog modela za logaritmirani adiponektin stepwise metodom

ULAZ:

sex age duration waste WHR fBG ppBG HbA1c CRP HCY FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST ALT GGT MK C1 FLI GDR

IZLAZ:

sex age duration waste SBP HDL C1

Kod u SAS-u:

```
1  proc reg data = novabaza;
2     model lnApN = sex age duration waste WHR fBG ppBG HbA1c CRP HCY
3     FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST
4     ALT GGT MK C1 FLI GDR / selection = stepwise;
5  run;
```

Tablica 8.15: Rezultati sažetka stepwise metode za odabir modela za logaritmirani adiponektin (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnApN

Number of Observations Read	59
Number of Observations Used	46
Number of Observations with Missing Values	13

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	5.07908	0.72558	11.68	<.0001
Error	38	2.36149	0.06214		
Corrected Total	45	7.44057			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.20381	0.53242	0.00911	0.15	0.7040
sex	-0.18953	0.07952	0.35305	5.68	0.0223
age	0.02284	0.00406	1.96198	31.57	<.0001
duration	-0.01128	0.00598	0.22064	3.55	0.0672
waste	0.00459	0.00303	0.14235	2.29	0.1384
SBP	-0.00497	0.00170	0.53313	8.58	0.0057
HDL	0.65323	0.12396	1.72565	27.77	<.0001
C1	0.21159	0.10557	0.24966	4.02	0.0522

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	age		age	1	0.2982	0.2982	20.6734	18.70	<.0001
2	HDL		HDL	2	0.1650	0.4632	7.9368	13.22	0.0007
3	duration		duration	3	0.0761	0.5394	3.1366	6.94	0.0117
4	SBP		SBP	4	0.0432	0.5826	1.2760	4.25	0.0457
5	sex		sex	5	0.0400	0.6226	-0.2928	4.24	0.0462
6	C1		C1	6	0.0409	0.6635	-1.9479	4.74	0.0355
7	waste		waste	7	0.0191	0.6826	-1.6565	2.29	0.1384

Stepwise metoda istaknula je varijablu spol kao potencijalnu nezvisnu varijablu pa će prvi korak biti primijeniti stepwise metodu posebno za žene i posebno za muškarce. Procjena dobivenog modela dana je u sljedećoj tablici.

Kod u SAS-u:

```
proc reg data = novabaza;
   model lnApN = sex age duration waste WHR fBG ppBG HbA1c CRP HCY
   FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST
   ALT GGT MK C1 FLI GDR / selection = rsquare adjrsq rmse cp aic;
run;
```

Tablica 8.16: Rezultati procjene modela dobivenog stepwise metodom za logaritmirani adiponektin (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
7	0.6826	0.6242	-1.6565	-120.5900	0.24929	sex age duration waste SBP HDL C1

Odabir linearnog regresijskog modela za logaritmirani adiponektin stepwise metodom raščlanjenom po spolu

Žene (kodirano sa 1)

ULAZ:

age duration waste WHR fBG ppBG HbA1c CRP HCY FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST ALT GGT MK C1 FLI GDR

IZLAZ:

age duration Cys_C HDL MK

Kod u SAS-u:

```
1  proc sort data=novabaza;
2     by sex;
3  run;
4  proc reg data = novabaza;
5     model lnApN = sex age duration waste WHR fBG ppBG HbA1c CRP HCY
6     FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST
7     ALT GGT MK C1 FLI GDR / selection = stepwise;
8     by sex;
9  run;
```

Tablica 8.17: Rezultati sažetka stepwise metode za odabir modela za logaritmirani adiponektin za žene (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnApN

sex=1

Number of Observations Read	24
Number of Observations Used	17
Number of Observations with Missing Values	7

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4.05904	0.81181	29.09	<.0001
Error	11	0.30695	0.02790		
Corrected Total	16	4.36599			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.73817	0.32158	0.14703	5.27	0.0424
age	0.01153	0.00449	0.18405	6.60	0.0261
duration	-0.02518	0.00556	0.57222	20.51	0.0009
Cys_C	1.25115	0.18413	1.28836	46.17	<.0001
HDL	1.79286	0.18964	2.49406	89.38	<.0001
MK	-0.00438	0.00079453	0.84980	30.45	0.0002

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	HDL		HDL	1	0.3155	0.3155	.	6.91	0.0190
2	Cys_C		Cys-C	2	0.2446	0.5601	.	7.78	0.0145
3	MK		MK	3	0.2327	0.7928	.	14.60	0.0021
4	duration		duration	4	0.0947	0.8875	.	10.11	0.0079
5	age		age	5	0.0422	0.9297	.	6.60	0.0261

Uočimo da su sve potencijalne nezavisne varijable dobivene stepwise metodom značajne na razini značajnosti od 5%. Procjena dobivenog modela dana je u sljedećoj tablici.

Kod u SAS-u:

```
proc reg data = novabaza;
   model lnApN = age duration waste WHR fBG ppBG HbA1c CRP HCY
   FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST
   ALT GGT MK C1 FLI GDR / selection = rsquare adjrsq rmse cp aic;
   by sex;
run;
```

Tablica 8.18: Rezultati procjene modela dobivenog stepwise metodom za logaritmirani adiponektin za žene (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
5	0.9125	0.8728	.	-52.5320	0.18631	duration HCY Cys_C HDL MK

U odnosu na model gdje smo modelirali žene i muškarce zajedno, ovaj model ima puno veći R^2 i R^2_{adj}.

Muškarci (kodirano sa 2)

ULAZ:

age duration waste WHR fBG ppBG HbA1c CRP HCY FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST ALT GGT MK C1 FLI GDR

IZLAZ:

age waste fBG HbA1c IL_6 AER HDL

Tablica 8.19: Rezultati sažetka stepwise metode za odabir modela za logaritmirani adiponektin za muškarce (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnApN

sex=2

Number of Observations Read	35
Number of Observations Used	29
Number of Observations with Missing Values	6

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	1.87205	0.26744	14.34	<.0001
Error	21	0.39174	0.01865		
Corrected Total	28	2.26379			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.42782	0.40217	0.02111	1.13	0.2995
age	0.01797	0.00294	0.69924	37.48	<.0001
waste	0.01049	0.00255	0.31550	16.91	0.0005
fBG	0.04237	0.01485	0.15183	8.14	0.0095
HbA1c	-0.11417	0.03025	0.26563	14.24	0.0011
IL_6	0.02056	0.00626	0.20101	10.78	0.0036
AER	-0.00006650	0.00004385	0.04290	2.30	0.1443
HDL	0.37588	0.07741	0.43985	23.58	<.0001

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	age		age	1	0.3524	0.3524	.	14.70	0.0007
2	HDL		HDL	2	0.1799	0.5324	.	10.00	0.0039
3	waste		waste	3	0.0831	0.6155	.	5.40	0.0285
4	IL_6		IL-6	4	0.0821	0.6975	.	6.51	0.0175
5	HbA1c		HbA1c	5	0.0573	0.7549	.	5.38	0.0296
6	fBG		fBG	6	0.0531	0.8080	.	6.09	0.0219
7	AER		AER	7	0.0189	0.8270	.	2.30	0.1443

Uočimo da je varijabla *AER* jedina varijabla koja nije značajna na razini značajnosti od 5%.Procjena dobivenog modela dana je u sljedećoj tablici.

Kod u SAS-u:

```
proc reg data = novabaza;
   model lnApN = sex age duration waste WHR fBG ppBG HbA1c CRP HCY
   FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST
   ALT GGT MK C1 FLI GDR / selection = rsquare adjrsq rmse cp aic;
   by sex;
run;
```

Tablica 8.20: Rezultati procjene modela dobivenog stepwise metodom za logaritmirani adiponektin za muškarce (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
7	0.8270	0.7693	.	-108.8291	0.13658	age waste fBG HbA1c IL_6 AER HDL

Uspoređujući R^2 i R^2_{adj} ovog modela za muškarce i zajedničkog modela uočavamo porast u njihovim vrijednostima.

Višestruka linearna regresija za logaritmirani adiponektin za žene varijabli odabranih stepwise metodom (MODEL 1.1)

ULAZ:

age duration Cys_C HDL MK

Kod u SAS-u:

```
1  data novabaza_restricted_for_ApN_sex1;
2     set novabaza;
3     where lnApN and sex and age and duration and Cys_C and HDL and MK
4     is not missing;
5  run;
6
7  proc reg data = novabaza_restricted_for_ApN_sex1;
8     model lnApN = age duration Cys_C HDL MK;
9     by sex;
10 run;
```

Tablica 8.21: Rezultati višestruke linearne regresije za logaritmirani adiponektin za žene varijabli odabranih stepwise metodom (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnApN

sex=1

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.67243	0.53449	4.72	0.0063
Error	18	2.03798	0.11322		
Corrected Total	23	4.71041			

Root MSE	0.33648	R-Square	0.5673
Dependent Mean	2.12246	Adj R-Sq	0.4472
Coeff Var	15.85350		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.36703	0.52733	0.70	0.4953
age	age	1	0.01206	0.00790	1.53	0.1442
duration	duration	1	-0.02533	0.00999	-2.54	0.0207
Cys_C	Cys-C	1	0.54835	0.29601	1.85	0.0804
HDL	HDL	1	0.89723	0.25945	3.46	0.0028
MK	MK	1	-0.00173	0.00127	-1.36	0.1897

Nakon ponovljene višestruke regresije sa nezavisnim varijablama dobivenih stepwise metodom uočavamo da varijable MK i age nisu više značajne na razini značajnosti od 5%. Sljedeći korak će biti izbacivanje varijable MK iz dobivenog modela. Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
1  proc reg data = novabaza_restricted_for_ApN_sex1;
2     model lnApN = age duration Cys_C HDL MK
3     / selection = rsquare adjrsq rmse cp aic;
4     by sex;
5  run;
```

Tablica 8.22: Rezultati procjene višestruke linearne regresije za logaritmirani adiponektin za žene varijabli odabranih stepwise metodom (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
5	0.5673	0.4472	6.0000	-47.1863	0.33648	age duration Cys_C HDL MK

Vrijednosti R^2 i R^2_{adj} su drastično pale, a ostale vrijednosti procjenitelja modela su se povećale. Ovo možemo pripisati razlici broja podataka nad kojima smo provodili modeliranje. Naime, u stepwise proceduri $N = 17$, a sada je $N = 24$.

Višestruka linearna regresija za logaritmirani adiponektin za žene (MODEL 1.2 = MODEL 1.1 - MK)

ULAZ:

age duration Cys_C HDL

Kod u SAS-u:

```
1  proc reg data = novabaza_restricted_for_ApN_sex1;
2     model lnApN = age duration Cys_C HDL;
3     by sex;
4  run;
```

Tablica 8.23: Rezultati višestruke linearne regresije modela 1.2 za logaritmirani adiponektin za žene (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnApN

sex=1

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2.46210	0.61552	5.20	0.0053
Error	19	2.24831	0.11833		
Corrected Total	23	4.71041			

Root MSE	0.34399	R-Square	0.5227
Dependent Mean	2.12246	Adj R-Sq	0.4222
Coeff Var	16.20737		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.09790	0.49988	0.20	0.8468
age	age	1	0.01545	0.00766	2.02	0.0582
duration	duration	1	-0.02277	0.01003	-2.27	0.0350
Cys_C	Cys-C	1	0.24892	0.20282	1.23	0.2347
HDL	HDL	1	0.77478	0.24884	3.11	0.0057

Vidimo da sada po značajnosti jako "odskače" varijabla Cys_C. Sljedeći korak je njezino izbacivanje iz modela. Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
1   proc reg data = novabaza_restricted_for_ApN_sex1;
2      model lnApN = age duration Cys_C HDL
3      / selection = rsquare adjrsq rmse cp aic;
4      by sex;
5 run;
```

Tablica 8.24: Rezultati procjene višestruke linearne regresije modela 1.2 za logaritmirani adiponektin za žene (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
4	0.5227	0.4222	5.0000	-46.8290	0.34399	age duration Cys_C HDL

Vrijednost R^2 je nešto niža, ali je R^2_{ajd} viši što opravdava izbacivanje varijable MK iz prethodnog modela. Vrijednost C_p je isto malo niža.

Višestruka linearna regresija za logaritmirani adiponektin za žene (MODEL 1.3 = MODEL 1.2 - Cys_C) - KONAČAN MODEL

ULAZ:

age duration HDL

Kod u SAS-u:

```
1  proc reg data =novabaza_restricted_for_ApN_sex1;
2     model lnApN=age duration HDL;
3     by sex;
4  run;
```

Tablica 8.25: Rezultati višestruke linearne regresije modela 1.3 za logaritmirani adiponektin za žene (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnApN

sex=1

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2.28386	0.76129	6.27	0.0035
Error	20	2.42655	0.12133		
Corrected Total	23	4.71041			

Root MSE	0.34832	R-Square	0.4849
Dependent Mean	2.12246	Adj R-Sq	0.4076
Coeff Var	16.41121		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.29348	0.47975	0.61	0.5476
age	age	1	0.01921	0.00711	2.70	0.0138
duration	duration	1	-0.02457	0.01004	-2.45	0.0238
HDL	HDL	1	0.65821	0.23289	2.83	0.0104

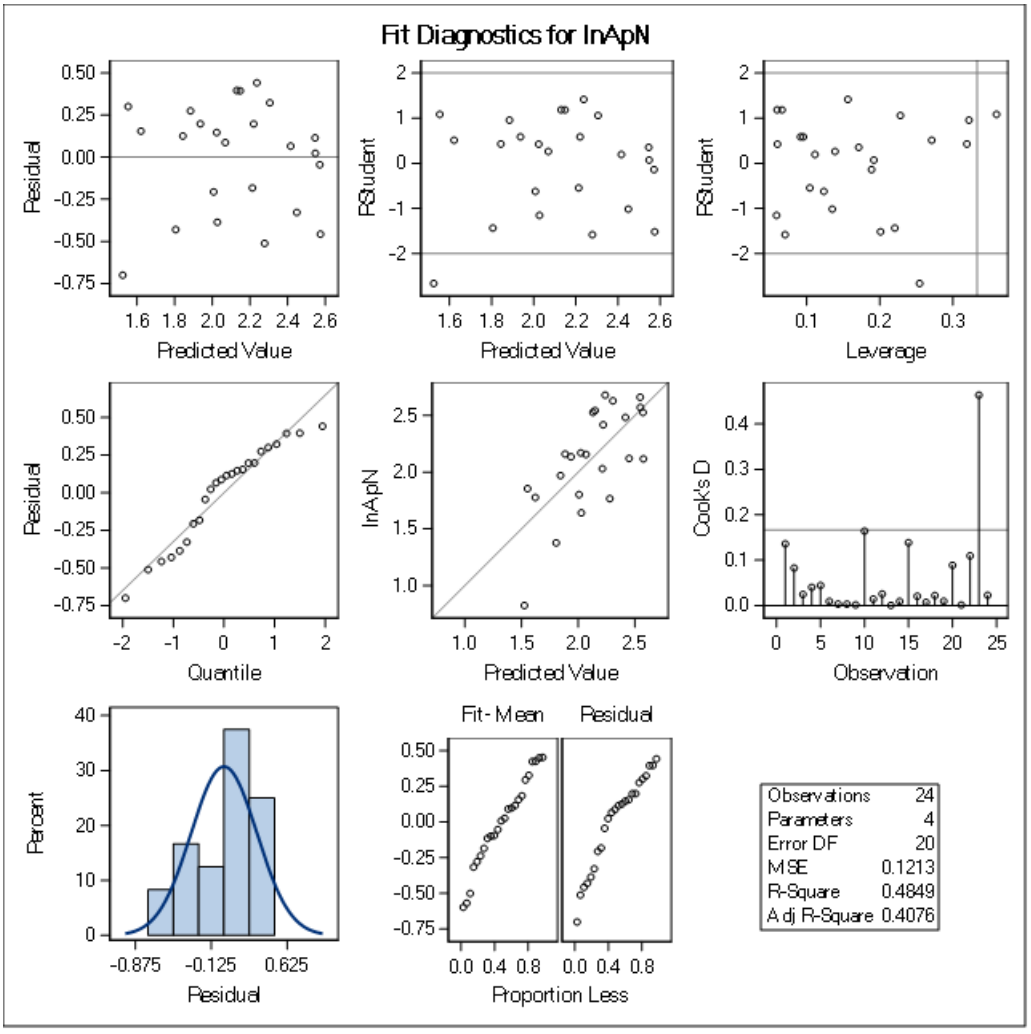
Sve navedene varijable su značajne na razini značajnosti od 5% pa tu stajemo s modeliranjem. Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
1  proc reg data=novabaza_restricted_for_ApN_sex1;
2     model lnApN = age duration HDL
3     / selection = rsquare adjrsq rmse cp aic;
4     by sex;
5  run;
```

Tablica 8.26: Rezultati procjene višestruke linearne regresije modela 1.3 za logaritmirani adiponektin za žene (ispis iz SAS-a)

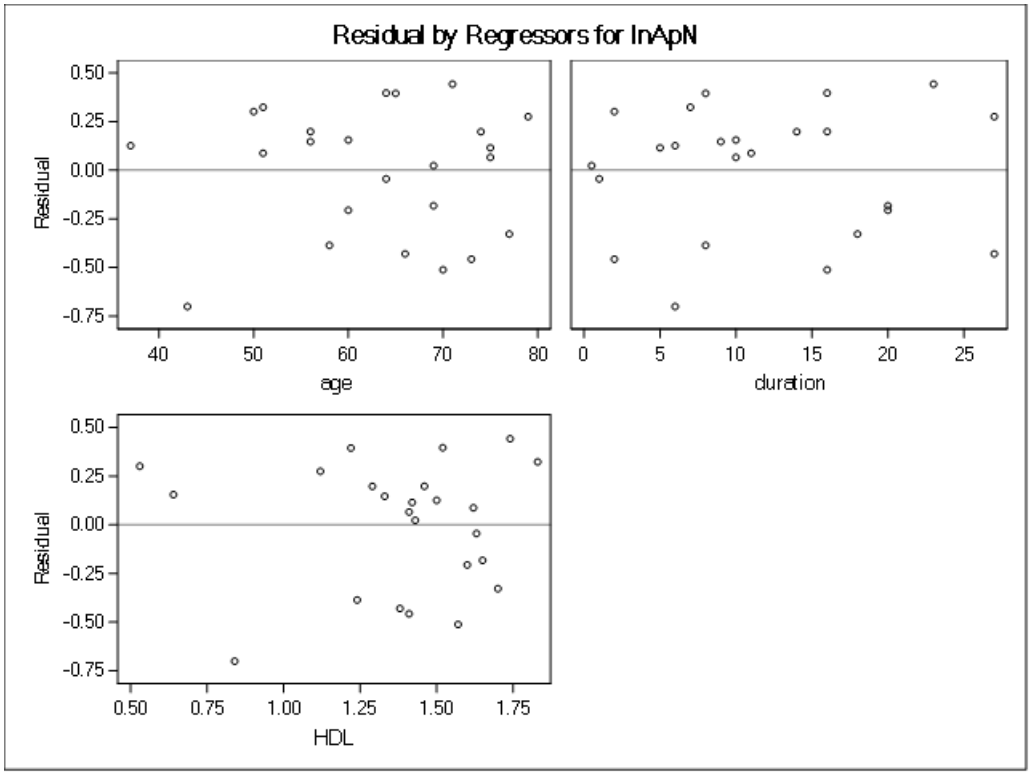
Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
3	0.4849	0.4076	4.0000	-46.9980	0.34832	age duration HDL

Vrijednost R^2, R^2_{ajd} i C_p su malo niže, dok su ostale vrijednosti približno jednake.



Slika 8.12: Dijagnostika višestrukog linearnog regresijskog konačnog modela za logaritmirani adiponektin za žene (ispis iz SAS-a)

Dijagnostika dobivenog modela zadovoljava kriterije modeliranja. Jedino graf $PredictedValues-lnApN$ ukazuje na blago raspršenje oko pravca $lnApN = PredictedValues$



Slika 8.13: Analiza reziduala nezavisnih varijabli višestrukog linearnog regresijskog modela za logaritmirani adiponektin za žene (ispis iz SAS-a)

Reziduali ne pokazuju očiti uzorak pa zaključujemo da je model dobar. Naravno, modeli se uvijek u suradnji sa strukom mogu doraditi do još "boljeg".

Dakle, dobiven je sljedeći regresijski model za žene:

$$ln(ApN) = 0.29348 + 0.01921age - 0.02457duration + 0.65821HDL\,.$$

Višestruka linearna regresija za logaritmirani adiponektin za muškarce varijabli odabranih stepwise metodom (MODEL 2.1)

ULAZ:

age waste fBG HbA1c IL_6 AER HDL

Kod u SAS-u: Kod u SAS-u:

```
data novabaza_restricted_for_ApN_sex2;
   set novabaza;
   where lnApN and sex and age and waste and fBG and HbA1c and AER
   and HDL and IL_6 is not missing;
run;
proc reg data=novabaza_restricted_for_ApN_sex2;
   model lnApN=age waste fBG HbA1c IL_6 AER HDL;
   by sex;
run;
```

Tablica 8.27: Rezultati višestruke linearne regresije za logaritmirani adiponektin za muškarce varijabli odabranih stepwise metodom (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnApN

sex=2

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	1.86566	0.26652	14.66	<.0001
Error	22	0.39988	0.01818		
Corrected Total	29	2.26554			

Root MSE	0.13482	R-Square	0.8235
Dependent Mean	1.86239	Adj R-Sq	0.7673
Coeff Var	7.23907		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.41090	0.39617	-1.04	0.3109
age	age	1	0.01821	0.00288	6.33	<.0001
waste	waste	1	0.01023	0.00249	4.11	0.0005
fBG	fBG	1	0.04321	0.01460	2.96	0.0073
HbA1c	HbA1c	1	-0.11473	0.02985	-3.84	0.0009
IL_6	IL-6	1	0.02099	0.00615	3.41	0.0025
AER	AER	1	-0.00006536	0.00004325	-1.51	0.1450
HDL	HDL	1	0.36671	0.07517	4.88	<.0001

Uočavamo da varijabla *AER* nije značajna na razini značajnosti od 5%. Ovo se donekle poklapa sa rezultatima dobivenim stepwise metodom za muškarce. Razlog tome je da u ovoj bazi imamo samo jednu observaciju više, tj. sada ih je $N = 30$, a prije ih je bilo $N = 29$. Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
1  proc reg data=novabaza_restricted_for_ApN_sex2;
2     model lnApN = age waste fBG HbA1c IL_6 AER HDL
3     / selection = rsquare adjrsq rmse cp aic;
4     by sex;
5  run;
```

Tablica 8.28: procjene višestruke linearne regresije za logaritmirani adiponektin za muškarce varijabli odabranih stepwise metodom (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
7	0.8235	0.7673	8.0000	-113.5337	0.13482	age waste fBG HbA1c IL_6 AER HDL

Dobivene vrijednosti procjenitelja modela vrlo dobre i nastavljamo sa daljnjim modeliranjem.

Višestruka linearna regresija za logaritmirani adiponektin za muškarce (MODEL 2.2 = MODEL 2.1 - AER) - KONAČAN MODEL

ULAZ:

age waste fBG HbA1c IL_6 HDL

Kod u SAS-u: Kod u SAS-u:

```
1  proc reg data=novabaza_restricted_for_ApN_sex2;
2     model lnApN=age waste fBG HbA1c IL_6 HDL;
3     by sex;
4  run;
```

Tablica 8.29: Rezultati višestruke linearne regresije modela 2.2 za logaritmirani adiponektin za muškarce (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnApN

sex=2

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1.82415	0.30402	15.84	<.0001
Error	23	0.44139	0.01919		
Corrected Total	29	2.26554			

Root MSE	0.13853	R-Square	0.8052
Dependent Mean	1.86239	Adj R-Sq	0.7543
Coeff Var	7.43833		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.31662	0.40200	-0.79	0.4390
age	age	1	0.01724	0.00288	5.98	<.0001
waste	waste	1	0.00970	0.00253	3.83	0.0009
fBG	fBG	1	0.03718	0.01444	2.58	0.0169
HbA1c	HbA1c	1	-0.10912	0.03044	-3.59	0.0016
IL_6	IL-6	1	0.02223	0.00626	3.55	0.0017
HDL	HDL	1	0.38024	0.07669	4.96	<.0001

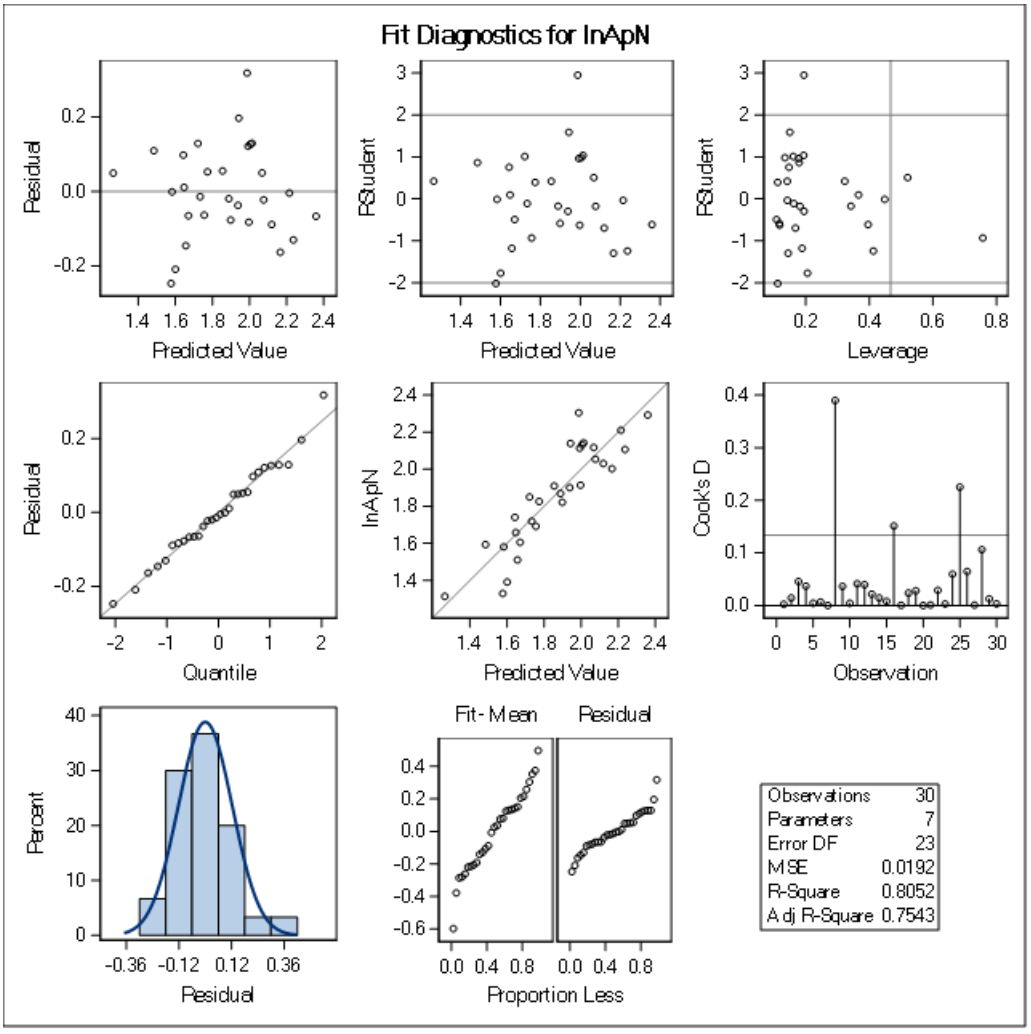
Sve nezavisne varijable u ovom modelu su značajne na razini značajnosti od 5%. Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
1  proc reg data=novabaza_restricted_for_ApN_sex2;
2     model lnApN = age waste fBG HbA1c IL_6 HDL
3     / selection = rsquare adjrsq rmse cp aic;
4     by sex;
5  run;
```

Tablica 8.30: Rezultati procjene višestruke linearne regresije modela 2.2 za logaritmirani adiponektin za muškarce (ispis iz SAS-a)

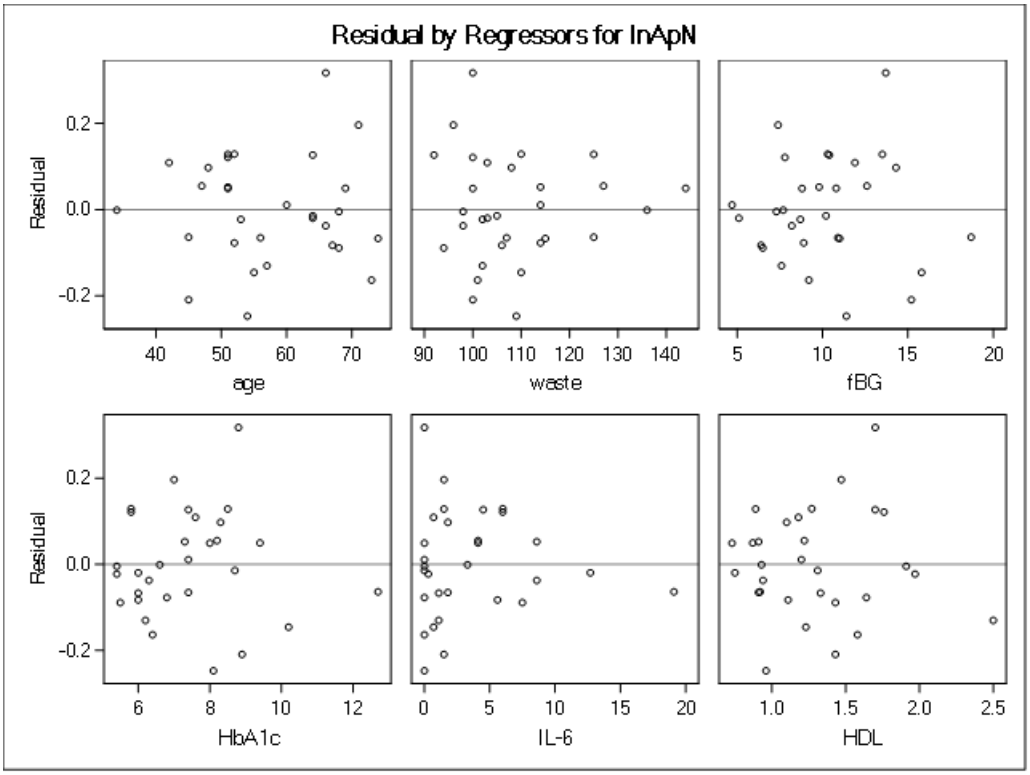
Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
6	0.8052	0.7543	7.0000	-112.5709	0.13853	age waste fBG HbA1c IL_6 HD

Vrijednosti R^2, R^2_{adj} i C_p su niže, AIC je nešto viši, a MSE je približno jednak. Zaključujemo da nismo puno izgubili izbacivanjem varijable AER iz modela.



Slika 8.14: Dijagnostika višestrukog linearnog regresijskog konačnog modela za logaritmirani adiponektin za muškarce (ispis iz SAS-a)

Dijagnostika dobivenog modela zadovoljava kriterije modeliranja. Jedino graf $PredictedValues - lnApN$ ukazuje na blago raspršenje oko pravca $lnApN = PredictedValues$.



Slika 8.15: Analiza reziduala nezavisnih varijabli višestrukog linearnog regresijskog modela za logaritmirani adiponektin za muškarce (ispis iz SAS-a)

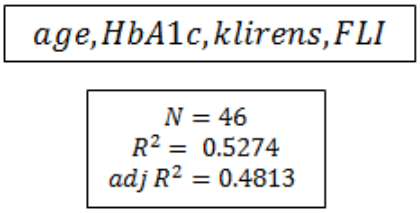
Kod reziduala možemo naslutiti blagi uzorak, ali model uvijek možemo poboljšati u suradnji sa strukom.

Dakle, dobiven je sljedeći regresijski model za muškarce:

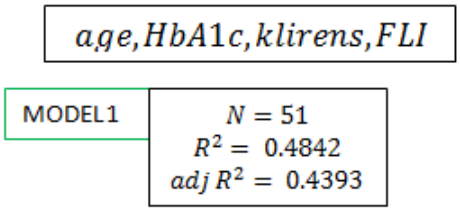
$$ln(ApN) = -0.31662 + 0.01724age + 0.00970waste + 0.03718fBG + 0.10912HbA1c + 0.02223IL\_6 + 0.36024HDL.$$

8.5.2 Odabir linearnog regresijskog modela za recipročnu vrijednost drugog korijena iz homocisteina, *HCY*

- *Stepwise metoda:*



- *Višestruka linearna regresija:*



Slika 8.16: Proces modeliranja linearnog regresijskog modela za recipročnu vrijednost drugog korijena iz homocisteina

Odabir linearnog regresijskog modela za recipročnu vrijednost drugog korijena iz homocisteina stepwise metodom

ULAZ:

sex age duration waste WHR fBG ppBG HbA1c ApN CRP FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST ALT GGT MK C1 FLI GDR

IZLAZ:

age HbA1c klirens FLI

Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model invsqrtHCY =sex age duration waste WHR fBG ppBG HbA1c ApN
3     CRP FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG
4     AST ALT GGT MK C1 FLI GDR/ selection=stepwise;
5  run;
```

Tablica 8.31: Rezultati sažetka stepwise metode za odabir modela za recipročnu vrijednost drugog korijena iz homocisteina (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: invsqrtHCY

Number of Observations Read	59
Number of Observations Used	46
Number of Observations with Missing Values	13

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.07617	0.01904	11.44	<.0001
Error	41	0.06825	0.00166		
Corrected Total	45	0.14442			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.30961	0.06083	0.04313	25.91	<.0001
age	-0.00134	0.00065351	0.00696	4.18	0.0474
HbA1c	0.01109	0.00442	0.01047	6.29	0.0162
klirens	0.03417	0.00921	0.02293	13.77	0.0006
FLI	-0.00125	0.00029338	0.03000	18.02	0.0001

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Cys_C		Cys-C	1	0.2859	0.2859	-9.9285	17.61	0.0001
2	klirens		klirens	2	0.0751	0.3610	-11.303	5.06	0.0297
3	DBP		DBP	3	0.0583	0.4192	-11.919	4.21	0.0464
4	age		age	4	0.0384	0.4577	-11.645	2.91	0.0958
5	FLI		FLI	5	0.0413	0.4990	-11.499	3.30	0.0770
6	HbA1c		HbA1c	6	0.0609	0.5599	-12.235	5.40	0.0254
7		Cys_C	Cys-C	5	0.0142	0.5456	-13.596	1.26	0.2681
8		DBP	DBP	4	0.0182	0.5274	-14.777	1.61	0.2125

Uočimo da su sve nezavisne varijable u ovom modelu značajne na razini značajnosti od 5%. Vidimo i da su tokom stepwise procedure iz modela izašle varijable Cys_C i DBP. Najčešći razlog je unutarnja korelacijska struktura.

Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
proc reg data=novabaza;
   model invsqrtHCY=sex age duration waste WHR fBG ppBG HbA1c ApN
   CRP FIB IL_6 Cys_C Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG
   AST ALT GGT MK C1 FLI GDR/selection=rsquare adjrsq rmse cp aic;
run;
```

Tablica 8.32: Rezultati procjene modela dobivenog stepwise metodom za recipročnu vrijednost drugog korijena iz homocisteina (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
4	0.5274	0.4813	-14.7767	-289.6077	0.04080	age HbA1c klirens FLI

Višestruka linearna regresija za recipročnu vrijednost drugog korijena iz homocisteina varijabli odabranih stepwise metodom (MODEL 1) - KONAČAN MODEL

ULAZ:

age HbA1c klirens FLI

Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model invsqrtHCY=age HbA1c klirens FLI;
3  run;
```

Tablica 8.33: Rezultati višestruke linearne regresije za recipročnu vrijednost drugog korijena iz homocisteina varijabli odabranih stepwise metodom (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: invsqrtHCY

Number of Observations Read	59
Number of Observations Used	51
Number of Observations with Missing Values	8

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.07520	0.01880	10.79	<.0001
Error	46	0.08011	0.00174		
Corrected Total	50	0.15532			

Root MSE	0.04173	R-Square	0.4842
Dependent Mean	0.28271	Adj R-Sq	0.4393
Coeff Var	14.76160		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.31702	0.06031	5.26	<.0001
age	age	1	-0.00144	0.00065606	-2.19	0.0337
HbA1c	HbA1c	1	0.00871	0.00426	2.04	0.0467
klirens	klirens	1	0.03366	0.00927	3.63	0.0007
FLI	FLI	1	-0.00100	0.00028033	-3.57	0.0008

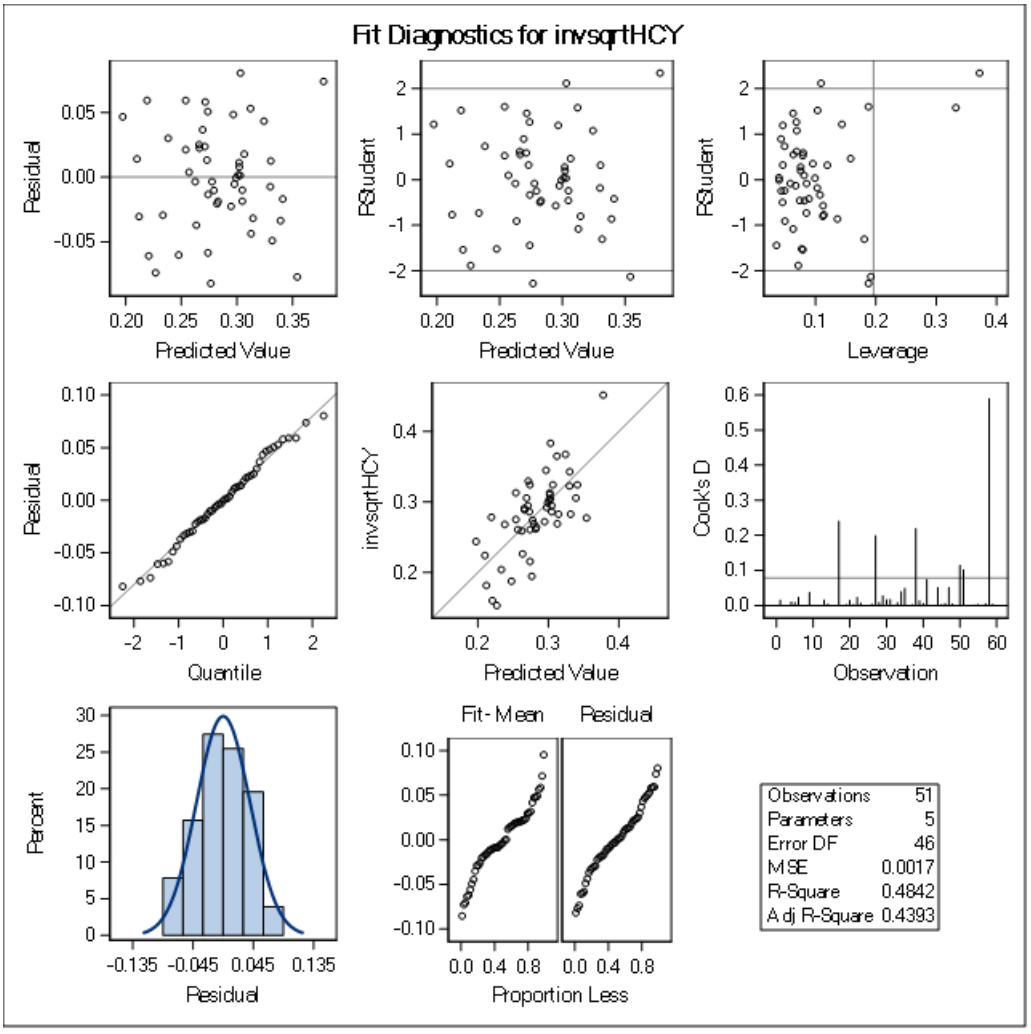
Ovo modeliranje je potvrdilo rezultate stepwise metode usprkos dodatnih 5 podataka. Dakle, sve varijable su i dalje značajne na razini značajnosti od 5%. Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model invsqrtHCY=age HbA1c klirens FLI
3     /selection=rsquare adjrsq rmse cp aic;
4  run;
```

Tablica 8.34: Rezultati procjene višestruke linearne regresije za recipročnu vrijednost drugog korijena iz homocisteina varijabli odabranih stepwise metodom (ispis iz SAS-a)

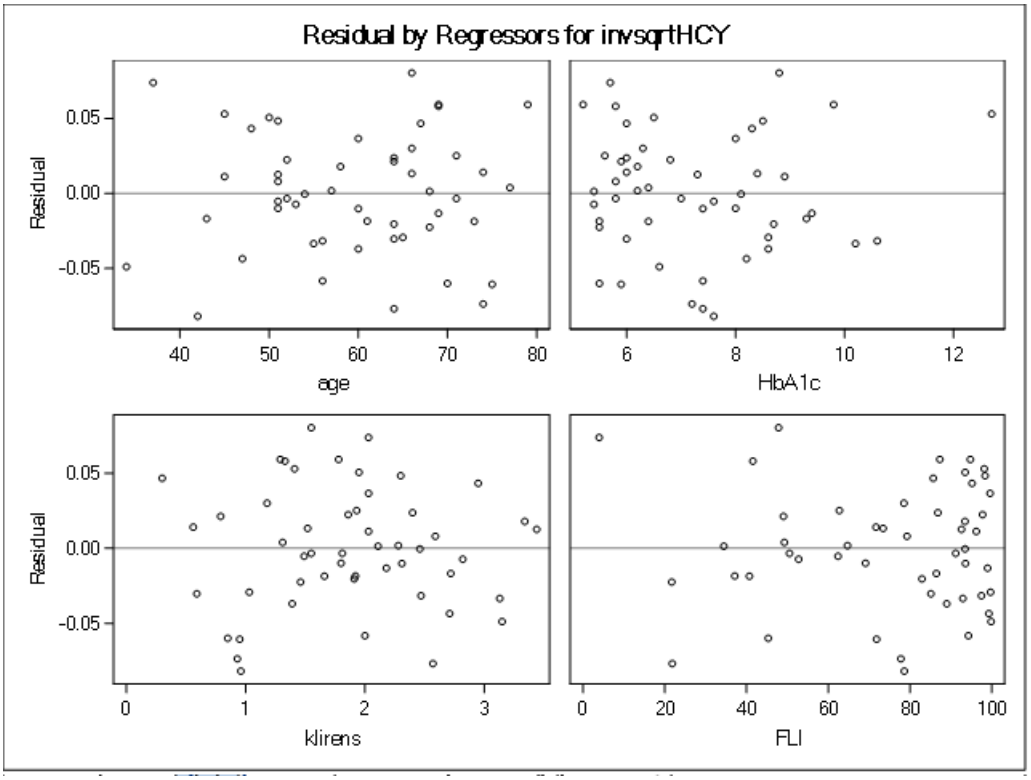
Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
4	0.4842	0.4393	5.0000	-319.2628	0.04173	age HbA1c klirens FLI

Uočimo da je *AIC* drastično pao, dok se vrijednosti ostalih statistika nisu značajno promijenile.



Slika 8.17: Dijagnostika višestrukog linearnog regresijskog konačnog modela za recipročnu vrijednost drugog korijena iz homocisteina (ispis iz SAS-a)

Dijagnostika dobivenog modela zadovoljava kriterije modeliranja. Jedino graf $PredictedValues - lnApN$ ukazuje na blago raspršenje oko pravca $lnApN = PredictedValues$.



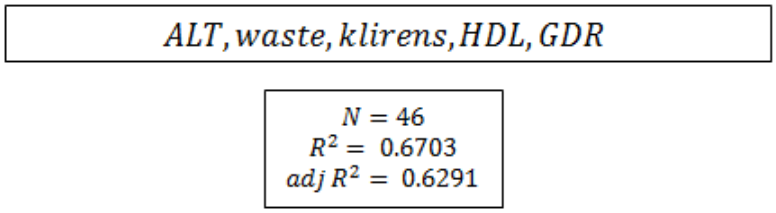
Slika 8.18: Analiza reziduala nezavisnih varijabli višestrukog linearnog regresijskog modela za recipročnu vrijednost drugog korijena iz homocisteina (ispis iz SAS-a)

Analiza reziduala daje dobre rezultate. Naravno, model uvijek možemo poboljšati u suradnji sa strukom.
Dakle, dobiven je sljedeći regresijski model:

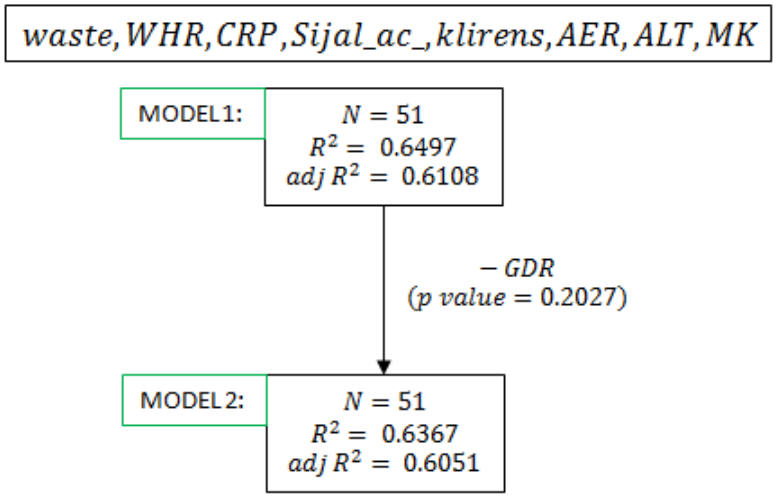
$$\frac{1}{\sqrt{HCY}} = 0.31702 - 0.00144age + 0.00871HbA1c + 0.3366klirens - 0.00100FLI.$$

8.5.3 Odabir linearnog regresijskog modela za inverz cistacina C, Cys_C

- *Stepwise metoda:*



- *Višestruka linearna regresija:*



Slika 8.19: Proces modeliranja linearnog regresijskog modela za inverz cistacina C

Odabir linearnog regresijskog modela za inverz cistacina C stepwise metodom

ULAZ:

sex age duration waste WHR fBG ppBG HbA1c ApN CRP HCY FIB IL_6 Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST ALT GGT MK C1 FLI GDR

IZLAZ:

ALT waste klirens HDL GDR

Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model invCys_C=sex age duration waste WHR fBG ppBG HbA1c ApN CRP
3     HCY FIB IL_6 Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST ALT
4     GGT MK C1 FLI GDR/selection=stepwise;
5  run;
```

Tablica 8.35: Rezultati sažetka stepwise metode za odabir modela za inverz cistacina C (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: invCys_C

Number of Observations Read	59
Number of Observations Used	46
Number of Observations with Missing Values	13

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3.12521	0.62504	16.26	<.0001
Error	40	1.53725	0.03843		
Corrected Total	45	4.66245			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.15186	0.35685	0.40041	10.42	0.0025
waste	-0.01072	0.00254	0.68654	17.86	0.0001
klirens	0.19084	0.04056	0.85068	22.14	<.0001
HDL	0.19515	0.08618	0.19703	5.13	0.0291
ALT	0.01289	0.00246	1.05582	27.47	<.0001
GDR	0.02410	0.01523	0.09632	2.51	0.1213

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	HCY		HCY	1	0.2617	0.2617	54.2598	15.59	0.0003
2	ALT		ALT	2	0.0919	0.3535	44.2814	6.11	0.0175
3	waste		waste	3	0.1653	0.5189	24.7286	14.43	0.0005
4	klirens		klirens	4	0.1016	0.6204	13.4870	10.97	0.0019
5		HCY	HCY	3	0.0171	0.6034	13.7116	1.84	0.1820
6	HDL		HDL	4	0.0463	0.6496	9.6786	5.42	0.0250
7	GDR		GDR	5	0.0207	0.6703	8.9852	2.51	0.1213

Uočimo da jedino varijabla *GDR* nije značajna na razini značajnosti od 5%. Također, tijekom stepwise procedure varijabla *HCY* je izašla iz modela. Najčešći razlog takvog izlaska varijable iz modela je unutarnja korelacijska struktura.

Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
proc reg data=novabaza;
    model invCys_C=sex age duration waste WHR fBG ppBG HbA1c ApN CRP
    HCY FIB IL_6 Sijal_ac_ klirens AER SBP DBP WBC HDL LDL TG AST ALT
    GGT MK C1 FLI GDR/selection=rsquare adjrsq rmse cp aic;
 run;
```

Tablica 8.36: Rezultati procjene modela dobivenog stepwise metodom za inverz cistacina C (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
5	0.6703	0.6291	8.9852	-144.3379	0.19604	waste klirens HDL ALT GDR

Višestruka linearna regresija za inverz cistacina C varijabli odabranih stepwise metodom (MODEL 1)

ULAZ:

ALT waste klirens HDL GDR

Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model invCys_C=ALT waste klirens HDL GDR;
3  run;
```

Tablica 8.37: Rezultati višestruke linearne regresije za inverz cistacina C varijabli odabranih stepwise metodom (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: invCys_C

Number of Observations Read	59
Number of Observations Used	51
Number of Observations with Missing Values	8

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3.16452	0.63290	16.69	<.0001
Error	45	1.70628	0.03792		
Corrected Total	50	4.87080			

Root MSE	0.19472	R-Square	0.6497
Dependent Mean	1.11921	Adj R-Sq	0.6108
Coeff Var	17.39828		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.06218	0.33532	3.17	0.0028
ALT	ALT	1	0.01220	0.00234	5.21	<.0001
waste	waste	1	-0.00971	0.00239	-4.06	0.0002
klirens	klirens	1	0.19178	0.03970	4.83	<.0001
HDL	HDL	1	0.21143	0.07771	2.72	0.0092
GDR	GDR	1	0.01905	0.01474	1.29	0.2027

Uočimo da se potvrđuje zaključak da varijabla GDR nije značajna na razini značajnosti od 5% pa ju izbacujemo u sljedećem modelu.
Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model invCys_C=ALT waste klirens HDL GDR
3     /selection=rsquare adjrsq rmse cp aic;
4  run;
```

Tablica 8.38: Rezultati procjene višestruke linearne regresije za inverz cistacina C varijabli odabranih stepwise metodom (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
5	0.6497	0.6108	6.0000	-161.2731	0.19472	ALT waste klirens HDL GDR

Jedino je vrijednost od AIC-a porasla, dok su se sve ostale vrijednosti malo spustile.

Višestruka linearna regresija za inverz cistacina C (MODEL 2 = MODEL 1 - GDR) - KONAČAN MODEL

ULAZ:

ALT waste klirens HDL

Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model invCys_C=ALT waste klirens HDL;
3  run;
```

Tablica 8.39: Rezultati višestruke linearne regresije modela 2 za inverz cistacina C (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: invCys_C

Number of Observations Read	59
Number of Observations Used	51
Number of Observations with Missing Values	8

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3.10115	0.77529	20.15	<.0001
Error	46	1.76965	0.03847		
Corrected Total	50	4.87080			

Root MSE	0.19614	R-Square	0.6367
Dependent Mean	1.11921	Adj R-Sq	0.6051
Coeff Var	17.52476		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.29481	0.28500	4.54	<.0001
ALT	ALT	1	0.01110	0.00220	5.05	<.0001
waste	waste	1	-0.01056	0.00231	-4.57	<.0001
klirens	klirens	1	0.19103	0.03999	4.78	<.0001
HDL	HDL	1	0.21612	0.07819	2.76	0.0082

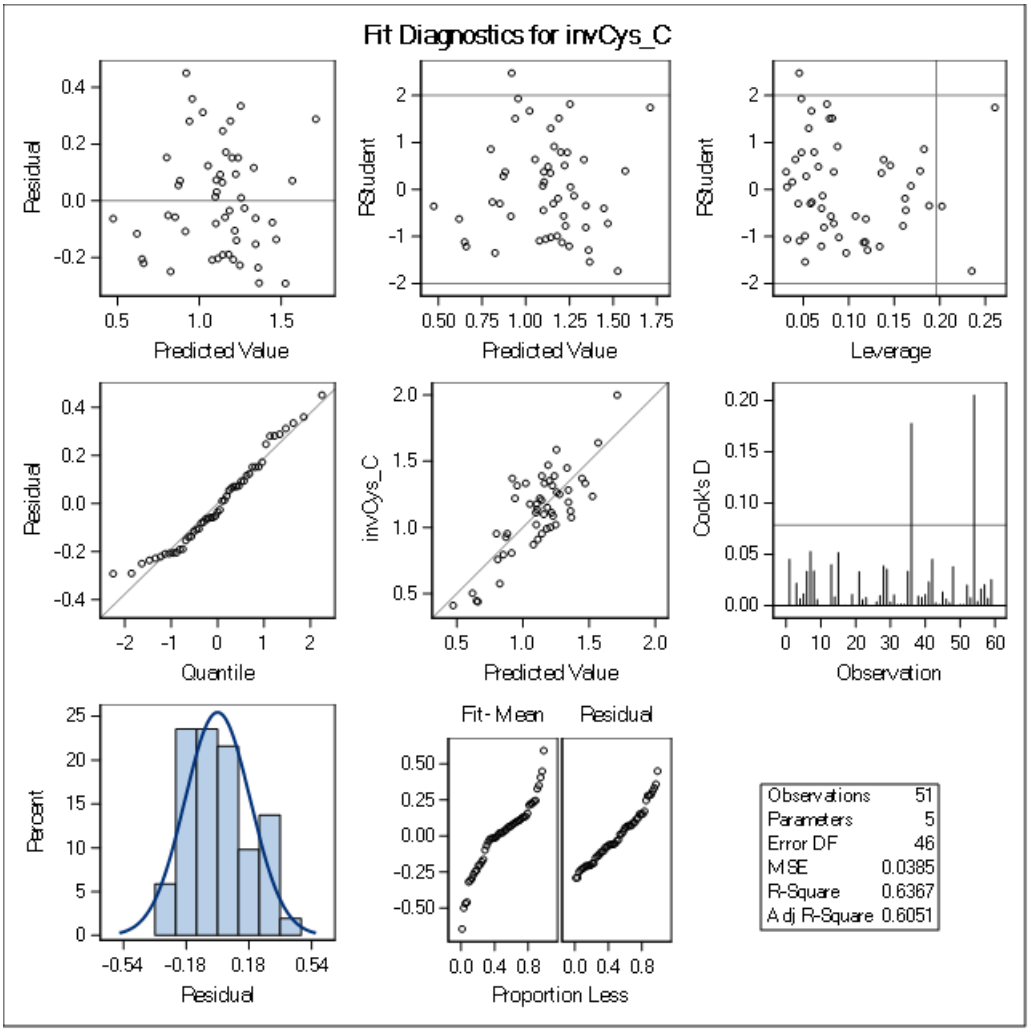
Sve nezavisne varijable su statističke nezavisne na razini značajnosti od 5%. Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model invCys_C=ALT waste klirens HDL
3     /selection=rsquare adjrsq rmse cp aic;
4  run;
```

Tablica 8.40: Rezultati procjene višestruke linearne regresije modela 2 za inverz cistacina C (ispis iz SAS-a)

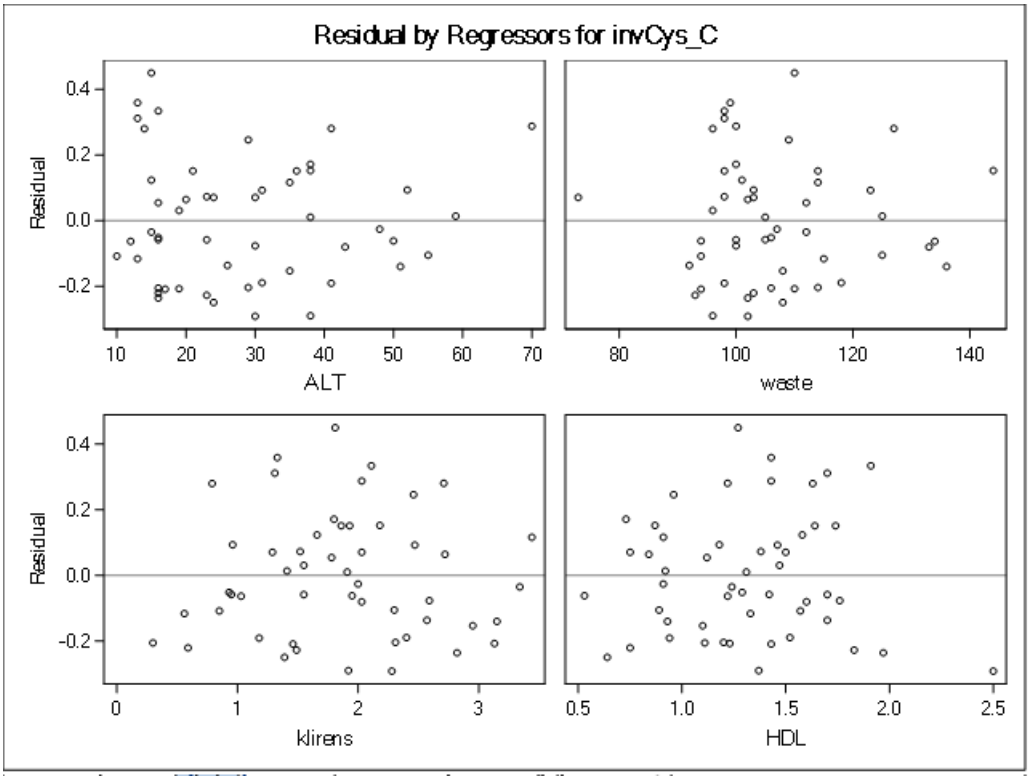
Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
4	0.6367	0.6051	5.0000	-161.4133	0.19614	ALT waste klirens HDL

Vrijednosti svih statistika su ostale približno iste.



Slika 8.20: Dijagnostika višestrukog linearnog regresijskog konačnog modela za inverz cistacina C (ispis iz SAS-a)

Dijagnostika dobivenog modela zadovoljava kriterije modeliranja. Jedino graf $PredictedValues - lnApN$ ukazuje na blago raspršenje oko pravca $lnApN = PredictedValues$.



Slika 8.21: Analiza reziduala nezavisnih varijabli višestrukog linearnog regresijskog modela za inverz cistacina C (ispis iz SAS-a)

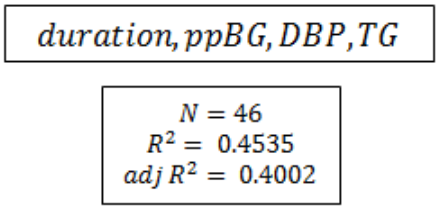
Kod reziduala možemo naslutiti blagi uzorak, ali model uvijek možemo poboljšati u suradnji sa strukom.

Dakle, dobiven je sljedeći regresijski model:

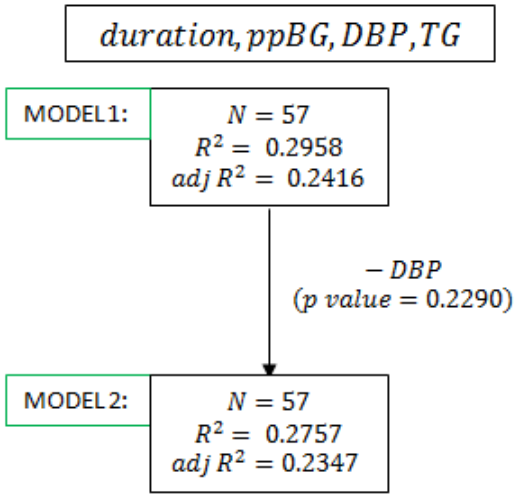
$$\frac{1}{Cys_C} = 1.29481 + 0.01110ALT - 0.01058waste + 0.19103klirens + 0.21612HDL\ .$$

8.5.4 Odabir linearnog regresijskog modela za dvostruko logaritmiranu ekskreciju albumina, AER

- *Stepwise metoda:*



- *Višestruka linearna regresija:*



Slika 8.22: Proces modeliranja linearnog regresijskog modela za dvostruko logaritmiranu ekskreciju albumina

Odabir linearnog regresijskog modela za dvostruko logaritmiranu ekskreciju albumina stepwise metodom

ULAZ:

sex age duration waste WHR fBG ppBG HbA1c ApN CRP HCY FIB IL_6 Sijal_ac_ klirens SBP DBP WBC HDL LDL TG AST ALT GGT MK Cys_C C1 FLI GDR

IZLAZ:

duration ppBG DBP TG

Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model lnlnAER=sex age duration waste WHR fBG ppBG HbA1c ApN CRP
3     HCY FIB IL_6 Sijal_ac_ klirens SBP DBP WBC HDL LDL TG AST ALT GGT
4     MK Cys_C C1 FLI GDR / selection = stepwise;
5  run;
```

Tablica 8.41: Rezultati sažetka stepwise metode za odabir modela zadvostruko logaritmiranu ekskreciju albumina (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnlnAER

Number of Observations Read	59
Number of Observations Used	46
Number of Observations with Missing Values	13

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5.96168	1.49042	8.51	<.0001
Error	41	7.18337	0.17520		
Corrected Total	45	13.14505			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.98347	0.54033	0.58042	3.31	0.0760
duration	0.01930	0.00870	0.86201	4.92	0.0321
ppBG	0.04057	0.01530	1.23204	7.03	0.0113
DBP	-0.01239	0.00626	0.68709	3.92	0.0544
TG	0.18877	0.05814	1.84692	10.54	0.0023

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	TG		TG	1	0.2338	0.2338	9.4877	13.42	0.0007
2	duration		duration	2	0.1024	0.3362	4.6082	6.63	0.0135
3	ppBG		ppBG	3	0.0651	0.4013	2.2333	4.57	0.0385
4	DBP		DBP	4	0.0523	0.4535	0.7209	3.92	0.0544

Uočimo da je varijabla *DBP* čija značajnost je bila oko 5% sada više nije značajna na puno većem nivou značajnosti. U sljedećem modelu ju izbacujemo.

Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
proc reg data=novabaza;
   model lnlnAER=sex age duration waste WHR fBG ppBG HbA1c ApN CRP
   HCY FIB IL_6 Cys_C Sijal_ac_ klirens SBP DBP WBC HDL LDL TG AST
   ALT GGT MK C1 FLI GDR/selection=rsquare adjrsq rmse cp aic;
run;
```

Tablica 8.42: Rezultati procjene modela dobivenog stepwise metodom za dvostruko logaritmiranu ekskreciju albumina (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
4	0.4535	0.4002	0.7209	-75.4162	0.41857	duration ppBG DBP TG

Višestruka linearna regresija za dvostruko logaritmiranu ekskreciju albumina (MODEL 1)

ULAZ:

duration ppBG DBP TG

Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model lnlnAER=duration ppBG DBP TG;
3  run;
```

Tablica 8.43: Rezultati višestruke linearne regresije za dvostruko logaritmiranu ekskreciju albumina varijabli odabranih stepwise metodom (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnlnAER

Number of Observations Read	59
Number of Observations Used	57
Number of Observations with Missing Values	2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4.99942	1.24986	5.46	0.0010
Error	52	11.90255	0.22890		
Corrected Total	56	16.90197			

Root MSE	0.47843	R-Square	0.2958
Dependent Mean	1.06799	Adj R-Sq	0.2416
Coeff Var	44.79733		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.87247	0.55833	1.56	0.1242
duration	duration	1	0.01813	0.00904	2.01	0.0501
ppBG	ppBG	1	0.04027	0.01551	2.60	0.0122
DBP	DBP	1	-0.00793	0.00651	-1.22	0.2290
TG	TG	1	0.07402	0.02651	2.79	0.0073

Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model lnlnAER=duration ppBG DBP TG
3     /selection=rsquare adjrsq rmse cp aic;
4  run;
```

Tablica 8.44: Rezultati procjene višestruke linearne regresije za dvostruko logaritmiranu ekskreciju albumina varijabli odabranih stepwise metodom (ispis iz SAS-a)

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
4	0.2958	0.2416	5.0000	-79.2790	0.47843	duration ppBG DBP TG

Vrijednosti procjenitelja modela su drastično pale što možemo pripisati razlici od 11 novih podataka u novoj bazi.

Višestruka linearna regresija za inverz cistacina C (MODEL 2 = MODEL 1 - DBP) - KONAČAN MODEL

ULAZ:

duration ppBG TG

Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model lnlnAER=duration ppBG TG;
3  run;
```

Tablica 8.45: Rezultati višestruke linearne regresije modela 2 za dvostruko logaritmiranu ekskreciju albumina (ispis iz SAS-a)

The REG Procedure
Model: MODEL1
Dependent Variable: lnlnAER

Number of Observations Read	59
Number of Observations Used	57
Number of Observations with Missing Values	2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4.66030	1.55343	6.73	0.0006
Error	53	12.24167	0.23097		
Corrected Total	56	16.90197			

Root MSE	0.48060	R-Square	0.2757
Dependent Mean	1.06799	Adj R-Sq	0.2347
Coeff Var	45.00038		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.24242	0.21023	1.15	0.2540
duration	duration	1	0.01987	0.00896	2.22	0.0310
ppBG	ppBG	1	0.03668	0.01530	2.40	0.0200
TG	TG	1	0.07163	0.02656	2.70	0.0094

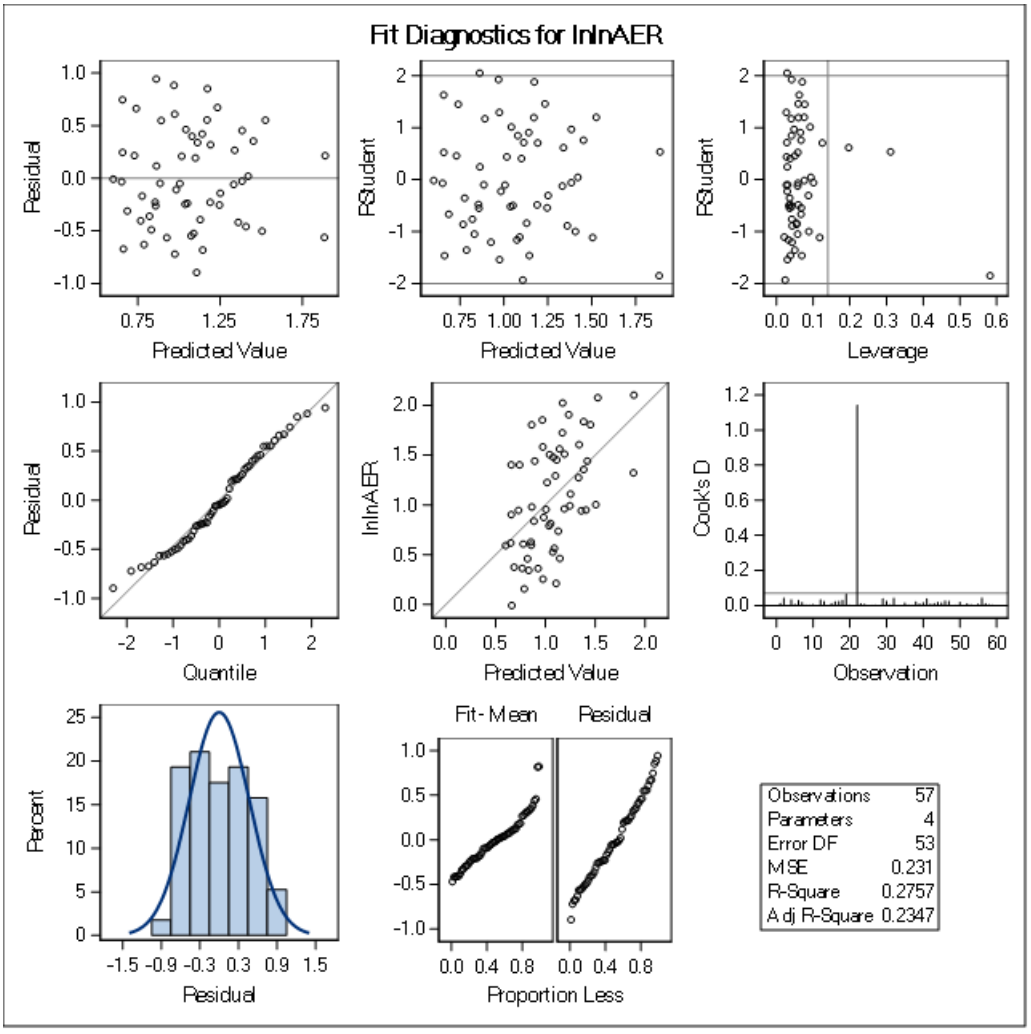
Sve nezavisne varijable su značajne na razini značajnosti od 5%.
Procjena dobivenog modela dana je u sljedećoj tablici.
Kod u SAS-u:

```
1  proc reg data=novabaza;
2     model lnlnAER=duration ppBG TG
3     /selection=rsquare adjrsq rmse cp aic;
4  run;
```

Tablica 8.46: Rezultati procjene višestruke linearne regresije modela 2 za dvostruko logaritmiranu ekskreciju albumina (ispis iz SAS-a)

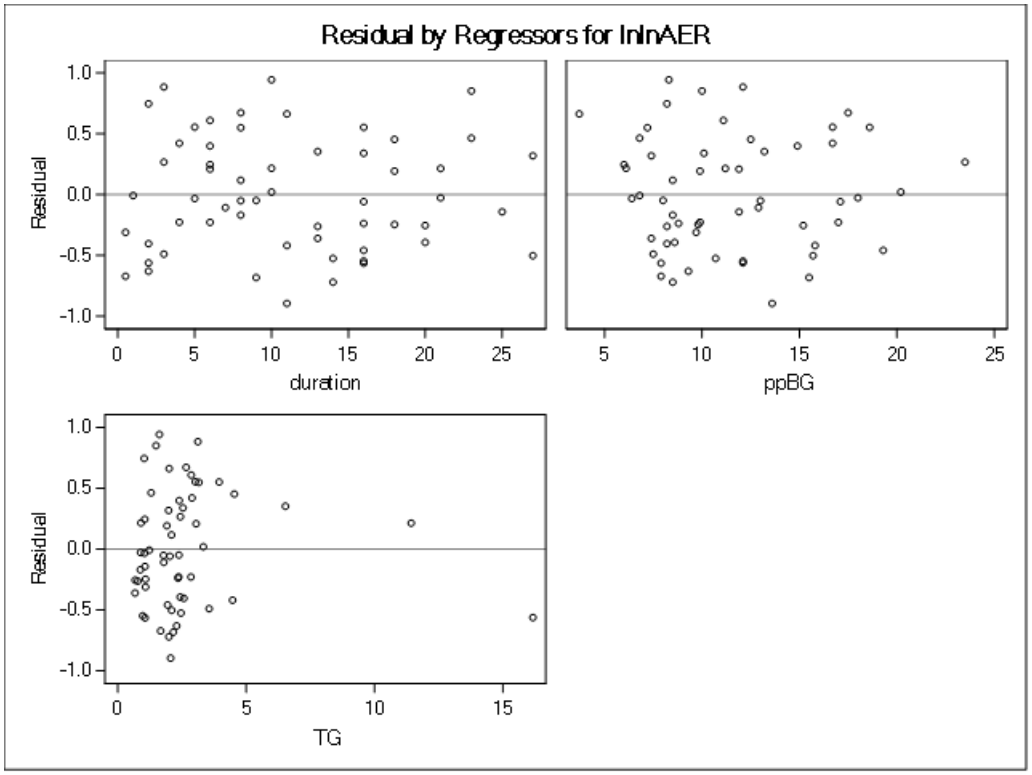
Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Root MSE	Variables in Model
3	0.2757	0.2347	4.0000	-79.6777	0.48060	duration ppBG TG

Vrijednosti statistika su se malo promijenile pa bismo mogli zaključiti da je izbacivanje varijable *DBP* bilo opravdano.



Slika 8.23: Dijagnostika višestrukog linearnog regresijskog konačnog modela za dvostruko logaritmiranu ekskreciju albumina (ispis iz SAS-a)

Dijagnostika dobivenog modela zadovoljava kriterije modeliranja. Jedino graf $PredictedValues - lnApN$ ukazuje na raspršenost i malo drugačiji položaj točaka od pravca $lnApN = PredictedValues$.



Slika 8.24: Analiza reziduala nezavisnih varijabli višestrukog linearnog regresijskog modela za dvostruko logaritmiranu ekskreciju albumina (ispis iz SAS-a)

Kod reziduala varijable TG možemo naslutiti blagi uzorak, ali model uvijek možemo poboljšati u suradnji sa strukom.

Dakle, dobiven je sljedeći regresijski model:

$$ln(ln(AER)) = 0.24242 + 0.02987duration + 0.03668ppBG + 0.07163TG.$$

Bibliografija

[1] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey
Introduction to Linear Regression Analysis, fifth edition.
John Wiley & Sons, Inc., Hoboken, New Jersey, 2012.

[2] Analytic software and solutions
http://www.sas.com/en_us/home.html
SAS University Edition:
http://www.sas.com/en_us/software/university-edition.html

[3] Anamarija Jazbec
Materijali s predavanja iz Odabranih statističkih metoda iz biomedicine.
Zagreb, 2014./2015.

Sažetak

Naglasak ovog rada je na regresijskoj analizi - teorijskoj podlozi i njezinoj praktičnoj primjeni. Regresijska analiza je jedna od najraširenijih statističkih metoda upravo zbog svoje fleksibilnosti u primjeni. Ideja je jednadžbom opisati ponašanje varijable koja nas zanima (zavisna varijabla) pomoću jednog ili više prediktora (nezavisna varijabla). Dodatne prednosti ove metode su jednostavne pretpostavke i dobro razrađena teorijska podloga.

U prvom poglavlju naznačeno je da se u radu koristi linearna regresijska analiza, tj. veza između zavisne i nezavisne varijable je linearna. U drugom poglavlju objašnjena je teorijska podloga jednostruke linearne regresijske analize, tj. način na koji pravcem možemo opisati vezu između jedne zavisne varijable i jedne nezavisne varijable. Pritom smo se dotakli problematike procjene parametara modela jednostruke linearne regresijske analize, testiranja hipoteza vezanih za slobodni član i koeficijent smjera koristeći t test i F test za analiziranje varijance (ANOVA) te procjene pouzdanih i egzaktnih intervala. Dobiveni rezultati su poopćeni u trećem poglavlju u obliku višestruke linearne regresije, tj. modeliranju jedne zavisne varijable u ovisnosti o nekoliko nezavisnih varijabli. U četvrtom poglavlju navedeni su testovi za testiranje normalnosti, dok su u petom poglavlju navedene prikladne transformacije za linearizaciju modela i stabilizaciju varijance. U svrhu provjere pretpostavki modela u šestom poglavlju je predstavljena analiza reziduala. Na samom kraju teorijskog dijela rada dotičemo se kriterija za procjenu modela te poznatijih algoritama odabira varijabli unutar modela: sve moguće regresije, forward, backward i stepwise procedura.

Osmo poglavlje sastoji se od primjene obrađene teorije na konkretan problem. Podaci se sastoje od 59 pacijenata koji boluju od dijabetesa tipa 2 te su za svakog izmjerene 29 vrijednosti. Prilikom obrade dobivenih podataka koristili smo statistički program SAS (University Edition) koji je ubrzao i olakšao proces modeliranja. Velika prednost SAS-a je u tome što je većina metoda kao što su jednostruka i višestruka linearna regresija, te metode odabira podskupa iz cijelog skupa nezavisnih varijabli već implementirane. Zadatak je bilo modelirati zavisne varijable adiponektin (ApN), homocistein (HCY), cistacin C (Cys_C) i ekskreciju albumina u urinu (AER) pomoću dane baze podataka.

Summary

In this thesis the emphasis is on regression analysis - teoretical background and its application. Regression analysis is one of the most widely used statistical method because of its flexibility in apllication. The idea is to describe behaviour of variable of interest (dependent variable) with one or more predictors (independent variable) as equation. The main advantages of this method are simple underlying assumptions and elegant mathematical theory.

In the first chapter the use of linear regression analysis is denoted, i.e. the relationship between dependent and independent variable is linear. In the second chapter theoretical background of univariate linear regression is decribed , i.e. the way we describe relationship between one dependent and one independent variable using a straight line. We have discussed the estimation of parameters from the model in univariate linear regression analysis, testing hypothesis for intercept and slope using t test and F test for the analysis of variance (ANOVA) and estimation of confidence and prediction intervals. These results are generalized in the third chapter to describe multiple regression analysis, i.e. modeling one dependent variable with several independent variables. In the fourth chapter we described testing normality of variables, while the theme of the fifth chapter are transformations of variables for linearization of model and stabilization of variance. In the sixth chapter analysis of residuals to check underlying assumptions. In the end we introduce criteria for validation of models and algorithms for variable selection and model building: all possible regressions, forward selection, backward elimination and stepwise regression.

The eighth chapter consists of application of this theory on a concrete problem. The data is from 59 pacients with diabetes type II and, for each of them, there are 29 values measured. While working with this data we use the statistical program SAS (University Edition) which makes it quicker and easier to model this problem. The main advantage of SAS is implementation of whole methods as univariate and multiple linear regression and variable selection methods. Our task was to model dependent variables Adiponektin (ApN), Homocistein (HCY), Cystacin C (Cys_C) i Albumin Ekskrection in urine (AER) using this data base.

Životopis

Zovem se Valentina Šeketa. Rođena sam 26. srpnja 1992. u Karlovcu gdje sam i odrasla.

Školske godine 1999./2000 upisala sam Osnovnu školu Dragojle Jarnević. Nakon njezinog uspješnog završetka nastavila sam daljnje srednjoškolsko obrazovanje u Gimnaziji Karlovac, opći smjer.

Nakon završetka srednje škole put me odveo u Zagreb gdje sam akademske godine 2011./2012. upisala Prirodoslovno-matematički fakultet, matematički odjel. Akademske godine 2013./2014. završila sam preddiplomski inženjerski studij te upisala diplomski studij Matematička statistika na istoimenom fakultetu.