Ekonometrija – I deo

Doktorske studije

Predavač: Aleksandra Nojković

Beograd, školska 2024/25

Struktura predavanja

- Klasični višestruki linearni regresioni model-posebne teme:

- Multikolinearnost
  - pojam i posledice
  - metodi otkrivanja i otklanjanja

- Veštačke promenljive

- Narušavanje pretpostavki KLRM: Slučajna greška nema normalnu raspodelu

Pretpostavke KLRM (višestrukog)

1. $E(\varepsilon_i)=0$, *za svako i.*
2. $Var\ (\varepsilon_i)=E(\varepsilon_i^2)=\sigma^2$, *za svako i.*
3. $Cov(\varepsilon_i,\varepsilon_j)=E(\varepsilon_i\varepsilon_j)=0$, *za svako i,j, tako da* $i\neq j$.
4. $E(\varepsilon_i X_i)=0$, *za svako i.*
5. $\varepsilon_i \sim N(0, \sigma^2)$.
6. Ne postoji tačna linearna zavisnost izmedu objašnjavajućih promenljivih (tj. jedna objašnjavajuća promenljiva nije linearna funkcija druge).

Multikolinearnost

- Jedna od pretpostavki KLRM: odsustvo linearne zavisnosti između objašnjavajućih promenljivih.
- U slučaju **perfektne linearne zavisnosti** nije moguće dobiti ocene parametara metodom ONK.
- Ako posmatramo populacionu regresionu jednačinu:

$$E(Y)=\beta_0+\beta_1X_1+\beta_2X_2,$$

koja se ocenjuje na bazi uzorka u kome važi:

$$X_{1i}=l_0 + l_1X_{2i} \text{ za svako i.}$$

- Ocene parametara β_0, β_1 i β_2 nisu jednoznačne (dobija se sistem od dve jednačine sa tri nepoznata parametra, **pokazati...**).

Dvostruki KLRM

- U modelu sa dve objašnjavajuće promenljive:

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

osnovni pokazatelj korelisanosti između objašnjavajućih promenljivih je koeficijent korelacije:

$$r = \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sqrt{\sum_{i=1}^{n} x_{1i}^2 x_{2i}^2}}.$$

- Efekat dve ekstremne vrednosti za *r (0 ili 1)* se jasno uočava iz izraza za ocenu b_1:

$$b_1 = \frac{\sum_{i=1}^{n} x_{1i} y_i \sum_{i=1}^{n} x_{2i}^2 - \sum_{i=1}^{n} x_{1i} x_{2i} \sum_{i=1}^{n} x_{2i} y_i}{\sum_{1=1}^{n} x_{1i}^2 \sum_{1=1}^{n} x_{2i}^2 - \left(\sum_{i=1}^{n} x_{1i} x_{2i}\right)^2} = \frac{r_{yx_1} - r_{y_{x_2}} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2} \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_{1i}^2}}.$$

Dvostruki KLRM: visoka multikolinearnost

- Ako je |r| blizu vrednosti 1, smatra se da je multikolinearnost visoka.

- Ocene ONK se mogu dobiti, ali se dovodi u pitanje njihova preciznost:

$$s_{b_1}^2 = \frac{s^2}{\left(1-r^2\right)\sum_{i=1}^{n} x_{1i}^2}.$$

- Posledice: povećanje standardnih greški ocena, proširenje intervala poverenja (neprecizne ocene), smanjuju se *t*-odonosi (neopravdano prihvatanje H_0).

Visoka multikolinearnost

- U praksi se gotovo nikad ne sreću dva pomenuta ekstrema, **odnosno izvestan stepen korelisanosti** između objašnjavajućih promenljivih **uvek postoji.**

- **Problem** nastaje onda kada je **korelisanost značajno izražena** (nedostatak nezavisnih varijacija promenljivih na desnoj strani jednačine).

- Postoji više razloga za pojavu multikolinearnosti – objasniti!

Neke od činjenica vezane za prisustvo multikolinearnosti

- Ocene ONK **ostaju NLNO**, ali značajnost ocenjenih parametara značajno opada.
- Nema jasnih kriterijuma koji nivo linearne zavisnosti je štetan za preciznost ocena regresionih korficijenata.
- **Pitanje stepena**, a ne postojanja (ne pravi se razlika između prisustva i odsustva multikolinearnosti).
- Isti nivo multikolinearnosti može imati različite efekte na rezultate ocenjivanja, u zavisnosti od opšte valjanosti modela.
- Odnosi se na stanje objašnjavajućih promenljivih, koje se u opštem slučaju smatraju nestohastičkim; ***karakteristika je uzorka, a ne populacije*** (meri se u svakom uzorku od interesa istražača).

Problem multikolinearnosti?

- Nije pitanje „ima ili nema“ multikolinearnosti.
- Greene: Više je pitanje stepena, tj. „crvenila“

- Podaci u uzorku su uvek manje ili više lin. zavisni, pa je pitanje „nijansi“.

Posledice visoke multikolinearnosti

- **Ocene** regresionih parametara mogu biti **neprecizne**, u smislu većih standardnih greški i širih intervala poverenja.
- **Niže vrednosti t-statistika** (pogrešan zaključak o potrebi izostavljanja pojedinih promenljivih iz modela).
- Visoka vrednost F-statistike je praćena niskim vrednostima t-statistika (uticaj regresora se ne može precizno razdvojiti).
- **Ocene vrlo nestabilne**, osetljive na promenu uzorka, moguće je dobiti i **pogrešan znak** regresionog koeficijenta (široki inter. poverenja).
- Ocene su vrlo osetljive na isključivanje pojedinih promenljivih (zbog visokih kovarijansi ocena).

Utvrđivanje postojanja multikolinearnosti

- Nije posledica svojstava osnovnog skupa, tako da **ne postoje formalni testovi** za njeno utvrđivanje (statistički testovi se zasnivaju na hipotezama o određenim vrednostima parametara osnovnog skupa).

1) **Vrednost koeficijenta korelacije**

a) U dvostrukoj regresiji:

- veće vrednosti koeficijenta korelacije r (0,7 ili 0,8)
- korisno je poređenje r^2 i R^2 (izražena je multikolinearnost za r^2 veće od R^2 , odnosno r veće od r_{yx1} i r_{yx2}.

b) U višestrukoj regresiji (r nije pouzdan pokazatelj):

- statistička značajnost pomoćnih regresija jedne objašnjavajuće promenljive na ostale u modelu (Klein-ovo pravilo, kor.$R^2 \geq$ kor.R_j^2).

Utvrđivanje postojanja multikolinearnosti (nastavak)

2) **Faktor rasta varijanse** (FRV; *eng. Variance-Inflation Factor, VIF*).

a) Za dvostruki linearni regresioni model (**pokazati: prirast varijanse** zbog pojave ($1-r^2$) u brojiocu za varijansu ocene **dvostruke u poređenju sa jednostavnom** regresiji):

$$FRV = \frac{1}{1-r^2}.$$

b) Za višestruki regresioni model (slično, prisrast se javlja zbog pojave ($1-R_j^2$) u brojiocu izraza za varijansu ocene višestruke u poređenju sa jednostavnom regresiom):

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

gde je R^2_j koeficijent determinacije u modelu u kome je objašnjavajuća promenljiva X_j regresirana na ostale objašnjavajuće promenljive.

R_j^2 i odgovarajuće vrednosti FRVj

R_j^2	0	0.5	0.8	0.9	0.95	0.975	0.99	0.995	0.999
FRVj	1	2	5	10	20	40	100	200	1000

Tumačenje izračunatih vrednosti za FRV

- Za r=1 (odnosno $R_j^2=1$) vrednost nije moguće odrediti.
- FRV je jednak 1 za objašnjavajuće promenljive koje su ortogonalne (za r=0, odnosno $R_j^2=0$).
- Vrednost FRV je veća za izraženiju multikolinearnost (visoka za vrednosti preko 10).

Šta raditi?

- U otklanjanju visoke mulikolinearnosti treba voditi računa o cilju istraživanja.
- ništa ne preduzimati ako su *t-odnosi* veći od 2.
- ako je **cilj istraživanja previđanje**:važnije je minimizirati *s*, od preciznog ocenjivanja parametara.

Moguća rešenja:

- Povećanje obima uzorka (raste vrednost $\sum x_{1i}^2$).
- Korišćenje spoljnih ocena (opravdanih eksternih ograničenja).
- Transformacija polaznih promenljivih.
- Izostavljanje iz modela one promenljive za koju se sumnja da je glavni uzrok visoke korelacije.
- Metod glavnih komponenata (*engl. Principal Component Analysis*; PCA)

Veštačke promenljive

- Koriste se da opišu uticaj kvantitativno nemerljivih faktora na kretanje izabrane zavisne promenljive
  - U podacima preseka: potrošnja može zavisiti od starosnih, polnih, regionalnih, verskih i drugih razlika.
  - U podacima vremenskih serija: sezonski efekti, efekti intervencija i strukturnog loma.
- Definišu se tako da uzimaju vrednost 1 za jedan modalitet i 0 za drugi modalitet.

Veštačke promenljive (primeri primene)

- Najčešće obuhvataju uticaje neekonomske prirode: kvalitataivne faktore (pol, bračno stanje, zanimanje, članstvo u sindikatu, pripadnost određenoj rasi, religijske i kulturne razlike) ili privremene efekte (promene u institucionalnom i političkom okruženju, ratni periodi, sezonski efekti).

- Međutim, mogu obuhvatati i šire grupe kvantitativnih efekata (dohodak ili godine starosti, kada je dovoljno odabrati nekoliko karakterističnih, širih grupa: npr. potrošači do i preko 35 godina ili oni sa dohotkom do 40000 din, između 40000-60000 din. i preko 60000 din.).

Načini uvođenje u model

- Ispitujemo zavisnost potrošnje datog proizvoda (Y) od dohotka (X_1) prema uzorku koji se sastoji od gradskih i seoskih domaćinstava):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Razlika se može ispoljiti u promeni:

1. vrednosti odsečka – slobodnog člana (β_0)
2. vrednost nagiba – marginalne sklonosti ka potrošnji (β_1) i
3. vrednost odsečka i nagiba (β_0 i β_1).

Promena vrednosti odsečka (β_0)

- Model koji obuhvata regionalne razlike u nivou potrošnje uključuje veštačku promenljivu V, definisanu kao:

$$V = \begin{cases} 0, & \text{za seoska dom.} \\ 1, & \text{za gradska dom.} \end{cases}$$

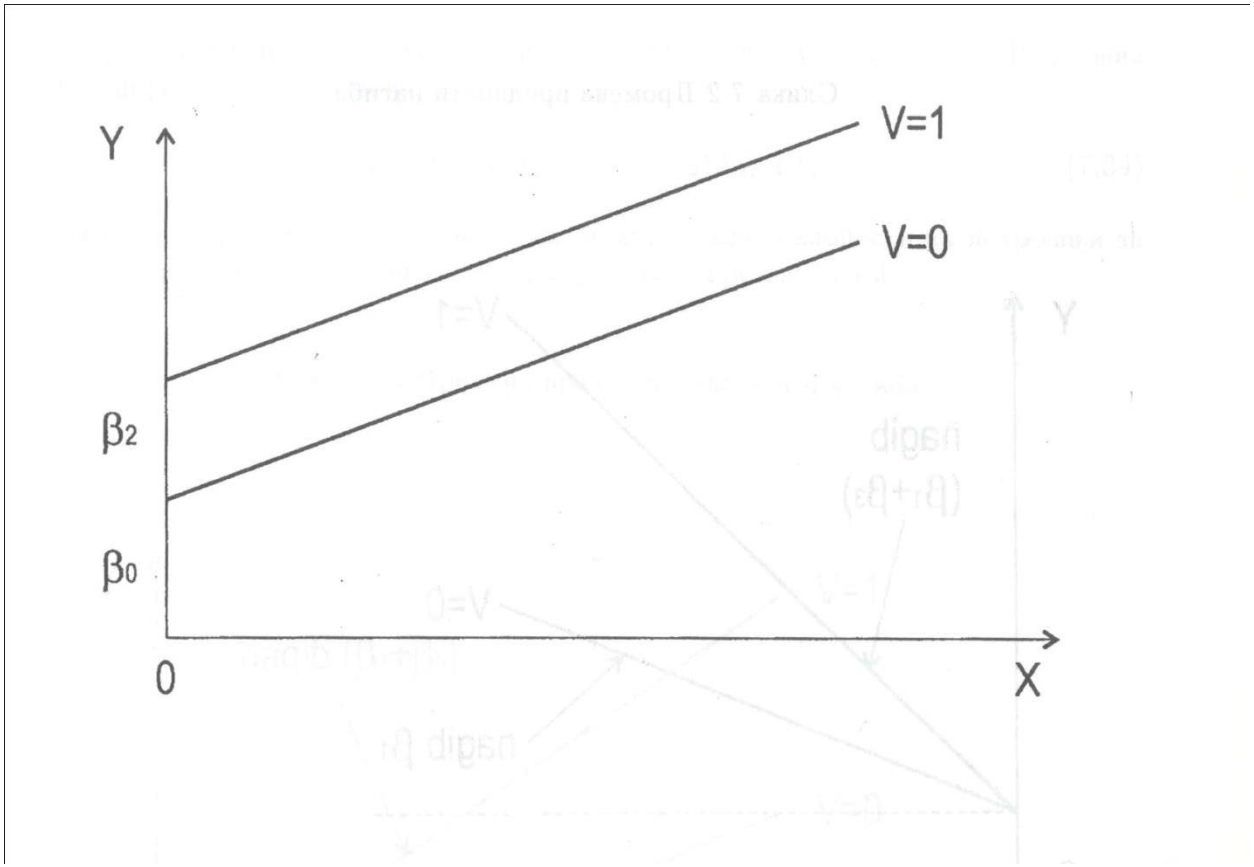
- Polazni model postaje:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 V + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

- Odnosno model postaje:

- za V = 0 (seoska dom.): $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$.
- za V = 1 (gradska dom.): $Y_i = (\beta_0 + \beta_2) + \beta_1 X_{1i} + \varepsilon_i$.

Grafički prikaz promene vrednosti odsečka (β_0)



Promena vrednosti nagiba (β_1)

- Ako pretpostavimo da postoji značajna razlika u marginalnoj slonosti ka potrošnji gradskih i seoskih domaćinstava, veštačka promenljiva se uvodi kao:

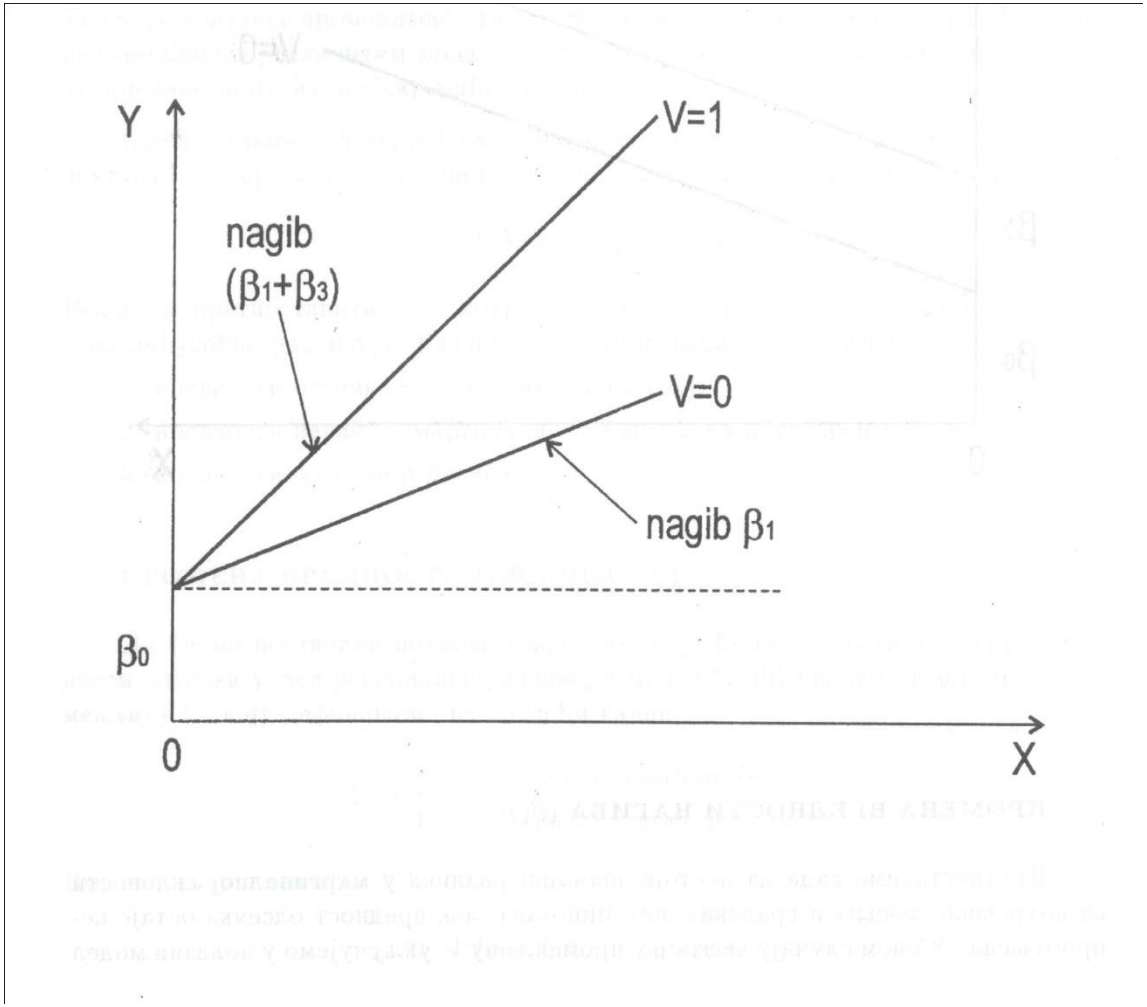
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 V X_{1i} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

- Ovom relacijom obuhvaćena su dva modela:

- za V = 0 (seoska dom.): $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$.

- za V = 1 (gradska dom.): $Y_i = \beta_0 + (\beta_1 + \beta_3) X_{1i} + \varepsilon_i$.

Grafički prikaz promene vrednosti nagiba (β_1)



Promena vrednosti odsečka i nagiba (β_0 i β_1)

- Polaznu funkciju proširujemo sa dve promenljive V i VX_i i model postaje:

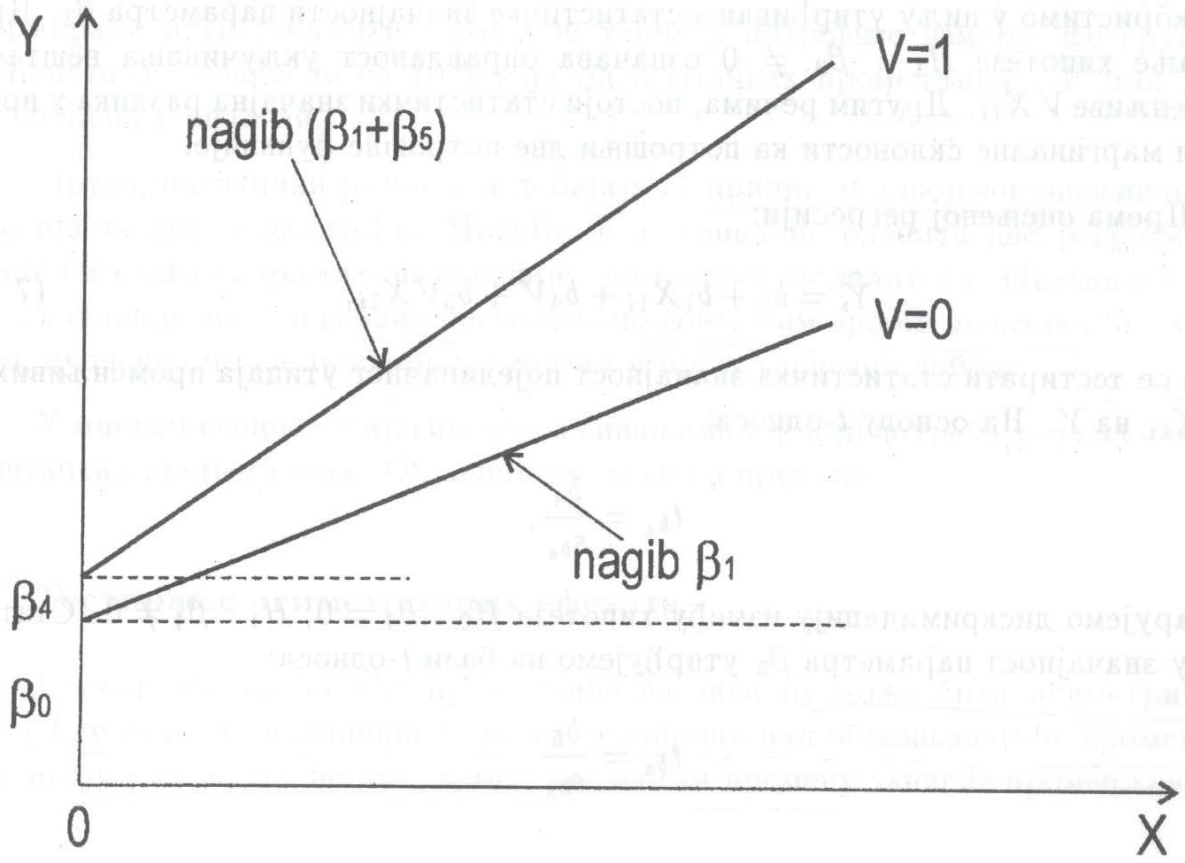
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_4 V + \beta_5 V X_{1i} + \varepsilon_i.$$

- Jednačinu je moguće raščlaniti na dve funkcije:

$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$, za seoska domać.

$Y_i = (\beta_0+\beta_4) + (\beta_1+\beta_5) X_{1i} + \varepsilon_i$, za gradska domać.

Grafički prikaz promene vrednosti odsečka (β_0) i nagiba (β_1)



Interakcija različitih faktora

- Za istarživanje interakcije različitih faktora formiraju se nove veštačke promenljive kao proizvodi već definisanih veštačkih promenljivih.
- Ako pretpostavimo da se ocenjuje uticaj pola i dve kategorije domaćinstava (gradska i seoska) na potrošnju nekog proizvoda, onda model postaje:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 V_1 + \beta_3 V_2 + \beta_4 V_3 + \varepsilon_i,$$

gde su veštačke promenljive definisane kao:

$$V_1 = \begin{cases} 0, & \text{za seoska dom.} \\ 1, & \text{za gradska dom.} \end{cases} \quad \text{i} \quad V_2 = \begin{cases} 0, & \text{za muškarce} \\ 1, & \text{za žene} \end{cases}$$

dok je interakcija $V_3 = V_1 V_2 =$ $\begin{cases} 1, & \text{za žene u gradskim dom.} \\ 0, & \text{za ostale kategorije stan.} \end{cases}$

Pravila ocenjivanja modela sa veštačkim promenljivima

- Dodeljivanje vrednosti 0 i 1 za pojedine modalitete potpuno je proizvoljno i ne menja konačne zaključke.
- Broj veštačkih promenljivih uvedenih u model uvek je za jedan MANJI od broja modeliteta (izbegavamo " zamku veštačke promenljive ").
- Identični rezultati dobijaju se ocenjivanjem dve odvojene regresije, kada raspolažemo dovoljnim brojem podataka.

Testiranje sezonskih efekata

- Izražena sezonska priroda pojedinih ekonomskih promenljivih modelira se uvođenjem sezonskih veštačkih promenljivih.
- Na primer, potrošnja sladoleda *pre capita* (Y) zavisi od realnog dohotka (X_1), relativne cene (X_2) i godišnjeg doba, što predstavljamo modelom:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 S_1 + \beta_4 S_2 + \beta_5 S_3 + \varepsilon_i,$$

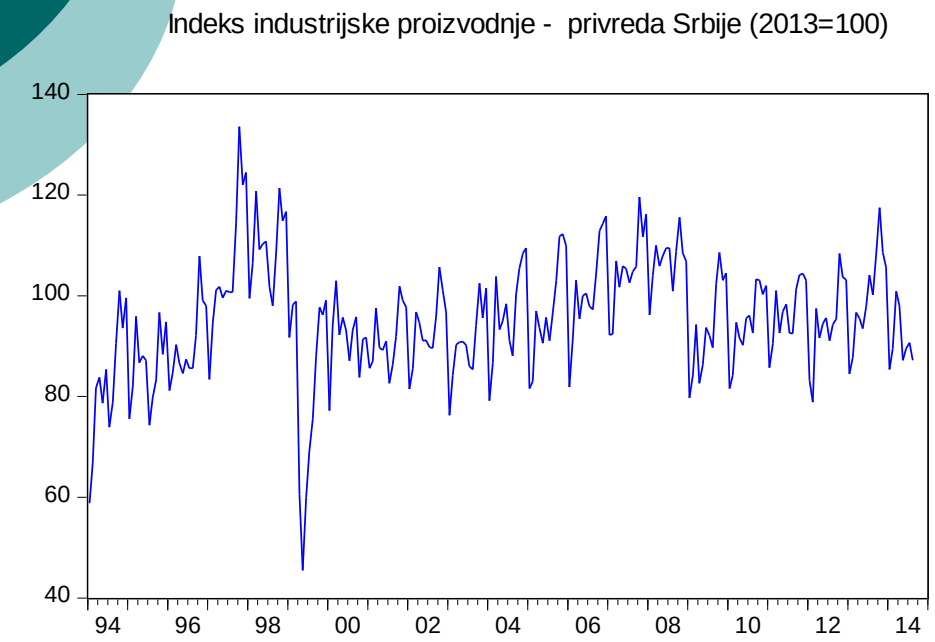
gde smo sa S_1, S_2 i S_3 označili sezonske veštačke promenljive definisane kao:

$$S_i = \begin{cases} 1, & \text{za opservacije } i-\text{tog kvartala } (i = 1, 2 \text{ ili } 3) \\ 0, & \text{u ostalim kvartalima} \end{cases}$$

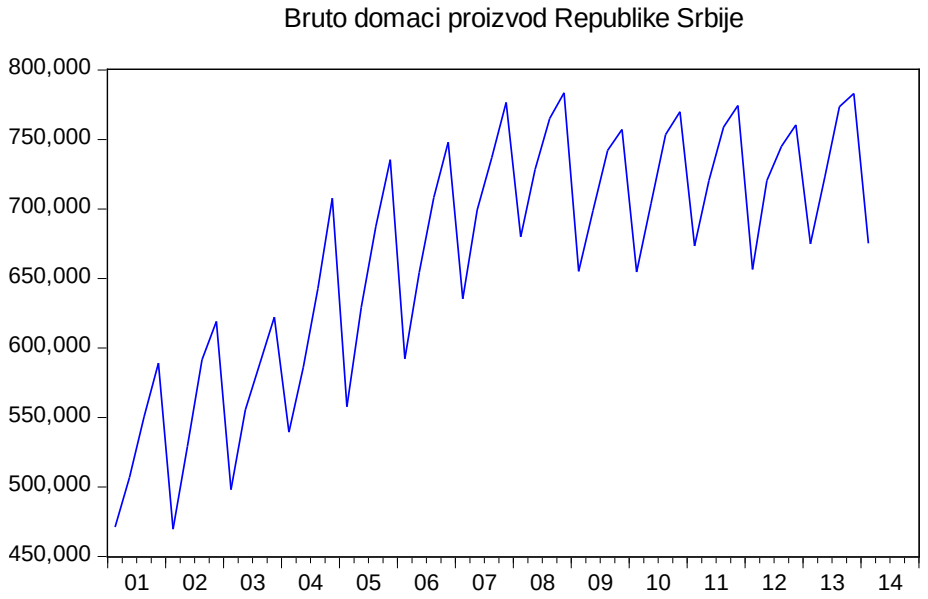
- Dovoljne su TRI veštačke promenljive za obuhvatanje ČETIRI modaliteta !

Sezonski karakter vremenskih serija srpske privrede

Mesečni podaci



Kvartalni podaci



Sta kada je zavisna promenljiva veštačka?

Mikroekonometrijski modeli kvalitativene (diskretne) zavisne promenljive: LMV, probit i logit (nisu predmet razmatranja ovog kursa).

KLRM pretpostavka 5: Slučajna greška ima normalnu raspodelu

- Ukoliko je samo ova pretpostavka narušena primenom metoda ONK se dobijaju najbolje linearne nepristrasne ocene.
- Postupak statističkog zaključivanja je pogrešan
- **Testiranje hipoteza je nepouzdano.**

Kako se proverava pretpostavka da slučajna greška ima normalnu raspodelu?

- Neformalni (grafički) metodi – Analiza histograma
- Formalno testiranja - Žark-Bera (engl. Jarque-Bera) test normalnosti

Koeficijenti kojima se opisuju svojstva raspodela

- Empirijska raspodela se opisuje sa dva koeficijenta: asimetrije i spljoštenosti.

- Koeficijent asimetrije meri stepen u kojem raspodela nije simetricna oko srednje vrednosti (simetricna raspodela, asimetricna u levo ili u desno), α_3:N(0, 6/n).

- Koeficijent spljoštenosti meri debljinu repa raspodele, α_4:N(3, 24/n).

- Kada postoje ekstremni dogadaji tada su repovi teži od repova normalne raspodele

- Veca spljoštenost – repovi su lakši
- Manja spljoštenost – repovi su teži.

JB test statistika

- Način izračunavanja:

$$JB = z_3^2 + z_4^2 = \frac{T}{6}\left[\hat{\alpha}_3{}^2 + \frac{(\hat{\alpha}_4 - 3)^2}{4}\right] : \chi_2^2$$

- Postupak testiranja:

H0: serija ima normalnu raspodelu $(\alpha_3 = 0\, i\, \alpha_4 = 3)$.

H1: serija nije normalno raspodeljena $(\alpha_3 \neq 0\, i / ili\, \alpha_4 \neq 3)$.

- Kritična vrednost na nivou značajnosti 5% je 5.99 (važi asimptotski!)

Šta raditi u slučaju da raspodela odstupa od normalne?

- Ne postoji jedinstveno rešenje.
- Mogu se koristiti metode testiranja koje ne pretpostavljaju normalnost, ali su one izuzetno komplikovane i njihova svojstva nisu poznata.
- Najčešće se modifikuje polazna specifikacija uključivanjem promenljivih kojima će **se eksplicitno modelirati ekstremni događaji.** Takve promenljive se nazivaju **veštačke promenljive**.