Primjena statističkih metoda u kreditnom riziku

Varga, Matko

Master's thesis / Diplomski rad

2023

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:217:962502

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Matko Varga

PRIMJENA STATISTIČKIH METODA U KREDITNOM RIZIKU

Diplomski rad

Voditelj rada:
Izv. prof. dr. sc. Nikola Sandrić

Zagreb, ožujak 2023.

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. _____

Zahvaljujem mentoru izv. prof. dr. sc. Nikoli Sandriću na dostupnosti, savjetima i vremenu uloženom u pisanje ovog diplomskog rada.
Najveće hvala obitelji, prijateljima i djevojci na podršci, ljubavi i vjeri u mene. Bez vaše pomoći i riječi ohrabrenja, ostvarenje ovog cilja ne bi bilo moguće.

Sadržaj

Uvod

Osnovni princip zarade financijskih institucija poput banaka jest posuđivanje novca klijentima i profitiranje na temelju naplaćenih kamatnih stopa. Međutim, može se dogoditi da pojedini klijent, bio to pojedinac ili društvo, počne kasniti sa vraćanjem posuđenog iznosa. U konačnici, klijent možda ne ispoštuje uvjete i ne isplati dogovoreni iznos natrag. Ovo stanje poznato je kao stanje ulaska u *default*. U tom je slučaju posuđivač na gubitku jer mu uložena investicija nije vraćena. Dakle, u trenutku sklapanja ugovora, posuđivač je izložen riziku jer postoji mogućnost da uložena sredstva više nikada "neće vidjeti". Kreditni rizik odnosi se na sve vrste takvog rizika. Očiti primjer kreditnog rizika jest poslovanje banaka i njihovo izdavanje kredita klijentima banke. Upravljanje kreditnim rizikom od ogromne je važnosti svim financijskim institucijama jer se svi žele ograditi od loših klijenata i potencijalnih gubitaka. Statističke metode kojima se mogu ocijeniti klijenti, odnosno njihova rizičnost, i svrstati u dobre ili loše bit će glavna tema ovog rada. Važnost danih metoda jest automatizacija procesa, odnosno zamjena za ljudsku procjenu.

U prvom poglavlju navodimo osnovne matematičke definicije i alate potrebne za analizu, kao i glavne pojmove iz svijeta kreditnog rizika. U drugom poglavlju navodimo elemente *generaliziranog linearnog modela*, u čiju klasu upada i model logističke regresije. Logistička regresija najpopularnija je metoda za modeliranje kreditnog rizika iz nekoliko razloga. Prvo, zavisna varijabla može se označiti brojem 0 ili 1, što predstavlja klijenta s uspješnom otplatom kredita, odnosno klijenta ušlog u *default*, respektivno. Nadalje, poznato je da banke prikupljaju puno informacija o svojim klijentima, i te informacije mogu poslužiti kao dobar indikator dobrih ili loših klijenata. Logistička regresija daje vjerojatnost da osoba s određenim informacijama jednog dana uđe u *default*. Banka time može dobiti procjenu rizičnosti klijenta i na temelju nje donijeti odluku odobriti kredit klijentu ili ne. Nakon definicije modela logističke regresije, navodimo glavne pretpostavke i način na koji određujemo utječe li pojedina varijabla značajno na vjerojatnost ulaska u *default* ili ne. Donosimo intrepretaciju modela i niz koraka kojih se može pridržavati pri odabiru finalnog modela. Krajnje, donosimo osnovne veličine za usporedbu više modela i određivanje njihovih prediktivnosti. Najbitniji dio rada čini treće poglavlje, gdje radimo primjer izrade modela logističke regresije s ciljem izračuna pouzdanosti klijenta, na način

kako bi to zaista moglo i izgledati u nekoj konkretnoj banci. Objašnjavamo kako odabrati varijable (odnosno, informacije o klijentu) koje najviše utječu na klijentov ulazak u *default* i kako interpretirati dobivene rezultate. Podaci korišteni za izradu primjera preuzeti su sa stranice Kaggle. Primjer je rađen u programskom jeziku R (RStudio).

Poglavlje 1

Kreditni rizik

1.1 Kreditni rizik

Pretpostavimo da dvije stranke, zajmodavac i zajmoprimac, žele sklopiti Ugovor o zajmu. Zajmodavac posuđuje zajmoprimcu određenu svotu novca, a zajmoprimac se Ugovorom obvezuje isplatiti cjelokupan iznos posuđenog novca natrag, u dogovorenom roku, plus određenu kamatu. Što ako zajmoprimac, iz bilo kojeg razloga, ne vrati dogovorenu svotu ili ju ne vrati na vrijeme? Primjerice, zajmoprimac može bankrotirati i uopće ne biti u stanju ispuniti svoju obvezu. Gubitak zajmodavca može biti ukupan (iznos posuđene svote) ili djelomičan (zajmoprimac je uspio vratiti dio posuđenog novca), ali u svakom slučaju zajmodavac gubi uloženi novac. Kako bi pokušao smanjiti rizik kojem je izložen, prije sklapanja Ugovora zajmodavac može napraviti neku vrstu sigurnosne provjere nad posuđivačem (npr. pogledati njegove povijesne podatke o novčanim transakcijama) kako bi procijenio njegovu pouzdanost. Matematički način na koji to može učiniti bit će srž ovog rada.

U literaturi [3], kreditni rizik se definira kao mogućnost gubitka novca zbog nesposobnosti, nevoljkosti ili nepravovremenosti druge stranke u ispunjenju financijskih obaveza. Kada god postoji šansa da druga stranka ne plati iznos duga ili prekrši novčanu obvezu, postoji kreditni rizik. Pod nesposobnost ubrajamo insolventnost (zaduženikove obveze su veće od njegove imovine), neplaćanje (eng. *default*) i bankrot. U većini slučajeva, gubitak nastao iz kreditnog rizika je upravo zbog obveznikove nesposobnosti plaćanja. Kreditni gubitci mogu nastati i iz obveznikove nevoljkosti, iako puno rjeđe. Ova situacija se može dogoditi zbog rasprave oko valjanosti ugovora i gubitci su obično puno manji. Kroz rad sa transakcijama koje uključuju kreditni rizik, ključna su pitanja:

1. Kolika je izloženost kreditnom riziku i koliki je ukupan trošak u slučaju obveznikova neplaćanja?

2. Kolika je vjerojatnost da obveznik ne ispuni svoje obveze? Zbog statističke pozadine ovog pitanja, ono će biti glavni fokus ovog rada. Više o tome u Poglavljima 2 i 3.

3. Koliko, i u kojem vremenskom okviru, može biti nadoknađeno u slučaju bankrota?

Tko je izložen kreditnom riziku?

Htjeli to ili nehtjeli, sve institucije i individualci izloženi su kreditnom riziku. Bitno je naglasiti da izloženost kreditnom riziku nije nužno negativna stvar. Štoviše, kreditne institucije, banke, investicijski fondovi i sl. postoje i profitiraju zbog kreditnog rizika, dokle god se njime upravlja na ispravan način. Kako su od svih institucija kreditnom riziku najviše izložene one financijske, fokusirat ćemo se samo na njih:

- Banke - zbog prirode poslovanja banaka, one imaju najveće kreditne portfelje i posjeduju najsofisticiranije organizacije za upravljanje rizikom. Unatoč tome, njihovi apetiti za kreditnim rizikom su u silaznoj putanji, zbog niskih marži i visokih zahtjeva za regulatornim kapitalom. Regulatorni kapital označuje iznos kapitala koji banka mora posjedovati, a propisan je od strane regulatora (npr. središnje banke). Nedavne aktivnosti regulatora sugeriraju suzdržanost prema izloženosti kreditnom riziku i u budućnosti. Najveći udio kreditnog rizika za banke proizlazi iz zajmova i kredita. Kako bi se dodatno osigurale, banke često koriste pozajmljivanje na temelju imovine. Ako obveznik nije u stanju vratiti posuđeno, banka će si naplatiti gubitak prodajom obveznikove imovine, npr. nekretnine koju posjeduje. Zbog potencijalno velikih gubitaka, banke upošljavaju grupe stručnjaka za rizik čiji je posao analizirati moguće posuđivače i rizik koji oni predstavljaju za banku;

- Agencije za upravljanje imovinom - investitori se često obraćaju ovakvim agencijama kako bi upravljale njihovim novcem i daljnje ga ulagale, dok ne ostvare povrat koji investitor želi. Posljedično, takvi agenti su izloženi velikoj količini kreditnog rizika;

- Investicijski fondovi - njihovi investitori obično imaju veće apetite za rizik, ali stoga zahtijevaju i veće povrate. Oni su, dakle, agresivniji nego obični investitori i voljni ulagati u rizičnije financijske instrumente, što ih čini izloženim većoj količini kreditnog rizika;

- Osiguravajuća društva - njihov posao svodi se na skupljanje premija osiguranika (u zamjenu za isplatu u slučaju dogovorenog događaja) i daljnje investiranje tog novca. Stoga nije neuobičajeno da im isplate osiguranicima nadmašuju uplaćene premije, ali oni i dalje ostvaruju profit na temelju investicija. Zbog visoke izloženosti, mnoge osiguravajuće kuće prakticiraju reosiguranje. Primjerice, osiguravajuća kuća može

izdati mnogo polica osiguranja protiv prirodnih katastrofa, bazirano na modelu koji predviđa malu vjerojatnost potresa. Ako se potres kojim slučajem dogodi, kuća ne bi bila u stanju isplatiti sve iznose bez reosiguranja;

- Mirovinski fondovi.

Iako je izloženost kreditnom riziku kod individualaca puno manja, ona i dalje postoji. Primjerice, obitelj koja želi renovirati kuću i unaprijed plati majstorima. Tko im garantira da će majstori zaista ispuniti obećanja i završiti sav posao? Fizičke osobe se također bave investiranjem, koje nosi rizik kao i kod institucija. Osobe koje stavljaju svoj novac u banku, na štednju, također su izložene. Srećom, u većini država postoji zaštita protiv ovakvih vrsta gubitaka. U RH, Hrvatska agencija za osiguranje depozita štiti depozite (između ostalog i) fizičkih osoba u svim kreditnim institucijama koje su dobile odobrenje za rad od Hrvatske narodne banke.

Zašto je bitno upravljati kreditnim rizikom?

Treba reći da se kreditni rizik može kontrolirati i da on ne proizlazi iz ničega. Stoga razumijevanje izvora kreditnog rizika i njegovo predviđanje daje mogućnost uspješnog upravljanja kreditnim rizikom. On je također produkt ljudskog ponašanja, tj. donošenja ljudskih odluka. Važan aspekt upravljanja kreditnim rizikom čini anticipacija dioničara i njihovih motivacija. Loše upravljanje kreditnim rizikom može dovesti do velikih gubitaka, pa čak i bankrota. Potrebna je dovoljna količina kapitala da se prežive veliki i neočekivani gubitci. Iako uvijek postoji mogućnost takvih gubitaka, vođenjem kreditnog portfelja mogu se regulirati svi manji gubitci. Stoga sve firme trebaju posvetiti veliku pažnju i resurse upravljanju kreditnim rizikom, zbog vlastitog opstanka i ostvarenja profita.

1.2 Modeliranje kreditnog rizika

Kako bi se prethodni zadatak mogao što bolje obavljati, kreditni rizik je potrebno kvantificirati. U tu svrhu, uvedimo [9] osnovne matematičke alate koji će nam pomoći u analizi.

Definicija 1.2.1. *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor i* $\mathcal{B}$ *σ-algebra Borelovih skupova na* $\mathbb{R}$*. Slučajna varijabla je funkcija* $X : \Omega \to \mathbb{R}$ *takva da je* $X^{-1}(B) \in \mathcal{F}$ *za svaki* $B \in \mathcal{B}$*.*

Drugim riječima, slučajna varijabla je funkcija sa Ω u $\mathbb{R}$ izmjeriva u paru σ-algebri $(\mathcal{F}, \mathcal{B})$.

Definicija 1.2.2. *Neka je X slučajna varijabla na* $(\Omega, \mathcal{F}, \mathbb{P})$*. X ima matematičko očekivanje ako vrijedi* $\int_\Omega |X(\omega)|\, d\mathbb{P}(\omega) < \infty$*. Tada definiramo očekivanje od X kao*

$$\mathbb{E}X = \int_\Omega X\, d\mathbb{P}. \tag{1.1}$$

Jasno, u slučaju diskretne slučajne varijable,

$$X \sim \begin{pmatrix} a_1 & a_2 & ... \\ p_1 & p_2 & ... \end{pmatrix},$$

vrlo generalan izraz iz (1.1) svodi se na $\mathbb{E}[X] = \sum_j a_j p_j$, pri čemu vrijedi $\sum_j p_j = 1$. U nastavku će nam od interesa biti *Bernoullijeva* slučajna varijabla. Prisjetimo se, *Bernoullijeva slučajna varijabla* s parametrom $p \in [0, 1]$ je $X : \Omega \to \{0, 1\}$ takva da je $\mathbb{P}(X = 1) = p$ i $\mathbb{P}(X = 0) = 1 - p = q$. Pišemo

$$X \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}.$$

Tako možemo izračunati očekivanje *Bernoullijeve* slučajne varijable kao

$$\mathbb{E}[X] = q \cdot 0 + 1 \cdot p = p \tag{1.2}$$

Definicija 1.2.3. *Neka je X slučajna varijabla na* $(\Omega, \mathcal{F}, \mathbb{P})$ *takva da je* $\mathbb{E}[|X|^r] < \infty, r > 0$*. R-ti centralni moment od X je* $\mathbb{E}[(X - \mathbb{E}X)^r]$*, a* $\mathbb{E}[|X - \mathbb{E}X|^r]$ *je r-ti apsolutni centralni moment od X. Varijanca od X je drugi centralni moment od X, dakle*

$$Var(X) = \mathbb{E}[(X - \mathbb{E}X)^2]. \tag{1.3}$$

Primijetimo da za $r > 0$ i $\mathbb{E}[|X|^r] < \infty$ vrijedi i $\mathbb{E}[|X|^s] < \infty$ za svaki s, $0 < s < r$. Zaista, $|X|^s \leq 1 + |X|^r$, pa iz monotonosti i linearnosti očekivanja slijedi $\mathbb{E}[|X|^s] \leq 1 + \mathbb{E}[|X|^r] < \infty$. Varijanca od X postoji ako i samo ako postoji $\mathbb{E}[X^2]$, tj. $\mathbb{E}[X^2] < \infty$. Tada dakle, postoji i $\mathbb{E}X$ te vrijedi

$$\begin{aligned} Var(X) &= \mathbb{E}[(X - \mathbb{E}X)^2] \\ &= \mathbb{E}[X^2 - 2X \cdot \mathbb{E}X + \mathbb{E}(X)^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}X \cdot \mathbb{E}X + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \end{aligned} \tag{1.4}$$

Koristeći (1.2) i (1.4) direktno slijedi da je varijanca *Bernoullijeve* slučajne varijable X

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p) = pq \tag{1.5}$$

Neka je $X = (X_1, ..., X_n)$ n-dimenzionalan slučajan vektor na $(\Omega, \mathcal{F}, \mathbb{P})$ takav da je $\mathbb{E}X_i < \infty$ i $\mathbb{E}[X_i^2] < \infty$ za $i = 1, ..., n$. Po Schwartz-Cauchyjevoj nejednakosti postoje realni brojevi

$$\mu_{ij} = Cov(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)]. \tag{1.6}$$

Za $i \neq j$ $Cov(X_i, X_j)$ zovemo *kovarijanca* slučajnih varijabli X_i i X_j, a $Cov(X_i, X_i) = Var(X_i)$. Jasno, iz (1.6) slijedi

$$Cov(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}X_i \mathbb{E}X_j. \tag{1.7}$$

Kovarijancu shvaćamo kao mjeru linearnog odnosa između slučajnih varijabli X i Y. Kako bismo utvrdili snagu linearne ovisnosti, promatramo *Pearsonov koeficijent korelacije*

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \in [-1, 1], \tag{1.8}$$

pri čemu je $\sigma_X = \sqrt{Var(X)}$ i $\sigma_Y = \sqrt{Var(Y)}$. Za slučajne varijable X i Y kažemo da su *nekorelirane* ako vrijedi $Cov(X, Y) = 0$. Primijetimo da nezavisnost slučajnih varijabli povlači njihovu *nekoreliranost*. Međutim, obrat generalno ne vrijedi. Ako je $X \sim Unif(-1, 1)$ i $Y = X^2$, X i Y očito nisu nezavisne, ali po (1.7) vrijedi $Cov(X, X^2) = \mathbb{E}[X \cdot X^2] - \mathbb{E}X \cdot \mathbb{E}[X^2] = \mathbb{E}[X^3] - \mathbb{E}X \cdot \mathbb{E}[X^2] = 0 - 0 \cdot \mathbb{E}[X^2] = 0$.

Definicija 1.2.4. *Neka je $\Theta \subseteq \mathbb{R}^k$ parametarski prostor, $\tau : \Theta \to \mathbb{R}^k$ izmjeriva funkcija, $k \geq 1$ i $\tau(\theta) = \theta$, $\theta \in \Theta$ (želimo procijeniti parametar θ). Statistika $T = t(X_1, ..., X_n)$ je nepristran procjenitelj za parametar θ ako vrijedi*

$$\mathbb{E}[T] = \theta. \tag{1.9}$$

O *nepristranom procjenitelju* intuitivno treba razmišljati kao o procjenitelju koji sustavno niti podcijenjuje niti precijenjuje pravu vrijednost željenog parametra.

Definicija 1.2.5. *Neka su $\mathcal{G}$,$\mathcal{H}$ i $\mathcal{K}$ σ-podalgebre od $\mathcal{F}$. $\mathcal{G}$ i $\mathcal{H}$ su uvjetno nezavisne uz dano $\mathcal{K}$ ako vrijedi*

$$\mathbb{P}(G \cap H \mid \mathcal{K}) = \mathbb{P}(G \mid \mathcal{K}) \cdot \mathbb{P}(H \mid \mathcal{K}), \quad \text{za sve } G \in \mathcal{G}, H \in \mathcal{H}.$$

Očekivani gubitak

Povijest sugerira da čak i pouzdani klijenti ponekad ne ispune svoje financijske obaveze. Stoga banke ne traže samo osiguranje protiv kritičnih, nego svih zajmova u kreditnom portfelju. Osnovna ideja je sljedeća. Npr., cijena isplate zdrastvenog osiguranja za nekoliko bolesnih klijenata je pokrivena ukupnom sumom uplata premija svih klijenata. Dakle, cijena

premije koju tridesetogodišnjak plaća za zdravstveno osiguranje će odražavati *očekivani* trošak osiguravajuće kuće za takvu skupinu klijenata. Za bankovne kredite, filozofija je ista. Naplaćivanje i skupljanje prikladnih premija za svaki kredit stvorit će *očekivanu pričuvu* u kojoj će biti dovoljno kapitala za pokriće gubitaka proizašlih iz nevraćenih kredita. Međutim, jedan broj nije dovoljan za odrediti dobru ili lošu transakciju. Stoga je potrebno promatrati veći broj parametara, od kojih su najbitniji:

- Vjerojatnost neplaćanja (eng. *probability of default*, kratica *PD*) - klijentova vjerojatnost ulaska u *default*;
- Gubitak u slučaju *defaulta* (eng. *loss given default*, kratica *LGD*) - postotak izgubljene imovine u trenutku *defaulta*;
- Izloženost u slučaju *defaulta* (eng. *exposure at default*, kratica *EAD*) - ukupan potencijalni gubitak u trenutku *defaulta*.

Pokazuje se da su ovi parametri, skupno gledano, dobar indikator rizika i tvore osnovu na temelju koje se donose odluke kod odobravanja transakcija. Gubitak svakog obveznika je zatim dan [2] *varijablom gubitka*

$$\tilde{L} = LGD \cdot EAD \cdot L \quad uz \quad L = \mathbf{1}_D, \tag{1.10}$$

gdje je $D = \{$obveznik je ušao u *default* u određenom vremenskom periodu$\}$. Uobičajeno je za vremenski period uzimati godinu dana. Dakle, L je *Bernoullijeva* slučajna varijabla sa parametrom uspjeha PD, $L \sim Bern(PD)$. Sada je prirodno definirati očekivani gubitak (eng. *expected loss*, kratica *EL*) svakog obveznika preko očekivanja pripadajuće *varijable gubitka* $\tilde{L}$

$$EL = \mathbb{E}[\tilde{L}] = LGD \cdot EAD \cdot \mathbb{P}(D) = LGD \cdot EAD \cdot PD. \tag{1.11}$$

Iz gornje jednadžbe je jasno da o *LGD* i *EAD* razmišljamo deterministički, tj. da se radi o konstantama. U praksi postoje scenariji i kada se ta dva parametra modeliraju kao slučajne varijable. U tom je slučaju *EL* ponovno dan formulom (1.11) jer su *LGD* i *EAD* *nepristrani procjenitelji* za očekivanja pripadajućih slučajnih varijabli.

Pogledajmo sada kako se u praksi parametrima iz (1.11) može pristupiti pojedinačno. Uvjerljivo je najkompleksnije modelirati *PD* te će stoga cjelokupno Poglavlje 2 biti posvećeno samo ovom parametru. *LGD* obično može biti izračunat kao $1 - stopa\ oporavka$ (eng. *recovery rate*). Ako je dug klijenta 300 000€ u trenutku *defaulta*, a banka proda klijentov stan u vrijednosti 200 000€ , $LGD = 1 - \frac{200\,000}{300\,000} = \frac{1}{3}$. Procjena ovakvih gubitaka naravno nije jednostavna. *LGD* modeli su obično bazirani na metodama linearne regresije, a jedan od mogućih izbora je *metoda cenzuriranih najmanjih kvadrata* (eng. *censored*

least squares). Razlika u odnosu na običnu metodu najmanjih kvadrata je *cenzura* zavisne varijable, tj. njena vrijednost poznata je samo parcijalno, npr. u nekom intervalu. Ako nam je poznato n povijesnih *LGD* podataka $\boldsymbol{LGD^*} = (LGD_1^*, ..., LGD_n^*)$, koeficijenti $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)$ su dani rješenjem minimizacijskog problema

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (LGD_i^* - y_i(\boldsymbol{\beta}, \boldsymbol{X}))^2 \tag{1.12}$$

sa

$$y_i(\boldsymbol{\beta}, \boldsymbol{X}) = \beta_0 + \sum_{j=1}^{m} \beta_j \cdot x_{ij} \tag{1.13}$$

pri čemu je

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & ... & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & ... & x_{nm} \end{bmatrix}$$

matrica dizajna. Naravno, rješenje problema je

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{LGD^*}. \tag{1.14}$$

LGD za i-ti zajam tada je procijenjen kao

$$\hat{LGD}_i(\hat{\boldsymbol{\beta}}, \boldsymbol{X}) = \max(0, \min(1, y_i(\hat{\boldsymbol{\beta}}, \boldsymbol{X}))). \tag{1.15}$$

Još jedan mogući model za *LGD* je *model napuhane beta regresije* (eng. *inflated beta regression*). Funkcija gustoće *LGD*-ja često ima vrhove oko nule i jedinice, što nazivamo bimodalnost. *Napuhana beta distribucija* proširenje je beta distribucije koja dodijeljuje vjerojatnost nulama i jedinicama (koristeći *Bernoullijevu distribuciju*) dok ostale opservacije modelira beta distribucijom. Prisjetimo se, beta distribucija $B(\alpha, \beta)$ ima funkciju gustoće

$$f_X(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad y \in (0, 1) \tag{1.16}$$

pri čemu je $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$ gama funkcija. Ako je $X \sim B(\alpha, \beta)$, tada je $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$ i $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. *LGD* model *napuhane beta regresije* bazira se na funkciji gustoće

$$f_X(x; \mu, \sigma, \zeta, \xi, p_0, p_1) = \begin{cases} p_0, & x = 0 \\ (1 - p_0 - p_1) \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & x \in (0, 1) \\ p_1, & x = 1 \end{cases} \tag{1.17}$$

pri čemu je $\alpha = \frac{\mu(1-\sigma^2)}{\sigma^2}$, $\beta = \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}$, $\alpha, \beta > 0$, $p_0, p_1 \in (0, 1)$ i $0 < p_0 + p_1 < 1$. Dakle, parametri distribucije su

$$\mu = \frac{\alpha}{\alpha+\beta}, \quad \sigma = (\alpha+\beta+1)^{-1/2}, \quad \zeta = \frac{p_0}{1-p_0-p_1}, \quad \xi = \frac{p_1}{1-p_0-p_1},$$

μ,$\sigma \in (0,1)$ i ζ,$\xi > 0$. Nadalje, očekivanje i varijanca su dani kao

$$\mathbb{E}[X] = \frac{\mu+\xi}{1+\zeta+\xi} \tag{1.18}$$

$$Var(X) = \frac{\sigma^2\mu(1-\mu)+\mu^2+\xi+(\mu+\xi)^2(1+\zeta+\xi)^{-1}}{1+\zeta+\xi} \tag{1.19}$$

O prva dva parametra, μ i σ, razmišljamo kao o parametrima lokacije i raspršenosti, dok ζ i ξ određuju asimetriju i zaobljenost. Model dakle pretpostavlja da je zavisna varijabla *LGD* distribuirana kao napuhana beta slučajna varijabla pri čemu svih 6 parametara iz (1.17) ima svoje koeficijente $\boldsymbol{\beta}$, u oznaci $\boldsymbol{\beta}^{(\mu,\sigma,\zeta,\xi,p_0,p_1)}$ i koji su aditivna funkcija prediktora (vrijednosti iz matrice dizajna). Ideja je zatim predvidjeti cijelu distribuciju *LGD*-jeva, iz koje ćemo procijeniti *LGD* za *i*-ti zajam kao očekivanje te distribucije:

$$\mathbb{E}[LGD_i] = \int_0^1 x \cdot f_X(x;\cdot)dx = p_1^i + \frac{1-p_0^i-p_1^i}{1+exp(-(\beta_0^{(\mu)} + \sum_{j=1}^m \beta_j^{(\mu)} \cdot x_{ij}))}. \tag{1.20}$$

Bitno je naglasiti da odabir modela i njegova prediktivna snaga najviše ovisi o izboru (pa i dostupnosti) prediktivnih varijabli, što su obično informacije koje banka ima o zajmoprimcu. O ovim, i drugim metodama modeliranja *LGD*-ja, više u [12].

EAD se može podijeliti u dva dijela, nepodmireni dio (eng. *outstandings*, kratica $OUTST$) i buduće obveze (eng. *commitments*, kratica $COMM$). Primjerice, ako klijent digne kredit od 100 000€ i za dvije godine, u trenutku kada duguje još 75 000€ uđe u *default*, nepodmireni dio je 75 000€. Buduća obveza izdaje se klijentu u obliku pisma, a obećaje kredit koji stupa na snagu u nekom trenutku u budućnosti, po klijentovoj želji. Dakle, kredit još nije izdan, ali je odobren. Buduće obveze dijelimo na podignuti dio i nepodignuti dio. Označimo sa $OUTST_t$ nepodmireni dio u sadašnjem trenutku, $OUTST_d$ nepodmireni dio u trenutku *defaulta*, $COMM_p$ podignuti dio u sadašnjem trenutku i $COMM_{nep} = COMM_p - OUTST_t$ nepodignuti dio u sadašnjem trenutku. Definiramo očekivani udio budućih obaveza podignutih prije *defaulta* kao

$$\gamma = \begin{cases} \frac{max(OUTST_d - OUTST_t, 0)}{COMM_{nep}}, & COMM_{nep} > 0 \\ 0, & COMM_{nep} \leq 0 \end{cases} \tag{1.21}$$

Tada definiramo *EAD* kao

$$EAD = OUTST_t + \gamma \cdot COMM_{nep} \tag{1.22}$$

Modeliranje *EAD*-ja je stogo usko povezano s modeliranjem γ. Pokazalo se da u praksi γ također slijedi bimodalnu distribuciju sa vrhovima oko nule i jedinice, pa se stoga svi modeli *LGD*-ja mogu primijeniti i za modeliranje γ. Ako se pak odluči na direktno modeliranje *EAD*-ja, potrebno je napraviti jednu promjenu u odnosu na model *napuhane beta regresije*. Naime, *EAD* nije ograničen odozgo sa 1, pa nam je potreban samo jedan parametar, p, koji dodijeljuje vjerojatnost nulama. Također je moguće procijeniti distribuciju *EAD*-ja nekom drugom poznatom distribucijom. Pokazalo se da je najbolji fit za ne-negativne vrijednosti *EAD*-ja gama distribucija, pa je tako uveden *ZAGA* (eng. *zero-adjusted gamma model*) model sa funkcijom gustoće

$$f_X(x;\mu,\sigma,p) = \begin{cases} p, & x = 0 \\ (1-p)\underbrace{\dfrac{1}{(\sigma^2\mu)^{1/\sigma^2}}\dfrac{x^{1/\sigma^2-1}e^{-x/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)}}_{=Gamma(x,\mu,\sigma)}, & x > 0 \end{cases} \tag{1.23}$$

sa očekivanjem $\mu > 0$, raspršenjem $\sigma > 0$ i $p = \mathbb{P}(EAD = 0) \in (0,1)$. Vrijedi da su očekivanje i varijanca ovakve distribucije

$$\mathbb{E}[X] = (1-p)\mu \tag{1.24}$$

$$Var(X) = (1-p)\mu^2(p+\sigma^2) \tag{1.25}$$

Parametri iz (1.23) ponovno će biti modelirani kao funkcije prediktora. Procjena za i-ti zajam tada je dana maksimiziranjem *funkcije vjerodostojnosti*

$$\begin{aligned} L &= \prod_{i=1}^{n} f_X(x_i;\cdot) \\ &= \prod_{x_i=0} p_i \prod_{x_i>0}(1-p_i)\cdot Gamma(\mu_i,\sigma_i) \end{aligned} \tag{1.26}$$

Treba reći da se za jedan od prediktora često uzima vrijeme do *defaulta*. Ova varijabla je naravno u praksi uvijek nepoznata i stoga nema utjecaj na prediktivnu snagu modela. Kako bi se dopustio model sa vremenom do *defaulta* kao prediktornom varijablom, kao dopuna *EAD* modelima predložen je model *analize doživljenja*. Ako sa T označimo slučajnu varijablu koja predstavlja vrijeme *defaulta*, možemo odrediti očekivani *EAD* kao

$$EAD = \sum_{t=1}^{12}\left(\frac{S(t-1)-S(t)}{1-S(12)}EAD(t)\right), \tag{1.27}$$

pri čemu je $S(t) = \mathbb{P}(T > t)$ funkcija doživljenja u trenutku t ($S(t-1) - S(t)$ dakle daje vjerojatnost *defaulta* u t-tom mjesecu) i $EAD(t)$ je procjena *EAD* modela za *EAD* uvjetno na t, vrijeme do *defaulta*. Više o ovim modelima u [10].

Neočekivani gubitak

Prethodno smo uveli pojam *očekivanog gubitka* transakcije kao osiguranja protiv gubitaka koje banka očekuje iz povijesnih iskustava. Kao dodatak očekivanoj rezervi, banka bi trebala štedjeti novac koji bi pokrio i *neočekivani gubitak* koji nadmašuje prosječne očekivane gubitke iz prošlosti. Kao mjeru varijacije gubitaka iz EL-a, prirodno je promatrati standardnu devijaciju *varijable gubitka* $\tilde{L}$ iz (1.10). Kao što je ranije spomenuto, u kontekstu formule (1.11) LGD i EAD bile su konstante. Od sada nadalje, uvodimo slučajnu varijablu *ozbiljnost gubitka* (eng. *severity of loss*, kratica SEV) čije je pripadno očekivanje LGD. EAD i dalje smatramo konstantom. Stoga definiramo *neočekivani gubitak* (eng. *unexpected loss*, kratica UL) kao

$$UL = \sqrt{Var(\tilde{L})} = \sqrt{Var(SEV \cdot EAD \cdot L)} \tag{1.28}$$

Koristeći (1.4) i (1.5) imamo sljedeći rezultat:

Propozicija 1.2.6. *Ako su SEV i D nekorelirani, vrijedi*

$$UL = EAD \cdot \sqrt{Var(SEV) \cdot PD + LGD^2 \cdot PD(1 - PD)} \tag{1.29}$$

Napomenimo da pretpostavka *nekoreliranosti* u prethodnoj propoziciji nije najrealnija, pa (1.29) često služi samo kao aproksimacija.

Do sada smo promatrali analizu kreditnog rizika na razini jednog zajma. S obzirom da su banke suočene s modeliranjem rizika kompletnog portfelja koji se sastoji od n zajmova, $n \in \mathbb{N}$, proširit ćemo dosadašnje definicije. *Varijabla gubitka* tako postaje

$$\tilde{L}_i = SEV_i \cdot EAD_i \cdot L_i \quad uz \quad L_i = \mathbf{1}_{D_i},\ \mathbb{P}(D_i) = PD_i,\ \ i = 1, ..., n. \tag{1.30}$$

Gubitak portfelja tada definiramo kao

$$\tilde{L}_{PF} = \sum_{i=1}^{n} \tilde{L}_i = \sum_{i=1}^{n} SEV_i \cdot EAD_i \cdot L_i \tag{1.31}$$

Sada možemo dobiti *očekivani gubitak portfelja* i *neočekivani gubitak portfelja* uzimanjem očekivanja, odnosno standardne devijacije slučajne varijable $\tilde{L}_{PF}$. Iz linearnosti očekivanja slijedi

$$EL_{PF} = \mathbb{E}[\tilde{L}_{PF}] = \sum_{i=1}^{n} \mathbb{E}[\tilde{L}_i] = \sum_{i=1}^{n} LGD_i \cdot EAD_i \cdot PD_i \tag{1.32}$$

Za varijancu linearnost vrijedi ako su $\tilde{L}_i$ po parovima *nekorelirane*. U praksi to nikada nije slučaj pa nam za definiciju neočekivanog gubitka portfelja treba sljedeći teorem:

Teorem 1.2.7. *Neka su $X_1, ..., X_n$ slučajne varijable na $(\Omega, \mathcal{F}, \mathbb{P})$ takve da postoji $\mathbb{E}X_i^2, i = 1, ..., n$. Tada vrijedi*

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} Cov(X_i, X_j) = \sum_{i,j=1}^{n} Cov(X_i, X_j)$$

Za dokaz vidjeti [9]. Sada možemo definirati *neočekivani gubitak portfelja* kao

$$UL_{PF} = \sqrt{Var(\tilde{L}_{PF})} = \sqrt{\sum_{i,j=1}^{n} EAD_i \cdot EAD_j \cdot Cov(SEV_i \cdot L_i, SEV_j \cdot L_j)} \tag{1.33}$$

Direktno iz definicije izlazi rezultat:

Propozicija 1.2.8. *Ako pretpostavimo poseban slučaj gdje su SEV_i konstante, vrijedi*

$$UL_{PF}^2 = \sum_{i,j=1}^{n} LGD_i \cdot LGD_j \cdot EAD_i \cdot EAD_j \cdot \sqrt{PD_i(1 - PD_i)PD_j(1 - PD_j)}\rho_{i,j}$$

pri čemu je $\rho_{i,j} = Corr(L_i, L_j) = Corr(\mathbf{1}_{D_i}, \mathbf{1}_{D_j})$ koeficijent korelacije defaulta zajmova i i j.

Kako bismo pobliže shvatili značenje gornje formule, pogledajmo posve jednostavan primjer gdje se portfelj sastoji od 2 zajma takva da je $LGD_1 = LGD_2 = EAD_1 = EAD_2 = 1, \rho = Corr(L_1, L_2)$ i $p_i = PD_i$, $i = 1, 2$. U tom slučaju je

$$UL_{PF}^2 = p_1(1 - p_1) + p_2(1 - p_2) + 2\rho\sqrt{p_1(1 - p_1)}\sqrt{p_2(1 - p_2)} \tag{1.34}$$

Mogući slučajevi su:

- $\rho = 0$. U ovom slučaju UL_{PF} postiže minimum s obzirom da nestane treći član u (1.34). Naime, investiranje u različite vrste imovine smanjuje svukupan rizik portfelja jer je manje vjerojatno da svi zajmovovi uđu u *default* odjednom. Što manje dva zajma imaju zajedničko, to će manji biti utjecaj *defaulta* jednog klijenta na ekonomsku budućnost drugog klijenta. Stoga je slučaj *potpune nekoreliranosti* zajmova najbolji za UL_{PF};

- $\rho > 0$. U ovom slučaju su dva zajma povezana na način da *default* jednog klijenta povećava vjerojatnost da i drugi klijent uđe u *default*:

$$\begin{aligned} \mathbb{P}(L_2 = 1 \mid L_1 = 1) &= \frac{\mathbb{P}(L_2 = 1, L_1 = 1)}{\mathbb{P}(L_1 = 1)} = \frac{\mathbb{E}[L_1 L_2]}{p_1} \\ &= \frac{p_1 p_2 + Cov(L_1, L_2)}{p_1} = p_2 + \frac{Cov(L_1, L_2)}{p_1} \end{aligned} \tag{1.35}$$

S obzirom da je $Cov(L_1, L_2) > 0$ slijedi da je $\mathbb{P}(L_2 = 1 \mid L_1 = 1) > p_2 = \mathbb{P}(L_2 = 1)$. Dakle, pozitivna korelacija znači veću vjerojatnost *defaulta* drugog klijenta pod uvjetom *defaulta* prvog klijenta nego običnu vjerojatnost *defaulta* samo drugog klijenta. U specijalnom slučaju $\rho = 1$ (tada je i $p_1 = p_2 = p$) imamo $UL_{PF} = 2\sqrt{p(1-p)}$, što znači da je rizik portfelja sveden na rizik jednog zajma, ali će iznositi dvostruko;

- $\rho < 0$. Ova će situacija biti jednaka prethodnoj osim u specijalnom slučaju $\rho = -1$. Tada je po (1.34) naravno $UL_{PF} = 0$ što možemo intrepretirati kao sljedeće: investiranje u zajam 2 koji je u *potpunoj negativnoj koreliranosti* sa već postojećim zajmom 1 u potpunosti neutralizira rizik zajma 1.

Ako banka pretpostavlja da će gubitci nadmašiti EL_{PF} za više od jedne standardne devijacije *gubitka portfelja* $\tilde{L}_{PF}$, okrenut će se pojmu *ekonomskog kapitala* (eng. *economic capital*, kratica *EC*), koji je u literaturi još poznat kao i *Value-at-Risk* (kratica *VaR*). Za odabranu razinu značajnosti $\alpha \in (0, 1)$ definiramo

$$EC_\alpha = q_\alpha - EL_{PF} \tag{1.36}$$

pri čemu je q_α α-ti kvantil od $\tilde{L}_{PF}$

$$q_\alpha = \inf\{q > 0 : \mathbb{P}(\tilde{L}_{PF} \leq q) \geq \alpha\}. \tag{1.37}$$

Distribucija gubitka

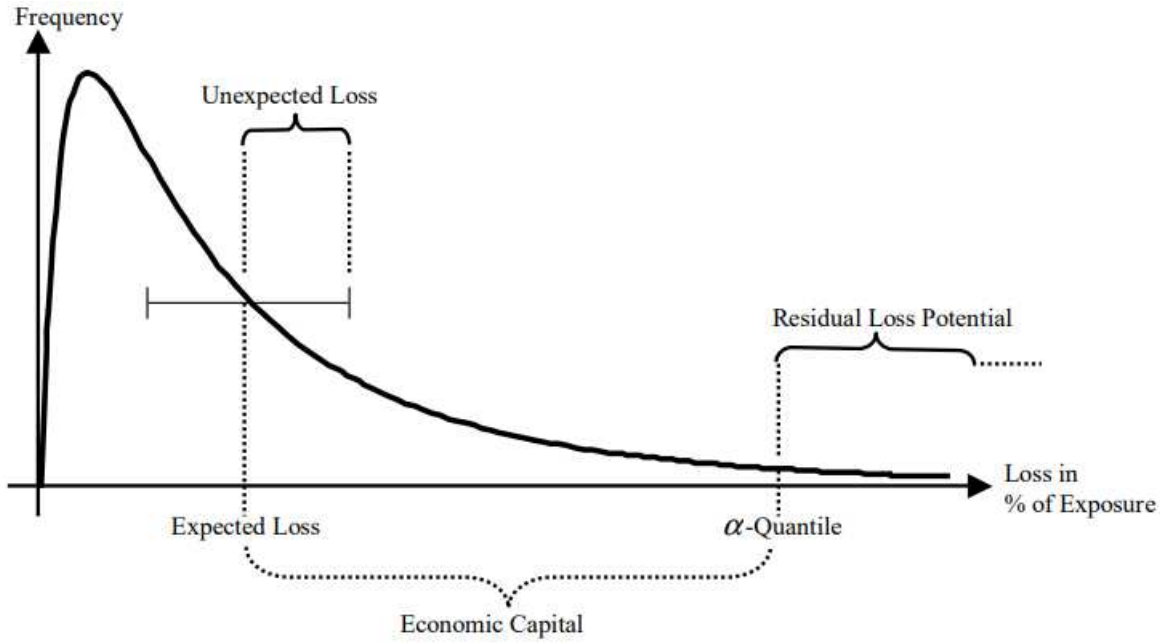
Vidjeli smo da sve prethodno definirane veličine bitne za rizik portfelja ovise o $\tilde{L}_{PF}$. Stoga ključnu ulogu u upravljanju kreditnim rizikom igra distribucija $\tilde{L}_{PF}$-a, tzv. *distribucija gubitka* (eng. *loss distribution*) portfelja. Slika 1.1 [2] prikazuje da svi kvantifikatori rizika portfelja koje smo do sada uveli mogu biti prikazani preko *distribucije gubitka*. Stoga je bitno primijetiti da u slučaju empirijskog generiranja *distribucije gubitka* izračunate statistike služe samo kao aproksimacija tih kvantifikatora. Empirijsko generiranje *distribucije gubitka* poznato je kao *Monte Carlo simulacija*.

U *Monte Carlo simulaciji*, funkciju gustoće gubitka aproksimiramo empirijskom funkcijom gustoće, odnosno histogramom. Uzimajući u obzir distribucije pojedinačnih varijabli gubitka $\tilde{L}_i$ i njihove korelacije, simuliramo *m gubitaka portfelja*, $m \in \mathbb{N}$, $\tilde{L}_{PF}^{(1)}, ..., \tilde{L}_{PF}^{(m)}$. Empirijska funkcija gustoće gubitka portfelja tada je dana

$$\hat{F}(x) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}_{[0,x]}(\tilde{L}_{PF}^{(j)}). \tag{1.38}$$

Kako bismo dobili procjenu α-tog kvantila *distribucije gubitka*, sortiramo simulirane podatke

$$\tilde{L}_{PF}^{(j_1)} \leq \tilde{L}_{PF}^{(j_2)} \leq ... \leq \tilde{L}_{PF}^{(j_m)}.$$



Slika 1.1: *Distribucija gubitka* portfelja

Tada će α-ti kvantil q_α empirijske funkcije gubitka biti

$$\hat{q}_\alpha = \begin{cases} \alpha\tilde{L}_{PF}^{(j_{[m\alpha]})} + (1-\alpha)\tilde{L}_{PF}^{(j_{[m\alpha]+1})}, & m\alpha \in \mathbb{N} \\ \tilde{L}_{PF}^{(j_{[m\alpha]})}, & m\alpha \notin \mathbb{N} \end{cases} \tag{1.39}$$

pri čemu je

$$[m\alpha] = \min\{k \in \{1, ..., m\} : m\alpha \leq k\}. \tag{1.40}$$

Procjena *ekonomskog kapitala* zatim je

$$\hat{EC}_\alpha = \hat{q}_\alpha - \frac{1}{m}\sum_{j=1}^{m} \tilde{L}_{PF}^{(j)}. \tag{1.41}$$

Još jedan način generiranja *disribucije gubitka* je *analitičkom aproksimacijom*. Neformalno, *analitička aproksimacija* preslika postojeći portfelj s nepoznatom *distribucijom gubitka* u ekvivalentan portfelj s poznatom *distribucijom gubitka* koja tada služi kao aproksimacija stvarnoj *distribuciji gubitka*. Ideja je sljedeća: odabrati familiju distribucija koje karakterizira njihov prvi i drugi moment. Zatim procijeniti prvi (EL) i drugi (UL) moment postojećeg portfelja te izabrati distribuciju koja najbolje odgovara tim procjenama.

Tu distribuciju zatim ćemo proglasiti *distribucijom gubtitka*. Primjerice, banka procijeni svoj portfelj na $EL = 30$ *b.b.* (*b.b.* je kratica za bazni bod, stoti dio od 1%), $UL = 22.5$ *b.b* i želi aproksimirati *distribuciju gubitka* sa $X \sim B(\alpha, \beta)$:

$$0.003 = \mathbb{E}[X] = \frac{\alpha}{\alpha + \beta}$$

$$0.00225^2 = Var(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Rješavanjem sustava, slijedi da će *distribucija gubitka* biti $B(1.76944, 588.045)$. Pokazuje se da u praksi postoji prirodan izbor za željenu familiju distribucija. Zadatak je pokazati da kada broj zajmova na portfelju $n \to \infty$, *distribucija gubitka* konvergira prema nekoj graničnoj distribuciji u zatvorenoj formi. Uvodimo slučajne varijable $Y \sim N(0, 1)$ koju shvaćamo kao trenutno *stanje gospodarstva* i $r_i \sim N(0, 1)$, što su log-povrati imovine klijenta i. Pretpostavit ćemo da je korelacija između r_i , u oznaci ϱ, uniformna na cijelom portfelju. Tada je $L_i \sim Bern(p_i(Y))$ pri čemu je

$$p_i(Y) = \Phi\left(\frac{\Phi^{-1}(p_i) - \sqrt{\varrho}Y}{\sqrt{1-\varrho}}\right), \qquad i = 1, ..., n \tag{1.42}$$

i pri čemu su $LGD_i \cdot L_i$ *uvjetno nezavisne* uz danu realizaciju $Y = y$, a Φ je funkcija distribucije standardne normalne razdiobe. Definiramo *relativan gubitak portfelja* sa n zajmova kao

$$\tilde{L}^{(n)}_{rel} = \sum_{i=1}^{n} w_i \cdot LGD_i \cdot L_i, \qquad w_i = \frac{EAD_i}{\sum_{j=1}^{n} EAD_j} \tag{1.43}$$

uz tehničku pretpostavku

$$\sum_{j=1}^{n} EAD_j \nearrow \infty \text{ kada } n \to \infty \ i$$

$$\sum_{n=1}^{\infty} \left(\frac{EAD_n}{\sum_{j=1}^{n} EAD_j}\right)^2 < \infty.$$

Primijetimo da je ovo sasvim realna pretpostavka: Ako je $EAD_i \in [a, b]$ za neki $0 < a \leq b$ i svaki $i = 1, ..., n$, tada je

$$\sum_{j=1}^{n} EAD_j \geq na \nearrow \infty \text{ kada } n \to \infty,$$

$$\sum_{n=1}^{\infty} \left(\frac{EAD_n}{\sum_{j=1}^{n} EAD_j}\right)^2 \leq \sum_{n=1}^{\infty} \frac{b^2}{n^2 a^2} = \frac{b^2}{a^2} \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.$$

Teorem 1.2.9. *Uz gornje pretpostavke vrijedi* $\tilde{L}_{rel}^{(n)} \xrightarrow{g.s.} \mathbb{E}[\tilde{L}_{rel}^{(n)} \mid Y]$ *kada* $n \to \infty$*, odnosno*

$$\mathbb{P}\left[\lim_{n\to\infty}\left(\tilde{L}_{rel}^{(n)} - \mathbb{E}[\tilde{L}_{rel}^{(n)} \mid Y]\right) = 0\right] = 1.$$

Dokaz. Neka je $y \in \mathbb{R}$. Definiramo vjerojatnosnu mjeru

$$\mathbb{P}_y(\cdot) = \mathbb{P}(\,\cdot \mid Y = y)$$

i niz slučajnih varijabli

$$X_k = EAD_k(LGD_k \cdot L_k - \mathbb{E}[LGD_k \cdot L_k \mid Y]), \quad k = 1, 2, ...$$

Niz $(X_k)_{k\geq 1}$ je nezavisan obzirom na $\mathbb{P}_y$. Definiramo $\tau_n = \sum_{j=1}^{n} EAD_j$. Po pretpostavci je niz $(\tau_n)_{n\geq 1}$ pozitivan, strogo rastući i zbog uniformne ograničenosti od $(LGD_k \cdot L_k - \mathbb{E}[LGD_k \cdot L_k \mid Y])$ vrijedi

$$\sum_{k=1}^{\infty} \frac{1}{\tau_k^2}\mathbb{E}[X_k^2] \leq \sum_{k=1}^{\infty} \frac{4EAD_k^2}{\tau_k^2} < \infty$$

Po *Kolmogorovljevom uvjetu za jaki zakon velikih brojeva* (vidi [9]) slijedi

$$\lim_{n\to\infty} \frac{1}{\tau_n} \sum_{k=1}^{n} X_k = 0 \quad \mathbb{P}_y - g.s.$$

Sada možemo napisati da za svaki $y \in \mathbb{R}$ vrijedi

$$\mathbb{P}\left[\lim_{n\to\infty}\left(\tilde{L}_{rel}^{(n)} - \mathbb{E}[\tilde{L}_{rel}^{(n)} \mid Y]\right) = 0 \;\middle|\; Y = y\right] = 1.$$

No sada vidimo da ova vjerojatnost vrijedi i bezuvjetno jer je

$$\mathbb{P}\left[\lim_{n\to\infty}\left(\tilde{L}_{rel}^{(n)} - \mathbb{E}[\tilde{L}_{rel}^{(n)} \mid Y]\right) = 0\right] = \int \mathbb{P}\left[\lim_{n\to\infty}\left(\tilde{L}_{rel}^{(n)} - \mathbb{E}[\tilde{L}_{rel}^{(n)} \mid Y]\right) = 0 \;\middle|\; Y = y\right] d\mathbb{P}_Y(y) = 1$$

□

Korolar 1.2.10. *Ako su* $LGD_i \cdot L_i$ *uvjetno nezavisne (uz dano* $Y = y$*) i jednako distribuirane, postoji izmjeriva funkcija* $p : \mathbb{R} \to \mathbb{R}$ *takva da za* $n \to \infty$ *vrijedi* $\tilde{L}_{rel}^{(n)} \xrightarrow{g.s.} p \circ Y$*. Štoviše,* $p \circ Y = \mathbb{E}[LGD_1 \cdot L_1 \mid Y]$ *g.s.*

Dokaz. $\mathbb{E}[\tilde{L}_{rel}^{(n)} \mid Y]$ je po definiciji $\sigma(Y)$-izmjerivo, pa po *Doob-Dynkinovoj lemi* postoji izmjeriva funkcija $p : \mathbb{R} \to \mathbb{R}$ takva da je $\mathbb{E}[\tilde{L}_{rel}^{(n)} \mid Y] = p \circ Y$. Tvrdnja sada slijedi iz prethodnog teorema i jednake distribuiranosti. □

Iz ovih rezultata možemo zaključiti da relativan gubitak portfelja $\tilde{L}_{rel}^{(n)}$ u limesu ovisi samo o slučajnosti slučajne varijable Y. Povećavajući broj zajmova na portfelju banka može ukloniti sav rizik koji proizlazi iz pojedinačnih zajmova. Tako će jedini preostali rizik biti onaj koji proizlazi iz volatilnosti gospodarstva. Kako bismo iskoristili gornji korolar, pretpostavimo $p_i = p$ i $LGD_i = 1$ za sve $i = 1, ..., n$ i izračunajmo $\mathbb{E}[\tilde{L}_{rel}^{(n)} \mid Y]$:

$$\mathbb{E}[\tilde{L}_{rel}^{(n)} \mid Y] = \sum_{i=1}^{n} w_i \cdot \mathbb{E}[L_i \mid Y] = \Phi\left(\frac{\Phi^{-1}(p) - \sqrt{\varrho}Y}{\sqrt{1-\varrho}}\right) = p(Y) \tag{1.44}$$

Sada po Teoremu 1.2.9 slijedi

$$\tilde{L}_{rel}^{(n)} \xrightarrow{g.s.} \Phi\left(\frac{\Phi^{-1}(p) - \sqrt{\varrho}Y}{\sqrt{1-\varrho}}\right), \quad \text{kada } n \to \infty. \tag{1.45}$$

Kako bismo dobili funkciju distribucije $F_{p,\varrho}$, računamo za $x \in [0,1]$

$$\begin{aligned} F_{p,\varrho}(x) &= \mathbb{P}[p(Y) \leq x] \\ &= \mathbb{P}\left[-Y \leq \frac{1}{\sqrt{\varrho}}\left(\Phi^{-1}(x)\sqrt{1-\varrho} - \Phi^{-1}(p)\right)\right] \\ &= \Phi\left[\frac{1}{\sqrt{\varrho}}\left(\Phi^{-1}(x)\sqrt{1-\varrho} - \Phi^{-1}(p)\right)\right] \end{aligned} \tag{1.46}$$

Pripadnu funkciju gustoće zatim dobivamo deriviranjem (1.46):

$$\begin{aligned} f_{p,\varrho}(x) &= \frac{\partial F_{p,\varrho}(x)}{\partial x} \\ &= \sqrt{\frac{1-\varrho}{\varrho}} \cdot exp\left(-\frac{1}{2\varrho}\left((1-2\varrho)(\Phi^{-1}(x))^2 - 2\sqrt{1-\varrho}\cdot\Phi^{-1}(x)\cdot\Phi^{-1}(p) + (\Phi^{-1}(p))^2\right)\right) \\ &= \sqrt{\frac{1-\varrho}{\varrho}} \cdot exp\left(\frac{1}{2}(\Phi^{-1}(x))^2 - \frac{1}{2\varrho}\left(\Phi^{-1}(p) - \sqrt{1-\varrho}\cdot\Phi^{-1}(x)\right)^2\right) \end{aligned} \tag{1.47}$$

Primijetimo da je funkcija $p(\cdot)$ iz (1.44) strogo padajuća pa za $L \sim F_{p,\varrho}$ vrijedi

$$\mathbb{P}[L \leq p(-z_\alpha)] = \mathbb{P}[p(Y) \leq p(-z_\alpha)] = \mathbb{P}[Y \geq -z_\alpha] = \mathbb{P}[-Y \leq z_\alpha] \tag{1.48}$$

pri čemu je z_α α-ti kvantil od $N(0,1)$. Iz ovoga slijedi da je α-ti kvantil q_α od $F_{p,\varrho}$ potreban za izračun *ekonomskog kapitala* iz (1.41) jednak

$$q_\alpha = \Phi\left(\frac{\Phi^{-1}(p) + \sqrt{\varrho}\cdot z_\alpha}{\sqrt{1-\varrho}}\right). \tag{1.49}$$

Propozicija 1.2.11. *Prvi i drugi momenti slučajne varijable* $L \sim F_{p,\varrho}$ *dani su*

$$\mathbb{E}[L] = p \qquad Var(L) = \Phi_2(\Phi^{-1}(p), \Phi^{-1}(p); \varrho) - p^2$$

pri čemu je $\Phi_2(\cdot, \cdot; \varrho)$ *funkcija distribucije bivarijantne normalne razdiobe s koeficijentom korelacije* ϱ*.*

Za dokaz vidjeti [2].

1.3 Kreditni rejting

Vrlo česta praksa svake banke jest kreditni rejting, odnosno kreditni skoring svakog klijenta na portfelju. Pojam rejtinga odnosi se na svrstavanje klijenta u određeni razred s obzirom na njegovu kreditnu sposobnost, a koju banka može procijeniti na temelju dostupnih informacija o tom klijentu. Postupak dodjeljivanja pojedinačnog broja (umjesto razreda) svakom klijentu koji će označiti njegovu kreditnu sposobnost potom je poznat kao i kreditni skoring. U ovom potpoglavlju fokus će biti na kreditnom rejtingu kako proces skoringa čini njegov podskup. Deklaracijom rejtinga bave se ovlaštene rejting agencije kao što su Moody's ili Standard and Poor's [7], a na čije se očitanje oslanjaju gotovo svi sudionici tržišta. Najpoznatije svjetske agencije za dodjeljivanje kreditnog rejtinga koriste sustav razreda $\{AAA, AA, A, BBB, ..., C, D\}$ pri čemu su u razred AAA svrstani klijenti s najvećom kreditnom sposobnošću, a u razred D klijenti ušli u *default*. Generalno dakle, što je klijent svstan u veći razred, to je veća njegova kreditna sposobnost, odnosno manja vjerojatnost ulaska u *default*. U praksi postoje mnogi načini za modeliranje ovakvih sustava, a samo jedan od njih jest pomoću *Markovljevih lanaca*.

Definicija 1.3.1. *Neka je* S *prebrojiv skup (kojeg nazivamo prostor stanja) i* $X = (X_n : n \geq 0)$ *slučajan proces definiran na* $(\Omega, \mathcal{F}, \mathbb{P})$ *s vrijednostima u* S*.* X *je Markovljev lanac ako vrijedi*

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i) \tag{1.50}$$

za sve $n \geq 0$ *i sve* $i_0, ..., i_{n-1}, i, j \in S$ *za koje su obje strane u* (1.50) *dobro definirane.*

Svojstvo (1.50) naziva se *Markovljevim svojstvom* i kaže da su neposredna budućnost i prošlost *uvjetno nezavisne* uz danu sadašnjost. Dakle, ponašanje Markovljevog lanca u idućem vremenskom koraku ovisi samo o njegovom trenutnom stanju i ono je dovoljno za predviđanje stanja koje će lanac sljedeće posjetiti.

Definicija 1.3.2. *Matrica* $P = (p_{ij} : i, j \in S)$ *je stohastička matrica ako vrijedi*

(a) $p_{ij} \geq 0$ *za sve* $i, j \in S$*;*

(b) $\sum_{j\in S} p_{ij} = 1$ *za sve* $i \in S$.

Elemente p_{ij} nazivamo *prijelaznim vjerojatnostima* iz stanja i u stanje j. Od interesa će nam posebno biti i još nešto jače svojstvo nego *Markovljevo*, a to je da *prijelazne vjerojatnosti* ne ovise niti o trenutku $n \geq 1$.

Definicija 1.3.3. *Neka je* $\lambda = (\lambda_i : i \in S)$ *vjerojatnosna distribucija na* S *i* $P = (p_{ij} : i, j \in S)$ *stohastička matrica. Slučajan proces* $X = (X_n : n \geq 0)$ *definiran na* $(\Omega, \mathcal{F}, \mathbb{P})$ *s vrijednostima u skupu stanja* S *je vremenski homogen Markovljev lanac s početnom distribucijom* λ *i matricom prijelaza* P *ako vrijedi*

(a) $\mathbb{P}(X_0 = i) = \lambda_i$ *za sve* $i \in S$*;*

(b) $\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = p_{ij}$ *za sve* $n \geq 0$ *i sve* $i_0, ..., i_{n-1}, i, j \in S$.

U nastavku pretpostavljamo da je $\lambda_i > 0$, odnosno $\mathbb{P}(X_0 = i) > 0$ i pri tome koristimo oznaku $\mathbb{P}_i(\cdot) = \mathbb{P}(\cdot \mid X_0 = i)$. Za $B \subset S$ definiramo *prvo vrijeme pogađanja* skupa B kao

$$T_B = \min \{n \geq 0 : X_n \in B\} \tag{1.51}$$

uz konvenciju $\min \emptyset = \infty$. Za stanje $j \in S$ kažemo da je *dostižno* iz stanja $i \in S$ ako vrijedi $\mathbb{P}_i(T_j < \infty) > 0$ i pišemo $i \longrightarrow j$. Stanja $i, j \in S$ *komuniciraju* ako vrijedi $i \longrightarrow j$ i $j \longrightarrow i$, u oznaci $i \longleftrightarrow j$. Markovljev lanac X je *ireducibilan* ako vrijedi $i \longleftrightarrow j$ za sve $i, j \in S$. Za $C \subset S$ kažemo da je *zatvoren* ako za svako stanje $i \in C$ vrijedi $\mathbb{P}_i(T_{S\setminus C} = \infty) = 1$. Dakle, ako se nalazimo u *zatvorenom* skupu, s vjerojatnošću 1 nikada nećemo stići u njegov komplement. S druge strane, lanac može ući u *zatvoren* skup. Stanje $j \in S$ zovemo *apsorbirajućim* ako je $\{j\}$ *zatvoren* skup. Nakon što lanac jednom uđe u takvo stanje, ostat će tamo zauvijek.

Primjer 1.3.4. *Primjer je rađen po uzoru na [11]. Radi jednostavnosti, neka je skup stanja* $S = \{A, B, C, D\}$ *i matrica prijelaza*

$$P = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} \frac{8}{10} & \frac{1}{10} & \frac{1}{10} & 0 \\ \frac{1}{5} & \frac{3}{5} & \frac{1}{10} & \frac{1}{10} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{2} & \frac{3}{10} \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Za vremensku jedinicu ponovno uzimamo godinu dana. Primijetimo da je stanje D apsorbirajuće. Neka od pitanja koja nas mogu zanimati jesu:

(a) Kolika je vjerojatnost da klijent koji trenutno ima rejting A ikada uđe u default?

(b) Ako klijent trenutno ima rejting C, kolika je vjerojatnost da mu rejting postane A?

(c) Koliko je očekivano vrijeme klijenta sa trenutnim rejtingom C do ulaska u default?

Prije nego što odgovorimo na ova pitanja, uvedimo oznake $h_i^B = \mathbb{P}_i(T_B < \infty)$ i $g_i^B = \mathbb{E}_i(T_B)$ za $i \in S,\ B \subset S$ i teorem čiji se dokaz može pronaći u [11].

Teorem 1.3.5. *Vektor vjerojatnosti pogađanja $h^B = (h_i^B : i \in S)$ je minimalno nenegativno rješenje sustava*

$$\begin{cases} h_i^B = 1, & i \in B \\ h_i^B = \sum_{j \in S}(p_{ij} \cdot h_j^B), & i \notin B. \end{cases} \tag{1.52}$$

Vektor očekivanog vremena pogađanja $g^B = (g_i^B : i \in S)$ je minimalno nenegativno rješenje sustava

$$\begin{cases} g_i^B = 0, & i \in B \\ g_i^B = 1 + \sum_{j \in S}(p_{ij} \cdot g_j^B), & i \notin B. \end{cases} \tag{1.53}$$

Vratimo se sada na pitanja iz primjera.

(a) Zanima nas dakle $h_A = \mathbb{P}_A(T_D < \infty)$ uz $h_D = 1$. Rješavamo sustav

$$\begin{aligned} h_A &= p_{AA} \cdot h_A + p_{AB} \cdot h_B + p_{AC} \cdot h_C + p_{AD} \cdot h_D = \frac{8}{10}h_A + \frac{1}{10}h_B + \frac{1}{10}h_C \\ h_B &= p_{BA} \cdot h_A + p_{BB} \cdot h_B + p_{BC} \cdot h_C + p_{BD} \cdot h_D = \frac{1}{5}h_A + \frac{3}{5}h_B + \frac{1}{10}h_C + \frac{1}{10} \\ h_C &= p_{CA} \cdot h_A + p_{CB} \cdot h_B + p_{CC} \cdot h_C + p_{CD} \cdot h_D = \frac{1}{10}h_A + \frac{1}{10}h_B + \frac{1}{2}h_C + \frac{3}{10} \end{aligned}$$

Rješenje sustava je $(h_A, h_B, h_C, h_D) = (1, 1, 1, 1)$. Dakle, klijent, krenuvši s rejtingom A, ući će u *default* g.s.

(b) S obzirom da je $\{D\}$ *apsorbirajuće* stanje, zapravo nas zanima $h_C = \mathbb{P}_C(T_A < T_D)$. U ovom kontekstu primijetimo također da vrijedi $h_A = 1$ i $h_D = 0$. Dobivamo sustav

$$\begin{aligned} h_B &= p_{BA} \cdot h_A + p_{BB} \cdot h_B + p_{BC} \cdot h_C + p_{BD} \cdot h_D = \frac{1}{5} + \frac{3}{5}h_B + \frac{1}{10}h_C \\ h_C &= p_{CA} \cdot h_A + p_{CB} \cdot h_B + p_{CC} \cdot h_C + p_{CD} \cdot h_D = \frac{1}{10} + \frac{1}{10}h_B + \frac{1}{2}h_C \end{aligned}$$

Rješenje je $(h_A, h_B, h_C, h_D) = (1, \frac{11}{19}, \frac{6}{19}, 0)$. Dakle, vjerojatnost da klijent sa rejtingom C napreduje do rejtinga A je $\frac{6}{19}$.

(c) Zanima nas $g_C = \mathbb{E}_C(T_D)$ uz $g_D = 0$ pa rješavamo sustav

$$g_A = 1 + p_{AA} \cdot g_g + p_{AB} \cdot g_B + p_{AC} \cdot g_C + p_{AD} \cdot g_D = 1 + \frac{8}{10} g_A + \frac{1}{10} g_B + \frac{1}{10} g_C$$
$$g_B = 1 + p_{BA} \cdot g_A + p_{BB} \cdot g_B + p_{BC} \cdot g_C + p_{BD} \cdot g_D = 1 + \frac{1}{5} g_A + \frac{3}{5} g_B + \frac{1}{10} g_C$$
$$g_C = 1 + p_{CA} \cdot g_A + p_{CB} \cdot g_B + p_{CC} \cdot g_C + p_{CD} \cdot g_D = 1 + \frac{1}{10} g_A + \frac{1}{10} g_B + \frac{1}{2} g_C$$

Rješenje je $(g_A, g_B, g_C, g_D) = (\frac{100}{7}, \frac{80}{7}, \frac{50}{7}, 0)$. Znači, klijentu sa rejtingom C potrebno je u prosjeku oko 7.14 godina (zaokruženo na 7 godina i 52 dana) do ulaska u *default*.

Poglavlje 2

Modeliranje *PD*-ja

U prethodnom poglavlju, vidjeli smo da se parametar *PD*, vjerojatnost ulaska u *default*, nalazi u svim bitnim formulama za pristupanje kreditnom riziku. Naravno, njegova će vrijednost biti drugačija za svakog klijenta na portfelju i ovisit će o informacijama koje banka ima o tom klijentu. Kada novi klijent aplicira za kredit, banka želi procijeniti njegov *PD* (ponovno, na temelju njegovih podataka), i na temelju dobivene procjene odlučiti hoće li mu odobriti kredit ili neće. Banka se u svakom slučaju želi ograditi od potencijalno "loših" klijenata (odnosno onih sa potencijalno niskom kreditnom sposobnošću) pa je ključno pitanje u analizi kako odvojiti "dobre" klijente od onih "loših". Krajnje, procjene *PD*-jeva potrebne su i za formiranje prijelaznih matrica iz Potpoglavlja 1.3. Jedna klasa modela koja može pomoći u procjeni, poznata je kao klasa *generaliziranih linearnih modela*.

2.1 Generalizirani linearni modeli

Općenito, linearni modeli su oblika [4]

$$\mathbb{E}[Y_i] = \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{2.1}$$

pri čemu su Y_i međusobno nezavisne i $Y_i \sim N(\mu_i, \sigma^2)$, $\boldsymbol{x}_i^T$ je i-ti redak matrice dizajna $\boldsymbol{X}$ i $\boldsymbol{\beta}$ je vektor koeficijenata. Međutim, možemo se nalaziti u problemu gdje zavisna varijabla nije normalno distribuirana. Štoviše, zavisna varijabla ne mora uopće biti numerička već kategorijska. Nama će od posebnog interesa biti slučaj kada je zavisna varijabla kategorijska i poprima samo dvije vrijednosti: 0 - klijent nije ušao u *default* i 1 - klijent je ušao u *default*. Također, veza između zavisne varijable i prediktora ne mora nužno biti linearna kao u (2.1). Generalno, promatrat ćemo slučajne varijable $Y_1, ..., Y_n$ sa distribucijom koja pripdada tzv. *eksponencijalnoj familiji*. Također, linearna komponenta $\boldsymbol{x}_i^T \boldsymbol{\beta}$ bit će povezana sa zavisnom varijablom preko neke nelinearne funkcije g, koju nazivamo *funkcijom veze*:

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{2.2}$$

Definicija 2.1.1. *Neka je Θ parametarski prostor i Y slučajna varijabla čija distribucija ovisi o $\theta \in \Theta$. Kažemo da distribucija pripada eksponencijalnoj familija ako se pripadna funkcija gustoće može zapisati u obliku*

$$f(y;\theta) = h(y) \cdot C(\theta) \cdot exp(t(y) \cdot Q(\theta)) \tag{2.3}$$

pri čemu je $h : \mathbb{R}^d \to [0,\infty)$, $C : \Theta \to (0,\infty)$, $t : \mathbb{R}^d \to \mathbb{R}$ Borelova funkcija i nije konstanta i $Q : \Theta \to \mathbb{R}$. $Q(\theta)$ nazivamo prirodnim parametrom distribucije.

Zbog simetričnosti y i θ (2.3) možemo zapisati kao

$$f(y;\theta) = exp(t(y) \cdot Q(\theta) + a(\theta) + b(y)) \tag{2.4}$$

uz $h(y) = exp(b(y))$ i $C(\theta) = exp(a(\theta))$. Primjeri distribucija koje spadaju u *eksponencijalnu familiju* jesu normalna, eksponencijalna, gama, chi-kvadrat, beta, *Bernoullijeva*, *Poissonova*, binomna itd. Kao što je već rečeno, nama će od interesa posebno biti binomna distribucija kojom ćemo označavati broj klijenta ušlih u *default*. Ako je $Y \sim Bin(n,p)$, pri čemu je $n \in \mathbb{N}$ broj klijenata na portfelju i p vjerojatnost ulaska u *default*, tada Y ima funkciju gustoće

$$f(y;p) = \binom{n}{y} p^y (1-p)^{n-y}. \tag{2.5}$$

Kako bi se vidjelo da binomna distribucija pripada *eksponencijalnoj familiji*, (2.5) možemo zapisati kao

$$f(y;p) = exp\left[y \cdot ln(p) - y \cdot ln(1-p) + n \cdot ln(1-p) + ln\binom{n}{y}\right] \tag{2.6}$$

uz

$$Q(p) = ln(p) - ln(1-p) = ln\left(\frac{p}{1-p}\right). \tag{2.7}$$

U modelu binarne logističke regresije, za *funkciju veze* g uzimat ćemo upravo (2.7). Sada možemo definirati sve komponente *generaliziranog linearnog modela*:

- Zavisne varijable $Y_1, ..., Y_n$ međusobno su nezavisne i jednako distribuirane, sa distribucijom članom *eksponencijalne familije*;
- Vektor koeficijenata $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ i prediktorne varijable zapisane u matricu dizajna

$$\boldsymbol{X} = \begin{bmatrix} 1 & \boldsymbol{x}_1^T \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & ... & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & ... & x_{nm} \end{bmatrix};$$

- Monotonu i diferencijabilnu funkciju g (*funkcija veze*) takva da vrijedi

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}. \tag{2.8}$$

Primijetimo da ako za funkciju g uzmemo identitetu, dobivamo najobičniji model linearne regresije. Generalno, *funkcija veze* razlikovat će se obzirom na pretpostavku distribucije na $Y_1, ..., Y_n$. Jasno, nama će vrijediti $Y_i \sim Bern(p_i)$, a u tom kontekstu je najpopularniji izbor *funkcije veze* već spomenuta *logit funkcija*

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) \qquad p_i \in (0,1) \tag{2.9}$$

ili *probit funkcija*

$$probit(p_i) = \Phi^{-1}(p_i) \qquad p_i \in (0,1). \tag{2.10}$$

Propozicija 2.1.2. *Neka $Y_1, ..., Y_n$ zadovoljavaju pretpostavke generaliziranog linearnog modela. Tada vrijedi*

$$E(Y_i) = \mu_i = \frac{-a'(\theta_i)}{Q'(\theta_i)} \;\; i \;\; Var(Y_i) = \frac{Q''(\theta_i)\cdot a'(\theta_i) - a''(\theta_i)\cdot Q'(\theta_i)}{(Q'(\theta_i))^3} \tag{2.11}$$

uz oznake kao u (2.4).

Dokaz vidjeti u [4].

2.2 Procjena parametara

Za procjenu vektora koeficijenata $\boldsymbol{\beta}$, koristi se *metoda maksimalne vjerodostojnosti* (eng. *maximum likelihood estimation*, kratica *MLE*). Ideja je maksimizirati *funkciju vjerodostojnosti* tako da su, pod pretpostavkom statističkog modela, opaženi podaci najvjerojatniji. Jasno, u kontekstu modela linearne regresije, *MLE* procjena svodi se na *metodu najmanjih kvadrata*.

Neka je $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ slučajan uzorak sa zajedničkom funkcijom gustoće $f(\boldsymbol{Y}; \boldsymbol{\theta})$ koja ovisi o parametrima $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)^T$ i neka je $\boldsymbol{y} = (y_1, ..., y_n)^T$ realizacija pripadnog slučajnog uzorka. Definiramo *funkciju vjerodostojnosti* (eng. *likelihood function*) $L : \Theta \to \mathbb{R}$ kao

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}). \tag{2.12}$$

Dakle, *funkcija vjerodostojnosti* algebarski je ekvivalentna zajedničkoj gustoći, ali je ona funkcija od $\boldsymbol{\theta}$ uz fiksirani $\boldsymbol{y}$. *MLE* procjenitelj od θ je vrijednost $\hat{\boldsymbol{\theta}}$ koja maksimizira *funkciju vjerodostojnosti*, dakle

$$L(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) \geq L(\boldsymbol{\theta}; \boldsymbol{y}) \qquad za\ sve\ \boldsymbol{\theta} \in \Theta. \tag{2.13}$$

Obično se promatra i *funkcija log-vjerodostojnosti* $l(\boldsymbol{\theta}; \boldsymbol{y}) = logL(\boldsymbol{\theta}; \boldsymbol{y})$. S obzirom da je $log(\cdot)$ monotono rastuća, vrijedi

$$l(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) \geq l(\boldsymbol{\theta}; \boldsymbol{y}) \qquad za\ sve\ \boldsymbol{\theta} \in \Theta. \tag{2.14}$$

MLE procjenitelj $\hat{\boldsymbol{\theta}}$ zatim se dobiva rješavanjem sustava jednadžbi

$$\frac{\partial l(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \theta_j} = 0 \qquad j = 1, ..., m. \tag{2.15}$$

Potrebno je provjeriti da je Hesseova matrica

$$\frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \theta_j \partial \theta_k}$$

izračunata u $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ negativno definitna kako bi se uvjerilo da je riječ o maksimumu od $l(\boldsymbol{\theta}; \boldsymbol{y})$. Rješavanje sustava (2.15) često se provodi numeričkim metodama. U kontekstu *generaliziranog linearnog modela*, želimo procijeniti parametre $\boldsymbol{\beta}$. Uz oznake kao u (2.4), *log-vjerodostojnost* je

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} l(\theta_i; y_i) = \sum_{i=1}^{n} y_i \cdot Q(\theta_i) + \sum_{i=1}^{n} a(\theta_i) + \sum_{i=1}^{n} b(y_i). \tag{2.16}$$

Označimo $l_i = l(\theta_i; y_i)$. Za *MLE* procjenu od β_j, trebamo izračunati

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} \right). \tag{2.17}$$

Pristupimo članovima s desne strane pojedinačno:

$$\frac{\partial l_i}{\partial \theta_i} = y_i \cdot Q'(\theta_i) + a'(\theta_i) = Q'(\theta_i)(y_i - \mu_i) \tag{2.18}$$

što slijedi iz Propozicije 2.1.2. Nadalje, deriviranjem izraza za očekivanje iz Propozicije 2.1.2. slijedi

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{-a''(\theta_i)}{Q'(\theta_i)} + \frac{a'(\theta_i) \cdot Q''(\theta_i)}{(Q'(\theta_i))^2} = Q'(\theta_i) \cdot Var(Y_i). \tag{2.19}$$

Uz oznaku $\boldsymbol{x}_i^T\boldsymbol{\beta} = \eta_i$ imamo

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}. \tag{2.20}$$

Dakle,

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{Var(Y_i)} x_{ij} \cdot \frac{\partial \mu_i}{\partial \eta_i} \right) := U_j \tag{2.21}$$

i spremamo kovarijacijsku matricu U_j-ova u *matricu informacija* $\mathcal{I}$:

$$\begin{aligned}
\mathcal{I}_{jk} &= \mathbb{E}[U_j U_k] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{Var(Y_i)} x_{ij} \cdot \frac{\partial \mu_i}{\partial \eta_i} \right) \cdot \sum_{l=1}^{n} \left(\frac{Y_l - \mu_l}{Var(Y_l)} x_{lk} \cdot \frac{\partial \mu_l}{\partial \eta_l} \right) \right] \\
&= [Y_i \text{ međusobno nezavisne}] \\
&= \sum_{i=1}^{n} \frac{\mathbb{E}[(Y_i - \mu_i)^2] \cdot x_{ij} x_{ik}}{(Var(Y_i))^2} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\
&= \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{Var(Y_i)} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 .
\end{aligned} \tag{2.22}$$

Procjena $\hat{\boldsymbol{\beta}}$ koeficijenata $\boldsymbol{\beta}$ zatim je dana

$$\hat{\boldsymbol{\beta}}^{(r)} = \hat{\boldsymbol{\beta}}^{(r-1)} + \left(\mathcal{I}^{(r-1)} \right)^{-1} \boldsymbol{U}^{(r-1)} \tag{2.23}$$

pri čemu je $\hat{\boldsymbol{\beta}}^{(r-1)}$ vektor procijenjenih koeficijenata $\hat{\beta_0}, ..., \hat{\beta_m}$ u $(r-1)$-oj iteraciji i $\boldsymbol{U}^{(r-1)}$ vektor s elementima iz (2.21) izračunatima u $\hat{\boldsymbol{\beta}}^{(r-1)}$. Za distribucije iz *eksponencijalne familije*, $\mathcal{I}$ će biti regularna pa je (2.23) dobro definiran. Ako pomnožimo obje strane s lijeva sa $\mathcal{I}^{(r-1)}$, dobivamo

$$\mathcal{I}^{(r-1)} \hat{\boldsymbol{\beta}}^{(r)} = \mathcal{I}^{(r-1)} \hat{\boldsymbol{\beta}}^{(r-1)} + \boldsymbol{U}^{(r-1)} \tag{2.24}$$

pa iz (2.22) slijedi

$$\mathcal{I} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \tag{2.25}$$

pri čemu je $\boldsymbol{W} \in M_{n,n}$ dijagonalna matrica s elementima

$$w_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 . \tag{2.26}$$

Kada se izračuna, desna strana iz (2.24) postaje vektor s elementima

$$\sum_{k=0}^{m}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{Var(Y_i)}\cdot\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\hat{\beta}_k^{(r-1)}+\sum_{i=1}^{n}\frac{(y_i-\mu_i)x_{ij}}{Var(Y_i)}\cdot\frac{\partial\mu_i}{\partial\eta_i} \tag{2.27}$$

izračunatima u $\hat{\boldsymbol{\beta}}^{(r-1)}$. Stoga zapisujemo desnu stranu iz (2.24) kao

$$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z} \tag{2.28}$$

uz $\boldsymbol{z}$ sa elementima

$$z_i=\sum_{k=0}^{m}x_{ik}\hat{\beta}_k^{(r-1)}+(y_i-\mu_i)\cdot\frac{\partial\mu_i}{\partial\eta_i}. \tag{2.29}$$

Sve skupa, (2.24) postaje

$$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(r)}=\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}. \tag{2.30}$$

Statistički programi potom koriste numeričke algoritme bazirane na (2.30) koji započinju početnom aproksimacijom $\hat{\boldsymbol{\beta}}^{(0)}$ i vrte se sve dok razlika dvaju uzastopnih iteracija $\left|\hat{\boldsymbol{\beta}}^{(r-1)}-\hat{\boldsymbol{\beta}}^{(r)}\right|$ nije dovoljno mala. Za nama zanimljivu *logit funkciju veze*, $\eta_i=ln\left(\frac{p_i}{1-p_i}\right)$, vrijedi

$$\frac{\partial\eta_i}{\partial p_i}=\frac{1}{p_i(1-p_i)}\quad i\quad\frac{\partial\mu_i}{\partial\eta_i}=\frac{\partial p_i}{\partial\eta_i}=p_i(1-p_i) \tag{2.31}$$

pa se (2.21) u našem slučaju pojednostavljuje na

$$\sum_{i=1}^{n}(y_i-p_i)x_{ij}. \tag{2.32}$$

2.3 Model binarne logističke regresije

Označimo sa

$$y_i=\begin{cases}1, & i\text{-ti klijent je ušao u } default\\ 0, & i\text{-ti klijent nije ušao u } default.\end{cases} \tag{2.33}$$

Ovdje nam je y_i realizacija slučajne varijable $Y_i\sim Bern(p_i)$ pri čemu je $p_i=PD_i$ vjerojatnost ulaska u *default* i-tog klijenta. Kada novi klijent aplicira za kredit, njegov *PD* je jasno, nepoznat. Banka, na temelju već postojećih podataka o postojećim klijentima, želi razviti model koji će što bolje predvidjeti *PD* novog klijenta, i tako ocijeniti njegovu pouzdanost. Dakle, želimo ocijeniti utjecaj *kovarijata* $\boldsymbol{x}_i^T$ na p_i. Najjednostavniji mogući model bio bi

$$p_i=\boldsymbol{x}_i^T\boldsymbol{\beta}. \tag{2.34}$$

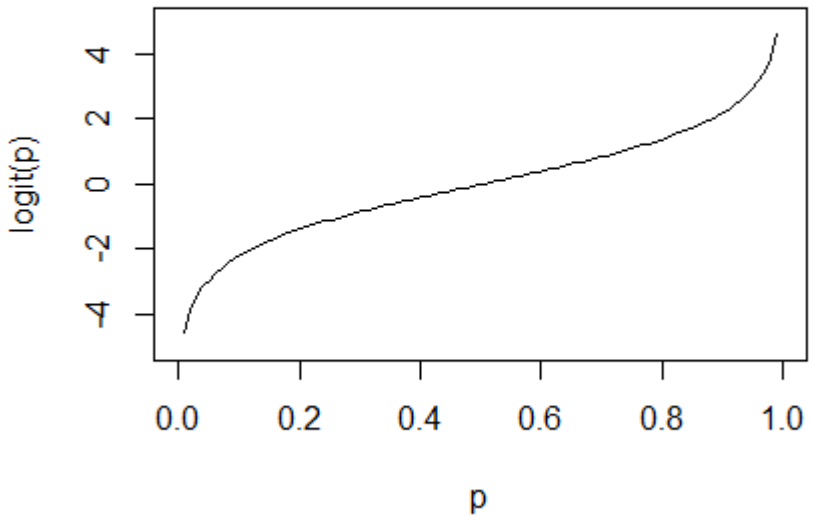
Međutim, problem ovakvog pristupa je što je p_i ograničen odozdo s 0 i odozgo s 1, dok je linearna kombinacija prediktora bilo koji realni broj. Rješenje je transformirati vjerojatnost da uklonimo granice i zatim modelirati transformaciju kao linearnu kombinaciju prediktora. Umjesto vjerojatnosti, definiramo *šansu* (eng. *odds*) kao

$$odds_i = \frac{p_i}{1 - p_i} \qquad p_i \in (0, 1) \tag{2.35}$$

čime smo uklonili gornju granicu. Npr. ako je vjerojatnost događaja $\frac{1}{2}$, $odds = 1$, pa kažemo da postoje jednaki izgledi da se događaj dogodi i da se događaj ne dogodi. Zatim uzimamo prirodni logaritam, čime uklanjamo donju granicu, pa dobivamo funkciju *logit* : $(0, 1) \to \mathbb{R}$

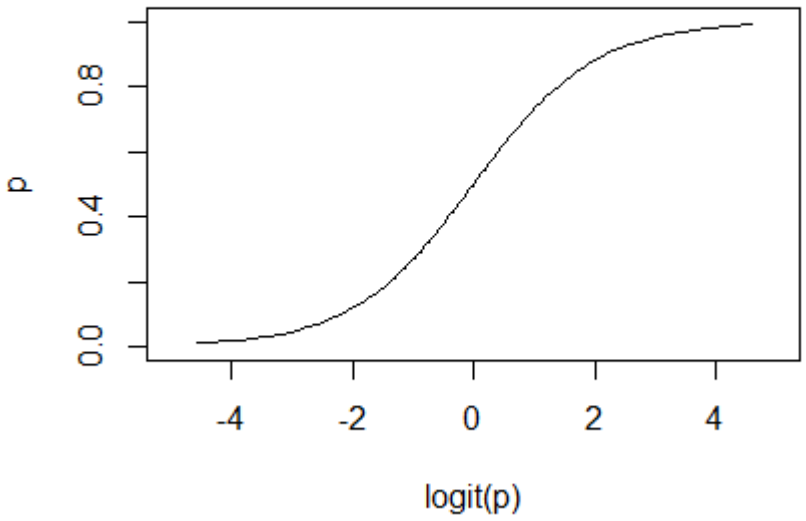
$$logit(p_i) = ln\left(\frac{p_i}{1 - p_i}\right) \qquad p_i \in (0, 1). \tag{2.36}$$

Grafički prikaz *logit* funkcije prikazan je na Slici 2.1. Primijetimo, ako je vjerojatnost



Slika 2.1: *Logit* funkcija

događaja $p = \frac{1}{2}$, $logit(p) = 0$. S obzirom da je $logit(\cdot)$ monotono rastuća, možemo zaključiti da negativne *logit* vrijednosti odgovaraju vjerojatnostima ispod $\frac{1}{2}$, a pozitivne *logit*



Slika 2.2: *Logit* transformacija

vrijednosti vjerojatnostima iznad $\frac{1}{2}$, što možemo vidjeti na Slici 2.2. Sada definiramo model binarne logističke regresije kao

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}. \tag{2.37}$$

Primijetimo da (2.37) zadovoljava sve uvjete *generaliziranog linearnog modela*. Ako želimo dobiti izraz za pojedinačne vjerojatnosti, eksponenciramo obje strane u (2.37) i dobijemo

$$p_i = \frac{exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}. \tag{2.38}$$

Pretpostavke modela

Uz uobičajene pretpostavke na *generalizirani linearni model*, dodatno ćemo još i pretpostavitljati:

- Nema *multikolinearnosti* među nezavisnim varijablama (prediktorima) - pojam *multikolinearnosti* javlja se kada postoji velika *koreliranost* među određenim kovarija-

tama, što znači da one imaju otprilike jednak utjecaj na zavisnu varijablu. Pojava *multikolinearnosti* može imati veliki utjecaj na procijenjene koeficijente, što uvelike otežava interpretaciju dobivenog modela. Postoji nekoliko načina za otkrivanje *multikolinearnosti*. Najjednostavniji je provjeriti matricu korelacija, koja računa *Pearsonov koeficijent korelacije* između svake kombinacije dvaju (numeričkih) prediktora. Ako želimo *korelaciju* između dvaju kategorijskih varijabli X i Y, računamo *Cramerov V*:

$$V(X,Y) = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}} \in [0,1] \tag{2.39}$$

pri čemu je χ^2 chi-kvadrat statistika iz *Pearsonovog testa za nezavisnost*, a k i r broj stupaca, odnosno redaka iz kontigencijske tablice koju formiraju X i Y. Još jedan način detekcije jest provjeriti $det(\boldsymbol{X}^T\boldsymbol{X})$. Naime, $\boldsymbol{X}^T\boldsymbol{X}$ bit će singularna ako sadrži linearno zavisne stupce ili retke. Dakle, ako je $det(\boldsymbol{X}^T\boldsymbol{X}) \sim 0$, možemo zaključiti da postoji *multikolinearnost* među prediktorima, ali ne i kojima. Nadalje, možemo računati *Farrarov* χ^2 [6]:

$$\chi^2 = -\left(n - 1 - \frac{1}{6(2m+5)}\right) \cdot ln(det(\boldsymbol{X}^T\boldsymbol{X})) \tag{2.40}$$

Multikolinearnost će postojati ako je $\chi^2 > \chi^2(\frac{1}{2}m(m-1))$. Međutim, ova mjera također neće otkriti koje varijable su u pitanju. Ako želimo pristupiti varijablama pojedinačno i otkriti koji prediktor uzrokuje *multikolinearnost*, najpopularnija mjera jest VIF (kratica za eng. *variance inflation factor*) faktor, $VIF_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$, $j = 1, ..., m$. Generalno, $VIF > 10$ ukazuje na visoku *koreliranost*. Kada detektiramo koje varijable rade problem, *multikolinearnost* rješavamo izbacivanjem jedne od njih iz modela, transformacijom (uzimanjem logaritma, korijenovanjem i sl.) jedne od njih ili korištenjem treće varijable umjesto njih. Primjerice, ako postoji visoka *koreliranost* između varijabli *tjelesna visina* i *tjelesna težina*, za novi prediktor možemo uzeti *indeks tjelesne mase*;

- Linearnost između $logit(\boldsymbol{p})$ i numeričkih nezavisnih varijabli - ovu pretpostavku najlakše je ispitati Box-Tidwellovim testom: ako imamo model

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}, \tag{2.41}$$

promatramo novi model

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \beta'_1 x'_{i1} + \cdots + \beta'_m x'_{im}, \tag{2.42}$$

pri čemu su $x'_{ij} = x_{ij} \cdot ln(x_{ij}),\ j = 1, ..., m$ i pri čemu su $x_{ij},\ j = 1, ..., m$ numeričke. Pretpostavka će vrijediti ako su koeficijenti $\beta'_1, ..., \beta'_m$ statistički neznačajni (o testiranju značajnosti koeficijenata više u sljedećem paragrafu). Ako bi neki β'_j bio statistički značajan, to bi značilo da varijabla $\boldsymbol{x}_j$ nema linearan odnos sa $logit(\boldsymbol{p})$. Problem se ponovno može rješiti transformacijom varijable $\boldsymbol{x}_j$;

- Nema utjecajnih točaka u modelu - potencijalno utjecajne točke čine *točke visoke poluge* (eng. *high leverage points*) i *odskočnici* (eng. *outliers*). *Točke visoke poluge* odnose se na ekstremne vrijednosti ulaznih podataka spremljenih u $\boldsymbol{X}$, dok su *odskočnici* ekstremne vrijednosti zavisne varijable Y_i. S obzirom da su nama Y_i nule i jedinice, ne moramo se previše brinuti oko *outliera*. S druge strane, *točke visoke poluge* mogu imati veliki utjecaj na procijenjene parametre $\hat{\boldsymbol{\beta}}$, tj. $\hat{\boldsymbol{\beta}}$ se mogu znatno razlikovati s obzirom na to jesu li *točke visoke poluge* uključene u model ili nisu. Krajnje, je li točka utjecajna ili ne, ovisit će o tome je li *odskočnik* i je li *točka visoke poluge*. Definiramo reziduale $\boldsymbol{e} = (e_1, ..., e_n)$ kao

$$e_i = y_i - \hat{p}_i \tag{2.43}$$

i promatramo matricu $\boldsymbol{H}$ definiranu sa

$$\boldsymbol{H} = \boldsymbol{V}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{1/2} \tag{2.44}$$

pri čemu je $\boldsymbol{V}$ dijagonalna sa elementima

$$v_{ii} = \hat{p}_i(1 - \hat{p}_i). \tag{2.45}$$

Dakle, i-ti dijagonalni element od $\boldsymbol{H}$ je

$$h_{ii} = \hat{p}_i(1 - \hat{p}_i)\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}_i. \tag{2.46}$$

Definiramo *Cookovu udaljenost* kao

$$D_i = \frac{\sum_{i=1}^{n}(\hat{p}_i - \hat{p}_i^{(-i)})^2}{(m+1) \cdot (\boldsymbol{e}^T\boldsymbol{e})/(n-m)} \tag{2.47}$$

pri čemu je $\hat{p}_i^{(-i)}$ procijenjena vrijednost za p_i u modelu bez i-te opservacije. Većina literature sugerira da će i-ta opservacija biti proglašena utjecajnom ako vrijedi $D_i > 4/n$. Još jedna mjera za utjecaj na model koju možemo koristiti je $DFFITS$ (kratica za eng. *difference in fit*) [8] definirana sa

$$DFFITS_i = \frac{\hat{p}_i - \hat{p}_i^{(-i)}}{v_{ii}^{(-i)}\sqrt{h_{ii}}}. \tag{2.48}$$

Tada ćemo i-tu opservaciju proglasiti utjecajnom ako vrijedi $DFFITS_i > 2\sqrt{m/n}$. Generalno, utjecajne točke možemo izbaciti iz modela, zamijeniti ih srednjom vrijednosti (aritmetičkom sredinom ili medijanom) ili ih ostaviti u modelu i naznačiti to prilikom predstavljanja rezultata. Iste opcije vrijede i u situaciji gdje podatak x_{ij} iz nekog razloga nedostaje (eng. *missing value*), što se često događa u praksi. Jasnih pravila o tome što učiniti nema i postupak će uvijek ovisiti o situaciji. Uz brojeve, za odluku ponekad treba koristiti i zdrav razum, ali kako god postupili, potrebno je dati opravdanje.

Testiranje značajnosti i interpretacija koeficijenata

Jednom kada napravimo model (2.37), želimo vidjeti koji prediktori imaju značajan utjecaj na PD. To će biti oni prediktori čiji je pripadni β_j statistički značajan. Drugim riječima, testiramo hipotezu

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0. \end{aligned} \tag{2.49}$$

Prva klasa testova koji testiraju ovakve hipoteze jesu *testovi omjera vjerodostojnosti*, koji rade usporedbu procijenjenih vrijednosti $\hat{p}_i$ u modelu gdje je prediktor $\boldsymbol{x}_j$ uključen i gdje $\boldsymbol{x}_j$ nije uključen, a temeljeni su na *funkciji log-vjerodostojnosti*. Testna statistika od interesa je D [5], definirana sa

$$\begin{aligned} D &= -2ln\left(\frac{vjerodostojnost \text{ punog modela}}{vjerodostojnost \text{ saturiranog modela}}\right) \\ &= -2\sum_{i=1}^{n}\left[y_i \cdot ln\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\cdot ln\left(\frac{1-\hat{p}_i}{1-y_i}\right)\right] \end{aligned} \tag{2.50}$$

pri čemu se saturirani model odnosi na onaj model koji ima onoliko parametara koliko i podataka, dakle $m = n$. Stoga je saturirani model savršeni fit za podatke i u takvom će modelu vrijediti $\hat{p}_i = y_i$. Koristeći funkciju gustoće *Bernoullijeve* distribucije dobivamo da je *vjerodostojnost* saturiranog modela

$$L_{sat}(\boldsymbol{p};\boldsymbol{y}) = \prod_{i=1}^{n} y_i^{y_i}\cdot(1-y_i)^{1-y_i} = 1 \tag{2.51}$$

iz čega slijedi

$$D = -2ln(vjerodostojnost \text{ punog modela}). \tag{2.52}$$

Testna statistika D naziva se *devijanca* (eng. *deviance*), i ima istu ulogu kao i suma kvadrata reziduala u običnoj linearnoj regresiji. U kontekstu testiranja (2.49), testna statistika je

$$\begin{aligned} G &= D(\text{model bez } \boldsymbol{x}_j) - D(\text{model sa } \boldsymbol{x}_j) \\ &= -2ln\left(\frac{vjerodostojnost \text{ bez } \boldsymbol{x}_j}{vjerodostojnost \text{ sa } \boldsymbol{x}_j}\right) \sim \chi^2(1). \end{aligned} \tag{2.53}$$

Dakle, p-vrijednost testa (2.49) je $pv = \mathbb{P}(\chi^2(1) > G)$. Ako je $pv < \alpha$ za unaprijed odabrani $\alpha \in (0,1)$, odbacujemo H_0 u korist alternative i parametar β_j proglasujemo statistički značajnim na razini značajnosti α. U tom slučaju možemo zaključiti da prediktor $\boldsymbol{x}_j$ značajno pospješe model i da ima značajan utjecaj na PD. U praksi, obično će početni model sadržavati m prediktora, od kojih želimo selektirati samo one statistički značajne. Razumno je pretpostaviti da neće svih m prediktora biti takvih, pa će krajnji model obično sadržavati manje varijabli, recimo k, $k < m$. U tom slučaju, želimo testirati postoji li značajna razlika između početnog i krajnjeg modela. Slično, koristit ćemo testnu statistiku G

$$G = -2ln\left(\frac{vjerodostojnost \text{ modela s } k \text{ varijabli}}{vjerodostojnost \text{ modela s } m \text{ varijabli}}\right) \sim \chi^2(m-k). \tag{2.54}$$

Ako se ispostavi da ne postoji statistička značajna razlika između ta dva modela, opredijelit ćemo se, jasno, za uži model s k varijabli jer možemo zaključiti da one jednako dobro (ili barem ne statistički značajno lošije) opisuju dane podatke kao i svih m varijabli. Dodatno, koristeći manje varijabli izbjegavamo *overfitting* i smanjujemo prostornu i vremensku kompleksnost računalnih programa.

Još jedan test koji testira hipoteze (2.49) je Waldov test, sa testnom statistikom

$$W = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \sim N(0,1) \tag{2.55}$$

pri čemu je $\widehat{SE}(\hat{\beta}_j)$ procijenjena standardna pogreška od $\hat{\beta}_j$. Pojedini programi koriste i $W^2 \sim \chi^2(1)$. Pokazalo se da Waldov test često čini pogrešku 2. vrste, ne odbacujući H_0 u slučajevima kada se β_j pokazao značajnim koristeći *test omjera vjerodostojnosti*. Kakogod, vrijednosti statistika G i W^2 su se obično pokazale vrlo sličnima. Ipak, ako nema razloga za drugačiji odabir, bit ćemo privrženi pouzdanijem *testu omjera vjerodostojnosti*.

Sada nas zanima kako interpretirati procijenjene koeficijente $\hat{\boldsymbol{\beta}}$. Želimo zaključiti što konkretno $\hat{\beta}_j$ govori o utjecaju prediktora $\boldsymbol{x}_j$ na zavisnu varijablu PD. Podijelimo situaciju u 3 slučaja:

- $\boldsymbol{x}_j$ je binarna, tj. $x_{ij} \in \{0,1\}$ - za p_i koristimo oznaku $p_i(0)$ kada je pripadni $x_{ij} = 0$ i $p_i(1)$ ako je pripadni $x_{ij} = 1$. Definiramo *omjer šansi* (eng. *odds ratio*, kratica *OR*)

kao

$$OR(x_{ij}=1, x_{ij}=0) = \frac{\frac{p_i(1)}{1-p_i(1)}}{\frac{p_i(0)}{1-p_i(0)}} = [(2.38)] = \frac{\left(\frac{exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_j\cdot 1+\cdots+\beta_m x_{im})}{1+exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_j\cdot 1+\cdots+\beta_m x_{im})}\right)}{\left(\frac{1}{1+exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_j\cdot 1+\cdots+\beta_m x_{im})}\right)}$$
$$\frac{\left(\frac{exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_j\cdot 0+\cdots+\beta_m x_{im})}{1+exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_j\cdot 0+\cdots+\beta_m x_{im})}\right)}{\left(\frac{1}{1+exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_j\cdot 0+\cdots+\beta_m x_{im})}\right)}$$
$$= \frac{exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_j+\cdots+\beta_m x_{im})}{exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im})} = exp(\beta_j). \quad (2.56)$$

Dakle, prelazak iz kategorije 0 u kategoriju 1 povećava *omjer šansi* za ulazak u *default* za $exp(\hat{\beta}_j)$. Drugim riječima, ulazak u *default* $exp(\hat{\beta}_j)$ je puta izgledniji (u terminu *šansi*) među klijentima u kategoriji 1 nego klijentima u kategoriji 0;

- $\boldsymbol{x}_j$ je kategorijska sa k mogućih kategorija - u ovom slučaju x_{ij} modeliramo tzv. *dummy* varijablama uz kodiranje

$$0 = 0.\text{ kategorija}$$
$$1 = 1.\text{ kategorija}$$
$$\vdots$$
$$k = k\text{-ta kategorija.}$$

Bez smanjenja općenitosti, neka je $\boldsymbol{x}_j$ m-ti prediktor u modelu. Dobivamo model $ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{m-1} x_{im-1} + \beta_{m_1} x_{im_1} + \cdots + \beta_{m_k} x_{im_k}$. Kategoriju kodiranu sa 0 proglašavamo referentnom kategorijom (eng. *baseline*) i procijenjene parametre vezane uz ostale kategorije interpretiramo u odnosu na referentnu kategoriju. Izraz za *omjer šansi* jednak je kao i u (2.56). Primjerice, klijenti svrstani u 1. kategoriju imaju $exp(\hat{\beta}_{m_1})$ puta veći *omjer šansi* ulaska u *default* nego klijenti u *baseline* kategoriji, dok klijenti u k-toj kategoriji imaju $exp(\hat{\beta}_{m_k})$ puta veći *omjer šansi* ulaska u *default* nego klijenti u *baseline* kategoriji. Potrebno je naglasiti da se može desiti da pripadni koeficijent uz neku od k kategorija ispadne statistički značajan, dok za ostale kategorije može ispasti neznačajan. U tom slučaju u modelu ostavljamo sve varijable $x_{im_1}, ..., x_{im_k}$. Naime, nema smisla izbaciti samo neke *dummy* varijable jer se sve odnose na isti prediktor $\boldsymbol{x}_j$. Ako su svi $\beta_{m_1}, ..., \beta_{m_k}$ statistički neznačajni, iz modela izbacujemo sve *dummy* varijable, a time i kompletni prediktor $\boldsymbol{x}_j$;

- $\boldsymbol{x}_j$ je numerička, tj. neprekidna - u ovom je slučaju *omjer šansi* ponovno

$$OR(x_{ij} = x+1, x_{ij} = x) = exp(\beta_j). \quad (2.57)$$

U slučaju neprekidnog tipa podatka, može nas zanimati i pomak od c jedinica:

$$OR(x_{ij} = x+c, x_{ij} = x) = exp(c \cdot \beta_j). \quad (2.58)$$

Primjerice, ako je x_{ij} plaća i-tog klijenta, povišicom od c €, *omjer šansi* ulaska u *default* i-tog klijenta povećao se za $exp(c \cdot \hat{\beta}_j)$.

Primijetimo da se može dogoditi i da koeficijent $\hat{\beta}_j$ uz prediktor $\boldsymbol{x}_j$ bude negativan. U tom je slučaju interpretacija slična: pomak za jednu jedinicu od x_{ij} znači $exp(\hat{\beta}_j)$ puta veći *omjer šansi* ulaska u *default* i-tog klijenta. Međutim, u ovom je slučaju $exp(\hat{\beta}_j) < 1$, što zapravo znači: pomak za jednu jedinicu od x_{ij} znači $\frac{|exp(\hat{\beta}_j)-1|}{exp(\hat{\beta}_j)}$ puta manji *omjer šansi* ulaska u *default* i-tog klijenta.

Selekcija optimalnog modela

Prethodno, diskutirali smo statističku značajnost koeficijenta vezanih uz pojedini prediktor. U ovom paragrafu želimo razviti strategiju za odabir prediktora koji će biti dio krajnjeg modela. Cilj je dobiti model koji se na neki način ponaša "najbolje" unutar danih podataka. Banke generalno skupljaju mnogo podataka o svojim klijentima, čime raste broj potencijalnih varijabli. Rastom broja potencijalnih varijabli raste i broj interakcija među njima (a koje potencijalno također mogu biti statistički značajne), a time i kompleksnost procesa razvoja modela. Redoslijed dodavanja i eliminiranja varijabli također može utjecati na statističku značajnost koeficijenata vezanih uz pojedine varijable. Stoga nije dovoljno odabrati samo one prediktore koji se "na prvu" čine značajnima, već je poželjno razviti plan kojim ćemo odabirati varijable prikladne za model. Pod statistički značajnim, jasno, misli se na p-vrijednost testnih statistika G i W.

Jedna vrsta razvoja modela jest manualna, tj. ona u kojoj analitičar sam procijenjuje koje varijable uvrstiti u model, umjesto da to za njega učini neka statistička procedura. Plan se sastoji od nekoliko koraka [5]:

1. Univarijatna analiza svake nezavisne varijable, tj. analiza modela

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} \tag{2.59}$$

   pri čemu za $\boldsymbol{x}_1$ uzimo sve redom potencijalne prediktore. U ovom koraku, za kandidate ćemo uzeti sve varijable čija je p-vrijednost manja od 0.25. Naime, zbog jednostavnosti ovakvog modela, pokazalo se da se korištenjem standardne razine značajnosti od 5% u ovom koraku često odbace varijable koje se u konačnosti pokažu bitnima;

2. Razvoj multivarijatnog modela sa svakom od varijabli izabranih u koraku 1 i ispitivanje značajnosti pripadnih koeficijenata koristeći p-vrijednost testne statistike W.

Varijable koje se ne pokažu značajnima na standradnoj razini značajnosti od 5% odbacujemo i radimo novi, uži multivarijatni model sa značajnim varijablama. Zatim testiramo razlikuje li se značajno podmodel od prvog, šireg modela koristeći *test omjera vjerodostojnosti* i p-vrijednost testne statistike G iz (2.54);

3. U slučaju da smo prihvatili uži model, promatramo vrijednosti procijenjenih koeficijenata $\hat{\beta}_j^{(new)}$ i uspoređujemo ih sa vrijednostima procijenjenih koeficijenata $\hat{\beta}_j^{(old)}$ iz prvog, šireg modela. Od interesa će nam biti varijable čiji se $\hat{\beta}_j^{(new)}$ i $\hat{\beta}_j^{(old)}$ razlikuju za više od 25% po mjeri $\triangle\hat{\beta}_j$:

$$\triangle\hat{\beta}_j = 100\% \cdot \left| \frac{\hat{\beta}_j^{(new)} - \hat{\beta}_j^{(old)}}{\hat{\beta}_j^{(old)}} \right|. \tag{2.60}$$

Varijabla čiji je $\triangle\hat{\beta}_j > 25\%$ sugerira da se efekt te varijable morao previše prilagoditi izostankom jedne od eliminiranih varijbli. Eliminirane varijable koje imaju takav utjecaj trebale bi biti dodane nazad u model;

4. U model izabran pri kraju koraka 3 dodajemo sve varijable označene kao neznačajne u koraku 1, jednu po jednu, i provjeravamo p-vrijednost pripadajućeg koeficijenta koristeći statistike W i G. Ovaj korak je vitalan za identifikaciju varijabli koje su same po sebi neznačajne, ali postaju značajne u prisustvu drugih varijabli;

5. Za odabrane varijable, još jednom provjeravamo sve pretpostavke modela opisane ranije;

6. Ako su pretpostavke zadovoljene, provjeravamo statističku značajnost interakcija odabranih varijabli koristeći univarijatni model. Drugim riječima, ako smo na kraju koraka 5 odabrali m značajnih varijabli $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$, promatramo $\binom{m}{2}$ univarijatnih modela

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1(x_{ij} \cdot x_{ik}) \qquad j, k = 1, ..., m,\ j \neq k. \tag{2.61}$$

One interakcije koje se pokažu značajnima dodajemo u model s kraja koraka 5, i ponavljamo korak 2, pri čemu eliminiramo samo neznačajne interakcije;

7. Prije nego što model proglasimo finalnim, moramo provjeriti njegovu adekvatnost, valjanost i prediktivnu snagu (sljedeća sekcija).

Druga vrsta selekcije modela jest stepenastim (eng. *stepwise*) metodama, koje su za razliku od prethodno opisane metode, automatizirane. Ovakav algoritam u svakom koraku dodaje (ili oduzima) varijable iz modela na osnovu nekog unaprijed definiranog kriterija, od kojih

se najčešće koristi *Akaike informacijski kriterij* (eng. *Akaike Information Criterion*, kratica *AIC*) definiran sa

$$AIC = -2l(\boldsymbol{p}; \boldsymbol{y}) + 2m. \tag{2.62}$$

Na neki način, *AIC* preko *log-vjerodostojnosti* ocijenjuje i valjanost modela, ali dodaje i kaznu kao rastuću funkciju broja procijenjenih parametara m. Generalno, zahtijevat ćemo da krajnji model ima što manji *AIC*. Komponente *stepwise* regresije jesu:

- Selekcija unaprijed (eng. *forward selection*) - metoda koja započinje praznim modelom i u svakom koraku dodaje onu varijablu koja je statistički najznačajnija. Proces se nastavlja sve dok niti jedna ne odabrana varijabla ne može poboljšati statističku značajnost modela;
- Eliminacija unatrag (eng. *backward elimination*) - metoda koja započinje punim modelom tj. sa svim potencijalnim varijablama i testira izbacivanje svake od njih. Iteracije prestaju kada niti jedna od preostalih varijabli preostalih ne može biti izbačena, a da model pri tome ne izgubi na statističkoj značajnosti;
- Dvosmjerna eliminacija (eng. *bidirectional elimination*) - kombinaciju dvaju gornjih metoda, tj. ona koja u svakom koraku testira koje varijable mogu biti dodane, a koje eliminirane.

Valja napomenuti da nam niti jedna metoda ne garantira da smo dobili univerzalno najbolji model. I manualan odabir modela i stepenaste metode nam u konačnici daju jedan krajnji model, dok u stvarnosti može postojati nekoliko jednako dobrih modela. Također, drugačiji skup podataka možda bi pridonio potpuno drugačijim rezultatima. Opisane metode ono su najbolje što možemo učiniti u kontekstu danog problema i dane metodologije, ali kao i uvijek u statistici, nema garancije da će buduća predviđanja biti apsolutno točna.

2.4 Valjanost modela (eng. *Goodness of Fit)*

Jednom kada smo se odlučili za krajnji model, želimo ocijeniti koliko se procijenjene vjerojatnosti $\hat{p}_1, ..., \hat{p}_n$ podudaraju sa stvarnim opservacijama $y_1, ..., y_n$ na skupu podataka korištenima u izradi modela. Ovakve mjere poznate su kao mjere valjanosti modela (eng. *goodness of fit*). Nadalje, želimo da model ima dobru prediktivnu snagu, tj. da procijenjeni *PD* novog klijenta dobro odražava hoće li taj klijent jednog dana zaista ući u *default* ili neće. Osim već spomenutog *AIC* kriterija, uvest ćemo osnovne mjere koje će nam reći koliko je model valjan i koliko je prediktivan. Prije svega, označimo sa I broj različitih vrijednosti od $\boldsymbol{x}_i^T$ (za što u nastavku koristimo termin "uzorak kovarijata"). Ako se podaci nekih klijenata podudaraju (što nije nemoguće, posebno ako su prediktori kategorijske varijable), vrijedi $I < n$. Označimo broj takvih klijenata (dakle, kod kojih vrijedi $\boldsymbol{x}^T = \boldsymbol{x}_i^T$) sa

m_i, $i = 1, ..., I$. Primijetimo da vrijedi $\sum_{i=1}^{I} m_i = n$. Prva skupina mjera valjanosti bazira se na rezidualima. Procijenjena vjerojatnost za i-ti uzorak kovarijata bit će

$$m_i\hat{p}_i = [(2.38)] = m_i \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_m x_{im})}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_m x_{im})}, \quad i = 1, ..., I. \tag{2.63}$$

Pearsonov rezidual jest

$$r_i = \frac{y_i - m_i\hat{p}_i}{\sqrt{m_i\hat{p}_i(1 - \hat{p}_i)}}, \tag{2.64}$$

a *Pearsonova chi-kvadrat statistika*

$$\chi^2 = \sum_{i=1}^{I} r_i^2 \sim \chi^2(I - m - 1). \tag{2.65}$$

Rezidual devijance jest

$$d_i = \pm\left(2\left[y_i \cdot ln\left(\frac{y_i}{m_i\hat{p}_i}\right) + (m_i - y_i) \cdot ln\left(\frac{m_i - y_i}{m_i(1 - \hat{p}_i)}\right)\right]\right)^{1/2}, \tag{2.66}$$

pri čemu za predznak uzimamo predznak od $(y_i - m_i\hat{p}_i)$. Slično, za mjeru valjanosti uzimamo

$$D = \sum_{i=1}^{I} d_i^2 \sim \chi^2(I - m - 1). \tag{2.67}$$

Obje mjere mogu biti korištene za testiranje hipoteza

$$\begin{aligned} H_0 &: \text{model je adekvatan} \\ H_1 &: \text{model nije adekvatan.} \end{aligned} \tag{2.68}$$

Problem radi činjenica da χ^2 distribucija nije najpreciznija kada $I \approx n$. U tom slučaju, ideja je grupirati opservacije s obzirom na sortirane procijenjene vjerojatnosti i tako dobiti slučaj $I < n$. Uobičajeno je koristiti $g = 10$ grupa, pa će tako 1. grupi biti pridodani klijenti čiji je $\hat{p}_i \leq 0.1$, dok će 10. grupi biti pridodani klijenti čiji je $\hat{p}_i > 0.9$. Grupiranje može biti izvršeno i s obzirom na percentile sortiranih procijenjenih vjerojatnosti (prvih $\frac{n}{10}$ klijenata s najmanjim $\hat{p}_i$ idu u 1. grupu itd.) umjesto korištenja fiksiranih vrijednosti za $\hat{p}_i$. Definiramo *Hosmer-Lemeshow statistiku valjanosti C* kao

$$C = \sum_{k=1}^{g} \left(\frac{(O_{1k} - E_{1k})^2}{E_{1k}} + \frac{(O_{0k} - E_{0k})^2}{E_{0k}}\right), \tag{2.69}$$

pri čemu je

$$\begin{aligned} O_{1k} &= \sum_{i=1}^{c_k} y_i \cdot \mathbf{1}_{\{y_i=1\}} \\ E_{1k} &= \sum_{i=1}^{c_k} m_i \hat{p}_i \cdot \mathbf{1}_{\{y_i=1\}} \\ O_{0k} &= \sum_{i=1}^{c_k} (m_i - y_i) \cdot \mathbf{1}_{\{y_i=0\}} \\ E_{0k} &= \sum_{i=1}^{c_k} m_i (1 - \hat{p}_i) \cdot \mathbf{1}_{\{y_i=0\}}, \end{aligned} \tag{2.70}$$

a c_k je broj uzoraka kovarijata u k-toj grupi. Pod pretpostavkom H_0 iz (2.68), vrijedi $C \sim \chi^2(g-2)$. Za $n > 1000$, preporučuje se uzeti broj grupa g na način

$$g = \max\left(10, \min\left[\frac{n_1}{2}, \frac{n-n_1}{2}, 2 + 8 \cdot \left(\frac{n}{1000}\right)^2\right]\right), \tag{2.71}$$

pri čemu je $n_1 = \sum_{i=1}^{n} y_i \cdot \mathbf{1}_{\{y_i=1\}}$. Ulogu koeficijenta determinacije R^2 iz linearne regresije preuzima *pseudo* R^2, od kojih postoji nekoliko verzija, a najpopularniji je *McFaddenov* R^2 definiran s

$$R^2_{McF} = 1 - \frac{l_M(\boldsymbol{p}; \boldsymbol{y})}{l_0(\boldsymbol{p}; \boldsymbol{y})} \in [0, 1], \tag{2.72}$$

pri čemu je l_M *log-vjerodostojnost* modela, a l_0 *log-vjerodostojnost* modela bez nezavisnih varijabli, dakle samo s interceptom $\hat{\beta}_0$.

Klasifikacijsku točnost našeg modela može nam reći površina ispod *ROC* (kratica za eng. *receiver operating characteristic*) krivulje. Primjerice, želimo klasificirati klijenta i s obzirom na njegov $\hat{p}_i$. Ako je $\hat{p}_i > w$ za neki prag $w \in (0, 1)$, klasificirat ćemo ga kao klijenta ušlog u *default*, a u suprotnom, nećemo. Možemo formirati kontigencijsku tablicu

	$y_i = 1$	$y_i = 0$
$\hat{p}_i > w$	a	b
$\hat{p}_i < w$	c	d

i nadati se što manjem broju pogrešno klasificiranih b i c. U terminologiji medicinskih dijagnostičkih testova, udio stvarno pozitivnih u ukupnom broju pozitivno klasificiranih $\frac{a}{a+c}$ poznat je kao i *senzitivnost* testa, dok je udio stvarno negativnih u ukupnom broju negativno klasificiranih $\frac{d}{b+d}$ *specifičnost* testa. *ROC* krivulja grafički prikazuje omjer stvarno

pozitivnih (*senzitivnost* na y osi) i lažno pozitivnih (1−*specifičnost* na x osi) za sve vrijednosti praga w. Što je model bolji klasifikator, to će površina ispod *ROC* krivulje biti bliže 1, pa i njegova prediktivna snaga veća.

Poglavlje 3

Primjer kroz R

Podaci su preuzeti sa [1]. Sve skupa imamo podatke za $n = 700$ klijenata jedne banke. Pogledajmo kako izgleda prvih nekoliko podataka:

	age	ed	employ	income	debtinc	creddebt	othdebt	default
1	41	3	17	176	9.3	11.359392	5.008608	1
2	27	1	10	31	17.3	1.362202	4.000798	0
3	40	1	15	55	5.5	0.856075	2.168925	0
4	41	1	15	120	2.9	2.658720	0.821280	0
5	24	2	2	28	17.3	1.787436	3.056564	1
6	41	2	5	25	10.2	0.392700	2.157300	0

Nezavisne varijable su:

- **age** $\in \mathbb{N}$ - dob klijenta izražena u godinama;
- **ed** $\in \{1, ..., 5\}$ - kategorijska varijabla koja označava razinu obrazovanja klijenta;
- **employ** $\in \mathbb{N}$ - radno iskustvo klijenta izraženo u godinama;
- **income** $\in \mathbb{N}$ - godišnja plaća klijenta izražena u tisućama USD;
- **debtinc** $\in [0, 100]$ - omjer ukupnog mjesečnog duga i ukupnih mjesečnih primanja klijenta izražen u postotku (eng. *debt to income ratio*);
- **creddebt** $\in [0, 100]$ - omjer potrošenog iznosa na revolving kreditnim karticama i ukupnog iznosa kredita dostupnog klijentu izražen u postotku (eng. *debt to credit ratio*);
- **othdebt** $\in [0, 100]$ - postotak ukupnog mjesečnog duga klijenta koji otpada na "ostale dugove" (dakle, koji nisu otplaćivanje kuće ili stana, automobila, ostalih kredita i sl.).

Zavisna varijabla je:

- **default** $\in \{0, 1\}$ - je li klijent ušao u *default* (1) ili nije (0).

Struktura podataka vidljiva je i iz sljedećeg ispisa:

```
'data.frame': 700 obs. of  8 variables:
 $ age     : int  41 27 40 41 24 41 39 43 24 36 ...
 $ ed      : Factor w/ 5 levels "1","2","3","4",..: 3 1 1 1 ...
 $ employ  : int  17 10 15 15 2 5 20 12 3 0 ...
 $ income  : int  176 31 55 120 28 25 67 38 19 25 ...
 $ debtinc : num  9.3 17.3 5.5 2.9 17.3 10.2 30.6 3.6 24.4 19.7 ...
 $ creddebt: num  11.359 1.362 0.856 2.659 1.787 ...
 $ othdebt : num  5.009 4.001 2.169 0.821 3.057 ...
 $ default : int  1 0 0 0 1 0 0 0 1 0 ...
```

Prije svega, uvjerimo se da nema *missing* vrijednosti (provjeravanjem je li neki podatak NA - kratica za eng. *not available*) i da imamo predstavnike iz obje kategorije zavisne varijable:

```
FALSE
 5600


  0   1
517 183
```

Kao što je opisano u prethodnom poglavlju, krećemo od univarijatne logističke regresije za svaku od nezavisnih varijabli:

```
model_1=glm(default~age,data=podaci,family=binomial(link="logit"))
```

Funkcijom **glm** određujemo da se radi o *generaliziranom linearnom modelu*, dok u argumentima navodimo željenu formulu, naziv *data frame-a* gdje su spremljeni podaci, željenu distribuciju zavisne varijable i željenu *funkciju veze*. Sve najbitnije informacije o modelu dohvaćamo naredbom **summary**:

```
Call:
glm(formula = default ~ age, family = binomial(link = "logit"),
    data = podaci)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9916  -0.8240  -0.7148   1.4120   2.0158
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.36693    0.39217   0.936 0.349452
age         -0.04105    0.01138  -3.609 0.000307 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 790.73  on 698  degrees of freedom
AIC: 794.73

Number of Fisher Scoring iterations: 4
```

Prvo možemo vidjeti karakterističnu petorku *reziduala devijance* iz (2.66). Uz varijablu **age**, nalazi se pripadni procijenjeni parametar, procijenjena standardna pogreška, testna statistika (2.55) i pripadajuća p-vrijednost, iz koje možemo zaključiti da varijabla **age** značajno utječe na vjerojatnost ulaska u *default*. Null deviance odnosi se na vrijednost (2.52) sa samo interceptom, dok se Residual deviance odnosi na vrijednost (2.52) sa varijablom **age** u modelu. Konačno, ispisana je i vrijednost *Akaike informacijskog kriterija*. Number of Fisher Scoring iterations označuje broj iteracija (2.30) potrebnih programu za procjenu parametara modela. Iz gornjeg ispisa također možemo iščitati i jednadžbu dobivenog modela, koja glasi

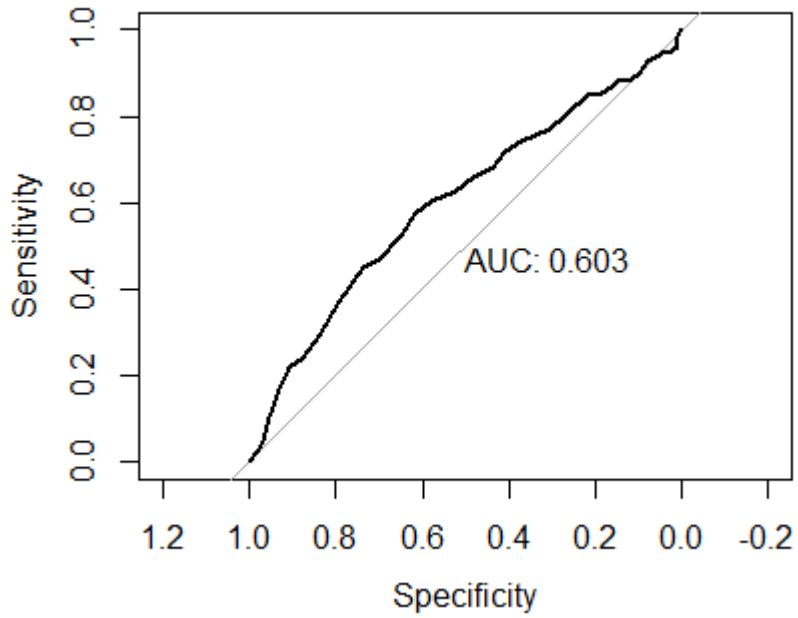
$$ln\left(\frac{p_i}{1-p_i}\right) = 0.36693 - 0.04105 \cdot \text{age}_i. \tag{3.1}$$

Primjerice, vjerojatnost da osoba od 25 godina uđe u *default* je po (2.38)

$$\frac{exp(0.36693 - 0.04105 \cdot 25)}{1 + exp(0.36693 - 0.04105 \cdot 25)} = 0.34089. \tag{3.2}$$

Krajnje, možemo zaključiti povećanje dobi za 1 godinu povećava *omjer šansi* ulaska u *default* $exp(-0.04105) = 0.96$ puta, odnosno ga smanjuje $\frac{|0.96-1|}{0.96} = 0.0416$ puta ili 4.16%. *ROC* krivulja prikazana je na Slici 3.1, zajedno sa površinom ispod krivulje (*AUC* - kratica za eng. *area under curve*). Sličnu analizu nastavljamo i za sve ostale nezavisne varijable:

```
model_2=glm(default~ed,data=podaci,family=binomial(link="logit"))
model_3=glm(default~employ,data=podaci,family=binomial(link="logit"))
```



Slika 3.1: *ROC* krivulja Modela 1

```
model_4=glm(default~income,data=podaci,family=binomial(link="logit"))
model_5=glm(default~debtinc,data=podaci,family=binomial(link="logit"))
model_6=glm(default~creddebt,data=podaci,family=binomial(link="logit"))
model_7=glm(default~othdebt,data=podaci,family=binomial(link="logit"))
```

Sljedeći ispis prikazuje procijenjene koeficijente i pripadajuće p-vrijednosti iz gore navedenih univarijatnih modela:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
age         -0.04105    0.01138  -3.609 0.000307 ***
ed2          0.45379    0.20054   2.263  0.02364 *
ed3          0.66887    0.25874   2.585  0.00974 **
ed4          0.77173    0.35940   2.147  0.03177 *
ed5         -0.07557    1.12520  -0.067  0.94645
employ      -0.12469    0.01744  -7.150 8.69e-13 ***
income      -0.005349   0.002872  -1.863   0.0625 .
```

```
debtinc      0.13162    0.01424    9.24   <2e-16 ***
creddebt     0.25726    0.04676   5.501 3.77e-08 ***
othdebt      0.09059    0.02463   3.678 0.000235 ***
```

Kao što smo objasnili, u ovom trenutku za razinu značajnosti uzimamo 25%. Vidimo da su sve varijable na toj razini statistički značajne (kao i barem neke od *dummy* varijabli vezanih uz razinu obrazovanja **ed**, pa je time i cjelokupna varijabla **ed** statistički značajna). Stoga u sljedećem koraku radimo multivarijatni model sa svim varijablama:

```
model_8=glm(default~.,data=podaci,family=binomial(link="logit"))
```

Od sada nadalje, koristimo uobičajenu razinu značajnosti od 5%. Promotrimo sada p-vrijednosti svih koeficijenata:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.789543   0.576927  -1.369   0.1711
age         -0.009270   0.015079  -0.615   0.5387
ed2          0.241534   0.246816   0.979   0.3278
ed3          0.253006   0.329657   0.767   0.4428
ed4         -0.182774   0.458806  -0.398   0.6904
ed5          0.815577   1.271628   0.641   0.5213
employ      -0.236841   0.031229  -7.584 3.35e-14 ***
income      -0.007980   0.007319  -1.090   0.2756
debtinc      0.070012   0.029184   2.399   0.0164 *
creddebt     0.573724   0.107708   5.327 1.00e-07 ***
othdebt      0.041019   0.072915   0.563   0.5737
```

U ovakvom modelu, možemo zaključiti da varijable **employ**, **debtinc** i **creddebt** značajno utječu na *PD*. Stoga nam je sljedeći potencijalni model onaj sa te 3 varijable:

```
model_9=glm(default~employ+debtinc+creddebt,data=podaci,
            family=binomial(link="logit"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.22661    0.23120  -5.305 1.12e-07 ***
employ      -0.24427    0.02726  -8.960  < 2e-16 ***
debtinc      0.08765    0.01815   4.830 1.36e-06 ***
creddebt     0.50275    0.08086   6.218 5.04e-10 ***
```

Kao što vidimo, sve 3 varijable su nam i dalje značajne. Sljedeći korak nam je utvrditi postoji li statistički značajna razlika između Modela 8 i Modela 9, jasno, koristeći (2.54). Dohvatimo *log-vjerodostojnosti* oba modela:

```
> logLik(model_8)
'log Lik.' -285.9243 (df=11)

> logLik(model_9)
'log Lik.' -287.8181 (df=4)
```

Sada imamo

$$G = -2(-287.8181 - (-285.9243)) = 3.7876 \sim \chi^2(11-4). \tag{3.3}$$

Pripadnu p-vrijednost sada dobivamo kao:

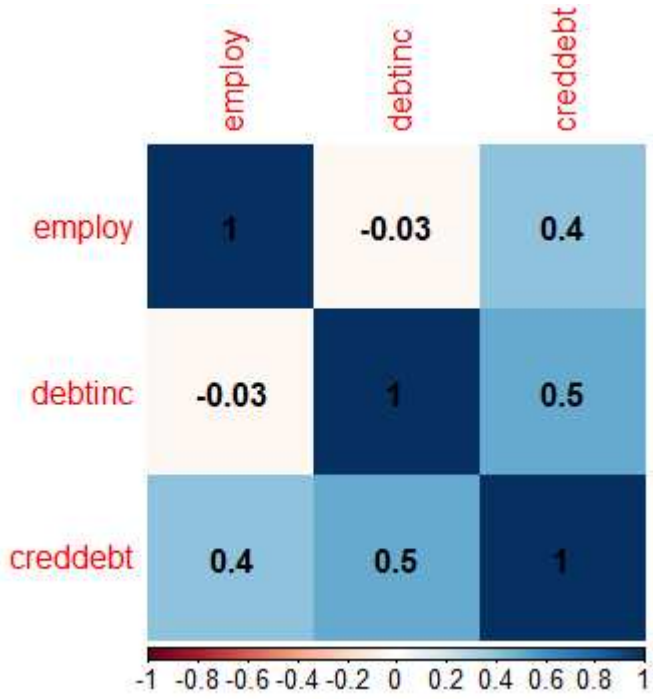
```
> 1-pchisq(3.7876,7)
[1] 0.803891
```

Dakle, ne postoji statistički značajna razlika između ta 2 modela. Drugim riječima, izbacivanjem varijabli **age**, **ed**, **income** i **othdebt** nismo izgubili na značajnosti modela, a dobili smo model sa puno manje varijabli. Možemo zaključiti da su **employ**, **debtinc** i **creddebt** dovoljne za procjenu *PD*-ja. Idemo usporediti vrijednosti njihovih koeficijenata u Modelu 9 sa onima u Modelu 8, kao što je opisano u (2.60).

```
              Model 8   Model 9   delta_beta
employ      -0.236841  -0.24427   0.03136952
debtinc      0.070012   0.08765   0.25194196
creddebt     0.573724   0.50275   0.12371305
```

Vidimo da nam promjena koeficijenta vezanog uz **debtinc** prelazi zadanu granicu od 25%. Međutim, s obzirom na malu magnitudu tih vrijednosti i činjenicu da se ne radi o velikom prekoračenju, radi jednostavnosti ćemo za sada prihvatiti Model 9. S obzirom da su se u univarijatnoj analizi sve varijable pokazale značajnima, preskačemo korak 4 i prelazimo na provjeravanje pretpostavki Modela 9. Prvo ćemo provjeriti postoji li *multikolinearnost* između **employ**, **debtinc** i **creddebt**. Matrica *korelacija* prikazana je na Slici 3.2, iz koje možemo vidjeti da nema pretjerano *koreliranih* varijabli, što je dobar znak. Nadalje, dijagnostičke mjere *multikolinearnosti* pokazuju sljedeće:

```
Call:
omcdiag(mod = model_9)
```



Slika 3.2: Matrica *korelacija* Modela 9

```
Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.5717         0
Farrar Chi-Square:       389.8678         1
Red Indicator:             0.3723         0
Sum of Lambda Inverse:     4.5207         0
Theil's Method:            0.4376         0
Condition Number:          5.8496         0

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Vidimo da $det(\boldsymbol{X}^T\boldsymbol{X})$ nije blizu 0, a većina ostalih mjera također sugerira da ne postoji *multikolinearnost*, što piše pod "detection". Jedina mjera koja bi potencijalno ukazala na to jest *Farrarov* χ^2. Idemo stoga provjeriti VIF svih varijabli:

```
Call:
```

```
imcdiag(mod = model_9, method = "VIF")

 VIF Multicollinearity Diagnostics

            VIF detection
employ   1.3089         0
debtinc  1.4642         0
creddebt 1.7476         0

NOTE:  VIF Method Failed to detect multicollinearity

0 --> COLLINEARITY is not detected by the test
```

S obzirom na male VIF vrijednosti, dobru matricu *korelacija* i dijagnostičke mjere, možemo zaključiti kako nemamo problema s *multikolinearnosti*. Sljedeću pretpostavku koju provjeravamo jest linearnost između $logit(\boldsymbol{p})$ i **employ**, **debtinc** i **creddebt**. Radimo, dakle, Box-Tidwellow test. Prvo kreiramo 3 pomoćne varijable:

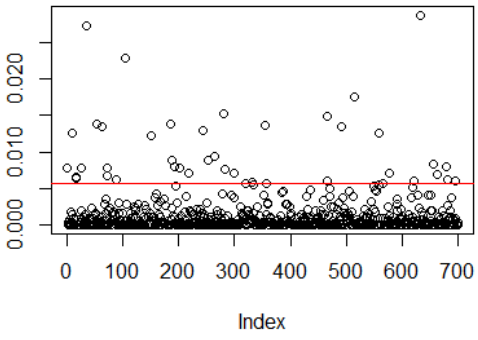
```
podaci$trans_employ=podaci$employ*log(podaci$employ)
podaci$trans_debtinc=podaci$debtinc*log(podaci$debtinc)
podaci$trans_creddebt=podaci$creddebt*log(podaci$creddebt)
```

Sada promatramo model

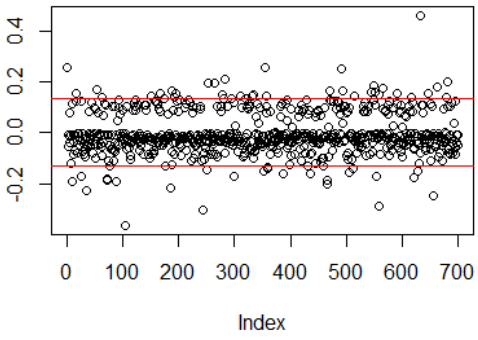
```
model_10=glm(default~employ+debtinc+creddebt+trans_employ
             +trans_debtinc+trans_creddebt,
             data=podaci,family=binomial(link="logit"))

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.3108126  0.7240407  -1.810   0.0702 .
employ         -0.1535943  0.1943720  -0.790   0.4294
debtinc         0.0969432  0.1951966   0.497   0.6194
creddebt        0.3072205  0.2947173   1.042   0.2972
trans_employ   -0.0336669  0.0651001  -0.517   0.6050
trans_debtinc   0.0008188  0.0540918   0.015   0.9879
trans_creddebt  0.0980695  0.1286546   0.762   0.4459
```

Svi koeficijenti uz transformirane varijable su statistički neznačajni. Prema tome, pretpostavka je ispunjena. Nadalje, provjeravamo postojanje utjecajnih točaka, dakle *Cookovu udaljenost* i $DFFITS$. Slike 3.3 i 3.4 pokazuju da ima nekoliko podataka koji nadmašuju zadane pragove. Idemo ih probati izbaciti iz modela i vidjeti kakve ćemo rezultate dobiti:



Slika 3.3: *Cookova udaljenost* - Model 9



Slika 3.4: *DFFITS* - Model 9

```
new_podaci=podaci[podaci$cd<4/700 & podaci$dffits<2*sqrt(3/700)
                  & podaci$dffits>-2*sqrt(3/700),]
```

Preostalo je ukupno $n = 628$ podataka.

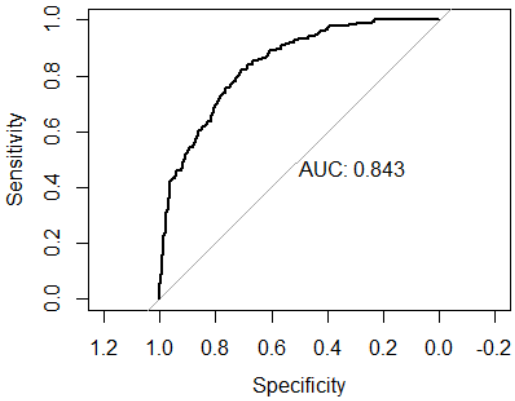
```
model_11=glm(default~employ+debtinc+creddebt,data=new_podaci,
             family=binomial(link="logit"))
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.03248    0.33670  -6.036 1.58e-09 ***
employ      -0.44860    0.04997  -8.978  < 2e-16 ***
debtinc      0.18553    0.03170   5.853 4.81e-09 ***
creddebt     0.81343    0.16408   4.958 7.14e-07 ***
```

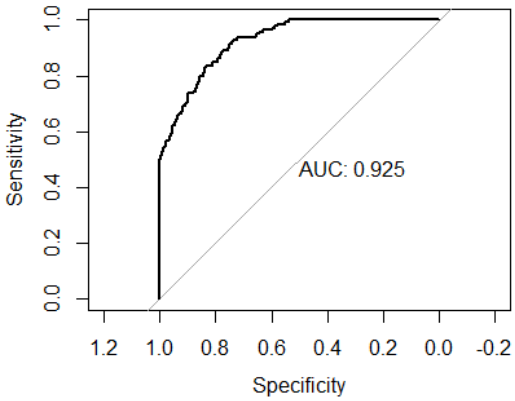
Sve varijable su i dalje statistički značajne. Idemo usporediti *AIC*, R^2_{McF} i površinu ispod *ROC* krivulje Modela 9 i Modela 11:

```
               Model 9   Model 11
AIC             583.64     344.38
McFadden    0.2843587  0.4952403
```

Vidimo da se u Modelu 11 *AIC* značajno smanjio, a R^2_{McF} značajno povećao. Na temelju Slika 3.5 i 3.6, Model 11 je model veće prediktivne snage. Zaključujemo da je izbacivanje utjecajnih točaka bio dobar izbor i prihvaćamo Model 11 boljim. Sada prelazimo na testiranje značajnosti mogućih interakcija između **employ**, **debtinc** i **creddebt**. S ozbirom da imamo 3 varijable, imamo $\binom{3}{2} = 3$ moguće kombinacije:



Slika 3.5: *ROC* krivulja Modela 9



Slika 3.6: *ROC* krivulja Modela 11

```
model_12=glm(default~employ:debtinc,data=new_podaci,
             family=binomial(link="logit"))
model_13=glm(default~employ:creddebt,data=new_podaci,
             family=binomial(link="logit"))
model_14=glm(default~debtinc:creddebt,data=new_podaci,
             family=binomial(link="logit"))
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
employ:debtinc -0.0006385  0.0010528  -0.606    0.544
employ:creddebt  0.003658   0.002129   1.718   0.0858 .
debtinc:creddebt 0.030084   0.004063   7.404 1.32e-13 ***
```

Na razini značajnosti 5%, interakcija između **debtinc** i **creddebt** je statistički značajna. Dodajmo ju stoga u Model 11:

```
model_15=glm(default~employ+debtinc+creddebt+debtinc*creddebt,
             data=new_podaci,family=binomial(link="logit"))
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.947680   0.435481  -4.472 7.73e-06 ***
employ          -0.449752   0.050145  -8.969  < 2e-16 ***
debtinc          0.178695   0.039112   4.569 4.90e-06 ***
creddebt         0.737393   0.290839   2.535   0.0112 *
```

```
debtinc:creddebt  0.005235   0.017448   0.300   0.7642
```

U multivarijatnom modelu interakcija se ipak ne pokazuje značajnom. Stoga proglašavamo Model 11 konačnim.

Sažetak odabranog modela

Odlučili smo se, dakle, za Model 11:

```
model_11=glm(default~employ+debtinc+creddebt,data=new_podaci,
             family=binomial(link="logit"))

Call:
glm(formula = default ~ employ + debtinc + creddebt,
    family = binomial(link = "logit"), data = new_podaci)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.58595  -0.44712  -0.13684  -0.01116   2.62466

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.03248    0.33670  -6.036 1.58e-09 ***
employ      -0.44860    0.04997  -8.978  < 2e-16 ***
debtinc      0.18553    0.03170   5.853 4.81e-09 ***
creddebt     0.81343    0.16408   4.958 7.14e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 666.42  on 627  degrees of freedom
Residual deviance: 336.38  on 624  degrees of freedom
AIC: 344.38

Number of Fisher Scoring iterations: 7
```

Jednadžba modela glasi

$$ln\left(\frac{p_i}{1-p_i}\right) = -2.03248 - 0.44860 \cdot \text{employ}_i + 0.18553 \cdot \text{debtinc}_i + 0.81343 \cdot \text{creddebt}_i. \quad (3.4)$$

Interpretacija je sljedeća:

- Povećanje radnog iskustva za 1 godinu znači $\frac{|exp(-0.44860)-1|}{exp(-0.44860)} = 0.56612$ puta (56.61%) manji *omjer šansi* ulaska u *default*;
- Povećanje omjera ukupnog mjesečnog duga i ukupnih mjesečnih primanja za 1% znači $exp(0.18553) = 1.20386$ puta (20.39%) veći *omjer šansi* ulaska u *default*;
- Povećanje omjera potrošenog iznosa na revolving kreditnim karticama i ukupnog iznosa dostupnog kredita za 1% znači $exp(0.81343) = 2.25563$ puta (125.56%) veći *omjer šansi* ulaska u *default*.

Primijetimo da su rezultati sukladni intuitivnim očekivanjima. Što klijent ima dulji radni staž, to mu je manji očekivani *PD*, što ga banci čini pouzdanijim. S druge strane, veći dugovi obično će značiti i veći *PD*, odnosno klijenta manjeg kredibiliteta. Idemo izračunati vjerojatnost ulaska u *default* na nekoliko primjera. Osoba zaposlena 25 godina sa debtinc = 5% i creddebt = 10% ima vjerojatnost ulaska u *default*

$$\frac{exp(-2.03248 - 0.44860 \cdot 25 + 0.18553 \cdot 5 + 0.81343 \cdot 10)}{1 + exp(-2.03248 - 0.44860 \cdot 25 + 0.18553 \cdot 5 + 0.81343 \cdot 10)} = 0.01499. \tag{3.5}$$

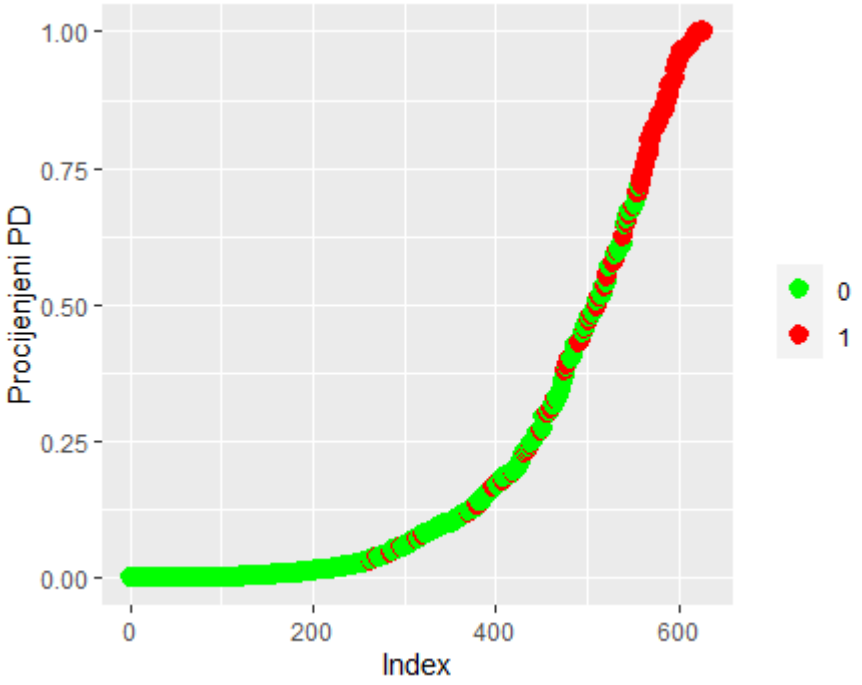
Nije nerazumno pretpostaviti da prosječan hrvatski student nema radnog iskustva, ali s druge strane, nema niti dugova. Stoga bi pri aplikaciji za svoj prvi kredit, npr. svježe diplomirani magistar/ica matematike (prema ovom modelu) ostvario procjenu *PD*-ja

$$\frac{exp(-2.03248)}{1 + exp(-2.03248)} = 0.11583, \tag{3.6}$$

čime bi najvjerojatnije i dobio željeni kredit. Nadalje, vrijednosti *Akaike informacijskog kriterija* i *McFaddenovog* R^2 Modela 11

$$\begin{aligned} AIC &= 344.38 \\ R^2_{McF} &= 0.49524 \end{aligned} \tag{3.7}$$

indiciraju vrlo dobar fit. Površina ispod *ROC* krivulje na Slici 3.6 govori nam da smo dobili model jako dobre prediktivne snage. Krajnje, graf procijenjenih vjerojatnosti skupa sa pravim vrijednostima *defaulta* vidljiv je na Slici 3.7.



Slika 3.7: Procijenjeni *PD*-jevi Modela 11

Za kraj, idemo vidjeti bi li stepenaste metode možda rezultirale drugačijim odabirom modela. Prvo kreiramo nulti model (samo s interceptom) i puni model sa svim varijablama:

```
model_null=glm(default~1,data=new_podaci[1:8],
               family=binomial(link="logit"))
model_full=glm(default~.,data=new_podaci[1:8],
               family=binomial(link="logit"))
```

Iz *data frame*-a new_podaci biramo samo prvih 8 stupaca iz razloga što su u daljnim stupcima sada spremljene transformirane varijable, *Cookove udaljenosti* i $DFFITS$ koje su bile potrebne za raniju analizu. Naredbom **stepAIC** radimo stepenastu regresiju koja vraća najbolji model na osnovu *AIC* kriterija. Za argumente navodimo želimo li selekciju unaprijed, eliminaciju unatrag ili dvosmjernu eliminaciju. Pod argumentom **scope** također navodimo minimalan dopustiv, odnosno maksimalan dopustiv model:

```
forward_model=stepAIC(model_null,direction="forward",
              scope=list(upper=model_full,lower=model_null))
```

```
backward_model=stepAIC(model_full,direction="backward",
               scope=list(upper=model_full,lower=model_null))

bidirectional_model=stepAIC(model_full,direction="both",
                    scope=list(upper=model_full,lower=model_null))
```

Ispisi su sljedeći:

```
summary(forward_model)

Call:
glm(formula = default ~ debtinc + employ + creddebt,
    family = binomial(link = "logit"), data = new_podaci[1:8])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.58595  -0.44712  -0.13684  -0.01116   2.62466

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.03248    0.33670  -6.036 1.58e-09 ***
debtinc      0.18553    0.03170   5.853 4.81e-09 ***
employ      -0.44860    0.04997  -8.978  < 2e-16 ***
creddebt     0.81343    0.16408   4.958 7.14e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 666.42  on 627  degrees of freedom
Residual deviance: 336.38  on 624  degrees of freedom
AIC: 344.38

Number of Fisher Scoring iterations: 7

summary(backward_model)
Call:
glm(formula = default ~ employ + debtinc + creddebt,
    family = binomial(link = "logit"), data = new_podaci[1:8])
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.58595  -0.44712  -0.13684  -0.01116   2.62466

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.03248    0.33670  -6.036 1.58e-09 ***
employ      -0.44860    0.04997  -8.978  < 2e-16 ***
debtinc      0.18553    0.03170   5.853 4.81e-09 ***
creddebt     0.81343    0.16408   4.958 7.14e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 666.42  on 627  degrees of freedom
Residual deviance: 336.38  on 624  degrees of freedom
AIC: 344.38

Number of Fisher Scoring iterations: 7

summary(bidirectional_model)
Call:
glm(formula = default ~ employ + debtinc + creddebt,
    family = binomial(link = "logit"), data = new_podaci[1:8])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.58595  -0.44712  -0.13684  -0.01116   2.62466

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.03248    0.33670  -6.036 1.58e-09 ***
employ      -0.44860    0.04997  -8.978  < 2e-16 ***
debtinc      0.18553    0.03170   5.853 4.81e-09 ***
creddebt     0.81343    0.16408   4.958 7.14e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 666.42  on 627  degrees of freedom
Residual deviance: 336.38  on 624  degrees of freedom
AIC: 344.38

Number of Fisher Scoring iterations: 7
```

Vidimo da sve metode daju jednake rezultate, a i koji se poklapaju sa rezultatima Modela 11 (jasno, ovo generalno neće uvijek biti slučaj). Možemo zaključiti da je po svemu sudeći Model 11 zaista najbolji izbor.

Bibliografija

[1] *Credit Risk Analysis for extending Bank Loans*, https://www.kaggle.com/datasets/atulmittal199174/credit-risk-analysis-for-extending-bank-loans.

[2] C. Bluhm, L. Overbeck i C. Wagner, *An Introduction to Credit Risk Modeling*, A CRC Press Company, 2003.

[3] S. Bouteille i Coogan-Pushner D., *The Handbook of Credit Risk Management*, John Wiley & Sons Inc., Hoboken, New Jersey, USA, 2013.

[4] A.J. Dobson i Barnett A.G., *An Introduction to Generalized Linear Models*, A CRC Press Company, 2008.

[5] D.W. Hosmer, S. Lemeshow i R.X. Sturdivant, *Applied Logistic Regression*, John Wiley & Sons Inc., 2013.

[6] M. Imdadullah, M. Aslam i S. Altaf, *An R Package for Detection of Collinearity among Regressors*, (2016), https://pdfs.semanticscholar.org/e27a/ebeed6b5a77892e128dee083285a5dd4475c.pdf.

[7] D. Jakovčević i I. Jolić, *Kreditni rizik*.

[8] A.A.M. Nurunnabi i M. Nasser, *Outlier Diagnostics in Logistic Regression*, https://www.researchgate.net/publication/237198212_Outlier_Diagnostics_in_Logistic_Regression_A_Supervised_Learning_Technique.

[9] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 1986.

[10] E.N.C. Tong, C. Mues, I. Brown i L.C. Thomas, *Exposure at default models with and without the credit conversion factor*, (2016), https://www.sciencedirect.com/science/article/pii/S0377221716001004.

[11] Z. Vondraček, *Markovljevi lanci*, PMF-MO, nastavni materijali, 2008.

[12] O. Yashkir i Y. Yashkir, *Loss given default modeling: A comparative analysis*, (2013), https://www.researchgate.net/publication/291321360_Loss_given_default_modeling_A_comparative_analysis.

Sažetak

U ovom radu objasnili smo što je kreditni rizik i način na koji su svi sudionici tržišta izloženi kreditnom riziku. Opisali smo zašto je bitno da financijske institucije analiziraju i upravljaju kreditnim rizikom na ispravan način i kako se to odražava na ostvarenje profita. Ključan pojam koji se provlači kroz cijeli diplomski rad jest pojam *defaulta*, klijentovog neplaćanja. Definirali smo *PD*, *LGD* i *EAD* kao ključne indikatore klijentove rizičnosti. S obzirom da se radi o slučajnim varijablama, uveli smo matematičke modele kojima se može pristupiti procjeni navedenih parametara. U praksi se *PD*, vjerojatnost ulaska u *default*, pokazao daleko najbitnijim parametrom, pa je glavni cilj ovog rada bio uvesti model kojim ćemo dobro moći pristupiti procjeni *PD*-ja. S tim razlogom, definirali smo *generalizirani linearni model* i objasnili *metodu maksimalne vjerodostojnosti*. Kao poseban slučaj *GLM*-a koji dobro može odrediti klijentov *PD*, naveli smo model logističke regresije. Pretpostavke koje trebaju biti zadovoljene u tom modelu jesu izostanak *multikolinearnosti* među prediktorima, izostanak utjecajnih točaka i linearnost između *funkcije veze* zavisne varijable i nezavisnih. Demonstrirali smo testiranje značajnosti koeficijenata i interpretaciju rezultata, u terminima *šansi*, *omjera šansi* i vjerojatnosti. Posebnu važnost u izboru krajnjeg modela i ocjeni prediktivnosti modela pokazale su veličine poput *Akaike informacijskog kriterija* i površine ispod *ROC* krivulje. Krajnje, sva teorija potkrijepljena je primjerom, koristeći bazu o klijentima jedne banke i programski jezik R. Na skupu podataka korištenima u primjeru, kao najznačajnijim varijablama, odnosno onima koje najviše utječu na klijentov ulazak u *default*, pokazale su se radno iskustvo, *debt to income ratio* i *debt to credit ratio*. Dulje radno iskustvo pozitivno je utjecalo na vjerojatnost ulaska u *default*, odnosno smanjivalo ju je. S druge strane, veći dugovi značili su i veću vjerojatnost *defaulta*, odnosno rizičnijeg i nepouzdanijeg klijenta.

Summary

In this thesis, it was explained what credit risk is and the means by which all market participants are exposed to it. Successful analysis and management of credit risk proved to be of crucial importance to all financial institutions as it reflects on their profits. Key term that found itself in all the stages of this thesis was *default*, client's failure to pay the lender per initial terms. *PD*, *LGD* and *EAD* were defined as main indicators of client's creditworthiness. Since they are random variables, several mathematical models were introduced as a way to model them. Main goal of this thesis was to establish method of *PD* assessment, as it proved to be of biggest importance among the three. For that reason, class of *generalized linear models* was defined and *maximum likelihood estimation* method explained. Logistic regression was proposed as a special case of *GLM* which can determine client's *PD* particularly well. Assumptions of logistic regression model to be satisfied included absence of *multicollinearity* among regressors, lack of influential data and linearity between *link function* of dependent variable and independent ones. Testing for significance of coefficients was demonstrated, as well as interpretation of results, in terms of *odds*, *odds ratio's* and probability. When deciding for a final model and determining its predictive power, values such as *Akaike information criterion* and area under *ROC* curve were essential. Lastly, to support the theory, example was given. Database with information on one bank's clients and programming language R were used. On this data set, variables work experience, *debt to income ratio* and *debt to credit ratio* showed to be of biggest influence on client's *default*. Longer work experience positively reflected on *PD*, lowering it. On the other side, bigger debts meant larger *PD*, making such a client more risky and unreliable.

Životopis

Rođen sam 27. ožujka 1998. godine u Virovitici. Nakon završetka Osnovne škole Vladimira Nazora u Daruvaru upisujem se u Gimnaziju Daruvar, gdje završavam prva dva razreda. Treći i četvrti razred gimnazijskog programa završavam u Birotehnici - centru za dopisno obrazovanje, nakon čega maturiram 2016. godine. Na temelju pune sportske (tenis) stipendije, akademske godine 2016./2017. upisujem Preddiplomski studij matematike i informatike na Radford University, Radford, Virginia, SAD u trajanju od 4 godine. Tijekom tih godina, član sam muške teniske ekipe na istom Sveučilištu te radim kao tutor za studente-sportaše. Od sportskih postignuća izdvaja se osvajanje Big South konferencije 2019. godine, a od akademskih nagrada ITA Scholar-Athlete 2018., Presidential Honor Roll 2020. i dospijeće na Dekanovu listu svih 8 semestara. 2020. godine stječem titulu prvostupnika znanosti (Bachelor of Science). Akademske godine 2020./2021. upisujem Diplomski sveučilišni studij Matematička statistika na Prirodoslovno-matematičkom fakultetu-Matematički odsjek u Zagrebu. Tijekom navedenog Diplomskog studija, stipendist sam Hrvatske narodne banke.