Analiza COVID-19 podataka analizom panel podataka

Gojević, Mia

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Mia Gojević

ANALIZA COVID-19 PODATAKA ANALIZOM PANEL PODATAKA

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, srpanj 2022.

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

1. ____________________________, predsjednik
2. ____________________________, član
3. ____________________________, član

Povjerenstvo je rad ocijenilo ocjenom _______________.

Potpisi članova povjerenstva:

1. ____________________________
2. ____________________________
3. ____________________________

Obitelji, Dini i prijateljima koji su sa mnom prolazili sve najbolje i najgore trenutke.

Sadržaj

Uvod

U većini znanstvenih radova postoji potreba za analizom neke baze podataka. Baze mogu biti raznolike, a u ovom radu će biti obrađena jedna od njih imena panel podaci. Naime, takva konstrukcija podataka je kombinacija vremenskih nizova i vremenskih presjeka. Zbog toga je vrlo popularna jer je njenom analizom moguće puno bolje objasniti dane podatke, nego samo vremenskim nizovima ili vremenskim presjecima. Uz to, moguće je napraviti preciznije matematičke modele koji mogu uključivati individualne i vremenske efekte. U svakoj analizi može doći do problema što bi se za panel podatke moglo dogoditi ukoliko baza nije kvalitetno pripremljena pa se narušava konzistentnost podataka na način da sve obzervacije nisu jednakovrijedne prilikom analize tj. u pravilnom mjerilu ili da neke vrijednosti nedostaju.
Za panel podatke moguće je konstruirati linearne modele poput modela konstantnih koeficijenata, modela fiksnih efekata i modela slučajnih efekata. Za svakog od njih bit će predstavljena teorijska podloga i konstrukcija u ovom radu. Budući da se radi o linearnim modelima, za bolje razumijevanje konstrukcije početno će biti predstavljen model jednostavne linearne regresije. Svi panel linearni modeli nisu jednako kvalitetni pri procjeni parametara prema tome bit će provedeni usporedni testovi modela F-test, Hausmanov test te Lagrangeov multiplikacijski test metodom Breusch-Pagan. Usporedbom njihovih ishoda bit će odlučeno koji linearni panel model najbolje opisuje panel podatke. U praksi, analiza panel podataka bit će provedena nad bazom o bolesti COVID-19 koja je proglašena epidemijom, a prisutna je u svijetu od 2019. godine do sadašnjeg trenutka. Modeli će biti konstruirani na način da je smrtnost uvjetovana primjerice brojem zaraženih osoba ili pak stupnjem kvalitete života u europskim državama Hrvatskoj, Italiji, Švicarskoj i Švedskoj.

Poglavlje 1

Analiza panel podataka

1.1 Panel podaci

Panel podaci čine kombinaciju vremenskih nizova i vremenskih presjeka [1]. Vremenski nizovi odnose se na analizu jednog objekta u vremenskim periodima t, dok se vremenski presjeci odnose na analizu više objekata u istom vremenskom periodu t.

S obzirom na količinu i izgled zadanih podataka panel podatke možemo kategorizirati na sljedeći način [12]:

1. Kratak ili dugačak panel
   - Kratak – velik broj objekata, vrlo mali broj vremenskih perioda
   - Dugačak – vrlo mali broj objekata, velik broj vremenskih perioda
2. Balansiran ili nebalansiran panel
   - Balansiran – svi objekti su promatrani u jednakom broju istih vremenskih perioda t te nema nedostajućih vrijednosti u bazi
   - Nebalansiran – svi objekti nisu promatrani u jednakom broju istih vremenskih perioda t te može nedostajati vrijednost u bazi
3. Fiksirani ili rotirajući panel
   - Fiksirani – u svakom vremenskom periodu t promatrani su isti objekti, nema promjene tijekom analize u smislu dodavanja ili uklanjanja objekata
   - Rotirajući – u svakom vremenskom periodu t nisu promatrani nužno isti objekti, može doći do promjene tijekom analize u smislu dodavanja ili uklanjanja objekata

Sljedećom tablicom 1.1 prikazan je primjer jednog dugačkog, balansiranog, fiksiranog panela.

Tablica 1.1: Primjer dijela panela iz 2. poglavlja

	Hrvatska										
vremenski period	5	6	7	8	9	10	11	12	13	14	15
broj slucajeva zaraze (ppm)	39	130	565	1236	1514	7960	18852	19476	5003	2561	6999
broj smrtnih slucajeva (ppm)	7	1	9	10	23	64	300	504	261	116	100
broj hospitaliziranih osoba (ppm)	33	5	30	35	71	138	425	667	453	237	228
	Švicarska										
vremenski period	5	6	7	8	9	10	11	12	13	14	15
broj slucajeva zaraze (ppm)	133	97	388	776	1249	11522	19829	13932	7920	3423	5006
broj smrtnih slucajeva (ppm)	206	1	2	2	7	53	300	311	181	53	31
broj hospitaliziranih osoba (ppm)	34	10	11	14	16	80	384	316	235	125	99
	Švedska										
vremenski period	5	6	7	8	9	10	11	12	13	14	15
broj slucajeva zaraze (ppm)	1627	2816	833	689	821	3026	11690	17384	12754	8893	14525
broj smrtnih slucajeva (ppm)	171	92	37	8	8	4	73	201	282	122	63
broj hospitaliziranih osoba (ppm)	188	127	45	20	14	27	139	244	245	138	156
	Italija										
vremenski period	5	6	7	8	9	10	11	12	13	14	15
broj slucajeva zaraze (ppm)	424	122	112	354	740	5997	14780	8055	7018	6035	10710
broj smrtnih slucajeva (ppm)	86	21	6	6	7	45	277	295	230	147	189
broj hospitaliziranih osoba (ppm)	205	59	14	16	40	143	549	499	412	347	467

*ppm - person per million

Ukoliko se analiza može svesti na problem jedne zavisne i jedne ili više nezavisnih varijabli radi se pogodnoj situaciji za analizu podataka metodom regresije. Regresijski model je linearan ako je matematička funkcija koja opisuje model svedena na operaciju zbrajanja između nezavisnih varijabli tj. ako je zavisna varijabla svedena na linearnu kombinaciju nezavisnih varijabli. Ukoliko je takvim modelom uspostavljen odnos između jedne zavisne i jedne nezavisne varijable metoda analize naziva se jednostavna linearna regresija, a ukoliko je zavisnoj varijabli modelom pridruženo više nezavisnih varijabli tada se radi o višestrukoj ili multivarijatnoj linearnoj regresiji.

1.2 Jednostavna linearna regresija

Jednostavnom linearnom regresijom bit će objašnjena osnovna struktura i princip analize regresijskog modela. Pri korištenju ovog modela ne uzima se u obzir više različitih vremenskih vremenski perioda t, već je konstruiran model za točno jedan specifičan vremenski period t. [13]

Neka su zadane slučajne varijable X i Y. Uređenim parovima $(x_1, y_1), ..., (x_n, y_n)$ dane su realizacije tih varijabli koje možemo grafički prikazati u Kartezijevom koordinatnom sustavu. Uz pretpostavku da se radi o linearnoj vezi između varijabli, konstruiran je opći oblik modela jednostavne linearne regresije sljedećom jednadžbom

$$y = \alpha + \beta x \tag{1.1}$$

gdje su α i β nepoznati parametri koje želimo procijeniti.
Da bi model bio ispravno konstruiran potrebno je zadovoljiti sljedeće pretpostavke:

- Slučajne varijable $X_1, ..., X_n$ imaju realizacije $x_1, ..., x_n$ koje ne sadrže pogreške te su precizno određene.
- Slučajne varijable $Y_1, ..., Y_n$ imaju realizacije $y_1, ..., y_n$ koje sadrže međusobno nezavisne pogreške $e_1, ..., e_n$. Uz to, $Y_1, ..., Y_n$ su međusobno nezavisne s očekivanjem $E(Y_i) = E(\alpha + \beta X_i) = \alpha + \beta E(X_i)$ za svaki $i = 1, ..., n$.

Ukoliko su pretpostavke usvojene, model se dograđuje uvodeći slučajne pogreške pa slijedi

$$y_i = \alpha + \beta x_i + e_i \qquad i = 1, ..., n, \tag{1.2}$$

a moguće je koristiti i model

$$y_i = \alpha + \beta(x_i - \bar{x}) + e_i \qquad i = 1, ..., n \tag{1.3}$$

Metoda najmanjih kvadrata

Osim zapisa jednadžbom, moguć je i grafički prikaz odnosa promatranih varijabli. Glavni cilj metode najmanjih kvadrata jest originalnim vrijednostima ucrtanim u dijagram rasipanja $(x_1, y_1), ..., (x_n, y_n)$, koje su dalje zvane točkama, pridružiti pravac naziva regresijski pravac ili pravac najmanjih kvadrata.
Među točkama je moguće ucrtati beskonačno mnogo pravaca, a regresijski pravac će odgovarati upravo pravcu kojim su minimizirane udaljenosti svih upisanih točaka. Slučajna pogreška e_i za svaki $i = 1, ..., n$ predstavlja udaljenost točke (x_i, y_i) za svaki $i = 1, ..., n$ od pravca regresije. Najbolje procjene parametara α i β određuju se iz uvjeta minimizacije tj.

$$min \rightarrow SSE = \sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[(y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})]^2 \tag{1.4}$$

gdje se SSE naziva suma kvadrata pogreške ili neobjašnjeni dio varijance (eng. sum of squared errors), y_i predstavlja originalnu vrijednost, a $\hat{y}_i$ procijenjenu vrijednost regresijskom jednadžbom. Iz nužnog i dovoljnog uvjeta za postojanje ekstrema funkcije SSE=SSE($\hat{\alpha}$, $\hat{\beta}$)

$$\frac{\partial SSE}{\partial \alpha} = 0 \qquad \frac{\partial SSE}{\partial \beta} = 0$$

slijede vrijednosti parametara $\hat{\alpha}$ i $\hat{\beta}$.
Uz oznake

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \qquad \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

vrijedi

$$\hat{\alpha} = \bar{y}$$

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

odakle slijedi regresijski pravac koji je dan jednadžbom

$$\hat{y} = \hat{\alpha} + \hat{\beta}(x - \bar{x})$$

odnosno

$$\hat{y} = \hat{\beta}x + (\hat{\alpha} - \hat{\beta}\bar{x}) \tag{1.5}$$

te prolazi točkom $(\bar{x}, \bar{y})$.

Analiza regresijskog modela

Osnovni statistički pokazatelji i metode koje služe za analizu regresijskih modela su rezidualna odstupanja, koeficijent determinacije, koeficijent korelacije, standardna greška regresije, analiza varijance, testiranje značajnosti regresijskih koeficijenata, intervali pouzdanosti regresijskih koeficijenata, intervali pouzdanosti prognoziranih vrijednosti i testiranje razine značajnosti koeficijenata korelacije. U radu je stavljena koncentracija na samo neke od njih, a to su rezidualna odstupanja i koeficijent determinacije.

Rezidualna odstupanja tj. slučajne pogreške, definiraju se sljedećom jednadžbom:

$$e_i = y_i - \hat{y}_i \qquad i = 1, ..., n. \tag{1.6}$$

U modelima linearne regresije polazi se od pretpostavke da su rezidualna odstupanja normalno distribuirana s očekivanjem $\mu = 0$ i konstantnom varijancom. To svojstvo je kraće zapisano oznakom $e_i \sim N(0, \sigma^2)$. Ukoliko svi podaci imaju konstantnu varijancu kaže se da posjeduju svojstvo homoskedastičnosti, no ako za neke podatke to ne vrijedi kaže se da podaci imaju svojstvo heteroskedastičnosti.

Koeficijent determinacije (oznaka R^2) je vrijednost unutar intervala [0,1] koja tumači koliko je dobro ili loše regresijski model prilagođen originalnim podacima. Što je ta vrijednost bliža 1 kaže se da je model regresije reprezentativniji, a počinje se smatrati dobrim ako postiže vrijednost barem 0.5. Koeficijent R^2 računa se primjenom formule

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2} = \frac{SSR}{SS_{yy}}$$

gdje SSR označava objašnjeni dio varijance tj. predstavlja varijacije koje su objašnjene regresijskim modelom, a SS_{yy} označava ukupne varijacije te pritom vrijedi uvjet

$$SS_{yy} = SSE + SSR$$

1.3 Linearni modeli panel podataka

Za razliku od osnovnog regresijskog modela koji je baziran na samo jednom vremenskom periodu, za definiciju linearnog modela za panel podatke potrebno je linearnom regresijskom modelu pridružiti dimenziju koja će se odnositi na više različitih vremenskih perioda.

Najopćenitiji oblik linearnog panel modela definiran je zapisom

$$Y_{it} = \alpha_{it} + \beta_{1,it}X_{1,it} + ... + \beta_{K,it}X_{K,it} + \epsilon_{it} \qquad i = 1, ..., N,\ t = 1, ..., T \tag{1.7}$$

što je u matričnom zapisu

$$Y_{it} = \alpha_{it} + \beta_{it}X_{i_t}^T + \epsilon_{it} \qquad i = 1, ..., N,\ t = 1, ..., T \tag{1.8}$$

gdje je N broj promatranih jedinica, a T ukupan broj vremenskih perioda.
Za jedinku i u vremenskom periodu t vrijedi

- α_{it} je slobodan član
- β_{it} je vektor nepoznatih regresijskih parametara dimenzije K
- X_{it} vektor nezavisnih slučajnih varijabli dimenzije K
- Y_{it} je zavisna slučajna varijabla
- ϵ_{it} je rezidual s očekivanjem 0 i varijancom σ_ϵ^2

Ovakav zapis modela vrlo je generalan pa je potrebno uvesti restikcije na model tako da regresijski parametri budu što bolje procijenjeni kako bi se konstruirali modeli primjenjivi u praksi. Neki od modela su združeni model, model fiksnih efekata i model slučajnih efekata.

Združeni model (*Pooled OLS model*)

Drugi hrvatski naziv za združeni model glasi model konstatnih koeficijenta. Upravo iz tog naziva slijedi glavna ideja modela. Naime, pretpostavka modela je da su svi regresijski parametri konstantni jer se zanemaruje činjenica da se radi o panel podacima. Cijeli set podataka uzima se u analizu kao uzorak od NT podataka i modeliranje se dalje svodi na multivarijatnu linearnu regresiju. U skladu s time model je definiran jednadžbom

$$Y_{it} = \alpha + \beta_1 X_{1,it} + ... + \beta_K X_{K,it} + \epsilon_{it} \qquad i = 1, ..., N, \ t = 1, ..., T \tag{1.9}$$

što se matrično zapisuje

$$Y_{it} = \alpha + \beta X_{it}^T + \epsilon_{it} \qquad i = 1, ..., N, \ t = 1, ..., T \tag{1.10}$$

gdje je N broj promatranih jedinica, a T ukupan broj vremenskih perioda. Kao i u općenitom panel modelu, za jedinku i u vremenskom periodu t vrijedi

- α je fiksan slobodan član
- β je fiksan vektor nepoznatih regresijskih parametara dimenzije K
- X_{it} vektor nezavisnih slučajnih varijabli dimenzije K
- Y_{it} je zavisna slučajna varijabla
- e_{it} je rezidual s očekivanjem 0 i varijancom σ_ϵ^2

Upravo činjenica da su α i β fiksni čini restrikciju od osnovnog linearnog panel modela 1.8. Kako bi združeni model bio primjenjiv naredne pretpostavke moraju biti zadovoljene:

1. Nema heterogenosti između jedinica promatranja $\alpha_i = 0$.
2. $\mathbb{E}(\epsilon_{it}) = 0$, $Var(\epsilon_{it}) = \sigma_\epsilon^2$
3. Stroga egzogenost - rezidual nije koreliran s nezavisnim varijablama u istom vremenskom periodu.

$$\mathbb{E}[X_{it}\epsilon_{it}] = 0 \qquad \forall i = 1, ..., N, \qquad \forall t = 1, ..., T$$

4. Stroga homoskedastičnost

$$\mathbb{E}[\epsilon_{it}^2] = \sigma^2 \qquad \forall i = 1, ..., N, \qquad \forall t = 1, ..., T$$

5. Slučajna greška nije autokorelirana.

$$\mathbb{E}[\epsilon_{it}\epsilon_{js}] = 0 \qquad i \neq j,\ s \neq t \qquad \forall i, j = 1, ..., N, \qquad \forall s, t = 1, ..., T$$

6. Nema multikolinearnosti - matrica X je punog ranga.

$$r(X) = K + 1 < NT$$

Jednadžbom 1.11 definiran je općeniti OLS procjenitelj

$$\hat{\beta}_{OLS} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i^T X_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} X_i^T Y_i\right) = (X^T X)^{-1} X^T Y \tag{1.11}$$

gdje je $X = (X_1^T, ..., X_N^T)$ matrica dimenzije $NT \times K$ i $Y = (Y_1^T, ..., Y_N^T)^T$ vektor dimenzije NT. Ako je

$$\sum_{i=1}^{N} X_i^T X_i = \sum_{i=1}^{N}\sum_{i=1}^{T} X_{it}^T X_{it} \qquad \sum_{i=1}^{N} X_i^T Y_i = \sum_{i=1}^{N}\sum_{i=1}^{T} X_{it}^T Y_{it}$$

tada je jednadžbom 1.12 definiran najbolji združeni procjenitelj

$$\hat{\beta}_{združeni} = \left(\sum_{i=1}^{N}\sum_{i=1}^{T} X_{it}^T X_{it}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{i=1}^{T} X_{it}^T Y_{it}\right) \tag{1.12}$$

Model fiksnih efekata (*Fixed effect (FE) model*)

Za razliku od združenog modela, FE model uzima u obzir slobodni član iz jednadžbe 1.8 kao individualne efekte α_i koji nisu fiksni za cijeli uzorak već samo za i-ti odabir, dok vektor β ostaje fiksan u cijelom modelu [11]. Takav model dan je jednadžbom

$$Y_{it} = \alpha_{it} + \beta_1 X_{1,it} + ... + \beta_K X_{K,it} + \epsilon_{it} \qquad i = 1, ..., N,\ t = 1, ..., T \tag{1.13}$$

tj. matrično zapisano

$$Y_{it} = \alpha_i + \beta X_{it}^T + \epsilon_{it} \qquad i = 1, ..., N,\ t = 1, ..., T \tag{1.14}$$

Ukoliko osim individualnih efekata postoji potreba za proučavanjem i vremenskih efekata model 1.14 se proširuje u oblik

$$Y_{it} = \alpha_i + \gamma_t + \beta X_{it}^T + \epsilon_{it} \qquad i = 1, ..., N,\ t = 1, ..., T \tag{1.15}$$

gdje za jedinku i u vremenskom periodu t vrijedi

- α_i je slučajna varijabla koja opisuje individualni efekt i-te jedinke
- γ_t je slučajna varijabla koja opisuje vremenski efekt i-te jedinke
- β je fiksan vektor nepoznatih regresijskih parametara dimenzije K
- X_{it} vektor nezavisnih slučajnih varijabli dimenzije K
- Y_{it} je zavisna slučajna varijabla
- e_{it} je normalno distribuiran rezidual s očekivanjem 0 i varijancom σ_ϵ^2

Da bi model bio ispravno primijenjen potrebno je početno zadovoljiti pretpostavke koje su većinom slične pretpostavkama združenog modela 1.3:

1. $\mathbb{E}(\epsilon_{it}) = 0$, $Var(\epsilon_{it}) = \sigma_\epsilon^2$
2. Reziduali nisu međusobno korelirani.

$$Cov(\epsilon_{it}, \epsilon_{js}) = \mathbb{E}[\epsilon_{it}\epsilon_{js}] = 0 \qquad i \neq j,\ s \neq t \qquad \forall i, j = 1, ..., N, \qquad \forall s, t = 1, ..., T$$

3. Reziduali i nezavisne varijable nisu korelirane u istom vremenskom periodu.

$$Cov(X_{it}, \epsilon_{it}) = \mathbb{E}[X_{it}\epsilon_{it}] = 0 \qquad \forall i = 1, ..., N, \qquad \forall t = 1, ..., T$$

4. Individualni efekti i nezavisne varijable su korelirane.

$$Cov(X_{it}, \alpha_i) = \mathbb{E}[X_{it}\alpha_i] \neq 0 \qquad \forall i = 1, ..., N, \qquad \forall t = 1, ..., T$$

5. Matrica X je punog ranga.

$$r(X) = K$$

Ovakav tip modela koristi se u slučaju kada postoji velik broj jedinki, a malo ili nikoliko vremenskih efekata u poželjno što većem broju vremenskih perioda. Za procjenu vektora β potrebno je konstruirati model s pomoćnim varijablama ili model unutar grupa.

Model s pomoćnim varijablama (*Least squares dummy variable (LSDV) model*)

Kako bi se izračunao procjenitelj $\hat{\beta_{FE}}$ potrebno je proširiti model 1.13 u oblik

$$Y_{it} = \sum_{j=1}^{N} \alpha_i D_{ji} + \beta_1 X_{1,it} + ... + \beta_K X_{K,it} + \epsilon_{it} \qquad i = 1, ..., N, \ t = 1, ..., T \tag{1.16}$$

pri čemu je D_{ji} indikatorska varijabla definirana

$$D_{ji} = \begin{cases} 1, & \text{za j=i} \\ 0, & \text{inače} \end{cases} \qquad j = 1, ..., N$$

koja sadrži efekt specifičan za i-tu jedinicu promatranja [2]. Upravo eliminacijom jedne inidikatorske varijable bit će izbjegnut problem multikolinearnosti u modelu 1.16 pa je potrebno modificirati sumu indikatorskih varijabli na N-1 član. Stoga, konačan model glasi

$$Y_{it} = \sum_{j=1}^{N-1} \alpha_i D_{ji} + \beta_1 X_{1,it} + ... + \beta_K X_{K,it} + \epsilon_{it} \qquad i = 1, ..., N, \ t = 1, ..., T. \tag{1.17}$$

Procjenitelj modela 1.17 je LSDV procjenitelj te će njegova formula biti jednaka formuli modela unutar grupa pa će biti prikazana naknadno 1.21.

Model unutar grupa (*Within model*)

Jednadžbom 1.14 dan je matrični zapis FE modela tj.

$$Y_{it} = \alpha_i + \beta X_{it}^T + \epsilon_{it} \qquad i = 1, ..., N, \ t = 1, ..., T. \tag{1.18}$$

Vremenski prosječni model FE modela 1.18 dan je jednadžbom

$$\overline{Y_i} = \alpha_i + \beta \overline{X_i}^T + \overline{\epsilon_i} \qquad i = 1, ..., N. \tag{1.19}$$

Kako bi se izbjeglo uvođenje pomoćnih varijabli za računanje procjenitelja kao u LSVD modelu potrebno je napraviti razliku jednadžbi 1.18 i 1.19. Time je kreiran novi model oblika

$$Y_{it} - \overline{Y_i} = \beta(X_{it} - \overline{X_i})^T + (\epsilon_{it} - \overline{\epsilon_i}) \qquad i = 1, ..., N \ t = 1, ..., T. \tag{1.20}$$

kojime su eliminirani svi individualni efekti. Primjenom metode najmanjih kvadrata dobiven je procjenitelj parametara $\hat{\beta}_{within}$ za FE model koji glasi

$$\hat{\beta}_{within} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \overline{X_i})^T(X_{it} - \overline{X_i})\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \overline{X_i})^T(Y_{it} - \overline{Y_i})\right] \tag{1.21}$$

Kako je ranije napomenuto $\hat{\beta}_{LSDV}$ dan je istom formulom kao $\hat{\beta}_{within}$. Dodatni uvjeti koji moraju vrijediti da bi procjenitelj bio ispravno izračunat:

1. $\mathbb{E}[\epsilon_{it} - \overline{\epsilon_i} \mid X_{it} - \overline{X_i}] = 0 \qquad \forall i = 1, ..., N, \qquad \forall t = 1, ..., T$
2. $\mathbb{E}[\epsilon_{it} | \, X_{i1}, ..., X_{iN}] = 0 \qquad \forall i = 1, ..., N$

Ukoliko je potrebno procijeniti i individualne efekte moguće je jednadžbom

$$\hat{\alpha}_i = \overline{Y_i} - \hat{\beta}_{within}\overline{X_i}^T \qquad i = 1, ..., N. \tag{1.22}$$

Procjenitelji $\hat{\alpha}_i$ su nepristrani i bit će konzistentni jedino ako je zadan velik panel takav da $T \to \infty$, dok će procjenitelj $\hat{\beta}_{within}$ biti konzistentan ako vrijedi $NT \to \infty$.

Model slučajnih efekata (*Random effect (RE) model*)

U ovom modelu polazi se od pretpostavke da je individualni efekt α_i slučajna varijabla koja je dio slučajne greške. Model je prikazan oblikom

$$Y_{it} = \alpha_i + \beta_1 X_{1,it} + ... + \beta_K X_{K,it} + \epsilon_{it} \qquad i = 1, ..., N, \; t = 1, ..., T \tag{1.23}$$

tj. matrično

$$Y_{it} = \beta X_{it}^T + u_{it} \qquad i = 1, ..., N, \; t = 1, ..., T \tag{1.24}$$

gdje je

$$u_{it} = \alpha_i + \epsilon_{it} \qquad i = 1, ..., N, \; t = 1, ..., T \tag{1.25}$$

takvi da $\alpha_i \sim N(0, \sigma_\alpha^2)$ i $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$.
Da bi model bio ispravno primijenjen potrebno je zadovoljiti pretpostavke:

1. Komponente greške su nekorelirane.

$$Cov(\epsilon_{it}, \epsilon_{js}) = \mathbb{E}[\epsilon_{it}\epsilon_{js}] = 0 \qquad \forall i, j = 1, ..., N, \; \forall s, t = 1, ..., T$$

$$Cov(\alpha_i, \alpha_j) = \mathbb{E}[\alpha_i \alpha_j] = 0 \qquad \forall i, j = 1, ..., N$$

$$Cov(\alpha_i, \epsilon_{it}) = \mathbb{E}[\alpha_i \epsilon_{it}] = 0 \qquad \forall i = 1, ..., N, \; t = 1, ..., T$$

2. Komponente greške nisu korelirane s nezavisnim varijablama.

$$Cov(X_{it}, \alpha_i) = \mathbb{E}[X_{it}\alpha_i] = 0 \qquad \forall i = 1, ..., N, \qquad \forall t = 1, ..., T$$

$$Cov(X_{it}, \epsilon_{it}) = \mathbb{E}[X_{it}\epsilon_{it}] = 0 \qquad \forall i = 1, ..., N, \qquad \forall t = 1, ..., T$$

Koristeći prethodne pretpostavke za u_{it} vrijedi:

1. Očekivanje
$$\mathbb{E}[u_{it}] = \mathbb{E}[\alpha_i] + \mathbb{E}[\epsilon_{it}] = 0$$

2. Varijanca

$$\mathbb{E}[u_{it}^2] = \mathbb{E}[\alpha_i^2] + 2\mathbb{E}[\alpha_i\epsilon_{it}] + \mathbb{E}[\epsilon_{it}^2] = \sigma_\alpha^2 + \sigma_\epsilon^2$$

$$Var(u_{it}) = \mathbb{E}[u_{it}^2] - (\mathbb{E}[u_{it}])^2 = \sigma_\alpha^2 + \sigma_\epsilon^2$$

3. Kovarijanca

$$Cov(u_{it}, X_{it}) = 0$$

$$Cov(u_{it}, u_{js}) \begin{cases} \sigma_\alpha^2 + \sigma_\epsilon^2, & i = j, \ t = s \\ \sigma_\alpha^2, & i = j, \ t \neq s \\ 0, & \text{inače} \end{cases}$$

Budući da slučajne greške u_{it} ovise o vremenskom periodu t te vremenski nepromijenjivoj komponenti α_i one postaju korelirane. Zbog toga procjenitelj dobiven metodom najmanjih kvadrata više neće biti efikasan pa je potrebno provesti GLS metodu koja je učinkovitija.

Poopćena metoda najmanjih kvadrata (*Generalized least squares (GLS) method*)

Jednadžbom 1.24 dan je matrični zapis RE modela tj.

$$Y_{it} = \beta X_{it}^T + u_{it} \qquad i = 1, ..., N, \ t = 1, ..., T. \tag{1.26}$$

Model je potrebno proširiti tako da se uvede neslučajan skalar η kako bi slučajni efekti bili normalizirani s očekivanjem nula

$$Y_{it} = \eta + \beta X_{it}^T + u_{it} \qquad i = 1, ..., N, \; t = 1, ..., T. \tag{1.27}$$

Potrebno je eliminirati korelaciju među greškama tako da se model uprosječi po vremenskom periodu tj.

$$\overline{Y_i} = \eta + \beta \overline{X_i}^T + u_i \qquad i = 1, ..., N, \tag{1.28}$$

te da mu se pridruži skalar $\hat{\lambda}$ konzistentan s

$$\lambda = 1 - \sqrt{\frac{\sigma_\epsilon^2}{T\sigma_\alpha^2 + \sigma_\epsilon^2}} \in [0, 1] \tag{1.29}$$

pa konačno slijedi

$$\hat{\lambda}\overline{Y_{it}} = \hat{\lambda}\eta + \hat{\lambda}\beta \overline{X_{it}}^T + \hat{\lambda}u_{it} \qquad i = 1, ..., N, \; t = 1, ..., T. \tag{1.30}$$

Sada oduzimanjem modela 1.28 od modela 1.27 slijedi novi model oblika

$$Y_{it} - \hat{\lambda}\overline{Y_i} = (1 - \hat{\lambda})\eta + \beta(X_{it} - \hat{\lambda}\overline{X_i})^T + (1 - \hat{\lambda})\alpha_i + (\epsilon_{it} - \hat{\lambda}\overline{\epsilon_i}) \qquad i = 1, ..., N. \tag{1.31}$$

Procjenitelj slučajnih efekata $\hat{\beta}_{RE}$ dobiva se procjenom modela 1.31 metodom najmanjih kvadrata i dan je formulom

$$\hat{\beta}_{RE} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \hat{\lambda}\overline{X_i})^T(X_{it} - \hat{\lambda}\overline{X_i})\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \hat{\lambda}\overline{X_i})^T(Y_{it} - \hat{\lambda}\overline{Y_i})\right] \tag{1.32}$$

Posebno, ako je $\hat{\lambda} = 1$ vrijedi β_{RE}=β_{FE}, a za $\hat{\lambda} = 0$ vrijedi β_{RE}=$\beta_{združeni}$.

Glavna razlika između RE i FE modela je u odnosu prema vremenski nepromijenjivim varijablama. FE model ne dopušta da nezavisne varijable budu konstantne jer to uzrukuje multikolineranost, dok RE model nema takvih problema. Cilj oba modela je pronaći nepristrane i konzistentne procjenitelje.

1.4 Usporedni testovi modela

Za isti set panel podataka moguće je kreirati modele iz prethodnog potpoglavlja. Za odabir jednog od tih modela koji najbolje opisuje podatke potrebno je provesti testiranja.

Hausmanov test

Kako bi se odredilo koji je model prikladniji u slučaju kada postoje individualni efekti koristi se Hausmanov test. Ideja testa jest usporediti procijenitelje RE i FE modela za parametar β i ustanoviti koliko su različiti na danoj razini značajnosti. Razlika dolazi iz koreliranosti α_i i $X_{k,it}$ te je test definiran u obliku

$$\begin{cases} H_0 : Cov(\alpha_i, X_{k,it}) = 0 & \forall k = 1, ..., K \\ H_a : Cov(\alpha_i, X_{k,it}) \neq 0 & \text{za neki k} \end{cases}$$

Test statistika dana je formulom

$$H = \left[\hat{\beta}_{FE} - \hat{\beta}_{RE}\right]^T (Var\hat{\beta}_{FE} - Var\hat{\beta}_{RE})^{-1} \left[\hat{\beta}_{FE} - \hat{\beta}_{RE}\right]$$

gdje su $\hat{\beta}_{FE}$ i $\hat{\beta}_{RE}$ procjene parametara za redom fiksne efekte pomoću metode unutar grupa te slučajne efekte pomoću GLS metode. Ukoliko je vrijednost H statistike manja od $\chi^2(K)$, gdje je K broj procijenjenih parametara, tada se prihvaća nul hipoteza. U tom slučaju odabran je model slučajnih efekata kao prikladniji. U suprotnom, odabran će biti model fiksnih efekata kao bolji tj. prema alternativnoj hipotezi slijedit će da procijenitelji RE modela nisu konzistentni.

F-test

Korištenjem F-testa ispituje se postojanje fiksnih efekata u panel podacima. Glavni zaključak testa svodi se na odabir boljeg modela između modela fiksnih efekata i združenog modela. Pretpostavke testa zapisuju se u obliku

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = ... = \alpha_{N-1} = 0 \\ H_a : \text{postoji barem jedan } \alpha_i \neq 0 \end{cases}$$

Neka su $SSE_{združeni}$ i SSE_{fiksni} kvadratne rezidualne sume iz zduženog modela i modela fiksnih efekata. Tada je test statistika dana formulom [11]

$$F = \frac{SSE_{združeni} - SSE_{fiksni}}{SSE_{fiksni}} \frac{N(T-1) - K}{N-1}$$

Ukoliko se odbacuje nulta hipoteza prikladniji je model s individualnim efektima tj. slijedi zaključak da je model fiksnih efekata bolji za danu razini značajnosti. No, ako je vrijednost F testa manja od $F(N-1, NT-N-K)$ za danu razinu značajnosti, prihvaća se nul hipoteza i slijedi da je združeni model prikladnije koristiti.

Lagrangeov multiplikacijski test uz metodu Breusch-Pagan (LMBP)

Slično kao i F-testom, LMBP testom ispituje se postojanje individualnih efekata pa su pretpostavke testa zapisane u obliku

$$\begin{cases} H_0 : \text{Ne postoje individualni efekti.} \\ H_a : \text{Postoje individualni efekti.} \end{cases}$$

Za razliku od F-testa, umjesto FE modela uzima se RE model u usporedbu sa združenim modelom. Neka su e_{it} reziduali združenog modela. Tada je test statistika dana formulom

$$LM = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^{N} (\sum_{t=1}^{T} e_{it})^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2} - 1 \right]^2$$

čime je dan omjer sume kvadrata reziduala po jedinkama sa sumom kvadrata svih reziduala. Inače, Breusch-Pagan test koristi se za provjeru heteroskedastičnosti reziduala pa je test moguće definirati i sljedećim zapisom:

$$\begin{cases} H_0 : \sigma_\alpha^2 = 0 \\ H_a : \sigma_\alpha^2 \neq 0 \end{cases}$$

pri čemu su pretpostavke RE modela $\mathbb{E}(\alpha_i) = 0$ i $Var(\alpha_i) = \sigma_\alpha^2 > 0$.
Ako je vrijednost LMBP testa veća od $\chi^2(1)$ za danu razinu značajnosti, odbacuje se nul hipoteza pa je RE model prikladnije koristiti jer su prisutni individualni efekti. U protivnom, odabire se združeni model [10].

U tablici 1.2 prikazan je redoslijed testiranja radi donošenja lakšeg zaključka koji model je najbolji.

Tablica 1.2: Redoslijed provedbe testova za odabir najboljeg modela

<table>
<tr><td>Test 1</td><td colspan="4">Hausmanov test</td></tr>
<tr><td>Odabir boljeg modela</td><td colspan="2">RE</td><td colspan="2">FE</td></tr>
<tr><td>Prijelaz na novi test</td><td colspan="2">↓</td><td colspan="2">↓</td></tr>
<tr><td>Test 2</td><td colspan="2">LMBP</td><td colspan="2">F-test</td></tr>
<tr><td>Odabir boljeg modela</td><td>združeni</td><td>RE</td><td>združeni</td><td>FE</td></tr>
<tr><td>Najbolji model za panel podatke</td><td colspan="4">Bolji model nakon provedbe drugog testa</td></tr>
</table>

Poglavlje 2

Primjer: Analiza COVID-19 podataka u R-u

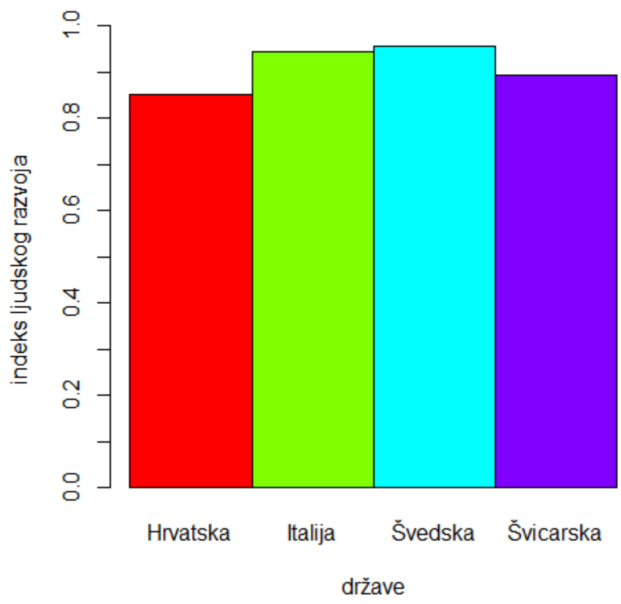
Krajem 2019. te u prvoj polovici 2020. godine s kontinenta Azija proširio se na ostatak svijeta virus SARS-CoV-2 koji je uzročnik bolesti COVID-19. Problem kod bolesti je visok stupanj zaraznosti jer se infekcija prenosi s pojedinca na pojedinca sitnim kapljicama iz usta ili nosa koje se izbacuju kad oboljela osoba kiše, kašlje ili govori. Ovisno o zdravstvenom stanju zaraženog simptomi bivaju različiti: asimptomatsko stanje, blagi simptomi poput kašlja i gubitka okusa i mirisa te jaki simptomi poput respiratornih problema koji znaju rezultirati smrtnim ishodima. Zbog velikog broja oboljelih, zdravstveni sustavi svih zemalja svijeta bili su pod velikim opterećenjem zbog nepostojanja lijeka ili cjepiva za ozdravljenje i zaštitu. Preventivne mjere temeljene su na fizičkom razmaku među ljudima, pojačanoj higijeni ruku i prostora te respiratornoj higijeni uz nošenje maske za lice koja smanjuje rizik prijenosa kapljica [3].

U ovom poglavlju provedena je u R-u analiza baze podataka [4] [5] o bolesti COVID-19 koristeći teorijsku podlogu iz prethodnih poglavlja. Koncentracija je usmjerena na europske države Hrvatsku, Švicarsku, Italiju i Švedsku u razdoblju od svibnja 2020. godine do studenog 2021. godine. Zbog negativnih pandemijskih utjecaja na stanovništvo odabran je broj smrtnih slučajeva kao zavisna varijabla s obzirom na nezavisne varijable:

- broj slučajeva zaraze u mjerilu jedan naprema milijun
- broj hospitaliziranih osoba u mjerilu jedan naprema milijun
- broj cijepljenih osoba barem jednom dozom cjepiva u mjerilu jedan naprema sto
- indeks strogosti mjera u rasponu od 0 do 100
- indeks ljudskog razvoja u rasponu od 0 do 1

2.1 Deskriptivna statistika

Indeks ljudskog razvoja predstavlja tri dimenzije razvoja čovjeka - dug život i zdravlje, znanje i obrazovanje te kvalitetan životni standard.



Slika 2.1: Grafički prikaz indeksa ljudskog razvoja prema državama

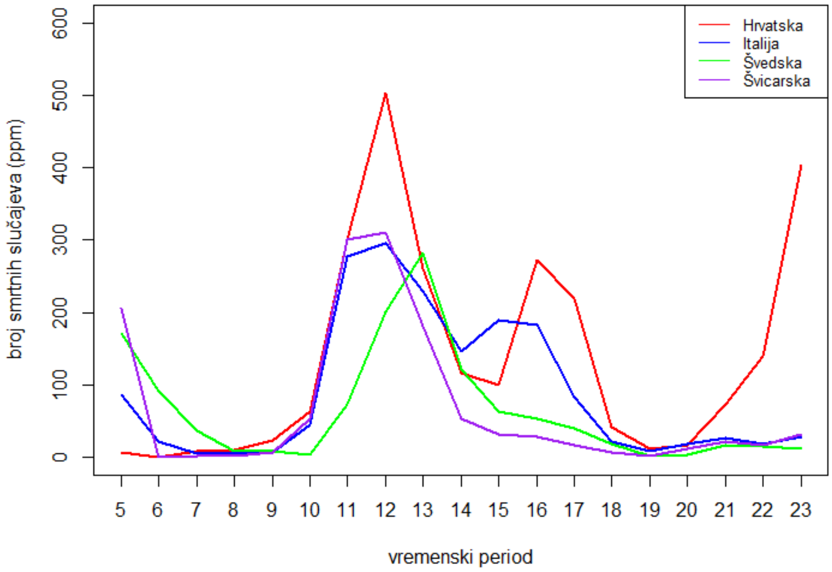
Moguće je uočiti da je u svim državama iz analize kvaliteta života skoro pa odlična jer im je iznos indeksa nedaleko od vrijednosti 1 koja predstavlja maksimum izvrsnosti.

Za lakše snalaženje u vremenskom razdoblju od svibnja 2020. godine do studenog 2021. godine sljedećom tablicom prikazan je tumač za nadolazeće grafove. Svakom paru kojeg čine mjesec i godina pridružena je vrijednost u rasponu brojeva od 5 do 23 koja se naziva vremenski period.

Tablica 2.1: Podjela vremenskog perioda

	SIJEČANJ	VELJAČA	OŽUJAK	TRAVANJ	SVIBANJ	LIPANJ	SRPANJ	KOLOVOZ	RUJAN	LISTOPAD	STUDENI	PROSINAC
vremenski period					5	6	7	8	9	10	11	12
godina					2020.	2020.	2020.	2020.	2020.	2020.	2020.	2020.
vremenski period	13	14	15	16	17	18	19	20	21	22	23	
godina	2021.	2021.	2021.	2021.	2021.	2021.	2021.	2021.	2021.	2021.	2021.	

Sljedećim tablicama i grafovima prikazana je usporedba vrijednosti svake varijable pojedine države u vremenskom periodu.

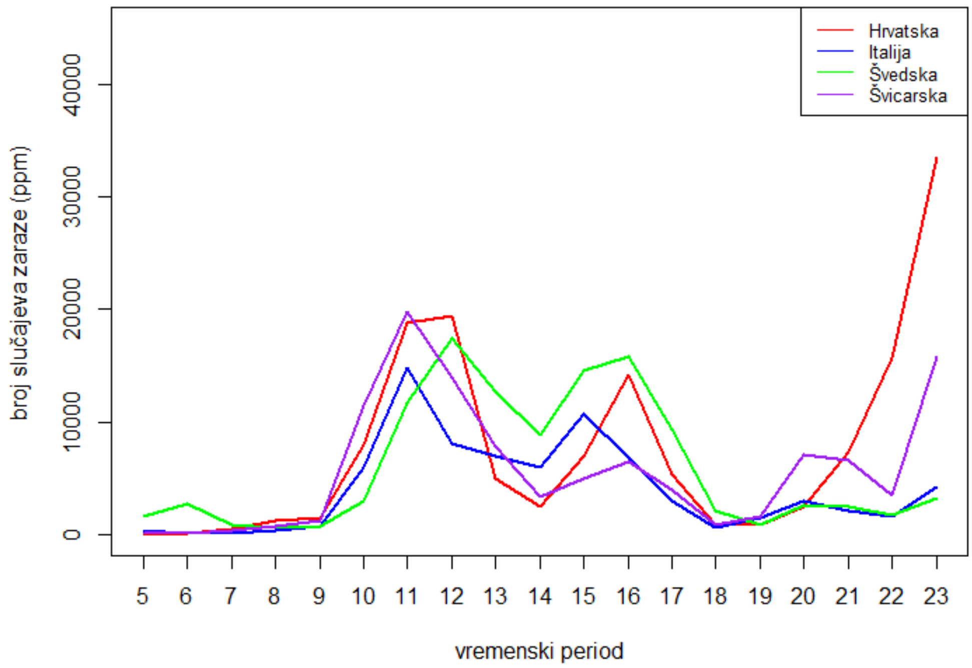



```
Descriptive Statistics
`broj smrtnih slucajeva (ppm)` by drzava
Data Frame: novipodaci
N: 19

                 Hrvatska   Italija   Švedska   Švicarska
------------- ---------- --------- --------- -----------
         Mean     135.63     89.37     64.37       67.47
      Std.Dev     149.88     98.65     78.22      101.39
          Min       1.00      6.00      2.00        1.00
       Median      74.00     28.00     37.00       21.00
          Max     504.00    295.00    282.00      311.00
```

Slika 2.2: Grafički prikaz i deskriptivna statistika broja smrtnih slučajeva u državama s obzirom na vremenski period

Moguće je primjetiti da se u svim državama značajno smanjuje broj smrtnih slučajeva u ljetnim mjesecima naspram zimskih mjeseci. Hrvatska ima najveće oscilacije tijekom vremena. Italija, Švedska i Švicarska postižu vrlo sličan maksimum smrtnosti, dok je Hrvatskoj maksimum otprilike 68% veći od njih.

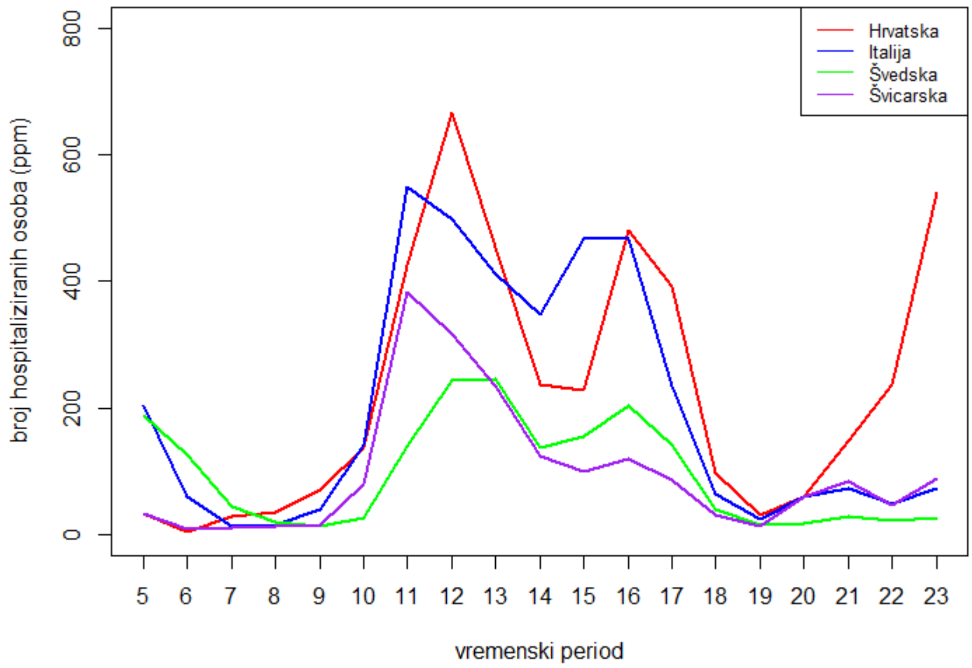



```
Descriptive Statistics
`broj slucajeva zaraze (ppm)` by drzava
Data Frame: novipodaci
N: 19

                 Hrvatska    Italija    Švedska   Švicarska
-------------- ---------- ---------- ---------- -----------
          Mean    7618.05    4069.74    5966.74     5808.63
       Std.Dev    8950.66    4074.94    5799.28     5767.46
           Min      39.00     112.00     689.00       97.00
        Median    5003.00    3018.00    2816.00     3989.00
           Max   33499.00   14780.00   17384.00    19829.00
```

Slika 2.3: Grafički prikaz i deskriptivna statistika broja slučajeva zaraze u državama s obzirom na vremenski period

U razdoblju od devetog do osamnaestog vremenskog perioda kretanje varijable broja slučajeva zaraze je vrlo slično za sve države. Uz to, od devetog do jedanaestog vremenskog perioda može se primjetiti da se radi o rastu svih krivulja koje su na tom intervalu vrlo sličnog nagiba.

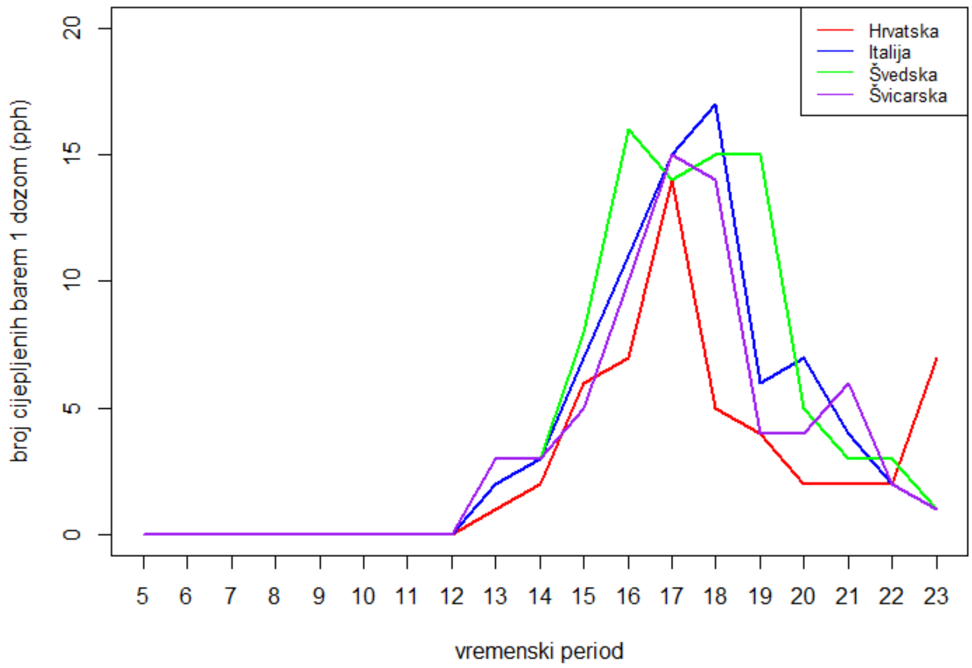



```
Descriptive Statistics
`broj hospitaliziranih osoba (ppm)` by drzava
Data Frame: novipodaci
N: 19

                 Hrvatska   Italija   Švedska   Švicarska
------------- ---------- --------- --------- -----------
         Mean     226.79    200.05     96.95       97.63
      Std.Dev     205.15    191.69     83.01      105.14
          Min       5.00     14.00     14.00       10.00
       Median     149.00     74.00     45.00       80.00
          Max     667.00    549.00    245.00      384.00
```

Slika 2.4: Grafički prikaz i deskriptivna statistika broja hospitaliziranih osoba u državama s obzirom na vremenski period

Slično kao za varijablu broj smrtnih slučajeva u ljetnim mjesecima je broj hospitaliziranih osoba najmanji. Generalno najmanji broj hospitaliziranih u cijelom periodu ima Švedska što je moguće zbog politike obrane od pandemije imunitetom "krda".

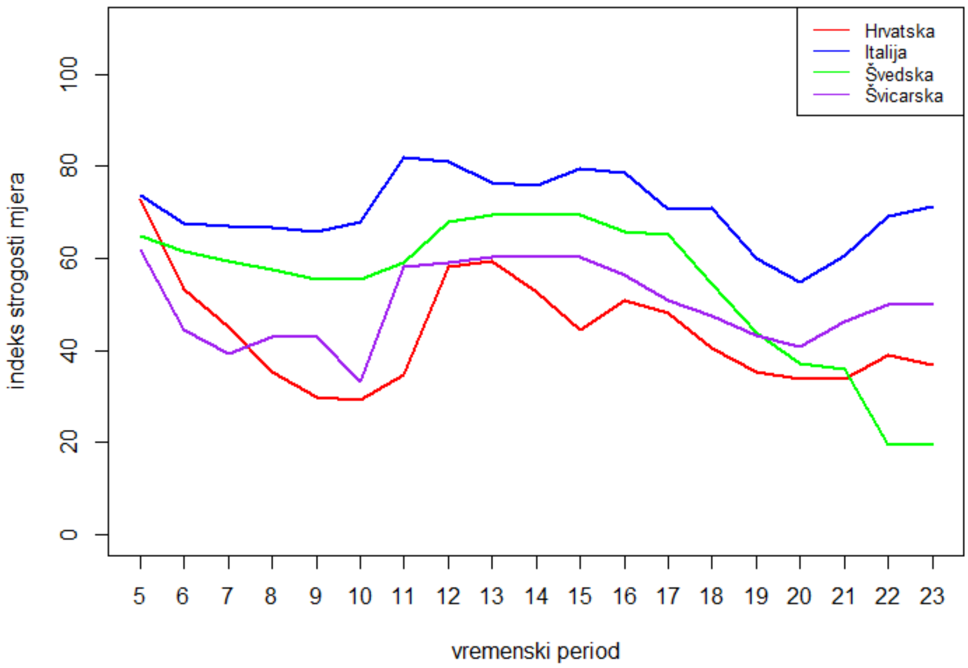



```
Descriptive Statistics
`broj cijepljenih barem 1 dozom (pph)` by drzava
Data Frame: novipodaci
N: 19

                Hrvatska   Italija   Švedska   Švicarska
------------- ---------- --------- --------- -----------
         Mean       2.74      3.95      4.53        3.53
      Std.Dev       3.69      5.31      5.96        4.72
          Min       0.00      0.00      0.00        0.00
       Median       2.00      2.00      3.00        2.00
          Max      14.00     17.00     16.00       15.00
```

Slika 2.5: Grafički prikaz i deskriptivna statistika broja cijepljenih barem 1 dozom u državama s obzirom na vremenski period

Cjepiva za obranu od virusa počela su se koristiti krajem 2020. godine. Moguće je biti cijepljeno s jednom, dvije i "booster" dozom, no ne odjednom već u razmacima od nekoliko mjeseci od primitka prethodne doze. U skladu s time, porast cijepljenih se ostvaruje od kraja 2020. godine, a počinje padati nakon otprilike pet mjeseci. Razlog može biti zbog toga što su pojedinci odlučili ne uzeti sljedeću dozu ili se pak uopće nisu htjeli cijepiti.



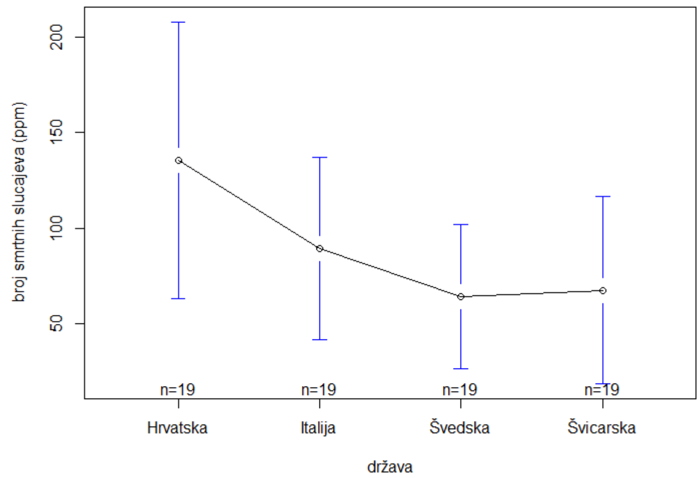

```
Descriptive Statistics
`indeks strogosti mjera` by drzava
Data Frame: novipodaci
N: 19

                 Hrvatska   Italija   Švedska   Švicarska
------------- ---------- --------- --------- -----------
         Mean      43.85     70.48     54.23       49.87
      Std.Dev      11.63      7.36     15.85        8.56
          Min      29.24     54.75     19.44       33.19
       Median      40.47     70.55     59.01       50.00
          Max      72.79     81.79     69.44       61.74
```

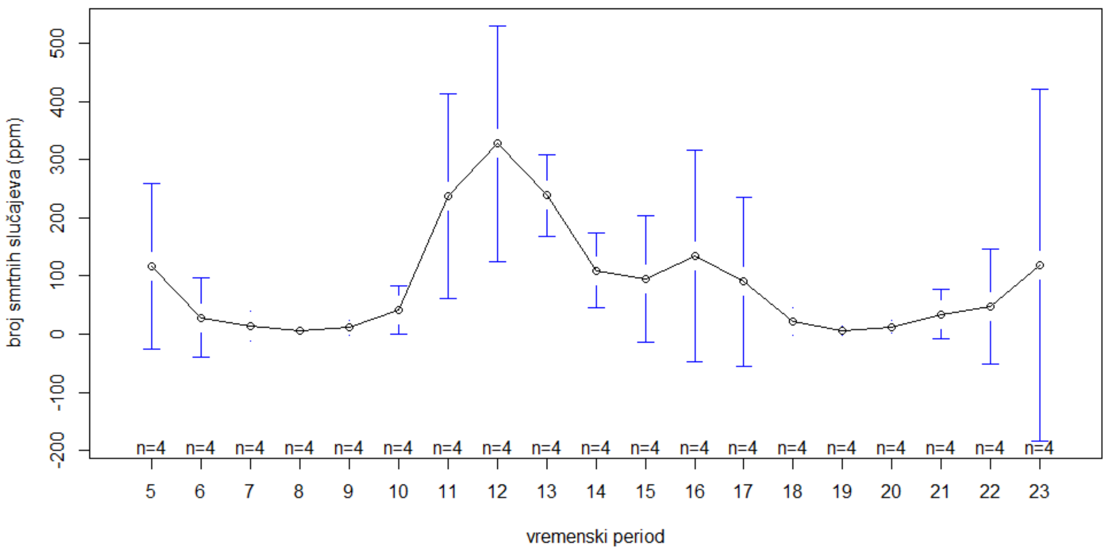
Slika 2.6: Grafički prikaz i deskriptivna statistika indeksa strogosti mjera u državama s obzirom na vremenski period

Indeks strogosti mjera predstavlja ograničenja kretanja pojedinaca poput zabrane okupljanja i kretanja među velikim brojem ljudi, zatvaranja škola i tvrtki ili zabrane putovanja. Italija je po tom pitanju bila najstroža u cijelom periodu.

Za dodatni grafički uvid u zavisnu varijablu broj smrtnih slučajeva na slikama 2.7 i 2.8 prikazane su aritmetičke sredine i 95% interval pouzdanosti zavisne varijable s obzirom na podjele prema vrsti države i vremenskom periodu.



Slika 2.7: Aritmetička sredina i 95% interval pouzdanosti broja smrtnih slučajeva u mjerilu $1 : 10^6$ prema državama za 19 vremenskih perioda



Slika 2.8: Aritmetička sredina i 95% interval pouzdanosti broja smrtnih slučajeva u mjerilu $1 : 10^6$ prema vremenskom periodu za 4 države

2.2 Rezultati analize

Konstruirani model dan je jednadžbom

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \beta_4 x_{4,it} + \beta_5 x_{5,it} \tag{2.1}$$

gdje je

- y = broj smrtnih slučajeva u mjerilu $1 : 10^6$
- x_1 = broj zaraženih osoba u mjerilu $1 : 10^6$
- x_2 = broj hospitaliziranih osoba u mjerilu $1 : 10^6$
- x_3 = broj cijepljenih osoba barem jednom dozom cjepiva u mjerilu $1 : 10^2$
- x_4 = indeks strogosti mjera u intervalu $[0, 100]$
- x_5 = indeks ljudskog razvoja u intervalu $[0, 1]$

za $i = 1, 2, 3, 4$ tako da

- 1 = Hrvatska
- 2 = Švedska
- 3 = Švicarska
- 4 = Italija

u vremenskom periodu $t = 5, 6, 7, ..., 23$ objašnjenom u tablici 2.1.

Varijable je potrebno transformirati kako bi bile normalizirane. U skladu s time, varijable broj smrtnih slučajeva, broj zaraženih osoba, broj hospitaliziranih osoba i indeks strogosti mjera bit će transformirane funkcijom $ln(x)$. Varijabla broj cijepljenih osoba barem jednom dozom će biti transformirana funkcijom $ln(x + 1)$. Indeks ljudskog razvoja neće biti transformiran.

Tablica 2.2: Varijable modela i njihove transformacije

	VARIJABLA	**TRANSFORMIRANA VARIJABLA**	**NOVO IME VARIJABLE**
y	broj smrtnih slučajeva	ln(broj smrtnih slučajeva)	umrli
x_1	broj zaraženih osoba	ln(broj zaraženih osoba)	zaraženi
x_2	broj hospitaliziranih osoba	ln(broj hospitaliziranih osoba)	hospitalizirani
x_3	broj cijepljenih osoba barem jednom dozom	ln(broj cijepljenih osoba barem jednom dozom + 1)	cijepljeni
x_4	indeks strogosti mjera	ln(indeks strogosti mjera)	indstrogosti
x_5	indeks ljudskog razvoja	/	indljudskirazvoj

Analiza je provedena prema modelima objašnjenima u prethodnom poglavlju.

Model konstantnih koeficijenata

Združeni model dobiven je linearnom kombinacijom svih nezavisnih varijabli na skupu originalnih podataka ne uzimajući u obzir o kojoj se državi radi.

Sljedećom tablicom 2.3 prikazan je rezultat analize koristeći paket plm u R-u.

Tablica 2.3: Ispis združenog modela iz konzole R-a

```
Pooling Model

Call:
plm(formula = umrli ~ zarazeni + hospitalizirani + cijepljeni +
    indstrogosti + indljudskirazvoj, data = novipodaci, model = "pooling",
    index = c("drzava"))

Balanced Panel: n = 4, T = 19, N = 76

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-1.284533 -0.234578 -0.042401  0.220976  2.433022

Coefficients:
                  Estimate Std. Error t-value  Pr(>|t|)
(Intercept)      -3.690598   1.472516 -2.5063  0.014526 *
zarazeni         -0.130157   0.076701 -1.6969  0.094152 .
hospitalizirani   1.447026   0.103670 13.9579 < 2.2e-16 ***
cijepljeni       -0.165286   0.059402 -2.7825  0.006929 **
indstrogosti     -0.365081   0.236105 -1.5463  0.126551
indljudskirazvoj  3.795186   1.707502  2.2227  0.029471 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    188.03
Residual Sum of Squares: 17.682
R-Squared:      0.90596
Adj. R-Squared: 0.89925
F-statistic: 134.879 on 5 and 70 DF, p-value: < 2.22e-16
```

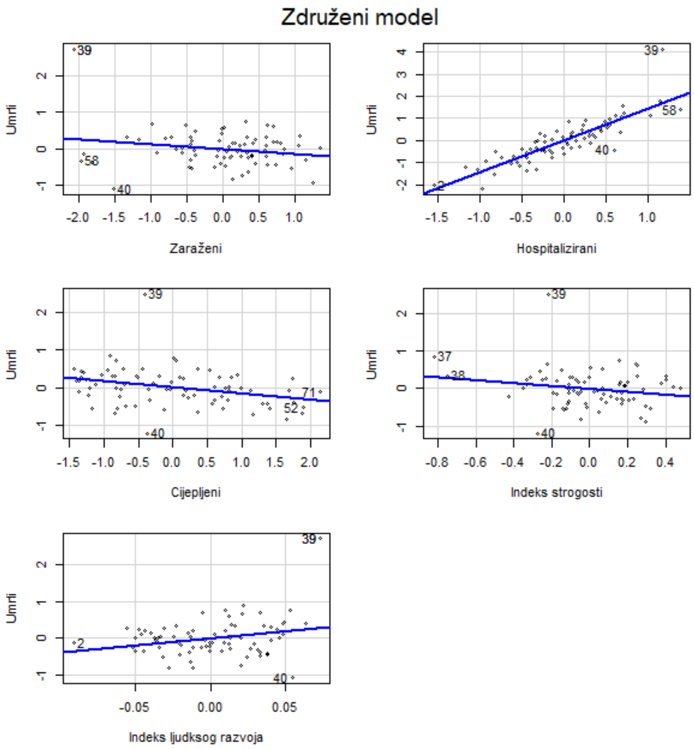
Iz rezultata dijagnostike modela slijedi da je model statistički značajan uz vrijednost F(5,70)=134.879; $p < 2.2 \times 10^{-16}$. Takav rezultat ukazuje na to da je ovim modelom uspješno objašnjeno 134.879 puta više varijabilnosti zavisne varijable umrli nego što je ostalo neobjašnjeno. Koeficijent determinacije R^2 iznosi 0.90596, dok prilagođeni koeficijent determinacije $Adj.R^2$ iznosi 0.89925 iz čega je vidljivo da su skoro jednaki. Zbog takvih koefcijenata slijedi zaključak da je model dobro prilagođen originalnim podacima jer se njegov koeficijent nalazi blizu desnog ruba intervala [0,1] te da je oko 90% ukupne

varijabilnosti zavisne varijable objašnjeno ovim modelom.
Iz tablice 2.3 slijedi da je združeni model dan jednadžbom:

$$\begin{aligned} umrli_{it} = -3.690598 - 0.130157 zara\check{z}eni_{it} + 1.447026 hospitalizirani_{it} \\ -0.165286 cijepljeni_{it} - 0.365081 indstrogosti_{it} \\ +3.795186 indljudskirazvoj_{it} \end{aligned}$$

u vremenskom periodu $t = 5, 6, 7, ..., 23$ za $i = 1, 2, 3, 4$.

Procijenjeni parametri modela ukazuju koliko se zavisna varijabla promijeni ukoliko se nezavisna varijabla poveća za jednu mjernu jedinicu. Primjerice, uz povećanje hospitaliziranih za 1 broj umrlih se poveća za 1.447026. Iz rezultata združenog modela 2.3 vidljivo je da povećanje hospitaliziranih statistički značajno povećava i smrtnost, dok povećan broj cijepljenih smanjuje. Statistički značajna se još pokazala i varijabla indeks ljudskog razvoja. Varijabla zaraženi statistički je značajna na razini značajnosti od 9.5% i ukazuje da povećanje zaraženih smanjuje smrtnost što može biti posljedica imunizacije "krda". Grafički prikaz utjecaja nezavisnih varijabli na zavisnu varijablu umrli vidljiv je na slici 2.9.



Slika 2.9: Utjecaj nezavisnih varijabli na zavisnu varijablu umrli u združenom modelu

Model fiksnih efekata (FE)

Model fiksnih efekata dobiven je linearnom kombinacijom svih nezavisnih varijabli na skupu originalnih podataka uzimajući u obzir podjelu podataka prema državama te vodeći računa o individualnih efektima. Za početak, bio je postavljen model koristeći sve nezavisne varijable, no prilikom računanja javljao se problem multikolinearnosti. Varijabla indeks ljudskog razvoja je vrijednost koja je konstantna za svaku državu i ona se manifestira kroz slobodni član za svaku pojedinu državu.

Radi iščitavanja konačnih modela i značajnosti varijabli bit će korištena tablica 2.4 za LSDV model i 2.5 za model unutar grupa koje prikazuju vrijednosti iz konzole R-a.

Tablica 2.4: Ispis LSDV modela iz konzole R-a

```
Call:
lm(formula = umrli ~ zarazeni + hospitalizirani + cijepljeni +
    indstrogosti + factor(drzava) - 1, data = novipodaci)

Residuals:
     Min       1Q   Median       3Q      Max
-1.17420 -0.23273 -0.04973  0.27975  2.51060

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
zarazeni                  -0.15288    0.07615  -2.008   0.0487 *
hospitalizirani            1.46917    0.10371  14.166   <2e-16 ***
cijepljeni                -0.16458    0.05838  -2.819   0.0063 **
indstrogosti              -0.32590    0.28066  -1.161   0.2496
factor(drzava)Hrvatska    -0.51478    1.09083  -0.472   0.6385
factor(drzava)Italija     -0.46578    1.21449  -0.384   0.7025
factor(drzava)Švedska      0.02042    1.19277   0.017   0.9864
factor(drzava)Švicarska   -0.27230    1.17544  -0.232   0.8175
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4928 on 68 degrees of freedom
Multiple R-squared:  0.9854,    Adjusted R-squared:  0.9837
F-statistic: 574.5 on 8 and 68 DF,  p-value: < 2.2e-16
```

Model fiksnih efekata LSDV je statistički značajan uz vrijednost F(8,68)=574.5; $p < 2.2 \times 10^{-16}$. Koeficijent determinacije R^2 iznosi 0.9854, a procijenjeni koeficijent determinacije $Adj.R^2$ je 0.9837. Zbog takvih koefcijenata slijedi zaključak da je model dobro prilagođen originalnim podacima jer je blizu 1 te da je otprilike 98.5% ukupne varijabilnosti zavisne varijable objašnjeno ovim modelom.

Tablica 2.5: Ispis modela unutar grupa iz konzole R-a

```
Call:
plm(formula = umrli ~ zarazeni + hospitalizirani + cijepljeni +
    indstrogosti + indljudskirazvoj, data = novipodaci, model = "within",
    index = c("brojdrz"))

Balanced Panel: n = 4, T = 19, N = 76

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-1.174200 -0.232731 -0.049733  0.279749  2.510604

Coefficients:
                 Estimate Std. Error t-value Pr(>|t|)
zarazeni        -0.152878   0.076150 -2.0076  0.04866 *
hospitalizirani  1.469174   0.103711 14.1661  < 2e-16 ***
cijepljeni      -0.164583   0.058379 -2.8192  0.00630 **
indstrogosti    -0.325900   0.280659 -1.1612  0.24962
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    177.38
Residual Sum of Squares: 16.516
R-Squared:      0.90689
Adj. R-Squared: 0.8973
F-statistic: 165.577 on 4 and 68 DF, p-value: < 2.22e-16
> # 1=Hrvatska 2=Švedska 3=Švicarska 4=Italija
> factfi=fixef(fixed)
> factfi
        1         2         3         4
-0.514783  0.020424 -0.272304 -0.465780
```

Model fiksnih efekata unutar grupa je statistički značajan uz vrijednost F(4,68)=165.577; $p < 2.2 \times 10^{-16}$. Koeficijenti determinacije navode da je 90% ukupne varijabilnosti umrlih objašnjeno ovim modelom, što je manje od LSDV modela, ali je dalje dobar rezultat.

Budući da model LSDV i model unutar grupa moraju imati iste procijenjene parametre nezavisnih varijabli model fiksnih efekata dan je jedinstvenim zapisom jednadžbi što je vidljivo iz tablica 2.4 i 2.5 :

- Hrvatska i=1

$$umrli_{1t} = -0.51478 - 0.152878 zaraženi_{1t} + 1.469174 hospitalizirani_{1t}$$
$$-0.164583 cijepljeni_{1t} - 0.325900 indstrogosti_{1t}$$

- Švedska i=2

$$umrli_{2t} = 0.02042 - 0.152878 zaraženi_{2t} + 1.469174 hospitalizirani_{2t}$$
$$-0.164583 cijepljeni_{2t} - 0.325900 indstrogosti_{2t}$$

- Švicarska i=3

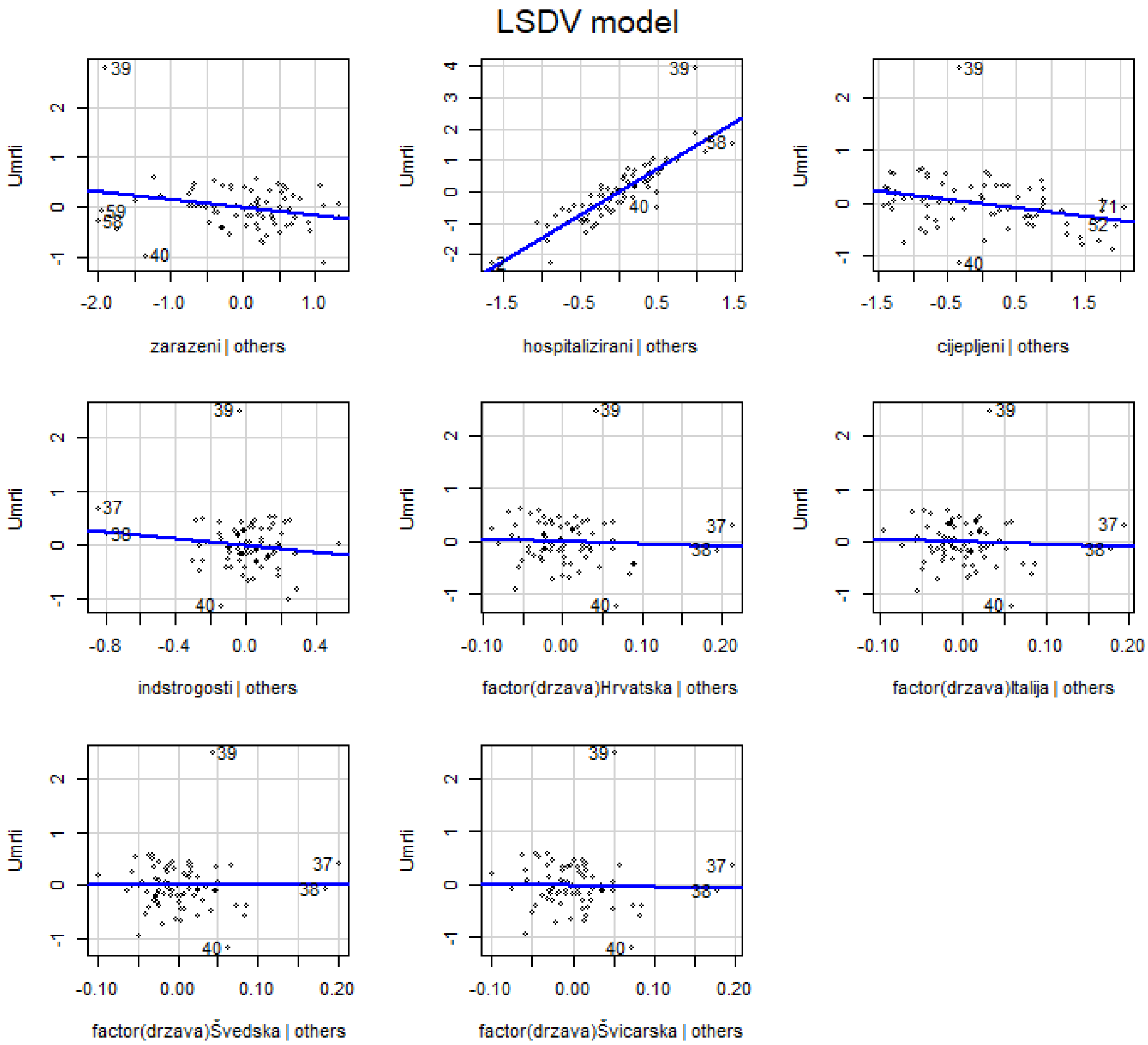
$$umrli_{3t} = -0.27230 - 0.152878 zaraženi_{3t} + 1.469174 hospitalizirani_{3t}$$
$$-0.164583 cijepljeni_{3t} - 0.325900 indstrogosti_{3t}$$

- Italija i=4

$$umrli_{4t} = -0.46578 - 0.152878 zaraženi_{4t} + 1.469174 hospitalizirani_{4t}$$
$$-0.164583 cijepljeni_{4t} - 0.325900 indstrogosti_{4t}$$

u vremenskom periodu $t = 5, 6, 7, ..., 23$.

Interpretacija procijenjenih parametara je vrlo slična kao i kod združenog modela samo što se mora voditi računa o razlikama među državama. Procijenjeni parametri pokazuju koliko se poveća zavisna varijabla tijekom vremena, kontrolirajući po državama kada se nezavisna varijabla poveća za 1. Kada se radi o značajnosti procijenjenih parametara varijable zaraženi, hospitalizirani i cijepljeni su statistički značajne i istog su smjera kao i u združenom modelu. Grafički prikaz utjecaja nezavisnih varijabli na varijablu umrli vidljiv je na slici 2.10.



Slika 2.10: Utjecaj nezavisnih varijabli na broj smrtnih slučajeva u LSDV modelu

Ukoliko je potrebno odlučiti koji je od danih modela fiksnih efekata bolji potrebno je odabrati LSDV zbog većeg iznosa R^2.

Dosad su bili prikazani načini kako se konstruira fiksni model bez međusobne usporedbe država. Naime, poželjno je napraviti još jedan LSDV model tako da se za referentnu državu odabere Hrvatska. Procijenjeni parametri uz svaku državu predstavljat će koliko se smanjio ili povećao iznos slobodnog koeficijenta naspram slobodnog koeficijenta Hrvatske. Osim toga, oni predstavljaju koliko iznosi prosječna razlika u smrtnosti između Hrvatske i svake države pojedinačno tijekom vremena.

Tablica 2.6: Ispis LSDV modela s Hrvatskom kao referentnom jedinicom iz konzole R-a

```
Call:
lm(formula = umrli ~ zarazeni + hospitalizirani + cijepljeni +
    indstrogosti + factor(drzava), data = novipodaci)

Residuals:
     Min       1Q   Median       3Q      Max
-1.17420 -0.23273 -0.04973  0.27975  2.51060

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.51478    1.09083  -0.472   0.6385
zarazeni                  -0.15288    0.07615  -2.008   0.0487 *
hospitalizirani            1.46917    0.10371  14.166   <2e-16 ***
cijepljeni                -0.16458    0.05838  -2.819   0.0063 **
indstrogosti              -0.32590    0.28066  -1.161   0.2496
factor(drzava)Italija      0.04900    0.20973   0.234   0.8160
factor(drzava)Švedska      0.53521    0.20215   2.648   0.0101 *
factor(drzava)Švicarska    0.24248    0.19249   1.260   0.2121
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4928 on 68 degrees of freedom
Multiple R-squared:  0.9122,    Adjusted R-squared:  0.9031
F-statistic: 100.9 on 7 and 68 DF,  p-value: < 2.2e-16
```

Iz rezultata modela 2.6 vidljivo je da su svi procjenitelji država pozitivni što znači da je smrtnost povećana u svakoj državi u usporedbi s Hrvatskom tijekom vremena. Najveća razlika je između Hrvatske i Švedske. Naime, Švedska je jedina imala politiku obrane od epidemije imunitetom krda. Uz tu činjenicu nije isključivo pitati se je li takva metoda obrane od epidemije zaslužna za činjenicu da je u Švedskoj smrtnost najveća.
Bez obzira je li model kreiran s ili bez referentne jedinice rezultati slobodnih koeficijenata će biti jednaki 2.7.

Tablica 2.7: Usporedba koeficijenata LSDV modela s i bez referentne jedinice

	Hrvatska	Švedska	Švicarska	Italija
slobodni koeficijent referentne države (skrd)	-0.51478	/	/	/
procijenjene usporedbe slobodnih koeficijenata s referentnom državom (pskrd)	/	0.53521	0.24248	0.04900
procijenjeni slobodni koeficijenti LSDV modela s referentnom državom (skrd+pskrd)	-0.51478	0.02042	-0.27230	-0.46578
procijenjeni slobodni koeficijenti LSDV modela bez referentne države	-0.51478	0.02042	-0.27230	-0.46578

Model slučajnih efekata (RE)

Model slučajnih efekata uključuje mogućnost varijabilnosti zavisne varijable između država i pretpostavlja da je ta varijabilnost slučajna i nekorelirana s nezavisnim varijablama. Za kreiranje modela slučajnih efekata korišten je plm paket u R-u prema metodi "walhus" koju su kreirali Wallace i Hussain 1969. godine. Za provedbu analize varijance korišten je "twoways" efekt kojim se uzima u obzir uz individualni efekt i efekt vremena. Tablica 2.8 prikazuje rezultate provedbe RE modela iz konzole R-a.

Tablica 2.8: Ispis modela slučajnih efekata iz konzole R-a

```
Twoways effects Random Effect Model
   (Wallace-Hussain's transformation)

Call:
plm(formula = umrli ~ zarazeni + hospitalizirani + cijepljeni +
    indstrogosti + indljudskirazvoj, data = novipodaci, effect = "twoways",
    model = "random", random.method = "walhus", index = c("drzava"))

Balanced Panel: n = 4, T = 19, N = 76

Effects:
                   var  std.dev share
idiosyncratic 0.259402 0.509315 0.996
individual    0.001068 0.032688 0.004
time          0.000000 0.000000 0.000
theta: 0.03697 (id) 0 (time) 0 (total)

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-1.28160 -0.23275 -0.04626  0.22115  2.43402

Coefficients:
                  Estimate Std. Error z-value  Pr(>|z|)
(Intercept)      -3.703896   1.512623 -2.4487  0.014339 *
zarazeni         -0.131809   0.076590 -1.7210  0.085258 .
hospitalizirani   1.448773   0.103553 13.9907 < 2.2e-16 ***
cijepljeni       -0.165280   0.059272 -2.7885  0.005295 **
indstrogosti     -0.364112   0.238135 -1.5290  0.126261
indljudskirazvoj  3.811438   1.747141  2.1815  0.029144 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    187.26
Residual Sum of Squares: 17.6
R-Squared:      0.90601
Adj. R-Squared: 0.8993
Chisq: 674.761 on 5 DF, p-value: < 2.22e-16
```

Iz dijagnostike modela slučajnih efekata vidljivo je da je model statistički značajan uz vrijednost Chisq(5)=674.761; $p < 2.2 \times 10^{-16}$. Koeficijenti determinacije iznose otprilike 0.9 pa zaključak o kvaliteti provedbe modela je isti kao i u združenom modelu i modelu unutar grupa.
U odjeljku Effects tablice 2.8 vidljivo je da je varijanca efekta vremena vrijednosti 0 što znači da nema takvih efekata, no individualni efekti hospitaliziranih i cijepljeni su i dalje statistički značajne varijable s istim smjerom kao i u prethodnim modelima. Varijabla zaraženi više nije statistički značajna na nivou značajnosti od 5% nego na 8.5%, dok je varijabla indeks ljudskog razvoja jednako značajna kao u združenom modelu.

Prema tablici 2.8 dobiven je model slučajnih efekata s jednadžbom

$$\begin{aligned} umrli_{it} = -3.70390 - 0.13181 zara\check{z}eni_{it} + 1.44877 hospitalizirani_{it} \\ -0.16528 cijepljeni_{it} - 0.36411 indstrogosti_{it} \\ +3.81144 indljudskirazvoj_{it} \end{aligned}$$

za $i = 1, 2, 3, 4$ u vremenskom periodu $t = 5, 6, 7, ..., 23$.
Tumačenje procijenjenih koeficijenata kod slučajnih modela je zahtjevnije budući da oni uključuju unutar entitetske i između entitetske učinke. Procijenjeni parametri predstavljaju prosječni učinak nezavisne varijable na zavisnu kada se nezavisna mijenja kroz vrijeme i između zemalja za jednu mjernu jedinicu.

Usporedba modela

Redoslijed provedbe testova zasniva se na tablici 1.2.

Prvim testom bit će odabran bolji model između modela slučajnih efekata i modela fiksnih efekata. Korišten je Hausmanov test pomoću naredbe phtest u R-u uz pretpostavke

$$\begin{cases} H_0 : \text{RE je bolji izbor} \\ H_1 : \text{FE je bolji izbor} \end{cases}$$

te slijede rezultati u tablici 2.9:

Tablica 2.9: Ispis rezultata Hausmanovog testa iz konzole R-a

```
	Hausman Test

data:  umrli ~ zarazeni + hospitalizirani + cijepljeni + indstrogosti +  ...
chisq = 7.6422, df = 4, p-value = 0.1056
alternative hypothesis: one model is inconsistent
```

Prema rezultatima slijedi da se H_0 ne odbacuje u korist H_1 na razini značajnosti od 5%. Dakle, bolji model je onaj u kojem postoje individualni efekti koji su nekorelirani s nezavisnim varijablama tako da su dio slučajne pogreške.

Nakon odabira boljeg modela s individualnim efektima potrebno je provjeriti je li takav model doista najbolji ili je možda bolje konstruirati model takav da inidividualnih efekata uopće nema. U skladu s tim pitanjem slijedi provedba LMBP-testa za odabir boljeg modela između združenog modela i modela slučajnih efekata pomoću naredbe plmtest u R-u. Pretpostavke su

$$\begin{cases} H_0 : \text{združeni model je bolji izbor - nema individualnih efekata} \\ H_1 : \text{RE je bolji izbor - ima individualnih efekata} \end{cases}$$

te slijede rezultati u tablici 2.10:

Tablica 2.10: Ispis rezultata LMBP-testa iz konzole R-a

```
        Lagrange Multiplier Test - (Breusch-Pagan) for balanced panels

data:  umrli ~ zarazeni + hospitalizirani + cijepljeni + indstrogosti +  ...
chisq = 0.086347, df = 1, p-value = 0.7689
alternative hypothesis: significant effects
```

Testiranjem je dobiveno da se H_0 ne može odbaciti na razini značajnosti od 5%. Dakle, model konstantnih koeficijenata najbolje opisuje dane panel podatke.

Unatoč činjenici što se mogu konstruirati modeli s postojanjem individualnih efekata ne mora se takvim modelima davati prednost pri odabiru najboljeg. Naime, koeficijenti determinacije svih modela su

$$\text{združeni model } R^2 = 0.90596 \qquad \text{model LSDV } R^2 = 0.9854$$
$$\text{model unutar grupa } R^2 = 0.90689 \qquad \text{RE model } R^2 = 0.90601$$

Usporedbom njihovih iznosa ne slijedi nužno da je onaj s najvećim iznosom najbolji što je upravo vidljivo iz konačnog rezultata usporednih testova.

Komparirati je moguće i procijenjene parametre za nezavisne varijable svih modela.

Tablica 2.11: Procijenjeni parametri nezavisnih varijabli združenog, FE i RE modela

tip modela	$\hat{\beta}_{zara\check{z}eni}$	$\hat{\beta}_{hospitalizirani}$	$\hat{\beta}_{cijepljeni}$	$\hat{\beta}_{indstrogosti}$	$\hat{\beta}_{indljudskirazvoj}$
združeni	0.130157	1.447026	0.165286	0.365081	3.795186
FE	0.152878	1.469174	0.164583	0.325900	/
RE	0.13181	1.44877	0.16528	0.36411	3.81144

U tablici 2.11 može se uočiti da su procjene sva tri modela za svaku nezavisnu varijablu vrlo približne vrijednosti istog predznaka, a slično vrijedi i za slobodni član u združenom modelu i modelu slučajnih efekata.

Bibliografija

[1] https://dokumen.tips/documents/analiza-panel-podataka.html?page=1.

[2] https://web.sgh.waw.pl/~jmuck/EconometricsOfPanelData.html.

[3] https://www.koronavirus.hr/sto-moram-znati/8.

[4] https://github.com/owid/covid-19-data/tree/master/public/data.

[5] https://ourworldindata.org/coronavirus.

[6] https://rstudio-pubs-static.s3.amazonaws.com/372492_3e05f38dd3f248e89cdedd317d603b9a.html#1_preliminaries.

[7] https://dss.princeton.edu/training/Panel101R.pdf.

[8] https://cran.r-project.org/web/packages/plm/vignettes/A_plmPackage.html.

[9] https://bookdown.org/ccolonescu/RPoE4/panel-data-models.html#organizing-the-data-as-a-panel.

[10] Blažević Andrijana, *Modeliranje rasta poduzeća u Hrvatskoj: analiza panel podataka*, Sveučilište J.J. Strossmayera u Osijeku Odjel za matematiku, 2020.

[11] Vašarević Filip, *Modeliranje rasta prihoda od prodaje iz panel podataka*, Sveučilište J.J. Strossmayera u Osijeku Odjel za matematiku, 2017.

[12] Park Hun Myoung, *Practical Guides To Panel Data Modeling: A Step by Step Analysis Using Stata*, Graduate School of International Relations International University of Japan, 2011.

[13] Šilj Marko, *Uvod u modernu poslovnu statistiku*, Ekonomski fakultet Sveučilišta u Mostaru, Nakladnička kuća Tonimir, Varaždinske toplice, 1998.

Sažetak

Ovaj rad baziran je na analizi statističkog i ekonometrijskog pojma naziva panel podaci. Analiza panel podataka vrlo je raširena u prirodoslovnim, društvenim i ekonomskim znanstvenim istraživanjima zbog mogućnosti praćenja više objekata putem više varijabli u više od jednog vremenskog perioda.

U prvom poglavlju definirani su panel podaci kao kombinacija vremenskih nizova i vremenskih presjeka te je dana njihova struktura. Kako bi bilo lakše razumijevanje linearnih modela s takvim podacima, za početak je objašnjena jednostavna linearna regresija te statistički pojmovi potrebni za analizu. Potom je prikazana osnovna konstrukcija linearnih modela za panel analizu te dana teorijska podloga za združeni model, model fiksnih efekata i model slučajnih efekata. Za odabir najboljeg modela od navedenih primijenjeni su F-test, Hausmann test i Lagrange Multiplier Test - (Breusch-Pagan). U drugom poglavlju provedena je analiza u programu za statističku analizu R nad panel podacima o COVID-19 bolesti u Hrvatskoj, Italiji, Švicarskoj i Švedskoj. U vremenskom razdoblju od svibnja 2020. godine do studenog 2021. godine promatrana je zavisna varijabla broj smrtnih slučajeva s obzirom na nezavisne varijable: broj zaraženih osoba, broj hospitaliziranih osoba, broj cijepljenih osoba barem jednom dozom cjepiva, indeks strogosti mjera te indeks ljudskog razvoja.

Summary

Fundamental part of this paper is analysis of statistical and econometrical term panel data. Panel analysis is used for scientific researches in fields like economy and nature and social sciences. It is very popular because it can handle multiple objects over multiple variables within more than one time period.

In first chapter, it is shown structure and definition of panel data as a combination of time series and cross-sectional data. Also, there is given an explanation of a simple linear regression and main statistical tools for validating linear models. Afterwards, basic structure of a linear panel model, pooled OLS model, fixed effects model and random model effects are presented. To determine which model fits best with the panel data there are conducted F-test, Hausman test and Lagrange Multiplier Test - (Breusch-Pagan). Second chapter gives overview of panel data analysis through example made in program for statistical analysis R. Panel is consisted of COVID-19 data from Croatia, Italy, Sweden and Switzerland from May 2020. until November 2021. Dependent variable is number of death cases opposite to independent variables that are number of total positive cases, number of patients in hospital, number of vaccinated patients with at least 1 dose, stringency index and human development index.

Životopis

Rođena sam 16. ožujka 1997. godine u Zagrebu. Svoje srednjoškolsko obrazovanje provela sam na općem smjeru u II. gimnaziji u Zagrebu. Odlučila sam nastaviti svoj obrazovni proces na Prirodoslovno-matematičkom fakultetu u Zagrebu upisavši preddiplomski nastavnički sveučilišni studij Matematike. Njegovim završetkom, upisala sam diplomski sveučilišni studij Matematičke statistike na istoimenom fakultetu te počela raditi na poziciji predavača iz Informatike i Matematike na Visokom učilištu Algebra.

Dodatak

U ovom dijelu rada priložen je kod iz R-a potreban za tablične rezultate svih modela i testova. Kombiniranjem izvora [6] [7] [8] [9] dobiven je kod za ovaj rad.

```
install.packages("summarytools")
install.packages("plm")
install.packages("car")
install.packages("lsdv")
install.packages("tidyverse")
install.packages("gplots")
install.packages("tseries")
install.packages("lmtest")
library(tidyverse) # Modern data science library
library(plm)       # Panel data analysis library
library(car)       # Companion to applied regression
library(gplots)    # Various programing tools for plotting data
library(tseries)   # For timeseries analysis
library(lmtest)    # For hetoroskedasticity analysis
library(readxl)

#ZDRUZENI MODEL
ols = plm(umrli~zarazeni+hospitalizirani+cijepljeni+indstrogosti+
    indljudskirazvoj,data = novipodaci,index = c("drzava"), model = "
    pooling")
summary(ols)

#LSDV MODEL
fixdum=lm(umrli ~ zarazeni+hospitalizirani+cijepljeni+indstrogosti+
    factor(drzava)-1, data = novipodaci)
summary(fixdum)

#LSDV MODEL Hrvatska referentna jedinica
fixhrv=lm(umrli ~ zarazeni+hospitalizirani+cijepljeni+indstrogosti+
    factor(drzava), data = novipodaci)
summary(fixhrv)

```

```
#MODEL UNUTAR GRUPA
fixed <- plm(umrli ~ zarazeni+hospitalizirani+cijepljeni+indstrogosti+
    indljudskirazvoj , data = novipodaci,index=c("brojdrz"), model = "
    within")
summary(fixed)
# 1=Hrvatska 2=Svedska 3=Svicarska 4=Italija
factfi=fixef(fixed)
factfi

#MODEL SLUCAJNIH EFEKATA
random = plm(umrli~zarazeni+hospitalizirani+cijepljeni+indstrogosti+
    indljudskirazvoj ,data= novipodaci,index = c("drzava"), model = "
    random", effect = "twoways", random.method ="walhus")
summary(random)

#HAUSMANOV TEST
phtest(fixed, random)

#LMBP TEST
plmtest(ols, type = c("bp"))
```