Regresijska analiza u mjerenju sezonskih varijacija prometa putnika u zračnim lukama Republike Hrvatske

Josip Arnerić[1], *Danijel Balek*[2]

Sažetak

Cilj ovog rada je primjena modela linearne regresije u mjerenju sezonskih varijacija, tj. u opisivanju periodičnih komponenti vremenskog niza prevezenih putnika u zračnim lukama Republike Hrvatske. Pri tome će se uz trend komponentu u model regresije uključiti i periodične komponente na dva načina: (a) koristeći trigonometrijske funkcije sinus i kosinus i (b) koristeći binarne, odnosno sezonske tzv. dummy varijable. Nakon procjene parametara navedenih modela, izabrat će se reprezentativniji, pomoću kojeg će se prognozirati broj prevezanih putnika u budućem razdoblju. Uzorak podataka je vremenski niz broja prevezenih putnika u RH od siječnja 2004. godine do prosinca 2010. godine (ukupno 84 opažanja).

Uvodni dio

Većina ekonomskih vremenskih nizova podložna je utjecaju sezone, odnosno fluktuacije pojave ponavljaju se na sličan način u razdoblju od godine dana ili kraćem. Sezonske varijacije vremenskog niza značajno utječu na njegovu ukupnu varijancu, pa će prognoze koje zanemaruju sezonsku komponentu, imati veliku prognostičku pogrešku. Zato je potrebno, u slučaju sezonskih varijacija pojave, empirijsku analizu provoditi nad desezoniranim podacima, to jest na temelju vremenskog niza iz kojeg su uklonjeni sezonski utjecaji (govorimo o desezoniranom vremenskom nizu). Metodama desezoniranja identificira se, procjenjuje i uklanja sezonski utjecaj, ali se ne analizira niti se objašnjava njegov uzrok. Naime, postupci desezoniranja pružaju važne informacije za prosudbu gospodarskih kretanja, odnosno za vođenje poslovne i gospodarske politike. Postupkom desezoniranja uklanja se uobičajeno odstupanje od trenda koje se javlja svake godine u istim mjesecima ili kvartalima. Međutim, desezoniranjem se ne uklanjaju ciklička odstupanja. Ciklička odstupanja od trenda su periodična kretanja vremenskog niza s periodom obnavljanja od više godina. Zato se ta komponenta niza teško identificira. U ovom radu se neće detaljno opisivati metode desezoniranja, kao što su pomični prosjeci ili eksponencijalno izglađivanje, već se želi opisati način na koji se periodične komponente, uz trend komponentu, mogu uključiti u model linearne regresije i time postići što točnija prognoza.

Dekompozicija vremenskog niza

Sezonske varijacije ili oscilacije su ono unutargodišnje sistemsko kretanje u vremenskim nizovima koje je najčešće prouzrokovano neekonomskim fenomenima. Temelj analize su modeli koji polaze od dekompozicije niza na trend-cikličku, sezonsku i iregularnu komponentu. U analizi se uobičajeno polazi od (1) aditivnog, (2)

[1] Autor je s Ekonomskog fakulteta u Zagrebu; e-pošta: `jarneric@efzg.hr`

[2] Autor je s Ekonomskog fakulteta u Splitu; e-pošta: `danijel.balek@efst.hr`

multiplikativnog modela i njegova lineariziranog (logaritamskog) oblika ili od (3) pseudoaditivnog modela. Opći oblik tih modela je:

(1) $Y = T + S + e$,

(2) $Y = T \cdot I_s \cdot I_\varepsilon \implies \log Y = \log T + \log I_s + \log I_e$,

(3) $Y = T + T(I_s + I_e - 1)$.

U navedenim izrazima Y predočuje vremenski niz, tj. promatranu pojavu, T je trend-ciklička komponenta, S je sezonska komponenta, dok je e iregularna (slučajna) komponenta. Budući da se ciklička komponenta teško identificira, promatra se zajedno s trend komponentom. Pri tome se utjecaji sezonske komponente i slučajne komponente često izražavaju indeksima (I_S, I_e) zbog jednostavnije interpretacije.

Svaki vremenski niz ne mora sadržavati sve ove komponente. S obzirom na složenost vremenskog niza, nije uvijek moguće strogo razlučiti pojedine komponente vremenskog niza, jer se može dogoditi da jedna komponenta prigušuje drugu.

Trend komponenta pretpostavlja da vrijednosti promatranog vremenskog niza imaju dugoročnu tendenciju rasta ili pada kroz vrijeme. Takvo kretanje se u statističkoj analizi vremenskih nizova može izraziti trend-modelima, koji mogu biti rastući ili padajući i linearni ili nelinearni, ovisno o tendenciji kretanja vremenskog niza.

Sezonska komponenta pretpostavlja sistemsko kretanje pojave unutar jedne godine, koje se onda ponavlja u svakoj narednoj godini. Takvo kretanje je u praksi najčešće uzrokovano neekonomskim faktorima, kao što su, na primjer, klimatske prilike i državni praznici. Sezonska komponenta može biti prisutna kod vremenskih nizova čije su frekvencije, odnosno vrijednosti pojave, mjerene mjesečno, kvartalno, tjedno ili čak dnevno. Na primjer, ako je vremenski niz broj turista na nekom području, on će biti veći u ljetnim i zimskim mjesecima zbog praznika i godišnjih odmora. Ta kretanja imaju sezonski karakter i ponavljaju se u svakoj godini.

Ciklička komponenta pretpostavlja period obnavljanja sistemskog kretanja promatrane pojave, koji je dulji od jedne godine. U praksi je ponekad teško pratiti cikličku komponentu, jer se često ne raspolaže s dovoljno dugim vremenskim nizovima. Ciklička kretanja su u praksi izražena u raznim granama privrede, kao što su građevinarstvo, stočarstvo i slično, a uzroci takvih kretanja su cikličke gospodarske aktivnosti. Ponekad ciklusi kretanja pojave nisu jasno izraženi, već su prigušeni trend komponentom, odnosno dugoročnom tendencijom kretanja.

Slučajna (nesistemska) komponenta predstavlja sve ostale utjecaje na promatranu pojavu, čiji učinak nije sistematski, već se ti utjecaji javljaju s nepredvidivim djelovanjem u nekim vremenskim razdobljima. Na primjer, mogu se javiti zbog nepredvidivosti prirode poslovnih i gospodarskih pojava, zbog iznenadne vremenske nepogode i slično.

U konkretnom slučaju primjene želi se definirati regresijska jednadžba koja uključuje trend i sezonsku komponentu.

Regresijski model s linearnim trendom

Prvo se procjenjuje regresijski model s trend komponentom, odnosno dugoročnom komponentom $Y_t = \hat{\alpha} + \hat{\beta}t + \hat{e}_t$. To znači da je zavisna varijabla u regresijskoj jednadžbi promet u zračnim lukama (varijabla Y_t), a nezavisna varijabla je vrijeme $(t = 1, 2, 3, \ldots, 84)$. Definirana jednadžba je linearni trend model pomoću kojeg se računaju regresijske (očekivane), tj. prognostičke vrijednosti varijable Y_t, nakon što se procijene nepoznati parametri ($\hat{\alpha}$ i $\hat{\beta}$). Razlika između stvarne i regresijske vrijednosti zavisne varijable naziva se rezidual. Oni su vrijednosti $\hat{e}_t = Y_t - \hat{Y}_t$.

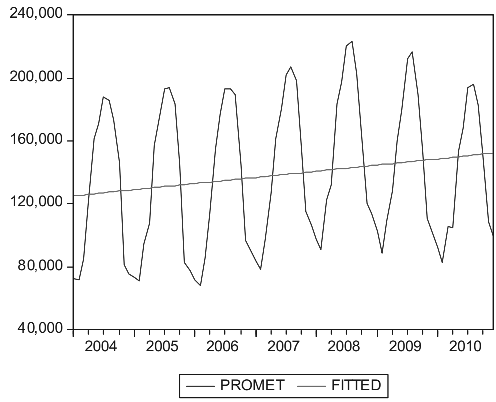
Regresijski koeficijent $\hat{\beta}$ opisuje dugoročnu tendenciju promatrane pojave. U regresijski model se uključuje i konstanta $\hat{\alpha}$ kao aditivni član. Procjene konstantnog člana i koeficijenta smjera u opisanoj linearnoj regresijskoj jednadžbi dobivene su metodom najmanjih kvadrata i to uz programsku podršku EViews verzija 7.0. Dio rezultata, kao i dijagnostički testovi pouzdanosti dobivene regresije prikazani su u tablici 1.

Tablica 1. Rezultati procjena i dijagnostički testovi pouzdanosti linearne regresijske jednadžbe.

Dependent Variable: PROMET
Included observations: 84

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	124600.4	10167.96	12.25422	0.0000
T	327.5856	207.8052	1.576407	0.1188
R-squared	0.029414	Mean dependent var		138522.8
Adjusted R-squared	0.017578	S.D. dependent var		46591.31
S.E. of regression	46180.01			

Na slici 1 uspoređuju se stvarne vrijednosti zavisne varijable PROMET (ravna linija) s regresijskim vrijednostima (FITTED). Regresijske vrijednosti nalaze se na pravcu pozitivnog nagiba što svjedoči o činjenici da postoji dugoročan rastući trend prometa u zračnim lukama. Jednadžba tog pravca (linearne regresijske jednadžbe), prema rezultatima iz tablice 1, se može zapisati analitički kao $\hat{Y}_t = 12\,4600.4 + 327.5856t$. Pri tome je $\hat{\alpha} = 124\,600.4$ i $\hat{\beta} = 327.5856$.



Slika 1. Usporedba stvarnih vrijednosti vremenskog niza s regresijskim vrijednostima.

Koeficijent smjera $\hat{\beta}$ pokazuje da se može očekivati porast prometa prevezenih putnika u zračnim lukama u RH za 327 putnika mjesečno. Ovaj podatak nije statistički značajan s obzirom da je p-vrijednost pripadajuće testne veličine veća od uobičajene razine značajnosti, tj. $0.1188 > 0.05$.

Iako postoji dugoročan rastući trend iz slike 1 je vidljivo da su prisutne i kratkoročne oscilacije oko trenda, odnosno periodična obnavljanja vremenskog niza koja pripisujemo sezonskom utjecaju.

Naime, vrijednost koeficijenta determinacije (R-squared, R^2) iz tablice 1 iznosi 0.029414 što je vrlo mala i nezadovoljavajuća razina protumačenosti odstupanja od prosjeka. To znači da se model dobro ne prilagođava podacima iz uzorka. Zato ga je potrebno proširiti za dodatne članove kojima se opisuje sezonska komponenta vremenskog niza. Komponente kojima se opisuje sezonalnost vremenskog niza u regresijski model se mogu uključiti pomoću trigonometrijskih funkcija sinus i kosinus ili kao dummy odnosno binarne varijable.

Regresijski model s linearnim trendom i sezonskom komponentom

Sezonska se komponenta može uključiti u regresijski model s linearnim trendom koristeći trigonometrijske funkcije sinus i kosinus na sljedeći način:

$$Y_t = \hat{\alpha} + \hat{\beta}t + \sum_{k=1}^{S/2}\left\{\hat{\gamma}_k \cos\left(\frac{2\pi kt}{S}\right) + \hat{\lambda}_k \sin\left(\frac{2\pi kt}{S}\right)\right\} + \hat{e}_t.$$

Kada su podaci mjesečni, tada je $S = 12$, pa je u modelu potrebno procijeniti ukupno 14 parametara $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_6, \hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_6)$. Drugim riječima, osim varijable vrijeme, kao nezavisne, u model se uključuje novih 12 varijabli. Vrijednosti novih nezavisnih varijabli su:

$x_1 = \cos\left(\frac{2\pi t}{12}\right) = \cos\left(\frac{\pi t}{6}\right)$	$z_1 = \sin\left(\frac{2\pi t}{12}\right) = \sin\left(\frac{\pi t}{6}\right)$
$x_2 = \cos\left(\frac{4\pi t}{12}\right) = \cos\left(\frac{\pi t}{3}\right)$	$z_2 = \sin\left(\frac{4\pi t}{12}\right) = \sin\left(\frac{\pi t}{3}\right)$
$x_3 = \cos\left(\frac{6\pi t}{12}\right) = \cos\left(\frac{\pi t}{2}\right)$	$z_3 = \sin\left(\frac{6\pi t}{12}\right) = \sin\left(\frac{\pi t}{2}\right)$
$x_4 = \cos\left(\frac{8\pi t}{12}\right) = \cos\left(\frac{2\pi t}{3}\right)$	$z_4 = \sin\left(\frac{8\pi t}{12}\right) = \sin\left(\frac{2\pi t}{3}\right)$
$x_5 = \cos\left(\frac{10\pi t}{12}\right) = \cos\left(\frac{5\pi t}{6}\right)$	$z_5 = \sin\left(\frac{10\pi t}{12}\right) = \sin\left(\frac{5\pi t}{6}\right)$
$x_6 = \cos\left(\frac{12\pi t}{12}\right) = \cos(\pi t)$	$z_6 = \sin\left(\frac{12\pi t}{12}\right) = \sin(\pi t)$

Budući da je $\sin(\pi t) = 0$, varijabla z_6 se može izostaviti iz modela pa se procjenjuje ukupno 13 parametara, a jednadžba regresije poprima oblik:

$$Y_t = \hat{\alpha} + \hat{\beta}t + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 x_2 + \hat{\gamma}_3 x_3 + \hat{\gamma}_4 x_4 + \hat{\gamma}_5 x_5 + \hat{\gamma}_6 x_6 + \hat{\lambda}_1 z_1 + \hat{\lambda}_2 z_2 + \hat{\lambda}_3 z_3 + \hat{\lambda}_4 z_4 + \hat{\lambda}_5 z_5 + \hat{e}_t.$$

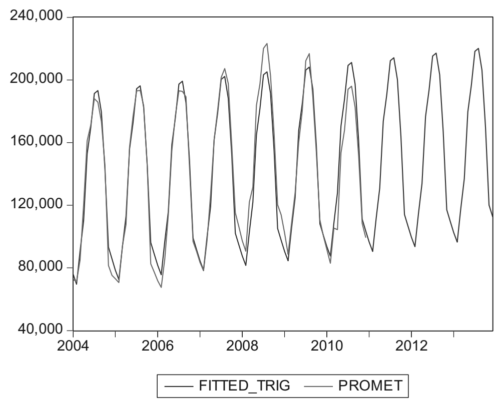
Također, umjesto linearnog trenda, model može uključivati parabolični trend drugog ili višeg stupnja. Moguća je i primjena multiplikativnog (eksponencijalnog) oblika modela. To ovdje nije potrebno. Rezultati procjene parametara prethodno definiranog modela su prikazani u tablici 2.

Dobivena "trigonometrijska regresija" značajno se bolje prilagođava stvarnim podacima kada se uspoređuju stvarne vrijednosti s regresijskim vrijednostima u uzorku ("in the sample") (slika 2). Vrijednost koeficijenta determinacije (R^2) iz tablice 2 je blizu 1 i iznosi 0.961041 što je vrlo velika i zadovoljavajuća razina protumačenosti odstupanja od prosjeka, tj. 96.1% zbroja kvadrata odstupanja prometa prevezenih putnika od prosječnog broja putnika je protumačeno "trigonometrijskom regresijom". To znači da je model reprezentativan i može se koristiti u prognostičke svrhe.

Tablica 2. Rezultati procjene trigonometrijske regresije.

Dependent Variable: PROMET				
Included observations: 84				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	127935.3	2206.286	57.98672	0.0000
T	249.1178	45.20310	5.511078	0.0000
X1	-47094.02	1534.910	-30.68195	0.0000
X2	-2918.892	1534.910	-1.901670	0.0613
X3	841.8822	1534.910	0.548490	0.5851
X4	1628.727	1534.910	1.061123	0.2922
X5	3470.567	1534.910	2.261089	0.0268
X6	-1105.785	1085.110	-1.019054	0.3116
Z1	-41112.16	1543.491	-26.63583	0.0000
Z2	4882.835	1536.240	3.178431	0.0022
Z3	5647.856	1534.910	3.679601	0.0005
Z4	1763.069	1534.466	1.148979	0.2544
Z5	3235.727	1534.292	2.108938	0.0385
R-squared	0.961041	Mean dependent var		138522.8
Adjusted R-squared	0.954456	S.D. dependent var		46591.31
S.E. of regression	9943.037			

Zato su na slici 2 prikazane i prognostičke vrijednosti van uzorka – "out of sample" (tamnija linija).



Slika 2. Usporedba stvarnih vrijednosti vremenskog niza s regresijskim vrijednostima i prognostičke vrijednosti van uzorka.

Sezonska se komponenta može uključiti u regresijski model s linearnim trendom pomoću indikatora, tj. binarnih ili tzv. dummy varijabli aditivnog tipa:

$$Y_t = \hat{\beta} t + \sum_{k=1}^{S} \hat{\varphi}_k d_k + \hat{e}_t.$$

Navedeni model ne sadrži konstantni član, dok za mjesečne podatke ($S = 12$) $d_1, d_2, d_3, \ldots, d_{12}$ su "dummy" ili binarne varijable za istoimene mjesece u godini. Pod pojmom "dummy" varijabla podrazumijevamo jednu umjetno konstruiranu varijablu, koja je rezultat postojanja ili nepostojanja nekog fenomena. Najčešće "dummy" varijabla poprima vrijednost 0 ukoliko taj fenomen ne postoji, a vrijednost 1 ako on postoji. Na taj se način u model mogu uključiti i kvalitativne varijable, npr. spol (M = 0, Ž = 1) ili bračno stanje (neoženjen = 0, oženjen = 1) itd. Dummy varijable mogu pomoći i kod analize sezonskih varijacija. Ako su podatci mjesečni, onda se za siječanj stavi vrijednost za dummy varijablu 1, a za sve ostale mjesece 0. Za veljaču se stavi vijednost 1, a za sve ostale mjesece 0, itd. Na taj se način formira matrica binarnih varijabli D. Dio matrice D se može prikazati tablično za jednu godinu promatranja.

Tablica 3. Vrijednosti sezonskih "dummy" varijabli za jednu godinu promatranja.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
siječanj	1	0	0	0	0	0	0	0	0	0	0	0
veljača	0	1	0	0	0	0	0	0	0	0	0	0
ožujak	0	0	1	0	0	0	0	0	0	0	0	0
travanj	0	0	0	1	0	0	0	0	0	0	0	0
svibanj	0	0	0	0	1	0	0	0	0	0	0	0
lipanj	0	0	0	0	0	1	0	0	0	0	0	0
srpanj	0	0	0	0	0	0	1	0	0	0	0	0
kolovoz	0	0	0	0	0	0	0	1	0	0	0	0
rujan	0	0	0	0	0	0	0	0	1	0	0	0
listopad	0	0	0	0	0	0	0	0	0	1	0	0
studeni	0	0	0	0	0	0	0	0	0	0	1	0
prosinac	0	0	0	0	0	0	0	0	0	0	0	1

Ukoliko se regresijska analiza radi s konstanim članom, onda bi matrica $(\mathrm{D}^\tau \mathrm{D})^{-1}$ bila singularna, te ne bi bilo moguće procijeniti parametre. Da bi se to izbjeglo, potrebno je ispustiti neku od dummy varijabli, npr. siječanj, koji je onda sadržan u konstantnom članu. Na taj način se ujedno izbjegava problem multikolinearnosti koji je tako često prisutan u slučaju više nezavisnih varijabli. S obzirom da se ne preporuča izostaviti konstantni član iz regresijske jednadžbe, jer to može uzrokovati pristranost preostalih parametara koji se procjenjuju, regresijski je model sezonske pojave s linearnim trendom i konstantnim članom, nakon što se izostavi "dummy" varijabla za siječanj, oblika:

$$\begin{aligned} Y_t = \hat{\alpha} + \hat{\beta} t + \hat{\varphi}_2 d_2 + \hat{\varphi}_3 d_3 + \hat{\varphi}_4 d_4 + \hat{\varphi}_5 d_5 + \hat{\varphi}_6 d_6 + \hat{\varphi}_7 d_7 \\ + \hat{\varphi}_8 d_8 + \hat{\varphi}_9 d_9 + \hat{\varphi}_{10} d_{10} + \hat{\varphi}_{11} d_{11} + \hat{\varphi}_{12} d_{12} + \hat{e}_t \end{aligned}$$

Nakon definiranja sezonskih "dummy" varijabli, navedeni model procjenjuje se metodom najmanjih kvadrata. Tom se metodom, kao u prethodnim slučajevima primjene, minimizira zbroj kvadrata rezidualnih odstupanja $\sum_{t=1}^{n} \hat{e}_t^2$, pri čemu je n veličina uzorka (broj opažanja vremenskog niza). Najčešće se izostavlja mjesec siječanj, čiji će utjecaj na promet u zračnim lukama biti sadržan upravo u konstantnom članu. Rezultati procjena tako definiranog regresijskog modela dani su u tablici 4.

Može se primijetiti da je model jednako dobro reprezentativan jer su koeficijent determinacije R^2 (R-squared) i standardna pogreška regresije (S.E. of regression) jednaki kao i u "trigonometrijskoj regresiji". Može se zaključiti da su dva modela ekvivalentna, s jednakim brojem procjenjenih parametara, a jedina je razlika što se češće u praksi koristi regresija s "dummy" varijablama zbog jednostavnosti i lakoće interpretacije regresijskih koeficijenata. Na primjer, regresijski koeficijent uz varijablu

d_6 je $\hat{\varphi}_6 = 92\,878.70$ što znači da je broj prevezenih putnika u lipnju u prosjeku veći nego u siječnju za 92 878 putnika.

Tablica 4. Rezultati procjene regresije sa sezonskim "dummy" varijablama, linearnim trendom i konstantnim članom.

Dependent Variable: PROMET Included observations: 84				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	75442.21	4113.482	18.34023	0.0000
T	249.1178	45.20310	5.511078	0.0000
D2	-6325.832	5314.969	-1.190192	0.2379
D3	14622.19	5315.546	2.750836	0.0075
D4	33577.79	5316.507	6.315762	0.0000
D5	76069.39	5317.852	14.30453	0.0000
D6	92878.70	5319.580	17.45978	0.0000
D7	114161.2	5321.693	21.45204	0.0000
D8	115794.2	5324.188	21.74870	0.0000
D9	101670.8	5327.065	19.08570	0.0000
D10	65083.37	5330.325	12.21002	0.0000
D11	15069.68	5333.965	2.825230	0.0061
D12	7315.561	5337.986	1.370472	0.1749
R-squared	0.961041	Mean dependent var		138522.8
Adjusted R-squared	0.954456	S.D. dependent var		46591.31
S.E. of regression	9943.037			

Također je važno napomenuti da će značajnost pojedinačnih "dummy" varijabli u modelu ovisiti o tome koja je "dummy" varijabla izostavljena.

U tablici 5 su dane prognostičke vrijednosti prevezenih putnika u zračnim lukama za tri godine od posljednjeg mjeseca opažanja, tj. za 2011., 2012. i 2013. godinu.

Tablica 5. Prognostičke vrijednosti prevezenih putnika u zračnim lukama.

godina	siječanj	veljača	ožujak	travanj	svibanj	lipanj
2011	96617.23	90540.51	111737.7	130942.4	173683.1	190741.5
2012	99606.64	93529.93	114727.1	133931.8	176672.5	193730.9
2013	102596.1	96519.34	117716.5	136921.2	179661.9	196720.3

	srpanj	kolovoz	rujan	listopad	studeni	prosinac
	212273.1	214155.2	200280.9	163942.7	114178.1	106673.1
	215262.5	217144.6	203270.4	166932.1	117167.5	109662.5
	218251.9	220134.1	206259.8	169921.5	120156.9	112651.9

Literatura

[1] E. Box, M. Jenkins, C. Reinsel, *Time series analysis, forecasting and control*, 2008.

[2] A. Rozga, *Statistika za ekonomiste*, Ekonomski fakultet Split, 2003.

[3] A. Rozga, B. Grčić, *Poslovna statistika*, Split 2009.

[4] *Statistički ljetopis*, Državni zavod za statistiku, 2012.

[5] I. Šošić, *Primijenjena statistika*, Školska knjiga, Zagreb, 2004.

[6] `www.dzs.hr`