Anita Vaš

Modeliranje višestrukom linearnom regresijom

Master rad

I VIŠESTRUKA LINEARNA REGRESIJA

1. Pojam regresije

Regresija predstavlja statističku metodu kojom se opisuje povezanost između različitih pojava. Značaj spomenute metode ogleda se u mogućnosti predviđanja ishoda određene pojave na osnovu saznanja o nekim drugim pojavama. Pojave na osnovu kojih se dobija predviđanje, $X_1, X_2,\ldots,X_k$, *populacije*, su *nezavisne (determinističke) promenljive* ili *faktori* a pojava koja zavisi od ovih promenljivih, Y zove se *zavisna (stohastička) promenljiva*. Zavisnost spomenutih pojava je data <u>*populacionim regresionim modelom*</u>:

$$Y = P_Y(X_{1,\ldots,}X_k)+E \quad \textit{svuda } P_Y \textit{ zameniti sa } \mu$$

gde je E slučajna greška.

Radi mogućnosti predviđanja potrebno je pronaći funkciju kojom bi se definisala međusobna zavisnost promenljivih. Za višestruku linearnu regresiju <u>*regresiona funkcija populacije*</u> je u obliku:

$$P_Y(X_1,X_2,..,X_k) = \beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k$$

gde su β_i, $i=0,\ldots,k$ nepoznati koeficijenti regresije,
X_j, $j=1,\ldots,k$ nezavisne deterministčke promenljive.

Razlika između stvarne vrednosti zavisnog faktora Y i predviđene vrednosti populaciji, predstavlja grešku predviđanja. Greške su skoro uvek različite od 0 pa je potrebno pronaći funkciju za koju su minimalne.

U opštem slučaju nemoguće je prikupiti podatke o celoj populaciji, stoga je nemoguće odrediti i tačan oblik regresione funkcije populacije. U tu svrhu se na osnovu uzorka prikupljaju podaci o faktorima. $X_{i,j}$, $i=1,\ldots,n$, $j=1,\ldots,k$ predstavlja vrednost j-tog faktora za i-ti elemenat u modelu, $x_{i,j}$, $i=1,\ldots,n$, $j=1,\ldots,k$ predstavlja realizovanu vrednost izabranog i-tog elementa u uzorku za j-ti faktor. Zavisnost faktora Y_i je data <u>uzoračkim regresionim modelom</u> u obliku:

$$Y_i=\hat{\beta}_0+\hat{\beta}_1X_{i,1}+\ldots+\hat{\beta}_kX_{i,k}+e_i, \quad i=1,2,\ldots,n$$

gde su $\hat{\beta}_j$, $j=0,\ldots,k$ ocene koeficijenata,
$X_{i,j}$, $i=1,\ldots,n$, $j=1,\ldots,k$ nezavisna deterministička promenljiva j za elemenat i,
e_i, $i=1,\ldots,n$ – reziduali (mogu se posmatrati kao ocene grešaka E_i).

<u>Uzoračka linija regresije</u> predstavlja funkciju kojom je izražena zavisnost promenljivih u uzorku:

$$\mu_Y(X_1,X_2,..,X_k)=\hat{\beta}_0+\hat{\beta}_1X_{i,1}+\hat{\beta}_2X_{i,2}+\ldots+\hat{\beta}_kX_{i,k}$$

gde su $\hat{\beta}_i$, $i=0,\ldots,k$ ocenjeni koeficijenti,
$X_{i,j}$, , $i=1,\ldots,n$, $j=1,\ldots,k$ nezavisna deterministička promenljiva *j* za *i*-ti elemenat u populaciji.

Razlika između stvarne vrednosti *i*-tog elementa u uzorku, Y_i i njegove predviđene vrednosti $\mu_Y(i)$, je skoro uvek različita od 0 i predstavlja grešku predviđanja za uzorak, e_i.

Naziv *regresija* uveo je statističar Sir Frances Galton (1822-1911) i znači vraćanje unatrag. Ispitivanjem veze između visina roditelja i dece ustanovio je da visoki roditelji teže da imaju visoko dete ali ne toliko koliko i oni sami. Otuda naziv vraćanje unatrag tj. regresija prema proseku. Rezultat ispitivanja je podatak da će dete skoro uvek biti visoko između visina roditelja, muška deca bliža očevoj visini a ženska deca bliža majčinoj visini.

Naziv *višestruka linearna regresija* znači:
višestruka- ima više nezavisnih promenljivih X
linearna- regresiona funkcija je linearna po koeficijentima β
regresija- koristi se regresiona funkcija kao najbolje predviđanje za Y na osnovu X_i, $i=1,..,n$

Cilj ispitivanja višestrukom linearnom regresijom je definisati uzoračku liniju regresije sa najmanjim mogućim rezidualima. U tu svrhu potrebno je oceniti nepoznate koeficijente regresije β_i, $i=0,\ldots,k$ na neki od sledećih načina:

1. *Tačkasta ocena* – Metodom tačkastog ocenjivanja dobija se ocena, statistika, $\hat{\beta}_i$, $i=0,\ldots,k$, za svaki nepoznati koeficijent β_i, $i=0,\ldots,k$. Na osnovu realizovanog uzorka dobijaju se realizovane vrednosti ovih statistika, b_i, $i=0,\ldots,k$. Ukoliko je ocenjena vrednost blizu stvarnoj vrednosti koeficijenta, navedena metoda predstavlja dobar način ocene. U daljem radu će biti objašnjeno par metoda za tačkasto ocenjivanje.

2. *Intervalna ocena*- Intervalnom ocenom dobija se interval poverenja koji sa verovatnoćom $1-\alpha$ sadrži željeni parametar, koeficijent. Realizovani interval na osnovu uzorka ili sadrži ili ne sadrži koeficijent, stoga se samo za slučajni interval kaže da sa $1-\alpha$ sigurnosti sadrži koeficijent. Postoje dvostrani i jednostrani intervali poverenja. Dvostrani su dati u obliku $C[L<\theta<U]= 1-\alpha$, gde je L donja granica a U gornja granica intervala poverenja i obično su u obliku

$$\hat{\theta} - \text{tablična vrednost} * SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + \text{tablična vrednost} * SE(\hat{\theta})$$

$\hat{\theta}$- ocenjeni parametar
$\mathrm{SE}(\hat{\theta})$ – standardna greška parametra.

3. *Testiranje hipoteza pomoću testova*-Ukoliko istražitelj želi da odredi da li je parametar tj koeficijent veći ili manji od određene vrednosti q, mogu se koristiti testiranje hipoteza i testovi. Postavljaju se dve hipoteze za parametar θ, nulta hipoteza NH i alternativna hipoteza AH koje mogu biti u obliku:

	NH	AH
1)	$\theta=q$	$\theta\neq q$
2)	$\theta\leq q$	$\theta>q$
3)	$\theta\geq q$	$\theta<q$

Cilj testiranja je utvrditi da li ima dokaza za odbacivanje hipoteze NH. U tu svrhu fiksira se veličina testa α (obično je *0.05* ili *0.01*) i p-vrednost: veličina kritične oblasti ako joj je granica registrovana vrednost test statitstike. Ukoliko je $p \leq \alpha$ odbacuje se hipoteza NH, a ako je $p > \alpha$ hipoteza NH se ne odbacuje.

2. Poželjne osobine ocena parametara

Za ocene koeficijenata β_i, $i=0,\ldots,k$ poželjno je da imaju sledeće osobine:

1. Nepristrasnost – Ocena je nepristrrasna ako je očekivana vrednost ocene $\hat{\beta}$ jednaka pravoj vrednosti β,

$$E(\hat{\beta})=\beta.$$

Ukoliko postoji razlika između ove dve vrednosti, $\hat{\beta}$ je pristrasna a $E(\hat{\beta})-\beta$ predstavlja pristrasnost ocene.

2. Efikasnost- Ocena je najefikasnija ako je nepristrasna I ima najmanju disperziju među svim ostalim nepristrasnim ocenama istog parametra. U zavisnosti na osnovu čega se posmatra, efikasnost može biti apsolutna i relativna. Ukoliko postoje dve ocene uzorka tada se za ocenu koja ima manju disperziju kaže da je relativno efikasnija. Za pronalaženje najefikasnije ocene koristi se teorema Rao-Cramera.

$$D(\beta) \geq \frac{1}{nE\left(\frac{\partial}{\partial \beta} lnL(X_i,\sigma^2)\right)^2}$$

n- veličina uzorka
$L(X_i, \sigma^2)$ - funkcija verodostojnosti.

U slučaju višestruke linearne regresije Y_i ima normalnu raspodelu

$$Y_i : \mathcal{N}(\beta_0+\beta_1 X_{i,1}+\ldots+\beta_k X_{i,k}, \sigma^2)$$

Funkcija verodostojnosti za Y_i je: $$L = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(\frac{-\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_{i1}-\beta_2 X_{i2}-\cdots-\beta_k X_{ik})}{\sigma}\right)^2}$$

a logaritam te funkcije:

$$lnL = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})^2$$

Sumirano - $\hat{\beta}$ je efiksan ocean za β ako je:

- Nepristrasna,
- Ima manju disperziju od svih ostalih ocena.

3. BLUE (best linear unbiased evaluator) –najbolja linearna nepristrasna ocena. Ocenna sa ovom osobinom treba da zadovolji sledeće:

- Ocenaje linearna funkcija opažanja.
- Ocena je nepristrasna.
- Ocena ima najmanju disperziju od svih ostalih ocean.

Navedene osobine važe samo u slučaju kada je uzorak konačan. Mogu se primenjivati i u slučaju malog uzorka (ispod 30 opažanja) jer i dalje daju primenljive rezultate. Ocena ima fiksnu raspodelu bez obzira na veličinu uzorka. U slučaju velikih uzoraka navedene osobine ne daju dobre rezultate nego se moraju primeniti drugačije osobine. Sledeće osobine važe kada se uzorak može beskonačno mnogo povećavati kako bi se našla tačna raspodela ocene tj data ocena ima različite raspodele u zavisnosti od veličine uzorka.

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

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

2. Konzistentnost- odrazumeva da raspodela ocenjivača s porastom uzorka posrne??????? u svoju pravu vrednost. U praksi taj uslov važi ako je

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

$MSE(\hat{\beta}) = E(\hat{\beta} - E(\hat{\beta}))^2 + (E(\hat{\beta}) - \beta)^2$- sredina kvadrata greške

3. Asimptotska efikasnost-podrazumeva da ocena koeficijenta ima sledeće osobine:

- Asimptotsku distribuciju s konačnom sredinom i varijansom.
- Konzistentnost .
- Najmanju asimptotsku varijansu ($\frac{1}{n}\lim_{n\to\infty} nMSE(\hat{\beta})$) od svih ostalih ocean.

3. Metode tačkastog ocenjivanja koeficijenata β_i, $i=0,\ldots,k$

Za dobijanje regresione linije uzorka potrebno je oceniti koeficijente β_i, $i=0,\ldots,k$ tačkastom metodom kako bi se dobijene konkretne vrednosti mogle uvrstiti u jednačinu. Različitim metodama se postižu različite osobine spomenutih koeficijenata. Najčešće korišćene tačkaste metode su:

- Metoda najmanjih kvadrata.
- Metoda maksimalne verodostojnosti.
- Metoda momenata.
- Najbolje linearno nepristrasno ocenjivanje.

3.1 Metoda najmanjih kvadrata

Metoda najmanjih kvadrata je jedna od najstarijih metoda. Suma kvadrata razlike zavisnog faktora i njegovog očekivanja za svaki elemenat uzorka treba da je što manja.

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})^2 \rightarrow \min$$

Minimum date sume se nalazi diferenciranjem sume po svakom koeficijentu β_i, $i=0,\ldots,k$ i izjednačavanjem s 0. Kada se reši tako dobijeni sistem od *k+1* jednačine dobijaju se ocene koeficijenata β_i, $i=0,\ldots,k$. Dobijeni sistem jednačina se naziva sistem normalnih jednačina. U matričnom zapisu ovaj sistem je dat sa

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

gde je $X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots\ldots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \ldots\ldots & x_{2,k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots\ldots & x_{n,k} \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$ - ocene parametara β_i, $i=0,\ldots,k$

U opštem slučaju ocenjeni parametri preko ove metode imaju nepoznate osobine pa se u mnogim slučajevima retko koristi. Kod višestruke linearne regresije ovom metodom se dobijaju ocene koeficijenata β_i, $i=0,\ldots,k$ ali ne i njihove varijanse ili varijansa zavisnog faktora. Ocene dobijene navedenom metodom imaju sve poželjne osobine (u daljem radu će biti objašnjeno zbog čega.)

3.2 Metoda maksimalne verodostojnosti

Polazi se od pretpostavke da svaki uzorak dolazi iz određene populacije i da je verovatnoća da je uzorak iz određene populacije najveća za populaciju kojoj uzorak pripada. Stoga se definiše funkcija maksimalne verodostojnosti kao raspodela populacije kojoj uzorak najverovatnije pripada. U slučaju višestruke linearne regresije to je normalna raspodela jer se pretpostavlja da zavisna promenljiva ima normalnu raspodelu.

$$L = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(\frac{-\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})}{\sigma}\right)^2}$$

Ponekad je lakše raditi sa logaritmom funkcije verodostojnosti, u konretnom slučaju se dobija:

$$lnL = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})^2$$

Diferencirajući gornju funkciju po koeficijentima β_i, $i=0,\ldots,k$ i odstupanju σ^2 i izjednačavajući s 0 dobija se sistem jednačina. Rešavanjem ovakvog sistema dobijaju se ocene za β_i, $i=0,\ldots,k$ ali i za σ^2. Ocene za β_i, $i=0,\ldots,k$ su iste kao i kod metode najmanjih kvadrata a ocena za σ^2 je:

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

Ocene parametara dobijene ovom metodom imaju osobine:

- Konzistentnost.
- Asimptotska efikasnost.

3.3 Metoda momenata

Ideja metode momenata je da se momenti raspodele populacije ocene preko momenata uzorka (na primer- sredina populacije se ocenjuje sredinom uzorka momentom prvog reda). U slučaju višestruke linearne regresije koriste se sledeći momenti:

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

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

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

Iz prve i treće jednačine dobijaju se ocene za koeficijente β_i, $i=0,\ldots,k$ a iz druge za σ^2. Ocene su potpuno identične onim dobijenim metodom maksimalne verodostojnosti.

Ocene parametara dobijene ovom metodom imaju osobine:

- Konzistentnost.
- Asimptotska normalnost.

3.4 Najbolja linearna nepristrasna ocena

Samo ime ove metode navodi da ocene dobijene ovom metodom su linearne po promenljivama, nepristrasne i imaju najmanju disperziju. U slučaju višestruke linearne regresije se dobija sledeći problem:

Pretpostavlja se da se traži ocena $\hat{\beta}_p$ za proizvoljno β_p, $p=0,\ldots,n$, (za svaki parametar se radi analogno). Ta ocena teba da ispunjava sledeće uslove:

- $\hat{\beta}_p = \sum_{i=1}^{n} a_i Y_i$ – ocena $\hat{\beta}_p$ je linearana po zavisnim promenljivama Y_i, $i= 1,\ldots,n$.

- Ocena je nepristrasna:

$$E(\hat{\beta}_p) = E\left(\sum_{i=1}^{n} a_i Y_i\right) = E\sum_{i=1}^{n} a_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \cdots + \beta_k X_{ik} + e_i)$$

$$= \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} a_i X_{i1} + \cdots + \beta_p \sum_{i=1}^{n} a_i X_{ip} + \cdots + \beta_k \sum_{i=1}^{n} a_i X_{ik} = \beta_p \ .$$

- Ocena je najefikasnija:

$$Var(\hat{\beta}_p) = \sum_{i=1}^{n} a_i^2 \, Var(Y_i) = \sigma^2 \sum_{i=1}^{n} a_i^2 \rightarrow \min.$$

Da bi ovi uslovi bili ispunjeni, $\hat{\beta}_p$ mora da važi:

$$\rightarrow \sum_{i=1}^{n} a_i = 0$$
$$\sum_{i=1}^{n} a_i X_{ij} = 0 \ , j \neq p$$
$$\sum_{i=1}^{n} a_i X_{ip} = 1$$

Zbog toga se ocena za $\hat{\beta}_p$ sa svim traženim osobinama dobija rRešavanjem sledećeg problema

$\sum_{i=1}^{n} a_i^2 \rightarrow min$
$\sum_{i=1}^{n} a_i = 0$
$\sum_{i=1}^{n} a_i X_{ip} = 1$
$\sum_{i=1}^{n} a_i X_{ij} = 0 \ , j \neq p$

Dati problem se može rešiti metodom Lagranžovih množitelja. Lagranžova funkcija je:

$$\mathcal{L} = \sum_{i=1}^{n} a_i^2 - \lambda_1 \sum_{i=1}^{n} a_i - \lambda_2 (\sum_{i=1}^{n} a_i X_{ip} - 1) - \lambda_m \sum_{i=1}^{n} a_i X_{ij}, \ m=3,\ldots,k+2, j \neq p$$

Diferencirajući Lagranžovu funkciju po svim koeficijentima a_i $i=1,\ldots,n$, i λ_j, $j=1,\ldots,k+2$ i izjednačavajući sa 0 dobija se sistem od $n+k+2$ jednačine.

Ocene koeficijenata β_i, $i=0,\ldots,k$ dobijene ovom metodom su identične ocenama dobijenim preko predhodnih metoda. Ovaj metod još pruža mogućnost da se dobije odstupanje ocene parametara β_i, $i=0,\ldots,k$.

Ocene dobijene ovom metodom imaju osobine:

- Linearnost.
- Nepristrasnost.
- Efikasnost.

Pošto su za višestruku linearnu regresiju dobijene identične ocene po bilo kom metodu zaključuje se da je dovoljno koristiti metod najmanjih kvadrata za dobijanje oceni sa svim poželjnim osobinama.

Teži se da svaki model poseduje ocene sa svim poželjnim osobinama kako bi se mogli izvoditi pouzdani zaključci.

Metod najmanjih kvadrata, pod određenim uslovima, daje takve ocene. Spomenuti uslovi su *osnovne pretpostavke* koje moraju biti ispunjene da bi navedena metoda davala dovoljno dobre ocene.

<u>*Osnovne pretpostavke*</u>:

- $E(E_i)=0$ -očekivana vrednost grešaka je 0
- $Var(E_i)=\sigma^2$ -sve greške imaju istu disperziju - homoskedastičnost
- $Cov(E_i,E_j)=0$ -greške su međusobno nezavisne
- $E_i : N(0, \sigma^2)$ -greške imaju normalnu raspodelu

Osnovne pretpostavke se odnose na populaciju pa ih je zbog toga teško proveriti. Poželjno je dodati i pretpostavke na sam uzorak, na primer da je *prost slučajan uzorak* kao i da su *vrednosti podataka uzete bez grešaka* . Sem nevaženja spomenutih pretpostavki može se javiti još niz poteškoća koje će biti objašnjenje u daljem radu.

4. Potencijalne poteškoće u modelu i njihovo prevazilaženje

4.1 Narušavanje osnovnih pretpostavki

- Ukoliko ne važi pretpostavka da je $E(E_i)=0$, $i=1,\dots,N$, ocene koeficijenata β_i, $i=1,..,k$ postaju pristrasne. Povećanjem uzorka može se smanjiti stepen pristrasnosti. Ukoliko greške zavise samo od nezavisnih promenljivih X_i, $i=1,..,k$ a ne i od drugih slučajnih uticaja tada ocene imaju sve poželjne asimptotske osobine.

- Ukoliko ne važi pretpostavk homoskedastičnosti, $Var(E_i)=\sigma^2$, $i=1,\dots,N$ greške su *heteroskedastične.* Većina dobrih osobina je očuvana međutim ne i efikasnosti, ni obična ni asimptotska, kao ni nepristrasnosti ocenjenih varijansi ocena, $D(\hat{\beta}_p)$, $p=0,\dots,k$. Zbog toga se ne mogu koristiti ni testovi ni intervali poverenja. Ukoliko postoji neka pretpostavka o obliku odstupanja varijansi grešaka σ_i^2 , $i=1,\dots,N$, može se napraviti model čije ocene koeficijenata β_i, $i=1,..,k$ imaju sve asimptotske osobine. U tom slučaju se koristi izmenjena metoda najmanjih kvadrata tj *metoda ponderiranh najmanjih kvadrata* koja uzima u obzir razlike u varijansama i na osnovu toga im daje veličinu uticaja.

$$\sum_{i=1}^{n} \omega_i (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})^2 \rightarrow \min$$
$$\omega_i = \frac{1}{\sigma_i^2} - \text{težine}$$

  Na ovaj način se dobijaju ocene koeficijenata β_i, $i=1,..,k$ koje imaju sve poželjne osobine i mogu se koristiti intervali poverenja kao i testovi.

- Ukoliko ne važi pretpostavka da je $Cov(E_i,E_j)=0$, $i=1,\dots,N$, $j=1,\dots,N$,došlo je do *autokoreliranosti*. Obično se javlja kod vremenskih serija kada greška jednog razdoblja utiče na greške narednih razdoblja. Ovo nije karakteristično za podatke uzete u jednom trenutku u vremenu, pa neće biti detaljnije objašnjeno.

- Ukoliko ne važi pretpostavka da $E_i : \mathcal{N}(0, \sigma^2)$, $i=1,\dots,N$, ostaju sve poželjne osobine osim efikasnosti. Efikasnost se može dobiti dobijanjem ocena preko *metode minimiziranja sume apsolutnih vrednosti*: $\sum_{i=1}^{n} |Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik}| \rightarrow \min$
  Ova metoda se koristi samo u slučaju kad je navedena pretpostavka narušena. U slučaju važenja pretpostavki ova metoda daje neefikasne ocenjivače.

Pošto se greške populacije ne mogu izmeriti nego samo oceniti na osnovu uzorka, pretpostavke se obično proveravaju na greškama uzorka tj rezidualima, e_i, koji su i sami ocene. U praksi se koriste

standardizovani reziduali koji se dobijaju standardizovanjem reziduala. Zbog standardizovanja se zna da imaju normalnu raspodelu sa sredinom 0 i odstupanjem 1.

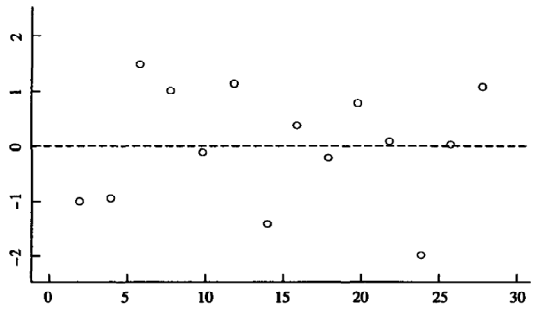
Standardizovani reziduali: $r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{i,i}}}$, $i=1,\ldots,n$ $h_{i,i}$ je dijagonalni elemenat matrice $H=X(X^TX)^{-1}X^T$

Postoje nekoliko grafičkih metoda provere koji su veoma praktični u konkretnim primerima. Mnogi statistički programi imaju ugrađene ove metode pa je stoga vrlo lako doći do njih. Poteškoće se mogu javiti prilikom tumačenja istih. Testovi daju jednoznačne rezultate međutim ponekad zahtevaju dodatne informacije koje nije moguće uvek prikupiti.

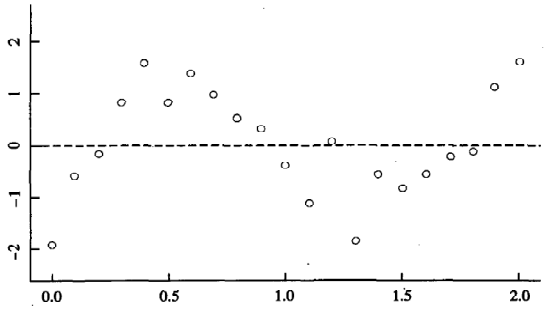
4.1.1 Grafičke metode

4.1.1.1 Upoređivanje standardizovanih reziduala i nezavisnih promenljivih

Ako se ucrtaju tačke s koordinatama $(r_i,x_{i,j})$, $i=1,\ldots,n$,$j=1,\ldots,k$ (za svako j se crta poseban grafik, r_i su obeleženi na *x*-osi a $x_{i,j}$ na *y*-osi) na grafiku tada se može oceniti da li su osnovne pretpostavke zadovoljene. Ako iscrtane tačke ne obrazuju konkretan šablon u odnosu na krivu $y=0$, pretpostavke su ispunjene (slika 1). Ako se vidi šablon tada pretpostavke verovatno ne važe (slika 2).



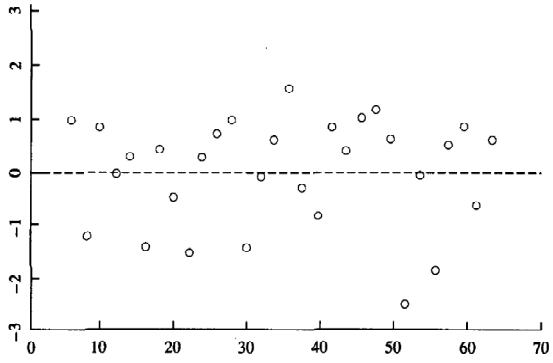
slika 1



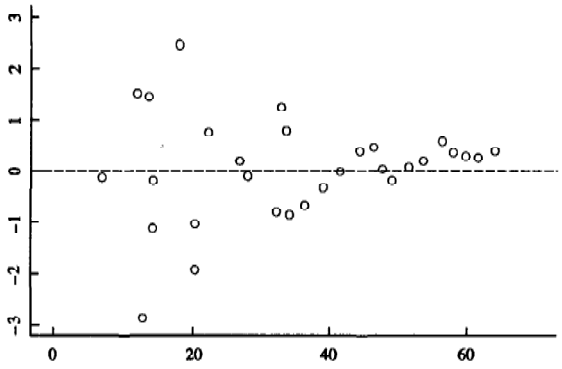
slika 2

4.1.1.2 Upoređivanje standardizovanih reziduala i ocenjenih vrednosti μ_Y

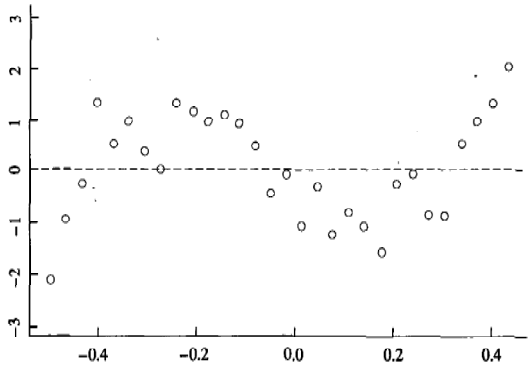
Ukoliko su zadovoljene osnovne pretpostavke, ocenjena vrednost zavisne promenljive Y_i, $i=1,\ldots,n$, $\mu_Y(i)$ je nezavisna od standardizovanih reziduala r_i , $i=1,\ldots,n$ što se manifestuje odsustvom šablona u odnosu na krivu $y=0$ (slika 3). Ukoliko nisu zadovoljene pretpostavke tada se može uočiti šablon. U odnosu na izgled šablona može se pretpostaviti koja pretpostavka nije ispunjena. Ukoliko je zgusnutost tačaka različita tada je verovatno došlo do heteroskedastičnosti (slika 4). Ako se primeti da tačke obrazuju nelinearnu krivu tada verovatno regresiona funkcija uopšte nije linearna (slika 5). U tom slučaju je najbolje promeniti oblik regresione funkcije radi dobijanja boljeg modela.



slika 3



slika 4

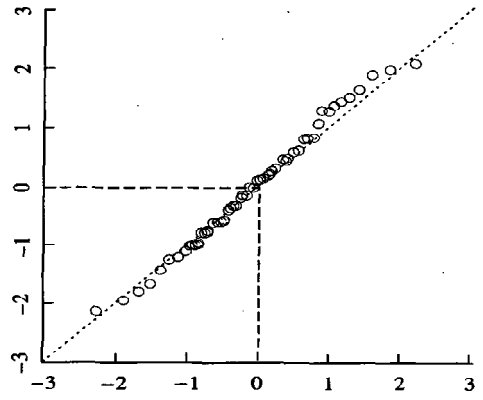


slika 5

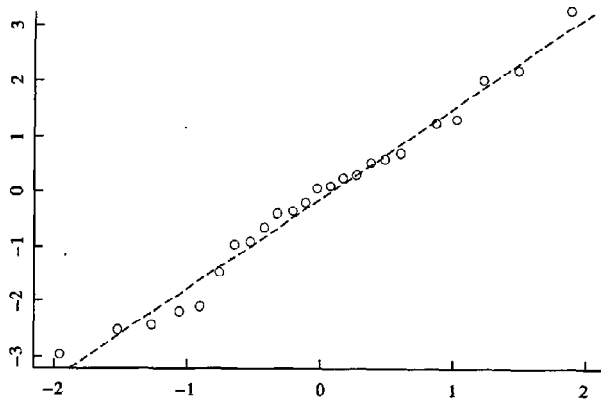
Na slici 4 se primeti da su varijanse veće za manje vrednosti promenljive Y odnosno manje za veće vrednosti promenljive Y. U ovom slučaju se može pretpostaviti oblik odstupanja pa bi se mogla iskoristiti ta informacija radi dobijanja efikasnih ocena.

4.1.1.3 Upoređivaje standardizovanih reziduala sa normalnom raspodelom

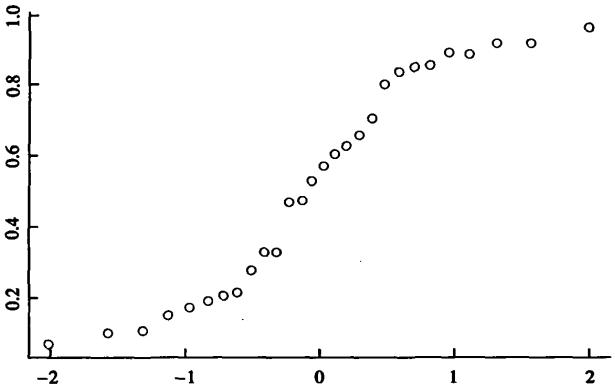
Jedna od pretpostavki je da greške E_i, $i=1,\ldots,N$ imaju normalnu raspodelu. Ukoliko to važi onda su standardizovani reziduali prost slučajan uzorak standardizovane normalne raspodele. To se može grafički proveriti upoređujući standardizovane reziduale s normalnom raspodelom. Potrebno je poređati standardizovane reziduale od najmanjeg do najvećeg i tako ih uporediti sa z vrednostima, koje su takođe poređane od najmanje do najveće, i koje imaju normalnu raspodelu. Tačke (r_i,z_i) $i=1,\ldots,n$ ukoliko zadovoljavaju normalnu raspodelu se nalaze na pravoj liniji sa nagibom 1 koja prolazi kroz koordinatni početak (slika 6). Može se desiti da postoji normalna raspodela ali sa drugačijom sredinom i odstupanjem (slika 7). U nekim slučajevima uopšte ni nema normalne raspodele (slika 8).Postoji ugrađena funkcija plot-fitting u statistici u kojoj se nalaze i *z* vrednosti. Sem spomenute funkcije može se crtati stubičast grafik radi provere normalnosti.



slika 6



slika 7



Slika 8

Na slici 7 je dat uzorak sa normalnom raspodelom sa sredinom 2 i odstupanjem 3.

4.1.1.4 <u>Upoređivanje vrednosti zavisnog faktora Y sa normalnom raspodelom</u>

Promenljiva Y može da se posmatra kao slučajna promenljiva koja ima normalnu raspodelu

$$\mathcal{N}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \cdots + \beta_k X_{ik}, \sigma^2)$$

Ako je uzorak reprezentativan, vrednosti y_i, $i=1,\ldots,n$ bi trebale da potiču iz normalne raspodele. Vrednosti y_i, $i=1,\ldots,n$ se upoređuju sa z vrednostima kao što je objašnjeno u prethodnoj metodi. Dobijeni grafici se interpretiraju na način objašnjen kod prethodne metode.

Prednost ovih metoda je lako dobijanje rezultati a potencijalna mana zaključak rezultata na osnovu mišljenja istraživača. Potrebno je iskustvo za određivanje stepena postojanja šablona.

Izvršenjem što većeg broja testova povećava se verovatnoća tačnog zaključka o ispunjenosti osnovnih pretpostavki.

4.1.2 Testovi

- Za proveravanje pretpostavke o normalnosti grešaka koristi se test poređenja. Koeficijenti β_i, $i=0,..,k$ se ocene pomoću metode najmanjih kvadrata i sume apsolutnih vrednosti. Ukoliko su ocene koeficijenata približno jednake onda nije narušena pretpostavka o normalnosti grešaka.

- Za proveravanje slučaja heteroskedastičnosti koriste se sledeći testovi:

  Goldfeld-Quandtov test se koristi kada se znaju vrednosti varijansi svakog opažanja. U tom slučaju se elementi poređaju po rastućem nizu njihovih varijansi grešaka. Iz tog niza se iz sredine izbaci šestina elemenata i od preostalih elemenata se uporedi varijansa prvog dela uzorka i drugog. Ukoliko varijanse nisu značajno različite onda nije došlo do heteroskedastičnosti. Ova metoda je dobra samo ukoliko postoji saznanje o obliku ili vrednostima varijanse.

  Breusch-Paganov test se bazira na pretpostavci da će se različitim metodama dobiti iste ocene ukoliko nije došlo do heteroskedastičnosti. Na primer, ako se pomoću metode maksimalne verodostojnosti krenu tražiti ocene koeficijenata β_i, $i=0,..,k$ uz pretpostavku da postoji heteroskedastičnost i u derivacije se ubace ocene iz metode najmanjih kvadrata pa i dalje jednačine ne budu signifikantno različite od 0 tada nije došlo do heteroskedastičnosti.

  Whiteov test se bazira na varijansama dobijenim pod pretpostavkom homoskedastičnosti i heteroskedastičnosti. Ukoliko se dobijene varijanse signifikantno ne razlikuju tada nije došlo do heteroskedastičnosti.

U većini slučajeva ukoliko se uzme dovoljno veliki uzorak mogu se prevazići ove poteškoće.Dokazano je da sa povećanjem broja uzorak teži normalnoj raspodeli. Osim ako zaista sama populacija nije adekvatna, povećanjem uzorka i odabirom odgovarajuće metode za ocenu koeficijenata dobiće se odgovarajući model.

4.2 Ostale poteškoće

4.2.1 Izbacivanje bitnog nezavisnog faktora

Bilo da je svesno ili nesvesno izbačena nezavisna promenljiva X_p, dolazi do pristrasnosti ocena ostalih koeficijenata β_i, $i=0,..,k$, $i\neq p$. Ukoliko je izbačen nezavisan faktor koreliran sa nekim od nezavisnih faktora u modelu tada može doći do nekonzistentnosti. Neizčezavanje ove veze sa povećanjem uzorka dovodi do nekonzistentnosti. U tom slučaju povećanje uzorka ne pomaže već samo mogućnost ubacivanja bitnog faktora. Nedostatak bitnog faktora u modelu se može manifestovati nedovoljnom adekvatnošću modela (na primer nizak $\bar{R}^2$-definicija data dalje u radu). Ukoliko se pokaže da za nedovoljnu adekvatnost nije zaslužen uzorak tada je verovatno izostavljena nezavisna promenljiva.

4.2.2 Uključivanje nebitnog nezavisnog faktora

Ukoliko je uključen nebitan nezavisan faktor u model ne dolazi do velikih poteškoća kao kod prethodnog slučaja. Ocene koeficijenata β_i, $i=0,..,k$ imaju sve poželjne osobine osim efikasnosti što znači da će imati znatno veće disperzije. Međutim mogu se i dalje koristiti intervali poverenja kao i

testovi značajnosti. U praktičnom radu se vrlo lako ispituje da li je određeni nezavisan faktor nebitan. Među uspešnim proverama spadaju *t-test* i upoređivanje $\bar{R}^2$ modela sa i bez spornog nezavisnog faktora. Spomenute provere biće detaljnije objašnjene u daljem radu.

4.2.3 Pogrešan oblik regresijske jednačine

Ukoliko je izabran pogrešan oblik regresijske jednačine model neće biti dovoljno adekvatan. Varijacije faktora Y neće se moći dovoljno dobro objasniti varijacijama nezavisnih faktora. Pojam adekvatnog modela biće detaljnije analiziran u daljem radu. Za sada je dovoljno naznačiti da se odabirom pogrešnog oblika jednačine dobija manje dobar model.

Do sada su prikazane poteškoće koje se mogu javiti prilikom modelovanja sa izabranim uzorkom. Kao opšte rešenje je par puta spomenuto povećanje veličine uzorka. Zanimljivo je da postoje elementi koji imaju ogroman uticaj na model. Od tih elemenata uzorka ponekad zavisi ceo model pa je bolje i izbaciti ih nego dozvoliti da model striktno zavisi od njih. Ove vrednosti su *outlieri* i *značajni elementi*.

4.2.4 Outlier

Ukoliko je konkretna vrednost zavisnog faktora y_i, $i=1,\dots,n$ za neki elemenat značajno različita od ocenjene vrednosti $\mu_Y(i)$, $i=1,\dots,n$ tada se za taj elemenat kaže da je outlier. Ova vrednost može da se javi zbog različitih razloga : pogrešan zapis vrednosti, pogrešan podatak ili stvarna neuobičajena vrednost (na primer: ljudska visina je obično u opsegu od 160-200cm, ljudi koji su viši tj niži od ovih vrednosti spadaju u outliere). Ukoliko se ovakav elemenat javi u uzorku potrebno je analizirati zbog čega se pojavio outlier. Ako je došlo do greške u unosu ili prilikom merenja, ovaj podatak se može izbaciti. Ako nije moguće otkriti uzrok prave se modeli sa i bez spomenutog elementa. Ukoliko su modeli jednaki, elemenat ne utiče bitno. Međutim ako se modeli bitno razlikuju potrebno je daljnje ispitivanje, još pokušati pronaći razloga za pojavu outliera ili uzrti novi uzork. Nakon analize se donosi odluka da li zadržati dati elemenat ili izbaciti ga iz daljeg modeliranja. Ako se zna da podaci potiču iz populacije sa normalnom raspodelom $\mathcal{N}(m,\sigma^2)$ tada su svi elementi van intervala $(m-3\sigma^2, m+3\sigma^2)$ outlieri. Obično samo 1% podataka se nalazi van ove granice i oni se uzimaju u dalju analizu modela. Najpoznatiji načini za otkrivanje outliera su:

- Standardizovani reziduali: $r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{i,i}}}$, $i=1,\dots,n$ $h_{i,i}$ je dijagonalni elemenat matrice H
  Ako je $|r_i|>2$ tada je i-ti elemenat mogući outlier.

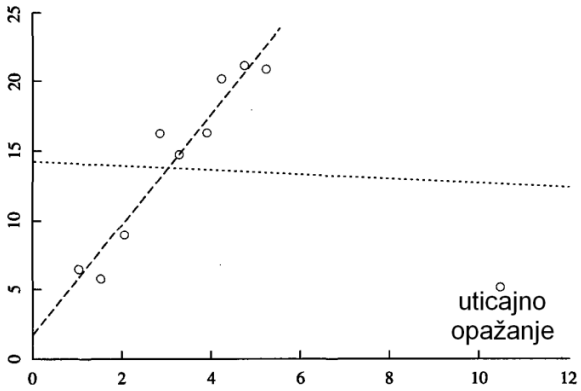
- Studentovi reziduali T_i : elemenat i se izbaci iz modela i napravi model pomoću ostalih $n-1$ elemenata. Na osnovu takvog modela predvidi se tj oceni vrednost zavisnog faktora y_i, $i=1,\dots,n$ za opažanje i, $\mu_Y(i)$, $i=1,\dots,n$. Neka je $\hat{\sigma}_i$, $i=1,\dots,n$ odstupanje modela bez i-tog elementa.

  Studentov residual za elemenat i je $T_i = \frac{Y-\hat{Y}_i}{\hat{\sigma}_i\sqrt{1-h_{i,i}}} \sim t_{n-k-2}$

  Ukoliko je $|T_i|>3$ tada je i-ti elemenat mogući outlier.

4.2.5 Značajni elementi

Pojedini elementi imaju veliki uticaj na model. Ukoliko outlier bitno utiče na model tj modeli sa i bez njega se **signifikantno ---- znacajno** razlikuju tada se za takav elemenat kaže da je značajan (slika 9).



Na slici 9 je data jednostruka linearna regresija pošto se višestruka regresija ne može predstaviti u 2 dimenzije. U slučaju jednostruke regresije se grafički vidi koliko značajan elemenat može da deluje na regresionu jednačinu. Bez značajnog elementa linija ima striktno pozitivan nagib dok sa elementom linija ima blago negativan nagib.
slika 9

Stoga je bitno pronaći značajne elemente i obratiti posebnu pažnju na njih. Postoji nekoliko načina za njihovo otkrivanje:

1. Matrica $H=X(X^TX)^{-1}X^T$. Dijagonalu H matrice čine elementi h_{ii}, $i=1,..,n$ čija je prosečna vrednost $h_{ii}=\frac{k+1}{n}$. Ukoliko je vrednost elementa $h_{ii}>2\frac{k+1}{n}$ tada se *i*-ti elemenat smatra značajnim.

2. *Cook's distance* (Kukova razdaljina)

$$c_i = \frac{1}{k+1}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_i^2,$$ ako je $c_i > F_{0.5:k+1,n-k-1}$ tada je *i*-ti elemenat značajan.

$k+1$ – broj koeficijenata β_i, $i=0,...,k$
h_{ii}, $i=1,...,n$ - vrednost elemenata na dijagonali H matrice
r_i, $i=1,...,n$ – standardizovan rezidual za *i*-to opažanje

3. DFFITS (*difference in the fitted value-snadardized*) metod- $DFFITS_i = \frac{\mu_Y(i)-\mu_Y(-i)}{\hat{\sigma}_i\sqrt{h_{i,i}}}$,

$\mu_Y(i)$ -predviđena vrednost *i*-tog elementa kada se model pravi preko celog uzorka,
$\mu_Y(-i)$ - predviđena vrednost *i*-tog elementa kada se model pravi preko uzorka bez *i*-tog Elementa,
$\hat{\sigma}_i$ - odstupanje modela bez *i*-to elementa,
h_{ii} - vrednost dijagonalnog elementa matrice H kada je ceo uzorak upotrebljen.

Alternativna formula za računanja $DFFITS_i = T_i\sqrt{\frac{h_{ii}}{1-h_{ii}}}$

Ukoliko je $DFFITS_i > 2\sqrt{\frac{k+1}{n}}$ tada je *i*-ti elemenat značajan.

4.2.6 Multikolinearnost

Ponekad iako je uzorak dobro izabran ipak dolazi do poteškoća. Može se desiti da dva ili više nezavisnih faktora zavise jedan od drugog. U tom slučaju je došlo do multikolinearnosti. U stvarnosti su retki slučajevi nepostojanja korelacija između nezavisnih promenljivih. U slučaju postojanja visoke korelacije odstupanja koeficijenata β_i, $i=0,\dots,k$ su velika pa su same ocene netačne. Ako je došlo do savršene multikolinearnosti matrica X^TX postaje singularna pa je nemoguće uopšte odrediti ocene koeficijenata β_i, $i=0,\dots,k$. Na postojanje visokog stepena multikolinearnosti upućuju:

- Determinanta matrice X^TX je približno 0.
- *F-test* se signifikantno razlkuje od 0 dok svi *t-testovi* se ne razlikuju od 0.
- $VIF_i = \frac{1}{1-\hat{\rho}^2_{X_i}}$, ako je *VIF (variance inflation faktor)* >10, $\hat{\rho}^2_{X_i} = \frac{SST-SSE(X_1,\dots,X_{i-1},X_{i+1},\dots X_k)}{SST}$

??????????????????

t-test: daje dokaz za prihvatanje tj odbacivanje pretpostavke da je pojedinačni koeficijent β_i, $i=0,\dots,k$ jednak nekom broju, u većini slučajeva 0.

Nulta hipoteza	$H_0: \beta_i=0,\ i=0,\dots,k$
Uopštija alternativna	$H_1: \beta_i\neq 0,\ i=0,\dots,k$
Test veličina	$\frac{\widehat{\beta_i}}{s_{\widehat{\beta}_i}}$

Područje prihvatanja nulte hipoteze za dvostruku test na nivou poverenja α

$$-t_{n-k-1;\alpha/2} \leq \frac{\widehat{\beta_i}}{s_{\widehat{\beta}_i}} \leq t_{n-k-1;\alpha/2}\ ,\ i=0,\dots,k.$$

Ukoliko se prihvati nulta hipoteza tada se odgovarajući koeficijent $\beta_i=0$, $i=0,\dots,k$ ne razlikuje signifikantno od 0 pa odgovarajući faktor X_i, $i=1,\dots,k$ nema značajnog uticaja na zavisan faktor Y.

F-test: daje dokaz za prihvatanje tj odbacivanje pretpostavke da su svi koeficijenti $\beta_i=0$, $i=0,\dots,k$ signifikantno različiti od 0.

Nulta hipoteza	$H_0: \beta_1=\beta_2=\dots=\beta_k=0$
Uopštena alternativna	$H_1: \beta_1\neq \beta_2\neq\dots\neq \beta_k\neq 0.$
Test veličina	$\frac{SSR/_k}{SSE/_{n-k-1}} = \frac{n-k-1}{k}\frac{R^2}{1-R^2}.$

Područje prihvatanja nulte hipoteze na nivou poverenja α sa k i $n-k-1$ stepeni slobode je

$$\frac{SSR/_k}{SSE/_{n-k-1}} \leq F_{k,n-k-1}.$$

F-test ne mora da daje iste zaključke o koeficijentima $\beta_i=0,\ i=0,\ldots,k$ kao *t-testovi*. Ukoliko su svi *t-testovi* jednaki nuli a *F-test* je signifikantno različit od nule tada je uticaj svakog pojedinačnog faktora mali ali zajednički uticaj faktora je enorman.

Moguća rešenja za otklanjanje multikolinearnosti su:

- Povećanje uzorka
- Ocenjivanje preko ocenjivača: $\tilde{\beta} = (X^T X + kI)^{-1} X^T y$, *k>0* tako da važi trMSE($\tilde{\beta}$) < trMSE($\hat{\beta}$)

4.2.7 Greške u računanju

Model može da podbaci i zbog neadekvatnog računanja. Zbog zaokruživanja vrednosti faktora dolazi do velikih odstupanja prilikom ocene. Ako je na matricu X uticalo zaokruživanje vrednosti onda je ona *loše-uslovljena*. Ako podaci zaokruživanjem dođu blizu 0 stvoriće se multikolinearnost. Računanje bez zaokruživanja (programi imaju tu opciju) će izbeći ovaj potencijalni problem. U današnje vreme ova poteškoća je lako prevaziđena upotrebom računara ali potrebno je imati na umu u slučaju manualnog računananja.

5. Ocenjivanje modela

Model se smatra adekvatnim ako se pomoću njega preko poznatih nezavisnih faktora X_i , $i=1,\ldots,k$ može predvideti vrednost i uticati na zavisnu promenljivu Y. Neka je data regresiona jednačina $\mu_Y(X_1,X_2,..,X_k)=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+...+\hat{\beta}_kX_k$. Koliko je dobar model dat ovom jednačinom može se proveriti preko nekoliko kriterijuma:

1. Neka je $s=\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n^- k-1}}$. U slučaju da važe osnovne pretpostavke s je ocena za odstupanje σ. Što je *s* manje to je bolji model.

2. *Koeficijent determinacije* R^2 – da bi se izačunao najpre se objašnjava:

   Rastavljanje varijacije promenljive Y iz uzorka

   Promena zavisnog faktora Y zavisi od promena nezavisnih faktora $X_i, i=1,\ldots,k$ ali i od nepredvidljivih faktora. Što su manje uticajni nepredvidljivi faktori to je predviđanje bolje.

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2+\sum_{i=1}^{n}e_i^2$$

SST *SSR* *SSE*

SST-ukupna suma kvadrata
SSR-regresijska suma kvadrata
SSE-suma kvadrata grešaka

Suma *SSR* predstavlja delovanje nezavisnih faktora X_i, $i=1,\ldots,k$ na promeljivu Y a *SSE* delovanje grešaka. Dok se na *SSR* može delovati *SSE* je potpuno van mogućnosti kontrole.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$0 < R^2 < 1$$

Što je veće R^2 to je model bolji. Za R^2=1 se dobija model koji je bez uticaja grešaka tj ne postoje slučajne greške.

3. *Prilagođeni koeficijent determinacije* $\bar{R}^2$ - primećeno je da u višestrukoj regresiji dodavajem bilo kakvog novog faktora, čak i nebitnog, dolazi do povećanja koeficijenta determinacije. Zbog toga je uzet u obzir broj faktora, *k+1* kao i veličina uzorka, *n* za bolju ocenu modela.

$$\bar{R}^2 = 1 - \frac{SSE/_{n-k-1}}{SST/_{n-1}} = 1 - \frac{n-1}{n-k-1}\frac{SSE}{SST}$$

Za razliku od običnog koeficijenta determinacije, prilagođeni koeficijent deteminacije ne mora da bude veći od 0. Važi isto rezonovanje kao i za običan koeficijent determinacije, što je prilagođeni koeficijent veći to je bolji model.

4. *Mallow-ov kriterijum*, C_{k+1} ,uzima u obzir veličinu uzorka *n*, broj faktora ali i ocenu varijanse modela $\hat{\sigma}^2$

$$C_{k+1} = \frac{SSE}{\hat{\sigma}^2} + 2(k+1) - n$$

Što je manje C_{k+1} to se smatra model boljim.

Ukoliko su unapred data 2 modela od kojih je potrebno izabrati bolji postoje naredni kriterijumi. Posmatraju se 2 slučaja odvojeno ugnježđeni i neugnježđeni slučaj.

5.1 Ugnježđeni slučaj

Neka je zavisna promenljiva Y određena nezavisnim promenljivama $X_1,X_2,\ldots,X_k$. Skup ovih promenljivih sa odgovarajućom regresionom funkcijom se zove model A. Srednja vrednost tj regresiona funkcija se označava sa μ_Y^A $(X_1,X_2,..,X_k)$ a standardno odstupanje sa σ_A. Ako je zavisna promenljiva određena nezavisnim promenljivima $X_1,X_2,\ldots,X_m$, $m<k$, tada se regresiona funkcija označava sa μ_Y^B $(X_1,X_2,..,X_m)$ a standardno odstupanje sa σ_B. Skup ovih nezavisnih i zavisne promenljive sa odgovarajućim parametrima zove se model B.

Regresione funkcije za modele A i B su date u sledećem obliku:

$\mu_Y^A\ (X_1,X_2,..,X_k)= \beta_0^A +\beta_1^A X_1+\beta_2^A X_2+...+\beta_k^A X_k$ – za model A

$\mu_Y^B\ (X_1,X_2,..,X_k)= \beta_0^B +\beta_1^B X_1+\beta_2^B X_2+...+\beta_m^B X_m$- za model B

Višestruko-parcijalni koeficijent determinacije služi za određivanje stepena adekvatnosti regresije $\mu_Y^A (X_1,X_2,..,X_k)$ za Y u odnosu na $\mu_Y^B\ (X_1,X_2,..,X_m)$ sa faktorima $X_{m+1},...X_k$ kada je $X_1,...,X_m$ fiksno.

Označava se sa $\rho^2_{Y(X_{m+1},...,X_k)|X_1,...X_m}$ i definisan je na sledeći način:

$$\rho^2_{Y(X_{m+1},...,X_k)|X_1,...X_m} = \frac{\sigma^2_{Y|X_1,...,X_m} - \sigma^2_{Y|X_1,...,X_k}}{\sigma^2_{Y|X_1,...,X_m}} = \frac{\sigma_B^2 - \sigma_A^2}{\sigma_B^2} = 1 - \frac{1}{(\sigma_B/\sigma_A)^2}$$

Tačkasta ocena za višestruko-parcijalni koeficijent determinacije je

$$\hat{\rho}^2_{Y(X_{m+1},...,X_k)|X_1,...X_m} = \frac{SSE(X_1,...X_m) - SSE(X_1,...X_k)}{SSE(X_1,...X_m)} = \frac{SSE(B) - SSE(A)}{SSE(B)}$$

Ovaj koeficijent je uvek između 0 i 1. Ukoliko je jednak 0 tada su β_{m+1}^A=...= β_k^A=0 a ukoliko je jednak 1 tada je model A savršeno predviđanje. Sve ostale vrednosti koeficijenta pokazuju koliko je model A bolji od modela B.

5.2 Neugnježdeni slučaj

U praksi se česće javlja slučaj kada nijedan skup nezavisnih nije podskup onog drugog. Ovakva situacija se zove *neugnježden slučaj višestrukih regresija*. Ako su data dva modela, za model A regresiona funkcija je $\mu_Y^A (X_1,X_2,..,X_r)$ i zavisi od *r* nezavisnih (*r*<*k*), $X_1^A,...,X_r^A$, za model B regresiona funkcija je $\mu_Y^B\ (X_1,X_2,..,X_m)$ koja zavisi od *m* nezavisnih (*m*<*k*), $X_1^B,...,X_m^B$. Skupovi nezavisnih mogu da imaju zajedničke faktore ali nijedan nije podskup drugog. $X_1,...,X_k$ je unija oba skupa tj sadrži sve nezavisne koje su ili u modelu A ili u modelu B. Cilj je da se odredi koji je model bolji.

Regresiona funkcija za model A je data u obliku: $\mu_Y^A\ (X_1,X_2,..,X_r)= \beta_0^A+\beta_1^A X_1+\beta_2^A X_2+...+\beta_r^A X_r$

Regresiona funkcija za model B je data u obliku: $\mu_Y^B\ (X_1,X_2,..,X_k)= \beta_0^B+\beta_1^B X_1+\beta_2^B X_2+...+\beta_m^B X_m$

Neka je σ_A standardno odstupanje za model A, σ_B standardno odstupanje za model B. Ukoliko je $\sigma_A<\sigma_B$ regresiona funkcija $\mu_Y^A\ (X_1,X_2,..,X_k)$ je bolje predviđanje. Sledi da je potrebno izračunati obe devijacije i uporediti ih. Moguće ih je oceniti tačkasto i preko intervala poverenja. Takođe se može izračunati interval poverenja za odnos σ_B/σ_A koji dobro dođe za praktično odlučivanje.

Dvostrani interval poverenja za σ_B/σ_A sa koeficijentom poverenja većim ili jednakim sa $1-\alpha$ koji koristi Bonferroni-ev metod je dat u nastavku:

- Neka r i m predstavljaju broj nezavisnih promenljivih u modelima A i B respektivno a neka je n veličina uzorka.

- Predvideti Y pomoću nezavisnih u modelu A i izračunati sumu kvadrata grešaka- *SSE(A).*
- Predvideti Y pomoću nezavisnih u modelu B i izračunati sumu kvadrata grešaka- *SSE(B).*
- Izračunati 1-α/2 dvostrani interval poverenja za σ_A koji je dat sa

$$C[L_A \leq \sigma_A \leq U_A] = 1-\alpha/2$$

gde je $L_A = \sqrt{\frac{SSE(A)}{\chi^2_{1-\frac{\alpha}{4}:n-r-1}}}$ *i* $U_A = \sqrt{\frac{SSE(A)}{\chi^2_{\frac{\alpha}{4}:n-r-1}}}$, $\chi^2_{p:n}$ je kvantil hi-kvadrat raspodele

- Izračunati 1-α/2 dvostrani interval poverenja za σ_B koji je dat sa

$$C[L_B \leq \sigma_A \leq U_B] = 1-\alpha/2$$

gde je $L_B = \sqrt{\frac{SSE(B)}{\chi^2_{1-\frac{\alpha}{4}:n-r-1}}}$ *i* $U_B = \sqrt{\frac{SSE(B)}{\chi^2_{\frac{\alpha}{4}:n-r-1}}}$, $\chi^2_{p:n}$ je kvantil hi-kvadrat raspodele

- Interval poverenja za σ_B/σ_A je:

$$C[L_B/U_A \leq \sigma_B/\sigma_A \leq U_B/L_A] \geq 1-\alpha$$

- Interval poverenja za σ_A/σ_B je:

$$C[L_A/U_B \leq \sigma_A/\sigma_B \leq U_A/L_B] \geq 1-\alpha$$

- Ukoliko je dovoljan jednostrani interval dat je sa:

$$C[L_B/U_A \leq \sigma_B/\sigma_A] \geq 1-\alpha/2$$

Ili

$$C[\sigma_B/\sigma_A \leq U_B/L_A] \geq 1-\alpha/2.$$

5.3 Odabir nezavisnih faktora

Najrealnija je situacija da se skupljaju podaci za raznovrsne faktore. Nakon toga se odlučuje koliko je faktora dovoljno ubaciti u model da bi se smatrao dovoljno dobrim. U predhodnom delu rada je objašnjeno kako izabrati bolju regresiju od dve ponuđenje. Dalje će se ukratko objasniti kako naći najbolju višestruku linearnu regresiju na osnovu raspoloživih faktora. Koristi se statistika :

$$Akaikin\ informacioni\ kriterijum\ (Akaike\ Information\ Criterion\ (AIC)) = nlog(SSE/n)+2(k+1)$$

Manja vrednost AIC upućuje na bolji model.

Postoje nekoliko načina traženja najboljeg modela:

- Svi podskupovi
- Selekcija unapred
- Selekcija unazad

Svi podskupovi: Kod ove metode traži se AIC za sve regresione funcije od svih podskupova faktora. Najbolji model je onaj koji ima najmanji AIC, a u većini slučajeva to je baš model sa najviše faktora. Ukoliko ima *k* faktora potrebno je izračunati 2^k AIC statistika što je vrlo nepraktično ako je broj faktora velik.

Selekcija unapred: Kod ove metode se prvo računa AIC za svaku jednofaktorsku regresionu funkciju. Potom se računa AIC za sve dvofaktorske regresione funkcije kod kojih je jedan faktor najbolji dobijen u prošlom računu, a drugi su redom ostali. Ako je AIC veći nego za jednofaktorsku, prekida se i najbolji model je jednofaktorski. Ukoliko nije taj slučaj nastavlja se dalje analogno. Traže se sve trofaktorske regresione funkcije kod koje su dva faktora najbolja iz predhodne analize a treći su redom svi ostali. Oko $k^2/2$ modela trebaju da se ispitaju dok se ne dobije najbolji. Za veliko *k* ovo je znatno manji broj analiza nego za predhodni način.

Selekcija unazad: Prvo se računa AIC statistika za model sa svim faktorima. Potom se računa AIC za sve modele sa jednim faktorom manje. Ako nijedan AIC od modela sa faktorom manje nije manji od AIC za model sa svim faktorima onda je model sa svim faktorima najbolji. Ukoliko to nije slučaj, računa se AIC za sve modele sa dva faktora manje i upoređuje sa AIC za modele sa jednim faktorom manje. Dalje rezonovanje je analogno.

II PRIMENA VIŠESTRUKE LINEARNE REGRESIJE

Pravljenje modela

Model predstavlja pojednostavljene međusobne odnose pojava iz realnog sveta. Pravljenja modela započinje definisanjem cilja. U ovom radu cilj je u većini primera predvideti očekivanu vrednost određene pojave tj zavisnog faktora preko ostalih pojava tj nezavisnih faktora. Svaki faktor, nezavisan ili zavisan pripada određenoj populaciji sa određenim karakteristikama tj raspodelama. Zavisnost između zavisnog i nezavisnih faktora je data u linearnom obliku. Tačan oblik ove povezanosti se može dobiti ispitivanjem cele populacije. Zbog praktičnih razloga ispitivanje se svodi na manji deo populacije-uzorak. Uzorak se dobija ličnim prikupljanjem podataka ili preuzimanjem već sakupljenih. Na osnovu uzorka dobija se oblik povezanosti tj regresiona jednačina uzorka. Model višestruke linearne regresije se sastoji iz skupa faktora, regresione jednačine i svih osobina jednačine, faktora i uzorka datih u prvom delu.

Komponente modela:

1. *POPULACIJA*-grupa tj skup ljudi, objekata ili pojava za koju je istražitelj zainteresovan, na kojoj proučava željenu pojavu tj osobinu. Može biti *realna*-ako je data grupa u sadašnjosti i može se direktno na njoj ispitivati; *konceptualna*-ako nije direktno moguće vršiti ispitivanja na njoj (ukoliko će postojati tek u budućnosti ili postoji samo kao misaona stvar). Skup može biti *konačan* ili *beskonačan*. *Ciljana populacija*-populacija na kojoj želi da se istražuje, obično je to moguće ako je populacija realna. *Studijska populacija*-ukoliko je populacija imaginarna ili je

previše skupo proučavati je, uzima se populacija slična ciljanoj za koju se smatra da će dati približno iste rezultate.

2. *MODEL*-osobina populacije, kvalitativna ili kvatitativna, koju istražitelj uzima u obzir.U većini slučajeva posmatrana osobina je *gausovska*.

3. *PARAMETRI*-karakteristike populacije koju istražitelj želi da oceni. Obično su to *srednja vrednost* i *standardno odstupanje*.

4. *UZORAK*-deo populacije na kojoj se vrši istraživanje. Odabir se vrši na više načina: prost slučajni uzorak, sistematski uzorak, stratifikovan i klaster uzorak...U ovom radu se posebno koriste *prost slučajni uzorak*-svaki elemenat ima istu verovatnoću da bude izabran; i *slučajni uzorak sa unapred određenim vrednostima od X*-istražitelj određuje vrednosti nezavisnih promenljivih $X_1,...,X_n$ i svaki skup ovih vrednosti predstavlja podpopulaciju. Iz svake podpopulacije se uzima prost slučajni uzorak.

5. *OBLIK VEZE*- jednačina preko koje zavisan faktor može da se predstavi preko nezavisnih. U zavisnosti od broja faktora, veza može biti *jednostruka* (jedan nezavisan faktor) ili *višestruka* (više zavisnih faktora). U zavisnosti u kom su obliku nezavisni faktori može biti *linearna* i *nelinearna* zavisnost.

Naredni primeri su uzeti radi demonstracije adekvatnosti modela dobijenog sa realnim podacima. Mogućnost uticanja na zavisne faktore predstavlja imperativ analize zbog važnosti primera. Svaki primer je detaljno analiziran metodama spomenutim u prethodnom delu i naglašen je njegov značaj u svakodnevnom, realnom životu.

Primer 1: Visina subjekta

Sir Frances Galton definisao je pojam regresije ispitivanjem veza između visina roditelja i deteta. Na osnovu toga prvi primer koji će biti razmatran je visina čoveka i faktori koji utiču na njega. Dobijeni su podaci za muške subjekte (prilog-tabela 1) na osnovu kojih se želi napraviti model. Iz svakodnevnog života se zna da će visina deteta zavisiti od visine njegovih roditelja. Sir F. Galton je došao do zaključka da će visina deteta obično biti između visina roditelja.Zanimljiva je činjenica da su dati i podaci o visinama baba i deda u datom uzorku. U stvarnosti ovaj podatak je vrlo teško dobiti pa bi bilo idealno ukoliko ti faktori nisu bitni za model.

	Regression Summary for Dependent Variable: Visina subjekta (Visine 2 R= .95763262 R²= .91706023 Adjusted R²= .86867870 F(7,12)=18.955 p<.00001 Std.Error of estimate: 2.5493					
N=20	Beta	Std.Err. of Beta	B	Std.Err. of B	t(12)	p-level
Intercept			-198.802	68.48454	-2.90288	0.013255
duzina pri rodjenju	0.264483	0.100384	1.372	0.52067	2.63471	0.021786
visina majke	0.416249	0.105995	0.782	0.19924	3.92706	0.002009
visina oca	0.831612	0.107421	1.051	0.13581	7.74161	0.000005
visina majcine majke	########	0.092611	-0.120	0.17173	-0.69826	0.498323
visina majcinog oca	0.064536	0.091837	0.091	0.13012	0.70272	0.495634
visina oceve majke	0.054906	0.100266	0.088	0.16133	0.54760	0.594003
visina ocevog oca	########	0.103254	-0.102	0.15490	-0.65684	0.523683

tabela 1

Iz tabele 1 se vidi da je regresiona jednačina:

$$Y=-198.8+1.372X_1+0.782\ X_2+1.05\ X_3-0.12\ X_4+0.091\ X_5+0.088\ X_6-0.102\ X_7$$

Pošto se iz analize vidi da su samo prva 3 faktora bitna (dužina pri rođenju, visina majke i visina oca) traži se nova regresiona jednačina uzorka. Prvi model se odbacuje bez ikakve daljne analize zbog traženja novog koji će biti praktičniji za upotrebu tj zahtevaće manje faktora a biće podjednako dobar. Još jedna bitna činjenica koja omogućava izbacivanje faktora je odsustvo bitne multikolinearnosti (najveći stepen korelacije je 0.48) (tabela 2). Iako ovo nije dovoljno za zaključivanje o stepenu multikolinearnosti činjenica da postoje značajni faktori omogućava zaključak da stepen nije visok.

tabela 2

Correlations of Regression Coefficients B; DV: Visina subjekta (Spreadsheet2)

	duzina pri r.	visina majke	visina oca	visina m. maj.	visina m. oca	visina o. maj.	visina o. o.
duzina pri rođ	1,000000	-0,459234	0,155015	-0,301068	-0,013757	-0,162414	0,200478
visina majke	-0,459234	1,000000	-0,105755	0,114586	0,254896	0,306270	-0,336052
visina oca	0,155015	-0,105755	1,000000	0,237747	0,285906	-0,429238	0,479512
visina m. majke	-0,301068	0,114586	0,237747	1,000000	0,085974	0,050097	0,049793
visina m. oca	-0,013757	0,254896	0,285906	0,085974	1,000000	-0,111867	0,121726
visina o. majke	-0,162414	0,306270	-0,429238	0,050097	-0,111867	1,000000	-0,445282
visina o. oca	0,200478	-0,336052	0,479512	0,049793	0,121726	-0,445282	1,000000

Regression Summary for Dependent Variable: Visina subjekta (Spreadsheet2)
R= .95129323 R²= .90495880 Adjusted R²= .88713858
F(3,16)=50.783 p<.00000 Std.Error of estimate: 2.3634

N=20	Beta	Std.Err. of Beta	B	Std.Err. of B	t(16)	p-level
Intercept			-198,712	33,62785	-5,90914	0,000022
duzina pri rodjenju	0,260335	0,086266	1,350	0,44745	3,01780	0,008170
visina majke	0,368391	0,085213	0,692	0,16017	4,32319	0,000525
visina oca	0,872015	0,078365	1,102	0,09908	11,12755	0,000000

tabela 3

Nova regresiona jednačina dobijena nakon izbacivanja nebitnih faktora je (tabela 3):

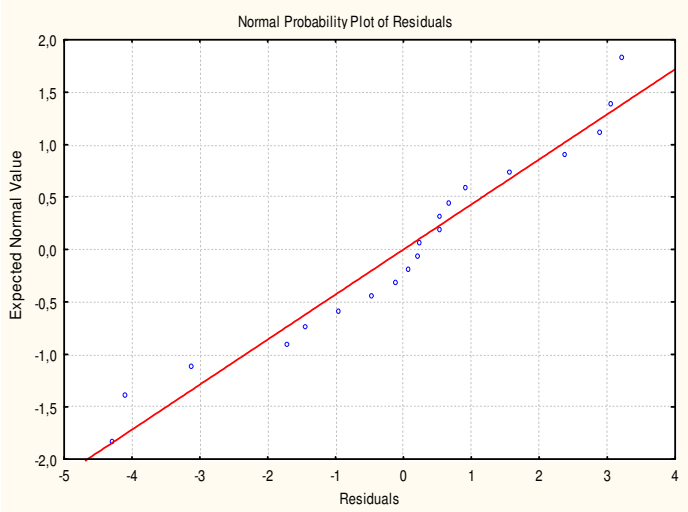
$$Y=-198.7+1.35X_1+0.69\ X_2+1.102\ X_3$$

Pošto je $\bar{R}^2$ prvog modela manji od $\bar{R}^2$ drugog modela (0.87<0.89) i standardna devijacija drugog modela je manja nego za prvi (2.36<2.55) dolazi se do zaključka da je drugi model bolji. Pošto su svi koeficijenti signifikatno različiti od 0 ne sumnja se u postojanje visokog stepena multikolinearnosti.Ne postoje outlieri u uzorku. Potom treba proveriti da li se mogu koristiti intervali poverenja i testovi.

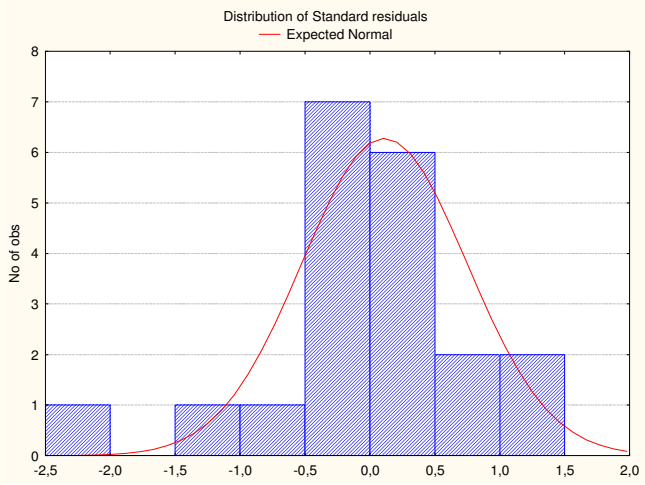
Ukoliko bi se izbacio i faktor dužine pri rođenju dobija se model:

$$Y=-155.45+0.89X_1+1.06X_2,\quad \bar{R}^2=0.83,\quad \sigma=2.87$$

Slede provere osnovnih pretpostavki za model sa 3 nezavisne promenljive.



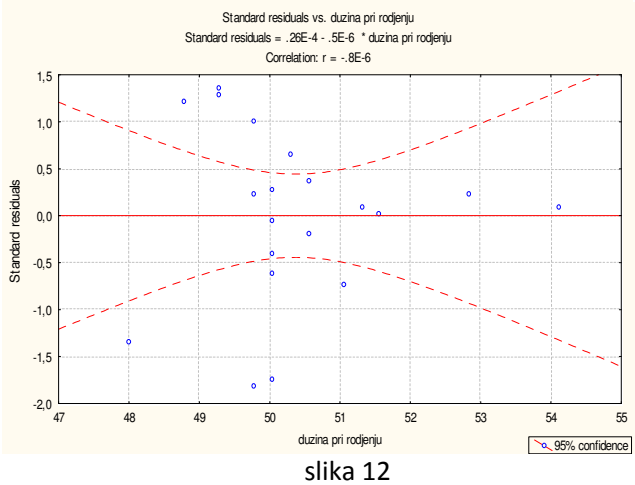
slika 10



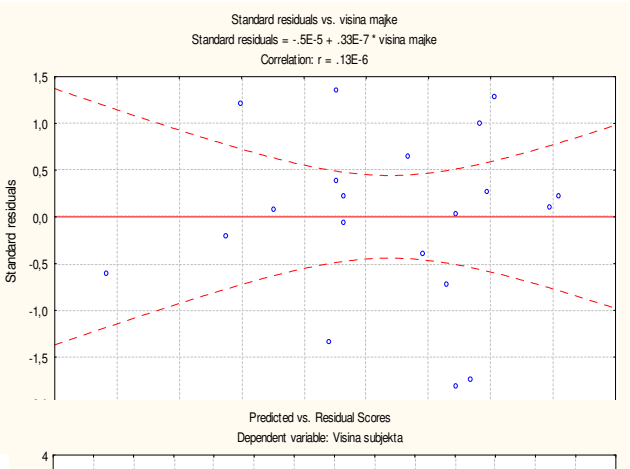
slika 11

Standardizovani reziduali nemaju normalnu raspodelu (slika 10 i slika 11). Testovi i intervali poverenja se mogu koristiti u slučaju da je uzorak velik. Pošto konkretni uzorak nije velik (preko 30) potrebno je dalje ispitivanje.

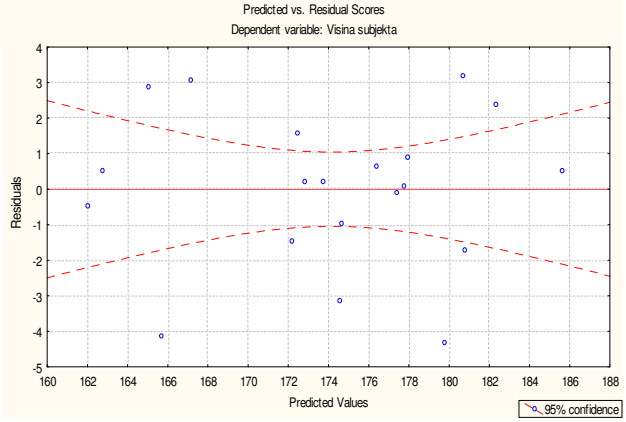
Iako se ne može konstatovati konkretan šablon u analizi standardizovanih reziduala i nezavisnih faktora dužina pri rođenju (slika 12), visina majke (slika 13) i visina oca (slika 14) raspored tačaka navodi na zaključak da je došlo do heteroskedastičnosti. Ovu pretpostavku podržava analiza predviđenih vrednosti i standardizovanih reziduala (slika 15) gde se primećuje da manje i veće vrednosti imaju veću disperziju u odnosu na one srednje.

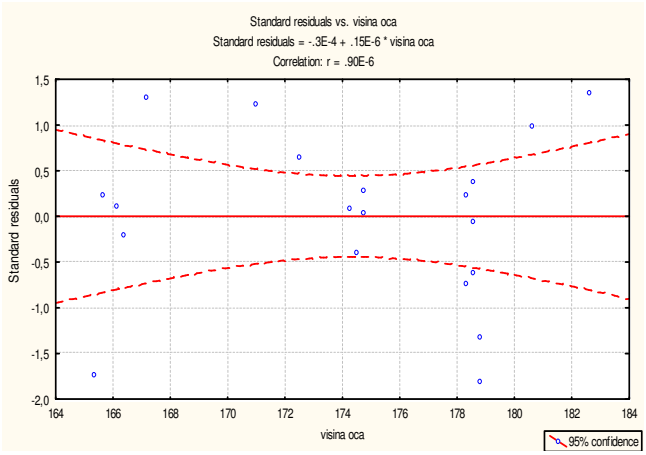


slika 12



slika 13





slika 14

slika 15

Zaključak: Ne mogu se koristiti ni intervali poverenja ni testovi značajnosti. U ovom modelu je došlo do: odstupanja od normalnosti, sredine od 0 i do heteroskedastičnosti. Ukoliko bi se znao oblik odstupanja mogao bi se dobiti efikasan model. Nažalost u ovom slučaju nije moguće saznati ovaj podatak. Stoga ovaj model se trenutno ne može popraviti.

Iz drugog izvora su preuzeti podaci sa interneta takođe za muške ispitanike (prilog-tabela 2). Da li bi ovi podaci mogli da se iskoriste za novi model?

S obzirom da su već nađena tri faktora koja bitno utiču (među njima su i dva iz ovog primera) očekuje se model u kome će biti izostavljen bitan faktor. Koeficijenti takvog modela nemaju sve poželjne osobine i nemogu se koristiti intervali poverenja ni testovi. Analiza je dala sledeće rezultate (tabela 4):

N=20	Regression Summary for Dependent Variable: Visina subjekta (Visine 3 R= .45931747 R²= .21097254 Adjusted R²= .11814578 F(2,17)=2.2728 p<.13344 Std.Error of estimate: 6.6497					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(17)	p-level
Intercept			79.21605	44.73941	1.770610	0.094554
Visina majke	0.383303	0.216480	0.37323	0.21079	1.770615	0.094553
Visina oca	0.218287	0.216480	0.18281	0.18129	1.008346	0.327426

tabela 4

Ne samo da model sam po sebi nema dovoljno faktora nego je uzorak takav da ni faktori koji su značajni se preko ovog uzorka ne razlikuju od nule, što pokazuju i *t-testovi* i *F- test* (2.27 <3.59). $\bar{R}^2$ je svega 0.12 što potvrđuje činjenicu da nezavisni faktor više zavisi od slučajnih faktora nego od nezavisnih faktora. Jedini outlier je prvo opažanje. Njegovim izbacivanjem faktor visine majke postaje značajan i $\bar{R}^2$ se povećava na 0.26. Standardno odstupanje modela se smanjilo sa 6.65 na 5. 56 (tabela 5). Ipak ni ovaj model nije dovoljno adekvatan.

Regresiona jednačina modela : $Y = 67.96 + 0.42X_1 + 0.2X_2$

N=19	Regression Summary for Dependent Variable: Visina subjekta (Visine 3.sta R= ,58218652 R²= ,33894114 Adjusted R²= ,25630878 F(2,16)=4,1018 p<,03647 Std.Error of estimate: 5,5579					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(16)	p-level
Intercept			67,95970	37,59632	1,807616	0,089503
Visina majke	0,487869	0,204139	0,42307	0,17702	2,389884	0,029507
Visina oca	0,275750	0,204139	0,20495	0,15172	1,350795	0,195557

tabela 5

Uzorak iz prvog modela je prikupljen u ruralnoj sredini dok je drugi prikupljen u metropoli. Srednja vrednost visine subjekta prvog uzorka je 174cm dok je od drugog uzorka 172.7cm tj 173.4cm (bez outliera). Može se zaključiti da sredina utiče na visinu.

Model dobijen preko uzorka sa ženskim podacima (prilog-tabela 3) je znatno bolji nego sa uzorkom sa muškim podacima.

$$Y = 18.93 + 0.7X_1 + 0.16X_2,\ \bar{R}^2 = 0.64,\ \sigma = 4.93$$

Da li postoji još neki faktor koji bitno utiče na visinu? Do sada je nađeno 3 faktora koji bitno utiču na visinu. Ne može se tačno saznati šta sve bitno utiče dok se ne ispita za što više faktora.

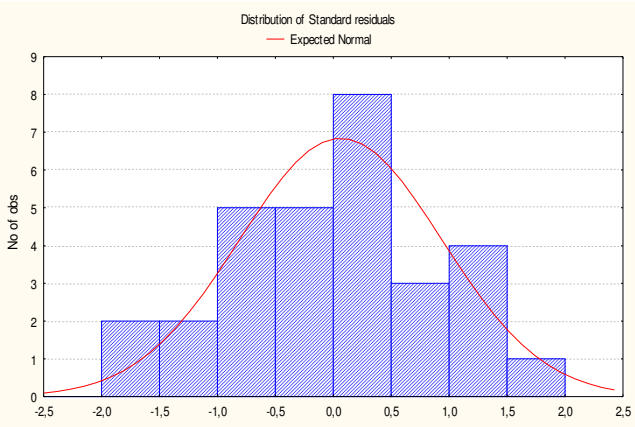
Prikupljeni su podaci i o fizičkim aktivnostima i načinu ishrane (prilog-tabela 4). Svima je poznata majčina izreka "Jedi puno voća i povrća da porasteš visok i zdrav". Naravno tu su i mnogobrojne aktivnosti na koje su verovatno svi išli kao mala deca da bi "bili zdravi i razvijeni". Za očekivati je da ti faktori imaju bitnu ulogu u visini subjekta. U tom slučaju bi dosadašnji najbolji model postao model kome nedostaju bitni faktori. Na osnovu datog uzorka napravljen je model (tabela 6).

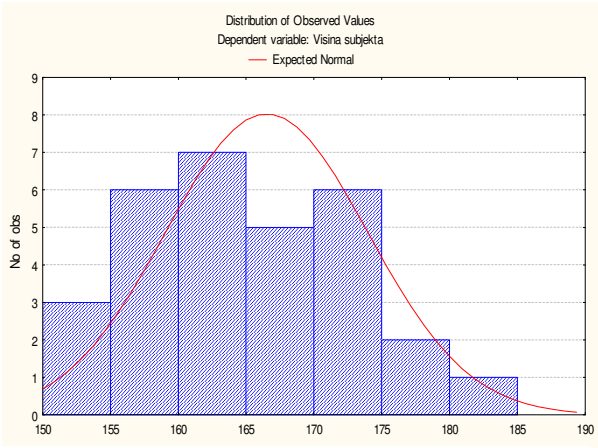
	Regression Summary for Dependent Variable: Visina subjekta (Visine) R= .63962214 R²= .40911648 Adjusted R²= .31457512 F(4,25)=4.3274 p<.00851 Std.Error of estimate: 6.3513					
N=30	Beta	Std.Err. of Beta	B	Std.Err. of B	t(25)	p-level
Intercept			54,30384	36,56638	1,485076	0,150025
Visina majke	0,132040	0,178564	0,18673	0,25253	0,739456	0,466517
Visina oca	0,495880	0,184319	0,41406	0,15391	2,690338	0,012535
nedeljno bavljenje aktivnostima	0,136830	0,172284	0,57885	0,72883	0,794214	0,434545
Nedeljno konzumiranje voca i povrca	0,191007	0,173158	1,01060	0,91617	1,103078	0,280499

tabela 6

Ponovo se javlja sumnja da uzorak nije reprezentativan kao da ni osnovne pretpostavke ne važe. Opet se bitan faktor pojavio kao nebitan a nova dva faktora takođe nemaju uticaj. $\bar{R}^2$ je 0.31 a standardno odstupanje modela 6.35 (veće dva puta nego kod prva dva modela).

Zavisna promenljiva nema normalnu raspodelu (slika 16) kao ni standardizovani reziduali (slika 17).

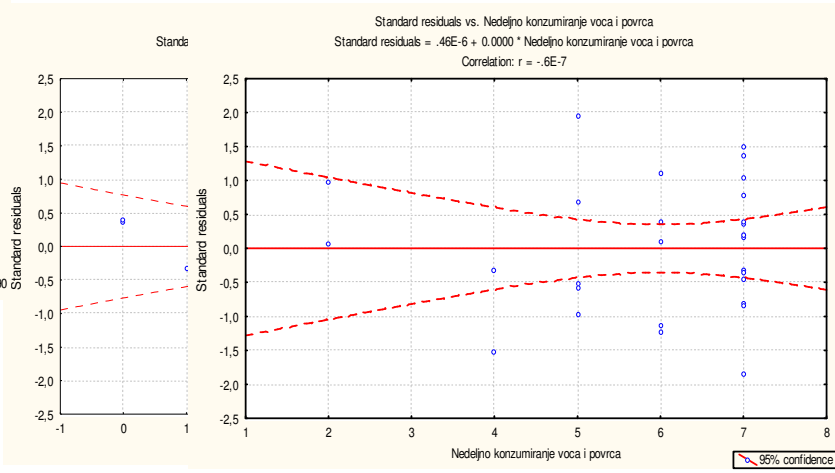




slika 16

slika 17

Odstupanja za faktore bavljenja sportom (slika 18) i konzumiranja voća i povrća (slika 19) su heterogena.



slika 18

slika 19

Iako određivanje visine ne predstavlja faktor od životne važnosti mnogi ljudi su znatiželjni, pogotovo budući roditelji. Svako dete je bar jednom poželelo da postane košarkaš tj manekenka kad poraste. Rađena su mnogobrojna istraživanja međutim savršeno predviđanje ne postoji. Za sada je pokazano da visina zavisi 80% od genetike (otkriveno je da 180 genetskih varijacija utiče na visinu) a 20% od sredine (načina ishrane, pogotovo neuhranjenosti, bolesti i fizičkih aktivnosti). []

Najjednostavniji modeli zahtevaju visinu roditelja i pol subjekta dok oni složeniji koriste i visine subjekta tokom različitih perioda života. Najčešća naučna predviđanja su data u tabeli 7 [].

Za muškarce	Za žene
2*visina muškog subjekta u dobu od 2 godine	2*visina ženskog subjekta u dobu od 18 meseci
(Visina muškog subjekta u dobu od 3 godine +55.88cm)*1.27	(Visina ženskog subjekta u dobu od 3 godine+43.18cm)*1.29
(visina majke+visina oca)/2+5.08cm	(visina majke+visina oca)/2-6.35cm
(visina majke+visina oca+5cm)/2	(visina majke+visina oca-5cm)/2

tabela 7

Tokom svog života čovek raste promenljivim tempom: []

- Od rođenja (tipična dužina je oko 50 cm) pa do druge godine rast je ubrzan, u ovom periodu dete poraste za 35.5 cm.
- Od druge godine do puberteta rast je ujednačen, oko 6.35 cm za svaku godinu.
- Ulaskom u pubertet rast se ubrzava, od 7.62 do 12.7cm godišnje.
- Izlaskom iz puberteta prestaje se sa rastom. Redovno bavljenje sportom može da doda 1-2 cm visine dok razne bolesti mogu da smanje.
- Oko 40-50 godine dolazi do smanjenja visine kao deo životnog procesa

Devojčice u proseku ulaze u pubertet 2 godine pre dečaka. Nagli rast obično počinje ekstremitetima pa odatle ideja da veličina stopala može da odredi visinu čoveka. Nažalost ovaj faktor nije dobar u stvarnosti. Istina je da visina zavisi od najduže kosti u čovekovom telu, to je femur koji se nalazi u gornjem delu noge. Sve dok kost raste, otprilike do 17-e godine, raste čovek. Uz adekvatnu

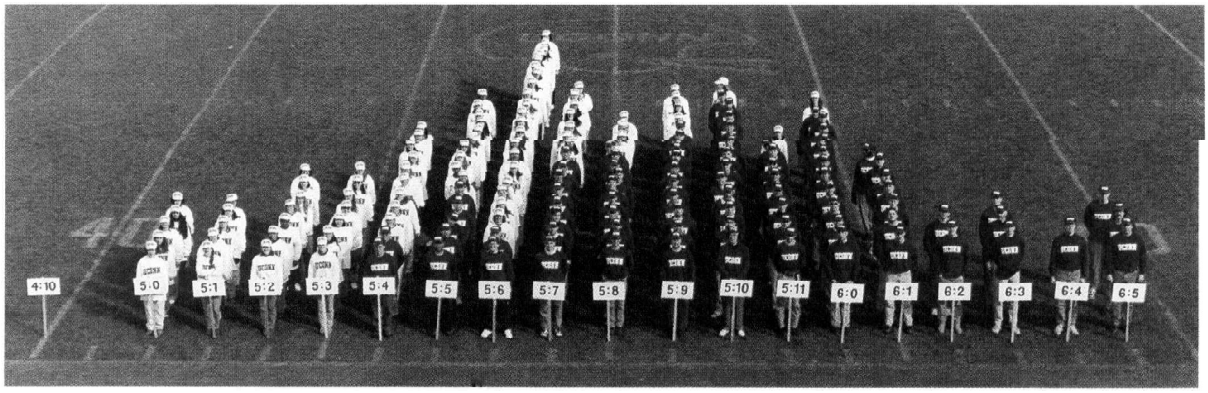
zdravu ishranu i način života čovek će dostići svoju maksimalnu visinu, koja se nalazi između visina majke i oca. Ukoliko u detinjstvu dođe do neadekvatnih uslova visina deteta kad odraste će biti znatno manja od moguće.Na visinu utiču i geografski faktori tj pripadnici različitih nacija imaju različitu visinu (tabela 8)

zemlja	Prosečna visina muškarca	Prosečna visina žena
Australia	174.8	161.42
Brazil	169	158
Canada	174	161
Colombia	170.64	158.65
Finland	178.2	164.7
Gambia	168	157.8
Japan	171.51	158.04
United Kingdom	176.8	163.79
United States	176.2	162.5

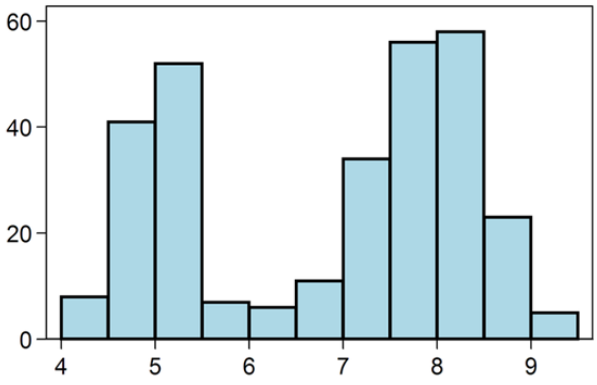
Tabela 8

Poznato je da visina čoveka ima normalnu raspodelu, to važi za svaki pol. Međutim kad se posmatraju visine muškaraca i žena zajedno tada se dobija bimodalna raspodela (slika 10). Bimodalna raspodela se sastoji iz dve unimodalne raspodele, u ovom slučaju normalne.

Postoje mnogobrojni primeri u prirodi ove raspodele: vreme između erupcija pojedinih gejzira, boje galaksije, veličina radnika zelenih mrava (slika 20).



Slika 20



Slika 21

Na slici 20 se nalaze 143 studenta univerzitet u Konektikiju. Slika datira iz 1996 godine. Cilj je bio da se živim dijagramom pokaže bimodalna raspodela visina.

Na slici 21 je data dužina radnika zelenih mrava i njihova učestalost. Ova vrsta mrava može se naći u Africi, južnim delovima Indije i Azije i u Australiji.

U tabeli 9 je izračunata predviđena visina na više načina za autora ovog rada

Formula za žene	Predviđena visina	Stvarna visina
$18.93 + 0.7X_1 + 0.16X_2$	160.4	164
$54.3 + 0.19X_1 + 0.41X_2 + 0.56X_3 + X_4$	168.4	
$2X_5$	160	
$55.7 + 1.29X_6$	170.5	
$-6.35 + 0.5X_1 + 0.5X_2$	164.2	
$-2.5 + 0.5X_1 + 0.5X_2$	168	

Tabela 9

X_1 – visina majke $X_1 = 161$
X_2 – visina oca $X_2 = 180$
X_3 – nedeljno bavljenje fizičkim aktivnostima $X_3 = 5$
X_4 – nedeljno konzumiranje voća i povrća $X_4 = 7$
X_5 – visina subjekta u dobu od 18 meseci $X_5 = 80$
X_6 – visina subjekta u dobu od 3 godine $X_6 = 89$

Primer 2 : Telesna masa

Gojaznost je postala jedna od najznačajnijih bolesti modernog doba. U Evropi kao i u Americi je, po rezultatima anketiranja, van kontrole. Procenjuje se da je skoro pola ljudi sa previše kilograma dok je skoro 20% preterano gojazno. Određivanje stepena uhranjenosti se vrši preko *indexa telesne mase BMI* (body mase index) koji predstavlja indikator ukupne telesne masti. Za većinu ljudi je vrlo efikasan pokazatelj dok za sportiste bi mogao da preceni nivo masti dok za starije ljude ili ljude bez mišića da potceni. On se računa na sledeći način: []

$$BMI = \frac{masa\ u\ kilogramima}{(visina\ u\ metrima)^2}$$

Skala je data u tabeli 10:

Stepen uhranjenosti	BMI
Neuhranjenost	*< 18.5*
Normalna težina	*[18.5 ,25)*
Prekomerna težina	*[25 ,30)*
Gojaznost	*≥30*

tabela 10

Uz ovaj index može se koristiti i obim struka kao efikasan pokazatelj. Ukoliko je veći od 102 cm kod muškaraca i 88 cm kod žena smatra se da je nakupljena mast oko struka opasna po zdravlje.

Primedba: Uz male modifikacije jednačina *BMI*-a se može napisati u linearnom obliku pa se dobija:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

gde su:

$Y = ln(BMI)$
$\beta_0 = 0$
$\beta_1 = 1$
$X_1 = ln(masa\ u\ kilogramima)$

β_2=-2
X_2=ln(visina u metrima)

Uzroci debljine su mnogobrojni a prvenstveno se odnose na moderan način življenja. Neki od uzroka su:

- Neadekvatna ishrana
- Nedovoljno fizičke aktivnosti
- Stres
- Hormonski poremećaj

Zdrava ishrana- prvenstveno podrazumeva pravilan balans različitih grupa namirnica i redovne obroke. Svakodnevna ishrana bi trebala da sadrži: []
Ugljene hidrate-40%
Belančevine-30%
Masti-30%

- Ugljeni hidrati- Tokom varenja se skrobovi i šećeri (glavne vrste ugljenih hidrata) razbijaju u glukozu. Ovaj produkt služi kao energija za mozak i centralni nervni sistem. Ukoliko je uneto previše ugljenih hidrata deo njih neće biti pretvoren u glukozu nego će se skladištiti kao salo. Ipak oni su jako bitni za pravilno funkcionisanje organizma pa postoji minimalna dnevna potreba za njima što iznosi 50 g. Ukoliko se ne ispoštuje, proteini će krenuti da se raspadaju u pokušaju stvaranja energije i dolazi do stanja ketoze. Ova bolest se često javlja kod dijabetičara koji nisu ni svesni da to jesu ili koji se ne čuvaju dovoljno. Telo kreće samo sebe da rastvara u pokušaju stvaranja energije. Ukoliko se ne otkrije na vreme može biti smrtonosno. Voće i žitarice predstavljaju najbolje namirnice koje sadrže ugljene hidrate.

- Proteini-Proteini su neophodni su za stvaranje mišića u telu. Dnevne potrebe zavise od godina i kilaže. U tabeli 11 je dat merni pokazatelj. Kad se odredi ovaj pokazatelj na osnovu godina pomnoži se sa brojem kila.

Godine starosti	1-3	4-6	7-10	11-14	15-18	>19
Merni pokazatelj	0.82	0.68	0.55	0.45	0.40	0.36

tabela 11

I u ovom slučaju postoji minimalna količina koja iznosi oko 45g. Osobe koje bi želele da povećaju svoju masu tj da im se mišići ne razgrade potrebno je da dnevno unesu gramažu proteina ekvivalentnu telesnom broju kilograma. Preporučuje se da gramaža proteina ne prelazi 15g po jednom obroku, radi boljeg varenja i iskorišćavanja. Za doručak je preporučljivo uzeti proteine. Mogu se naći u namirnicama kako životinjskog tako i biljnog porekla. Najzastupljeniji su pileće i ćureće meso, riba, jaja, mlečni proizvodi...

- Masti- Dokazano je da ljudska dnevna ishrana treba sadržati 30% masti. Postoje nezasićene i zasićene masti. Nezasićene masti se nalaze u namirnicama biljnog porekla i izuzetno su zdrave. Njih treba svakodnevno unositi u organizam., pogotovo omega-3-kiseline koje se nalaze u ribama. Deluju na zdravlje čitavog organizma a za razliku od pojedinih masti koje organizam može sam da proizvede, one se moraju direktno uneti. Zasićene masti se nalaze u namirnicama životinjskog porekla i njih treba u umerenim dozama konzumirati.Holesterol je složeni lipid (vrsta masti) iz grupe sterola, koji se nalazi samo u namirnicama životinjskog porekla. Ljudski organizam može takodje proizvesti holesterol u jetri. Njegova funkcija u organizmu je da služi kao osnovna sirovina za sintezu

seksualnih hormona, žucnih soli i membrana ćelija. Holesterol je neophodna supstanca u ljudskom organizmu, ali kada njegova stopa u krvi poraste, on se taloži na zidovima arterije smanjujuci time njihov obim: to je ateroskleroza. Otuda povišena stopa holesterola može dovesti do infarkta miokarda,tromboze nedovoljne cirkulacije krvi u ekstremitetima. Holesterol se kreće kroz krv spojen sa supstancama koje se zovu lipoproteini. Upravo ti lipoproteini dele holesterol na dve vrste:

· LDL holesterol. Ovaj holesterol protiče krvotokom spojen sa lipoproteinima male gustine. Taj holesterol priblizno iznosi 75% od ukupnog holesterola u krvi. LDL holesterol pospešuje stvaranje ateroskleroze. On se takodje naziva "lošim holesterolom".

· HDL holesterol. On protiče krvotokom vezan za lipoproteine velike gustine. Ovaj holesterol nazvan je "dobrim" jer sprečava aterosklerozu. Njegova stopa u krvi trebalo bi da bude veća.

U tabeli 12 data je kaloričnost svake grupe po gramu .

1g	Broj kalorija u 1 g
Proteina	4
Ugljenih hidrata	4
Masti	9

Tabela 12

Fizičke aktivnosti- Analize Britanskih istraživača pokazuju da redovna umerena do intenzivna fizička aktivnost smanjuje rizik od srčanih bolesti i ishemijskog i hemoragijskog moždanog udara (šloga). Povećanje fizičke aktivnosti takođe može da smanji rizik od nekih vrsta raka, osteoporoze, dijabetesa, depresije, gojaznosti i hipertenzije. Pacijenti koji su već oboleli od raka, takođe, mogu imati koristi od fizičke aktivnosti, jer je ona povezana sa manjim rizikom od smrti i boljim oporavkom posle bolesti.

Da bi štitilo od bolesti vežbanje treba da ubrza puls na 50 do 70% od maksimalnog. Maksimalni se izračunava kada se od 220 oduzme broj godina. Maksimalan puls za osobu od 60 godina je 160. Izračunato je ovako: 220 - 60 = 160. Polovina od toga tj. 50% je 80, dok 70% iznosi 112. Znači da puls za osobu od šezdeset godina treba da bude brži od 80 otkucaja u minuti, ali sporiji od 112. Neutrenirane osobe treba da počnu sa manje intenzivnim vežbama, koje sasvim malo ubrzavaju puls. []

Stres- Stres je odgovor organizma na situaciju koju osoba doživljava kao ugrožavajuću, bilo fizičku ili psihičku. Organizam se u stresu priprema na brzu reakciju i zaštitu - npr. pojačava se rad srca i pluća, povišava se krvni pritisak, količina šećera u krvi, mišićna napetost itd. Psihičke reakcije su strah, teskoba, promene pozornosti, rasuđivanja.Kako je u stresu ravnoteža organizma poremećena, telo ulaže dodatne napore da bi ponovo uspostavilo ravnotežu. Ako je organizam, odnosno osoba, često izložena stresu ili je stanje stresa dugotrajno, stalno ulaganje napora da se vrati u ravnotežu iscrpljuje organizam i čini ga podložnijim nastanku raznih zdravstvenih poremećaja i bolesti. []

Posledice prekomerne težine su smanjenje fizičke pokretljivosti, mehaničko opterećenje kostiju, a posebno zglobova, na kojima se javljaju promene. Stoga je česta posledica debljine artroza koja uzrokuje bol i ukočenost zglobova i mišića. Iz istih razloga dolazi i do bolova i oticanja u nogama, proširenih vena, problema s kičmom, a javljaju se i reumatična oboljenja. Debljina može biti rizični faktor za pojavu osteoartritisa, koji obeležava propadanje hrskavice u zglobovima, a najčešće u zglobovima kičme, kuka i kolena, koji su posebno opterećeni suvišnim kilogramima.

Posledica debljine mogu biti i poremećaji i obolenja metabolizma, pa čak i nastanak dijabetesa tipa II i povišenih vrednosti masnoća u krvi. Posebno su ugroženi i srce i krvotok, pa debljina može pospešiti ili izazvati povišeni i visoki krvni pritisak, oslabljenje funkcija srca, srčani infarkt i moždani udar. Debljina predstavlja rizik i kada je riječ o disanju i organima za disanje, jer se bitno smanjuje vitalni

kapacitet pluća, pa ugrožene osobe imaju osjećaj da nemaju dovoljno vazduha, posebno pri fizičkom naporu.

Debljina predstavlja rizični faktor i kada je reč o kancerogenim oboljenjima materice, dojke, žučnog mehura, creva i prostate. Uz to, ona predstavlja i vrlo visoki faktor rizika prilikom podvrgavanja različitim operativnim zahvatima. Naravno, kao posledica debljine mogu se javiti i psihički problemi i bolesti, kao što su osećaj manje vrijednosti i osamljenosti, ali i depresija i socijalna izolacija, do čega dolazi posebno stoga što višak kilograma u savremenom društvu predstavlja i veliki estetski problem.

Kod gojaznih osoba se opaža čitav niz hormonskih promena.Suprotno uvaženim pretpostavkama da su hormonske promene uzrok debljine, one su najčešće njena uobičajena komplikacija. []

Zbog svega gore navedenog jasno je da je od životne važnosti otkriti kako regulisati težinu. Postaviće se model čiji će cilj biti definisanje veze između težine i godina, visine, ishrane, aktivnosti i stresa subjekta. Očekuje se da će težina značajno zavisiti od svih ovih faktora.Iako se na neke faktore ne može uticati, kao što su godine i visina, očekuje se da će ostali faktori biti dovoljno značajni da utiču na faktor težine.

Korišćenjem podataka (prilog-tabela 5) dobijeni su podaci iz tabele 13

N=32	Regression Summary for Dependent Variable: tezina (Spreadsheet2 R= .70030744 R²= .49043052 Adjusted R²= .39243638 F(5,26)=5.0047 p<.00242 Std.Error of estimate: 6.4649					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(26)	p-level
Intercept			-75,4850	29,23441	-2,58206	0,015807
godine	0,134843	0,150781	0,1018	0,11383	0,89430	0,379367
visina	0,716674	0,165869	0,8001	0,18518	4,32074	0,000202
ishrana	0,048291	0,174013	0,2810	1,01267	0,27751	0,783580
fizicka aktivnost	-0,090150	0,152660	-0,4216	0,71395	-0,59053	0,559933
stres	-0,044632	0,157396	-0,4609	1,62527	-0,28356	0,778991

tabela 13

Regresijska jednačina uzorka je $Y=-75.48+0.1X_1+0.8\,X_2+0.28\,X_3-0.42\,X_4-0.46\,X_5$

Nažalost model ne pokazuje očekivane rezultate. Godine, visina i ishrana su u srazmernom odnosu sa težinom. Za ishranu se to ne bi moglo pretpostaviti s obzirom da unošenje više zdrave i niskokalorične hrane bi trebalo obrnuto da utiče. S druge strane fizička aktivnost i stres obrnuto utiču što je i logično za fizičku aktivnost ali ne i za stres. Kao posledica stresa češće se javlja kompulsivno prejedanje nego nejedenje. Nažalost jedino visina značajno utiče u ovom modelu na težinu.

Jedino uticajno opažanje je broj 29. Model bez spomenutog opažanja je (tabela 14):

N=31	Regression Summary for Dependent Variable: tezina (Spreadsheet21.st R= ,77964574 R²= ,60784748 Adjusted R²= ,52941697 F(5,25)=7,7501 p<,00016 Std.Error of estimate: 5,4169					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(25)	p-level
Intercept			-74,1534	24,49840	-3,02687	0,005660
godine	0,106287	0,135208	0,0752	0,09568	0,78610	0,439197
visina	0,938043	0,156318	0,9839	0,16396	6,00085	0,000003
ishrana	-0,408887	0,186644	-2,6020	1,18773	-2,19074	0,038009
fizicka aktivnost	-0,286975	0,147138	-1,2579	0,64497	-1,95038	0,062428
stres	-0,320108	0,159538	-3,1418	1,56582	-2,00647	0,055738

tabela 14

Regresijska jednačina uzorka je $Y=-74.15+0.08X_1+0.98\,X_2-2.6X_3-1.26\,X_4-3.14\,X_5$

S obzirom da nisu svi koeficijenti značajno jednaki nuli i da nijedna korelacija između faktora nije prevelika (ne prelazi 0.58, tabela 15) može se tvrditi da nema visokog stepena multikolinearnosti.

Variable	Correlations of Regression Coefficients B; DV: tezina (Spreadsheet21.st				
	godine	visina	ishrana	fizicka aktivnost	stres
godine	1,000000	0,219380	0,112758	0,054219	0,154690
visina	0,219380	1,000000	-0,496201	-0,236674	-0,413735
ishrana	0,112758	-0,496201	1,000000	0,515967	0,578124
fizicka aktivnost	0,054219	-0,236674	0,515967	1,000000	0,220107
stres	0,154690	-0,413735	0,578124	0,220107	1,000000

tabela 15

Variable	Redundancy of Independent Variables; DV: tezina (Spreadsheet21.st R-square column contains R-square of respective variable with all other independent variables			
	Toleran.	R-square	Partial Cor.	Semipart Cor.
godine	0,858050	0,141950	0,155312	0,098454
visina	0,641941	0,358059	0,768266	0,751571
ishrana	0,450285	0,549715	-0,401316	-0,274377
fizicka aktivnost	0,724542	0,275458	-0,363406	-0,244273
stres	0,616292	0,383708	-0,372425	-0,251298

tabela 16

Nijedan R^2 nije blizu 1 (tabela 16) pa se da zaključiti da zaista ne postoji visok stepen multikolinearnosti u modelu. Ipak ukoliko bi se iskoristio ovaj model rezultati bi bili zbunjujući.

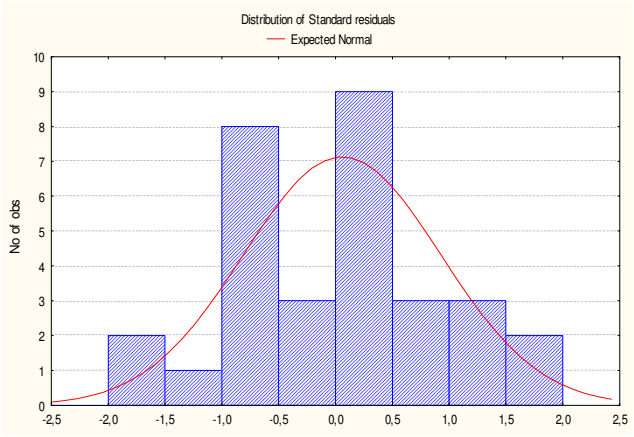
Zanimljivo je analizirati sam uzorak. Srednja težina uzorka je 59.4 (tabela 17). Iako je nivo stresa vrlo visok, 3.8, izgleda da ne utiče na povećanje težine. Može biti da zbog srednje starosti uzorka, 28.3, još uvek se ne osete posledice stresa na težinu. Iako 50% populacije ima prekomernu težinu (BMI je preko 25), u datom uzorku od 31 ženske osobe, 29 ima normalnu težinu (u granicama od 18.5 do 24.9), jedna ženska osoba ima BMI 16,18 dok jedna ženska osoba ima BMI 25. Po ovom uzorku se ne može zaključiti da ženska populacija ima problema sa prekomernom težinom.

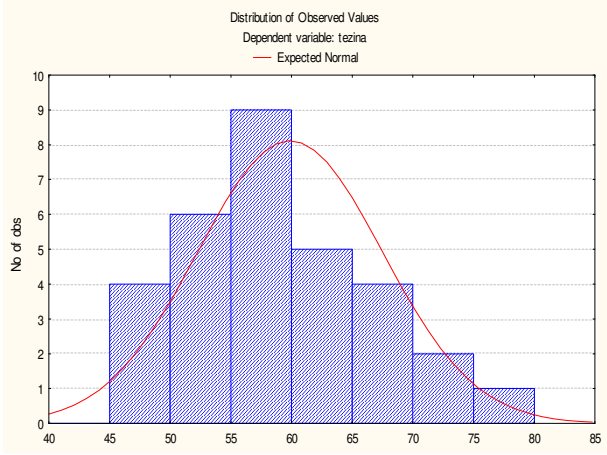
Variable	Means and Standard Deviations (Spreadsheet21.st		
	Means	Std.Dev.	N
godine	28,3871	11,15849	31
visina	166,2903	7,52858	31
ishrana	6,1613	1,24088	31
fizicka aktivnost	3,3871	1,80143	31
stres	3,7742	0,80456	31
tezina	59,4419	7,89649	31

Tabela 17

Za osobu od 23 godine, visoke 164 cm koja svaki dan konzumira voće i povrće, 5 puta nedeljno se bavi fizičkim aktivnostima i retko se nervira predviđena težina je 58kg. 95% interval poverenja je (52.4,64)kg , 99% interval poverenja (50.4,66)kg. U stvarnosti spomenuta osoba ima 49kg. Ova vrednost čak ne upada ni u 99% interval poverenja.
Da li je dati model neupotrebljiv pokazaće dalja analiza.

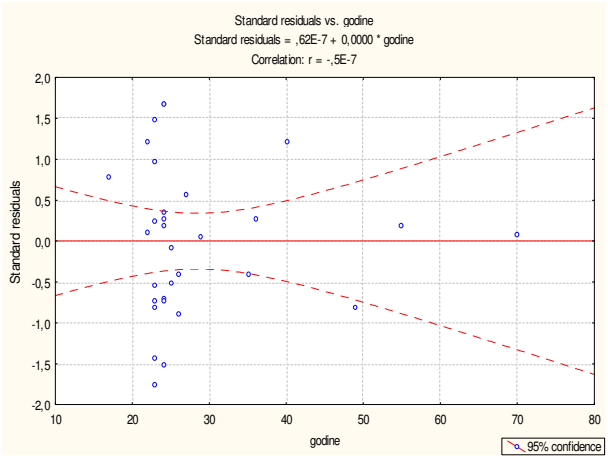
Daljom analizom proverava se narušenost pretpostavki I mogućnost korišćenja interval poverenja I testova značajnosti.



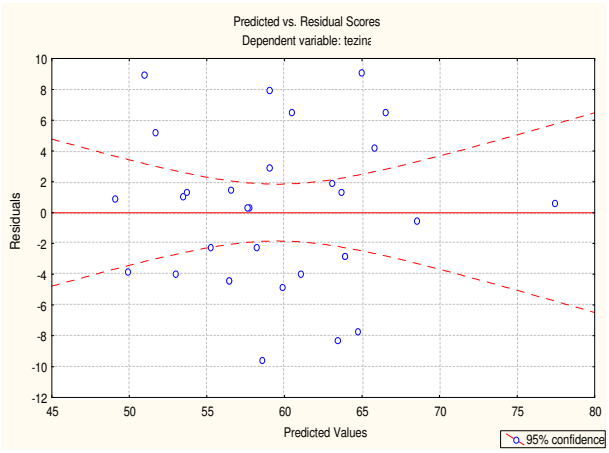


Standardizovani reziduali nemaju normalnu raspodelu (slika 22) dok zavisan faktor ima približno normalnu raspodelu (slika 23).

slika 22



slika 23



slika 24 slika 25

Kod većine faktora nije došlo do heteroskedastičnosti, jedino se ističe kod faktora godina (slika 24). Uopšteno u modelu nepostoji signifikantna sumnja da je došlo do heteroskedastičnosti ili pogrešnog oblika jednačine regresije (slika 25). Iako je došlo do rezultata različitih od stvarnih, analizom modela se može doći do pretpostavke da je možda subjekt outlier. Ova pretpostavka će se proveriti u daljem radu.

Drugi uzorak (prilog-tabela 6) sadrži samo podatke o godinama, visini i težini. Model pokazuje da su i godine i visina bitni faktori za težinu (tabela 18) ali nažalost se ne može uopšte koristiti jer se ne mogu menjati ni godine ni težina. Ovakav model bi mogao da posluži za određivanje poželjne težine na osnovu godina i visine.S obzirom da varijacije zavisnog faktora ne zavise previše od nezavisnih ($\bar{R}^2$ =0.38) ovaj model ne bi bio dobar kontrolni model. Stoga nije potrebna dalja analiza.

	Regression Summary for Dependent Variable: tezina (Spreadsheet3 R= .63806876 R²= .40713175 Adjusted R²= .37508482 F(2,37)=12.704 p<.00006 Std.Error of estimate: 13.490					
N=40	Beta	Std.Err. of Beta	B	Std.Err. of B	t(37)	p-level
Intercept			-76.6385	49.84453	-1.53755	0.132667
godine	0.527430	0.127469	0.7227	0.17466	4.13770	0.000194
visina	0.302365	0.127469	0.7411	0.31241	2.37206	0.023002

tabela 18

Kako onda uticati na težinu?

Jedna od metoda koja se najčešće upotrebljava je brojanje kalorija. Prostom matematikom se dolazi do zaključka da povećana potrošnja kalorija u odnosu na unos dovodi do smanjenja kilograma. Za izračunavanje potrebnog broja kalorija koristi se *BMR* (*Basal Metabolic Rate*) koji predstavlja broj kalorija potreban za održavanje vitalnih funkcija. Izračunava se na osnovu pola, godina, visine i težine (tabela 19). Precizno izračunavanje se vrši uz pomoć kliničke analize potrošnje vazduha. Ovakvi testovi nisu neophodni jer je razlika između njih i izračunavanja preko tabele 22 najviše 100 kalorija. []

Za žene	$Y = 655 + (9.6 \times X_1) + (1.8 \times X_2) - (4.7 \times X_3)$
Za muškarce	$Y = 666 + (13.7 \times X_1) + (5 \times X_2) - (6.8 \times X_3)$

tabela 19

gde su
Y *– dnevni unos kalorija-BMR*
X_1*- težina u kilogramima*
X_2*-visina u centimetrima*
X_3*-godine*
Ove dve jednačine predstavljaju višestruku linearnu regresiju sa 3 nezavisna faktora.

Pošto je ova količina kalorija neophodna samo za održavanje života tj u stanju mirovanja potrebno je prilagoditi svakodnevnim životnim aktivnostima (rad, vežbanje, igra sa decom, hobi). Korigovanjem BMR formule dobija se *Hariss Benedict-ova formula* koja uzima u obzir bavljenje fizičkim aktivnostima. Računanje ove formule je data u tabeli 20: []

Intenzitet fizičkih aktivnosti	Dnevna količina kalorija
Nizak-bez bavljenja fizičkih aktivnostima	BMR x 1.2

Umeren-1-3 puta nedeljno umerene vežbe	BMR x 1.375
Osrednji-3-5 puta nedeljno fizičke aktivnosti	BMR x 1.55
Aktivno- 6-7 puta nedeljno fizičke aktivnosti	BMR x 1.725
Intenzivan- 2 puta dnevno fizičke aktivnosti	BMR x 1.9

tabela 20

Ukoliko je *BMI* van normalnog potrebno je skinuti višak kilograma. Na osnovu *BMI* se izračuna kolika bi trebala biti željena težina. U jednom kilogramu viška nalazi se 7000 kcal. Preporučen dnevni gubitak kalorija je od 500 do 950, što bi značilo da se nedeljno gubi najviše 1 kg. Nakon organizovanja fizičkih aktivnosti izaračuna se BMR. Kada se od BMR-a oduzme željeni dnevni gubitak kalorija dobija se potrebna dnevna količina kalorija.

Primer: Osoba ima 80 kg i visoka je 170 cm. Njen BMI je 27.7 što ukazuje na prekomernu težinu. Potrebno je napraviti novi režim ishrane i aktivnosti. Ukoliko odluči da krene 3 puta nedeljno na aktivnosti potreban unos dnevnih kalorija je 2216 kalorija (1612 kalorija za funkcionisanje organizma i 604 za fizičke aktivnosti). Za lagani gubitak kilograma, pola kilograma nedeljno, potrebno je smanjiti unos kalorija za 500 svaki dan. Da bi osoba došla do gornje granice normalne težine potrebno je da dođe barem na 72 kg. Navedenim režimom ovaj gubitak kilograma se očekuje za 16 nedelja sa dnevnim konzumiranjem od 1716 kalorija.

Iako se ova metoda čini vrlo razumljivom i efektnom, rezultati nisu uvek zagarantovani. Kontrola samog unosa kalorija je vrlo problematična. U jednom istraživanju posmatrana je osoba prekomerne težine koja je trebala da broji dnevni unos kalorija. Osoba je nesvesno zaboravljala da zapisuje kalorije sitnih grickalica čija je ukupna vrednost na kraju dana bila oko 1000 kalorija. Iako je na većini ambalaža utisnuta energetska vrednost ona može da varira, pogotovo zbog tretiranja hrane. Takođe postoje i mnogobrojni mitovi o kalorijama koji su neistiniti. Neki od najuobičajenijih su: postoji hrana s negativnim kalorijama, ne moraju se brojati kalorije ako se broje ugljeni hidrati, niskomasni proizvodi imaju manje kalorija, telo sagoreva više kalorija tokom dijete s malo ugljenih hidrata, redovno vežbanje otklanja potrebu brojanja kalorija…

U okviru NCHS-a (National Center for Health Statistics) sprovedeno je Zdravstveno i Nutritivno istraživanje u cilju ispitivanja gojaznosti u SAD tokom 2005-2008 godine. Rezultati istraživanja su objavljeni u decembru 2010 godine: []

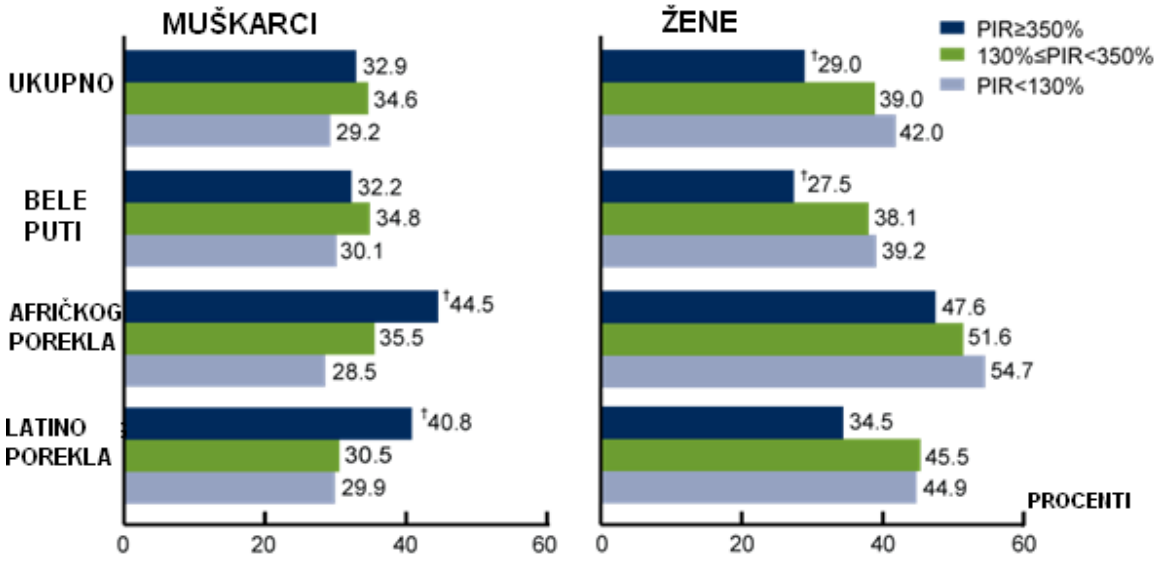
Muškarci:

- Uopšteno gojaznost nije povezana sa visinom zarade (slika 26).
- Kod afričko-američke populacije i američke populacije meksičkog porekla veća zarada je povezana sa većom kilažom (slika 26).
- Ne postoji značajna veza između stepena obrazovanja i težine (slika 27).

Žene:

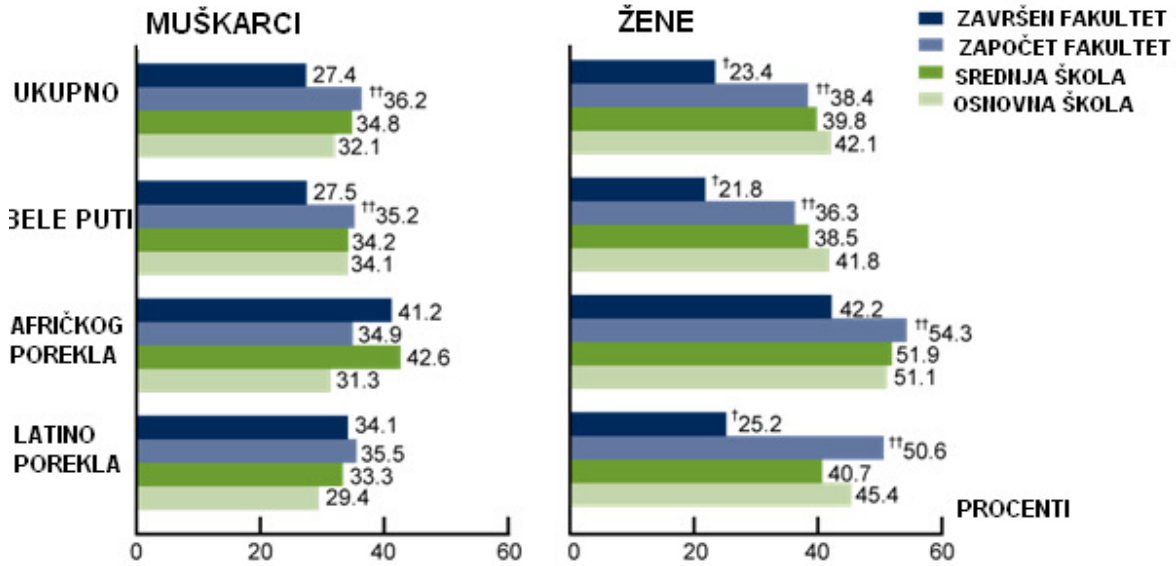
- Visina zarade utiče obrnuto srazmerno na težinu (slika 26).
- Visina zarade je značajan faktor za populaciju amerikanaca bele puti dok se i kod ostalih grupa primeti ovaj uticaj (slika 26).
- Obrazovanje je uticajan faktor, povećanjem obrazovanja smanjuje se težina (slika 27).

Došlo je do drastičnog porasta gojaznosti među svim grupama obrazovanja i veličine zarade od poslednjeg istraživanja 1988-1994 godine.



Slika 26

PIR (Poverty income ratio)- pokazatelj porodičnog prihoda kada se uzme u obzir broj članova i inflacija. Normalna primanja spadaju u PIR od 130%-350%.
†- signifikantan trend



slika 27

†-signifikantan trend
††-signifikantno različito u odnosu na završen fakultet

U tabeli 22 data su predviđanja težine i stvarne težine dva kontrolna ženska subjekta. Na prvom subjektu se već radilo predviđanje koje se nije poklapalo sa stvarnom težinom. *BMI* prvog subjekta je

18.2 što potpada u grupu blage neuhranjenosti. Stoga se može tvrditi da je dati subjekt outlier. Težina drugog subjekta spada u granice normalne težine. 99% interval poverenja za model dat u ovom primeru predstavlja podskup granica BMI-a. Na osnovu toga dobijeni model se može koristiti za predviđanje zdrave težine.

Formula	I subjekat	II subjekat
$-74.15+0.075X_1+0.98X_2-2.6X_3-1.26X_4-3.14X_5$	58.15 (50.37-65.92)-99% int	73.26 (65,8- 80.72)- 99% int
$-76.64+0.72X_1+0.74X_2$	61.28	73.12
BMI	49.76-67	59.94-80.68
Stvarna težina	49	69

tabela 22

	I subjekt	II subjekt
X_1 - godine	23	23
X_2 - visina	164	180
X_3 - broj nedeljnog konzumiranja voća i povrća	7	7
X_4 - broj nedeljnog bavljenja fizičkim aktivnostima	5	3
X_5 - učestalost nerviranja	2	3

Primer 3 : Visina krvnog pritiska

Jedna od posledica prekomerne težine je povišen krvni pritisak. Na vrednosti krvnog pritiska utiču i drugi faktori: način života (stres, ishrana, bavljenje fizičkih aktivnosti) kao i lične karakteristike (godine i visina). Krvni pritisak predstavlja silu kojom cirkulišuća krv deluje na jedinicu površine krvnog suda a nastaje zbog kontrakcije srca. Prilikom merenja dobijaju se dve vrednosti, sistolni i dijastolni pritisak. Sistolni tj gornji krvni pritisak predstavlja silu kad se krv izbacuje iz leve komore. Tada se stvara najveća sila koja polako opada između dve kontrakcije. U tom periodu dolazi do najmanje vrednosti sile koja se zove dijastolni tj donji krvni pritisak. Krv se tada uliva iz predkomore u komoru. Smatra se da je normalan krvi pritisak 120/80. Vrednosti veće od navedene spadaju u povišen krvni pritisak a ukoliko su više od 140/90 dolazi do *hipertenzije*. Vrednosti niže od normalne spadaju u niži krvni pritisak a ukoliko su niže od 100/60 dolazi do *hipotenzije*. Hipotenzija obično nije ozbiljan poremećaj mada su simtomi često vrlo nelagodni: umor, nesvestica,poremećaj sna, anksioznost...*Hipertenzija* se smatra ozbiljnim obolenjem i zahteva lečenje. Potrebna je svakodnevna terapija lekovima kao i promena životnih navika. Smatra se da preko 40 % stanovništva pati od povišenog pritiska. Ne preduzimanje nikakvih mera dovodi do ozbiljnih komplikacija pa čak i prevremene smrti. []

Zbog gore navedenih razloga bitno je otkriti način kontrolisanja krvnog pritiska. Koji sve faktori utiču na njega i koliko je moguće regulisati pritisak preko regulisanja ostalih faktora? U cilju odgovora na predhodno pitanje dobijeni su podaci na osnovu ankete (prilog-tabela 7). Momentalno iznenađuje činjenica da svi ispitanici imaju normalan ili čak šta niži krvni pritisak uprkos visokom stepenu stresa. Već na prvi pogled čini se da model dobijen pomoću ovih podataka neće biti adekvatan. Ipak pokušaće se.

N=32	Regression Summary for Dependent Variable: Sys (krvni pritisak.sta R= ,52049305 R²= ,27091302 Adjusted R²= ,09593214 F(6,25)=1,5482 p<,20379 Std.Error of estimate: 12,424					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(25)	p-level
Intercept			163,3464	62,97523	2,59382	0,015644
godine	-0,029061	0,186737	-0,0346	0,22209	-0,15562	0,877580
visina	-0,131319	0,265206	-0,2310	0,46647	-0,49516	0,624813
tezina	-0,006938	0,239231	-0,0109	0,37690	-0,02900	0,977093
ishrana	0,206609	0,212583	1,8942	1,94902	0,97190	0,340411
aktivnosti	0,110593	0,187466	0,8148	1,38124	0,58994	0,560529
stres	-0,417610	0,192295	-6,7937	3,12826	-2,17171	0,039570

Tabela 23

N=32	Regression Summary for Dependent Variable: Dias (krvni pritisak.st R= ,36365600 R²= ,13224569 Adjusted R²= ----- F(6,25)=,63500 p<,70109 Std.Error of estimate: 9,1193					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(25)	p-level
Intercept			94,12479	46,22329	2,03631	0,052444
godine	-0,077068	0,203722	-0,06167	0,16301	-0,37830	0,708403
visina	-0,025751	0,289330	-0,03047	0,34239	-0,08900	0,929789
tezina	-0,098645	0,260992	-0,10456	0,27664	-0,37796	0,708650
ishrana	0,119090	0,231920	0,73459	1,43056	0,51350	0,612113
aktivnosti	-0,050443	0,204518	-0,25005	1,01382	-0,24664	0,807200
stres	-0,284270	0,209786	-3,11135	2,29612	-1,35505	0,187521

Tabela 24

Tabele 23 i 24 daju rezultate za sistolni i dijastolni pritisak. Kod sistolnog pritiska samo stres se javlja kao bitan faktor, dok kod dijastolnog nijedan faktor se ne ističe kao značajan. Zanimljivo je da svi faktori osim ishrane deluju obrnuto na pritisak dok jedino ishrana ga podiže. Ova činjenica je u potpunoj kontradikciji u odnosu na činjenice koje su date na početku. $\bar{R}^2$ je za obe vrste pritiska izrazito mali (0.095 i 0.0001) što znači da njihove varijacije ne zavise značajno od promena zavisnih faktora. Takođe i *t-testovi* pokazuju da su svi koeficijenti jednaki 0 (osim stresa kod sistolnog pritiska) i *F-testovi* podržavaju taj zaključak (F(6,25)=2.49).Stoga se odmah uzima novi uzorak (prilog-tabela 8)i pravi drugi model .

N=40	Regression Summary for Dependent Variable: Sys (pritisak.sta) R= ,64234230 R²= ,41260363 Adjusted R²= ,32622181 F(5,34)=4,7765 p<,00206 Std.Error of estimate: 14,048					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(34)	p-level
Intercept			111,7513	53,60983	2,084530	0,044695
godine	0,048649	0,168172	0,0669	0,23110	0,289278	0,774125
visina	-0,145923	0,146284	-0,3587	0,35955	-0,997533	0,325552
tezina	0,658648	0,191124	0,6605	0,19167	3,446184	0,001531
puls	0,114783	0,145745	0,1572	0,19956	0,787561	0,436410
holesterol	-0,063301	0,156511	-0,0058	0,01440	-0,404454	0,688412

tabela 25

Opet nije dobijen zadovoljavajući model. $\bar{R}^2$ je svega 0.33 a jedini značajni faktor je težina. F(5,34)=2.5 pa *F-test* pokazuje veliki uticaj faktora. S obzirom da su svi *t-testovi* osim za težinu, pokazali da koeficijenti nisu signifikantno različiti od nule postoji velika verovatnoća da je došlo do visokog stepena multikolinearnosti. Nijedna korelacija između faktora se ne ističe kao prevelika (-0.53 je najveća)(tabela 26). U slučaju višestruke regresije ne može se osloniti samo na ovu proveru jer korelacija može biti i između više faktora, što se ne može videti na ovaj način. Nijedan R^2 nije blizu 1 (najveći 0.53)(tabela 27).

	Correlations of Regression Coefficients B; DV: Sys (pritisak.sta				
Variable	godine	visina	tezina	puls	holesterol
godine	1,000000	0,172669	-0,531369	-0,305546	0,052332
visina	0,172669	1,000000	-0,391083	-0,226850	0,127133
tezina	-0,531369	-0,391083	1,000000	0,205632	-0,438433
puls	-0,305546	-0,226850	0,205632	1,000000	-0,247686
holesterol	0,052332	0,127133	-0,438433	-0,247686	1,000000

tabela 26

	Redundancy of Independent Variables; DV: Sys (pritisak.sta R-square column contains R-square of respective variable with all other independent variables			
Variable	Toleran.	R-square	Partial Cor.	Semipart Cor.
godine	0,610862	0,389138	0,049550	0,038023
visina	0,807346	0,192654	-0,168626	-0,131115
tezina	0,472958	0,527042	0,508797	0,452965
puls	0,813326	0,186674	0,133850	0,103517
holesterol	0,705285	0,294715	-0,069197	-0,053161

tabela 27

Pošto dosadašnje analize nisu dale kokretan zaključak moraju se dalje vršiti provere za stepen multikolinearnosti. Matrica X^TX (matrica 1) nije dijagonalna i ima inverznu čiji su elementi vrlo blizu 0 (matrica 2). Det (X^TX)=1.5181e+0.22 što je blizu 0. Nakon svih provera može se tvrditi da postoji visok stepen multikolinearnosti.

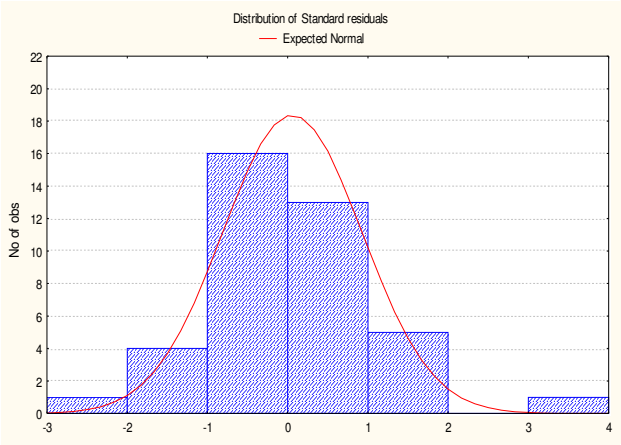
40	1329	6420	2652	3052	9635
1329	50205	213697	92770	103272	347425
6420	213697	103228	427293	490548	1552147
2652	92770	427293	187166	203508	699415
3052	103272	490548	203508	238960	759260
9653	347425	1552147	699415	759260	3669819

matrica 1

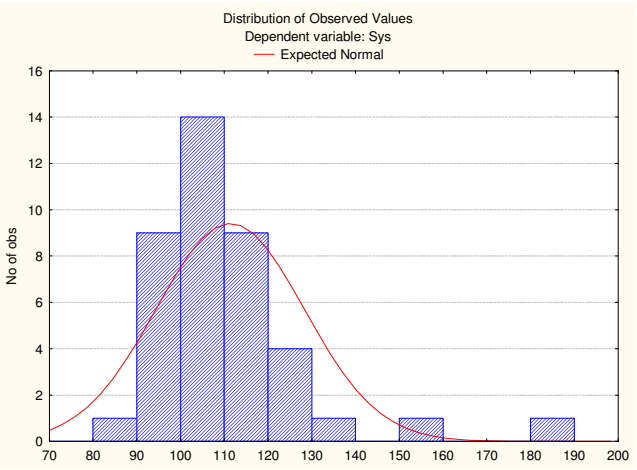
0.0250	0.0008	0.0002	0.0004	0.0003	0.0001
0.0008	0.0000	0.0000	0.0000	0.0000	0.0000
0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
0.0004	0.0000	0.0000	0.0000	0.0000	0.0000
0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
0.0001	0.0000	0.0000	0.0000	0.0000	0.0000

matrica 2

Dalja analiza pokazuje da ni standardni reziduali (slika 28) ni zavisan faktor (slika 29) nemaju normalnu raspodelu .

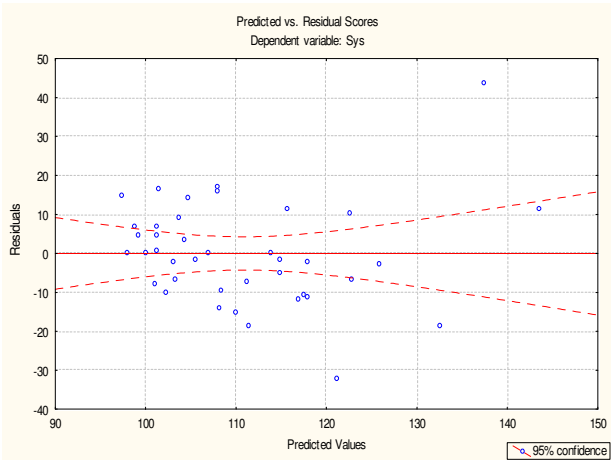


slika 28



slika 29

Postoji verovatnoća da je došlo do heteroskedastičnosti (slika 30). Takođе se sumnja da postoje značajna opažanja. Dalje se obavljaju testovi za otkrivanje značajnih opažanja (tabela 28).



slika 30

Outlieri su opažanja broj 11 i 15. Koliko bitno utiču na model videće se njihovim izbacivanjem. Ukoliko se dobije znatno bolji model onda ih je bolje izuzeti iz uzorka.

Standard Residual: Sys (pritisak.sta) Outliers

	Observed - Value	Predicted - Value	Residual	Standard - Pred. v.	Standard - Residual	Std.Err. - Pred.Val
11 . . . *. . .	89,0000	121,0313	-32,0314	0,930722	-2,28019	5,048296
15 . . . . * . .	181,0000	137,3957	43,6043	2,419353	3,10402	7,334036
Minimum . . . *.	89,0000	121,0313	-32,0314	0,930722	-2,28019	5,048296

Maximum	181,0000	137,3957	43,6043	2,419353	3,10402	7,334036
Mean . . . .* .	135,0000	129,2135	5,7865	1,675038	0,41192	6,191166
Median . . . .*	135,0000	129,2135	5,7865	1,675038	0,41192	6,191166

Tabela 28

Izbacivanjem ova 2 opažanja nije se dobio znatno bolji model (tabela 29). Standardno odstupanje se smanjilo sa 14 na 10.4 međutim $\bar{R}^2$ se poboljšao samo za 0.003.

N=38	Regression Summary for Dependent Variable: Sys (pritisak.sta R= ,64781445 R²= ,41966356 Adjusted R²= ,32898599 F(5,32)=4,6281 p<,00273 Std.Error of estimate: 10,394					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(32)	p-level
Intercept			120,2598	40,68454	2,95591	0,005812
godine	0,198175	0,181373	0,1975	0,18077	1,09264	0,282710
visina	-0,151973	0,143731	-0,2829	0,26757	-1,05734	0,298272
tezina	0,579088	0,201061	0,4673	0,16226	2,88016	0,007036
puls	-0,006378	0,151847	-0,0065	0,15556	-0,04200	0,966759
holesterol	-0,108808	0,160441	-0,0074	0,01088	-0,67818	0,502533

tabela 29

Na dijastolni pritisak jedino značajno utiču godine . $\bar{R}^2$ je nizak (0.36) tako da nezavisni faktori nemaju prevelik uticaj na pritisak. Pošto je korišćen isti uzorak kao i za sistolni pritisak došlo je do istih problema kao i u predhodnom modelu,postoji visok stepen multikorelacije .

Bez obzira na podbacivanje modela otkriveno je da na krvni pritisak utiču: []

- Ishrana
- Fizičke aktivnosti
- Težina

Ishrana- smanjenje pritiska se može postići *izbacivanjem soli* iz ishrane. So dovodi do zadržavanje vode u organizmu koja dovodi do podizanja krvnog pritiska. Preporučena dnevna doza je do 6g soli.Dokazano je da *voće i povrće* smanjuje pritisak. Preporučljivo je jesti 5 puta dnevno voće i povrće, po mogućstvu što manje prerađeno, bez soli i sosova i što pre upotrebljeno . Preradom kao i dugim stajanjem se mogu uništiti svi minerali i vitamin koje bijke poseduju. Celer i beli luk se ističu kao najbolji prirodni lek za smanjenje krvnog pritiska.*Alkohol* povišava krvni pritisak. Sadrži mnogo kalorija pa povećava težinu koja pak povećava krvni pritisak. Preporučljivo je po jedna manja čaša vina ili piva dnevno ali sve preko te doze može biti štetno.*Masti* mogu dvojaka uticati na krvni pritisak. Sve masti životinjskog porekla povisuju krvi pritisak dok masti biljnog porekla snižavaju. Iako su masti biljnog porekla zdrave imaju dosta kalorija pa se treba ograničiti unos zbog dobijanja na težini.

Fizičke aktivnosti-Bavljenjem fizičkim aktivnostima dovodi do smanjenja krvnog pritiska. Tokom rekreacije pritisak raste međutim posle prestanka rekreacije se smanjuje. Poželjne su sve aerobne vežbe: hodanje, trčanje, plivanje, vožnja biciklom, plesanje… Rekreacije kao što su dizanje tegova, padobranstvo i brzo trčanje imaju kontraefekat i svi ljudi sa visokim pritiskom bi trebali da ih izbegavaju. Ukoliko osoba ima krvni pritisak preko 140/90 trebalo bi da se konsultuje sa doktorom pe nego što krene sa vežbanjem.

Održavanje težine- povećana težina je povezana sa višim krvnim pritiskom. Doktori preporučuju da se težina drži u propisanim granicama BMI-a (dat je u primeru gojaznosti).Svim ljudima van granice normalne težine se preporučuje smanjnje kilograma. Najbolje je krenuti sa novim životnim navikama-zdravijom ishranom i fizičkim vežbama, uz nadzor lekara. Povećana težina se vezuje i za pojavu dijabetesa koji u većini slučajeva nestaje gubitkom suvišnih kilograma.

U nekim slučajevima promena životnih navika nije dovoljna za smanjenje pritiska tj neće ga smanjiti dovoljno. U tom slučaju postoji raznovrsan rang lekova za smanjenje pritiska. Bitno je konsultovati se sa lekarom pre primene bilo kojeg leka. Faktori rizika bitni za konzumiranje lekova su:

- Visok holesterol
- Pušenje
- Dijabetes
- Bolesti bubrega
- Porodično visok krvni pritisak
- Druge bolesti srca ili krvnih sudova

Postoje 4 vrste leka za smanjenje pritiska. Dele se po načinu delovanja: prve dve grupe deluju na hormone koji kontrolišu krvni pritisak, treća grupa opušta zidove arterija čime ih proširuje a četvrta grupa izbacuje suvišnu tečnost iz tela. Mogu se kombinovati više lekova iz različitih grupa. Na odluku koju vrstu leka treba koristiti utiču godine, etnička pripadnost, ostale bolesti i lekovi koji su predhodno konzumirani. Kad jednom krene da se uzima lek potrebno je konzumirati ga do kraja života. Još nije pronađen način da se krvni pritisak smanji dugoročno. Stoga je nužno redovno uzimati lekove i pridržavati se uputstava. Nažalost mogu se javiti propratni efekti među kojima su: mučnina, vrtoglavica, suvi kašalj… Koliko će lek delovati zavisi od osobe do osobe.

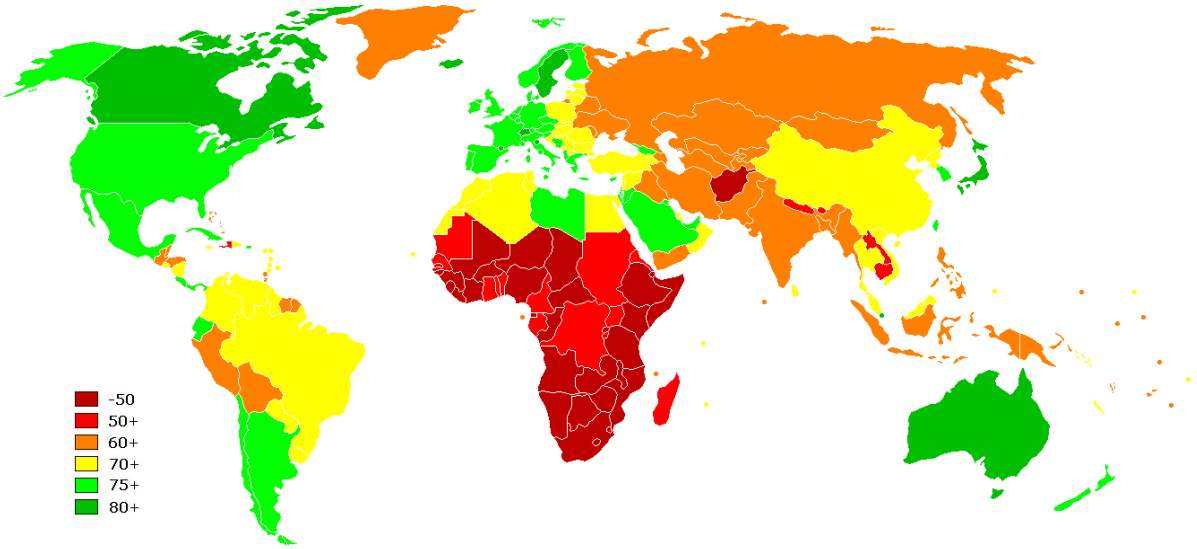
Primer 4: dužina životnog veka

Jedna od najbitnijih životnih stvari je sama dužina života. Otkako je počela ekspanzija lekova i moderne medicine životni vek se znatno produžio. S druge strane svakodnevni moderan život uzima svoj danak kroz savremene bolesti kao što su srčani udari i šlogovi. U današnje doba se istraživanjima došlo do zaključka da ljudi koji žive po principima prošlog veka imaju znatno duži životni vek nego ljudi 21. veka. U Bibliji se spominje da su ljudi oko doba Hrista živeli od 300 do 900 godina. Imajući sve to u vidu, mnogobrojni su činioci od kojih zavisi životni vek a za neke se ni ne zna na koji način doprinose.Na internetu se nalaze mnogobrojni kalkulatori za izračunavanje očekivane dužine života. U ovu svrhu koriste se faktori: pol, godine, težina, visina, stepen obrazovanja, nasledne bolesti, dužina života ostalih članova porodice, visina krvnog pritiska, varijacije u težini, nivo holesterola, konzumiranje alkohola i cigareta, način ishrane, vežbanje, način vožnje. U zavisnosti od vrste kalkulatora očekivana dužina autorovog života se kreće od 85 (http://www.uwic.ac.uk) do 102 godine (http://calculator.livingto100.com).U osiguranju je bitno imati tablice života radi računanja premije osiguranja. Najčešće se rade istraživanja po zemljama (tabela 30). U tu svrhu uzimaju se ekonomski faktori u zemljama (GDP, stopa nezaposlenosti, inflacija i sl). Očekivana dužina života u Srbiji je 74 godine. Današnji životni vek se kreće u rasponu od 40-80 godina, na slici 31 je prikazana prosečna dužina ljudskog života po državama (http://geography.about.com).

Zemlje s najdužom očekivanom dužinom života		Zemlje s najkraćom očekivanom dužinom života	
Andora	83.5	Svaziland	33.2

San Marino	82.1	Bocvana	33.9
Singapur	81.6	Lesoto	34.5

tabela 30



slika 31

Najduži životni vek je u Kanadi i Australiji (preko 80 godina) dok je najkraći u raznim delovima Afrike (ispod 50 pa negde i 40 godina). Očito da dužina prosečnog životnog veka čoveka zavisi od razvijenosti zemlje.

Postoji više načina predviđanja dužine života,pravljenje modela višestrukom regresijom je samo jedan od njih. Logičnije bi bilo koristiti nelinearan model i vremenske serije. Ipak zanimljivo je proveriti koliko bi se dobar model dobio preko višestruke linearne regresije. Dati su faktori (prilog-tabela 9):

TV- broj televizora na 100 osoba

Doktori-koliko ljudi leči 1 doktor

GDP-bruto domaći proizvod, predstavlja sumu bruto dodatih vrednosti svih aktivnosti jedne ekonomije, izražene u stalnim cenama, umanjenu za usluge finansijskog posredovanja indirektno merene i uvećanu za poreze minus subvencije na proizvode u stalnim cenama. GDP je pokazatelj jačine ekonomije bilo koje zemlje.

Na prvi pogled neki faktori su potpuno neočekivani. Pre nego što se krene s analizom modela pretpostaviće se šta bi ona mogla da pokaže.

Faktor TV bi trebao obrnuto da deluje na životni vek. Što ima manje televizora pretpostavlja se da se čovek bavi nečim drugim, npr zdravijim. S druge strane televizor može da bude pokazatelj ekonomije pa bi u tom slučaju delovao srazmerno. Ukoliko se to dogodi verovatno je onda ovaj faktor u velikoj korelaciji sa pokazateljom ekonomije tj GDP-om.

Faktor Doktor se čini da deluje srazmerno na životni vek. Što je veća medicinska pomoć to će ljudi biti zdraviji.

Faktor GDP bi takođe trebao biti srazmeran. Ovo je pokazivač jake ekonomije pa samim tim i standarda života. Što su bolji standardi i životni vek je duži.

Model dobijen preko ovih faktora se čini vrlo dobrim (tabela 31). $\bar{R}^2$ je 0.71. Sva 3 faktora su značajna. Samo faktor GDP deluje kako je pretpostavljeno. Može se smatrati da veći broj televizora predstavlja pokazatelj dobrog standarda dok manjak doktora utiče na bolje brinjenje o sebi.

N=119	Regression Summary for Dependent Variable: zivotni vek (Spreadsheet2 R= .84703339 R²= .71746556 Adjusted R²= .71009510 F(3,115)=97.343 p<0.0000 Std.Error of estimate: 5.6685					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(115)	p-level
Intercept			60.29627	0.974629	61.86587	0.000000
tv	0.206022	0.097375	0.11685	0.055226	2.11576	0.036523
doktor	########	0.055953	-0.00029	0.000042	-6.83536	0.000000
GDP	0.413736	0.097057	0.00071	0.000166	4.26282	0.000042

Tabela 31

Regresiona jednačina: $Y = 60.3 + 0.12X_1 - 0.0003X_2 + 0.0007X_3$

Faktori GDP i TV su u visokoj korelaciji (-0.82)(tabela 32). To pokazuje i R^2 za ova dva faktora zasebno (tabela 33).

Variable	Correlations of Regression Coefficients B; DV: zivotni vek (Spreadsheet2		
	tv	doktor	GDP
tv	1.000000	0.150063	########
doktor	0.150063	1.000000	0.126919
GDP	########	0.126919	1.000000

tabela 32

Variable	Redundancy of Independent Variables; DV: zivotni vek (zivot.st R-square column contains R-square of respective variable with all other independent variables			
	Toleran.	R-square	Partial Cor.	Semipart Cor.
tv	0,259109	0,740891	0,193564	0,104871
doktor	0,784750	0,215250	-0,537498	-0,338804
GDP	0,260808	0,739192	0,369395	0,211292

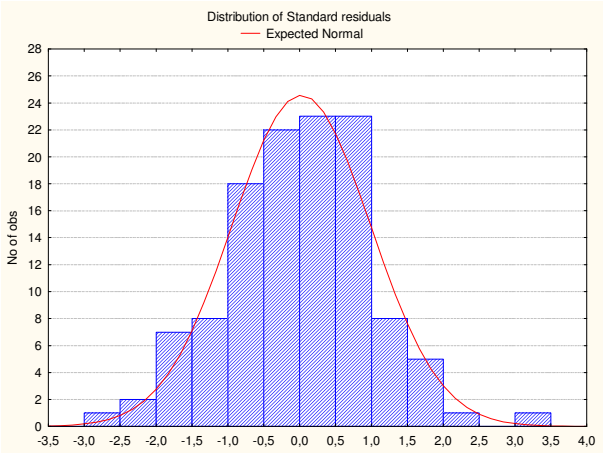
tabela 33

Logično je pretpostaviti da faktor TV zavisi od faktora GDP. Da bi se izbegao visok stepen multikolinearnosti pravi se novi model samo sa dva faktora: Doktor i GDP (tabela 34).

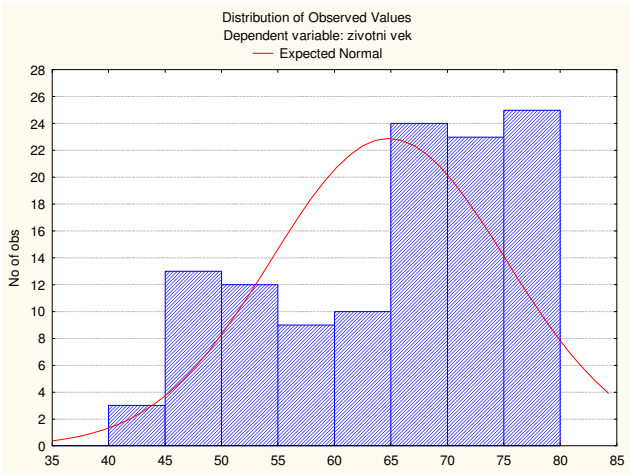
N=119	Regression Summary for Dependent Variable: zivotni vek (zivot.st R= ,84051635 R²= ,70646773 Adjusted R²= ,70140683 F(2,116)=139,59 p<0,0000 Std.Error of estimate: 5,7528					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(116)	p-level
Intercept			60,59755	0,978511	61,92832	0,000000
doktor	-0,400222	0,056142	-0,00030	0,000042	-7,12875	0,000000
GDP	0,582465	0,056142	0,00100	0,000096	10,37487	0,000000

tabela 34

$\bar{R}^2$ je 0.7, nije se mnogo promenio, model je i dalje dobar. Standardizovani reziduali imaju normalnu raspodelu (slika 32) . Zavisna promenljiva nema normalnu raspodelu (slika 33). To može značiti da je izabran pogrešan oblik regresijske jednačine.

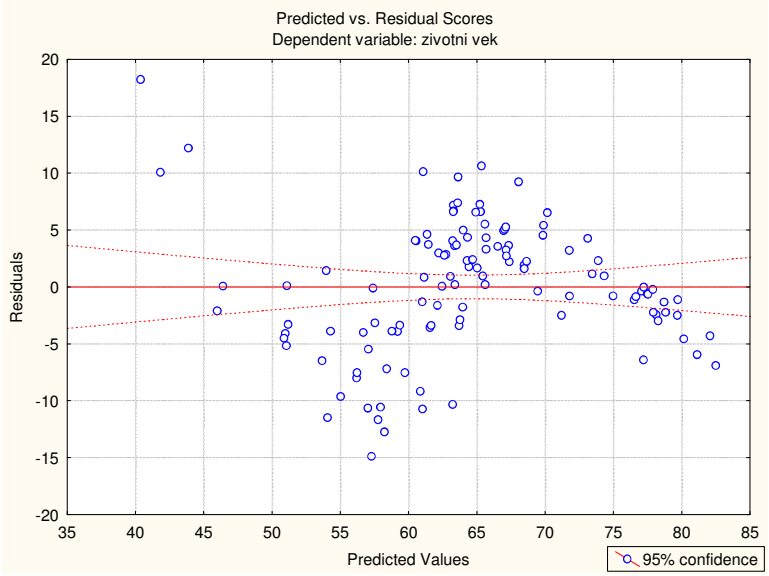


slika 32



slika 33

S obzirom da se vidi šablon na slici 34 regresiona jednačina bi trebala da bude u drugom obliku.



slika 34

Ukoliko bi se ozbacio faktor GDP a ostavio faktor Doktor, dobija se model:

$Y= 61.63 + 0.31X_1 - 0.0003X_2$, $\sigma^2=6.07$, $\bar{R}^2 = 0.67$, oba faktora su signifikantna.

Model dobijen sa ova dva faktora je lošiji od modela dobijenog s faktorima GDP i Doktor.

Naredni model je dobijen preko uzorka sa podacima za TV i Doktora (prilog-tabela 10).

Regresiona jednačina je: $Y= 70.25 - 0.024X_1 - 0.0004X_2$, $\sigma^2=6$, $\bar{R}^2 =0.4$, oba faktora su signifikantna.

Oba uzorka imaju podatke za iste faktore. Da li su stoga i regresione jednačine dobijene preko ovih modela jednake? Za proveru ove pretpostavke koristi se Chow-ov test:

$$\frac{(SSE-SSE_1-SSE_2)/k+1}{(SSE_1+SSE_2)/n+m-2(k+1)} \sim F_{k+1,n+m-2(k+1)}$$ gde su: za primer:

SSE - suma kvadrata odstupanja modela napravljenog od oba uzorka	8952,51
SSE_1 - suma kvadrata odstupanja modela napravljenog od prvog uzorka	4279,08
SSE_2 - suma kvadrata odstupanja modela napravljenog od drugog uzorka	1261,245
$k+1$ – broj koeficijenta	3
n – obim prvog modela	119
m – obim drugog modela	38

$$\frac{(SSE-SSE_1-SSE_2)/k+1}{(SSE_1+SSE_2)/n+m-2(k+1)} = \frac{1137,392}{38,47} = 29.56 \quad > \quad F_{k+1,n+m-2(k+1)} = F_{3,144} = 2,69$$

Modeli nisu jednaki. Moglo se pretpostaviti s obzirom da su podaci dobijeni iz različitih vremenskih perioda. Faktori se signifikantno menjaju kroz vreme. Stoga je verovatno bolje praviti model sa vremenskim serijama.

Primer 5 : Proizvodnja plastičnih kontejnera.

Plastika predstavlja veštački materijal proizveden od sintetčkih ili polusintetčkih smola i različitih dodataka (punila, omekšivača, stabilizatora i pigmenata) koji se u toku prerade nalaze bar povremeno u plastičnom stanju. Mogu se podeliti u dve glavne grupe: termoplastični materijali i termoreaktivni materijali ili duroplasti.

Termoplastične mase - grejanjem omekšaju, a hlađenjem se vraćaju u prvobitno stanje (npr. polivinilhlorid, polietilen, polistiren) . Sastoje se od vrlo dugih molekula s ravnim lancima (linearni polimeri).

Termoreaktivne plastične mase ili duroplasti-grejanjem ireverzibilno očvrsnu i kasnije se više ne mogu oblikovati (bakelit, aminoplasti). Imaju prostornu mrežastu strukturu.

Plastične mase prerađuju se valjanjem u folije, istiskivanjem pod pritiskom, ubrizgivanjem, tlačenjem itd.

Različitosti među sintetičkim polimerima, njihove najraznovrsnije karakteristike, ključne su za njihov uspeh. Polimerne stvari se retko upotrebljavaju u izvornom obliku, već im se prethodno dodaju razni dodaci (aditivi) koji bitno poboljšavaju jedno ili više njihovih svojstava, pa se tako dobivaju tehnički uportebljivi polimerni materijali.

Zajedinička im je:

- otpornost na hemikalije;
- odlična toplotna izlolacijska svojstva;
- elektroizolacijska svojstva;
- manja masa u odnosu na druge materijale sličnih svojstava. []

Jedan od mnogobrojnih proizvoda od ovog materijala je kontejner. Upotrebom različitog stepena zagrevanja i pritiska dobija se proizvod različite čvrstoće. Cilj je pronaći optimalnu temperature i pritisak za dobijanje najčvršćeg kontejnera. Dobijeni su uzorci, sa unapred određenim vrednostima nezavisnih faktora i prost (prilog-tabela 10 i tabela 11) za ispitivanje. Zanimljivo je primetiti da u ovom

primeru je nemoguće raditi analizu na celoj populaciji. Razlog za to je što su nezavisni faktori (temperature i pritisak) neprekidne promenljive pa ne bi se mogle ispitati sve različite vrednosti.

N=66	Regression Summary for Dependent Variable: cvrstoca (kontejneri.sta) R= ,97635323 R²= ,95326562 Adjusted R²= ,95178199 F(2,63)=642,52 p<0,0000 Std.Error of estimate: 1,4160					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(63)	p-level
Intercept			2,90909	1,585930	1,8343	0,071331
temperatura	0,817561	0,027236	0,16545	0,005512	30,0174	0,000000
pritisak	-0,533722	0,027236	-1,00000	0,051031	-19,5960	0,000000

Tabela 35

N=66	Regression Summary for Dependent Variable: cvrstoca (kontejneri 2.st R= ,97196407 R²= ,94471416 Adjusted R²= ,94295905 F(2,63)=538,27 p<0,0000 Std.Error of estimate: 1,6426					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(63)	p-level
Intercept			-3,62187	2,029360	-1,7847	0,079119
temperatura	0,841333	0,029629	0,19821	0,006980	28,3953	0,000000
pritisak	-0,470369	0,029629	-1,05245	0,066295	-15,8751	0,000000

Tabela 36

Regresiona jednačina sa unapred određenim vrednostima X: $Y = 2.91 + 0.16X_1\text{-}X_2$

Regresiona jednačina sa prostim slučajnim uzorkom: $Y = \text{-}3.62 + 0.2X_1\text{-}1.05X_2$

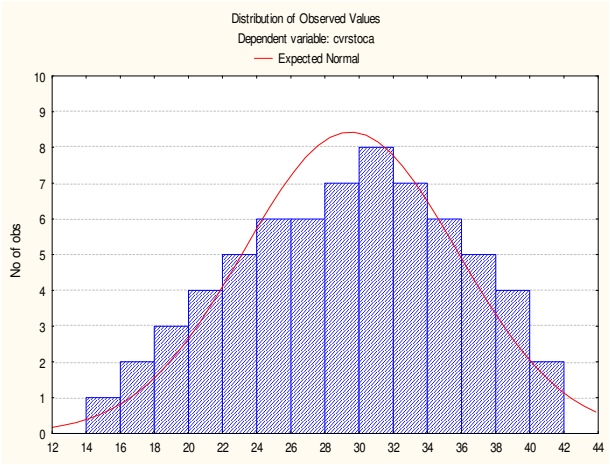
Oba faktora se pokazuju kao signifikantna (tabela 35 i tabela 36). $\bar{R}^2$ su respektivno 0.95 i 0.94 stoga su oba modela izuzetno dobra. Ne postoji nikakva korelacija između nezavisnih faktora (tabela 37 i 38). Izbegnut je visok stepen multikolinearnosti.

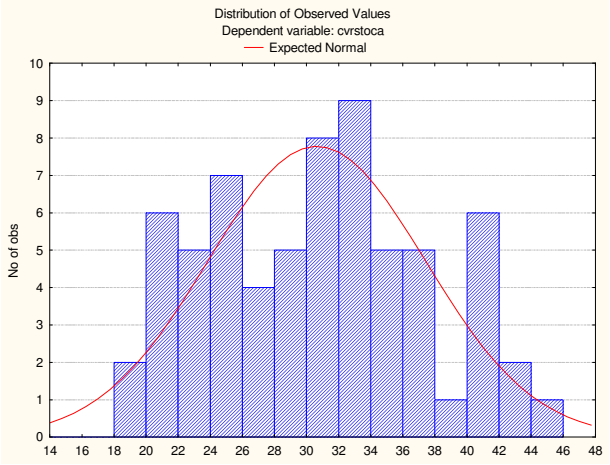
Variable	Correlations of Regression Coefficien	
	temperatura	pritisak
temperatura	1,000000	-0,000000
pritisak	-0,000000	1,000000

Variable	Correlations of Regression Coefficien	
	temperatura	pritisak
temperatura	1,000000	0,019743
pritisak	0,019743	1,000000

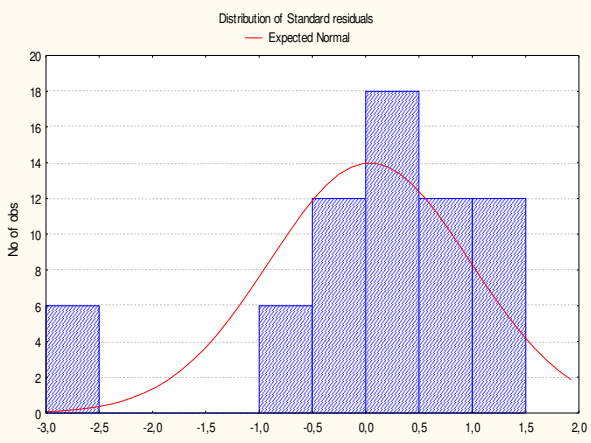
tabela 38

Zavisna promenljiva ima normalnu raspodelu (slika 35) kod uzorka sa unapred određenim vrednostima međutim kod prostog slučajnog uzorka nema (slika 36). Standardizovani reziduali nemaju normalnu raspodelu u oba modela (slika 37 i slika 38). Pošto se čini da je to jedini problem koji se javio u prvom modelu i dalje se pretpostavlja da je model dobar.

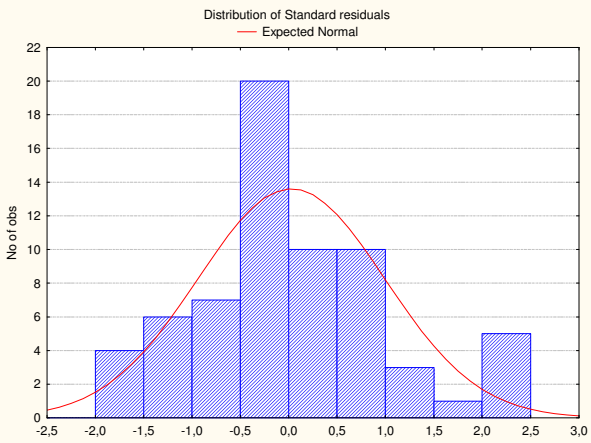


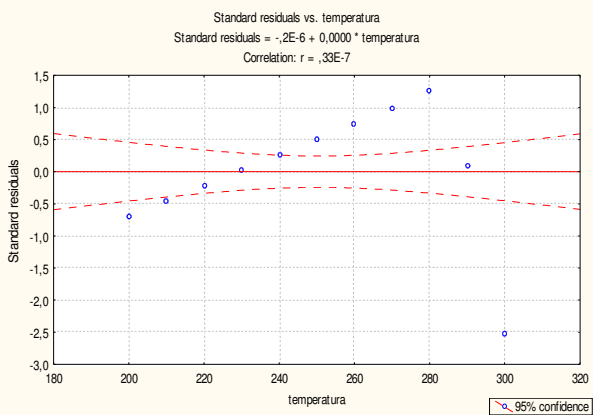


slika 35

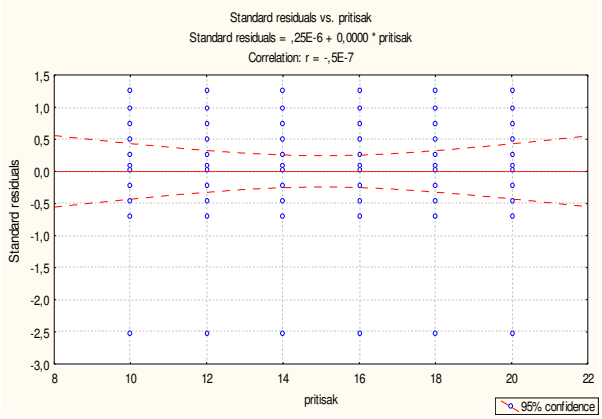


slika 36





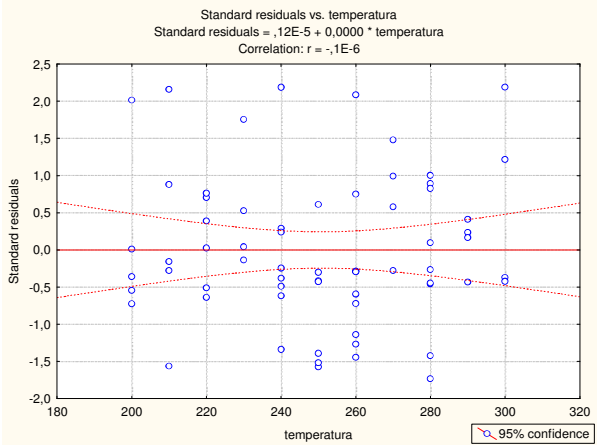
slika 37

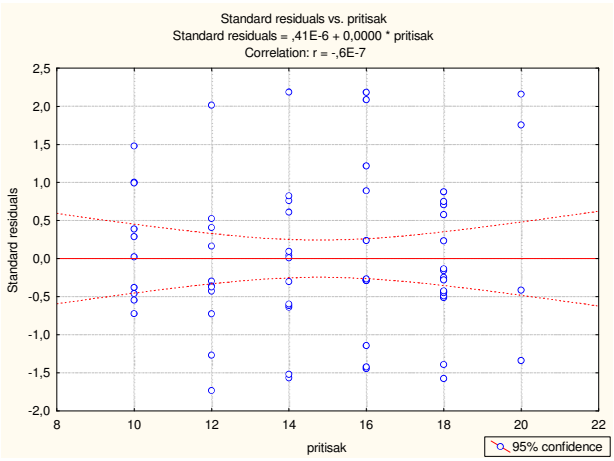


slika 38

slika 39

slika 40





slika 41

slika 42

Nije došlo do heteroskedastičnosti ni kod jednog modela (slike 39,40,41 i 42). Mogu se upotrebljavati i intervali poverenja i testovi značajnosti. Oba modela su vrlo dobra, ostaje još da se proveri da li ovi modeli imaju iste regresione jednačine, što se proverava preko Chow-og testa.

$$\frac{(SSE-SSE_1-SSE_2)/k+1}{(SSE_1+SSE_2)/n+m-2(k+1)} = \frac{(355.382-126.327-169.977)/3}{(126.327+169.977)/66+66-6} = 8.374 > 2.69 = F_{3,126}$$

Regresione jednačine nisu jednake. Po $\bar{R}^2$ i standardnom odstupanju bi se izabrala jednačina dobijena od uzorka sa unapred odabranim vrednostima X. Jednačina dobijena preko oba uzorka $Y=0.04+0.18X_1-1.03X_2$ ima $\bar{R}^2$= 0.938. Ovaj podatak pokazuje da povećanje uzorka ne mora uvek da poboljša model.

U tabeli 39 su data predviđanja po oba modela za nasumično odabrana opažanja.

temperatura	pritisak	čvrstoća	$Y = 2.91+0.16X_1-X_2$	$Y =-3.62 +0.2X_1-1.05X_2$	$Y=0.04+0.18X_1-1.03X_2$
220	18	22.5	21.3	21.04	21.2
260	14	32.6	31.9	33.2	35.6
280	14	36.6	35.2	37.1	36.2
300	10	45.3	42.5	45.3	43.9

tabela 39

Primer 6 : Otvaranje supermarketa

Otvaranje sopstvenog preduzeća je jedna od aktuelnih tema poslednjih par godina. Smatra se da je to trenutno jedini način sigurne zarade bez zavisnosti od mnogobrojnih državnih faktora. Prvi korak je naravno odlučiti se kakva bi to bila delatnost. Nakon toga ide podnošenje zahteva i sva druga papirologija pa samo otvaranje i poslovanje. Jedan od bitnih činilaca je sama lokacija. Na osnovu čega može da se zaključi da će upravo ta lokacija biti savršena za sopstvenu firmu? U slučaju otvaranja supermarketa lokacija ima vrlo važnu ulogu. Poslovanje neke prodavnice zavisi prvenstveno od njenih kupaca tj od njihovih primanja. Koliko tačno zavisi pokazaće naredna analiza. Na osnovu podataka (prilog-tabela 12) dobijen je model (tabela 40).

N=27	Regression Summary for Dependent Variable: kupovina (prodavnica.s R= ,96518805 R²= ,93158797 Adjusted R²= ,92266466 F(3,23)=104,40 p<,00000 Std.Error of estimate: 53,040					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(23)	p-level
Intercept			-324,608	51,16248	-6,34465	0,000002
prihod	0,787362	0,056988	0,105	0,00761	13,81630	0,000000
broj dece	0,673691	0,057687	96,601	8,27177	11,67843	0,000000
broj odraslih	0,285778	0,058283	110,761	22,58901	4,90331	0,000059

tabela 40

Regresiona jednačina: $Y = -324.61+0.105X_1+96.6X_2+110,76X_3$

Na sumu potrošenu na kupovinu utiče prihod porodice i raspodela odraslih i dece. Sva tri faktora bitno utiču na potrošenu sumu. Model je vrlo dobar ($\bar{R}^2$ je 0.92). Ne postoji visok stepen multikolinearnosti jer su svi parcijalni koeficijenti daleko od 1 (tabela 41) kao i zasebni koeficijenti determinacije (tabela 42).

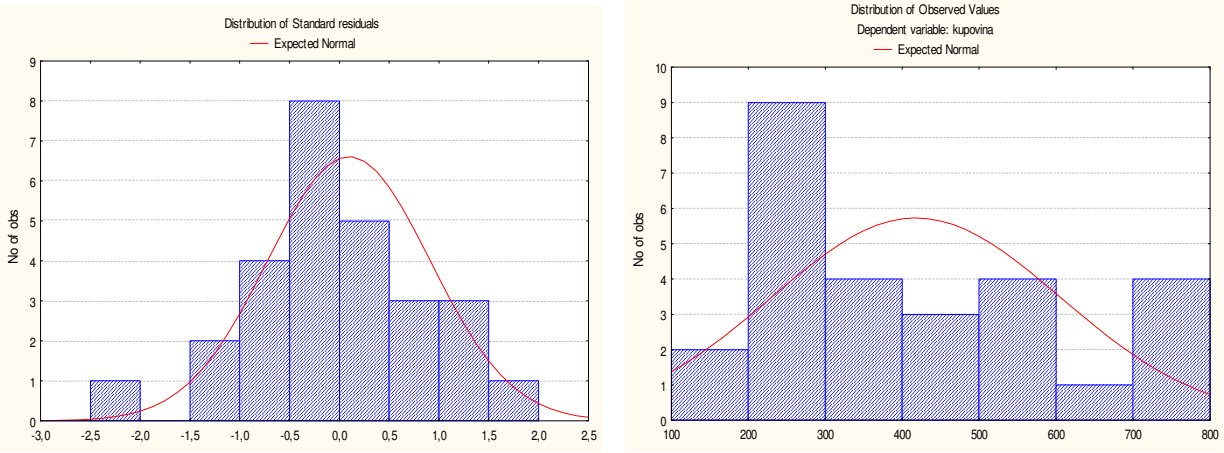
Variable	Correlations of Regression Coefficients B; DV: kupovina (prodavnica.s		
	prihod	broj dece	broj odraslih
prihod	1,000000	0,147429	0,204045
broj dece	0,147429	1,000000	-0,254392
broj odraslih	0,204045	-0,254392	1,000000

tabela 41

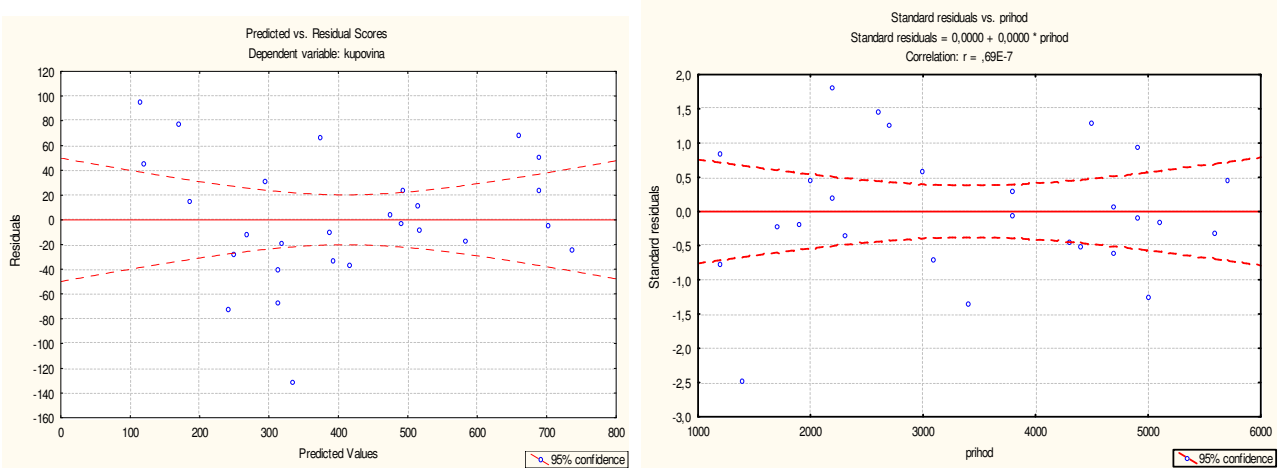
Variable	Redundancy of Independent Variables; DV: kupovina (prodavnica.s R-square column contains R-square of respective variable with all other independent variables			
	Toleran.	R-square	Partial Cor.	Semipart Cor.
prihod	0,915881	0,084119	0,944705	0,753519
broj dece	0,893824	0,106176	0,925038	0,636923
broj odraslih	0,875642	0,124358	0,714898	0,267419

tabela 42

Standardizovani reziduali imaju približno noramlnu raspodelu (slika 43) međutim zavisna promenljiva nema (slika 44). To može biti direktna posledica odsustva normalne raspodele, što u ovom primeru nije slučaj, ili je sam uzorak doveo do toga. Ako analiza pokaže da su sve ostale provere u redu onda je problem do izbora uzorka.



slika 43



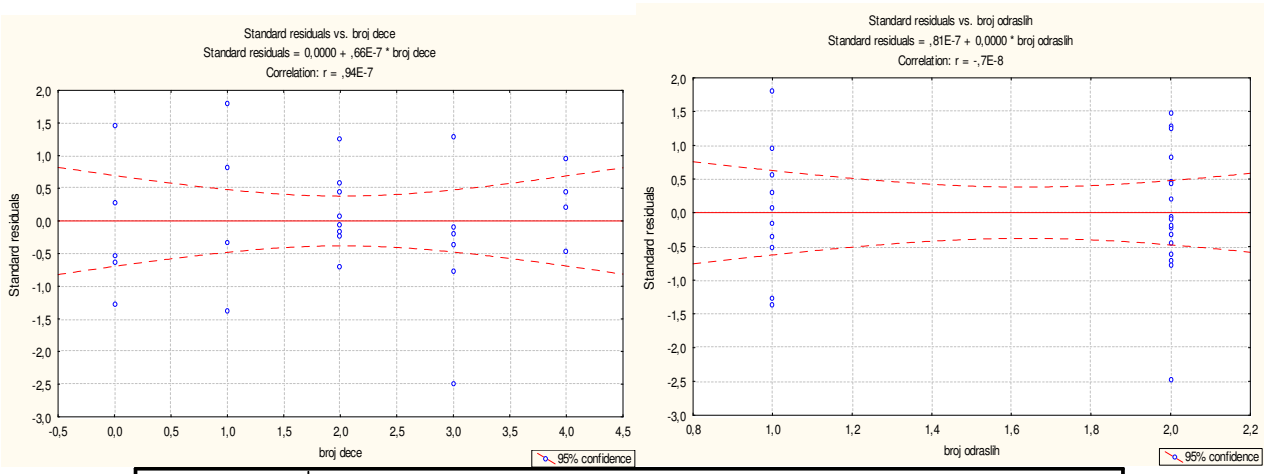
slika 44

slika 45 slika 46

slika 47 slika 48

Upoređujući standardizovane reziduale sa nezavisnim faktorima (slike 45, 46 ,47) ne primećuje se poseban šablon što dovodi do zaključka da nije došlo do narušavanja osnovnih pretpostavki. Može se zaključiti da odstupanja nisu svuda jednaka ali nisu signifikantna odnosno i dalje važi homoskedastičnost (slika 48).

Jedini outlier je opažanje broj 21. Izbacivanjem ovog opažanja dobija se još bolji model (tabela 43). $\bar{R}^2$ ovog modela je 0.94, standardno odstupanje se smanjilo sa 53 na 45.1.



N=26	Regression Summary for Dependent Variable: kupovina (prodavnica.s R= ,97483476 R²= ,95030282 Adjusted R²= ,94352593 F(3,22)=140,23 p<,00000 Std.Error of estimate: 45,127					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(22)	p-level
Intercept			-311,289	43,73807	-7,11713	0,000000
prihod	0,736298	0,049116	0,100	0,00668	14,99112	0,000000
broj dece	0,694360	0,049853	98,322	7,05928	13,92804	0,000000
broj odraslih	0,299283	0,050313	114,550	19,25737	5,94837	0,000005

tabela 43

Regresiona jednačina dobijena bez 21. opažanja je: $Y = -311.29 + 0.1X_1 + 98.3X_2 + 114.6X_3$

Primer 7 : Ponuda i potražnja hleba

Hleb je jelo napravljeno od testa koje se sastoji od brašna različitih vrsta (pšenično, ražano, kukuruzno), izmešanog sa drugim suvim ili tečnim sastojcima.Testo je obično kombinovano sa

sredstvom za podizanje, kao što je kvasac, izgnječeno i izmešano u jednoliku masu, zatim oblikovano u štruce i pečeno. Dodaje se i so, a u industrijskoj proizvodnji i emulgatori. Pri pravljenju peciva, pre pečenja, mogu se na testo dodati začini kao što su: veoma lekoviti čurokot, kim, susam i slično.

Prvi su hleb, i to od pšenice i ječma, pekli Egipćani 1500 godina pne., verovatno je bio beskvasni i zapravo pogača od brašna, soli, koja se pekla u žeravici i pod pepelom u peći, na što upućuju slike na zidovima grobnica i zapisi starih Egipćana. Hleb su pekli u najrazličitijim oblicima, često u obliku životinja ili ljudi, jer su služili za verske ili magijske potrebe i posipali ga raznim semenkama (često kimom).

U srednjem veku hleb pripremljen od belog brašna bio je privilegija bogatih, a raženi su dobijali zatvorenici, kojeg su ponekad jeli fratri kako bi pokazali svoju poniznost, a status se procenjivao po boji hleba, beli su jeli oni na vrhu društvene lestvice, a oni na dnu, crni hleb.

19. vek je bila značajna prekretnica u prehrani stanovništva Evrope, jer je započela masovna proizvodnja belog brašna, a svi nedostaci belog hleba kao simbola statusa pokazaće se tek u 20. veku.

Najpopularnije vrste hleba i peciva su: hleb pravljen od brašna različitih žitarica (pšenica, raž, ovas, ječam...), proja (hleb od kukuruznog brašna), kifle, buhtle, lepinja, pogača, đevrek, pereca, zemička, lakumić, pletenica, kroasan, baget, somun, dvopek, peksmit.[]

I u 21. veku hleb je neizostavno jelo svakodnevne balkanske trpeze. Može se reći da spada u osnovnu prehranu ovih prostora. Po toj karakteristici trebao bi da spada u osnovna dobra tj da njegova tražnja ne zavisi od cene. Da li je zaista tako pokazaće sledeća analiza. Prikupljeni su podaci (prilog-tabela 13) na osnovu kojih je napravljan model (tabela 44).

Regresiona jednačina: $Y = -0.015 + 0.012X_1 + 0.000002X_3 + 0.98X_4$

N=29	Regression Summary for Dependent Variable: kolicina (hleb.sta R= ,50050189 R²= ,25050214 Adjusted R²= ,16056240 F(3,25)=2,7852 p<,06162 Std.Error of estimate: 2,2054					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(25)	p-level
Intercept			-0,015034	2,625951	-0,005725	0,995477
cena	0,080541	0,173286	0,011535	0,024819	0,464788	0,646106
prihod	0,022495	0,225270	0,000002	0,000023	0,099859	0,921253
broj clanova	0,481519	0,225169	0,984293	0,460277	2,138479	0,042435

tabela 44

Iako model nije dovoljno dobar ($\bar{R}^2$ je svega 0.16) sasvim su logični rezultati nezavisnih faktora. Količina kupljenog hleba zavisi samo od broja članova porodice. Bez obzira na cenu hleba i na primanja porodice uvek će morati da se konzumira hleb, makar samo on sam.

Postoji sumnja da je došlo do određenog stepena multikolinearnosti između faktora prihoda i broja članova (tabela 45 i tabela 46) mada verovatno nije signifikantna.

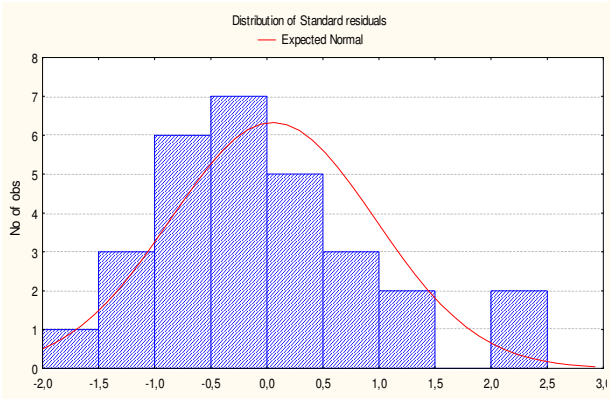
Variable	Correlations of Regression Coefficients B; DV: kolicina (hleb.sta		
	cena	prihod	broj clanova
cena	1,000000	0,029940	0,001251
prihod	0,029940	1,000000	-0,638966
broj clanova	0,001251	-0,638966	1,000000

tabela 45

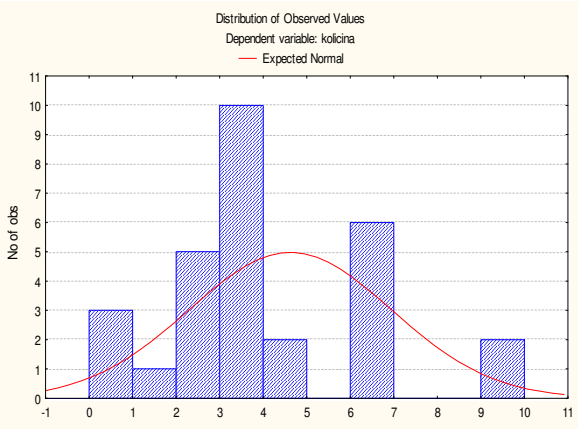
Variable	Redundancy of Independent Variables; DV: kolicina (hleb.st R-square column contains R-square of respective variable with all other independent variables			
	Toleran.	R-square	Partial Cor.	Semipart Cor.
cena	0,998402	0,001598	0,092559	0,080477
prihod	0,590778	0,409222	0,019968	0,017290
broj clanova	0,591307	0,408693	0,393239	0,370271

tabela 46

Standardizovani reziduali imaju normalnu raspodeu (slika 49) dok zavisan faktor nema (slika 50).

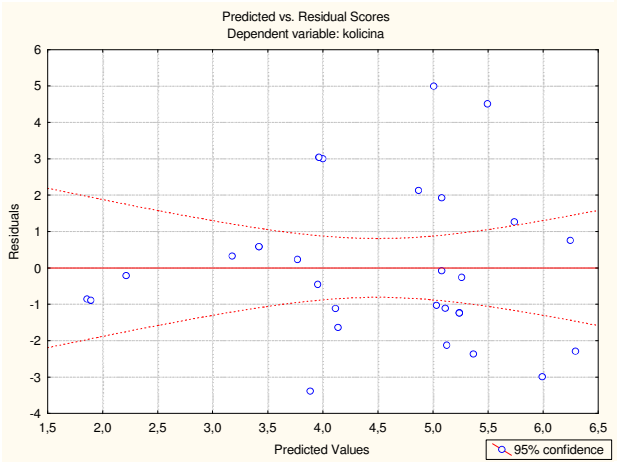


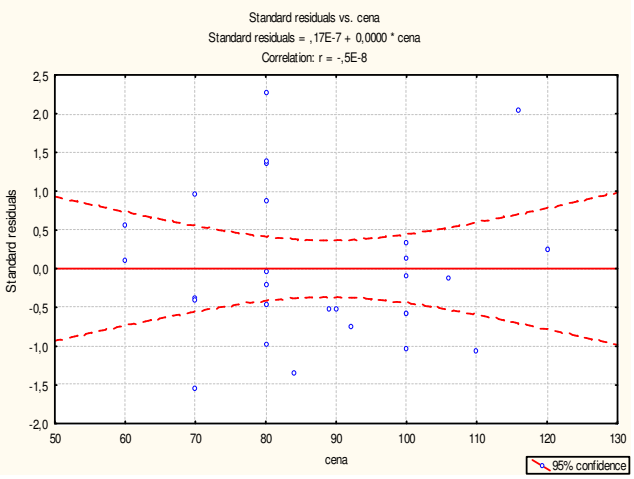
slika 49



slika 50

Postoji velika verovatnoća da je došlo do heteroskedastičnosti (slike 51,52,53 i 54).



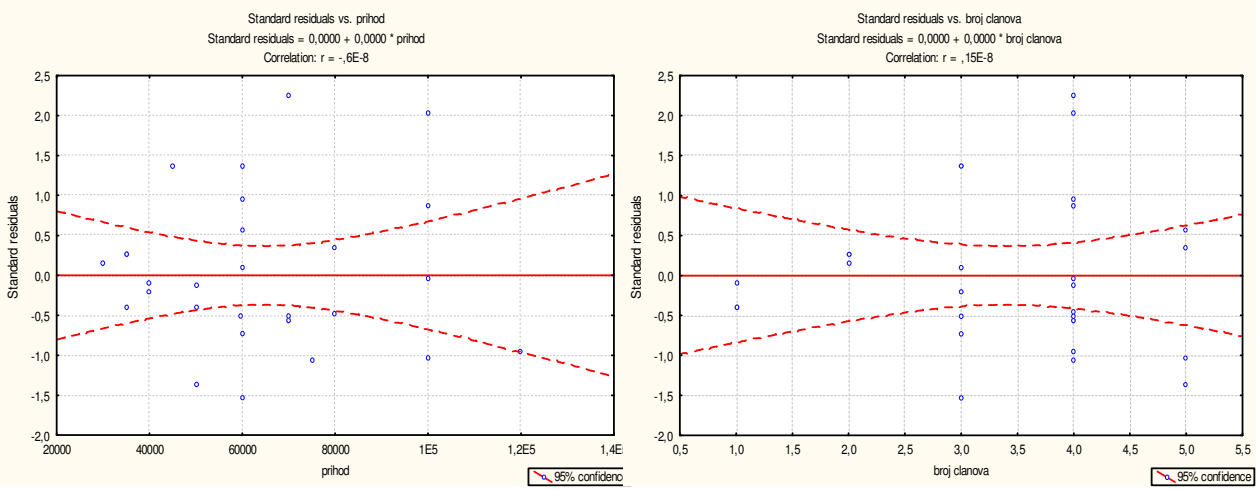


slika 51 slika 52
slika 53 slika 54

Model se ne može upotrebiti jer je došlo do visokog stepena multikolinearnosti i heteroskedastičnosti. Ipak model je ukazao na osobinu hleba kao osnovnog dobra.

S druge strane hleb se može smatrati inferiornim dobrom kao što je u Gifenovom paradoksu. Na početku 20-og veka Gifen je zapazio da potražnja hleba raste sa cenom istog.To je posledica činjenice da veliki deo stanovništva ima nizak realni dohodak pa zbog toga kupuju inferiorna dobra.Kada cena tih dobara raste stanovnici povećavaju svoju kupovinu kako bi kompenzovali još veće sužavanje mogućnosti kupovine superiornih dobara. []

Nedavno istraživanje je pokazalo da je potrošnja hleba u Srbiji i dalje na visokom nivou prvenstveno zbog niskog standarda života. Potrošnja po glavi stanovnika na godišnjem nivou iznosi preko 100 kg što je tri puta više nego u potrošnji zapadno- evropskih zemalja. Iako je veća potrošnja peciva i različitih vrsta hleba, dok je potrošnja standardnog hleba umanjena, proizvodnja peciva čini od



3% do 5% ukupne proizvodnje.[]

Primer 8 : Ponuda i potražnja mleka

Mleko predstavlja belu ili žućkastu tečnost koju luče mlečne žlezde ženki roda sisavaca. Mleko služi mladuncu, odnosno dojenčetu, kao jedina hrana u prvim danima života i potpuno osigurava normalan rast i razvoj mladog sisavca. Mleko sadrži sve potrebne prehrambene sastojke: mast, belančevine, ugljene hidrate, vitamine, minerale,enzime i antitela.

Uz ženino mleko se za prehranu najčešće koristi kravlje, ovčije i kozje mleko koje se dobija mužnjom. Od te tri vrste mleka se proizvode različiti mlečni proizvodi, od kojih je najpoznatiji sir.

Za ljudsku prehranu najčešće se koristi kravlje mleko i ono se najčešće priprema industrijski za čuvanje i dalju upotrebu. To je najčešće korišćeno mleko u razvijenim zemljama, dok se u siromašnijim zemljama (Afrika i delovi Azije) vrlo često koristi kozje mleko.

Kravlje mleko je dostupno u više varijanti. U nekim državama to su:

- punomasno
- poluobrano (s otprilike 1.5-1.8% masti)
- obrano (oko 0.1% masti)

Punomasno mleko ima oko 3-4% masti. Za poluobrano i obrano mleko, sva mast se uklanja i onda se (u slučaju poluobranog) delom vraća. Najprodavanija varijanta je poluobrano mleko, jer se u mnogim državama smatra da je punomasno manje zdravo, a obrano neukusno. Punomasno mleko se preporučuje da pruži dovoljno masti za razvoj male djece koja su prerasla majčino mleko.

Kritičari mleka tvrde da ono može na zdravlje da ima neželjene posledice koje nadjačavaju koristi. Oni ukazuju na znanstvene studije koje sugerišu da postoje veze između mleka i nekih zdravstvenih problema: raka jajnika , raka prostate , povećanim rizikom od ateroskleroze i srčanih oboljenja, nepodnošljivosti laktoze (javljaju se grčevi, proliv, nadutost), prouzrokovanje akni. []

Proizvodnja mleka u Srbiji je bila jedna od najvažnijih poljoprivrednih grana, trebalo bi da bude okosnica poljoprivrede i ruralnog razvoja Srbije kao dela Zajedničke agrarne politike EU, u budućnosti.

U Srbiji trenutno radi 201 mlekara iako je broj registrovanih mlekara daleko veći. Najveći deo su srednje mlekare, dok na velike industrijske mlekare otpada nešto više od 10% od ukupnog broja mlekara.

Iako potrošnja konstantno raste od 2000. godine ona je još uvak jako mala. Tako je ukupna proizvodnja i potrošnja mleka po stanovniku u 2007. godini iznosila je 207 litara. Ovo je izuzetno malo u odnosu na zemlje EU i okruženja. Tako na primer Danska troši 897 litara po glavi stanovnika, a Bugarska 283 litra.

Mlečnost po kravi u Srbiji iznosi oko 2600 litara godišnje, što je u proseku za 18% manje od svetskog proseka tj 40% manje u odnosu na evropski. U SAD prosečna mlečnost krava iznosi 5000 litara a u Izraelu 9000 litara. []

Mleko takođe spada u inferirona dobra. Najveća konzumacija je kod grupe dece. Da li će se ova činjenica potvrditi preko modela? Koliko primanja ,broj članova porodice po strukturi odrasli-deca i sama cena mleka utiču na njegovu prodaju?

Prihod porodice i broj dece u porodici se izdvajaju kao signifikantni faktori (tabela 47). Samo faktor cene deluje obrnuto srazmerno na količinu mleka. Iako mleko spada u osnovna dobra došlo je do uticaja primanja. Pošto je uzorak (prilog-tabela 14) dobijen od ispitanika, stanovnika Srbije, ovaj rezultat nije u potpunoj kontradikciji. U predhodnom delu rada je napomenuto da je u Srbiji potrošnja mleka na niskom nivou. Može se zaključiti da mleko se ne gleda kao tipično osnovno dobro.

N=27	Regression Summary for Dependent Variable: kolicina (mleko.sta R= ,74159697 R²= ,54996606 Adjusted R²= ,46814171 F(4,22)=6,7213 p<,00108 Std.Error of estimate: 2,8750					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(22)	p-level
Intercept			3,202884	3,616584	0,88561	0,385405
cena	-0,210438	0,145193	-0,062056	0,042815	-1,44938	0,161339
prihod	0,441431	0,197796	0,000073	0,000033	2,23175	0,036137
broj clanova	0,089799	0,209929	0,326753	0,763874	0,42776	0,672985
deca	0,426903	0,157229	3,020045	1,112288	2,71516	0,012641

tabela 47

Regresiona jednačina: $Y= 3.2 -0.06X_1+0.00007X_2+0.33X_3+3.02X_4$

$\bar{R}^2$ je 0.47, standardno odstupanje zavisne promenljive je 2.88.

Postoji verovatnoća da su prihod i broj članova porodice u korelaciji (tabela 48).Logično je smatrati da svaki odrastao čovek zarađuje pa samim tim prihodi porodice rastu s porastom odraslih članova. Ipak stepen korelacije nije izrazito visok (tabela 49).

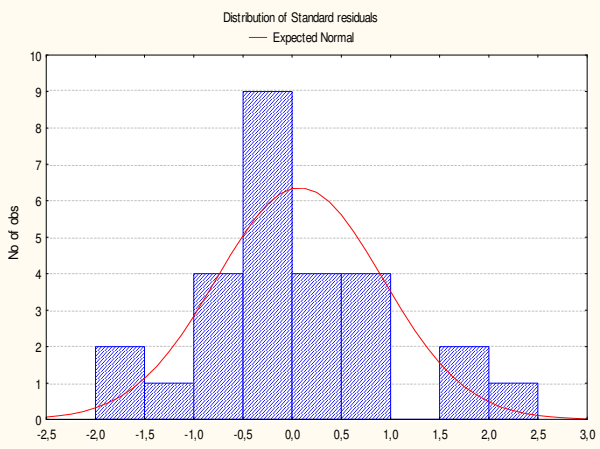
Variable	Correlations of Regression Coefficients B; DV: kolicina (mleko.sta			
	cena	prihod	broj clanova	deca
cena	1,000000	0,040549	-0,032490	-0,147639
prihod	0,040549	1,000000	-0,676410	0,099679
broj clanova	-0,032490	-0,676410	1,000000	-0,341779
deca	-0,147639	0,099679	-0,341779	1,000000

tabela 48

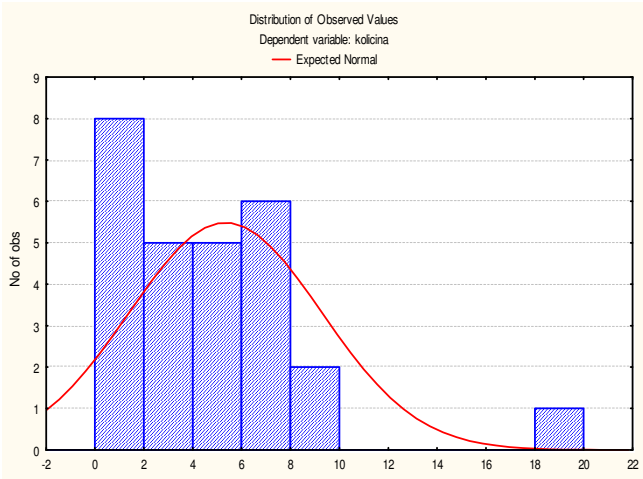
Variable	Redundancy of Independent Variables; DV: kolicina (mleko.sta R-square column contains R-square of respective variable with all other independent variables			
	Toleran.	R-square	Partial Cor.	Semipart Cor.
cena	0,970363	0,029637	-0,295234	-0,207297
prihod	0,522863	0,477137	0,429654	0,319195
broj clanova	0,464169	0,535831	0,090821	0,061180
deca	0,827478	0,172522	0,500989	0,388336

tabela 49

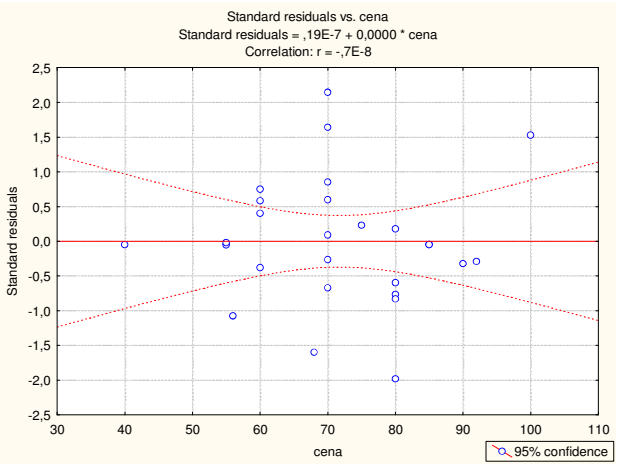
Nažalost nije ispunjena pretpostavka o normalnoj raspodeli, ni kod standardizovanih reziduala (slika 55), ni kod zavisne promenljive (slika 56). Takođe je skoro sigurno došlo do heteroskedastičnosti (slike 57, 58, 59). Nažalost nije očito kakav oblik ima odstupanje pa se ne može nastaviti sa analizom u tom pravcu.



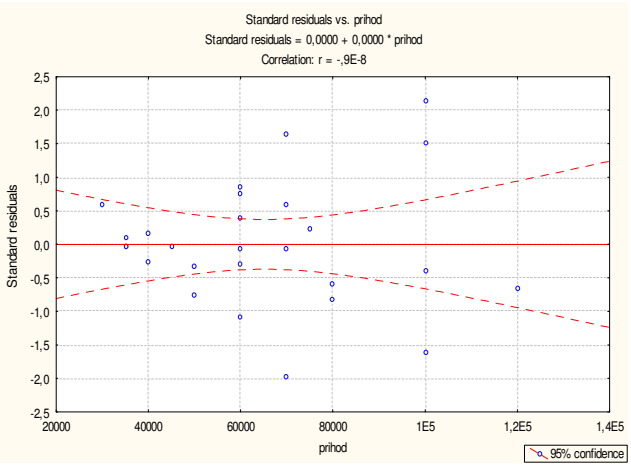
slika 55



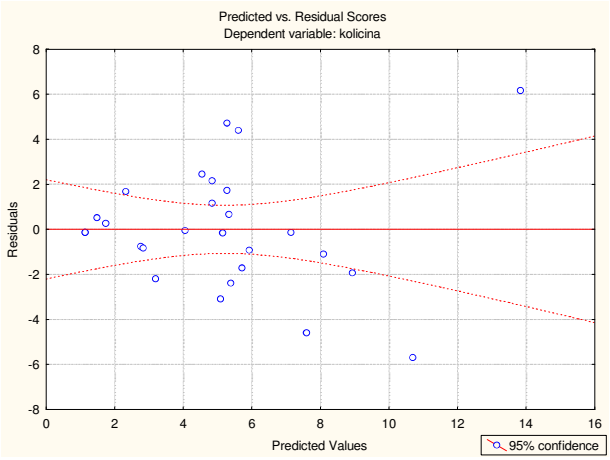
slika 56



slika 57



slika 58



slika 59

S obzirom da je došlo do heteroskedastičnosti, odstupanja od normalne raspodele kao i da je prilagođeni koeficijent determinacije relativno nizak (0.47) ovaj model nije za praktičnu upotrebu.

Primer 9 : Troškovi doma

Troškovi doma tj komunalije spadaju u svakodnevne životne obaveze. Visina ovih troškova zavisi od više faktora: veličine stana, mesta stanovanja, broja članova porodice… Najtipičnije komunalije su: grejanje, voda, struja… Verovatno zvuči čudno raspravljati o komunalijama nakon analize potražnje za mlekom i hlebom međutim ove tri stvari imaju nešto zajedničko, sve spadaju u dobra koja imaju kontrolne cene. Ova pojava je tipična u zemljama realsocijalizma i ispoljava se na dva načina:

- Određivanje najviših dozvoljenih prodajnih cena
- Određivanje najnižih mogućih otkupnih cena

Ovakva kontrola predstavlja prikrivene subvencije. Tržišne cene na osnovu slobodnog delovanja ponude i tražnje, bi se formirale na višem ravnotežnom nivou ali ih država prisilno zadržava na nižem nivou. U prilog takvom postupku obično se navode razlozi za zaštitu interesa potrošača. Ukoliko bi se dopustilo

formiranje ravnotežnih cena došlo bi do porasta istih i jedan deo potrošača sa niskim dohocima ne bi bio u stanju da pribavi ove robe. Zbog toga su ograničene cene hleba, mleka, komunalija…
Određivanje najnižih otkupnih cena ide u prilog poljoprivrednim proizvođačima gde im država to garantuje. Ovakva subvencija je opravdana jer takva proizvodnja je vrlo rižična, sezonskog je karaktera i zahteva kreditiranje. []

S obzirom da se cena komunalija ne definiše preko uobičajene ravnoteže, koji se onda faktori izdvajaju kao relevantnim ?

N=28	Regression Summary for Dependent Variable: troskovi (stan.st R= ,57541556 R²= ,33110307 Adjusted R²= ,17908104 F(5,22)=2,1780 p<,09362 Std.Error of estimate: 50053,					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(22)	p-level
Intercept			-3772,2	35595,46	-0,10597	0,916564
prihod	0,498142	0,235410	1,2	0,56	2,11606	0,045891
udaljenost	-0,279853	0,206178	-922,6	679,71	-1,35734	0,188437
starost	-0,097583	0,185140	-274,2	520,14	-0,52707	0,603417
kvadratura	0,378814	0,227367	329,8	197,95	1,66609	0,109871
broj clanova	-0,297574	0,247479	-13734,7	11422,49	-1,20242	0,241982

tabela 50

Nažalost jedino se primanja ističu kao bitan faktor (tabela 50) što je u kontradikciji sa prethodno navedenim. S obzirom da je $\bar{R}^2$ nizak (0.18) model nije dobar. Uzorak koji se koristio za dobijanje ovog modela nije reprezentativan (prilog-tabela 15).

U tipične komunalije spadaju računi za struju, vodu i grejanje. U Novom Sadu su za ove komunalije zaduženi Elektrovojvodina, Vodovod i kanalizacija i Informatika.

Svako domaćinstvo ima sopstveni strujomer koji evidentira potrošnju struje. Koliko će koje domaćinstvo potrošiti struje zavisi od njega samog. Bitni faktori su veličina doma, broj električnih aparata, broj članova.... Cena kWh se povećava kako se povećava potrošnja. Cilj ovakvog plaćanja je stimulisati građane na manju potrošnju. Ipak pošto se potrošnja povećava sa kvadraturom doma i članova ova optimalnost ponekad ne može biti dostignuta.

Potrošnja vode se obračunava na sledeći način: ukupna potrošnja vode u nekom objektu se deli srazmerno na broj manjih celina prema evidentiranom broju članova. Ovakva metoda može da ima brojne slabosti kao što su pogrešna evidencija broja članova i nesrazmerna potrošnja između članova.
Stoga se ne može predvideti koliki bi bio račun za vodu pojedinog domaćinstva jer su nepredviđeni faktori izuzetno uticajni na račun.

Grejanje se obračunava slično kao potrošnja vode. Ukupan račun za grejanje se deli srazmerno na površine celina unutar objekta. S obzirom da je cena grejanja konstantna (toplana reguliše protok tople vode pa samim tim i jačinu) račun za grejanje je predvidljiv. Varijacije se mogu javiti samo u slučaju smanjenja grejanja od samih korisnika čiji se uticaj smanjenja prenosi na ceo objekat u čiju celinu spadaju.

Pomoću novog uzorka (prilog-tabela 16) napravljen je model za predviđanje računa za struju (tabela 51).

N=34	Regression Summary for Dependent Variable: Potrošnja struje (Racu R= .92269941 R²= .85137420 Adjusted R²= .83651161 F(3,30)=57.283 p<.00000 Std.Error of estimate: 135.42					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(30)	p-level
Intercept			-358.442	198.7358	-1.80361	0.081344
mesečni prihod	0.242790	0.439760	0.075	0.1361	0.55210	0.584973
broj stanara	0.246939	0.130200	55.088	29.0452	1.89662	0.067544
površina	0.581374	0.467637	0.281	0.2261	1.24322	0.223416

tabela 51

Regresiona jednačina: $Y= -358.4+0.075X_1+55.1X_2+0.28X_3$

Iako se model čini vrlo dobrim, $\bar{R}^2$ je 0.84, skoro sigurno postoji visok stepen multikolinearnosti. Svi *t-testovi* ne pokazuju nijedan značajan faktor dok je *F-vrednost* daleko od kritične oblasti (57.283>2.92). Korelacija između nezavisnih faktora je izuzetno visoka (tabela 52) i R^2 za svaki faktor pojedinačno je jako blizu 1 (tabela 53). Stoga je sigurno zaključiti da postoji visok stepen multikolinearnosti.

Variable	Correlations of Regression Coefficients B; DV: Potrošnja struje (Racun.st		
	mesečni prihod	broj stanara	površina
mesečni prihod	1,000000	0,814040	-0,986838
broj stanara	0,814040	1,000000	-0,837664
površina	-0,986838	-0,837664	1,000000

tabela 52

Variable	Redundancy of Independent Variables; DV: Potrošnja struje (Racun.st R-square column contains R-square of respective variable with all other independent variables			
	Toleran.	R-square	Partial Cor.	Semipart Cor.
mesečni prihod	0,025618	0,974382	0,100290	0,038860
broj stanara	0,292249	0,707751	0,327212	0,133496
površina	0,022655	0,977345	0,221349	0,087505

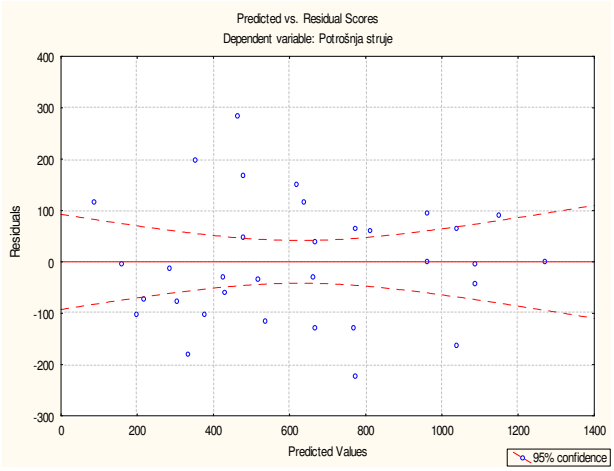
tabela 53

Samo se opažanje broj 7 ističe kao outlier.

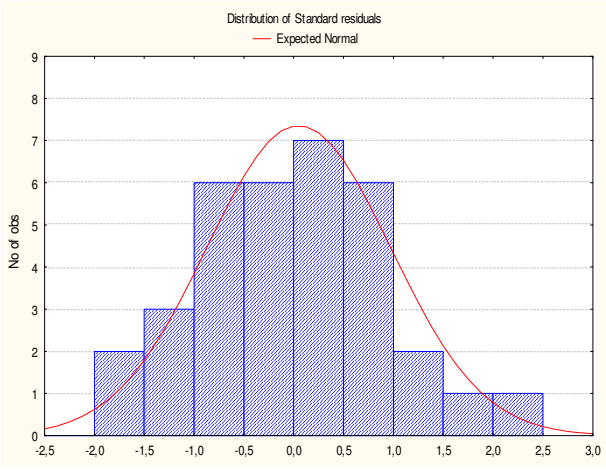
Izbacivanjem ovog opažanja poboljšava se model (tabela 54) koji je homoskedastičan (slika 60) i ima normalnu rasodelu (slika 61) ali i dalje postoji visok stepen multikolinearnosti.

N=33	Regression Summary for Dependent Variable: Potrošnja struje (Racun.st R= ,94035880 R²= ,88427468 Adjusted R²= ,87230309 F(3,29)=73,864 p<,00000 Std.Error of estimate: 121,02					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(29)	p-level
Intercept			-485,701	182,8427	-2,65639	0,012704
meseèni prihod	0,380416	0,384130	0,121	0,1226	0,99033	0,330203
broj stanara	0,297908	0,116348	67,315	26,2899	2,56049	0,015923
površina	0,454431	0,405450	0,227	0,2029	1,12081	0,271566

tabela 54



slika 60



slika 61

Nažalost nijedan od datih modela nije dovoljno dobar za praktičnu upotrebu, prvi zbog nedovoljne korelacije zavisnog i nezavisnih faktora a drugi zbog visokog stepena multikolinearnosti.

Primer 10 : Troškovi održavanja automobila

Pored troškova komunalija u najčešće troškove spadaju i troškovi održavanja automobila. Prilikom kupovine automobila presudni faktor je cena istog koja uključuje samu cenu mašine kao i buduće troškove održavanja. Ovi faktori u svakodnevnom životu bitno zavise od marke automobila međutim ona ne može da se ubaci u standardni model višestruke linearne regresije. Ostali faktori koji bi mogli bitno da utiču su starost, kilometraža, redovnost servisiranja i sl.

Nažalost model (tabela 55) dobijen preko datog uzorka (prilog-tabela 17) ne izdvaja nijedan od faktora starosti automobila, kilometraže, učestalosti servisiranja, broj osoba koje ga voze i njihova primanja kao bitan. $\bar{R}^2$ je 0.03 što pokazuje da se ne mogu predvideti troškovi održavanja automobila na ovaj način.

N=15	Regression Summary for Dependent Variable: godisnji troskovi (auto.st R= ,61338404 R²= ,37623998 Adjusted R²= ,02970664 F(5,9)=1,0857 p<,42962 Std.Error of estimate: 9719,1					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(9)	p-level
Intercept			22614,82	15135,91	1,49412	0,169353
prihod	-0,417044	0,347353	-0,17	0,14	-1,20063	0,260539
starost auta	0,279356	0,428939	390,66	599,84	0,65127	0,531147
predjena kilometraza	-0,309552	0,422524	-0,03	0,04	-0,73263	0,482430
godisnje servisiranje	0,380119	0,297624	5182,10	4057,46	1,27718	0,233502
broj osoba	0,359940	0,285815	3132,07	2487,06	1,25934	0,239585

tabela 55

Na osnovu drugog uzorka (prilog-tabela 18) dobijen je model sa bitnim faktorom kilometraža (tabela 56). $\bar{R}^2$ je 0.88 što pokazuje jaku zavisnost troškova održavanja automobile od cene i kilometraže. Mala je verovatnoća da je došlo do visokog stepena multikolinearnosti među nezavisnim faktorima (korelacija je -0.38, pojedinačni R^2 su 0.14).

N=30	Regression Summary for Dependent Variable: Mesecni troskovi odrzavanja (car.s R= ,94112451 R²= ,88571535 Adjusted R²= ,87724982 F(2,27)=104,63 p<,00000 Std.Error of estimate: 34,775					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(27)	p-level
Intercept			167,1485	28,63685	5,83683	0,000003
cena	-0,086440	0,070394	-0,0012	0,00094	-1,22794	0,230066
predjene milje	0,970736	0,070394	0,0344	0,00250	13,78997	0,000000

tabela 56

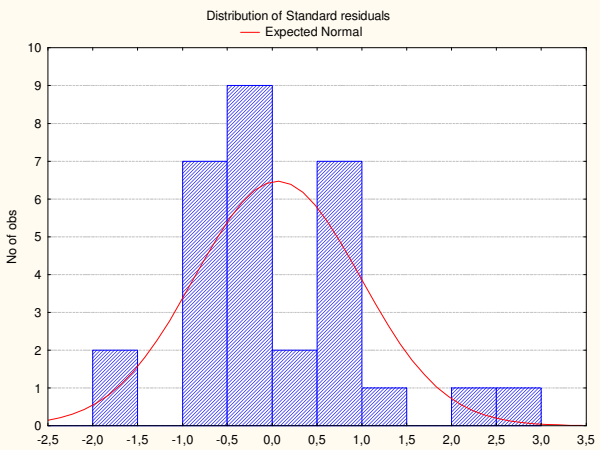
Regresiona jednačina : $Y = 167.15\text{-}0.001X_1 + 0.03X_2$

Outlieri su opažanja 14 i 29. Njihovim izbacivanjem dobija se bolji model ($\bar{R}^2$=0.92)(tabela 57) međutim model nije savršen. Upoređujući normalnost standardizovanih reziduala sa i bez outliera (respektivno slike 62 i 63), normalnost zavisne promenljive sa i bez outliera (respektivno slike 64 i 65), poređenje standardizovanih reziduala sa nezavisnim faktorima sa i bez outliera (respektivno slike 66,67,68 i 69) i poređenje standardizovanih reziduala i predviđenih vrednosti (respektivno slike 70 i 71) ne primeti se bitno poboljšanje modela bez outliera.

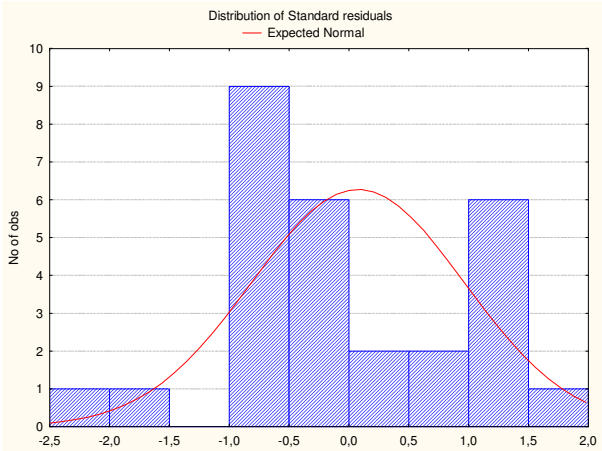
N=28	Regression Summary for Dependent Variable: Mesecni troskovi odrzavanja (car.s R= ,96246855 R²= ,92634571 Adjusted R²= ,92045336 F(2,25)=157,21 p<,00000 Std.Error of estimate: 25,423					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(25)	p-level
Intercept			140,5759	23,98154	5,86184	0,000004
cena	-0,103808	0,058778	-0,0012	0,00069	-1,76611	0,089587
predjene milje	0,997515	0,058778	0,0363	0,00214	16,97093	0,000000

tabela 57

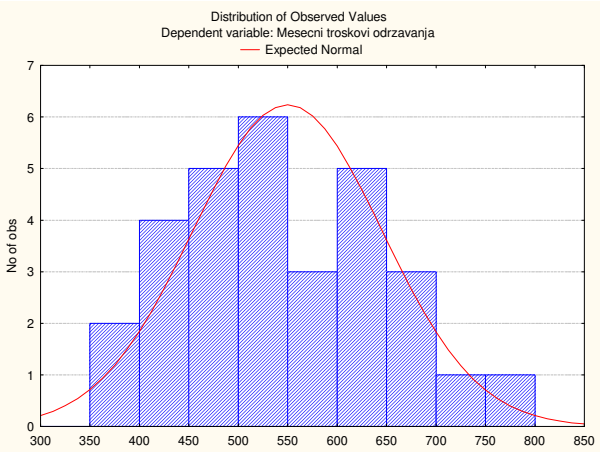
Regresiona jednačina: $Y=140.6\text{-}0.001X_1 + 0.04X_2$



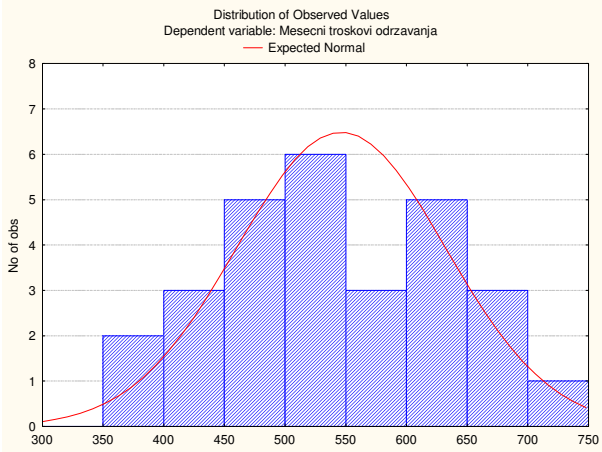
slika 62



slika 63

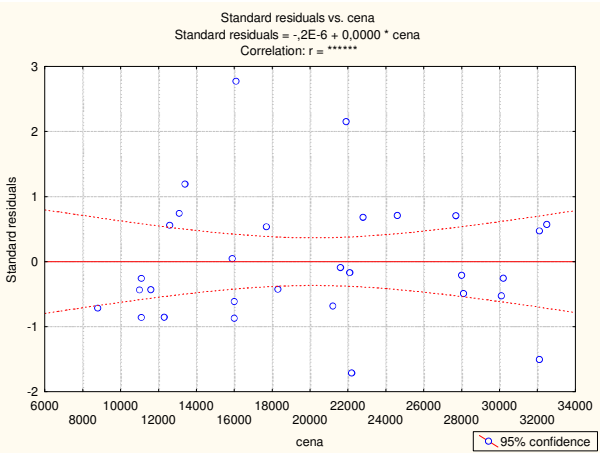


slika 64

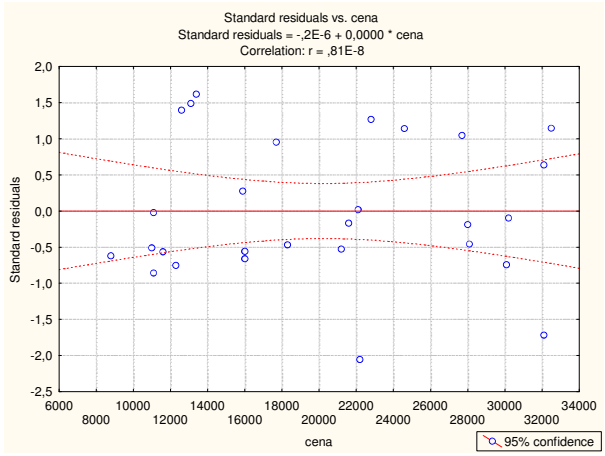


slika 65

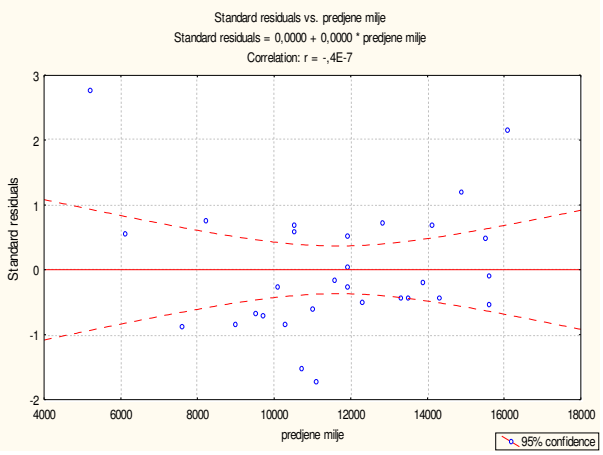
U oba modela se ne vidi bitno odstupanje od homoskedastičnosti.



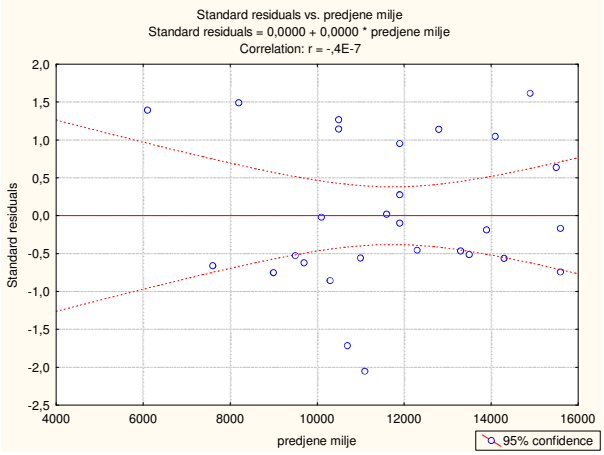
slika 66



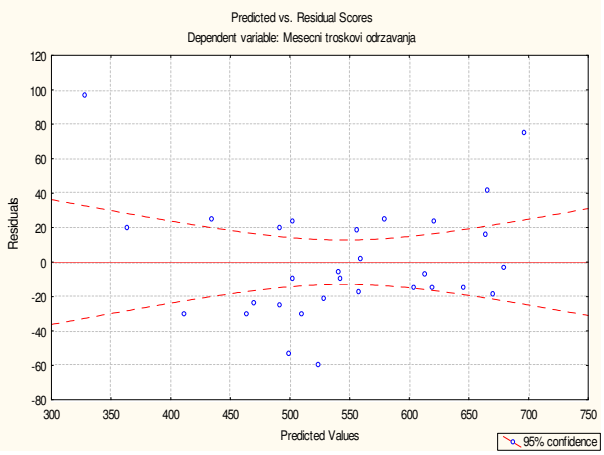
slika 67



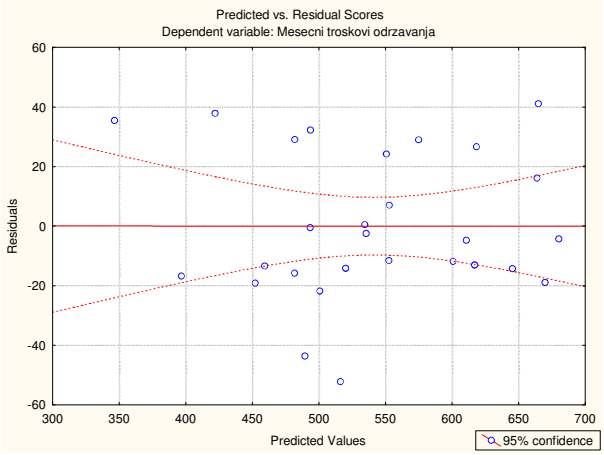
slika 68



slika 69



slika 70



slika 71

Proveravanjem outliera u novom modelu ističe se opažanja broj 11. Ponovnim pravljenjem novog modela bez opažanja broj 11 dobija se model sa većim $\bar{R}^2$ (0.93) ali se opažanje 19 ističe kao outlier. Izbacivanjem ovog opažanja i pravljenjem novog modela dobija se $\bar{R}^2$=0.94. Dobijeni model ima outlier, opažanje broj 3. Izbacivanjem spomenutog opažanja dobija se model sa $\bar{R}^2$=0.95. Dobijeni model nema više outliera. Dobijeni model je:

Regresiona jednačina: $155.2 - 0.00001X_1 + 0.03X_2$

$\bar{R}^2 = 0.95, \sigma = 19.08$

-Standardizovani reziduali nemaju normalnu raspodelu

-Zavisna promenljiva ima približno normalnu raspodelu

-Nije došlo do bitne heteroskedastičnosti

Model se može koristiti za predviđanje mesečnih troškova održavanja automobila.

Primer 11 : Plata

Jedan od glavnih ciljeva rada tj zaposlenja je obezbeđivanje egzistencije. Naknada za rad tj plata bi trebala da bude dovoljna za obezbeđivanje osnovnih potreba kao što su hrana, krov nad glavom i odeća a potom i želja. Normalna je težnja za što većom platom u cilju većeg komfora življenja. Koji su sve bitni faktori za obezbeđivanje što veće plate?

Na osnovu uzorka (prilog-tabela 19) dobijen je model (tabela 58).

N=14	Regression Summary for Dependent Variable: zarada (posao.st R= ,76282105 R²= ,58189595 Adjusted R²= ,45646473 F(3,10)=4,6392 p<,02789 Std.Error of estimate: 11416,					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(10)	p-level
Intercept			52889,6	32886,57	1,60824	0,138861
stepen spreme	-0,553383	0,253281	-10767,9	4928,43	-2,18486	0,053802
iskustvo	0,528141	0,279624	794,1	420,41	1,88876	0,088243
utisak	0,959045	0,282348	14130,9	4160,20	3,39668	0,006809

Tabela 58

Jedini bitan faktor u modelu je lični utisak o poslu. Nažalost ovaj faktor može dvojako da se interpretira: zadovoljstvo zbog obavljanja željenog posla ili zadovoljstvo zbog visine plate koja se dobija na osnovu obavljenog posla. Obično se kao bitni faktori ističu iskustvo i stepen stručne spreme. Zbog visoke korelacije između ta dva faktora dolazi do kontradiktornih rezultata, kao što se desilo i u ovom modelu, da stepen stručne spreme utiče obrnuto srazmerno na platu. Ipak u ovom modelu se ne vidi da je došlo do bitne korelacije između dva prethodno spomenuta faktora (tabela 59).

Variable	Correlations of Regression Coefficients B; DV: zarada (posao.st		
	stepen spreme	iskustvo	utisak
stepen spreme	1,000000	0,277646	-0,307915
iskustvo	0,277646	1,000000	0,507278
utisak	-0,307915	0,507278	1,000000

tabela 59

U model se dodaje još jedan faktor-pol. Nažalost u mnogim zemljama se ovaj faktor ističe kao značajan.[]

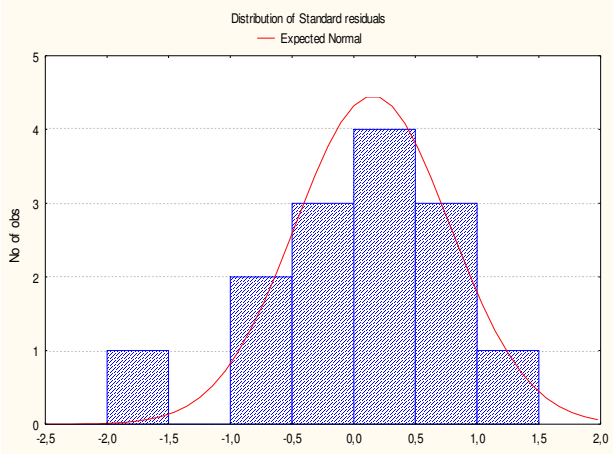
Novi model takođe izdvaja ovaj model kao signifikantan (tabela 60).

N=14	Regression Summary for Dependent Variable: zarada (posao.st R= ,87326149 R²= ,76258563 Adjusted R²= ,65706813 F(4,9)=7,2271 p<,00686 Std.Error of estimate: 9067,7					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(9)	p-level
Intercept			53417,1	26122,91	2,04484	0,071206
stepen spreme	-0,191927	0,244026	-3734,6	4748,35	-0,78650	0,451785
iskustvo	0,273794	0,242439	411,6	364,50	1,12933	0,287958
utisak	0,775634	0,234966	11428,4	3462,06	3,30105	0,009214
Pol	-0,612034	0,233852	-20215,0	7723,94	-2,61719	0,027940

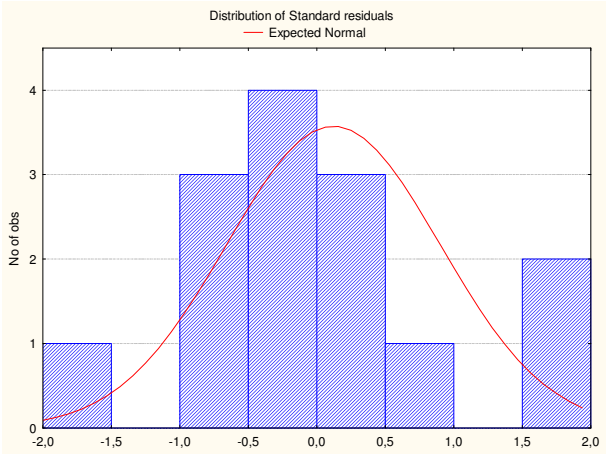
tabela 60

$\bar{R}^2$ ovog modela je znatno viši nego za predhodni (0.66>0.46) takođe ocenjeno standardno odstupanje je manje nego za predhodni (9067.7<11416).

Standardizovani reziduali imaju normalnu raspodelu kod novog modela (slika 72) dok kod starog nemaju (slika 73).

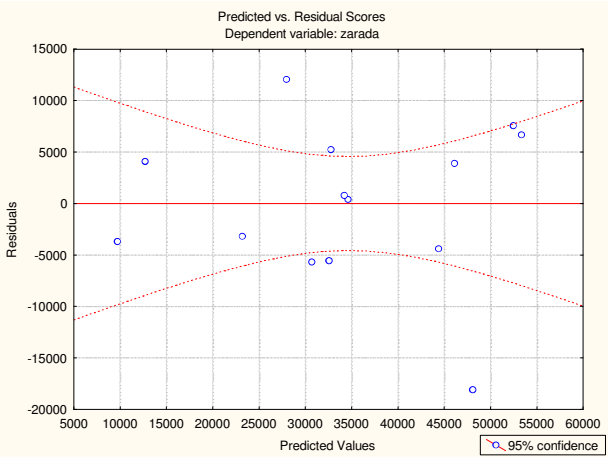


slika 72

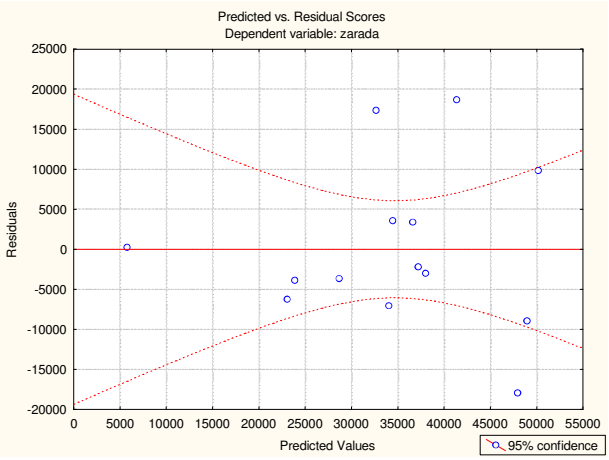


slika 73

Kod novog modela je znatno manja verovatnoća da postoji heteroskedastičnost (slika 74) nego kod predhodnog modela (slika 75).



slika 74



slika 75

Sem navedenih faktora naučnici su došli do još jednog značajnog faktora koji je naveden u sledećem članku:

Rezultati istraživanja su utemeljeni na 30-godišnjem praćenju dečaka u Gvatemali od njihova rođenja. Naime, istraživanje je pokazalo da deca koja su od rođenja uzimala kvalitetnu hranu prosečno zarađuju i do 50% više od onih koji nisu imali zdravu ishranu.Stručnjaci koji su radili na istraživanju smatraju da ovi rezultati mogu uticati na upravljanje zemalja u razvoju, kao i na socijalnu politiku u razvijenim zemljama. Iako na visinu plate utiče nekoliko faktora, od hrane, školovanja, ekonomskih instrumenta i socijalne politike, tokom istraživanja u gvatemalskom selu za varijablu istraživanja je uzeta samo kvalitetna ishrana.Tokom 70-tih, deo dece primao je dodatak u obliku hranjivih namirnica, dok je drugi deo dece dobijao manje hranjive namirnice. Rezultati se razlikuju između dečaka i devojčica.

Naime, nakon što su naučnici posle 30 godina želeli saznati da li je kvalitetnija ishrana uticala na nekadašnje dečake, danas odrasle muškarce, došli su do zaključka da grupa dečaka kojoj su bile dostupne dostatne količine hranjivih materija danas zarađuje i do 50% više novca po satu od ostalih meštana. Međutim, kod djevojčica nije zabeležen takav napredak, ali kao jedan od razloga je navedena manja mogućnost zapošljavanja u tom okruženju. []

Primer 12 : Uspeh tokom obrazovanja

Stepen stručne spreme utiče na visinu plate. Stoga bi bilo korisno odrediti kakav će uspeh dete postići tokom svog obrazovanja. Model dobijen pomoću uzorka (prilog-tabela 20) nije nimalo dobar jer je $\bar{R}^2$ =0.05 (tabela 61).

N=26	Regression Summary for Dependent Variable: uspeh u srednjoj (uspeh.st R= ,40657823 R²= ,16530585 Adjusted R²= ,05148393 F(3,22)=1,4523 p<,25479 Std.Error of estimate: ,35799					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(22)	p-level
Intercept			2,373603	1,174921	2,020224	0,055701
uspeh u osnovnoj	0,410064	0,214411	0,467421	0,244402	1,912513	0,068918
uspeh majke	-0,011320	0,222805	-0,009498	0,186946	-0,050807	0,959938
uspeh oca	0,015396	0,203181	0,009323	0,123034	0,075776	0,940282

tabela 61

Na osnovu drugog uzorka (prilog-tabela 21)dobijen je novi model (tabela 62).

N=20	Regression Summary for Dependent Variable: GPA (GPA 20 R= .92345740 R²= .85277358 Adjusted R²= .81351320 F(4,15)=21.721 p<.00000 Std.Error of estimate: .26851					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(15)	p-level
Intercept			0.161550	0.437532	0.369229	0.717117
SAT-math	0.537305	0.156218	0.002010	0.000584	3.439467	0.003650
SAT-eng	0.267071	0.117630	0.001252	0.000552	2.270444	0.038350
Math	0.259807	0.125992	0.189440	0.091868	2.062090	0.056966
Eng	0.068175	0.137416	0.087564	0.176496	0.496122	0.626999

tabela 62

Model je vrlo dobar, $\bar{R}^2$=0.81 i ispunjene su standardne pretpostavke. Nažalost može se koristiti samo u SAD pošto u Srbiji trenutno ne postoji ovakav sistem školovanja.