Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku

Ines Malić

Generalizirani linearni model na panel podatcima

Diplomski rad

Osijek, 2014.

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku

Ines Malić

Generalizirani linearni model na panel podatcima

Diplomski rad

Mentor: prof. dr. sc. Mirta Benšić
Komentor: prof. dr. sc. Nataša Šarlija

Osijek, 2014.

Sadržaj

1 Uvod

Panel podatci, također poznati kao longitudinalni podatci, podatci su u kojima se ponašanje entiteta promatra kroz vrijeme. Sastoje se od N subjekata (poduzeća, osoba, ...) koje promatramo kroz T vremenskih trenutaka. Mogli bismo reći da su panel podatci kombinacija "cross-sectional" podataka (N poduzeća) i vremenskog niza (T trenutaka). Ukupno onda imamo NT pojedinačnih opservacija. U idealnim situacijama podatci se bilježe u pravilnim vremenskim razmacima (godina, kvartal, mjesec) i nemamo nedostajućih podataka.

Ukoliko u panelu nemaju zabilježeni podatci za svaki entitet, u svakom vremenskom periodu, takav panel nazivamo *neuravnoteženim*. Kod takvih panela ne raspolažemo više s NT opservacija. *Uravnotežen* je onaj panel kod kojega raspolažemo s NT podataka, tj. svaki od N entiteta ima zabilježene vrijednosti varijabli u svakom od T trenutaka. Također, panel podatke možemo razvrstati prema broju vremenskih trenutaka, odnosno entiteta. Paneli kod kojih imamo puno entiteta, ali samo nekoliko vremenskih perioda ($N > T$) nazivaju se *kratki paneli*. Obratno, paneli koji bilježe podatke za malo entiteta, u većem broju vremenskih trenutaka, nazivaju se *dugi paneli* ($N < T$). U literaturi pojavljuju se i pojmovi: $fiksni$ i *rotirajući* paneli. Kod $fiksnih$ panela uvijek promatramo iste entitete, dok kod *rotirajućih* panela u nekim vremenskim trenutcima pojavljuju se novi entiteti umjesto nekoga od entiteta iz prijašnjeg vremenskog razdoblja. To se može dogoditi npr., ako neko od promatranih poduzeća propadne ili nastavi poslovati pod drugim imenom s novim vlasnikom i sl. Kako bismo što jednostavnije mogli modelirati podatke, priželjkujemo podatke koji dolaze iz kratkoga uravnoteženoga fiksnoga panela podataka.

U nastavku ovoga rada navedeni su neki problemi koji se mogu pojaviti kod modeliranja. Zatim, predstavljeni su linearni modeli za panel podatke, zajedno s procjeniteljima za parametre tih modela dobivene metodom najmanjih kvadrata. U idućim poglavljima koncentriramo se na složenije modele, modele diskretnog izbora, kod kojih ovisna varijabla može primiti samo diskretne vrijednosti. Kod takvih modela, za procjenu parametara predložena je Newton-Raphsonova iterativna metoda. Među modelima diskretnoga izbora izdvojili smo dva tipa modela, model fiksnih efekata i model slučajnih efekata, za koje smo izveli pripadajuće funkcije vjerodostojnosti čijom maksimizacijom dobivamo procjenitelje za nepoznate parametre. Objašnjene metode primijeniti ćemo na panel koji se sastoji od 1270 poduzeća koje promatramo kroz razdoblje od pet godina. Između dobivenih modela vršiti ćemo odabir prema AIC kriteriju. Zaključiti ćemo s interpretacijom rezultata i zaključivanjem o ekonomskoj utemeljenosti tih modela.

2 Mogući problemi kod modeliranja

Pogledajmo primjer vremenskoga niza:

$$y_t = \sum_{\tau=0}^{h} \beta_\tau x_{t-\tau} + u_t, \quad t = 1, ..., T,$$

gdje je x_t egzogena varijabla, u_t slučajna greška. Mogu se pojaviti stroge multikolinearnosti između $h+1$ neovisnih varijabli $x_t, x_{t-1}, ..., x_{t-h}$. Tada nemamo dovoljno informacija da bismo dobili precizne procjene za koeficijente.
Panel podatci lako mogu riješiti problem varijabli za koje nemamo izmjerenih podataka, a koje su korelirane s neovisnim varijablama. Pogledajmo jednostavan regresijski model:

$$y_{it} = \alpha + \beta' \mathbf{x}_{it} + \rho' \mathbf{z}_{it} + u_{it}, \quad i = 1, ..., N \qquad t = 1, ..., T,$$

gdje su $\mathbf{x}_{it}$ i $\mathbf{z}_{it}$ vektori egzogenih varijabli dimenzije $k_1 \times 1$, $k_2 \times 1$ redom. α, β, ρ su vektori konstanti dimenzija 1×1, $k_1 \times 1$, $k_2 \times 1$ redom. u_{it} je greška, jednako distribuirana s obzirom na i i t, s očekivanjem 0 i varijancom σ_u^2. Poznato je da regresija metodom najmanjih kvadrata za y_{it} na $\mathbf{x}_{it}$ i $\mathbf{z}_{it}$ dovodi do nepristranih i konzistentnih procjenitelja za α, β i ρ.
Pretpostavimo sada, da su vrijednosti od $\mathbf{z}_{it}$ nepoznate i da je kovarijanca između $\mathbf{x}_{it}$ i $\mathbf{z}_{it}$ različita od nule. Koeficijenti regresije dobiveni metodom najmanjih kvadrata biti će pristrani. Ako imamo na raspolaganju više promatranja za pojedinu individuu, možemo se riješiti efekta od $\mathbf{z}$. Primjerice, ako je $\mathbf{z}_{it} = \mathbf{z}_i$ za sve t možemo dobiti:

$$y_{it} - y_{i,t-1} = \beta'(\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) + (u_{it} - u_{i,t-1}), \quad i = 1, ..., N \quad t = 2, ..., T. \tag{2.1}$$

Slično, za $\mathbf{z}_{it} = \mathbf{z}_t$ za sve i možemo dobiti:

$$y_{it} - \bar{y}_t = \beta'(\mathbf{x}_{it} - \bar{\mathbf{x}}_t) + (u_{it} - \bar{u}_t), \quad i = 1, ..., N \quad t = 1, ..., T, \tag{2.2}$$

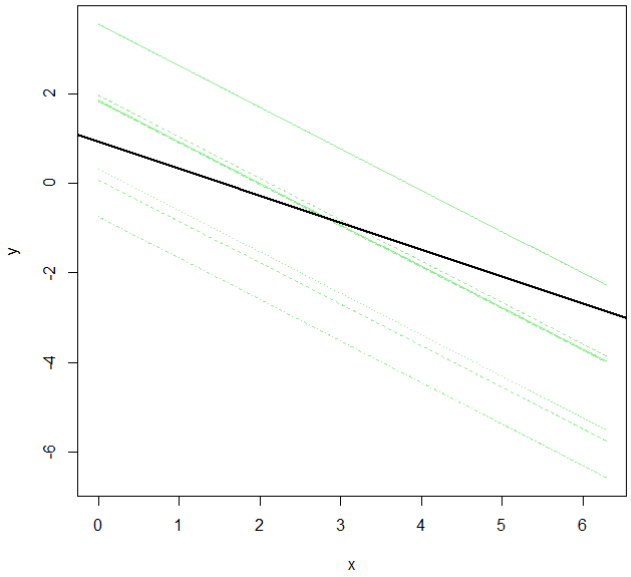
gdje je:

$$\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it} \qquad \bar{\mathbf{x}}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{it} \qquad \bar{u}_t = \frac{1}{N}\sum_{i=1}^{N} u_{it}.$$

Regresija metodom najmanjih kvadrata na 2.1 i 2.2 sada daje nepristrane i konzistente procjenitelje za β. Kada bismo raspolagali s "cross-sectional" podatcima ne bismo mogli primijeniti 2.1, odnosno 2.2 na vremenski niz.

2.1 Heterogena pristranost

Moć panel podataka proizlazi iz njihove teoretske sposobnosti da izoliraju efekte određenih aktivnosti, tretmana i sl. Ta teoretska sposobnost bazirana je na pretpostavci da su podatci generirani iz kontroliranih eksperimenata u kojima su ishodi slučajne varijable s vjerojatnosnom distribucijom, koja je glatka funkcija različitih varijabli koje opisuju uvjete eksperimenta. Ako su dostupni podatci generirani iz jednostavnog kontroliranog



Slika 2.1: Regresijski pravac modela zajedničkih koeficijenata i pravci regresije pojedinih entiteta uz $\alpha_i \neq \alpha_j$ i $\beta_i = \beta_j$

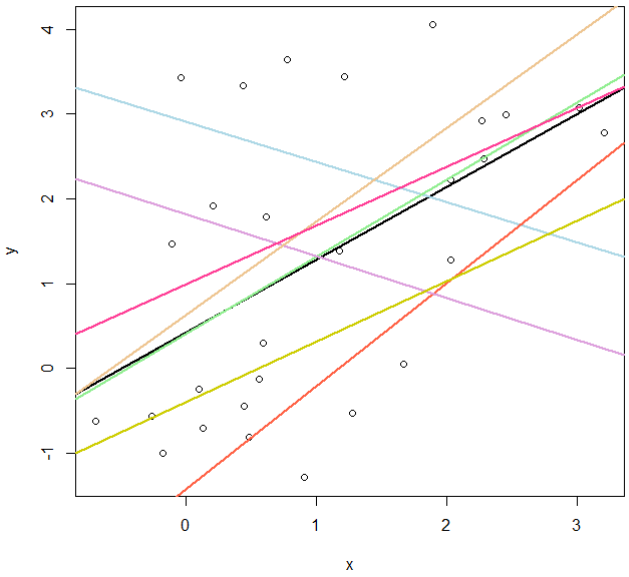
ekperimenta, mogu se primijeniti standardne statističke metode. Nažalost, većina panel podataka dolazi kao rezultat svakodnevnih kompliciranih ekonomskih procesa. Na jednog pojedinca djeluje jako puno različitih faktora. Za opisivanje ponašanja pojedinca ne možemo koristiti beskonačan broj utjecajnih čimbenika. Svrha modeliranja nije oponašati stvarnost, nego uočiti najbitnije elemente koji imaju najveći utjecaj na ishode.
Ignoriranje individualnoga ili vremenskoga fiksnog efekta, koji može postojati među podatcima, može dovesti do parametarske heterogenosti. Ako ju ignoriramo, to može dovesti do nekonzistentnih ili beznačajnih procjena parametara.
Promotrimo jednostavan model:

$$y_{it} = \alpha_i + \beta_i x_{it} + u_{it}, \quad i = 1, ..., N \quad t = 1, ..., T. \tag{2.3}$$

Ukoliko koristimo svih NT podataka za procjenu parametara, imati ćemo model konstantnih koeficijenata:

$$y_{it} = \alpha + \beta x_{it} + u_{it}, \quad i = 1, ..., N \quad t = 1, ..., T. \tag{2.4}$$

U programskom paketu **R** napravljena je simulacija panel podataka. Koristeći ugrađene funkcije, na jednostavan način dobivamo procjene i odgovarajuće grafove linearne regresije na svim podatcima i linearne regresije za pojedini entitet uz određene pretpostavke. Uzmimo prvo da je: $\alpha_i \neq \alpha_j$ i $\beta_i = \beta_j$. Na slici 2.1 zelene linije označuju regresijski

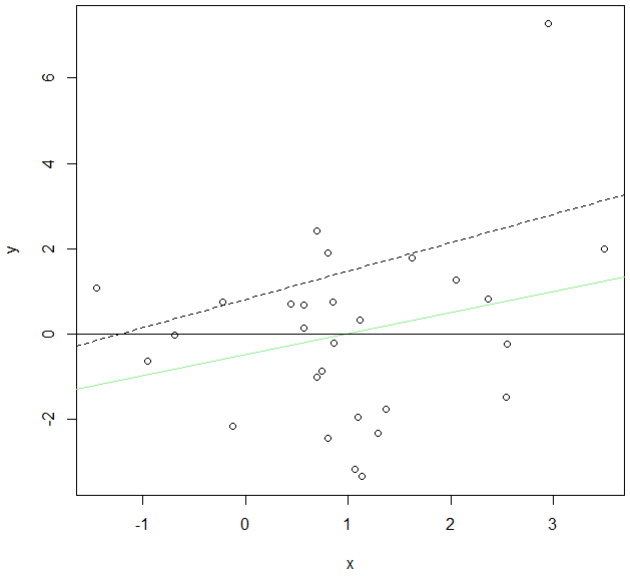


Slika 2.2: Regresijski pravac modela zajedničkih koeficijenata i pravci regresije pojedinih entiteta uz $\alpha_i \neq \alpha_j$ i $\beta_i \neq \beta_j$

pravac za pojedini entitet. Puna linija označuje regresijski pravac dobiven metodom najmanjih kvadrata, korištenjem svih NT podataka. Slika opisuje pristranost koja nastaje zbog heterogenih konstantnih članova kada primjenjujemo model 2.4. U tim slučajevima model zajedničkih koeficijenata, koji zanemaruje heterogene konstantne članove, ne bi se trebao koristiti.
Možemo promatrati slučaj kada je: $\alpha_i \neq \alpha_j$ i $\beta_i \neq \beta_j$. Na slici 2.2 model zajedničkih koeficijenata od svih NT podataka doveo bi do besmislenog modela zato što bi predstavljao prosjek koeficijenata koji se jako razlikuju za pojedinu individuu. Slične strukture pristranosti će se iskristalizirati ako konstantni član i koeficijent smjera variraju kroz vrijeme, a u određenom periodu jednaki su za sve individue.

2.2 Selektivna pristranost

Kada uzorak nije slučajno uzet iz populacije, to također može dovesti do pristranosti. Primjerice, ako istražujemo porez na prihod i izbacimo sve osobe iz istraživanja koje su imale prihode 1.5 puta veće od službeno utvrđene granice siromaštva, korisnici podataka koji koriste komponente zarade kao ovisne varijable stvoriti će selektivnu pristranost. S y bismo mogli označiti zaradu, a s x varijable koje ju opisuju, npr. obrazovanje, inteligenciju itd. Selektivnu pristranost ilustrirati ćemo na simuliranim podacima u **R**-u.



Slika 2.3: Ilustracija selektivne pristranosti

Nacrtati ćemo regresijske pravce kako bismo uvidjeli da korištenje neslučajnoga uzorka stvara pristrane LS procjene. Veza među podatcima može imati oblik:

$$y_i = \beta' \mathbf{x}_i + u_i, \quad i = 1, ..., N, \tag{2.5}$$

gdje je u_i nezavisno distribuiran s očekivanjem nula i varijancom σ_u^2. Pretpostavimo da su iz istraživanja izbačeni podatci kod kojih je vrijednost $y_i \leq 0$. Eksperiment biti će proveden na sljedeći način:

$$y_i = \beta' \mathbf{x}_i + u_i > 0, \quad \text{uključen u istraživanje}$$
$$y_i = \beta' \mathbf{x}_i + u_i \leq 0, \quad \text{isključen iz istraživanja}$$

Na slici 2.3 vidljiva je iscrtana linija regresije koja je procijenjena na "odrezanih" podataka i puna zelena linija koja označuje regresijski pravac koji bismo dobili da su svi podatci

uključeni u model.
Navedeni primjeri pokazuju da unatoč svim prednostima koje panel podatci mogu donijeti, ipak postoje dodatni problemi koji mogu proizaći iz njihova korištenja.

3 Linearni modeli za panel podatke

Pretpostavimo da imamo zabilježena promatranja N individua kroz T vremenskih perioda. S y_{it} označujemo ovisnu varijablu, a s x_{kit} varijable koje opisuju y, $i = 1, ..., N$, $t = 1, ..., T$, $k = 1, ..., K$. Zabilježene vrijednosti od y smatramo slučajnim ishodima eksperimenta, s distribucijom uvjetovanom na vektor karakteristika $\mathbf{x}$ i fiksnog broja parametara θ, $f(y|\mathbf{x}, \theta)$. Kada imamo panel podatke, jedan od ciljeva jest iskoristiti sve raspoložive podatke kako bismo mogli izvesti zaključke o θ. Možemo osvrnuti se na jednostavnu pretpostavku: da je y linearna funkcija od $\mathbf{x}$. Da bismo proveli regresiju metodom najmanjih kvadrata na svih NT promatranja, trebamo pretpostaviti da regresijski parametri primaju vrijednosti jednake za sve "cross-sectionalne" jedinice, za sve vremenske periode. Ako ta pretpostavka ne vrijedi, takvo modeliranje dovesti će nas do krivih zaključaka.
Prvi korak kod baratanja s takvim podatcima jest, odrediti ostaju li parametri koji opisuju slučajan ishod varijable y konstantni za sve i i t.
Linearni model koji je često korišten kako bismo obuhvatili efekte kvalitativnih i kvantitativnih faktora je dan s:

$$y_{it} = \alpha_{it} + \beta'_{it}\mathbf{x}_{it} + u_{it}, \quad i = 1, ..., N \quad t = 1, ..., T, \tag{3.1}$$

gdje su α_{it} i $\beta_{it} = (\beta_{1it}, \beta_{2it}, ..., \beta_{Kit})$ 1×1 i $K \times 1$ dimenzionalni vektori konstanti koji variraju po svim i i svim t, $\mathbf{x}_{it} = (x_{1it}, ..., x_{Kit})$ je $K \times 1$ dimenzionalan vektor neovisnih varijabli, a u_{it} označava grešku. Možemo testirati dva aspekta procjene regresijskih koeficijenata: homogenost koeficijenta β i homogenost konstantnog člana α. Ta procedura sadrži tri osnovna koraka:

1. testirati jesu li koeficijenti smjera i konstantni član istovremeno homogeni za sve individue, kroz sve vremenske trenutke

2. testirati jesu li koeficijenti smjera jednaki

3. testirati jesu li konstantni članovi jednaki

Model 3.1 ima jedino deskriptivnu vrijednost. Ne možemo za njega napraviti procjene parametara niti možemo pomoću njega generirati predikcije zato što je dostupan broj stupnjeva slobode, NT, manji od broja parametara, $NT(K + 1)+$ broj parametara koji opisuju distribuciju od u_{it}. Dodatne pretpostavke moraju biti uvedene. Možemo promatrati nekoliko različitih modela. Jedan od njih je:

$$y_{it} = \alpha_i + \beta'_i\mathbf{x}_{it} + u_{it}, \quad i = 1, ..., N \quad t = 1, ..., T. \tag{3.2}$$

Zvati ćemo ga: model bez restrikcija. Dodatne restrikcije daju nam još tri moguća modela:

$$y_{it} = \alpha_i + \beta' \mathbf{x}_{it} + u_{it}, \quad i = 1, ..., N \quad t = 1, ..., T, \tag{3.3}$$

$$y_{it} = \alpha + \beta_i' \mathbf{x}_{it} + u_{it}, \quad i = 1, ..., N \quad t = 1, ..., T, \tag{3.4}$$

$$y_{it} = \alpha + \beta' \mathbf{x}_{it} + u_{it}, \quad i = 1, ..., N \quad t = 1, ..., T. \tag{3.5}$$

Ponekad je besmisleno pitati se jesu li konstantni članovi jednaki kada su koeficijenti smjera različiti pa ćemo zanemariti tip jednadžbe označen s 3.4. Model 3.3 zvati ćemo: model individualnih očekivanja, a 3.5 zvati ćemo: model zajedničkih koeficijenata.
Neka su:

$$\bar{y}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it}, \tag{3.6}$$

$$\bar{\mathbf{x}}_i = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_{it}, \tag{3.7}$$

aritmetičke sredine od y i $\mathbf{x}$ za i-tu individuu. Procjenitelji dobiveni metodom najmanjih kvadrata za β_i i α_i, u modelu 3.2, dani su sa:

$$\hat{\beta}_i = W_{xx,i}^{-1} W_{xy,i}, \quad \hat{\alpha}_i = \bar{y}_i - \hat{\beta}_i' \bar{\mathbf{x}}_i, \quad i = 1, ..., N, \tag{3.8}$$

gdje su:

$$W_{xx,i} = \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)',$$
$$W_{xy,i} = \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(y_{it} - \bar{y}_i),$$
$$W_{yy,i} = \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2.$$

RSS (residual sum of squares) za i-tu grupu je $RSS_i = W_{yy,i} - W_{xy,i}' W_{xx,i}^1 W_{xy,i}$. Za model bez restrikcija RSS je:

$$S_1 = \sum_{i=1}^{N} RSS_i. \tag{3.9}$$

Metodom najmanjih kvadrata možemo dobiti procjenitelje za modele 3.3 i 3.5.
Pod pretpostavkom da su u_{it} nezavisni i normalno distribuirani za sve i i t, s očekivanjem nula i varijancom σ_u^2, možemo upotrijebiti F test kako bismo provjerili je li opravdano uvođenje restrikcija 3.3 i 3.5.
Analogno, možemo pretpostaviti da koeficijenti α i β ovise samo o vremenskom trenutku t i provesti analognu analizu.

4 Diskretni podatci

Statistički modeli, u kojima ovisne varijable primaju samo diskretne vrijednosti, nazivaju se diskretni, kategorijalni ili modeli kvalitativnoga izbora. Kod takvih modela, analitičar na raspolaganju ima uzorak od N idividua, za koje ima zabilježen podatak o prisutnosti/odsutnosti nekog događaja za svaki od T trenutaka.
Situacija iz stvarnog života, koju možemo opisati ovim modelom, može biti vrlo jednostavna, kao primjerice kupovina automobila. Svaki pojedinac prije kupovine automobila razmišlja o mogućim troškovima i koristima koji dolaze s tom investicijom. Nakon što osoba donese konačnu odluku, njegov potez možemo zabilježiti kao jedinicu ukoliko je kupovina ostvarena ili kao nulu ukoliko je odustao od kupovine automobila. Nakon što skupimo bazu od dovoljnog broja osoba koje su bile suočene s istom dilemom, možemo kreirati model s binarnom varijablom koja nam govori je li osoba kupila automobil (1) ili ne (0). Pitanja na koja nam model binarnoga izbora može dati odgovor, mogu izgledati ovako:
"Kolika je vjerojatnost da pojedinac s određenim karakteristikama posjeduje automobil?"
"Ako se neka varijabla X promijeni za jednu jedinicu, kako to utječe na vjerojatnost posjedovanja automobila?"
Promotrimo slučaj, u kome ovisna varijabla y može primiti samo dvije vrijednosti: 1 ako se promatrani događaj zbio i 0 inače. Pretpostavimo, da imamo skrivenu neprekidnu slučajnu varijablu y^*, koja je linearna funkcija vektora neovisnih varijabli, $\mathbf{x}$,

$$y^* = \beta' \mathbf{x}, \tag{4.1}$$

gdje je greška ν nezavisna od $\mathbf{x}$ s očekivanjem nula. Pretpostavimo, da umjesto skrivene (latentne) varijable y^* promatramo y,

$$y = \begin{cases} 1, & y^* > 0, \\ 0, & y^* \leq 0. \end{cases}$$

$$y^* = y + \nu$$

Zašto je y^* skrivena varijabla? Zamislimo da y^* predstavlja neto korist kupovine automobila. Pojedinac analizira odnos troškova i dobiti koje može prouzročiti kupnovina automobila i odlučuje obaviti kupovinu ako je korisnost pozitivna. Ne možemo egzaktno izmjeriti korisnost do koje je došao pojedinac svojim misaonim procesima, ali možemo zabilježiti njegovu konačnu odluku.

Očekivana vrijednost od y_i je vjerojatnost da će se promatrani događaj ostvariti:

$$\begin{aligned} E[y|\mathbf{x}] &= 1 \cdot P(\nu > -\beta'\mathbf{x}) + 0 \cdot P(\nu \leq -\beta'\mathbf{x}) \\ &= P(\nu > -\beta'\mathbf{x}) \\ &= P(y = 1|\mathbf{x}). \end{aligned}$$

Kada zakon vjerojatnosti za generiranje ν prati distribuciju od dvije točke $(1 - \beta'\mathbf{x})$ i $-\beta'\mathbf{x}$, s vjerojatnostima $\beta'\mathbf{x}$ i $(1 - \beta'\mathbf{x})$ redom, imamo linearni vjerojatnosni model:

$$y = \beta'\mathbf{x} + \nu, \tag{4.2}$$

uz $E\nu = \beta'\mathbf{x}(1 - \beta'\mathbf{x}) + (1 - \beta'\mathbf{x})(-\beta'\mathbf{x}) = 0$.
Kada ν ima standardnu normalnu funkciju gustoće,

$$\phi(\nu) = \frac{1}{\sqrt{2\pi}} e^{(\frac{-\nu^2}{2})},$$

imamo probit model:

$$P(y = 1|\mathbf{x}) = \int_{-\beta'\mathbf{x}}^{\infty} \phi(\nu)d\nu = \int_{-\infty}^{\beta'\mathbf{x}} \phi(\nu)d\nu = \Phi(\beta'\mathbf{x}).$$

Kada je vjerojatnosna funkcija gustoće standardna logistička,

$$f(\nu) = \frac{e^{\nu}}{(1 + e^{\nu})^2} = [(1 + e^{\nu})(1 + e^{-\nu})]^{-1},$$

onda imamo logit model:

$$P(y = 1|\mathbf{x}) = \int_{-\beta'\mathbf{x}}^{\infty} \frac{e^{\nu}}{(1 + e^{\nu})^2} d\nu = \frac{e^{\beta'\mathbf{x}}}{1 + e^{\beta'\mathbf{x}}}. \tag{4.3}$$

Neka je $F(\beta'\mathbf{x}) = E(y_i|\mathbf{x})$. Tri često korištena modela s binarnim izborom, možemo zapisati s jednim indeksom w na ovaj način:
Linearni model

$$F(w) = w \tag{4.4}$$

Probit model

$$F(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} e^{\frac{u^2}{2}} du = \Phi(w) \tag{4.5}$$

Logit model

$$F(w) = \frac{e^w}{1 + e^w} \tag{4.6}$$

Linearni vjerojatnosni model specijalan je slučaj modela linearne regresije s heteroskedastičnom varijancom, $\beta'\mathbf{x}(1-\beta'\mathbf{x})$. Može se procijeniti metodom najmanjih kvadrata ili MNK s težinama. Linearni model jednostavan je za procjenu i za tumačenje rezultata, no ima i neke nedostatke. Vrijednosti koje dobijemo za ovisnu varijablu ne moraju biti iz intervala $(0,1)$, što nije dobro ako želimo da nam taj rezultat predstavlja vjerojatnost nekog događaja (npr. kupovine automobila).
Zato što su to funkcije distribucije, kod logit i probit modela, ovisna varijabla prima vrijednosti između 0 i 1.
Za slučajan uzorak od N individua, $(y_i, \mathbf{x}_i)$, $i = 1, ..., N$, funkcija vjerodostojnosti za tri modela može se općenito zapisati kao:

$$L = \prod_{i=1}^{N} [F(\beta'\mathbf{x})]^{y_i} [1 - F(\beta'\mathbf{x})]^{1-y_i}. \tag{4.7}$$

Diferenciranjem logaritmirane funkcije vjerodostojnosti, dobivamo vektor prvih derivacija i matricu s derivacijama drugog reda:

$$\frac{\partial logL}{\partial \beta} = \sum_{i=1}^{N} \frac{y_i - F(\beta'\mathbf{x}_i)}{F(\beta'\mathbf{x}_i)[1 - F(\beta'\mathbf{x}_i)]} F'(\beta'\mathbf{x}_i)\mathbf{x}_i \tag{4.8}$$

i

$$\begin{aligned}\frac{\partial^2 logL}{\partial \beta \partial \beta'} = &\Bigg\{ - \sum_{i=1}^{N} \left[\frac{y_i}{F^2(\beta'\mathbf{x}_i)} + \frac{1 - y_i}{[1 - F(\beta'\mathbf{x}_i)]^2} \right] [F(\beta'\mathbf{x}_i)]^2 + \\ &\sum_{i=1}^{N} \left[\frac{y_i - F(\beta'\mathbf{x}_i)}{F(\beta'\mathbf{x}_i)[1 - F(\beta'\mathbf{x}_i)]} \right] F''(\beta'\mathbf{x}_i) \Bigg\} \mathbf{x}_i \mathbf{x}_i'.\end{aligned}$$

Ako je funkcija vjerodostojnosti 4.7 konkavna, u gore opisanim modelima jest, onda možemo upotrijebiti Newton-Raphsonovu metodu:

$$\hat{\beta}^{(j)} = \hat{\beta}^{(j-1)} - \left(\frac{\partial^2 logL}{\partial \beta \partial \beta'} \right)^{-1}_{\beta = \hat{\beta}^{(j-1)}} \left(\frac{\partial logL}{\partial \beta} \right)_{\beta = \hat{\beta}^{(j-1)}}$$

ili metodu skoringa:

$$\hat{\beta}^{(j)} = \hat{\beta}^{(j-1)} - \left[E \frac{\partial^2 logL}{\partial \beta \partial \beta'} \right]^{-1}_{\beta = \hat{\beta}^{(j-1)}} \left(\frac{\partial logL}{\partial \beta} \right)_{\beta = \hat{\beta}^{(j-1)}}$$

za traženje procjenitelja metodom maksimalne vjerodostojnosti (MLE) od β s proizvoljnim početnim vrijednostima $\hat{\beta}^{(0)}$, gdje $\hat{\beta}^{(j)}$ označava j-to iterativno rješenje.

4.1 Parametarski pristup statičkim modelima s heterogenosti

Panel podatci daju nam mogućnost razlikovanja modela individualnoga ponašanja od modela prosječnoga ponašanja grupe individua.
Radi jednostavnosti, pretpostavimo da je heterogenost među individuama neovisna o vremenu i da individualni efekt možemo prikazati kroz grešku modela ν_{it} iz 4.1 kao $\alpha_i + u_{it}$. Ako α_i tretiramo kao konstantu, vrijedi: $Var(\nu_{it}|\alpha_i) = Var(u_{it}) = \sigma_u^2$. Ako α_i smatramo slučajnima, onda pretpostavljamo da $E\alpha_i = Eu_{it} = E\alpha_i u_{it} = 0$ i $Var(\nu_{it}) = \sigma_u^2 + \sigma_\alpha^2$. Radi jednostavnosti, normaliziramo varijancu σ_u^2 od u tako da bude jednaka 1.

4.1.1 Modeli fiksnih efekata

Ako pretpostavimo da je individualni efekt α_i, fiksan, onda su i α_i i β nepoznati parametri koje trebamo procijeniti za model:

$$P(y_{it} = 1|\mathbf{x}_{it}, \alpha_i) = F(\beta'\mathbf{x}_{it} + \alpha_i).$$

Ako T teži prema beskonačnosti, MLE je konzistentan. No, T je obično mali kod panel podataka. Imamo samo ograničen broj promatranja za procjenu α_i. Bilo koja procjena α_i besmislena je ako planiramo koristiti procjenitelje zbog svojstava koje imaju za velike uzorke. Radi toga koncentrirati ćemo se na procjenu parametra β.
Općenito, ne postoji jednostavna transformacija za eliminaciju parametara α_i u nelinearnom modelu. MLE za α_i i β nisu nezavisni jedno od drugoga u modelima diskretnoga izbora. Kada je T fiksan, nekonzistentnost od $\hat{\alpha}_i$ prenosi se na MLE od β. Čak i kada N teži prema beskonačno, MLE od β ostaje nekonzistentan.
Demonstrirati ćemo nekonzistentnost od MLE za β, na primjeru logit modela. Logaritmirana funkcija vjerodostojnosti za taj model je:

$$logL = -\sum_{i=1}^{N}\sum_{t=1}^{T} log[1 + e^{\beta'\mathbf{x}_{it}+\alpha_i}] + \sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}(\beta'\mathbf{x}_{it} + \alpha_i). \tag{4.9}$$

Radi jednostavnosti, promatrati ćemo specijalan slučaj kada je $T = 2$ i s jednom neovisnom varijablom, $x_{i1} = 0$ i $x_{i2} = 1$. Jednadžbe prvih derivacija dane su sa:

$$\frac{\partial logL}{\partial \beta} = \sum_{i=1}^{N}\sum_{t=1}^{2}\Big[-\frac{e^{\beta x_{it}+\alpha_i}}{1 + e^{\beta' x_{it}+\alpha_i}} + y_{it}\Big]x_{it} = \sum_{i=1}^{N}\Big[-\frac{e^{\beta+\alpha_i}}{1 + e^{\beta+\alpha_i}} + y_{i2}\Big] = 0, \tag{4.10}$$

$$\frac{\partial logL}{\partial \alpha_i} = \sum_{t=1}^{2}\Big[-\frac{e^{\beta x_{it}+\alpha_i}}{1 + e^{\beta x_{it}+\alpha_i}} + y_{it}\Big] = 0. \tag{4.11}$$

Rješavanjem 4.11 imamo:

$$\hat{\alpha}_i = \begin{cases} \infty, & y_{i1} + y_{i2} = 2, \\ -\infty, & y_{i1} + y_{i2} = 0. \\ -\frac{\beta}{2}, & y_{i1} + y_{i2} = 1. \end{cases}$$

Uvrštavanjem gornjeg izraza u 4.10 i ako s n_1 označimo broj individua $y_{i1} + y_{i2} = 1$ i s n_2 broj individua $y_{i1} + y_{i2} = 2$ imamo:

$$\sum_{i=1}^{N} \frac{e^{\beta+\alpha_i}}{1+e^{\beta+\alpha_i}} = n_1 \frac{e^{\beta/2}}{1+e^{\beta/2}} + n_2 = \sum_{i=1}^{N} y_{i2}. \tag{4.12}$$

Prema tome:

$$\hat{\beta} = 2\left\{ log\left(\sum_{i=1}^{N} y_{i2} - n_2 \right) - log\left(n_1 + n_2 - \sum_{i=1}^{N} y_{i2} \right) \right\}. \tag{4.13}$$

Prema zakonu velikih brojeva izraz:

$$p \lim_{N\to\infty} \frac{1}{N} (\sum_{i=1}^{N} y_{i2} - n_2)$$

odgovara izrazu

$$\frac{1}{N} \sum_{i=1}^{N} P(y_{i1} = 0, y_{i2} = 1 | \beta, \alpha_i) = \frac{1}{N} \sum_{i=1}^{N} \frac{e^{\beta+\alpha_i}}{(1+e^{\beta+\alpha_i})(1+e^{\alpha_i})},$$

Isto vrijedi za:

$$p \lim_{N\to\infty} \frac{1}{N} (n_1 + n_2 - \sum_{i=1}^{N} y_{i2})$$

i

$$\frac{1}{N} \sum_{i=1}^{N} P(y_{i1} = 1, y_{i2} = 0 | \beta, \alpha_i) = \frac{1}{N} \sum_{i=1}^{N} \frac{e^{\alpha_i}}{(1+e^{\beta+\alpha_i})(1+e^{\alpha_i})}.$$

Supstitucijom $\hat{\alpha}_i = -\frac{\beta}{2}$ u gornja dva limesa dobivamo:

$$p \lim_{N\to\infty} \hat{\beta} = 2\beta, \tag{4.14}$$

što nije konzistentno.

Neyman i Scott (1948) predložili su princip za nalaženje konzistentnog procjenitelja parametra β, u prisutnosti parametra α_i. Postupak koji su definirali zvati ćemo *Neyman-Scott princip*. Ideja je pronaći K funkcija:

$$\Psi_{N_j}(\mathbf{y}_1, ..., \mathbf{y}_N | \beta), \quad j = 1, ..., K, \tag{4.15}$$

koje su neovisne o parametrima α_i i imaju svojstvo, da kada su β stvarne vrijednosti, $\Psi_{N_j}(\mathbf{y}_1, ..., \mathbf{y}_N|\beta)$ konvergiraju prema nuli po vjerojatnosti kada N teži u beskonačno. U navedenim jednadžbama pretpostavljamo da su y_{it}, odnosno $\mathbf{y}_i$, međusobno nezavisni za svaki t. Procjenitelj $\hat{\beta}$, kojega dobijemo rješavanjem $\Psi_{N_j}(\mathbf{y}_1, ..., \mathbf{y}_N|\beta) = 0$ biti će konzistentan pod određenim uvjetima. Na primjer, $\hat{\beta}^* = \frac{1}{2}\hat{\beta}$ za logit model fiksnih efekata 4.9 takav je procjenitelj.

U slučaju linearnoga vjerojatnosnoga modela, uzimanje prve diferencije po vremenskoj komponenti ili s obzirom na individualno očekivanje, eliminira individualni efekt. LS regresijom diferenciranih jednadžbi dobivamo konzistentan procjenitelja za β, kada N teži u beskonačno.

Kod nelinearnih modela, jednostavne funkcije za Ψ nije uvijek lako naći. Općenito, mi ne znamo vjerojatnosni limes od MLE za logit model fiksnih efekata. No ako minimalna dovoljna statistika (vidi [10]) τ_i za parametar α_i postoji i nezavisna je od strukturalnog parametra β, onda uvjetna gustoća

$$f^*(\mathbf{y}_i|\beta, \tau_i) = \frac{f(\mathbf{y}_i|\beta, \alpha_i)}{g(\tau_i|\beta, \alpha_i)} \quad za\ g(\tau_i|\beta, \alpha_i) > 0, \tag{4.16}$$

više ne ovisi o α_i. Andersen (1970,1973) pokazao je da maksimiziranjem uvjetne gustoće kao funkcije od $\mathbf{y}_1, ..., \mathbf{y}_N$ za dane $\tau_1, ..., \tau_N$,

$$\prod_{i=1}^{N} f^*(\mathbf{y}_i|\beta, \tau_i), \tag{4.17}$$

dobijemo uvjete $\Psi_{N_j}(\mathbf{y}_1, ..., \mathbf{y}_N|\hat{\beta}, \tau_1, ..., \tau_N) = 0$ za $j = 1, ..., K$. Rješavanje tih funkcija dati će konzistentan procjenitelj parametra β pod blagim uvjetima regularnosti.

Ilustrirati ćemo uvjetnu metodu maksimalne vjerodostojnosti na primjeru logit modela. Gustoća od $\mathbf{y}_i$ je:

$$p(\mathbf{y}_i) = \frac{exp\Big\{\alpha_i \sum_{t=1}^{T} y_{it} + \beta' \sum_{t=1}^{T} \mathbf{x}_{it} y_{it}\Big\}}{\prod_{t=1}^{T} [1 + exp(\beta' \mathbf{x}_{it} + \alpha_i)]}. \tag{4.18}$$

Jasno je da $\sum_{t=1}^{T} y_{it}$ je minimalna dovoljna statistika za α_i. Uvjetna vjerojatnost za $\mathbf{y}_i$, uz dani $\sum_{t=1}^{T} y_{it}$ je:

$$P\left(\mathbf{y}_i \Big| \sum_{t=1}^{T} y_{it}\right) = \frac{exp\left[\beta' \sum_{t=1}^{T} \mathbf{x}_{it} y_{it}\right]}{\sum_{D_{ij} \in \tilde{B}_i} exp\Big\{\beta' \sum_{t=1}^{T} \mathbf{x}_{it} y_{it}\Big\}}, \tag{4.19}$$

gdje je:

$$\tilde{B}_i = \left\{ D_{ij} = (d_{ij1}, ..., d_{ijT}) | d_{ijt} = 0 \text{ ili } 1 \text{ i } \sum_{t=1}^{T} d_{ijt} = \sum_{t=1}^{T} y_{it} = s, j = 1, 2, ..., \frac{T!}{s!(T-s)!} \right\},$$

skup svih mogućih različitih nizova $(d_{ij1}, ..., d_{ijT})$ koji zadovoljavaju

$$\sum_{t=1}^{T} d_{ijt} = \sum_{t=1}^{T} y_{it} = s.$$

Sa skupom $\tilde{B}_i$, koji se nalazi kod sume u jednadžbi 4.19, želimo pobrojati sve moguće kombinacije nula i jedinica s kojima se može realizirati ovisna varijabla y_{it} kroz T vremenskih trenutaka. Ako sumiramo elemente y_{it} po t, istu vrijednost sume možemo dobiti i za različite permutacije realizacija nula i jedinica. Iznos te sume ovisi o broju realizacije jedinica. Ako broj jedinica označimo sa s tada ukupan broj različitih redoslijeda realizacija s jednica i $T - s$ nula jednak je $\binom{T}{s}$.
Jednadžba 4.19 u uvjetnom je logit obliku, s alternativnim skupovima $(\tilde{\beta}_i)$, koji variraju po i. Oni ne ovise o parametru α_i. Prema tome, uvjetni MLE od β može se dobiti koristeći standardne ML logit programe i on je konzistentan uz neke blage uvjete.
Opisani postupak ilustrirati ćemo na pojednostavljenom primjeru. Pretpostaviti ćemo da je $T = 2$. Realizacija ovisnih varijabli od interesa dana je s $y_{i1} + y_{i2} = 1$. Radi ljepšeg zapisa definirati ćemo ω_i koji je jednak:

$$\omega_i = \begin{cases} 1, & (y_{i1}, y_{i2}) = (0, 1), \\ 0, & (y_{i1}, y_{i2}) = (1, 0). \end{cases}$$

Uvjetna vjerojatnost od $w_i = 1$, uz dani $y_{i1} + y_{i2} = 1$ je:

$$\begin{aligned} P(\omega_i = 1 | y_{i1} + y_{i2} = 1) &= \frac{P(\omega_i = 1)}{P(\omega_i = 1) + P(\omega_i = 0)} \\ &= \frac{exp[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]}{1 + exp[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]} \\ &= F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]. \end{aligned}$$

Gornja jednadžba u formi je binarne logit funkcije u kojoj su dva moguća ishoda $(0, 1)$ i $(1, 0)$, s neovisnim varijablama $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$. Uvjetna logaritmirana funkcija vjerodostojnosti glasi:

$$logL^* = \sum_{i \in \tilde{B}_1} \{\omega_i log F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})] + (1 - \omega_i) log(1 - F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})])\}, \tag{4.20}$$

gdje je $\tilde{B}_1 = \{i | y_{i1} + y_{i2} = 1\}$.
Za slučaj $T > 2$ ne gubimo na općenitosti biranjem niza $D_{i1} = (d_{i11}, ..., d_{i1T})$, $\sum_{t=1}^{T} d_{i1t} =$

$\sum_{t=1}^{T} y_{it} = s$, $1 \leq s \leq T-1$. Sada možemo zapisati uvjetnu vjerojatnost 4.19 kao:

$$P\left(\mathbf{y}_i \Big| \sum_{t=1}^{T} y_{it}\right) = \frac{exp\Big\{\beta'\{\sum\limits_{t=1}^{T} \mathbf{x}_{it}(y_{it} - d_{i1t})\}\Big\}}{1 + \sum\limits_{D_{ij} \in (\tilde{B}_i - D_{i1})} exp\Big\{\sum\limits_{t=1}^{T} \mathbf{x}_{it}(d_{ijt} - d_{i1t})\Big\}}. \tag{4.21}$$

Uvjetna logaritmirana funkcija vjerodostojnosti poprima oblik:

$$\begin{aligned} logL^* = \sum_{i \in C} \Bigg\{ & \beta' \sum_{t=1}^{T} \mathbf{x}_{it}(y_{it} - d_{i1t}) \\ & - log\left[1 + \sum_{D_{ij} \in (\tilde{B}_i - D_{i1})} exp\left\{\beta' \sum_{t=1}^{T} \mathbf{x}_{it}(d_{ijt} - d_{i1t})\right\}\right]\Bigg\}, \end{aligned}$$

gdje je $C = \{i | \sum_{t=1}^{T} y_{it} \neq T, \sum_{t=1}^{T} y_{it} \neq 0\}$.

Iako možemo pronaći jednostavne transformacije linearnih vjerojatnosti i logit modela koje će zadovoljavati Neyman-Scott princip, ne možemo pronaći jednostavne funkcije za parametre koji nas zanimaju, a koje su nezavisne od parametra α_i za probit modele. Čini se da ne postoji konzistentan procjenitelj za β, za probit modele fiksnih efekata.

4.1.2 Modeli slučajnih efekata

Kada su individualni specifični efekti α_i tretirani kao slučajni, još uvijek možemo dobiti procjenitelje fiksnih efekata za procjenu strukturalnog parametra β. Asimptotska svojstva tog procjenitelja ostaju nepromijenjena. No ako su α_i slučajni, ali ih tretiramo kao da su fiksni, posljedica je, u najboljem slučaju, gubitak efikasnosti u procjeni β, ali možemo dobiti procjenitelje koji nisu konzistentni.
Model slučajnih efekata dan je formulom:

$$y_{it} = \beta' \mathbf{x}_{it} + \alpha_i + u_{it}, \tag{4.22}$$

$$P(y_{it} = 1 | \mathbf{x}_{it}, \alpha_i) = F(\beta' \mathbf{x}_{it}, \alpha_i). \tag{4.23}$$

Ključne pretpostavke toga modela:

- α_i i $\mathbf{x}_{it}$ su nezavisni
- $\mathbf{x}_{it}$ strogo su exogeni (funkciju vjerodostojnosti možemo zapisati kao produkt pojedinačnih funkcija vjerodostojnosti)

- α_i ima normalnu distribuciju, s očekivanjem 0 i varijancom σ_α^2
- $y_{i1}, ..., y_{iT}$ nezavisni su uvjetno na $(\mathbf{x}_{it}, \alpha_i)$

Kada su α_i nezavisni od $\mathbf{x}_i$ i kada su slučajan uzorak iz distribucije G, indeksirane konačnim brojem parametara δ, log funkcija vjerodostojnosti postaje:

$$logL = \sum_{i=1}^{N} log \int \prod_{t=1}^{T} F(\beta' \mathbf{x}_{it} + \alpha)^{y_{it}} [1 - F(\beta' \mathbf{x}_{it} + \alpha)]^{1-y_{it}} dG(\alpha|\delta), \tag{4.24}$$

gdje je $F(\cdot)$ distribucija greške uvjetno na $\mathbf{x}_i$ i na α_i. Jednadžba 4.24 funkcija je konačnoga broja parametara (β', δ'). Prema tome, maksimiziranjem 4.24, pod slabim uvjetima regularnosti, dobiti ćemo konzistentne procjenitelje za β i δ kada N teži u beskonačno. Logaritmirana funkcija vjerodostojnosti za svih N entiteta dana je s:

$$logL = \sum_{i=1}^{N} logL_i(y_{i1}, ..., y_{iT}|\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}, \beta, \sigma_\alpha^2). \tag{4.25}$$

Ako je α_i koreliran s $\mathbf{x}_{it}$, maksimiziranjem 4.25 neće eliminirati pristranost izostavljene varijable. Da bismo dopustili zavisnost između α i $\mathbf{x}$, moramo specificirati distribuciju $G(\alpha|\mathbf{x})$ za α uvjetno na $\mathbf{x}$ i onda promatrati sljedeću funkciju vjerodostojnosti:

$$logL = \sum_{i=1}^{N} log \int \prod_{t=1}^{T} F(\beta' \mathbf{x}_{it} + \alpha)^{y_{it}} [1 - F(\beta' \mathbf{x}_{it} + \alpha)]^{1-y_{it}} dG(\alpha|\mathbf{x}). \tag{4.26}$$

Chamberlain predložio je da pretpostavimo:

$$\alpha_i = \sum_{t=1}^{T} \mathbf{a}_t' \mathbf{x}_{it} + \eta_i = \mathbf{a}' \mathbf{x}_i + \eta_i, \tag{4.27}$$

gdje je $\mathbf{a}' = (\mathbf{a}_1', ..., \mathbf{a}_T')$, $\mathbf{x}_i' = (\mathbf{x}_{i1}, ..., \mathbf{x}_{iT})$ i η_i je rezidual. Pretpostavljamo da je regresijska funkcija $E[\alpha_i|\mathbf{x}_i]$ linearna, da je η_i nezavisno od $\mathbf{x}_i$ i da η_i ima specifičnu vjerojatnosnu distribuciju.
Uzimajući u obzir navedene pretpostavke, log funkcija vjerodostojnosti za slučajne efekte glasi:

$$logL = \sum_{i=1}^{N} log \int \prod_{t=1}^{T} F(\beta' \mathbf{x}_{it} + \mathbf{a}' \mathbf{x}_i + \eta)^{y_{it}} [1 - F(\beta' \mathbf{x}_{it} + \mathbf{a}' \mathbf{x}_i + \eta)]^{1-y_{it}} dG^*(\eta), \tag{4.28}$$

gdje je G^* funkcija distribucije za η. Razlika između 4.28 i 4.25 jedino je u inkluziji termina $\mathbf{a}'\mathbf{x}_i$, da bismo uhvatili zavisnost između parametara α_i i $\mathbf{x}_i$. Prema tome, što se tiče procjene za 4.28 i 4.25, osnovne karakteristike jednake su. Zbog toga, raspraviti ćemo jedino proceduru za procjenu općenitijeg modela 4.28.

Maksimiziranje 4.28 uključuje integraciju T dimenzija, što može biti nezgrapno u računu. Alternativni pristup, koji pojednostavljuje izračun MLE, je pretpostaviti da su komponente greške $\nu_{it} = \alpha_i + u_{it}$ nezavisne i normalno distribuirane, s očekivanjem α_i i varijancom 1 i da imaju vjerojatnosnu gustoću $\phi(\nu_{it}|\alpha_i)$. Tada

$$\begin{aligned} P(y_{i1}, ..., y_{iT}) &= \int_{c_{i1}}^{b_{i1}} ... \int_{c_{iT}}^{b_{iT}} \prod_{t=1}^{T} \phi(\nu_{it}|\alpha_i) G(\alpha_i|\mathbf{x}_i) d\alpha_i d\nu_{i1} ... d\nu_{iT} \\ &= \int_{-\infty}^{\infty} G(\alpha_i|\mathbf{x}_i) \prod_{t=1}^{T} [\Phi(b_{it}|\alpha_i) - \Phi(c_{it}|\alpha_i)] d\alpha_i, \end{aligned}$$

gdje je $\Phi(\cdot|\alpha_i)$ kumulativna funkcija distribucije (cdf) od $\phi(\cdot|\alpha_i)$, $c_{it} = -\beta'\mathbf{x}_{it}$, $b_{it} = \infty$, ako je $y_{it} = 1$ i $c_{it} = -\infty$, $b_{it} = -\beta'\mathbf{x}_{it}$, ako je $y_{it} = 0$ i $G(\alpha_i|\mathbf{x}_i)$ je vjerojatnosna funkcija gustoće od α_i za dani $\mathbf{x}_i$. Ako pretpostavimo da je $G(\alpha_i|\mathbf{x}_i)$ normalno distribuirana s varijancom σ_α^2, gornji izraz reducira T-dimenzionalnu integraciju na jedan integral, čiji je integrand produkt jedne normalne gustoće i T razlika normalnih cdf-ja za koje imamo visoko pouzdane aproksimacije. Na primjer, Butler i Moffit (1982) predložili su Gaussovu kvadraturu kako bi poboljšali učinkovitost izračunavanja na računalu.

$$\pi^{-1/2} \sum_{m=1}^{M} w_m \prod_{t=1}^{T} [\Phi(\mathbf{x}_{it}\beta + \sqrt{2}\sigma_c g_m)]^{y_{it}} [1 - \Phi(\mathbf{x}_{it}\beta + \sqrt{2}\sigma_c g_m)]^{1-y_{it}} \tag{4.29}$$

gdje je M broj čvorova, w_m unaprijed određena težina i g_m unaprijed određeni (prespecified) čvor. Čvorove određujemo tako da dobijemo što bolju aproksimaciju normalne distribucije. Lista čvorova i M mogu se naći u literaturi.

5 Testiranje fiksnih i slučajnih efekata

Kako znamo postoji li fiksni ili slučajni efekt u našim podatcima?
Postojanje fiksnoga efekta možemo testirati F testom[1]. On uspoređuje model fiksnih efekata s OLS[2] regresijom kako bismo vidjeli koliko model fiksnih efekata može poboljšati "goodnes-of-fit".
Testovi navedeni u nastavku mogu se primijeniti kod linearnih modela za panel podatke, dok kod generaliziranih modela nisu primjenjivi. Postojanje slučajnog efekta testira se Breusch-Pagan-Lagrange multiplier (LM) testom. On uspoređuje model slučajnih efekata s OLS modelom.

F-test za fiksne efekte

U regresiji $y_{it} = \alpha + \mu_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}$ nulta je hipoteza da su svi parametri 0-1 varijabli, osim jednoga, jednaki nuli.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_{n-1} = 0$$

Alternativna hipoteza jest, da je barem jedan parametar 0-1 varijable različit od nule. F test baziran je na gubitku goodnes-of-fit. On uspoređuje LSDV s OLS modelom konstantnih koeficijenata i ispituje "goodnes-of-fit" mjere (SSE ili R^2) i njihovu promjenu.

$$F(n-1, nT-n-k) = \frac{R^2_{LSDV} - R^2_{pooled}(n-1)}{\frac{1-R^2_{LSDV}}{nT-n-k}} \tag{5.1}$$

Ako je nul hipoteza odbačena, barem jedan individualni konstantni član u_i nije nula, onda zaključujemo da postoji značajan fiksni efekt ili značajan porast u goodnes-of-fit, kod modela fiksnih efekata pa je taj model bolji izbor od OLS modela konstantnih koeficijenata.

Breusch-Pagan LM test za slučajne efekte

LM test ispituje jesu li individualne ili vremenske specifične komponente varijance jednake nuli:

$$H_0 : \sigma^2_u = 0.$$

[1] statistički test u kojem testna statistika ima F distribuciju
[2] ordinary least square

LM statistika ima χ^2 distribuciju s jednim stupnjem slobode.

$$LM_u = \frac{nT}{2(T-1)}\left[\frac{T^2\bar{e}'e}{e'e} - 1\right]^2 \sim \chi^2(1), \tag{5.2}$$

gdje je e' $n \times 1$ vektor srednjih vrijednosti reziduala modela konstantnih koeficijenata, $e'e$ SSE od modela konstantnih koeficijenata.
Ako je nulta hipoteza odbačena, možemo zaključiti da postoji značajan slučajan efekt u panel podatcima i da se model slučajnih efekata bolje nosi s heterogenosti nego model konstantnih koeficijenata.

Hausmanov test za usporedbu fiksnoga i slučajnoga efekta

Kako znamo koji je efekt, slučajni ili fiksni, značajniji u podatcima?
Hausmanov test uspoređuje modele fiksnih i slučajnih efekata pod nultom hipotezom, da su individualni efekti nekorelirani s bilo kojim od regresora u modelu. Ako nulta hipoteza o nekorelaciji nije odbačena, LSDV i GLS konzistentni su, ali LSDV nije efikasan. U suprotnom, LSDV je konzistentan, ali GLS je nekonzistentan i pristran. Statistika Hausmanovog testa ima χ^2 distribuciju sa k stupnjeva slobode:

$$LM = (b_{LSDV} - b_{random})'\hat{W}^{-1}(b_{LSDV} - b_{random}) \sim \chi^2(k),$$
$$\hat{W} = Var((b_{LSDV} - b_{random})) = Var((b_{LSDV}) - Var(b_{random}),$$

gdje je $\hat{W}$ razlika u procjenjenim matricama kovarijanci dvaju modela.
Konstantni član i 0-1 varijable trebali bi biti isključeni iz ovoga računa. Formula kaže, da Hausmanov test ispituje je li procjena slučajnoga efekta neznatno drugačija od nepristrane procjene fiksnoga efekta. Ako je nulta hipoteza o nekorelaciji odbačena, možemo zaključiti da individualni efekt u_i je značajno koreliran s barem jednim od regresora u modelu i zbog toga model slučajnih efekata je problematičan. Pametnije bi bilo uzeti u obzir model fiksnih efekata.
Negativna strana ovoga testa jest, može se dogoditi da razlika kovarijanci matrica W ne bude pozitivno definitna i tada možemo zaključiti da nula nije odbačena.

Chow test za "poolability"

"Poolability" ispituje jesu li koeficijenti smjera jednaki za sve individue ili kroz sve vremenske trenutke. Jednostavan test za "poolability" nadogradnja je Chow testa. Nulta hipoteza kaže: koeficijent smjera nekog regresora isti je bez obzira o kojoj se individui radi, za svih k regresora.

$$H_0 : \beta_{ik} = \beta_k$$

Sjetimo se, koeficijent smjera konstantan je u modelima fiksnih i slučajnih efekata. Kod ovih dvaju modela, jedino konstantni članovi i varijance greške mijenjaju se. Statistika ovog testa da je s:

$$F[(n-1)(k+1), n(T-k-1)] = \frac{e'e - \sum e_i' e_i (n-1)(k+1)}{(\sum e_i' e_i)/(n(T-k-1))}. \tag{5.3}$$

Ako je nulta hipoteza odbačena, svaka individua ima različiti koeficijent smjera za sve regresore. U tom slučaju, trebali bismo probati modele slučajnih koeficijenata ili hijerarhijski regresijski model.

Akaike informacijski kriterij (AIC[3])

Model koji dobijemo za neke podatke ne može savršeno precizno opisati stvarnost. Među svim modelima koje promatramo, tražimo one koji najbolje aproksimiraju stvarnost, tj. pokušavamo minimizirati gubitak informacija. Kullback i Leibler (1951) razvili su kriterij koji predstavlja gubitak informacija prilikom aproksimiranja stvarnosti. Nazivamo ga Kullback-Leibler information (KLI). Desetak godina kasnije, Akaike (1973) predložio je da se KLI koristi kod izbora modela. Njegov informacijski kriterij dan je s formulom:

$$AIC = 2k - 2ln(L), \tag{5.4}$$

gdje je k broj slobodnih parametara za procjenu, a L maksimizirana vrijednost funkcije vjerodostojnosti za promatrani model.
Sam za sebe, broj koji dobijemo iz formule 5.4 za pojedini model, ne znači nam ništa dok ga ne usporedimo s vrijednostima koje dobijemo za druge modele. Pomoću AIC-a moći ćemo odrediti model s najmanjim gubitkom informacija među unaprijed odabranim modelima. Bitno je da prije korištenja toga kriterija imamo već nekoliko modela za koje smatramo da dobro predstavljaju podatke. Također bitno je da svi modeli koje uspoređujemo imaju istu zavisnu varijablu. Prednost ovog kriterija su objetivnost, jednostavnost tumačenja i činjenica da je implementiran u većini statističkih softwarea. Kao nedostatak, možemo navesti činjenicu da rezultati odabira modela ovim kriterijem najviše ovise o modelima koje smo prethodno odabrali nekim drugim metodama kao najbolje.

Značenje fiksnoga i slučajnog efekta

Kako tumačimo te efekte? Spomenute efekte ilustrirati ćemo jednim primjerom. Želimo napraviti regresiju proizvodnje nekoliko poduzeća, s njihovim investicijama kao neovisnim varijablama. Fiksni efekt mogao bi se tumačiti kao početni proizvodni kapaciteti

[3]Akaike Information Criterion

kada nikakva investicija nije još napravljena. Slučajni efekt mogli bismo gledati kao stabilnost proizvodnje. Ako proizvodnja ima velike fluktuacije, npr. kod jednog poduzeća komponenta varijance je veća nego kod drugih, čak i kad produktivnost ostaje ista za to poduzeće, to nam može ilustrirati prisutnost slučajnoga efekta. Model slučajnih efekata ima "slučajni termin greške", sastavljen od slučajne greške i "slučajnoga konstantnog člana", koji mjeri koliko se konstantni član individue razlikuje od konstantnog člana koji se odnosi na sve podatke. Ključna razlika između fiksnoga i slučajnoga efekta nije u tome je li neopažena heterogenost pripisana konstantnom članu ili komponenti varijance nego u tome je li individualna specifična komponenta greške u vezi s regresorima.

6 Primjena teorije na podatcima

Ulaganje poduzeća u istraživanje i razvoj stvaraju uvjete da to poduzeće razvije inovaciju u narednih nekoliko godina. Inovacije mogu biti: razvijanje novih proizvoda, poboljšanje kvalitete dosadašnjih proizvoda ili uvođenje novih procesa proizvodnje koji smanjuju troškove toga poduzeća.
Hoće li neko poduzeće proizvesti inovaciju, pokušati ćemo modelirati pomoću podataka koji su korišteni na nastavi profesora Williama Greena na Odjelu za ekonomiju, Stern School of Business, New York University. Panel se sastoji od 1270 poduzeća koje promatramo kroz pet godina, od 1984.do 1988. Panel je uravnotežen, nemamo nedostajućih podataka, što znači da imamo ukupno 6350 promatranja.[4] Opis varijabli dostupan je na [11].
Takve baze podataka nastaju na način da, poduzeća se kontaktiraju putem elektroničke pošte, u kojoj je priložena anketa koju trebaju ispuniti. Poduzeća nisu obavezna ispuniti anketu pa je jasno da je odaziv na sudjelovanje u ovakvom istraživanju mali (prema [6] oko 25% poduzeća dostavi ispunjenu anketu). Problem kod prikupljanja podataka jest i taj, da bismo mogli sastaviti panel, moramo dobiti odgovor na anketu od istih poduzeća za nekoliko uzastopnih godina.

Među svim do sada spomenutim modelima, pokušat ćemo odabrati onaj koji najbolje odgovara podatcima koje imamo na raspolaganju. Za ovisnu varijablu uzeli smo **IP** koja označava pojavu novoga proizvoda ili inovacije ako se realizira s jedinicom, odnosno odsustvo inovacije ako se realizira s nulom. Poduzeća koja su opisana spomenutim varijablama dolaze iz četiriju grana djelatnosti, koje su naznačene u podatcima s 0-1 varijablama: prehrambena industrija (**FOOD**), industrija potrošačkih dobara (**CONSGOOD**), investicijske industrije (**INVGOOD**)i prerade sirovina (**RAWMTL**). U panelu imamo kategorijalnu varijablu koja označuje vremensko razdoblje (**YEAR**) i kategorijalnu varijablu koja označuje pojedino poduzeće (**FIRM**). Numeričke varijable među kojima biramo neovisne varijable za model su sljedeće:

- **EMPLP** - broj zaposlenika u poduzeću
- **IM** - vrijednost uvoza
- **IMUM** - udio uvoza u industriji
- **FDIUM** - udio direktnih stranih investicija u industrijskoj grani

[4] Preuzeto s: http://people.stern.nyu.edu/wgreene/Econometrics/probit-panel.txt

Oznaka varijable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
EMPLP	1	80	190	580	480	39750
IM	1161	6020	18900	17400	25460	45590
IMUM	0.05681	0.17570	0.20900	0.25280	0.29160	0.66940
FDIUM	0.002401	0.024770	0.043940	0.045810	0.049620	0.345200
PROD	0.04973	0.07667	0.08587	0.08962	0.09057	1.00000
LOGSALES	6.992	9.997	10.320	10.540	11.430	11.990

Tablica 6.1: Deskriptivna statistika neovisnih varijabli

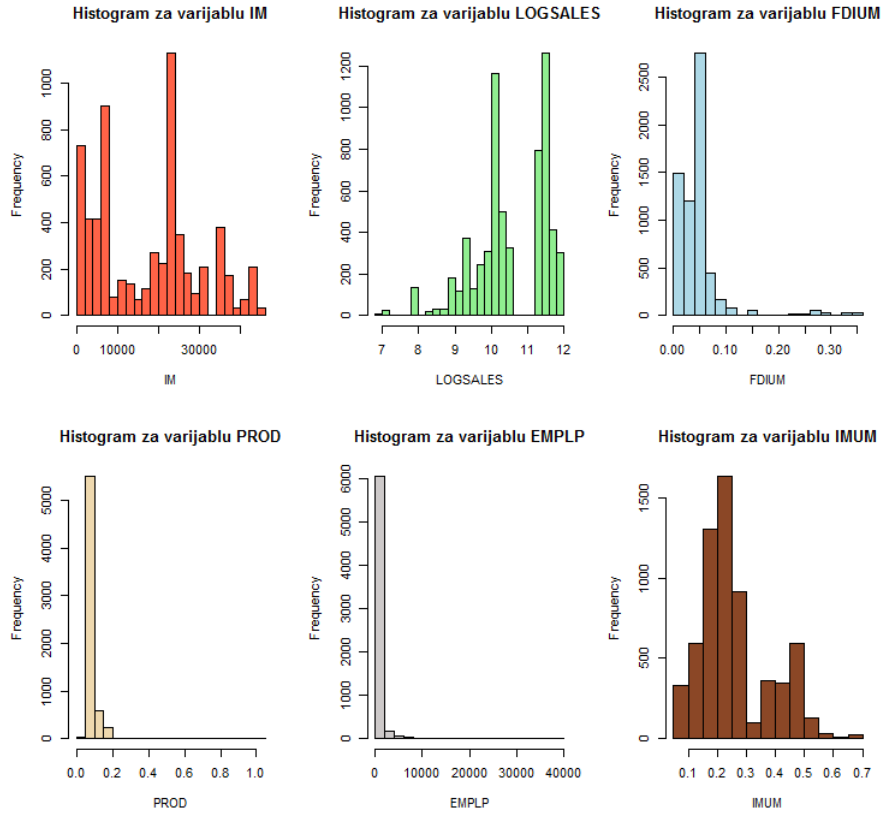
- **PROD** - mjera produktivnosti
- **LOGSALES** - logaritam prihoda od prodaje

Sve analize podataka kao i proces odabira modela rađeni su u statističkom softwareu **R**. Na numeričke varijable primijenila sam *t.test*[5] kako bismo testirali razlikuje li se očekivana vrijednost pojedine varijable za koju je **IP** == 1 od one koju povezujemo s **IP** == 0. Kod svih varijabli mogli smo odbaciti nul hipotezu o jednakom očekivanju na razini značajnosti 0.05, što znači da sve one mogu dobro razlikovati pojavu inovacije u poslovanju nekoga poduzeća. Za navedene varijable možemo promotriti deskriptivne statistike u tablici 6.1 i njihove histograme i kutijaste dijagrame na slikama 6.1 i 6.2.
Histogrami prikazuju frekvenciju pojavljivanja određenih vrijednosti za navedene varijable. Iz oblika histograma možemo pretpostaviti da varijable nisu normalno distribuirane, što smo i opovrgnuli Shapiro-Wilk testom normalnosti. Iz kutijastih dijagrama možemo zaključiti da imamo dosta stršećih vrijednosti, posebice kod varijabli **EMPLP** i **FDIUM**. Stršeće vrijednosti ostavili smo u podatcima, što znači da i dalje imamo uravnoteženi panel.

6.1 Odabir modela

Prijašnja istraživanja ([5]), u pogledu razvijanja inovacija kod poduzeća, kažu da su najpoznatije determinante inovacije veličina poduzeća i struktura tržišta. Osim njih, bitne su i diverzifikacija proizvoda, stupanj internacionalizacije, produktivnost poduzeća, dostupnost financijskih resursa te tehnološke sposobnosti. Pretpostavljaju pozitivan utjecaj uvoza na inovacije. Poduzeća koja su dio konglomerata imaju lakši pristup vanjskom kapitalu, pa za takvu pravnu zavisnost pretpostavljamo pozitivnu vezu s inovativnosti. U nekim granama industrije inovacije se jače potiču nego u drugim granama. Za poduzeća

[5]statistički test u kojem testna statistika ima Studentovu ili t distribuciju

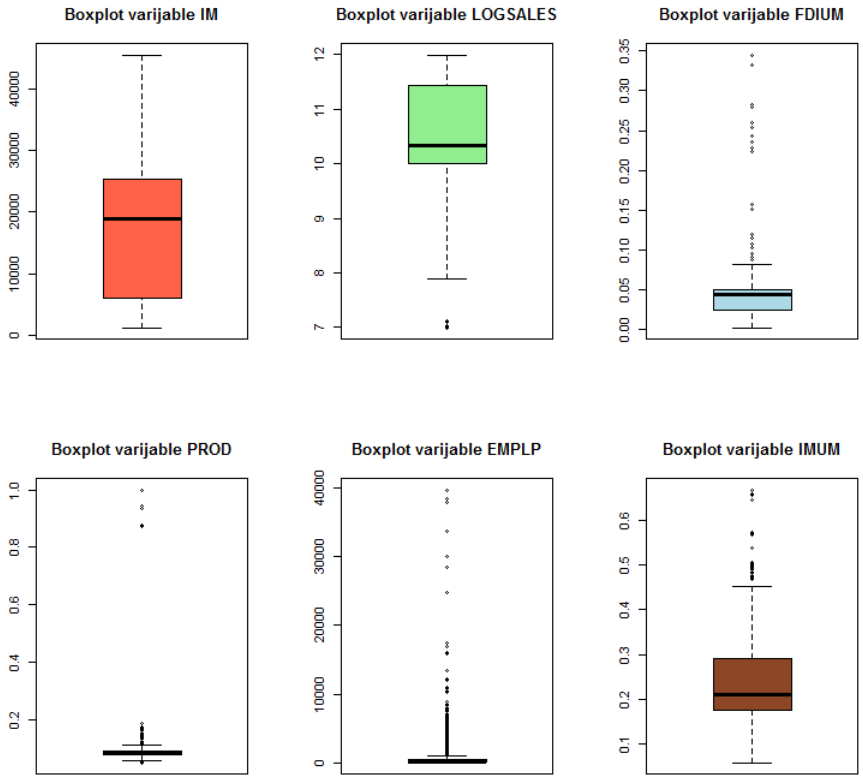


Slika 6.1: Histogrami neovisnih varijabli

koja su financijski ograničena manje je vjerojatno da će generirati inovaciju. Značajna aktivnost poduzeća na stranim tržištima u pozitivnoj je vezi s inovacijama. Također [6] navode da se dobrobiti od ulaganja u istraživanje i razvoj povećavaju s produktivnosti poduzeća i cijenama dionica. U članku [7] navodi se pozitivna veza s inovativnosti za slijedeće varijable: veličina tržišta, ljudski resursi, financijski resursi usmjereni na istraživanje i razvoj, suradnja sa stručnjacima izvan poduzeća.

Od navedenih determinanti u našem panelu pojavljuju se varijable koje opisuju stupanj internacionalizacije kroz uvoz (**IM**) i udio uvoza u grani djelatnosti (**IMUM**), mjere produktivnosti (**PROD**), financijske resurse (**LOGSALES**, **FDIUM**) i veličinu poduzeća (**EMPLP**). Navedenih šest varijabli uključivali smo u modele, zajedno s 0-1 varijablama **FIRM** i **YEAR** i isprobavali kombinacije varijabli kako bismo dobili što bolje modele. Rađen je logit model fiksnih efekata, s fiksnim efektom po svim poduzećima i logit model slučajnih efekata. U paketu *pglm*[6] statističkog softwarea **R** nije implementiran Hausmanov test, kojim bismo mogli olakšati donošenje odluke o izboru modela slučajnog ili fiksnog efekta. Prema [1], koji je radio istraživanje na ovim podatcima, bolji izbor je model slučaj-

[6]panel generalized linear model



Slika 6.2: Kutijasti dijagrami neovisnih varijabli

nih efekata. Također, primjenom Hausmanovog testa na linearne modele ne odbacujemo nultu hipotezu o nekoreliranosti regresora s individualnim efektom, što ide u korist modelu slučajnih efekata.

Kao kriterij odabira modela koristila sam Akaike informacijski kriterij (AIC) koji je dan formulom 5.4. Ovaj kriterij daje nam relativnu mjeru gubitka informacija kada dani model koristimo za opisivanje stvarnosti. Biramo one modele koji imaju što manju vrijednost AIC-a.

AIC: 7134.857	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	2.2582e-01	2.1104e-01	1.0700	0.284610
IM	3.6912e-05	5.5544e-06	6.6455	3.021e-11 ***
PROD	-6.4226e+00	2.0798e+00	-3.0881	0.002014 **
FDIUM	7.8293e+00	1.4809e+00	5.2869	1.244e-07 ***
sigma	2.9070e+00	1.1541e-01	25.1881	< 2.2e-16 ***

Tablica 6.2: Procijenjeni koeficijenti modela slučajnih efekata (1)

AIC: 7136.456	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	9.5018e-01	1.1602e+00	0.8190	0.412787
IM	4.1453e-05	9.0731e-06	4.5688	4.905e-06 ***
PROD	-6.3494e+00	2.0817e+00	-3.0502	0.002287 **
FDIUM	7.4703e+00	1.5835e+00	4.7175	2.388e-06 ***
LOGSALES	-7.5274e-02	1.1862e-01	-0.6346	0.525691
sigma	2.9075e+00	1.1552e-01	25.1677	< 2.2e-16 ***

Tablica 6.3: Procijenjeni koeficijenti modela slučajnih efekata (2)

AIC: 7143.773	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-4.945525	0.879710	-5.6218	1.890e-08 ***
IMUM	2.726137	0.608381	4.4810	7.431e-06 ***
FDIUM	8.596287	1.541882	5.5752	2.473e-08 ***
LOGSALES	0.428329	0.075986	5.6369	1.731e-08 ***
sigma	2.928800	0.116131	25.2199	< 2.2e-16 ***

Tablica 6.4: Procijenjeni koeficijenti modela slučajnih efekata (3)

Među procijenjenim modelima slučajnih efekata, odabrala sam tri koji su prikazani u tablicama 6.2, 6.3 i 6.4. Prva dva modela imaju najmanju vrijednost AIC kriterija i zato su nam ovdje zanimljivi. Dok treći model ima nešto lošiji AIC ali zato ima sve koeficijente značajno različite od nule. Primjećujemo da se kod prvog modela svi procijenjeni koeficijenti razlikuju od nule na razini značajnosti 0.05, osim konstantnog člana. U modelu 6.3 konstantni član i varijabla **LOGSALES** nisu statistički značajno različiti od nule. Kada uzmemo u obzir naučeno iz prethodnih istraživanja, prva dva modela se čine lošijima od zadnjeg. Varijabla **PROD** ima negativnu vezu s inovativnosti, što nema smisla prema [6]. U svojim modelima [1] je također dobio negativan koeficijent za varijablu koja opisuje produktivnost, ali nije detaljnije komentirao taj rezultat. Koeficijenti modela 6.4 nam govore da se šanse da poduzeće ostvari inovaciju povećavaju s povećanjem udjela uvoza u grani djelatnosti, s povećanjem direktnih stranih investicija i s povećanjem prihoda od prodaje. Ta činjenica se podudara sa zaključkom istraživanja [1], gdje je potvrđeno da uvoz i udio stranih investicija imaju pozitivnu vezu s inovativnošću.

Sažetak

Na samom početku rada upoznali smo se sa strukturom podataka nazvanom panel podatci. Naveli smo neke probleme koji se mogu pojaviti u modeliranju kao što su: heterogena i selektivna pristranost, koje smo objasnili pomoću modela dobivenih na simuliranim podatcima u R-u.
Zatim, opisali smo linearne modele za panel podatke i njihove procjenitelje dobivene metodom najmanjih kvadrata. U idućem poglavlju fokusiramo se na složeniji oblik modela. Promatramo diskretne modele, konkretnije, probit i logit modele za koje opisujemo metodu procjene parametara pomoću Newton-Raphsonove metode.
U nastavku detaljnije opisujemo diskretne modele, model fiksnih i model slučajnih efekata za panel podatke. Za procjenu modela fiksnih efekata, opisana je metoda zvana Neyman-Scottov princip.
Navedene metode koristimo u izradi modela za uravnoteženi panel od 1270 poduzeća koje promatramo kroz pet godina.

Title and summary

Generalized linear model for panel data

In this paper we introduced definition of panel data. Then, we discussed some problems that we can encounter during modeling procedure. In next chapter, we describe the simplest models for panel data, linear models. We describe discrete models for panel data focusing on fixed effects model and random effects model and methods of estimating parameters in these models. To conclude this paper we use statistical software for modeling panel data that consists of 1270 individuals across five time periods.

Literatura

[1] W. Greene, *Convenient Estimators for the Panel Probit Model: Further Results*, Department of Economics, Stern School of Business, New York University, 2002.

[2] C. Hsiao, *Analysis of Panel Data*, Cambridge University Press, 2003.

[3] M. J. Mazerolle, *Making sense out of Akaike's Information Criterion (AIC): its use and interpretation in model selection and inference from ecological data*, Centre de recherche en biologie forestie're, Pavillon Abitibi-Price, Faculté de Foresterie et de Géomatique, Université Laval, Québec, Québec G1K 7P4, Canada

[4] H. M. Park, *Practical Guides To Panel Data Modeling: A Step-by-step Analysis Using Stata*, Tutorial Working Paper, Graduate School of International Relations, International University of Japan, 2011.

[5] B. Peters, *Persistence of Innovation: Stylised Facts and Panel Data Evidence*, Centre for European Economic Research (ZEW), Departement of Industrial Economics and international Management, Germany, 2006.

[6] B. Peters, M. J. Roberts,V. A. Vuong, H. Fryges, *Firm R&D, Innovation, and Productivity in German Industry*, Centre for European Economic Research (ZEW), preliminary draft, 2013.

[7] A. G. Vieites, J. L. Calvo, *A Study on the Factors That Influence Innovation Activities of Spanish Big Firms*, Department of Economic Analysis I, Universidad Nacional de Educación a Distancia,Madrid, Spain, 2010.

[8] M. Söderbom, *ERSA Training Workshop, Lecture 5: Estimation of Binary Choice Models with Panel Data*, 2009.

[9] `http://en.wikipedia.org/wiki/Akaike_information_criterion`, [20.5.2014.]

[10] `http://en.wikipedia.org/wiki/Sufficient_statistic`, [24.5.2014.]

[11] `http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataSets.htm`, [24.5.2014.]

7 Životopis

Rođena sam 23.7.1989. godine u Osijeku. Osnovnoškolsko obrazovanje započela sam u OŠ "Retfala" u Osijeku, a završila u OŠ "Bilje". Osim što sam bila odlična učenica predstavljala sam školu na natjecanjima iz nekoliko predmeta: matematike, fizike, kemije. Zatim upisujem Prirodoslovno-matematičku gimnaziju u Osijeku, koju sam završila s odličnim uspjehom i bila oslobođena mature.
Na Odjelu za matematiku, u Osijeku 2008. godine, upisujem preddiplomski studij matematike, koji završavam s odličnim uspjehom. Zatim upisujem diplomski studij, smjera Financijska i poslovna matematika, na istom odjelu na kojem sam položila sve ispite s prosječnom ocjenom odličan.
Trenutno sam apsolventica spomenutog diplomskoga studija. Paralelno s nastavnim obvezama, do druge godine preddiplomskoga studija aktivno sam trenirala odbojku u klubu ŽOK "Bilje". 2011. godine položila sam ispit za županijskoga odbojkaškog suca. I danas se bavim odbojkom, ali u ulozi suca.