Analiza panel podataka

Pavić, Matea

Master's thesis / Diplomski rad

2019

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:399336

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Matea Pavić

ANALIZA PANEL PODATAKA

Diplomski rad

Voditeljica rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, srpanj 2019.

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. _____

Sadržaj

Sadržaj **iii**

Uvod **1**

1 Linearna regresija **2**
1.1 Linearni regresijski model 2

2 Analiza panel podataka **6**
2.1 Vrste podataka 6
2.2 Linearni modeli panel podataka 7
2.3 Testovi za odabir modela 18

3 Primjer u SAS-u **21**
3.1 Deskriptivna statistika 23
3.2 Modeli i analiza rezultata 32

Bibliografija **42**

Uvod

Analiza podataka čini osnovni okvir znanstvenog istraživanja, bez obzira na sadržaj ili pojavu koja se želi istražiti. Jedna od tehnika analiziranja jest i analiza panel podataka. Termin „panel podaci" odnosi se na združena opažanja vremenskog presjeka o nekoj razmatranoj jedinici prikupljena tijekom više vremenskih točaka. S obzirom na prirodu panel podataka, za njihovu analizu razvijene su posebne metode i modeli. Takvi modeli najčešće se koriste u analizi potrošnje, štednje ili tržišta, gdje jedinice promatranja mogu biti pojedinci, domaćinstva, poduzeća, države, čije je ponašanje međusobno i tijekom vremena različito.

Posljednjih nekoliko godina panel podaci sve se više koriste u empirijskim istraživanjima, što je posljedica sve lakše i šire dostupnosti informacija. Nekoliko primjera dobre prakse i korištenja panel modela dolazi iz Sjedinjenih Američkih Država, gdje je i započelo prvo organizirano prikupljanje panel podataka. Znanstvenici Instituta za društvena istraživanja na Sveučilištu Michigan pokrenuli su 1968. godine Panel istraživanje dinamike prihoda (Panel Study of Income Dynamics – PSID), a Centar za istraživanje ljudskih potencijala na Državnom sveučilištu Ohio već nekoliko desetljeća prikuplja podatke za Nacionalno longitudinalno istraživanje tržišta rada (National Longitudinal Surveys of Labor Market Experience – NLS). Projekt PSID prikuplja godišnje podatke na temelju nacionalnog uzorka koji se sastoji od približno 6000 obitelji i 15.000 pojedinaca. NLS je pokrenut 1960-ih, a čini ga pet odvojenih studija koje pokrivaju različite segmente tržišta radne snage. I mnoge europske države provode godišnja istraživanja, kao što su Nizozemski društveno-ekonomski panel, Njemački društveno-gospodarski panel i Britanski panel istraživanja kućanstava. Organizacija za gospodarsku suradnju i razvoj (OECD) objavljuje godišnja izvješća sa statističkim podacima o različitim gospodarskim aspektima raznih zemalja. Novi izvori podataka mogu se pronaći na internetskim tražilicama (npr. Google Flu Trends) te u podacima baza trgovačkih lanaca (Nielsen Datasets for Consumer Marketing). Podatke vezane za statistiku Europske unije i europskog područja mogu se pronaći na stranicama Eurostata koji nudi usporediv, pouzdan i objektivan prikaz promjena u Europi.

1 Linearna regresija

1.1 Linearni regresijski model

Jedan od osnovnih postupaka statističkog modeliranja jest regresijska analiza. Regresijskim postupkom želi se utvrditi kako se vrijednost zavisne varijable mijenja promjenom nezavisnih varijabli. Time se određuje funkcija nezavisnih varijabli, koja se naziva regresijskom funkcijom. Modeli regresije mogu se podijeliti na jednostruku i višestruku regresiju, ovisno o broju nezavisnih varijabli koje su unutar modela. Također se mogu podijeliti na linearne i nelinearne, ovisno o obliku matematičke funkcije kojom je model opisan.
Neka su $x_1, x_2, \ldots, x_K$ kontrolirane (neslučajne) varijable i y slučajna varijabla mjerenja u ovisnosti o $x = (x_1, x_2, \ldots, x_K)$, odnosno $y = y(x)$. Linearni model ovisnosti y o x zadan je sa

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \epsilon, \tag{1.1}$$

gdje je ϵ slučajna greška, a $\alpha, \beta_1, \ldots, \beta_K$ nepoznati parametri modela [8]. Varijablu y nazivamo zavisnom varijablom ili varijablom odgovora, dok $x_1, x_2, \ldots, x_K$ nazivamo nezavisnim varijablama ili regresorima. Ukoliko se linearna regresijska veza između y i $x_1, x_2, \ldots, x_K$ želi utvrditi na osnovi N opažanja, tada se jednadžba (1.1) može zapisati u obliku sustava jednadžbi

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{Ki} + \epsilon_i,\ i = 1, \ldots, N, \tag{1.2}$$

odnosno

$$\begin{aligned} y_1 &= \alpha + \beta_1 x_{11} + \beta_2 x_{21} + \ldots + \beta_k x_{K1} + \epsilon_1 \\ y_2 &= \alpha + \beta_1 x_{12} + \beta_2 x_{22} + \ldots + \beta_k x_{K2} + \epsilon_2 \\ &\vdots \\ y_N &= \alpha + \beta_1 x_{1N} + \beta_2 x_{2N} + \ldots + \beta_k x_{KN} + \epsilon_N. \end{aligned}$$

Gornji sustav možemo zapisati u matričnoj notaciji

$$\boldsymbol{y} = \boldsymbol{x\beta} + \boldsymbol{\epsilon}, \tag{1.3}$$

gdje su

$$\boldsymbol{y} = (y_1, \ldots, y_N)^T,$$

$$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T,$$
$$\boldsymbol{\beta} = (\alpha, \beta_1, \ldots, \beta_K)^T$$

vektori stupci, a $\boldsymbol{x}$ matrica

$$\begin{bmatrix} 1 & x_{11} & \ldots & x_{K1} \\ 1 & x_{12} & \ldots & x_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & \ldots & x_{KN} \end{bmatrix}$$

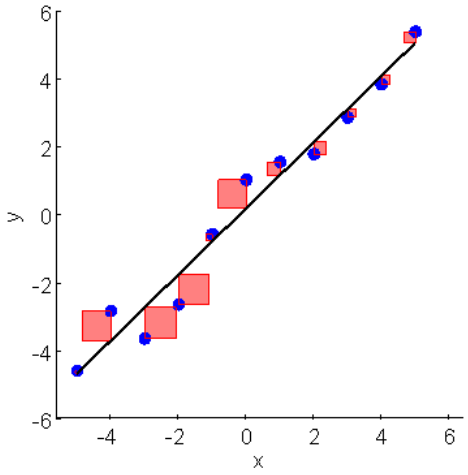
Prirodno se postavlja pitanje kako procijeniti parametre modela. Najčešće korištena metoda je metoda najmanjih kvadrata, koja procjene nepoznatih parametara računa minimizirajući sumu kvadrata reziduala. Rezidual je definiran kao

$$e_i = y_i - \hat{y}_i,$$

gdje je y_i opažena vrijednost od y, a $\hat{y}_i$ vrijednost od y predviđena regresijom. Dakle minimiziramo

$$\boldsymbol{e}^T\boldsymbol{e} = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - \alpha - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki})^2. \tag{1.4}$$

Na sljedećoj slici dan je primjer metode najmanjih kvadrata s jednom nezavisnom varijablom x (izvor: `https://theclevermachine.wordpress.com/`):



Slika 1.1: Primjer metode najmanjih kvadrata

Parcijalnim deriviranjem izraza (1.4) po parametrima modela i izjednačavanjem s nulom dobije se da se minimum postiže u točki

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}, \tag{1.5}$$

gdje se pretpostavlja da postoji inverzna matrica $(\boldsymbol{x}^T\boldsymbol{x})^{-1}$, to jest da matrica $\boldsymbol{x}^T\boldsymbol{x}$ nije singularna. Procijenjeni parametar $\hat{\boldsymbol{\beta}}$ iz (1.5) naziva se procjeniteljem metodom najmanjih kvadrata, odnosno OLS procjeniteljem (*ordinary least squares*). Općenito se procijenjeni parametar $\hat{\beta}_k,\ k = 1, \ldots, K$ interpretira kao promjena vrijednosti zavisne varijable za jedinični porast nezavisne varijable x_k, uz pretpostavku da su ostale nezavisne varijable ostale nepromijenjene. Konstantni član $\hat{\alpha}$ je vrijednost zavisne varijable kada sve nezavisne varijable poprimaju vrijednost nula, koji često nema suvislu interpretaciju.

Kako među svim procjeniteljima želimo odabrati one koji su, po nekim kriterijima, bolji od drugih, izdvojimo neka poželjna svojstva procjenitelja:

- nepristranost procjenitelja
  Kažemo da je $\hat{\beta}_k$ nepristrani procjenitelj parametra β_k ako je $\mathbb{E}(\hat{\beta}_k) = \beta_k$, to jest procjena parametra je u prosjeku jednaka njegovoj stvarnoj vrijednosti.

- efikasnost procjenitelja
  Nepristrani procjenitelj $\hat{\beta}_k$ je efikasan ako u skupu svih nepristranih procjenitelja ima najmanju varijancu, to jest ako je

  $$Var(\hat{\beta}_k) < Var(\hat{\theta}),$$

  za svaki nepristrani procjenitelj $\hat{\theta}$.

- konzistentnost procjenitelja
  Procjenitelj $\hat{\beta}_k$ je konzistentan procjenitelj parametra β_k ako povećanjem uzorka konvergira po vjerojatnosti prema stvarnoj vrijednosti, to jest ako vrijedi

  $$\lim_{N\to\infty} \mathbb{P}(|\hat{\beta}_k - \beta_k| < \epsilon) = 1,\ \forall \epsilon > 0.$$

OLS procjenitelj iz (1.5) je najbolji (u smislu efikasnosti) linearni nepristrani procjenitelj u slučaju da su ispunjene sljedeće pretpostavke modela [7]

- ϵ_i je slučajna greška s očekivanjem 0 i homogenom varijancom:

  $$\mathbb{E}(\epsilon_i) = 0,\ \mathbb{E}(\epsilon_i^2) = \sigma_\epsilon^2,\ i = 1, \ldots, N$$

- greške su međusobno nekorelirane:

$$Cov(\epsilon_i, \epsilon_j) = \mathbb{E}(\epsilon_i \epsilon_j) = 0,\ i \neq j,\ i, j = 1, \ldots, N$$

- greške su nekorelirane s nezavisnim varijablama:

$$Cov(x_{ki}, \epsilon_i) = \mathbb{E}(x_{ki}, \epsilon_i) = 0,\ k = 1, \ldots, K$$

- matrica $\boldsymbol{x}$ je punog ranga:

$$r(\boldsymbol{x}) = K + 1 < N$$

Dodatno se pretpostavlja da su greške normalno distribuirane $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ iz čega slijedi normalna distribuiranost procjenitelja $\hat{\beta}_k \sim N(\beta_k, \sigma_\epsilon^2(\boldsymbol{x}^T\boldsymbol{x})_{kk}^{-1})$, gdje (k, k) predstavlja element na glavnoj dijagonali matrice $(\boldsymbol{x}^T\boldsymbol{x})^{-1}$.

2 Analiza panel podataka

2.1 Vrste podataka

U empirijskim analizama najčešće korišteni podaci su podaci vremenskog niza, vremenskog presjeka i panel podaci.
Vremenski nizovi (*time series*) se sastoje od opažanja jedne ili više varijabli kroz vrijeme. Takvi se podaci mogu prikupljati dnevno (npr. cijene dionica), mjesečno (npr. stopa nezaposlenosti), godišnje (npr. državni proračun), desetogodišnje (npr. popis stanovništva). Podaci vremenskog presjeka (*cross section*) su podaci jedne ili više varijabli prikupljeni u jednoj vremenskoj točki. Ako kombiniramo vremenske nizove s vremenskim presjecima dobivamo združene podatke. Posebna vrsta združenih podataka, u kojima se kroz različite vremenske točke pojavljuju iste vremenski presječne jedinice (iste obitelji, iste države...) nazivaju se panel podaci [9].

Sljedećom tablicom dan je primjer panel podataka:

Tablica 2.1: Primjer panel podataka

Jedinica (i)	Vrijeme (t)	y	x_1	x_2	x_3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1
4	2000	6.1	7.2	5.9	1.9
4	2001	4.2	1.2	9.1	7.1
4	2002	8.8	3.5	5.1	2.4

gdje imamo četiri jedinice promatranja ($i = 1, 2, 3, 4$) kroz tri vremenske točke ($t = 2000, 2001, 2002$). Varijabla y predstavlja zavisnu varijablu, dok x_1, x_2 i x_3 predstavljaju

nezavisne varijable.

Analiza panel podataka svodi se na jednostavnu pretpostavku - veći broj podataka znači više informacija. Upravo to osigurava specifična struktura koja objedinjava vremenski presjek i vremenski niz čime se dobiva združeni vremenski presjek ili specifična struktura panel podataka. Promotrimo jedan primjer. Zanima nas kako izgradnja spalionice otpada utječe na cijenu nekretnina, u ovom slučaju obiteljskih kuća u neposrednoj blizini spalionice. Dakle, uzimamo u obzir cijenu kuća unutar područja spalionice i izvan nje. Unutar područja kuće su znatno jeftinije u odnosu na cijene kuća izvan tog područja. Ključno je pitanje: je li izgradnja spalionice doista u tom omjeru utjecala na cijenu nekretnina? Sljedeća regresijska jednadžba nam daje prosječne cijene kuća ($\hat{c}$) izvan i unutar područja spalionice:

$$\hat{c} = 101.307 - 30.688 \times S, \tag{2.1}$$

gdje je S pomoćna (*dummy*) varijabla koja poprima vrijednost 1 ako je kuća unutar područja, a 0 inače. Vidimo da prosječna cijena kuća izvan područja ($S = 0$) iznosi 101.307 eura, a cijene kuća unutar područja su za 30.688 eura niže od ostalih. Pitamo se da li tih 30.688 eura zaista reflektira stvarni efekt spalionice? Pretpostavimo da imamo cijene kuća godinu dana prije izgradnje spalionice na tom području:

$$\hat{c} = 82.517 - 18.824 \times S. \tag{2.2}$$

S je sada hipotetska vrijednost koja na neki način obuhvaća lokaciju kuća u blizini potencijalne spalionice (nazovimo to područjem A). Vidimo da su cijene kuća unutar područja A već za 18.824 eura niže od ostalih, što znači da ta lokacija već prije nije bila poželjna. Dakle, stvarni efekt spalionice nije -30.688 eura, nego $-30.688 - (-18.824) = -11.864$ eura. Ovim smo združili podatke za dvije vremenske točke i dobili stvarni efekt spalionice na cijene kuća.

Spomenimo još da prema kriteriju raspoloživosti podataka razlikujemo balansirane i nebalansirane panel podatke. U slučaju balansiranih panel podataka svaka jedinica promatranja ima isti broj opservacija vremenskih nizova, odnosno vremenski nizovi su iste duljine. Ukoliko se broj opservacija razlikuje od jedne do druge jedinice promatranja, onda se radi o nebalansiranim panel podacima [3]. Tablicom 2.1 dan je primjer balansiranih panel podataka. Njima ćemo se baviti u nastavku rada.

2.2 Linearni modeli panel podataka

U općem obliku linearnog modela panel podataka, promjene u zavisnoj varijabli, y, objašnjene su promjenama K nezavisnih varijabli, $x_1, x_2, \ldots, x_K$, i slučajnim promjenama kojima se

obuhvaća djelovanje drugih varijabli koje nisu eksplicitno uključene u model. Opći oblik linearnog panel modela glasi

$$y_{it} = \alpha_{it} + \beta_{1,it}x_{1,it} + \beta_{2,it}x_{2,it} + \ldots + \beta_{K,it}x_{K,it} + \epsilon_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{2.3}$$

gdje je

- N broj jedinica promatranja, T broj vremenskih točaka,
- y_{it} vrijednost zavisne varijable y za i-tu jedinicu promatranja u trenutku t,
- $x_{k,it}$ vrijednost nezavisne varijable x_k za i-tu jedinicu promatranja u trenutku t,
- α_{it} slobodni član za i-tu jedinicu promatranja u trenutku t,
- $\beta_{k,it}$ nepoznati regresijski parametar k-te nezavisne varijable za i-tu jedinicu promatranja u trenutku t
- ϵ_{it} slučajna greška s očekivanjem 0 i varijancom σ^2_ϵ za sve jedinice promatranja i i za svaki vremenski trenutak t.

Ovaj model podrazumijeva da za svaku jedinicu promatranja, i, postoji različita reakcija zavisne varijable na promjene u nezavisnim varijablama i da se ta reakcija razlikuje za svaku vremensku točku t. Prema tome, regresijski parametar svake jedinice promatranja je specifičan za svaku vremensku točku. Ovakav se model ne može procijeniti jer broj nepoznatih parametara, $NT(K + 1)$, nadmašuje broj podataka u uzorku, NT. Zbog toga uvodimo određene pretpostavke modela. Za početak pretpostavljamo da su regresijski parametri uz nezavisne varijable konstantni za svaku jedinicu promatranja u svakom vremenskom trenutku, to jest da je $\beta_{k,it} = \beta_k$, za sve i i t, dok su slobodni članovi varijabilni. Heterogenost između jedinica promatranja i vremenskih točaka obuhvaćena je pomoću efekata varijabli koje nisu eksplicitno uključene u model (α_{it}). Takve varijable su ili vremenski nepromjenjive (npr. spol, nacionalnost) ili nepromjenjive po jedinicama promatranja (npr. cijena proizvoda). Ako se pretpostavi da se slobodni članovi razlikuju samo po jedinicama promatranja, procjenjuje se model sljedećeg oblika

$$y_{it} = \alpha + \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \ldots + \beta_K x_{K,it} + \epsilon_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{2.4}$$

gdje je $\alpha + \alpha_i$ slobodni član i-te jedinice promatranja, α prosječna vrijednost slobodnog člana, dok α_i predstavlja veličinu odstupanja slobodnog člana od prosjeka α. Vrijednost α_i naziva se individualni efekt ili neuočena heterogenost, pomoću kojeg su u model uključene

heterogenosti između jedinica promatranja. Zanemarivanje heterogenosti može voditi nekonzistentnim procjenama i pogrešnom statističkom zaključivanju [2].
Procjena utjecaja regresora na zavisnu varijablu može se vršiti pomoću tri modela: združeni model, model fiksnih efekata i model slučajnih efekata [2].

Združeni model (*Pooled OLS model*)

Jedno od ograničenja na parametre modela (2.4) zasniva se na pretpostavci da su svi parametri modela konstantni, odnosno za sve i i t pretpostavlja se konstantan utjecaj nezavisnih varijabli na zavisnu. Takav model zanemaruje činjenicu da se radi o panel podacima te sve podatke združi u NT podataka vremenskog presjeka. Model je dan sljedećim izrazom

$$y_{it} = \alpha + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \ldots + \beta_K x_{K,it} + \epsilon_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{2.5}$$

ili matrično

$$\boldsymbol{y} = \boldsymbol{x\beta} + \boldsymbol{\epsilon}, \tag{2.6}$$

gdje je

- $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ $NT \times 1$ vektor vrijednosti zavisne varijable,
- $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T$ $NT \times 1$ vektor vrijednosti slučajne greške,
- $\boldsymbol{\beta} = (\alpha, \beta_1, \ldots, \beta_K)^T$ $(K+1) \times 1$ vektor parametara modela,

uz oznake $y_i = (y_{i1}, \ldots, y_{iT})^T$, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})^T$, $i = 1, \ldots, N$.

$\boldsymbol{x}$ je $NT \times (K+1)$ matrica

$$\begin{bmatrix} 1 & x_{1,11} & \ldots & x_{K,11} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,1T} & \ldots & x_{K,1T} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,N1} & \ldots & x_{K,N1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,NT} & \ldots & x_{K,NT} \end{bmatrix}$$

Vidimo da je združeni model isto što i klasični linearni model. Dakle, ovaj model ne uzima u obzir heterogenosti između jedinica promatranja ($\alpha_i = 0$), što bi značilo da sve

jedinice reagiraju na isti način.

Uz pretpostavku da nema heterogenosti između jedinica promatranja i standardne pretpostavke klasičnog linearnog modela:

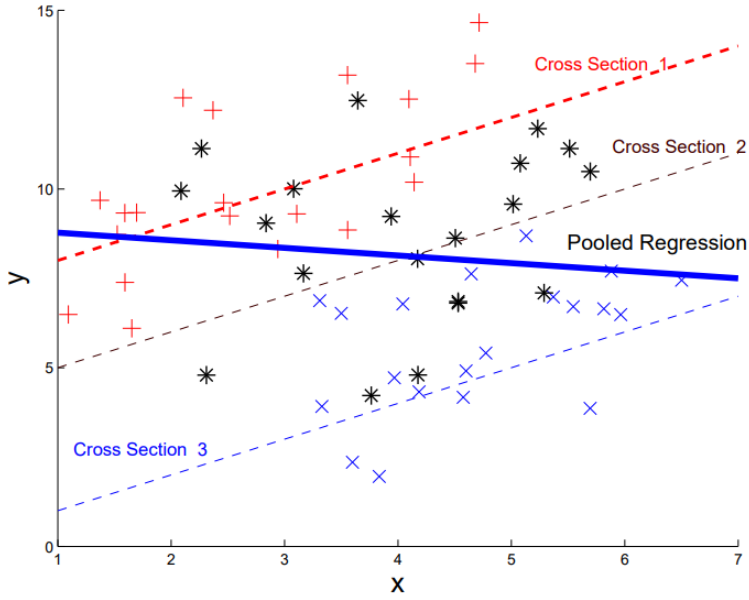
- $\mathbb{E}(\epsilon_{it}) = 0$, $Var(\epsilon_{it}) = \sigma_{\epsilon}^2$,
- $Cov(\epsilon_{it}, \epsilon_{js}) = \mathbb{E}(\epsilon_{it}\epsilon_{js}) = 0$,
- $Cov(x_{k,it}, \epsilon_{it}) = \mathbb{E}(x_{k,it}\epsilon_{it}) = 0$,
- $r(\boldsymbol{x}) = K + 1 < NT$

$$\hat{\boldsymbol{\beta}}_{združeni} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y} = \Big(\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}})^T\Big)^{-1}\Big(\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}})(y_{it} - \bar{y})\Big) \tag{2.7}$$

je najbolji linearni nepristrani procjenitelj, gdje je

$$\boldsymbol{x}_{it} = (x_{1,it}, \ldots, x_{K,it}), \quad \bar{\boldsymbol{x}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\boldsymbol{x}_{it}, \quad \bar{y} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}y_{it}.$$

Na slici je dan primjer krive upotrebe združenog modela, to jest zanemarivanje postojeće heterogenosti između jedinica promatranja [6]



Slika 2.1: Postojanje heterogenosti između jedinica promatranja

Model fiksnih efekata (*Fixed effects model*)

U modelu fiksnih efekata, individualni efekti direktno su uključeni u model kao fiksni parametri i to preko varijabilnih slobodnih članova. Jednadžba modela glasi

$$y_{it} = \alpha + \alpha_i + \beta_1 x_{1,it} + \ldots + \beta_K x_{K,it} + \epsilon_{it}, \; i = 1, \ldots, N, t = 1, \ldots, T, \tag{2.8}$$

gdje je ϵ_{it} slučajna greška koja je normalno distribuirana, s očekivanjem $\mathbb{E}(\epsilon_{it}) = 0$ i varijancom $Var(\epsilon_{it}) = \sigma_\epsilon^2$. Osim toga, pretpostavlja se nekoreliranost grešaka, to jest

$$Cov(\epsilon_{it}, \epsilon_{js}) = \mathbb{E}(\epsilon_{it}\epsilon_{js}) = 0$$

i nekoreliranost grešaka i nezavisnih varijabli

$$Cov(x_{k,it}, \epsilon_{it}) = \mathbb{E}(x_{k,it}\epsilon_{it}) = 0.$$

Model fiksnih efekata dopušta korelaciju individualnih efekata i nezavisnih varijabli

$$Cov(\alpha_i, x_{k,it}) = \mathbb{E}(\alpha_i x_{k,it}) \neq 0,$$

zbog čega bi OLS procjenitelj modela (2.8) bio pristran i nekonzistentan.

Postoji nekoliko načina za procjenu modela fiksnih efekata: korištenjem pomoćnih varijabli, transformacijom podataka unutar jedinica promatranja te metodom prvih razlika [5].

Model s pomoćnim varijablama (LSDV model)

U model (2.8) za jedinice promatranja i dodamo po jednu pomoćnu varijablu koja sadrži efekt specifičan za određenu jedinicu promatranja. Sada (2.8) poprima oblik

$$y_{it} = \alpha + \alpha_1 D_{1i} + \ldots + \alpha_{N-1} D_{(N-1)i} + \beta_1 x_{1,it} + \ldots + \beta_K x_{K,it} + \epsilon_{it},\ i = 1, \ldots, N, t = 1, \ldots, T, \quad (2.9)$$

gdje su D_j pomoćne varijable

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

Prethodni model se naziva LSDV model (*least squares dummy variable*). Da bi se izbjegao problem multikolinearnosti, eliminira se jedna pomoćna varijabla (umjesto N ima ih $N-1$), koja se koristi kao referentna jedinica, odnosno ona s kojom se ostale jedinice promatranja uspoređuju. Značenje parametara α_j uz pomoćnu varijablu jest udaljenost od intercepta α koji predstavlja parametar referentne jedinice. OLS procjenitelj modela (2.9) naziva se LSDV procjeniteljem. Jedan od problema koji nastaje prilikom procjene ovog modela odnosi se na procjenu velikog broja parametara čime se gubi velik broj stupnjeva slobode. Ako je broj jedinica promatranja velik, to jest N teži u beskonačnost, tada i broj koeficijenata uz pomoćne varijable teži u beskonačnost pa te procjene nisu konzistentne. Međutim, procjene za β_k su i dalje konzistentne [3].

Spomenimo da postoje još dva pristupa LSDV modelu. Jedan od njih ne obuhvaća konstantu α, nego uvodi N pomoćnih varijabli za svaku jedinicu promatranja, dok drugi sadrži i konstantu α i N pomoćnih varijabli, ali postavlja ograničenje $\sum_{i=1}^{N} \alpha_i = 0$ [4].

Model „unutar grupa" (*Within group model*)

Da bi se izbjeglo uvođenje pomoćnih varijabli, koristi se takozvana transformacija unutar grupa koja uklanja individualne efekte, za koje se pretpostavlja da su korelirani s nekim nezavisnim varijablama. Transformacija se vrši na sljedeći način:

- uprosječi se originalna jednadžba (2.8) po vremenu

$$\bar{y}_i = \bar{\alpha} + \bar{\alpha}_i + \beta_1 \bar{x}_{1i} + \ldots + \beta_K \bar{x}_{Ki} + \bar{\epsilon}_i,\ i = 1, \ldots, N, \quad (2.10)$$

  gdje je

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it},\ \bar{x}_{ki} = \frac{1}{T}\sum_{t=1}^{T} x_{k,it},\ \bar{\epsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it},\ \bar{\alpha} = \alpha,\ \bar{\alpha}_i = \alpha_i$$

- oduzme se (2.10) od (2.8), to jest centriraju se podaci (*demean data*)

$$y_{it} - \bar{y}_i = (\alpha - \alpha) + (\alpha_i - \alpha_i) + \beta_1(x_{1,it} - \bar{x}_{1i}) + \ldots + \beta_K(x_{K,it} - \bar{x}_{Ki}) + (\epsilon_{it} - \bar{\epsilon}_i) \quad (2.11)$$

Označi li se s $\ddot{y}_{it} = y_{it} - \bar{y}_i$, $\ddot{x}_{k,it} = x_{k,it} - \bar{x}_{ki}$ i $\ddot{\epsilon}_{it} = \epsilon_{it} - \bar{\epsilon}_i$, prethodni model može se kraće zapisati kao

$$\ddot{y}_{it} = \beta_1\ddot{x}_{1,it} + \beta_2\ddot{x}_{2,it} + \ldots + \beta_K\ddot{x}_{K,it} + \ddot{\epsilon}_{it}. \quad (2.12)$$

Iz zadnje jednadžbe vidimo da su uklonjeni individualni efekti. Novi podaci predstavljaju odstupanja pojedinih vrijednosti od aritmetičke sredine i-te jedinice promatranja. Provjerimo što vrijedi za model unutar grupa (2.12):

$$\mathbb{E}(\ddot{\epsilon}_{it}) = \mathbb{E}(\epsilon_{it} - \bar{\epsilon}_i) = 0 - 0 = 0,$$

$$Var(\ddot{\epsilon}_{it}) = \mathbb{E}(\ddot{\epsilon}_{it}^2) = \mathbb{E}[(\epsilon_{it} - \bar{\epsilon}_i)^2] = \sigma_\epsilon^2\left(1 - \frac{1}{T}\right),$$

$$Cov(\ddot{x}_{k,it}, \ddot{\epsilon}_{it}) = \mathbb{E}(\ddot{x}_{k,it}\ddot{\epsilon}_{it}) = \underbrace{\mathbb{E}(x_{k,it}\epsilon_{it})}_{=0} - \mathbb{E}(x_{k,it}\bar{\epsilon}_i) - \mathbb{E}(\bar{x}_{ki}\epsilon_{it}) + \mathbb{E}(\bar{x}_{ki}\bar{\epsilon}_i).$$

Da bi prethodni izraz bio jednak nuli, trebamo pretpostaviti

$$\mathbb{E}(x_{k,is}\epsilon_{it}) = 0, \ \forall t, s = 1, \ldots, T, \quad (2.13)$$

odnosno u modelu (2.8) trebamo dodatno pretpostaviti nekoreliranost grešaka i nezavisnih varijabli unutar svake jedinice promatranja i.

Za $i \neq j$ i $t \neq s$ iz $Cov(\epsilon_{it}, \epsilon_{js}) = 0$, slijedi

$$Cov(\ddot{\epsilon}_{it}, \ddot{\epsilon}_{js}) = 0.$$

Međutim, za istu jedinicu promatranja greške su korelirane:

$$Cov(\ddot{\epsilon}_{it}, \ddot{\epsilon}_{is}) = \mathbb{E}(\ddot{\epsilon}_{it}\ddot{\epsilon}_{is}) = \mathbb{E}(\epsilon_{it}\epsilon_{is}) - \mathbb{E}(\epsilon_{it}\bar{\epsilon}_i) - \mathbb{E}(\epsilon_{is}\bar{\epsilon}_i) + \mathbb{E}(\bar{\epsilon}_i^2) = 0 - \frac{1}{T}\sigma_\epsilon^2 - \frac{1}{T}\sigma_\epsilon^2 + \frac{1}{T}\sigma_\epsilon^2 = -\frac{1}{T}\sigma_\epsilon^2$$

pa procjenitelj dobiven metodom najmanjih kvadrata modela (2.12) neće biti efikasan.

Uz uvjet (2.13) i puni rang matrice $\ddot{\boldsymbol{x}}$ (koju analogno definiramo kao matricu $\boldsymbol{x}$, samo bez prvog stupca popunjenog jedinicama) metoda najmanjih kvadrata primijenjena na model (2.12) daje nam nepristrane i konzistentne procjenitelje $\hat{\beta}_1, \ldots, \hat{\beta}_K$ [3]. Dobiveni procjenitelj $\hat{\boldsymbol{\beta}}_{FE} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)$ naziva se procjeniteljem unutar grupa i jednak je onom kojeg

daje i LSDV metoda. Ukoliko želimo procijeniti i individualne efekte, koristimo sljedeću formulu

$$\hat{\alpha}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{1i} - \ldots - \hat{\beta}_K \bar{x}_{Ki}.$$

Procjenitelji $\hat{\alpha}_i$ su nepristrani, a konzistentni jedino u slučaju kad broj vremenskih točaka T teži u beskonačnost.

Procjenitelj unutar grupa dan je formulom

$$\hat{\boldsymbol{\beta}}_{FE} = \Big(\sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)^T \Big)^{-1} \Big(\sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)(y_{it} - \bar{y}_i) \Big), \tag{2.14}$$

gdje je $\boldsymbol{x}_{it} = (x_{1,it}, x_{2,it}, \ldots, x_{K,it})$ i $\bar{\boldsymbol{x}}_i = (\bar{x}_{1i}, \bar{x}_{2i}, \ldots, \bar{x}_{Ki})$.

Napomenimo još da se jednadžba (2.10) naziva modelom između grupa (*between group model*). Primijenimo li metodu najmanjih kvadrata, dobiveni procjenitelj nazivamo procjeniteljem između grupa. Za razliku od procjenitelja unutar grupa, on može procijeniti utjecaj varijabli koje nisu promjenjive kroz vrijeme, ali se gubi na preciznosti rezultata jer se uprosječivanjem podataka gubi vremenska komponenta. Također daje procjene i za individualne efekte, ali nije konzistentan kad su α_i korelirani s regresorima [3].

Model „prvih razlika" (*First difference model*)

Još jedan način uklanjanja individualnih efekata jest pomoću metode prvih razlika. Za istu jedinicu promatranja i, zapišimo model u dvije susjedne vremenske točke

$$y_{it} = \alpha + \alpha_i + \beta_1 x_{1,it} + \ldots + \beta_K x_{K,it} + \epsilon_{it},$$
$$y_{i(t-1)} = \alpha + \alpha_i + \beta_1 x_{1,i(t-1)} + \ldots + \beta_K x_{K,i(t-1)} + \epsilon_{i(t-1)}.$$

Oduzmemo li prethodne dvije jednadžbe dobijemo

$$y_{it} - y_{i(t-1)} = \beta_1 (x_{1,it} - x_{1,i(t-1)}) + \ldots + \beta_K (x_{K,it} - x_{K,i(t-1)}) + (\epsilon_{it} - \epsilon_{i(t-1)}) \tag{2.15}$$

ili kraće

$$\Delta y_{it} = \beta_1 \Delta x_{1,it} + \beta_2 \Delta x_{2,it} + \ldots + \beta_K \Delta x_{K,it} + \Delta \epsilon_{it}, \tag{2.16}$$

gdje je $\Delta y_{it} = y_{it} - y_{i(t-1)}$, $\Delta x_{k,it} = x_{k,it} - x_{k,i(t-1)}$, $\Delta \epsilon_{it} = \epsilon_{it} - \epsilon_{i(t-1)}$.

Uočimo da je sada broj opservacija $N(T-1)$. Model (2.16) nazivamo modelom prvih razlika, a za njega vrijedi:

$$\mathbb{E}(\Delta\epsilon_{it}) = \mathbb{E}(\epsilon_{it}) - \mathbb{E}(\epsilon_{i(t-1)}) = 0,$$

$$Var(\Delta\epsilon_{it}) = Var(\epsilon_{it}) + Var(\epsilon_{i(t-1)}) - 2\underbrace{Cov(\epsilon_{it}, \epsilon_{i(t-1)})}_{=0} = 2\sigma_\epsilon^2,$$

$$Cov(\Delta x_{k,it}, \Delta\epsilon_{it}) = \mathbb{E}(\Delta x_{k,it}\Delta\epsilon_{it}) = \underbrace{\mathbb{E}(x_{k,it}\epsilon_{it})}_{=0} - \mathbb{E}(x_{k,it}\epsilon_{i(t-1)}) - \mathbb{E}(x_{k,i(t-1)}\epsilon_{it}) + \underbrace{\mathbb{E}(x_{k,i(t-1)}\epsilon_{i(t-1)})}_{=0}.$$

Da bi vrijedio uvjet $Cov(\Delta x_{k,it}, \Delta\epsilon_{it}) = 0$, u polaznom modelu (2.8) moramo dodatno pretpostaviti

$$\mathbb{E}(x_{k,is}\epsilon_{it}) = 0, \tag{2.17}$$

za susjedne vremenske točke t i s. Nekoreliranost grešaka neće vrijediti zbog

$$Cov(\Delta\epsilon_{it}, \Delta\epsilon_{i(t-1)}) = \mathbb{E}(\Delta\epsilon_{it}\Delta\epsilon_{i(t-1)}) = -\sigma_\epsilon^2 \neq 0,$$

što znači da OLS procjenitelj modela (2.16) neće biti efikasan.

Uz uvjet (2.17) i puni rang matrice $\Delta\boldsymbol{x}$ (koju analogno definiramo kao matricu $\boldsymbol{x}$, samo bez prvog stupca popunjenog jedinicama) metoda najmanjih kvadrata primijenjena na model (2.16) daje nam nepristrane i konzistentne procjenitelje $\hat{\beta}_1, \ldots, \hat{\beta}_K$ [3]. Dobiveni procjenitelj $\hat{\boldsymbol{\beta}}_{FD} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)$ naziva se procjeniteljem prvih razlika, a dan je formulom

$$\hat{\boldsymbol{\beta}}_{FD} = \Big(\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta\boldsymbol{x}_{it}\Delta\boldsymbol{x}_{it}^T\Big)^{-1}\Big(\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta\boldsymbol{x}_{it}\Delta y_{it}\Big), \tag{2.18}$$

gdje je $\Delta\boldsymbol{x}_{it} = (\Delta x_{1,it}, \Delta x_{2,it}, \ldots, \Delta x_{K,it})$.

Lako se pokaže da je u slučaju $T = 2$ procjenitelj unutar grupa jednak procjenitelju prvih razlika, a za $T > 2$ procjenitelj unutar grupa je efikasniji od procjenitelja prvih razlika.

Glavni nedostatak modela fiksnih efekata je taj što briše sve vremenski nepromjenjive varijable, budući da su one konstantne kroz vrijeme za svaku jedinicu promatranja (npr. spol, religija). Dakle, vremenski nepromjenjive varijable ne mogu biti nezavisne varijable, inače bi se pojavio problem multikolinearnosti, to jest matrica vrijednosti nezavisnih varijabli ne bi bila punog ranga (imali bismo stupce popunjene nulama).

Model slučajnih efekata (*Random effects model*)

Model slučajnih efekata pretpostavlja da je individualni efekt α_i slučajna varijabla, odnosno dio slučajne greške

$$y_{it} = \alpha + \beta_1 x_{1,it} + \ldots + \beta_K x_{K,it} + v_{it},\ v_{it} = \alpha_i + \epsilon_{it},\ i = 1, \ldots, N, t = 1, \ldots, T, \tag{2.19}$$

gdje su α_i i ϵ_{it} normalno distribuirani s očekivanjem $\mathbb{E}(\alpha_i) = \mathbb{E}(\epsilon_{it}) = 0$ i varijancama $Var(\alpha_i) = \sigma_\alpha^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$. Zbog specifične dekompozicije slučajne greške v_{it}, isti model još se naziva i model komponenata slučajne greške (*error components model*). Ovaj model pretpostavlja [1]

nekoreliranost komponenata greške

$$\mathbb{E}(\alpha_i \alpha_j) = \mathbb{E}(\epsilon_{it}\epsilon_{js}) = \mathbb{E}(\alpha_i \epsilon_{it}) = 0,$$

te njihovu nekoreliranost s nezavisnim varijablama

$$Cov(\alpha_i, x_{k,it}) = 0 \text{ i } Cov(\epsilon_{is}, x_{k,it}) = 0.$$

Iz prethodnog slijedi da je $\mathbb{E}(v_{it}) = 0,\ Var(v_{it}) = \sigma_\alpha^2 + \sigma_\epsilon^2$, $Cov(v_{it}, x_{k,it}) = 0$ i

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

Naglasimo da za razliku od modela fiksnih efekata, model slučajnih efekata ne dopušta da individualni efekti budu u korelaciji s regresorima, to jest $Cov(\alpha_i, x_{k,it}) = 0$. Kako sada greške sadrže vremenski nepromjenjivu komponentu, α_i, one postaju korelirane pa OLS procjenitelj nije efikasan. Rješenje je poopćena metoda najmanjih kvadrata, to jest GLS metoda (*generalized least squares*) koja eliminira korelaciju među greškama iste jedinice promatranja i primjenjuje metodu najmanjih kvadrata na novi model:

- prvo se definira parametar λ (tzv. ponder, *weight*)

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

- originalna jednadžba se uprosječi po vremenu

$$\bar{y}_i = \alpha + \beta_1 \bar{x}_{1i} + \ldots + \beta_K \bar{x}_{Ki} + \bar{v}_i,\ i = 1, \ldots, N \tag{2.20}$$

- pomnoži se (2.20) s λ i oduzme se od (2.19) (kvazi-centriraju se podaci)

$$y_{it} - \lambda\bar{y}_i = \alpha(1-\lambda) + \beta_1(x_{1,it} - \lambda\bar{x}_{1i}) + \ldots + \beta_K(x_{K,it} - \lambda\bar{x}_{Ki}) + (v_{it} - \lambda\bar{v}_i) \quad (2.21)$$

Prethodni model (2.21) se procijeni metodom najmanjih kvadrata. Procjenitelj tog modela nazivamo GLS procjeniteljem ili procjeniteljem slučajnih efekata, a dan je formulom

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{RE} = \Big(\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)^T + \lambda \sum_{i=1}^{N} T(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^T \Big)^{-1} \times \\ \Big(\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)(y_{it} - \bar{y}_i) + \lambda \sum_{i=1}^{N} T(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{y}_i - \bar{y}) \Big).\end{aligned} \quad (2.22)$$

Uočimo da je za $\lambda = 0$ dobiveni procjenitelj jednak procjenitelju fiksnih efekata, dok je za $\lambda = 1$ jednak procjenitelju združenog modela [1].

Kako u praksi λ često nije poznat jer se sastoji od teorijskih varijanci, koristimo se dvostupanjskom poopćenom metodom najmanjih kvadrata, to jest FGLS metodom (*feasible generalized least squares*):

- za procjenu nepoznatih varijanci koristimo se rezidualima dobivenima iz združenog modela pa je procijenjeni λ dan s

$$\hat{\lambda} = 1 - \sqrt{\frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_\epsilon^2 + T\hat{\sigma}_\alpha^2}}$$

- u (2.21) λ zamijenimo s $\hat{\lambda}$ te primijenimo GLS metodu

Procjenitelj takvog modela nazivamo FGLS procjeniteljem.

Napomenimo da je bez obzira na uvjet $Cov(\alpha_i, x_{k,it}) = 0$, $\hat{\boldsymbol{\beta}}_{FE}$ uvijek konzistentan, dok u slučaju $Cov(\alpha_i, x_{k,it}) \neq 0$, $\hat{\boldsymbol{\beta}}_{RE}$ nije konzistentan. Ako $T \to \infty$, $N \to \infty$ i vrijedi $Cov(\alpha_i, x_{k,it}) = 0$, onda je $\hat{\boldsymbol{\beta}}_{RE}$ najbolji linearni nepristrani procjenitelj.

Kod procjene modela (2.19) mogli smo se koristiti spomenutim modelom između grupa:

$$\bar{y}_i = \alpha + \beta_1\bar{x}_{1i} + \ldots + \beta_K\bar{x}_{Ki} + \bar{v}_i, \; i = 1, \ldots, N, \quad (2.23)$$

gdje metodom najmanjih kvadrata dobijemo procjenitelj između grupa $\hat{\boldsymbol{\beta}}_{BE}$, koji je nepristran i konzistentan, ali nije efikasan

$$\hat{\boldsymbol{\beta}}_{BE} = \Big(\sum_{i=1}^{N}(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^T \Big)^{-1} \Big(\sum_{i=1}^{N}(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{y}_i - \bar{y}) \Big). \quad (2.24)$$

Sada vidimo da je procjenitelj slučajnih efekata $\hat{\boldsymbol{\beta}}_{RE}$ zapravo pondenirani prosjek (*weighted average*) procjenitelja fiksnih efekata i procjenitelja između grupa [1].

Prednost korištenja modela slučajnih efekata je u tome što čuva broj stupnjeva slobode budući da je jedini uključeni parametar varijanca individualnog efekta i u tome što dopušta da nezavisne varijable budu vremenski nepromjenjive. Glavni nedostatak tog modela je da u slučaju koreliranosti individualnih efekata i nezavisnih varijabli daje pristrane i nekonzistentne procjene.

2.3 Testovi za odabir modela

Kako bi se odabrao adekvatan model, potrebno je testirati postojanje individualnih efekata.

$F-$ test

Pomoću $F-$ testa možemo utvrditi postoje li fiksni (individualni) efekti, odnosno je li prikladniji model fiksnih efekata (LSDV model) ili združeni model. Dakle, testiramo

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_{N-1} = 0,$$

$$H_1 : \text{ barem jedan } \alpha_i \text{ je različit od } 0.$$

Statistika na osnovu koje se testira je dana sa

$$F = \frac{(SSE_{združeni} - SSE_{LSDV})/(N-1)}{(SSE_{LSDV})/(NT-N-K)}, \tag{2.25}$$

gdje je $SSE_{združeni}$ suma kvadrata reziduala združenog modela, a SSE_{LSDV} suma kvadrata reziduala LSDV modela. Ako je H_0 istinita, F statistika ima F distribuciju s $N-1$ i $NT-N-K$ stupnjeva slobode. Ukoliko testom utvrdimo da se H_0 ne odbacuje, to jest vrijednost F testa je manja od $F(N-1, NT-N-K)$ za danu razinu značajnosti, zaključujemo da je prikladnije koristiti združeni model. U suprotnom je prikladniji model fiksnih efekata.

Breusch-Paganov test

Opravdanost korištenja modela slučajnih efekata može se testirati pomoću testa koji polazi od Lagrangeovog multiplikatora. Jedan od njih je modificirani Breusch-Paganov *LM* test za testiranje postojanja individualnih efekata. Originalni Breusch-Paganov test koristi

se za otkrivanje heteroskedastičnosti slučajne greške. Znamo da je pretpostavka modela slučajnih efekata $\mathbb{E}(\alpha_i) = 0$ i $Var(\alpha_i) = \sigma_\alpha^2 > 0$. Test se definira sljedećim hipotezama:

$$H_0 : \sigma_\alpha^2 = 0,$$
$$H_1 : \sigma_\alpha^2 \neq 0.$$

Breusch-Paganova *LM* statistika glasi

$$LM = \frac{NT}{2(T-1)}\left[\frac{\sum_{i=1}^{N}(\sum_{t=1}^{T} e_{it})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2} - 1\right]^2, \tag{2.26}$$

gdje su e_{it} reziduali združenog modela. Ako je H_0 istinita, *LM* statistika ima χ^2 distribuciju s jednim stupnjem slobode. Ukoliko je vrijednost *LM* testa veća od $\chi^2(1)$ za danu razinu značajnosti, odbacujemo H_0, što znači da je model slučajnih efekata prikladan za procjenu parametara. Ukoliko se H_0 ne odbacuje, nema heterogenosti između jedinica promatranja pa je združeni model prikladniji.

Hausmanov test

Kada se prethodnim testovima utvrdi postojanje individualnih efekata α_i, postavlja se pitanje izbora modela panel podataka. Priroda takvih efekata (fiksna ili slučajna) testira se Hausmanovim testom. Hausmanov test definira se sljedećim hipotezama:

$$H_0 : Cov(\alpha_i, x_{k,it}) = 0, \ \forall k = 1, \ldots, K,$$
$$H_1 : Cov(\alpha_i, x_{k,it}) \neq 0, \text{ za neko } k.$$

Vrijednost Hausmanove statistike H računamo pomoću formule

$$H = (\hat{\boldsymbol{\beta}}_{RE} - \hat{\boldsymbol{\beta}}_{FE})^T [Var(\hat{\boldsymbol{\beta}}_{RE}) - Var(\hat{\boldsymbol{\beta}}_{FE})]^{-1} (\hat{\boldsymbol{\beta}}_{RE} - \hat{\boldsymbol{\beta}}_{FE}). \tag{2.27}$$

Vektor $\hat{\boldsymbol{\beta}}_{FE}$ je procijenjen modelom unutar grupa, što znači da ne sadrži konstantni član za svaku jedinicu promatranja, a vektor $\hat{\boldsymbol{\beta}}_{RE}$ je procijenjen GLS metodom i sadrži procijenjene parametre uz nezavisne varijable. Ukoliko je H_0 istinita, H statistika ima χ^2 distribuciju s K stupnjeva slobode, gdje je K broj procijenjenih parametara modela. Ako je vrijednost H statistike manja od $\chi^2(K)$ za zadanu razinu značajnosti, onda ne odbacujemo H_0 i zaključujemo da je model slučajnih efekata prikladniji od modela fiksnih efekata.

Sljedeća tablica nam daje sažeti prikaz za odabir pravog modela [4]

Tablica 2.2: Odabir modela

FE model ($F-$test)	RE model (LM test)	Naš odabir
H_0 nije odbačena	H_0 nije odbačena	Združeni model
H_0 odbačena	H_0 nije odbačena	FE model
H_0 nije odbačena	H_0 odbačena	RE model
H_0 odbačena	H_0 odbačena	Hausmanov test (FE ili RE)

3 Primjer u SAS-u

Nakon teorijskog dijela rada, u ovom poglavlju dat ćemo primjer analize panel podataka koristeći statistički program SAS. Podaci koji će se koristiti za analizu preuzeti su na stranici Eurostata `https://ec.europa.eu/eurostat/data/database`.

Kako je iseljavanje radno sposobnog stanovništva i obitelji društveni problem od velikog značaja, kao zavisna varijabla odabran je broj iseljenih, koji ćemo promatrati u 5 europskih država: Grčka, Hrvatska, Mađarska, Rumunjska i Slovenija, kroz vremensko razdoblje od 2008. godine do 2016. godine. Za polazišnu vremensku točku odabrana je 2008. godina, kao prijelomna godina svjetske gospodarske krize, a kao nezavisne varijable:

- minimalna mjesečna plaća u eurima
- postotak nezaposlenog stanovništva unutar skupine radno aktivnog stanovništva
- postotak stanovništva sa završenim fakultetom (od 15. do 64. godine)
- broj rastava na 100 brakova

Procjena utjecaja nezavisnih varijabli na zavisnu izvršit će se primjenom tri modela koja su obrađena u teorijskom dijelu (združeni model, model fiksnih efekata i model slučajnih efekata).

Tablica 3.1: Prikupljeni podaci: ispis iz SAS-a

Obs	drzava	godina	broj_iseljenih	broj_rast_100brakova	min_mj_pl	nezap_p_ac	zavrseni_faks_p
1	Grcka	1	43,044	24.6	794.02	7.8	19.8
2	Grcka	2	43,686	23.0	862.82	9.6	19.9
3	Grcka	3	62,041	23.6	862.82	12.7	20.9
4	Grcka	4	92,404	23.1	876.62	17.9	22.2
5	Grcka	5	124,694	29.9	683.76	24.5	22.9
6	Grcka	6	117,094	32.6	683.76	27.5	24.0
7	Grcka	7	106,804	27.2	683.76	26.5	24.6
8	Grcka	8	109,351	29.1	683.76	24.9	25.4
9	Grcka	9	106,535	22.2	683.76	23.6	26.4
10	Hrvatska	1	10,638	21.5	379.60	8.6	13.6
11	Hrvatska	2	12,355	22.7	386.91	9.3	14.5
12	Hrvatska	3	13,017	23.8	390.94	11.8	15.7
13	Hrvatska	4	12,699	28.0	380.18	13.7	15.4
14	Hrvatska	5	12,877	27.8	374.31	15.8	15.8
15	Hrvatska	6	15,262	31.3	400.67	17.4	17.0
16	Hrvatska	7	20,858	31.5	398.31	17.2	18.5
17	Hrvatska	8	29,651	30.3	399.05	16.1	19.7
18	Hrvatska	9	36,436	34.4	414.45	13.4	20.0
19	Madarska	1	9,591	62.7	293.08	7.8	16.4
20	Madarska	2	10,483	64.9	263.30	10.0	16.9
21	Madarska	3	13,365	67.2	256.99	11.2	17.1
22	Madarska	4	15,100	65.2	293.11	11.0	18.0
23	Madarska	5	22,880	60.4	323.17	11.0	19.0
24	Madarska	6	34,691	54.6	332.37	10.2	19.5
25	Madarska	7	42,213	50.5	328.16	7.7	20.2
26	Madarska	8	43,225	44.0	340.58	6.8	20.9
27	Madarska	9	39,889	37.7	350.09	5.1	20.6
28	Rumunjska	1	302,796	23.9	137.31	5.6	10.7
29	Rumunjska	2	246,626	24.1	142.61	6.5	11.2
30	Rumunjska	3	197,985	28.2	137.30	7.0	11.9
31	Rumunjska	4	195,551	33.9	157.89	7.2	12.9
32	Rumunjska	5	170,186	29.1	157.26	6.8	13.5
33	Rumunjska	6	161,755	26.5	179.36	7.1	13.8
34	Rumunjska	7	172,871	23.0	205.34	6.8	14.2
35	Rumunjska	8	194,718	25.1	238.38	6.8	15.0
36	Rumunjska	9	207,578	22.9	276.34	5.9	15.1
37	Slovenija	1	12,109	33.5	566.50	4.4	19.0
38	Slovenija	2	18,788	35.1	589.19	5.9	19.6
39	Slovenija	3	15,937	37.2	734.15	7.3	20.2
40	Slovenija	4	12,024	34.4	748.10	8.2	21.6
41	Slovenija	5	14,378	35.6	763.06	8.9	23.0
42	Slovenija	6	13,384	37.6	783.66	10.1	24.4
43	Slovenija	7	14,336	37.6	789.15	9.7	25.1
44	Slovenija	8	14,913	37.7	790.73	9.0	26.6
45	Slovenija	9	15,572	38.0	790.73	8.0	27.2

3.1 Deskriptivna statistika

Deskriptivna statistika za promatrane varijable po državama:

```
/*SAS kod*/
proc means data=diiplomski;
var broj_iseljenih  broj_rast_100brakova    min_mj_pl   zavrseni_faks_p;
by drzava;
run;
```

Tablica 3.2: Deskriptivna statistika za države: ispis iz SAS-a

drzava=Grcka

Variable	Label	N	Minimum	Median	Maximum	Mean	Std Dev
broj_iseljenih	broj_iseljenih	9	43044.00	106535.00	124694.00	89517.00	31629.18
broj_rast_100brakova	broj_rast_100brakova	9	22.2000000	24.6000000	32.6000000	26.1444444	3.6946282
min_mj_pl	min_mj_pl	9	683.7600000	683.7600000	876.6200000	757.2311111	90.0660708
nezap_p_ac	nezap_p_ac	9	7.8000000	23.6000000	27.5000000	19.4444444	7.6456051
zavrseni_faks_p	zavrseni_faks_p	9	19.8000000	22.9000000	26.4000000	22.9000000	2.3900837

drzava=Hrvatska

Variable	Label	N	Minimum	Median	Maximum	Mean	Std Dev
broj_iseljenih	broj_iseljenih	9	10638.00	13017.00	36436.00	18199.22	9056.63
broj_rast_100brakova	broj_rast_100brakova	9	21.5000000	28.0000000	34.4000000	27.9222222	4.4350247
min_mj_pl	min_mj_pl	9	374.3100000	390.9400000	414.4500000	391.6022222	12.7510105
nezap_p_ac	nezap_p_ac	9	8.6000000	13.7000000	17.4000000	13.7000000	3.2630507
zavrseni_faks_p	zavrseni_faks_p	9	13.6000000	15.8000000	20.0000000	16.6888889	2.2685042

drzava=Madarska

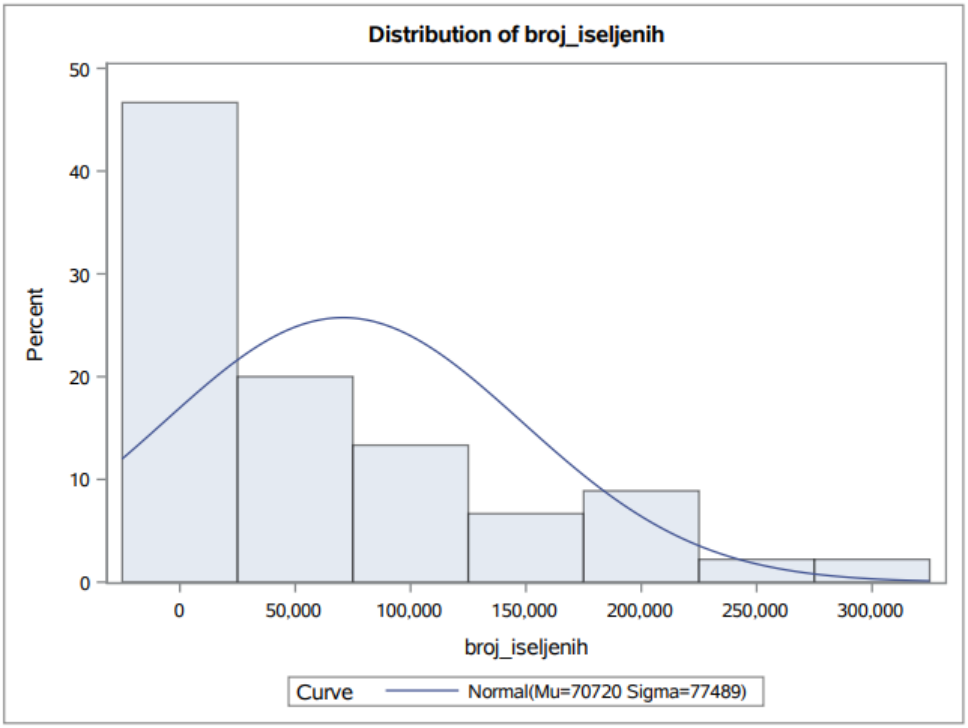
Variable	Label	N	Minimum	Median	Maximum	Mean	Std Dev
broj_iseljenih	broj_iseljenih	9	9591.00	22880.00	43225.00	25715.22	14254.34
broj_rast_100brakova	broj_rast_100brakova	9	37.7000000	60.4000000	67.2000000	56.3555556	10.3893107
min_mj_pl	min_mj_pl	9	256.9900000	323.1700000	350.0900000	308.9833333	33.7253651
nezap_p_ac	nezap_p_ac	9	5.1000000	10.0000000	11.2000000	8.9777778	2.1924745
zavrseni_faks_p	zavrseni_faks_p	9	16.4000000	19.0000000	20.9000000	18.7333333	1.6955825

drzava=Rumunjska

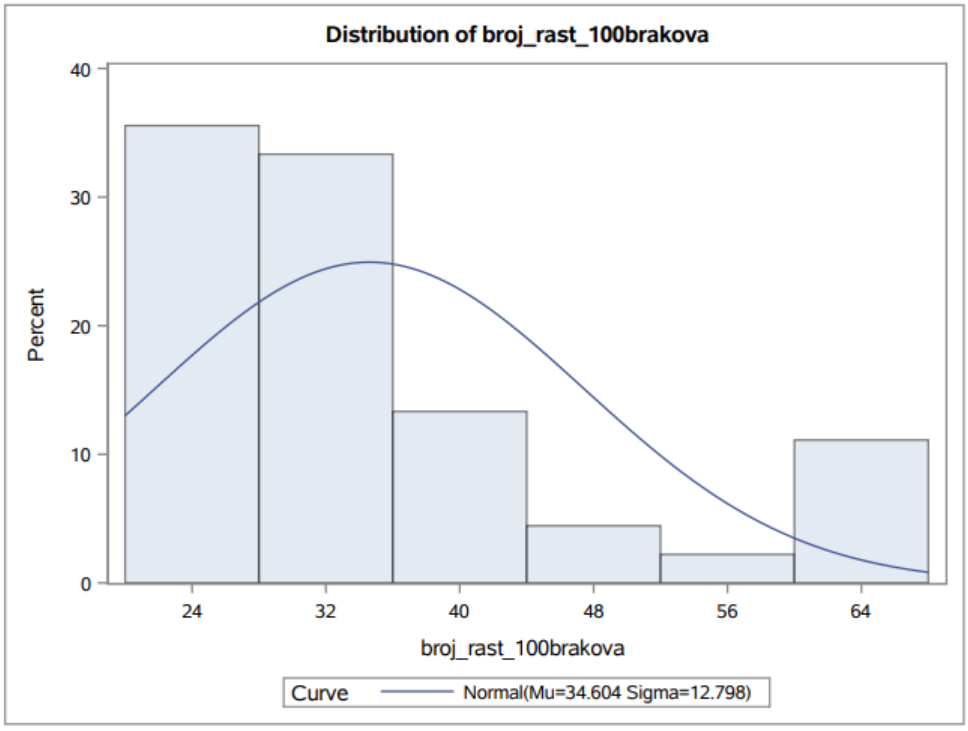
Variable	Label	N	Minimum	Median	Maximum	Mean	Std Dev
broj_iseljenih	broj_iseljenih	9	161755.00	195551.00	302796.00	205562.89	44242.52
broj_rast_100brakova	broj_rast_100brakova	9	22.9000000	25.1000000	33.9000000	26.3000000	3.6010415
min_mj_pl	min_mj_pl	9	137.3000000	157.8900000	276.3400000	181.3100000	49.2016237
nezap_p_ac	nezap_p_ac	9	5.6000000	6.8000000	7.2000000	6.6333333	0.5454356
zavrseni_faks_p	zavrseni_faks_p	9	10.7000000	13.5000000	15.1000000	13.1444444	1.5930404

drzava=Slovenija

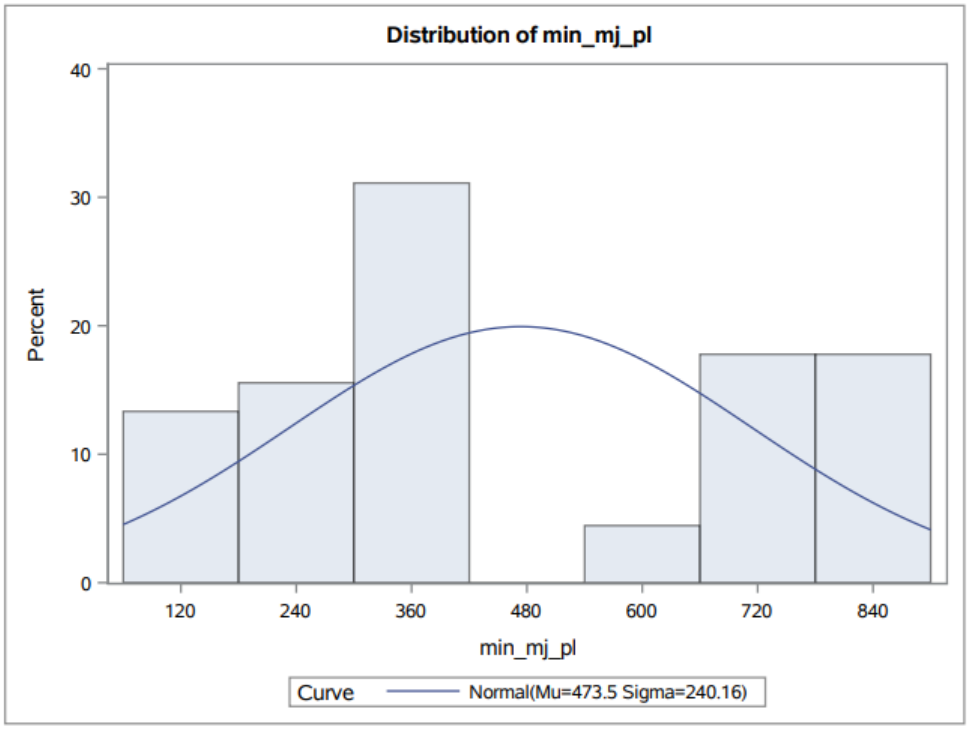
Variable	Label	N	Minimum	Median	Maximum	Mean	Std Dev
broj_iseljenih	broj_iseljenih	9	12024.00	14378.00	18788.00	14604.56	2085.90
broj_rast_100brakova	broj_rast_100brakova	9	33.5000000	37.2000000	38.0000000	36.3000000	1.6740669
min_mj_pl	min_mj_pl	9	566.5000000	763.0600000	790.7300000	728.3633333	87.8406071
nezap_p_ac	nezap_p_ac	9	4.4000000	8.2000000	10.1000000	7.9444444	1.8365124
zavrseni_faks_p	zavrseni_faks_p	9	19.0000000	23.0000000	27.2000000	22.9666667	3.0479501



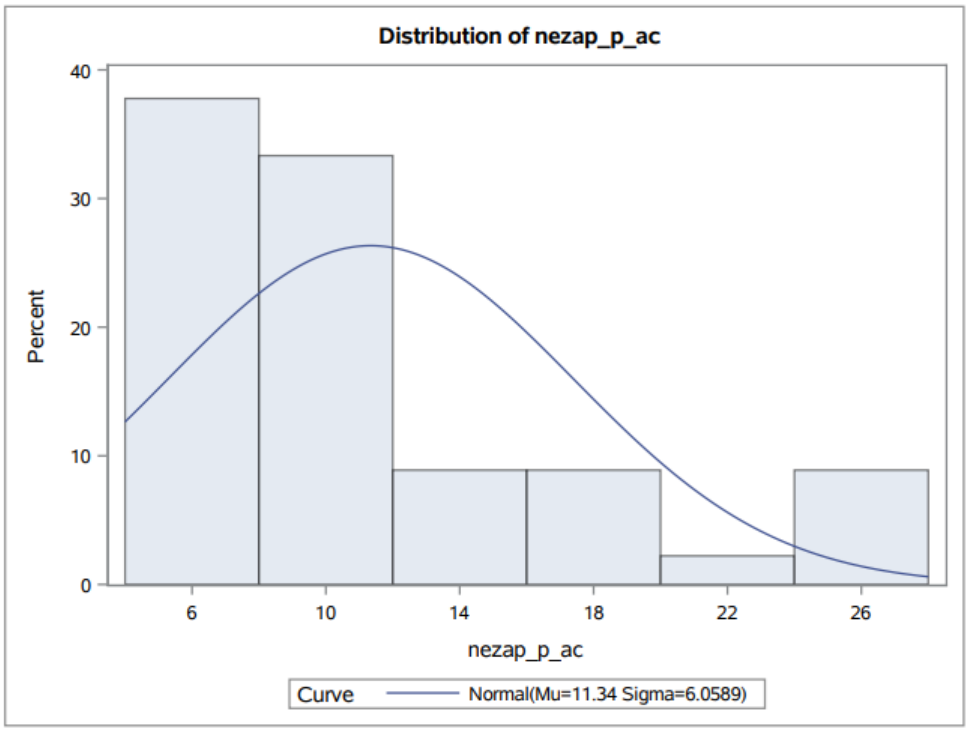
Slika 3.1: Histogram za broj iseljenih



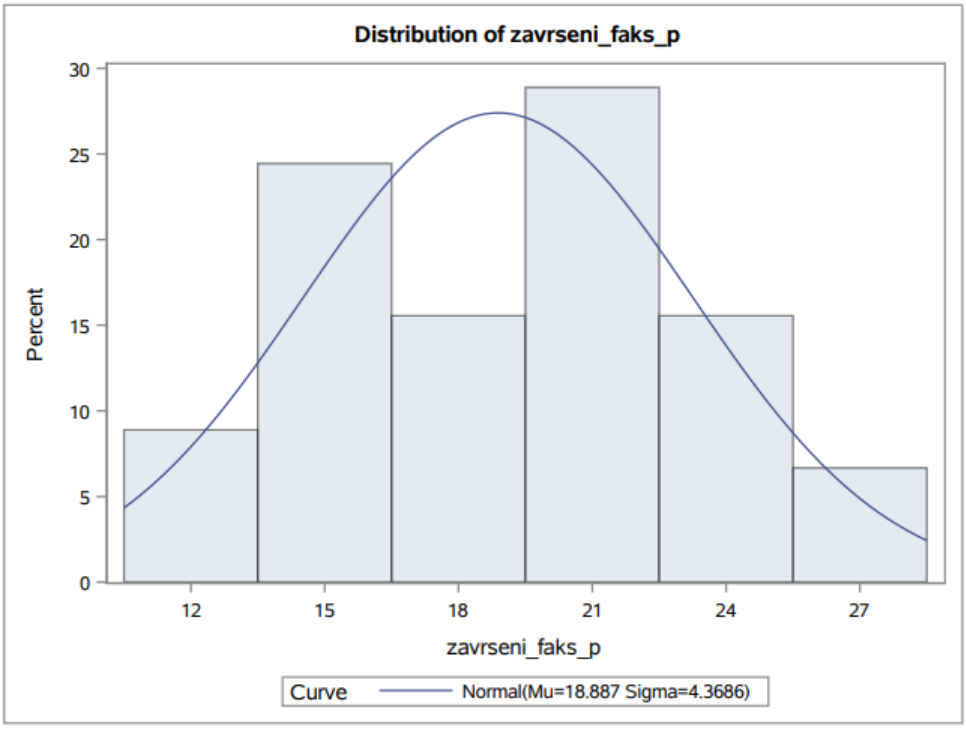
Slika 3.2: Histogram za broj rastava na 100 brakova



Slika 3.3: Histogram za minimalnu mjesečnu plaću u eurima



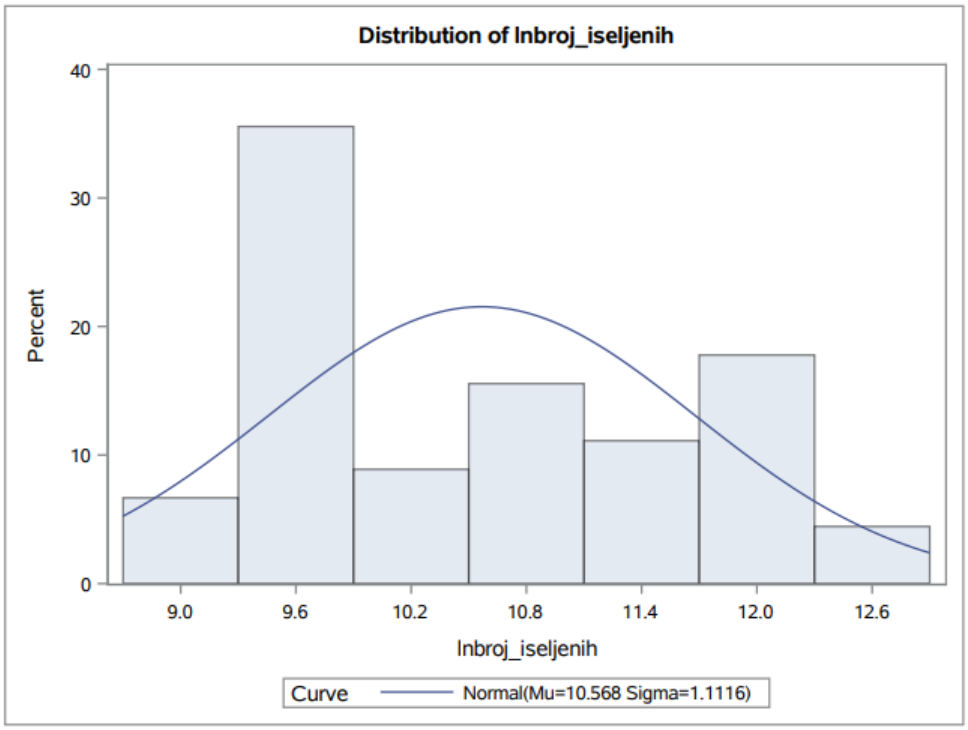
Slika 3.4: Histogram za postotak nezaposlenog stanovništva unutar skupine radno aktivnog stanovništva



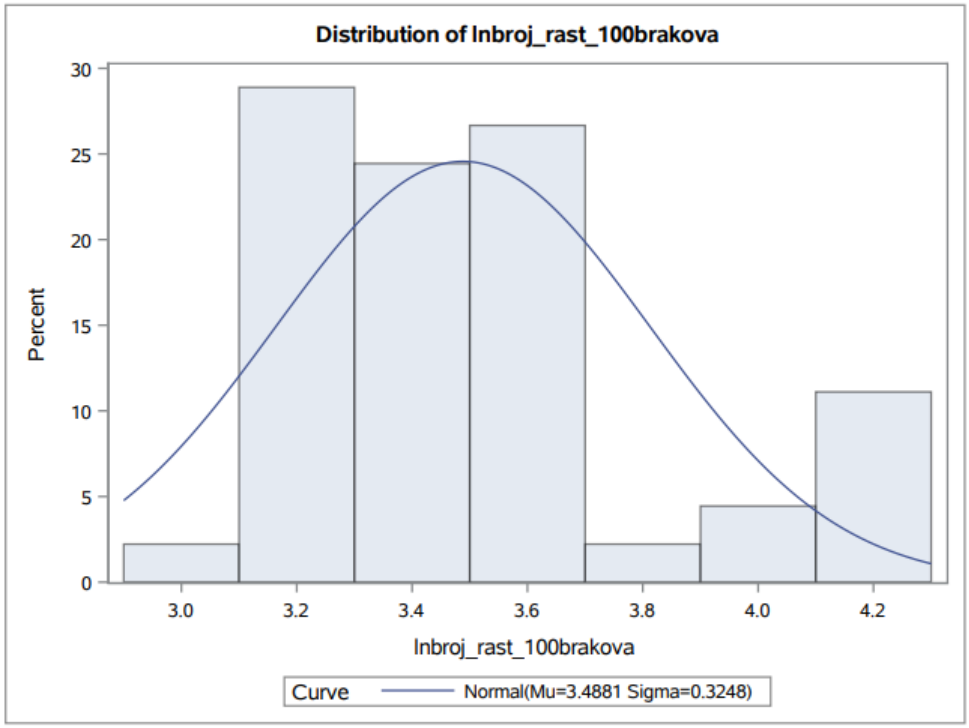
Slika 3.5: Histogram za postotak stanovništva sa završenim fakultetom

Sa prethodnih histograma vidimo da je neke varijable potrebno transformirati, to jest logaritmirati broj iseljenih, postotak nezaposlenog stanovništva unutar skupine radno aktivnog stanovništva i broj rastava na 100 brakova:

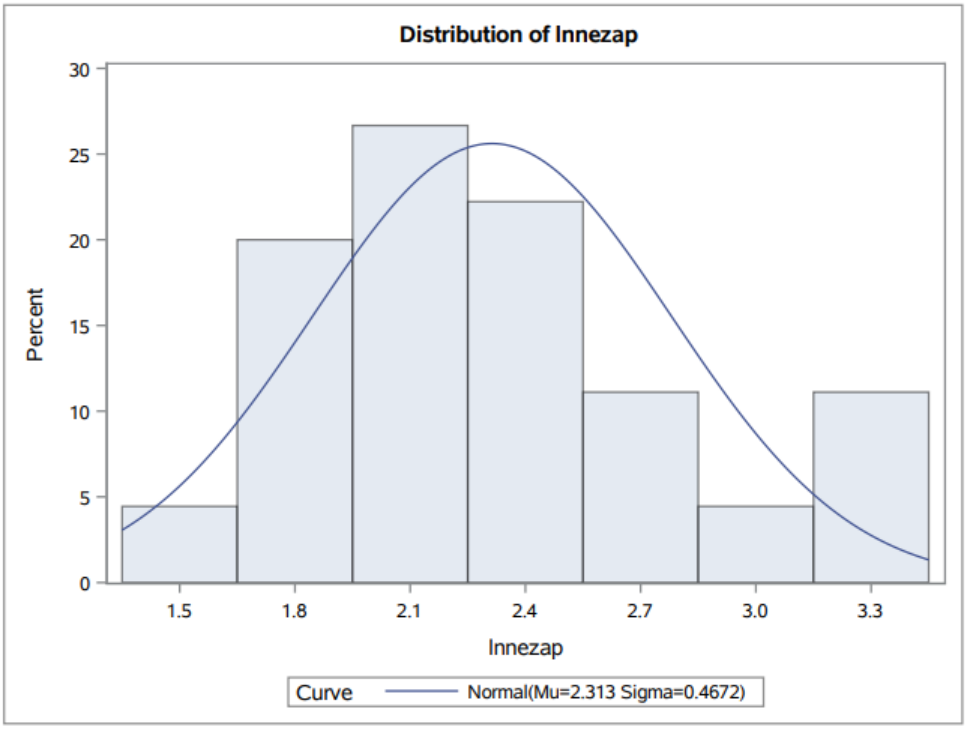
```
/*SAS kod*/
data diiplomski; set diiplomski;
lnbroj_iseljenih=log(broj_iseljenih);
lnnezap=log(nezap_p_ac);
lnbroj_rast_100brakova=log(broj_rast_100brakova);
run;
```



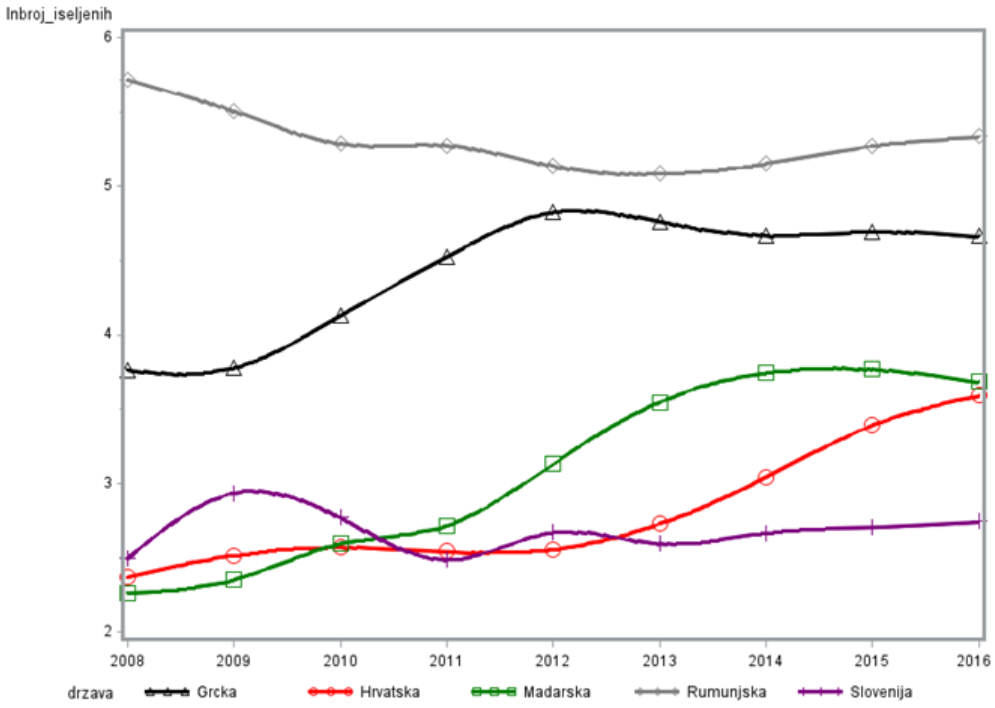
Slika 3.6: Histogram transformiranih podataka za broj iseljenih



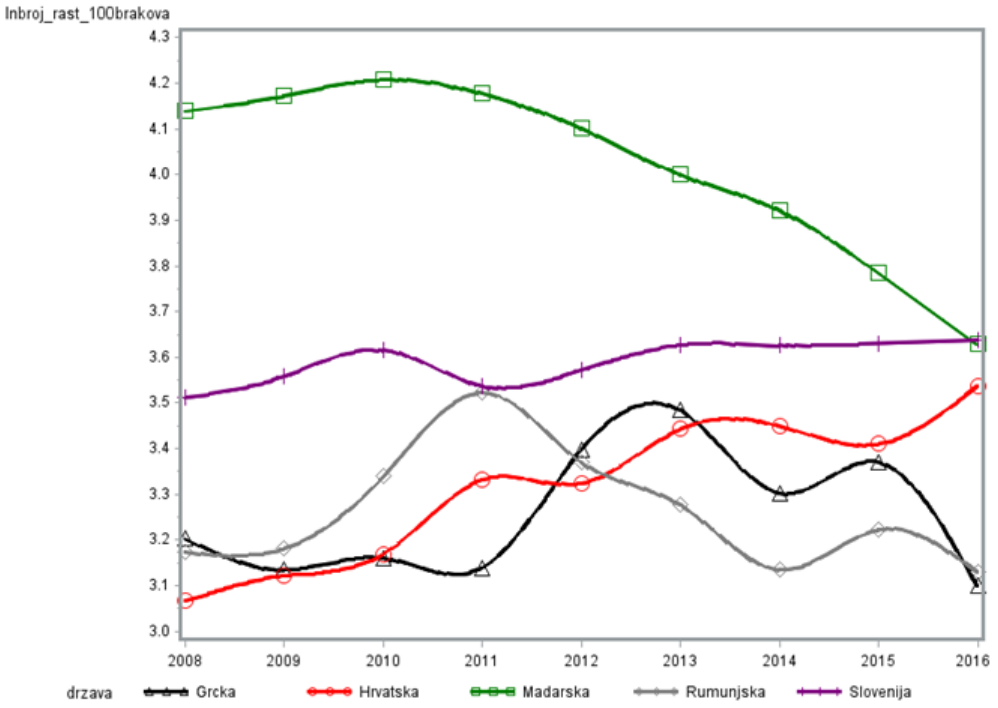
Slika 3.7: Histogram transformiranih podataka za broj rastava na 100 brakova



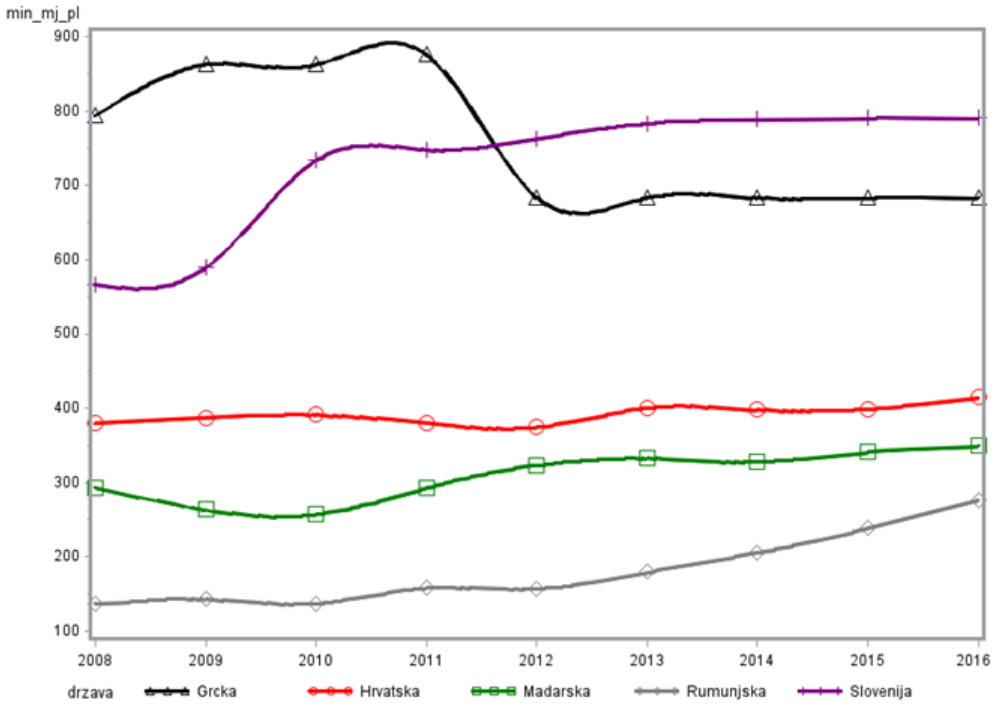
Slika 3.8: Histogram transformiranih podataka za postotak nezaposlenog stanovništva unutar skupine radno aktivnog stanovništva



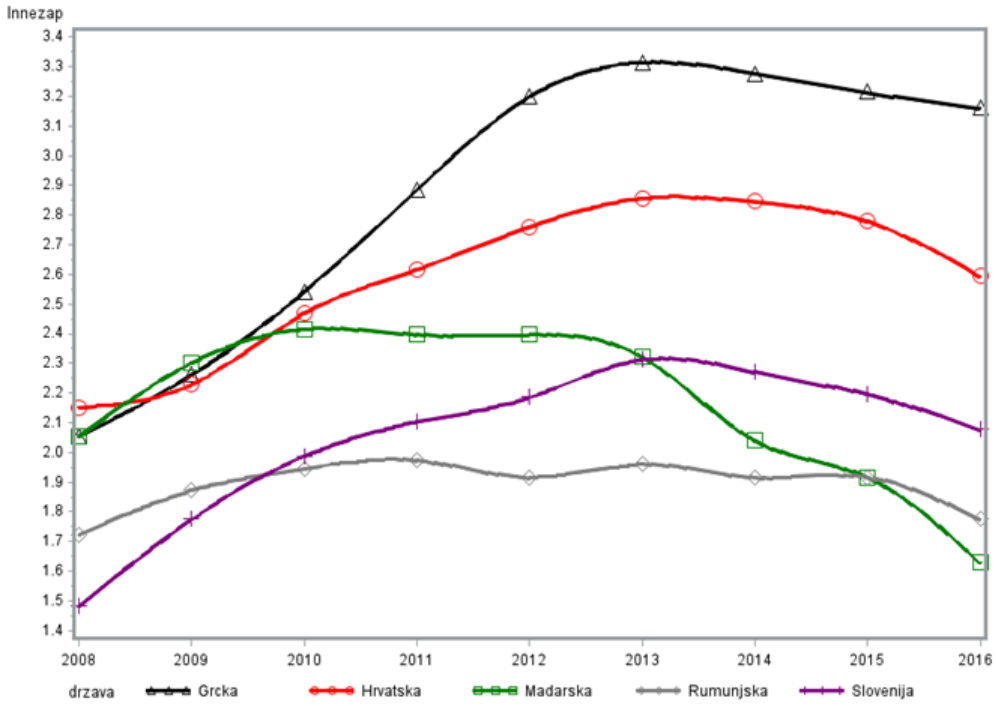
Slika 3.9: Distribucija za *ln*(broj iseljenih) kroz razdoblje 2008.-2016.



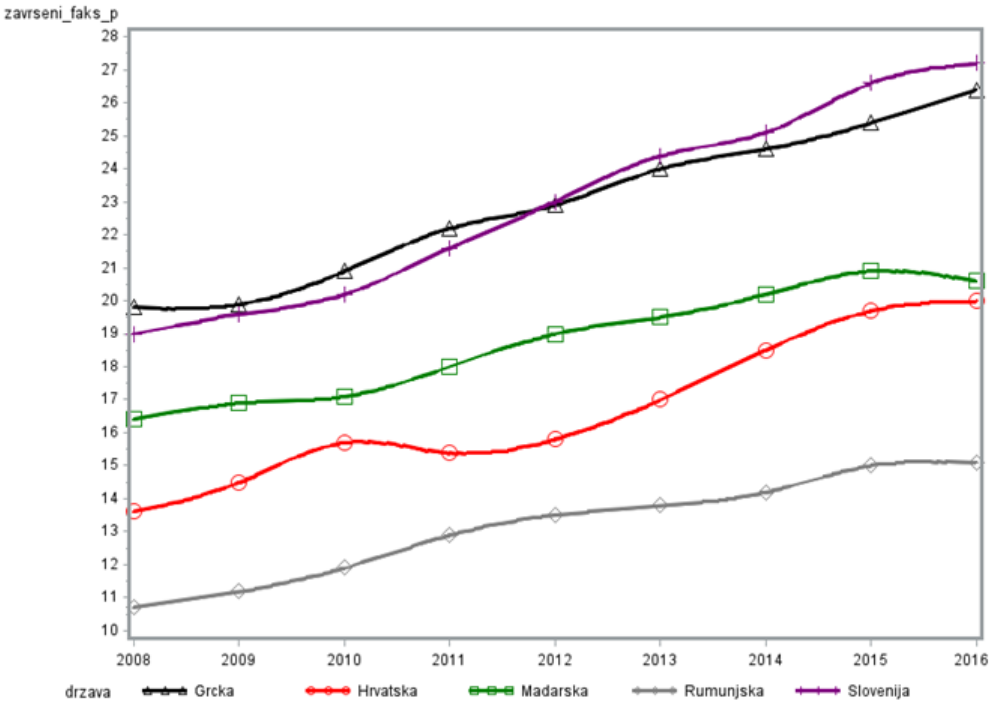
Slika 3.10: Distribucija za *ln*(broj rastava na 100 brakova) kroz razdoblje 2008.-2016.



Slika 3.11: Distribucija za minimalnu mjesečnu plaću kroz razdoblje 2008.-2016.



Slika 3.12: Distribucija za *ln*(postotak nezaposlenog stanovništva unutar skupine radno aktivnog stanovništva) kroz razdoblje 2008.-2016.



Slika 3.13: Distribucija za postotak stanovništva sa završenim fakultetom kroz razdoblje 2008.-2016.

3.2 Modeli i analiza rezultata

Polazni model dan je s:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \beta_4 x_{4,it},\ i = 1, 2, \ldots, 5,\ t = 1, 2, \ldots, 9, \tag{3.1}$$

gdje je

- $y = ln$(broj iseljenih)
- $x_1 = ln$(broj rastava na 100 brakova)
- x_2=minimalna mjesečna plaća u eurima,
- x_3=postotak stanovništva sa završenim fakultetom,
- $x_4 = ln$(postotak nezaposlenog stanovništva unutar skupine radno aktivnog stanovništva),

za $i = 1, 2, \ldots, 5$

- 1=Grčka,
- 2=Hrvatska,
- 3=Mađarska,
- 4=Rumunjska,
- 5=Slovenija,

a za $t = 1, 2, \ldots, 9$

- 1=2008. godina,
- 2=2009. godina,
  ⋮
- 9=2016. godina

Za združeni model, u SAS-u se mogu koristiti procedure REG i PANEL koje daju iste rezultate:

```
/*SAS kod*/
title"Zdruzeni model";
proc panel data=diiplomski;
id drzava godina;
model lnbroj_iseljenih= lnbroj_rast_100brakova  min_mj_pl   zavrseni_faks_p lnnezap /POOLED;
run;

/*ili*/

proc reg data=diiplomski;
model lnbroj_iseljenih= lnbroj_rast_100brakova  min_mj_pl   zavrseni_faks_p lnnezap;
run;
```

Tablica 3.3: Rezultati panel analize za združeni model: ispis iz SAS-a

Fit Statistics			
SSE	31.5506	DFE	40
MSE	0.7888	Root MSE	0.8881
R-Square	0.4197		

Parameter Estimates						
Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	18.2412	1.7388	10.49	<.0001	Intercept
lnbroj_rast_100brakova	1	-2.4474	0.5169	-4.73	<.0001	
min_mj_pl	1	-0.00406	0.00119	-3.40	0.0015	min_mj_pl
zavrseni_faks_p	1	0.15523	0.0704	2.20	0.0333	zavrseni_faks_p
lnnezap	1	-0.0629	0.3337	-0.19	0.8514	

Na temelju procjene združenog modela vidimo kako su sve varijable značajne, osim postotka nezaposlenog stanovništva unutar skupine radno aktivnog stanovništva. Nadalje, vidimo kako postotak stanovništva sa završenim fakultetom ima pozitivan utjecaj na broj iseljenih, dok ostale varijable imaju negativan utjecaj. Procijenjeni model dan je s:

$$y_{it} = 18.2412 - 2.4474x_{1,it} - 0.00406x_{2,it} + 0.1552x_{3,it} - 0.0629x_{4,it},\ i = 1, 2, \ldots, 5,\ t = 1, 2, \ldots, 9. \tag{3.2}$$

Kod združenog modela R^2 iznosi 0.4197 pa možemo posumnjati na postojanje individualnih efekata. Model fiksnih efekata u SAS-u možemo dobiti pomoću procedura TSCSREG i PANEL uz opciju /FIXONE (odnosno /FIXTWO ako se žele uključiti i vremenski efekti). Oni nam zapravo daju LSDV model, bez da posebno kreiramo pomoćne varijable i računamo udaljenosti od sredina. Također daju R^2 i $F-$ test za testiranje postojanja fiksnih efekata. Da bismo dobili i ispis fiksnih efekata, koristimo opciju /PRINTFIXED.

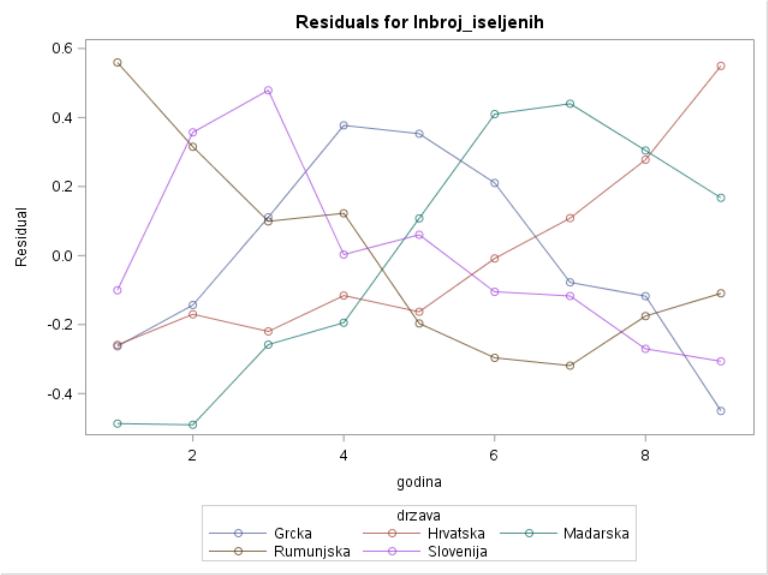
```
/*SAS kod*/
title"Model fiksnih efekata";
proc panel data=diiplomski;
id drzava godina;
model lnbroj_iseljenih= lnbroj_rast_100brakova  min_mj_pl   zavrseni_faks_p  lnnezap
/FIXONE printfixed;
run;

/*ili*/

proc tscsreg data=diiplomski;
id drzava godina;
model lnbroj_iseljenih= lnbroj_rast_100brakova  min_mj_pl   zavrseni_faks_p  lnnezap /FIXONE ;
run;
```

Tablica 3.4: Rezultati panel analize za model fiksnih efekata: ispis iz SAS-a

Fit Statistics			
SSE	3.5590	DFE	36
MSE	0.0989	Root MSE	0.3144
R-Square	0.9345		

F Test for No Fixed Effects			
Num DF	Den DF	F Value	Pr > F
4	36	70.79	<.0001

Parameter Estimates						
Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
CS1	1	1.65434	0.3844	4.30	0.0001	Cross Sectional Effect 1
CS2	1	0.073563	0.4437	0.17	0.8693	Cross Sectional Effect 2
CS3	1	0.333015	0.3851	0.86	0.3929	Cross Sectional Effect 3
CS4	1	2.496484	0.5325	4.69	<.0001	Cross Sectional Effect 4
Intercept	1	10.62568	1.7346	6.13	<.0001	Intercept
lnbroj_rast_100brakova	1	-0.6445	0.4712	-1.37	0.1799	
min_mj_pl	1	-0.00217	0.000901	-2.41	0.0213	min_mj_pl
zavrseni_faks_p	1	0.129121	0.0296	4.36	0.0001	zavrseni_faks_p
lnnezap	1	-0.05766	0.2730	-0.21	0.8339	



Slika 3.14: Reziduali za model fiksnih efekata: ispis iz SAS-a

SAS automatski uzima zadnju jedinicu (državu) kao referentnu jedinicu pa u ovom modelu intercept predstavlja stvarni efekt države 5, to jest Slovenije, dok su za ostale države dane udaljenosti od intercepta. Želimo li ostale države uspoređivati s Hrvatskom, u bazu dodamo pomoćnu varijablu di za svaku državu i. Sada u model uključimo još i pomoćne varijable, osim one za koju želimo da bude referentna jedinica.

Tablica 3.5: Izgled baze s dodanim pomoćnim varijablama: ispis iz SAS-a

Obs	godina	drzava	broj_iseljenih	broj_rast_100brakova	min_mj_pl	nezap_p_ac	zavrseni_faks_p	d1	d2	d3	d4	d5
1	1	Grcka	43,044	24.6	794.02	7.8	19.8	1	0	0	0	0
2	2	Grcka	43,686	23.0	862.82	9.6	19.9	1	0	0	0	0
3	3	Grcka	62,041	23.6	862.82	12.7	20.9	1	0	0	0	0
4	4	Grcka	92,404	23.1	876.62	17.9	22.2	1	0	0	0	0
5	5	Grcka	124,694	29.9	683.76	24.5	22.9	1	0	0	0	0
6	6	Grcka	117,094	32.6	683.76	27.5	24.0	1	0	0	0	0
7	7	Grcka	106,804	27.2	683.76	26.5	24.6	1	0	0	0	0
8	8	Grcka	109,351	29.1	683.76	24.9	25.4	1	0	0	0	0
9	9	Grcka	106,535	22.2	683.76	23.6	26.4	1	0	0	0	0
10	1	Hrvatska	10,638	21.5	379.60	8.6	13.6	0	1	0	0	0
11	2	Hrvatska	12,355	22.7	386.91	9.3	14.5	0	1	0	0	0
12	3	Hrvatska	13,017	23.8	390.94	11.8	15.7	0	1	0	0	0
13	4	Hrvatska	12,699	28.0	380.18	13.7	15.4	0	1	0	0	0
14	5	Hrvatska	12,877	27.8	374.31	15.8	15.8	0	1	0	0	0
15	6	Hrvatska	15,262	31.3	400.67	17.4	17.0	0	1	0	0	0
16	7	Hrvatska	20,858	31.5	398.31	17.2	18.5	0	1	0	0	0
17	8	Hrvatska	29,651	30.3	399.05	16.1	19.7	0	1	0	0	0
18	9	Hrvatska	36,436	34.4	414.45	13.4	20.0	0	1	0	0	0
19	1	Madarska	9,591	62.7	293.08	7.8	16.4	0	0	1	0	0
20	2	Madarska	10,483	64.9	263.30	10.0	16.9	0	0	1	0	0
21	3	Madarska	13,365	67.2	256.99	11.2	17.1	0	0	1	0	0
22	4	Madarska	15,100	65.2	293.11	11.0	18.0	0	0	1	0	0
23	5	Madarska	22,880	60.4	323.17	11.0	19.0	0	0	1	0	0
24	6	Madarska	34,691	54.6	332.37	10.2	19.5	0	0	1	0	0
25	7	Madarska	42,213	50.5	328.16	7.7	20.2	0	0	1	0	0

```
/*SAS kod*/
title "Fiksni efekti: referentna drzava Hrvatska";
proc reg data=diiplomski;
   model lnbroj_iseljenih = d1 d3 d4 d5 lnbroj_rast_100brakova  min_mj_pl   zavrseni_faks_p lnnezap ;
   label d1="Grcka";
   label d3="Madarska";
   label d4="Rumunjska";
   label d5="Slovenija";
run;
```

Tablica 3.6: Rezultati za model fiksnih efekata s Hrvatskom kao referentnom državom: ispis iz SAS-a

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	10.69925	1.40126	7.64	<.0001
d1	Grcka	1	1.58078	0.34126	4.63	<.0001
d3	Madarska	1	0.25945	0.47089	0.55	0.5850
d4	Rumunjska	1	2.42292	0.28150	8.61	<.0001
d5	Slovenija	1	-0.07356	0.44374	-0.17	0.8693
lnbroj_rast_100brakova		1	-0.64450	0.47120	-1.37	0.1799
min_mj_pl	min_mj_pl	1	-0.00217	0.00090054	-2.41	0.0213
zavrseni_faks_p	zavrseni_faks_p	1	0.12912	0.02959	4.36	0.0001
lnnezap		1	-0.05766	0.27298	-0.21	0.8339

Procijenjeni model za Hrvatsku je:

$$y_{it} = 10.69925 - 0.644x_{1,it} - 0.002x_{2,it} + 0.129x_{3,it} + 0.057x_{4,it},\ i = 1,2,\ldots,5,\ t = 1,2,\ldots,9, \tag{3.3}$$

za Grčku:

$$y_{it} = (10.69925 + 1.58078) - 0.644x_{1,it} - 0.002x_{2,it} + 0.129x_{3,it} + 0.057x_{4,it},\ i = 1,2,\ldots,5,\ t = 1,2,\ldots,9, \tag{3.4}$$

za Mađarsku:

$$y_{it} = (10.69925 + 0.25945) - 0.644x_{1,it} - 0.002x_{2,it} + 0.129x_{3,it} + 0.057x_{4,it},\ i = 1,2,\ldots,5,\ t = 1,2,\ldots,9, \tag{3.5}$$

za Rumunjsku:

$$y_{it} = (10.69925 + 2.42292) - 0.644x_{1,it} - 0.002x_{2,it} + 0.129x_{3,it} + 0.057x_{4,it},\ i = 1,2,\ldots,5,\ t = 1,2,\ldots,9, \tag{3.6}$$

a za Sloveniju:

$$y_{it} = (10.69925 - 0.07356) - 0.644x_{1,it} - 0.002x_{2,it} + 0.129x_{3,it} + 0.057x_{4,it},\ i = 1,2,\ldots,5,\ t = 1,2,\ldots,9. \tag{3.7}$$

S obzirom na to da je Hrvatska referentna država jedino Slovenija ima manji procijenjeni broj iseljenih (procijenjeni parametar modela je -0.07), iako taj rezultat nije statistički

značajan. Sve druge analizirane države imaju veći procijenjeni broj iseljenih s obzirom na Hrvatsku. Najveći broj ima Rumunjska (procijenjeni parametar je 2.42) i taj je rezultat statistički značajan. Nakon toga, Grčka (procijenjeni parametar je 1.58) i taj je rezultat statistički značajan, dok Mađarska također ima veći procijenjeni broj iseljenih, ali taj rezultat nije statistički značajan.

Prema procijenjenim parametrima združenog modela, jedini parametar koji se promijenio je broj rastavljenih brakova koji više nije statistički značajan, iako je smjer i dalje ostao isti. Zbog toga je za pretpostaviti da je to varijabla koja najviše ovisi o državi.

R^2 iznosi 0.93, a osim toga dana je i vrijednost $F-$testa. Kako je $p-$vrijednost jako mala, odbacujemo hipotezu H_0 o nepostojanju fiksnih efekata. Dakle može se zaključiti da među državama postoji heterogenost slobodnih članova.

U SAS-u TSCSREG i PANEL procedure imaju opciju /RANONE (odnosno /RANTWO) za prilagodbu podataka modelu slučajnih efekata. Obje procedure za procjenu modela koriste Fuller Battese (1974) metodu, a postoje još tri metode: Wansbeek Kapteyn, Wallace Hussain i Nerlove, koje se trebaju posebno uključiti. PROC PANEL ima i opciju /BP koja daje Breusch-Paganov LM test.

```
/*SAS kod*/
title"Model slucajnih efekata";
proc panel data=diiplomski;
id drzava godina;
model lnbroj_iseljenih= lnbroj_rast_100brakova  min_mj_pl   zavrseni_faks_p  lnnezap
/RANONE BP;
run;

/*ili*/

proc tscsreg data=diiplomski;
id drzava godina;
model lnbroj_iseljenih= lnbroj_rast_100brakova  min_mj_pl   zavrseni_faks_p  lnnezap /RANONE;
run;
```

Tablica 3.7: Rezultati panel analize za model slučajnih efekata: ispis iz SAS-a

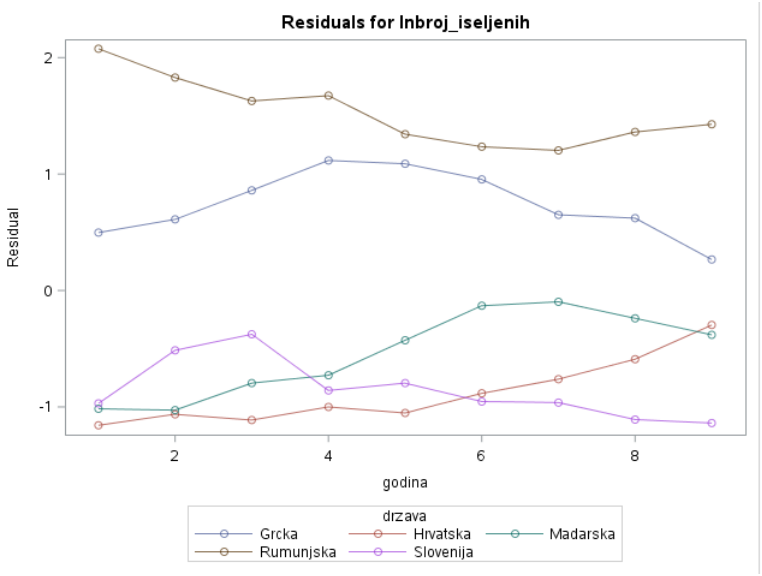
Fit Statistics			
SSE	3.8155	DFE	40
MSE	0.0954	Root MSE	0.3088
R-Square	0.4341		

Variance Component Estimates	
Variance Component for Cross Sections	1.839825
Variance Component for Error	0.098861

Hausman Test for Random Effects			
Coefficients	DF	m Value	Pr > m
4	4	0.62	0.9610

Breusch Pagan Test for Random Effects (One Way)		
DF	m Value	Pr > m
1	111.14	<.0001

Parameter Estimates						
Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	11.90872	1.5807	7.53	<.0001	Intercept
lnbroj_rast_100brakova	1	-0.74559	0.4460	-1.67	0.1024	
min_mj_pl	1	-0.00228	0.000850	-2.68	0.0107	min_mj_pl
zavrseni_faks_p	1	0.126455	0.0287	4.40	<.0001	zavrseni_faks_p
lnnezap	1	-0.02199	0.2591	-0.08	0.9328	



Slika 3.15: Reziduali za model slučajnih efekata: ispis iz SAS-a

Iz rezultata se može vidjeti da su procjene komponenata varijance $\hat{\sigma^2}_{\alpha} = 1.839825$ i $\hat{\sigma^2}_{\epsilon} = 0.098861$. Kod Breusch-Paganovog LM testa p-vrijednost je jako mala pa odbacujemo hipotezu H_0 o nepostojanju slučajnih efekata, to jest da je združeni model prikladniji. Dakle individualni efekti postoje, jer smo u oba slučaja odbacili hipotezu o prikladnosti združenog modela. Također opcija /RANONE nam daje i Hausmanov test iz kojeg vidimo da ne možemo odbaciti hipotezu o nepostojanju korelacije nezavisnih varijabli i individualnih efekata. Zbog moguće precijenjenog Hausmanovog testa, naša analiza je pokazala djelomice oprečne rezultate koji zahtijevaju daljnju analizu. Hausmanov test je pokazao da je model slučajnih efekata prikladniji, no R^2 je puno veći kod modela fiksnih efekata, te su grafovi reziduala puno bolji. Također, procjenitelj modela fiksnih efekata je uvijek konzistentan, bez obzira na to jesu li regresori korelirani s individualnim efektima, dok to nije slučaj kod procjenitelja slučajnih efekata. Stoga se „sigurnije" odlučiti za model fiksnih efekata.

Bibliografija

[1] Cheng Hsiao: *Analysis of Panel Data, Third Edition*, Cambridge University Press, 2014.

[2] Radmila Dragutinović Mitrović: *Analiza panel serija*, Zadužbina Andrejević

[3] Jeffrey M. Wooldridge: *Econometric Analysis of Cross Section and Panel Data, Second Edition*, The MIT Press, 2010.

[4] *https://www.iuj.ac.jp/faculty/kucc625/method/panel/panel_iuj.pdf*

[5] *http://statmath.wu.ac.at/~hauser/LVs/FinEtricsQF/FEtrics_Chp5.pdf*

[6] *https://feb.kuleuven.be/public/u0017833/courses/advanced_econometrics/panelshort.pdf*

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

[8] *https://web.math.pmf.unizg.hr/nastava/statpr/files/linearna.pdf*

[9] *https://fet.unipu.hr/images/50016022/Belullo-Ekonometrija.pdf*

Sažetak

U ovom radu izlažu se osnovne metode i teorijska podloga analize panel podataka kao vrlo važnog aspekta brojnih istraživanja u prirodnim i društveno-humanističkim znanostima. Njima se obuhvaćaju i analiziraju opažanja prikupljena tijekom različitih vremenskih točaka, to jest promatra se promjena varijabli i njihov utjecaj kroz vrijeme.

Na početku rad donosi prikaz linearne regresije koja čini osnovu statističkog modeliranja, a kojom se u vezu dovodi zavisna varijabla i funkcija nezavisnih varijabli. U sljedećem poglavlju dan je pregled vrsta podataka te se objašnjava specifična struktura panel podataka koja se svodi na objedinjavanje vremenskog presjeka i vremenskog niza. Nakon toga objašnjavaju se linearni modeli panel podataka: združeni model, model fiksnih efekata i model slučajnih efekata; te testovi za odabir određenog modela: $F-$ test, Breusch-Paganov test i Hausmanov test. U nastavku rada izloženi teorijski koncept panel podataka nadograđuje se primjerom analize podataka Statističkog ureda Europskih zajednica (Eurostata) u računalnom programu za statističku analizu SAS. Analiza se temelji na broju iseljenih osoba kao zavisnoj varijabli te podacima o mjesečnoj plaći, nezaposlenosti, stupnju obrazovanja i broju rastavljenih kao nezavisnim varijablama. Osim Hrvatske, odabrane su još četiri europske zemlje kao jedinice promatranja u vremenskom razdoblju od 2008. do 2016. godine.

Summary

This paper presents basic methods and theoretical background of panel data analysis as a very important aspect of numerous research topics in natural and social sciences. Panel data contain and analyze the observations of phenomena obtained over multiple time periods and explain influencing variables.

First, the paper presents the linear regression that forms the basis of statistical modeling, which links the dependent variable to the function of independent variables. The next chapter provides an overview of data type and explains specific structure of panel data as combination of cross-section and time series data. Afterwards, text explains various linear models of the panel data: pooled OLS model, fixed effects model and random effects model, as well as tests to evaluate how statistical model corresponds to the data: F-test, the Breusch-Pagan test and Hausman specification test. The theoretical concept of the panel data analysis presented in this paper is outlined through data aggregated by the European Statistical Office (Eurostat) and computed in the statistical analysis software SAS. Presented panel analysis is based on annual collections of statistics on international migration flows as dependent variable, and statistics on monthly wages, unemployment, educational attainment level and divorce rate as independent variables. Besides Croatia, four other European countries have been selected as observation units over the period of 2008 to 2016.

Životopis

Rođena sam 13. veljače 1993. godine u Splitu. Nakon završene Osnovne škole Strožanac, upisujem i završavam Prvu jezičnu gimnaziju u Splitu. Tada upisujem preddiplomski sveučilišni studij Matematike na Prirodoslovno-matematičkom fakultetu Sveučilišta u Splitu. Diplomski sveučilišni studij Matematičke statistike upisujem na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu.