UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Magistrsko delo

# Modeli za kategorične odzive

(Models for categorical response variables)

Ime in priimek: Maruša Pajk
Študijski program: Matematične znanosti, 2. stopnja
Mentor: izr. prof. dr. Mihael Perman

Koper, september 2015

# Ključna dokumentacijska informacija

Ime in PRIIMEK: Maruša PAJK

Naslov magistrskega dela: Modeli za kategorične odzive

Kraj: Koper

Leto: 2015

Število listov: 66 Število slik: 7 Število tabel: 18

Število referenc: 18

Mentor: izr. prof. dr. Mihael PERMAN

UDK: 519.876.2(043.2)

Ključne besede: logistična regresija, statistično modeliranje, analiza podatkov, kreditno tveganje, bonitetni model

Math. Subj. Class. (2010): 62-07, 62J02, 90C31, 91G40

**Izvleček:**
Regresijske metode so postale glavna komponenta skoraj vsake analize podatkov, ki se ukvarja s povezavo med odzivno spremenljivko in eno ali več pojasnjevalnimi spremenljivkami. Pogosto je izid diskretna spremenljivka, ki zavzame dve ali več možnih vrednosti. Logistična regresija je najpomembnejši izmed modelov za kategorične odzive. V zadnjem času je logistična regresija postala popularna tudi v bančnem svetu, predvsem za modeliranje kreditne sposobnosti komitenta. Namen banke je z uporabo bonitetnega modela minimizirati stroške, ki nastanejo iz naslova kreditnega tveganja, kar omogoči boljše upravljanje s kapitalskimi zahtevami. Cilj magistrske naloge je prikazati postopke in metode, kako dobiti take spremenljivke oziroma kazalnike poslovanja podjetja, ki v največji meri vplivajo na to, da bo komitent banke postal neplačnik in pridobiti formulo za izračun verjetnosti neplačila.

# Key words documentation

Name and SURNAME: Maruša PAJK

Title of master's thesis: Models for categorical response variables

Place: Koper

Year: 2015

Number of pages: 66 Number of figures: 7 Number of tables: 18

Number of references: 18

Mentor: Assoc. Prof. Mihael PERMAN, PhD

UDC: 519.876.2(043.2)

Keywords: logistic regression, statistical modelling, data analysis, credit risk, rating model

Math. Subj. Class. (2010): 62-07, 62J02, 90C31, 91G40

**Abstract:**
Regression methods have become the main component of almost every data analysis, which deals with the relationship between the response variable and one or more explanatory variables. The result is often a discrete variable that takes two or more possible values. Logistic regression is one of the most significant models for categorical responses and has recently become popular also in the banking world, especially for modelling client's creditworthiness. Banks are using rating models to minimize costs resulting from credit risk, which enables it to better manage capital requirements. The aim of the master's thesis is to show how to determine which indicators of the company largely influence the fact that the bank's customers will default on their loan and to obtain the formula to calculate the probability of default.

# Zahvala

Najprej bi se želela zahvaliti svojemu mentorju, izr. prof. dr. Mihaelu Permanu, za sprejeto mentorstvo in za strokovne nasvete ter usmerjanje pri izdelavi magistrskega dela.

Zahvala gre tudi moji družini za vso izkazano podporo, potrpežljivost, spodbudo in pomoč na vsakem življenjskem koraku.

Posebna zahvala pa gre mojemu partnerju Jerneju, ki mi je bil celotni čas pisanja magistrskega dela v oporo in mi je izkazoval veliko mero potrpežljivosti ter razumevanja.

# Kazalo vsebine

# Kazalo tabel

# Kazalo slik

# 1 Uvod in oznake

Statistika je veda, ki proučuje pojave, ki se kažejo v večjem številu, v določenem času in prostoru. S statističnim proučevanjem skušamo globlje razumeti množične pojave, odkrivati njihove zakonitosti in poskušamo napovedovati, kaj lahko pričakujemo.

Statistično proučevanje zakonitosti množičnih pojavov sestavljajo tri faze [11]:

- **zbiranje in urejanje podatkov**, ki opisujejo proučevani množični pojav;
- **analiza zbranih podatkov**, pri kateri uporabljamo statistične metode. Izbira metode je odvisna od namena analize in od vrste podatkov.
- **razlaga rezultatov**, kjer povemo, kaj smo o proučevanem množičnem pojavu izvedeli novega.

Vsaka od omenjenih faz je ključna za uspešnost statističnega proučevanja in mora biti narejena korektno, premišljeno in celovito.

Danes je statistika del mnogih naravoslovnih in družboslovnih ved, na primer za biološke in medicinske vede se uporablja biostatistika, za demografijo se uporablja demografska statistika, del ekonomije je ekonometrika in tako dalje.

Pri statističnem proučevanju množičnega pojava moramo najprej opredeliti *statistično populacijo*. To pomeni, da opredelimo koga ali kaj proučujemo, kje in kdaj. Populacijo sestavljajo *enote* in oznaka $N$ nam pove število enot v populaciji. Del populacije, katerega enote izbiramo z namenom, da ocenimo stanje v populaciji, imenujemo *vzorec*. Vzorcu, s katerim dobro povzamemo lastnosti celotne populacije, pravimo *reprezentativen vzorec*. Oznaka $n$ označuje število enot v vzorcu iz populacije, kjer je $n < N$.

Ena izmed glavnih stvari, ki nas zanimajo med proučevanjem pojavov, so lastnosti enot. Na primer, pri proučevanju uspešnosti študentov 1. letnika nas zanimajo naslednje lastnosti: leto rojstva, kraj rojstva, spol, srednja šola, smer študija in tako dalje. V statistiki opisuje posamezno lastnost enote *statistična spremenljivka*, ki ji bomo v nadaljevanju rekli kar spremenljivka. Natančneje pa si bomo pogledali, kaj so *kategorične spremenljivke*.

Kategorična spremenljivka ima za vrednosti kategorije. Na primer, pri določitvi spola novorojenčka imamo na voljo dve kategoriji: ženski spol in moški spol. Razvoj

metod, ki uporabljajo kategorične spremenljivke, se je začel z raziskavami v družboslovnih in biomedicinskih znanostih. V družboslovju se kategorične spremenljivke običajno uporabljajo za merjenje stališč in mnenj posameznika, v biomedicini pa za merjenje uspešnosti zdravljenja. Poznamo več vrst kategoričnih spremenljivk. V nadaljevanju bomo videli različne načine razvrščanja kategoričnih in tudi ostalih vrst spremenljivk. Statistične spremenljivke delimo na:

- **Odvisne/Neodvisne spremenljivke**

  Statistične analize najpogosteje razlikujejo med odvisnimi (ali odzivnimi) in neodvisnimi (ali pojasnjevalnimi) spremenljivkami. Odvisna spremenljivka je odvisna od drugih vplivov oziroma od neodvisnih spremenljivk. Gre za spremenljivko, ki jo lahko delno razložimo z eno ali več neodvisnimi spremenljivkami. Neodvisna spremenljivka pa je spremenljivka, ki se ne spreminja zaradi ostalih vplivov.

- **Nominalne/Ordinalne spremenljivke**

  Vrednosti kategoričnih spremenljivk so lahko razporejene v nekem zaporedju ali ne. Spremenljivkam, pri katerih kategorije niso razdeljene hierarhično temveč enakovredno, pravimo nominalne spremenljivke. Mogoče je določiti le, ali se dve spremenljivki med seboj razlikujeta, ne pa tudi katera izmed dveh je bolj pomembna. Primer nominalne spremenljivke je vrsta prevoznega sredstva (avtomobil, motor, kolo, avtobus ...).

  Poznamo pa tudi spremenljivke, katerih vrednosti lahko uredimo po določenem kriteriju. Takim spremenljivkam pravimo ordinalne in pri njih je možna razvrstitev vrednosti, ne pa tudi določitev razlike med njimi. Primer ordinalne spremenljivke je stanje pacienta (dobro, stabilno, resno, kritično).

  Spremenljivkam, pri katerih lahko primerjamo razliko med dvema vrednostma, pravimo *intervalne spremenljivke*. Intervalne spremenljivke imajo z vidika merske lestvice tudi boljše merske lastnosti kot ordinalne spremenljivke, saj poleg osnovnih izračunov in izračunov mediane omogočajo tudi izračune aritmetične sredine, standardnega odklona in standardne napake aritmetične sredine. Primeri take spremenljivke so nivo krvnega pritiska, teža, letni prihodek in podobno.

- **Zvezne/Diskretne spremenljivke**

  Spremenljivke razvrščamo med zvezne ali diskretne glede na število vrednosti, ki jih lahko zavzamejo. Zvezne spremenljivke zavzamejo vsako vrednost na nekem intervalu (na primer teža, višina, starost, čas, ...), diskretne spremenljivke pa lahko zavzamejo le določene končne, najpogosteje celoštevilčne vrednosti (na primer šolska ocena, število učencev, število prometnih nesreč, ...). Dejansko se

vse spremenljivke merijo na diskretni način zaradi omejitev natančnosti merilnih instrumentov. Zavedati se je potrebno, da ločnica med zveznimi in diskretnimi spremenljivkami ni vedno popolnoma jasna in v praksi pogosto razlikujemo med spremenljivkami, ki zavzamejo veliko vrednosti in spremenljivkami, ki zavzamejo le nekaj vrednosti. Statistiki pogosto obravnavajo diskretne intervalne spremenljivke, ki imajo veliko število vrednosti, kot zvezne in jih uporabljajo skupaj z metodami za zvezne odzive.

- **Kvalitativne/Kvantitativne spremenljivke**

  Spremenljivke ločujemo tudi glede na način izražanja njihove vrednosti. Kvalitativne spremenljivke (ali opisne spremenljivke) so spremenljivke, katerih vrednost je smiselno izraziti zgolj opisno, torej z besedami (na primer spol). Kvantitativne spremenljivke (ali numerične spremenljivke) pa so spremenljivke, katerih vrednost je smiselno izraziti s številom (na primer starost).

  Nominalne spremenljivke spadajo med kvalitativne, intervalne spremenljivke pa med kvantitativne. Razvrstitev ordinalnih spremenljivk glede na kvalitativno-kvantitativno klasifikacijo ni tako očitna. Analitiki jih pogosto obravnavajo kot kvalitativne z uporabo metod za nominalne spremenljivke, toda v mnogih pogledih so ordinalne spremenljivke bolj podobne intervalnim, saj imajo pomembne kvantitativne značilnosti: vsaka kategorija ima večji ali manjši vpliv kot druga kategorija in čeprav jih ni mogoče izmeriti, se v ozadju navadno skriva neka zvezna spremenljivka.

Verjetnostne poskuse in njihove izide opisujejo *slučajne spremenljivke*. Spremenljivka je slučajna spremenljivka, če nastopi kot rezultat poskusa, kjer je možnih več izidov in pri tem pojavitev katerekoli vrednosti iz danega območja predstavlja slučajno vrednost. Slučajna spremenljivka je določena z zalogo vrednosti in s porazdelitvenim zakonom, ki pove, kako je verjetnost porazdeljena po zalogi vrednosti. Zaloga vrednosti slučajne spremenljivke $X$ so vrednosti, ki jih $X$ zavzame in glede na zalogo vrednosti jih delimo na [11]:

- *diskretne* - imajo končno ali števno neskončno zalogo vrednosti;
- *nediskretne* - imajo neštevno zalogo vrednosti. Med njimi so najpomembnejše zvezne slučajne spremenljivke.

Pri statistiki je pomembno razumeti odnos med statistično spremenljivko in slučajno spremenljivko. Glede na vrsto sklepanja poznamo dve vrsti statistike, in sicer [13]:

- *opisno statistiko*, pri kateri se osredotočimo le na podatke, ki jih imamo ter poskusimo narediti smiseln povzetek;

- *inferenčno statistiko*, kjer gledamo podatke kot del nečesa večjega, česar ne poznamo v celoti. Na primer pri vzorcu, ki smo ga izbrali iz populacije, običajno vrednosti statistične spremenljivke na vzorcu poznamo, na celotni populaciji pa ne.

V splošnem gre pri inferenčni statistiki za to, da poznamo $X$, želeli pa bi statistično sklepati o $Y$. V teoriji $X$ in $Y$ predstavimo kot slučajni spremenljivki na istem verjetnostnem prostoru, ki pa nima nujno znane verjetnostne mere in temu pravimo *statistični model* [13]. Lahko rečemo, da statistični model opisuje, kako spremembe neodvisnih spremenljivk določajo spreminjanje odvisne spremenljivke.

Primer inferenčne statistike je tudi *regresijska analiza*, ki se ukvarja z napovedovanjem dogajanja v prihodnosti, na podlagi preteklih podatkov. Regresijske metode so postale glavna komponenta skoraj vsake analize podatkov, ki se ukvarja s povezavo med odzivno spremenljivko in eno ali večimi pojasnjevalnimi spremenljivkami. Pogosto je izid diskretna spremenljivka, ki zavzame dve ali več možnih vrednosti.

Magistrska naloga je iz matematično-finančnega področja in je zato v nadaljevanju razdeljena na dva dela. Prvi del predstavlja matematični vidik, v katerem so podrobneje predstavljeni modeli za kategorične odzive, s poudarkom na logistični regresiji. V drugem delu pa je najprej predstavljen finančni oziroma bančni vidik spopadanja s kreditnim tveganjem, nato pa še opis praktičnega primera razvoja modela za ocenjevanje verjetnosti nastanka neplačila v poslovni banki. Podatki v tem poglavju magistrskega dela so prirejeni in ne prikazujejo rezultatov modela v konkretni banki.

# 2 Modeli za kategorične odzive

Modeli za kategorične odzive sodijo med *posplošene linearne modele* (angl. generalized linear models - GLM). Ti modeli so razširjena oblika običajnih regresijskih modelov, ki zajemajo nenormalne porazdelitve odzivov in modeliranje funkcij povprečja. Posplošene linearne modele sestavljajo tri glavne komponente:

- **Slučajna komponenta** opredeli odzivno spremenljivko $Y$ in njeno verjetnostno porazdelitev, ki mora spadati v družino eksponentnih porazdelitev.
- **Sistematična komponenta** določi pojasnjevalne spremenljivke, ki jih uporabimo v linearni napovedni funkciji. Naj $x_{ij}$ predstavlja vrednosti $j$-te napovedne spremenljivke ($j = 1, 2, \ldots, p$) za subjekt $i$. Potem je linearna napovedna funkcija definirana kot
$$\eta_i = \sum_j \beta_j x_{ij}, \qquad i = 1, \ldots, N.$$
- **Povezovalna komponenta** definira funkcijo, ki povezuje slučajno komponento skupaj s sistematično komponento. Naj bo $\mu_i = E(Y_i), i = 1, ..., N$. Model povezuje $\mu_i$ in $\eta_i$ z enačbo $\eta_i = g(\mu_i)$, kjer je povezovalna funkcija $g$ monotona in odvedljiva funkcija. Tako $g$ povezuje $E(Y_i)$ s pojasnjevalnimi spremenljivkami preko formule
$$g(\mu_i) = \sum_j \beta_j x_{ij}, \qquad i = 1, ..., N.$$

Modelom za kategorične odzive, pri katerih zavzema odvisna spremenljivka dva izida, pravimo *modeli binarne izbire*. Izhodiščna metodološka pristopa pri izvedbi modelov binarne izbire sta *model logistične regresije* in *model probit*. Pri obeh metodoloških pristopih običajno ocenjujemo vpliv ene ali več pojasnjevalnih spremenljivk na eno odvisno spremenljivko z dvema izidoma.

Modelom za kategorične odzive, pri katerih zavzema odvisna spremenljivka več kot dva izida pravimo *modeli multiple izbire*. Pri teh modelih so izidi lahko razvrščeni na ordinalni merski lestvici ali pa niso razvrščeni. Tako kot pri modelih binarne izbire sta tudi pri izvedbi modelov multiple izbire izhodiščna metodološka pristopa model logistične regresije in model probit, od katerih sta najpogosteje v uporabi *multinomski logit model* in *multinomski model probit*.

V nadaljevanju magistrske naloge se bomo posvetili modelom binarne izbire, še posebej modelu logistične regresije. V zadnjem času je logistična regresija postala popularna tudi v poslovnem svetu, med drugim tudi za modeliranje verjetnosti poravnavanja obveznosti komitentov v banki. Za izračun verjetnosti, da bo komitent pravočasno poravnal dolg, potrebujemo kazalnike kot so višina zneska na računu, letni prihodek, poklic, kreditne obveznosti, odstotek pravočasno poravnanih računov in ostale podobne lastnosti komitentove plačilne zgodovine.

Preden se bolj poglobimo v logistično regresijo je pomembno razumeti, da je cilj uporabe te metode enak kot pri ostalih tehnikah za izgradnjo modelov v statistiki: želimo najti model, ki se bo najbolje prilegal podatkom in ne bo vseboval preveč parametrov za opis zveze med odvisno spremenljivko in množico neodvisnih spremenljivk. Najpogostejši primer modeliranja je z linearno regresijo, kjer domnevamo, da je odzivna spremenljivka zvezna. Kar razlikuje model logistične regresije od linearne regresije je ravno odzivna spremenljivka, ki je v primeru logistične regresije običajno *binarna*.

V nadaljevanju magistrske naloge si bomo podrobneje ogledali model logistične regresije. Kot glavni gradivi za teoretični del tega poglavja magistrske naloge sta bili uporabljeni [1] in [7].

## 2.1 Model logistične regresije

Naj $Y$ predstavlja binarno odzivno spremenljivko. Vsako opažanje ima enega od dveh možnih odzivov, označenega z 0 ali 1. Povprečje zapišemo kot $E(Y) = P(Y = 1)$, pri čemer $P(Y = 1)$ označimo s $\pi(\mathrm{x})$, kar predstavlja odvisnost od vrednosti napovednih spremenljivk, ki jih zapišemo kot $\mathrm{x} = (x_1, ..., x_p)$. Varianca za $Y$ je enaka

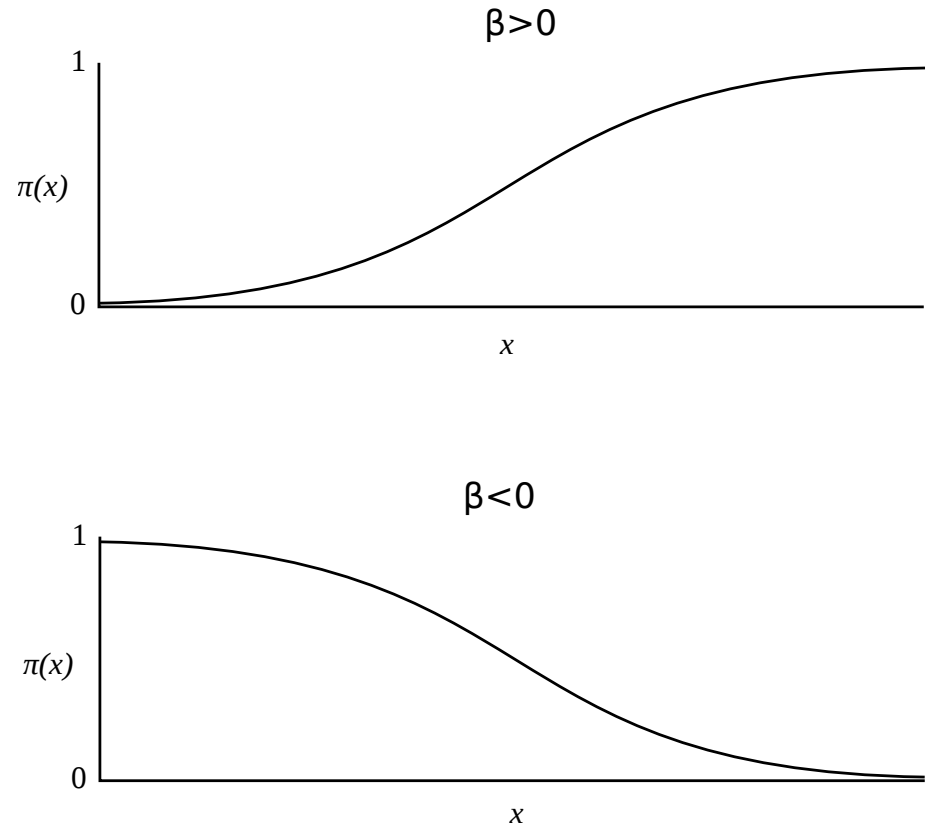
$$\mathrm{var}(Y) = \pi(\mathrm{x})[1 - \pi(\mathrm{x})].$$

Običajno binarni odzivi izhajajo iz nelinearnega odnosa med $\pi(x)$ in $x$. Fiksna sprememba v $x$ ima običajno manj vpliva, ko je $\pi(x)$ blizu 0 ali 1 kot pa, ko je $\pi(x)$ v bližini $\frac{1}{2}$. V praksi je nelinearna relacija med $\pi(x)$ in $x$ običajno monotona, kjer $\pi(x)$ zvezno narašča ali zvezno pada, ko $x$ narašča. Na sliki 1 lahko vidimo tipično S-obliko krivulje za funkcijo logistične regresije.

*Logistični regresijski model* z eno pojasnjevalno spremenljivko zapišemo kot

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}. \tag{2.1}$$

Predznak parametra $\beta$ določa ali je $\pi(x)$ naraščajoča ali padajoča, ko $x \to \infty$. Ko je $\beta < 0$, potem $\pi(x) \downarrow 0$, ko pa je $\beta > 0$ potem $\pi(x) \uparrow 1$, kar lahko vidimo na sliki 1. Hitrost naraščanja oziroma padanja se povečuje, ko $|\beta|$ narašča. Ko $\beta \to 0$, se




Slika 1: Funkcija logistične regresije.

krivulja izravna v vodoravno linijo. Ko je $\beta = 0$, je $Y$ neodvisen od $x$. Ko je $\beta > 0$, ima krivulja $\pi(x)$ obliko porazdelitvene funkcije logistične porazdelitve. Ker je gostota logistične regresije simetrična, se $\pi(x)$ približuje 1 z enako hitrostjo kot se približuje 0.

Poiščimo še povezovalno funkcijo, za katero je logistična regresija posplošeni linearni model. Razmerju med verjetnostjo, da bo prišlo do dogodka, ki ga obravnavamo (na primer neplačilo računa) in verjetnostjo, da do tega dogodka ne bo prišlo pravimo *obeti* (angl. odds). Iz enačbe (2.1) izrazimo *obete* kot

$$\frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta x}. \tag{2.2}$$

Za *logaritem obetov* (angl. log odds), ki ga imenujemo tudi *logit*, velja naslednja linearna zveza

$$logit[\pi(x)] = \ln \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x. \tag{2.3}$$

Modeli logistične regresije torej spadajo med *posplošene linearne modele* z binomsko slučajno komponento in povezovalno funkcijo *logit*. Modelom logistične regresije pravimo tudi *logit modeli*.

### 2.1.1 Multipla logistična regresija

Model logistične regresije lahko vsebuje več pojasnjevalnih spremenljivk. Recimo, da imamo na voljo $p$ neodvisnih spremenljivk predstavljenih z vektorjem $\mathrm{x} = (x_1, x_2, ..., x_p)$. Naj bo pogojna verjetnost, da je izid prisoten, označena kot $P(Y = 1|\mathrm{x}) = \pi(\mathrm{x})$. Logit za model multiple logistične regresije zapišemo kot

$$logit[\pi(\mathrm{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p. \tag{2.4}$$

Alternativna formula, ki izraža neposredno model logistične regresije, pa je

$$\pi(\mathrm{x}) = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2 \cdots+\beta_p x_p}}{1 + e^{\alpha+\beta_1 x_1+\beta_2 x_2 \cdots+\beta_p x_p}}. \tag{2.5}$$

Parameter $\beta_i$ izraža učinek $x_i$ na logaritem obetov pri $Y = 1$, ko so ostali $x_j$ nadzorovani. Na primer, $e^{\beta_i}$ predstavlja multiplikativni učinek na obete, pri povečanju $x_i$ za eno enoto, ko so ostali $x_j$ fiksni.

Če so katere izmed neodvisnih spremenljivk nominalne spremenljivke, kot na primer spol, rasa, poklic in podobno, je neprimerno, da jih v model vključimo in obravnavamo kot intervalne spremenljivke. Števila, ki predstavljajo različne kategorije teh nominalnih spremenljivk, so le identifikatorji in nimajo nobenega numeričnega pomena. V takem primeru uporabimo metodo, s katero ustvarimo zbirko *slamnatih spremenljivk* (angl. dummy variables). Predpostavimo, da imamo na razpolago neodvisno spremenljivko, ki predstavlja raso in za katero imamo na voljo vrednosti: »bela rasa«, »črna rasa« in »ostale rase«. V tem primeru sta potrebni dve slamnati spremenljivki. Ena možna strategija označevanja je, da bi v primeru, ko posameznik spada med »belo raso«, imeli dve slamnati spremenljivki, $D_1$ in $D_2$ z vrednostjo 0; ko posameznik spada med »črno raso«, bi $D_1$ bila enaka 1 in $D_2$ enaka 0; ko pa bi posameznik spadal med »ostale rase« bi uporabili $D_1 = 0$ in $D_2 = 1$. V tabeli 1 je prikazano označevanje slamnate spremenljivke za neodvisno spremenljivko, ki predstavlja raso.

Tabela 1: Primer označevanja slamnate spremenljivke, ki predstavlja raso.

| | Slamnata spremenljivka | |
|---|---|---|
| RASA | $D_1$ | $D_2$ |
| Bela rasa | 0 | 0 |
| Črna rasa | 1 | 0 |
| Ostale rase | 0 | 1 |

Če ima nominalna spremenljivka $k$ možnih vrednosti, potem v splošnem velja, da potrebujemo $k - 1$ slamnatih spremenljivk. To velja, ker imajo vsi modeli po eno kon-

stanto, če ni določeno drugače. Predpostavimo, da ima $j$-ta neodvisna spremenljivka $x_j$, $k$ kategorij. Potem bo $k-1$ slamnatih spremenljivk označenih kot $D_{jl}$ in koeficienti za te slamnate spremenljivke bodo enaki $\beta_{jl}$, kjer je $l = 1, 2, ..., k-1$. Iz tega sledi, da logit za model s $p$ spremenljivkami, med katerimi je $j$-ta spremenljivka nominalna, zapišemo kot

$$logit[\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \cdots + \sum_{l=1}^{k-1} \beta_{jl} D_{jl} + \beta_p x_p.$$

## 2.2 Probit model

Monotona regresijska funkcija kot je prva na sliki 1, ima obliko kumulativne porazdelitvene funkcije za zvezno slučajno spremenljivko. Iz tega izhaja model za binarni odziv z obliko $\pi(x) = F(x)$ za neko kumulativno porazdelitveno funkcijo $F$.

Z uporabo celotnega razreda kumulativnih porazdelitvenih funkcij, na primer normalne porazdelitvene funkcije z njenimi raznovrstnimi povprečji in variancami, dovolimo krivulji $\pi(x) = F(x)$ prožnost glede na stopnjo naraščanja in glede na območje, kjer se zgodi največ naraščanja. Naj $\Phi(\cdot)$ predstavlja standardno kumulativno porazdelitveno funkcijo razreda kot je $N(0, 1)$. Z uporabo $\Phi$ in zapisom modela kot

$$\pi(x) = \Phi(\alpha + \beta x) \tag{2.6}$$

zagotovimo enako prožnost. Oblike različnih kumulativnih porazdelitvenih funkcij v razredu se zgodijo, ko se $\alpha$ in $\beta$ spreminjata. Če zamenjamo $x$ z $\beta x$, s tem dovolimo krivulji, da narašča z različno stopnjo kot standardna kumulativna porazdelitvena funkcija (ali celo pada, če je $\beta < 0$). Sprememba parametra $\alpha$ pa vpliva na premik krivulje levo ali desno.

Ko je $\Phi$ strogo naraščajoča čez celotno realno premico, potem obstaja njena inverzna funkcija $\Phi^{-1}(\cdot)$ in iz enačbe (2.6) dobimo

$$\Phi^{-1}[\pi(x)] = \alpha + \beta x. \tag{2.7}$$

Za ta razred kumulativnih porazdelitvenih funkcij, je povezovalna komponenta za splošni linearni model enaka $\Phi^{-1}$. Povezovalna funkcija preslika območje verjetnosti $(0, 1)$ na $(-\infty, \infty)$, torej območje linearnih prediktorjev. Krivulja ima obliko normalne porazdelitve, ko je $\Phi$ standardna normalna porazdelitvena funkcija. Modelu iz enačbe (2.7) pravimo *probit model*. Ta krivulja ima podoben izgled kot krivulja logistične regresije.

Razlika med logit in probit modelom je v tem, da prvi predpostavlja logistično, drugi pa normalno kumulativno porazdelitev. Ali bomo izbrali logistični model ali model probit je odvisno predvsem od računalniške programske opreme, ki nam je na voljo.

Ob predpostavki, da je vzorec preučevanja dovolj velik, naj se rezultati obeh metodoloških pristopov ne bi bistveno razlikovali in enako velja tudi za postopek ocenjevanja parametrov izbranih modelov z metodo največjega verjetja [10].

## 2.3 Ocenjevanje parametrov modela logistične regresije

Predpostavimo, da imamo vzorec $n$ neodvisnih opazovanih parov $(x_i, y_i), i = 1, 2, ..., n$, kjer $y_i$ predstavlja vrednost binarne odzivne spremenljivke in $x_i$ vrednost neodvisne spremenljivke za $i$-ti subjekt. Predpostavimo še, da so možni izidi odzivne spremenljivke označeni kot 0 ali 1 in predstavljajo odsotnost ali prisotnost karakteristike, ki jo preučujemo. Da se bo model logistične regresije iz enačbe (2.1) dobro prilegal množici podatkov, moramo oceniti vrednosti $\alpha$ in $\beta$, ki sta neznana parametra.

Pri linearni regresiji, je najpogosteje uporabljena metoda za ocenjevanje neznanih parametrov *metoda najmanjših kvadratov*. Pri tej metodi izberemo za $\alpha$ in $\beta$ tiste vrednosti, ki minimizirajo vsoto kvadratov odstopanj opazovanih vrednosti $Y$ od napovedanih vrednosti modela. Glede na običajne predpostavke linearne regresije, nam metoda najmanjših kvadratov ponudi ocene s številnimi želenimi statističnimi lastnostmi. Na žalost pa nam metoda najmanjših kvadratov, če jo apliciramo na model z binarnim odzivom, ne da ocenjenih vrednosti z enakimi lastnostmi.

Za ocenjevanje parametrov logistične regresije se zato uporablja *metoda največjega verjetja*. V splošnem poda metoda največjega verjetja vrednosti za neznane parametre, ki maksimirajo *funkcijo verjetja*. Parametri ocenjeni z metodo največjega verjetja, so takšne končne ocene, ki najbolj sovpadajo z opazovanimi podatki. Vrednosti parametrov, ki maksimirajo funkcijo verjetja, maksimirajo tudi njen logaritem. Lažje je maksimirati logaritem verjetja, saj imamo vsoto namesto produkta izrazov.

Poglejmo si, kako označujemo predstavljene vrednosti in funkcije. Parameter, ki ga obravnavamo označimo z $\beta$, njegov ocenjen maksimum verjetja pa z $\hat{\beta}$. Funkcija verjetja je $l(\beta)$ in logaritem verjetja je $L(\beta) = \ln[l(\beta)]$. Za večino modelov ima $L(\beta)$ konkavno obliko in $\hat{\beta}$ je točka, v kateri je odvod logaritma verjetja enak 0. Ocena maksimuma verjetja je torej rešitev enačbe verjetja $\partial L(\beta)/\partial\beta = 0$. Pogosto je $\beta$ večdimenzionalna in jo pišemo kot $\boldsymbol{\beta}$, z $\hat{\boldsymbol{\beta}}$ pa označujemo rešitev množice enačb verjetja.

Naj SE označuje standardno napako za $\hat{\beta}$ in naj $\text{cov}(\hat{\boldsymbol{\beta}})$ označuje asimptotično kovariančno matriko ocen $\hat{\boldsymbol{\beta}}$. Pod ustrezno določenimi pogoji je $\text{cov}(\hat{\boldsymbol{\beta}})$ inverz *informacijske matrike*. Element $(j, k)$ v informacijski matriki je enak

$$-E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}\right).$$

Standardne napake so kvadratni koreni diagonalnih elementov inverza informacijske matrike. Večja kot je ukrivljenost logaritma verjetja, manjše so standardne napake. To je smiselno, saj večja ukrivljenost pomeni, da logaritem verjetja hitro pada, ko se $\boldsymbol{\beta}$ odmika od $\hat{\boldsymbol{\beta}}$. Zato obstaja veliko večja verjetnost, da se bodo podatki pojavili, če je $\boldsymbol{\beta}$ blizu $\hat{\boldsymbol{\beta}}$ kot pa, če je oddaljen od $\hat{\boldsymbol{\beta}}$.

Poglejmo si, kako dobimo ocenjene vrednosti parametrov za model logistične regresije. Če ima $y$ možna izida 0 in 1, potem nam izraz za $\pi(x)$ iz enačbe (2.1) da pogojno verjetnost, da je $Y$ enak 1 ob danem $x$, kar zapišemo kot $P(Y = 1|x)$. Iz tega sledi, da nam $1 - \pi(x)$ da pogojno verjetnost, da je $Y$ enak 0 ob danem $x$, kar zapišemo kot $P(Y = 0|x)$. Torej je za tiste pare $(x_i, y_i)$, kjer je $y_i = 1$, prispevek k funkciji verjetja enak $\pi(x_i)$ in za tiste pare, kjer je $y_i = 0$, je prispevek k funkciji verjetja enak $1-\pi(x_i)$. Prispevek k funkciji verjetja za par $(x_i, y_i)$ najlažje izrazimo kot

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{2.8}$$

Ker za opažanja predpostavimo, da so neodvisna, dobimo funkcijo verjetja kot produkt izrazov iz (2.8):

$$l(\alpha, \beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{2.9}$$

Princip največjega verjetja navaja, da za ocene parametrov $\alpha$ in $\beta$ uporabimo tiste vrednosti, ki maksimirajo izraz iz enačbe (2.9). Ker je matematično lažje operirati z logaritmom enačbe, najprej definirajmo *logaritem verjetja* kot:

$$L(\alpha, \beta) = \ln[l(\alpha, \beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}. \tag{2.10}$$

Da najdemo vrednosti za $\alpha$ in $\beta$, ki maksimirata $L(\alpha, \beta)$, moramo $L(\alpha, \beta)$ odvajati glede na $\alpha$ in $\beta$ ter dobljeni izraz enačiti z nič. Dobimo naslednji dve enačbi, poznani kot *enačbi verjetja*:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{2.11}$$

in

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0. \tag{2.12}$$

Pri logistični regresiji sta izraza v enačbah (2.11) in (2.12) nelinearna v $\alpha$ in $\beta$, zato potrebujemo posebne iterativne metode za njihovo rešitev. Vrednost $B = (\alpha, \beta)$, dana kot rešitev enačb (2.11) in (2.12), se imenuje *ocena maksimuma verjetja* in jo označimo kot $\hat{B}$. Zanimiva posledica enačbe (2.11) je, da velja

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i),$$

kjer je $\hat{\pi}(x_i)$ ocena največjega verjetja za $\pi(x_i)$. To pomeni, da je vsota opazovanih vrednosti $y$ enaka vsoti napovedanih (pričakovanih) vrednosti.

## 2.4 Testiranje statistične značilnosti koeficientov logistične regresije

Po zaključenem ocenjevanju koeficientov nas zanima, kako statistično značilne so spremenljivke modela. Ocena pomembnosti spremenljivk običajno vključuje oblikovanje in testiranje statistične hipoteze, s čimer preverimo ali so neodvisne spremenljivke v modelu značilno povezane z odzivno spremenljivko.

### 2.4.1 Statistični testi

Glavni princip za testiranje statistične značilnosti je, da primerjamo opazovane vrednosti odzivne spremenljivke z napovedanimi vrednostmi pridobljenimi iz modela, ki vsebuje in modela, ki ne vsebuje testirane spremenljivke. Pri logistični regresiji temelji primerjava med opazovanimi in napovedanimi vrednostmi na logaritmu funkcije verjetja, ki smo ga definirali v enačbi (2.10). Naj bo $L_1$ logaritem verjetja *nasičenega modela* (angl. saturated model), to je modela, ki vsebuje toliko parametrov kot je podatkovnih točk in naj bo $L_0$ logaritem verjetja *nameščenega modela* (angl. fitted model), to je modela z omejenim številom parametrom. Nameščeni model ima praviloma manj parametrov kot nasičen model. Testno statistiko zapišemo kot:

$$D = -2 \ln \left[ \frac{L_0}{L_1} \right]. \tag{2.13}$$

Model z več parametri se bo vedno prilegal podatkom vsaj tako dobro kot nameščen model. Vrednost znotraj oglatih oklepajev imenujemo *razmerje verjetij*, testu pa pravimo *test razmerja verjetij* (angl. likelihood ratio test). Razmerje verjetij bo vedno med 0 in 1. Manjša kot je vrednost testa, slabše je prileganje nameščenega modela. Z uporabo enačbe (2.10), postane enačba (2.13) enaka

$$D = -2 \sum_{i=1}^{n} \left[ y_i \ln(\frac{\hat{\pi}_i}{y_i}) + (1 - y_i) \ln(\frac{1 - \hat{\pi}_i}{1 - y_i}) \right], \tag{2.14}$$

kjer je $\hat{\pi}_i = \hat{\pi}(x_i)$.

Statistiki D iz enačbe (2.14) pravimo tudi *odklonskost* (angl. deviance) in ima isto vlogo kot vsota kvadratov ostankov (SSE) pri linearni regresiji. Pravzaprav je

odklonskost prikazana v enačbi (2.14), če jo izračunamo za linearno regresijo, identično enaka SSE.

Ko sta vrednosti binarne odzivne spremenljivke enaki 0 ali 1, je verjetje nasičenega modela enako 1. To sledi iz definicije nasičenega modela, da je $\hat{\pi}_i = y_i$ in verjetje zapišemo kot

$$l(\text{nasičen model}) = \sum_{i=1}^{n} y_i^{y_i} \times (1 - y_i)^{(1-y_i)} = 1.$$

Iz enačbe 2.13 nato sledi, da je odklonskost enaka

$$D = -2\ln(\text{verjetje nasičenega modela}). \tag{2.15}$$

Da bi ocenili pomembnost neodvisne spremenljivke, primerjamo vrednost $D$ za model z vključeno neodvisno spremenljivko in za model brez nje. Sprememba v vrednosti $D$, zaradi vključitve neodvisne spremenljivke v model, je izračunana kot:

$$G = D(\text{model brez spremenljivke}) - D(\text{model s spremenljivko}).$$

Ker je vrednost verjetja nasičenega modela skupna obema vrednostma $D$ iz zgornje enačbe, lahko $G$ izrazimo kot:

$$G = -2\ln\left[\frac{D(\text{model brez spremenljivke})}{D(\text{model s spremenljivko})}\right]. \tag{2.16}$$

Za poseben primer, ko imamo samo eno neodvisno spremenljivko, lahko enostavno pokažemo, da je v primeru, ko spremenljivke ni v modelu, ocena maksimuma verjetja za $\beta_0$ enaka $\ln(n_1/n_0)$, kjer je $n_1 = \sum y_i$ in $n_0 = \sum(1 - y_i)$ ter napovedana vrednost enaka konstanti $n_1/n$. V tem primeru, je $G$ vrednost enaka:

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{(1-y_i)}}\right] \tag{2.17}$$

ali

$$G = 2\{\sum_{i=1}^{n}[y_i\ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)] - [n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)]\}. \tag{2.18}$$

Izračun logaritma verjetja in test razmerja verjetij sta običajni funkciji vsake programske opreme za logistično regresijo. To omogoča enostavno preverjanje statistične značilnosti dodanih novih spremenljivk v model. V primeru, ko imamo samo eno neodvisno spremenljivko, najprej prilagajamo model, ki vsebuje le konstanto. Nato prilagajamo model, ki poleg konstante vsebuje še neodvisno spremenljivko. To zviša vrednost logaritma verjetja. Test razmerja verjetij dobimo tako, da pomnožimo razliko teh dveh vrednosti z $-2$.

Poznamo še dva podobna, statistično enakovredna testa. Imenujemo ju *Waldov test* in *test izida*. Predpostavke, ki so potrebne za ta dva testa so enake, kot tiste za test razmerja verjetij v enačbi (2.17).

Waldov test primerja oceno maksimuma verjetja naklonskega parametra $\hat{\beta}_1$ z oceno njegove standardne napake in ga zapišemo kot

$$W = \hat{\beta}_1 / \widehat{SE}(\hat{\beta}_1).$$

Končni kvocient, pod hipotezo $\beta_1 = 0$, aproksimativno sledi standardni normalni porazdelitvi.

Tako test razmerja verjetij ($G$) kot Waldow test ($W$) zahtevata izračun ocene maksimuma verjetja za $\beta_1$. Test statistične značilnosti spremenljivke, ki ne zahteva teh izračunov, se imenuje *test izida* (angl. score test). Ta test je osnovan na porazdelitveni teoriji odvodov logaritma verjetja. V univariatnem primeru, ta test temelji na pogojni porazdelitvi odvoda enačbe (2.17), ob danem odvodu enačbe (2.16). Test uporablja vrednost enačbe (2.17) pri $\beta_0 = \ln(n_1/n_0)$ in $\beta_1 = 0$. Ob teh vrednostih parametrov velja $\hat{\pi} = n_1/n = \bar{y}$. Tako postane leva stran enačbe (2.17) enaka $\sum x_i(y_i - \bar{y})$. Ocenjena varianca je potem $\bar{y}(1-\bar{y})\sum(x_i - \bar{x})^2$. Testna statistika za test izida (ST) je

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

Za velike vzorce nam ti trije testi običajno dajo podobne rezultate. Test razmerja verjetij je v praksi bolj zaželen za uporabo od Waldovega, saj uporabi več informacij. Ko je $|\beta|$ relativno velik, je Waldov test manj zanesljiv kot test razmerja verjetij in lahko pokaže zmotne rezultate.

### 2.4.2 Ocenjevanje intervalov zaupanja

Interval zaupanja za ocenjevanje parametrov je interval, v katerem lahko z veliko verjetnostjo trdimo, da se znotraj njega nahaja dejanska vrednost parametra. Spodnjo in zgornjo mejo intervala izračunamo iz ocenjene vrednosti parametra iz danega vzorca populacije ob predpostavki znane porazdelitve ocenjenih vrednosti iz vzorca. Osnova za določitev ocen intervala je enaka statistična teorija, ki smo jo uporabili za prej omenjene teste statistične značilnosti parametrov modela.

Naj $z_{1-a}$ predstavlja oceno $z$-testa s standardno normalno porazdelitvijo in stopnjo značilnosti $a$, pri kateri bomo ničelno hipotezo zavrnili (to predstavlja $100(1-a)\%$ percentil te porazdelitve). Ocene intervala zaupanja za naklon ($\beta_1$) in odsek ($\beta_0$) temeljijo na Waldovem testu. Mejne točke $100(1-a)\%$ intervala zaupanja za koeficient naklona so:

$$\hat{\beta}_1 \pm z_{1-a/2}\widehat{SE}(\hat{\beta}_1), \tag{2.19}$$

za odsek pa:

$$\hat{\beta}_0 \pm z_{1-a/2}\widehat{SE}(\hat{\beta}_0), \tag{2.20}$$

kjer je $z_{1-a/2}$ zgornja $100(1-a/2)\%$ točka iz standardne normale distribucije in $\hat{SE}(\cdot)$ predstavlja oceno standardne napake modela za pripadajočo oceno parametra.

Logit je linearni del modela logistične regresije in njegovo cenilko zapišemo kot:

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{2.21}$$

Cenilka variance cenilke logita iz zgornje enačbe (2.21) zahteva, da pridobimo varianco vsote:

$$\widehat{\mathrm{Var}}(\hat{g}(x)) = \widehat{\mathrm{Var}}(\hat{\beta}_0) + x^2\widehat{\mathrm{Var}}(\hat{\beta}_1) + 2x\widehat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_1). \tag{2.22}$$

Mejne točke $100(1-a)\%$ Waldovega intervala zaupanja za logit so:

$$\hat{g}(x) \pm z_{1-a/2}\widehat{SE}[\hat{g}(x)], \tag{2.23}$$

kjer je $\widehat{SE}[\hat{g}(x)]$ pozitiven kvadratni koren cenilke variance iz enačbe (2.22).

## 2.5 Interpretacija parametrov logistične regresije

Po tem, ko smo izbrali končni model in ocenili pomembnost ocenjenih koeficientov, nas zanima še interpretacija njihovih vrednosti. Ko imamo izbran končni model, se nam običajno pojavi vprašanje, kaj nam ocenjeni parametri v modelu povejo o problemu, zaradi katerega delamo raziskavo. Ocenjeni koeficienti za neodvisne spremenljivke predstavljajo naklon (tj. stopnjo spremembe) funkcije odvisne spremenljivke v eni enoti spremembe neodvisne spremenljivke. Interpretacija se sooča z dvema problemoma: določanjem funkcionalnega odnosa med odvisno in neodvisno spremenljivko ter primernim definiranjem enote spremembe pri neodvisni spremenljivki.

Zapišimo najprej povezovalno funkcijo logit iz enačbe (2.3) kot

$$g(x) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x.$$

Za model linearne regresije je koeficient naklona, $\beta_1$, enak razliki med vrednostjo odvisne spremenljivke v $x+1$ in vrednostjo odvisne spremenljivke v $x$, za vsak $x$. Na primer, če imamo $y(x) = \beta_0 + \beta_1 x$, iz tega sledi

$$\beta_1 = y(x+1) - y(x).$$

V tem primeru je interpretacija koeficienta $\beta_1$ precej enostavna in izraža spremembo vrednosti odvisne spremenljivke, ko se neodvisna spremenljivka spremeni za eno enoto. Če, na primer, z regresijo obravnavamo vpliv teže glede na višino mladoletnih moških

in je naklon enak 5, potem bi iz tega sklepali, da je 1 centimeter dodatne višine povezan s povečanjem teže za 5 kilogramov.

Pri logistični regresiji nam koeficient naklona predstavlja spremembo v logitu glede na spremembo neodvisne spremenljivke za eno enoto. To zapišemo kot

$$\beta_1 = g(x+1) - g(x).$$

Interpretacija koeficienta $\beta_1$ v modelu logistične regresije, je odvisna od pomena razlike med dvema logit funkcijama. V naslednjem podpoglavju si bomo najprej pogledali interpretacijo pomena koeficientov na primeru binarne neodvisne spremenljivke, ki je osnova za ostale vrste spremenljivk, nato pa še na primeru zvezne neodvisne spremenljivke.

### 2.5.1 Binarna neodvisna spremenljivka

Nominalnim neodvisnim spremenljivkam, ki imajo samo dve kategoriji, pravimo binarne in predstavljajo osnovo za interpretiranje koeficientov logistične regresije.

Predpostavimo, da neodvisna spremenljivka $x$ zavzame vrednosti 0 ali 1. Razlika v logitu pri $x = 1$ in $x = 0$ je enaka

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1.$$

Kot lahko vidimo je razlika enaka $\beta_1$. Da bomo lahko interpretirali ta rezultat, moramo pogledati razmerje obetov. Obete za izid, ki je prisoten med subjekti z $x = 1$, definiramo kot $\pi(1)/[1-\pi(1)]$. Podobno so obeti za izid, ki je prisoten med posamezniki z $x = 0$, definirani kot $\pi(0)/[1-\pi(0)]$. Razmerje obetov (OR) je definirano kot razmerje obetov pri $x = 1$ glede na obete pri $x = 0$, in ga zapišemo kot

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}. \tag{2.24}$$

Če enačbo (2.24) zapišemo z enačbo modela logistične regresije, dobimo

$$\begin{aligned} OR &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) / \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) / \left(\frac{1}{1+e^{\beta_0}}\right)} \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\ &= e^{(\beta_0+\beta_1)-\beta_0} \\ &= e^{\beta_1}. \end{aligned}$$

Torej, za logistično regresijo z binarno neodvisno spremenljivko, katere vrednosti sta enaki 0 ali 1, je zveza med razmerjem obetov in regresijskim koeficientom izražena kot

$$OR = e^{\beta_1}. \tag{2.25}$$

Ta enostavna zveza med koeficientom in razmerjem obetom je osnovni razlog, zakaj je logistična regresija tako uporabno orodje za analitične raziskave.

Razmerje obetov je mera, ki ima široko uporabnost, sploh v epidemiologiji, saj pove, koliko bolj (ali manj) bo preučevan izid prisoten pri tistih, ki imajo $x = 1$ kot pa pri tistih z $x = 0$. Na primer, če $y$ predstavlja prisotnost ali odsotnost raka na pljučih in $x$ predstavlja lastnost, da je oseba kadilec, potem nam $\hat{OR} = 2$ pove, da je dvakrat več možnosti, da bo pljučni rak prisoten med kadilci kot nekadilci v populaciji raziskave.

### 2.5.2 Zvezna neodvisna spremenljivka

Ko model logistične regresije vsebuje zvezno neodvisno spremenljivko, je interpretacija ocenjenih koeficientov odvisna od tega, kako je definirana enota spremenljivke ter kako je dodana v model. Za namen razvoja metode za interpretacijo koeficientov zvezne spremenljivke predpostavimo, da je logit spremenljivke linearen. Pod to predpostavko je enačba za logit enaka $g(x) = \beta_0 + \beta_1 x$. Iz tega sledi, da nam koeficient naklona, $\beta_1$, pove spremembo v logaritmu obetov, ko pride do povečanja za eno enoto v $x$, kar pomeni, da velja $\beta_1 = g(x + 1) - g(x)$, za vsak $x$.

Za pridobitev uporabne interpretacije koeficientov zveznih spremenljivk moramo razviti metodo za oceno točke in intervala za poljubno spremembo za »$c$« enot pojasnjevalne spremenljivke. Na primer, eno leto višja starost ali povečanje sistoličnega krvnega tlaka za $1\,mmHg$, je lahko premajhna sprememba, da bi jo šteli za pomembno, medtem ko se razlika za 10 let ali $10\,mmHg$ lahko izkaže za bolj uporabno. Po drugi strani, če $x$ zavzema vrednosti v razponu od 0 do 1, potem je sprememba za 1 prevelika in bo sprememba za $0,01$ mogoče bolj primerna. Logaritem obetov za spremembo za $c$ enot v $x$ dobimo iz razlike logita kot $g(x + c) - g(x) = c\beta_1$. Pripadajoče razmerje obetov pa dobimo, če uporabimo eksponentno funkcijo na razliki logita, in sicer kot

$$OR(c) = OR(x + c, x) = \exp(c\beta_1).$$

Oceno sedaj pridobimo s tem, da zamenjamo $\beta_1$ z njegovim ocenjenim maksimalnim verjetjem $\hat{\beta}_1$. Oceno za standardno napako, ki jo potrebujemo za interval zaupanja dobimo tako, da pomnožimo ocenjeno standardno napako z vrednostjo $c$. Tako so mejne točke $100(1 - a)\%$ intervala zaupanja za $OR(c)$ enake

$$\exp[c\hat{\beta}_1 \pm z_{1-a/2} c\, \hat{SE}(\hat{\beta}_1)].$$

Ker sta tako ocena enote kot mejne točke intervala zaupanja odvisni od izbire vrednosti $c$, mora biti ta še posebej jasno navedena v vseh izračunih in tabelah. To lahko včasih povzroča težave, saj je izbira vrednosti $c$ dokaj poljubne narave. Na primer, zakaj gledamo spremembo 10-ih let, ko bi lahko 5, 15 ali celo 20 let razlike

dalo isto dobre rezultate? Seveda lahko uporabimo katerokoli razumno število, vendar moramo imeti v mislih, da bralcu naših analiz jasno nakažemo, kako se tveganje, da je preiskovana lastnost prisotna, spreminja glede na spremenljivko, ki nam je dana. Zato so navadno spremembe večkratnikov števila 5 ali 10 najbolj smiselne in najlažje razumljive.

## 2.6 Izdelava modelov logistične regresije

Ko znamo interpretirati modele logistične regresije in znamo oceniti njihove parametre, si lahko pogledamo, kako začeti izdelavo takega modela. V realnih primerih smo pogosto v situaciji, ko imamo veliko neodvisnih spremenljivk, ki bi lahko bile vključene v model in zato je potrebno izdelati posebno strategijo za ravnanje s takšnimi kompleksnimi sistemi. Za uspešno modeliranje kompleksnih podatkov je delno potrebna znanost, delno statistične metode, delno izkušnje in pa zdrav razum. V nadaljevanju si bomo pogledali, kaj vse potrebujemo za izgradnjo modela logistične regresije.

### 2.6.1 Strategije za izbiro modela

Cilj vsake metode je izbira takšnih spremenljivk, ki nam dajo najboljši možen model v znanstvenem smislu problema. Da dosežemo takšen cilj, moramo imeti osnovni načrt za izbiro spremenljivk modela in pa nabor metod za ocenjevanje ustreznosti modela, tako v smislu ocenjevanja njegovih posameznih spremenljivk kot celotno prileganje podatkom. Izbira modela je težja, če imamo večje število pojasnjevalnih spremenljivk, zaradi povečanja možnih učinkov in interakcije. Model mora zadostovati dvema glavnima pogojema, in sicer [7]:

- biti mora dovolj kompleksen, da se dobro prilega podatkom in
- biti mora enostaven za interpretacijo.

Obstaja veliko postopkov izbire modela, vendar za nobenega ne moremo trditi, da bo vedno najboljši. Poznamo pa določene ukrepe, ki jih lahko naredimo, da izboljšamo model. Na primer, če imamo model z več pojasnjevalnimi spremenljivkami, imamo lahko probleme s preveliko korelacijo med njimi, kar pomeni, da določena spremenljivka nima dobre napovedne moči, ko so prisotne še ostale spremenljivke v modelu in v tem primeru, je odstranitev take spremenljivke lahko koristna, saj s tem zmanjšamo standardno napako ter ostale učinke, ki jih ocenjujemo.

### 2.6.2 Izbira spremenljivk

Kriterij za vključitev spremenljivk v model se med različnimi problemi lahko precej razlikuje. Enostavno rečeno, iščemo model s čim manjšim številom neodvisnih spremenljivk, ki še vedno dobro pojasni podatke. Razlog za minimiziranje števila spremenljivk je v temu, da je takšen model bolj numerično stabilen in ima manjšo standardno napako. Če imamo v modelu preveč spremenljivk, je lahko preveč odvisen od opazovanih podatkov in dobimo numerično nestabilne ocene (običajno dobimo nerealno visoke ocene koeficientov in standardnih napak) [7].

Za pomoč pri izbiri spremenljivk, ki jih bomo vključili v model logistične regresije, imamo na voljo več korakov, ki si jih bomo podrobneje ogledali v naslednjih podpoglavjih.

#### 2.6.2.1 Opisna analiza

Proces izbire se mora začeti z *opisno analizo*, kar pomeni, da vsako spremenljivko opišemo z uporabo ene ali več statistik, kot so frekvenčna porazdelitev vrednosti, aritmetična sredina, standardni odklon, mediana, normalnost porazdelitve spremenljivke, sploščenost, simetričnost in podobno. Za različno vrsto spremenljivk imamo različne analize, in sicer:

- za kategorične spremenljivke izdelamo kontingenčne tabele izida ali konstruiramo črtni graf razčlenjen po spremenljivki izida,
- za zvezne spremenljivke pa ocenimo povprečje in standardne odklone ali konstruiramo okvir z ročaji, ki nam zelo nazorno prikaže obliko porazdelitve spremenljivke, njene kvartile, variacijski razmik in kvartilni razmik, ki ga odčitamo v dolžini okvira.

Te analize nam dajo začetno idejo o povezavi med pojasnjevalnimi spremenljivkami in izidom. V tem koraku tudi izločimo nekatere spremenljivke, če imajo premajhno variabilnost ali preveliko število manjkajočih vrednosti.

#### 2.6.2.2 Univariatna analiza

Proces izbire spremenljivk nadaljujemo z *univariatno analizo*. Že iz imena lahko razberemo, da testiramo zvezo med eno pojasnjevalno spremenljivko in odzivom, brez upoštevanja ostalih spremenljivk. To je zelo pomemben korak, da pridobimo skrajšan seznam spremenljivk za kasnejšo multivariatno analizo. Iz nadaljnje analize izključimo tiste spremenljivke, ki individualno ne kažejo nobene značilne povezave z izidom in je najverjetneje ne bi niti po združitvi z ostalimi spremenljivkami [5].

Rezultati univariatne analize pri logistični regresiji vključujejo Waldovo chi-kvadrat testno statistiko, razmerje verjetij, P-vrednost, ocene parametrov, standardne napake, razmerje obetov in meje intervala zaupanja.

P-vrednost je najmanjša stopnja značilnosti, pri kateri še zavrnemo ničelno hipotezo pri danih podatkih. Nizka P-vrednost pomeni, da obstaja majhna verjetnost, da ničelna hipoteza drži. Z drugimi besedami to pomeni, da v kolikor imamo hipotezo, ki predpostavlja, da med pojasnjevalno spremenljivko in izidom ni značilne povezave, potem nam nizka P-vrednost pove, da obstaja statistično pomembna povezava med testirano spremenljivko in izidom ter lahko hipotezo zavrnemo.

Po opravljeni univariatni analizi moramo izbrati spremenljivke za multivariatno analizo. Kandidati so tiste spremenljivke, katerih univariatni test ima P-vrednost $< 0,25$ [7]. Mejo za P-vrednost postavimo višjo kot je običajna (navadno je $0,05$), saj se izkaže, da sicer lahko na tem koraku izgubimo nekatere spremenljivke, ki se kasneje izkažejo za pomembne. Zaradi tega razloga je pomembno, da vzamemo večje število spremenljivk, saj pri univariatni analizi ignoriramo možnost, da določena spremenljivka lahko pridobi na moči ob prisotnosti neke druge spremenljivke.

#### 2.6.2.3 Testiranje korelacij

Sedaj pride na vrsto *testiranje korelacij*. Če sta dve pojasnjevalni spremenljivki visoko korelirani med sabo, lahko povzročata probleme pri multivariatni analizi, ker pojasnjujeta skoraj enako variabilnost izida [5]. Zato je zelo pomembno, da preučimo soodvisnost oziroma korelacijo med pojasnjevalnimi spremenljivkami in izločimo po eno iz para visoko koreliranih spremenljivk pred izvedbo multivariatne analize.

#### 2.6.2.4 Multivariatna analiza

Glavni in najpomembnejši korak izgradnje modela je *multivariatna analiza*. V tem koraku preverjamo povezavo spremenljivk z izidom, ob upoštevanju in kombiniranju z ostalimi spremenljivkami in vplivi.

Cilj multivariatne analize je pridobiti čim manjšo podskupino pojasnjevalnih spremenljivk, ki bodo predstavljale maksimalno variabilnost v odzivu. Ena izmed poglavitnih procedur, ki se običajno uporablja za izbiro dokončnih kazalnikov za model, je *metoda postopne izbire* (angl. stepwise method). Ta metoda je kombinacija postopkov vključevanja spremenljivk v model in izločanja spremenljivk iz modela. Metoda *vnaprejšnje izbire* (angl. forward selection) na vsakem koraku doda spremenljivko, ki največ doprinese k prileganju modela. Eden izmed kriterijev za izbiro spremenljivk za model je, da gledamo minimalno P-vrednost. Metoda *vzvratne izbire* (angl. backward selection) začne s kompleksnim modelom in na vsakem koraku odstrani spremenljivko,

ki ima najmanjši učinek na model (ima najvišjo P-vrednost). Postopek izločanja se zaključi, ko vsaka nadaljnja izločena spremenljivka vodi do bistveno slabšega prileganja. Metoda postopne izbire v regresijski model vključuje tiste spremenljivke, za katere ugotovi, da statistično značilno vplivajo na odvisno spremenljivko. Na vsakem koraku je neodvisna spremenljivka vključena, če je njena statistična značilnost dovolj velika. Velja tudi obratno, spremenljivke so izločene, če je vrednost statistične značilnosti premajhna. Metoda se zaključi, ko ni več primernih spremenljivk za vključitev ali izločitev iz analize.

Potrebno je poudariti, da statistična značilnost ne sme biti edino merilo za vključitev/izključitev spremenljivke v model. Če imamo spremenljivko, ki je ključnega pomena za našo študijo, jo je smiselno dodati v model in poročati njen učinek, četudi ni statistično značilna. S tem bomo lahko zmanjšali pristranskost v ocenah učinkov ostalih napovednih spremenljivk in bo morda lažje primerjati rezultate z drugimi študijami, kjer je bil učinek te spremenljivke pomemben (morda tudi zaradi večje velikosti vzorca). Kljub algoritmičnim izbirnim postopkom, moramo pri gradnji modela še vedno skrbno premisliti o izboru spremenljivk.

#### 2.6.2.5 Izbirni kriteriji

Obstaja veliko kriterijev za pomoč pri izbiri spremenljivk med multivariatno analizo logistične regresije. Ko izbiramo model, se moramo zavedati, da nikoli ne moremo z gotovostjo trditi, da smo našli pravega. Poleg že predhodno omenjenih statističnih testov značilnosti, poznamo še nekatera ostala merila, ki lahko pomagajo pri izbiri dobrega modela. En izmed najbolj poznanih testov je *Akaikeov informacijski kriterij* (angl. Akaike Information Criterion), ki presodi model glede na to, kako blizu so prilegajoče vrednosti glede na prave vrednosti, v smislu neke pričakovane vrednosti. Čeprav so vrednosti enostavnega modela bolj oddaljene od pravih vrednosti kot bi bile vrednosti kompleksnejšega modela, pa ima enostavnejši model lahko prednost, saj zagotavlja boljše ocene določenih karakteristik modela. Akaikeov informacijski kriterij je definiran kot

$$\text{AIC} = -2(\text{maksimiran logaritem verjetja} - \text{število parametrov v modelu}).$$

Opazimo lahko, da kaznujemo model s prevelikim številom parametrov, kar pomeni, da imajo modeli z nižjo vrednostjo AIC prednost pred modeli z višjo vrednostjo. Kadar se ostanki porazdeljujejo normalno, postane prvi člen v informacijskem kriteriju ocenjena varianca ostankov in meri prileganje modela podatkom. Več napovednih spremenljivk kot vključimo, več regresijskih koeficientov imamo na razpolago in lahko model bolje prilagodimo podatkom. Ko dodana spremenljivka ne prispeva več k napovedovanju podatkov, ostane varianca ostankov nespremenjena. Običajno se s tem, ko dodajamo

pojasnjevalne spremenljivke, varianca ostankov manjša in dobimo lažen občutek, da imamo vedno boljši model. Če si predstavljamo, da v model vključimo maksimalno število spremenljivk, bo varianca ostankov zelo blizu nič, ampak to nam ne bo dalo najboljšega modela za generiranje podatkov, temveč le zelo kompleksno transformacijo podatkov. Zato je drugi člen kriterija AIC takšen, da kaznuje vključitev dodatne napovedne spremenljivke v model [1]. Če primerjamo dva modela, ki se razlikujeta v eni neodvisni spremenljivki, bo model z nižjo vrednostjo AIC vzet kot boljši.

#### 2.6.2.6 Testiranje interakcij

Ko imamo izbran model, je potrebno izvesti še *testiranje interakcij* med spremenljivkami. V kateremkoli modelu nam interakcija med dvema spremenljivkama pove, da učinek ene izmed spremenljivk ni konstanten v vseh ravneh druge spremenljivke. Na primer, interakcija med spolom in težo implicira, da je koeficient naklona za težo različen za moške in ženske. Končna odločitev, ali bomo interakcije vključili v model ali ne, mora temeljiti tako na statističnem kot tudi praktičnem razmišljanju. Vsaka interakcija v modelu pa mora imeti smisel iz kliničnega vidika. Klinično verodostojnost obravnavamo tako, da ustvarimo seznam možnih parov spremenljivk v modelu, ki imajo neke znanstvene osnove za medsebojno interakcijo. Spremenljivke interakcije so ustvarjene kot aritmetični produkt parov spremenljivk z največjim vplivom. Nato te interakcijske spremenljivke dodamo po eno naenkrat v model, ki vsebuje vse glavne učinke in ocenimo njihovo statistično značilnost z uporabo testa razmerja verjetij. Prav je, da tudi interakcije v modelu prispevajo neko statistično značilnost, saj vključitev interakcije, ki ni statistično značilna, običajno le poveča ocenjeno standardno napako in ne izboljša ostalih statistik [7].

Modelu, ki ga dobimo po tem koraku pravimo *predhodni končni model.* Preden začnemo model lahko uporabljati tudi v praksi, moramo oceniti še njegovo napovedno moč in oceniti prileganje podatkom. Te postopke si bomo pogledali v naslednjem poglavju.

## 2.7 Ocenjevanje ustreznosti modela logistične regresije

V poglavju 2.4 smo si že pogledali statistike za preverjanje ustreznosti modela v globalnem smislu. Ko pa izberemo predhodni končni model, dobimo boljši vpogled in dodatne podatke za podrobnejše analize.

Predpostavimo, da imamo izbrane spremenljivke, s katerimi smo zadovoljni in so primerne za model. Sedaj nas zanima, kako učinkovito naš model opisuje odzivno

spremenljivko. V nadaljevanju si bomo ogledali dva precej različna pristopa k temu problemu.

Prvi pristop za ocenjevanje ustreznosti modela je izračun *statistik ustreznosti prileganja modela* (angl. goodness-of-fit). To so testi ničelne hipoteze, s katero testiramo ali je vgrajeni model sprejemljiv, ki nam kot rezultat vrnejo P-vrednost. V tem primeru nam P-vrednost pod neko določeno mejo (običajno $\alpha = 0,5$) nakazuje, da naš model ni sprejemljiv.

Z drugim pristopom želimo dobiti statistike, ki merijo, kako dobro lahko napovemo odvisno spremenljivko na podlagi neodvisnih spremenljivk. Takšnim statistikam pravimo *mere napovedne moči*. Običajno se vrednosti nahajajo med 0 in 1, kjer 0 pomeni, da model nima napovedne moči in 1 pomeni popolno napoved modela. Višja vrednost torej pomeni boljši model, vendar pa imamo redko določeno natančno mejo, ki razlikuje sprejemljiv model od nesprejemljivega.

Potrebno se je zavedati, da omenjena pristopa merita zelo različne stvari in ni nenavadno, da bi model z zelo visoko mero napovedne moči imel nesprejemljive statistike ustreznosti prileganja modela ter obratno. Statistike ustreznosti prileganja modela ne testirajo, kako dobro lahko napovemo odvisno spremenljivko, ampak ali bi model s tem, da bi ga bolj zakomplicirali z dodajanjem interakcije ali spreminjanjem povezovalne funkcije, lahko izboljšali.

### 2.7.1 Statistike ustreznosti prileganja modela

Statistike ustreznosti prileganja modela so običajno na voljo kot rezultat vsakega končnega modela in nam dajo celostno sliko ustreznosti modela, ne dajo nam pa informacije o posameznih komponentah v modelu. Kot rezultat dobimo P-vrednost in če je ta vrednost nizka (recimo pod $0,05$), potem zavrnemo predlagan model, če pa je visoka, to pomeni, da je naš model prestal test [2].

Po končanem prilagajanju modela, lahko pridobimo opazovano število dogodkov in pričakovano število dogodkov za vsak profil. Profili so skupine primerov, ki imajo popolnoma enake vrednosti napovednih spremenljivk. V nadaljevanju si bomo pogledali dve dobro poznani statistiki, ki primerjata opazovano in pričakovano število. To sta *Pearsonova* $\chi^2$ *statistika* in *odklonskost*. Kasneje pa bo predstavljen še *Hosmer-Lemeshowov test*, ki nam pomaga v primeru, ko prej omenjeni statistiki nista uporabni.

#### 2.7.1.1 Pearsonova $\chi^2$ statistika, odklonskost in Hosmer-Lemeshowov test

Dve najpogostejši meri ustreznosti prileganja modela, ki sta dostopni v skoraj vsaki komercialni programski opremi, sta *Pearsonova* $\chi^2$ *statistika* in *odklonskost*. Preden ju podrobneje opišemo, si na kratko poglejmo, kaj so kovariatni vzorci. Množici vrednosti

pojasnjevalnih spremenljivk vsakega subjekta pravimo *kovariatni vzorec* in ga označimo z $j$ ($j = 1, 2, ..., J$). Če ima vsak subjekt v vzorcu unikatno množico vrednosti, potem je število kovariatnih vzorcev enako številu opazovanih subjektov iz vzorca ($J = n$). Tak kovariatni vzorec je pogost, ko so pojasnjevalne spremenljivke zvezne in zelo natančne pri analizi binarne logistične regresije. Če pa vsak subjekt iz vzorca nima unikatne množice vrednosti, potem je število kovariatnih vzorcev manjše od števila opazovanih subjektov iz vzorca ($J < n$). V sledečem primeru je število subjektov z lastnostmi $\mathbf{x} = \mathbf{x}_j$ označeno z $m_j$, $j = 1, 2, 3, ..., J$ in nam predstavlja število opažanj znotraj kovariatnega vzorca $j$. Sledi, da je $\sum_j^J m_j = n$. Število dogodkov za vsak $m_j$ označimo z $y_j$ in sledi, da je $\sum_j^J y_j$ enako celotnemu številu dogodkov.

Pri logistični regresiji imamo več možnih načinov merjenja razlik med opazovanimi in pričakovanimi vrednostmi. Da poudarimo dejstvo, da so pričakovane vrednosti pri logistični regresiji izračunane za vsak kovariatni vzorec in so odvisne od ocenjene verjetnosti tega kovariatnega vzorca, označimo pričakovano vrednost za $j-$ti vzorec kot

$$\hat{y}_j = m_j\hat{\pi}_j = m_j\frac{e^{\hat{g}(\mathbf{x}_j)}}{1 + e^{\hat{g}(\mathbf{x}_j)}},$$

kjer je $\hat{g}(\mathbf{x}_j)$ ocenjen logit kovariatnega vzorca $j$.

Začnemo z upoštevanjem dveh mer, ki merita razliko med opazovanimi in pričakovanimi vrednostmi: *Pearsonov ostanek* in *odklonski ostanek*. Za specifičen kovariatni vzorec je Pearsonov ostanek definiran kot:

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1 - \hat{\pi}_j)}}.$$

Statistika, katera sešteje vse te ostanke, je Pearsonova $\chi^2$ statistika, ki jo zapišemo kot

$$\chi^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2.$$

Odklonski ostanek je definiran kot

$$d(y_j, \hat{\pi}_j) = \pm\left\{2\left[y_j \ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\ln\left(\frac{m_j - y_j}{m_j(1 - \hat{\pi}_j)}\right)\right]\right\}^{1/2},$$

kjer je predznak ($+$ ali $-$) enak kot predznak izraza $(y_j - m_j\hat{\pi}_j)$. Za kovariatni vzorec, kjer velja $y_j = 0$ je odklonski ostanek enak

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j|\ln(1 - \hat{\pi}_j)|},$$

pri $y_j = m_j$ pa je odklonski ostanek enak

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j|\ln(\hat{\pi}_j)|}.$$

Statistiko, ki sešteje vse te ostanke imenujemo odklonskost in jo zapišemo kot

$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2.$$

Četudi bosta $\chi^2$ in $D$ imeli različni vrednosti, moramo postati previdni, ko je razlika med njima velika, saj nam taka situacija lahko nakazuje, da hi-kvadrat približek porazdelitve za $\chi^2$ in $D$ ni zadovoljiv [3].

Odklonskost je običajno bolj zaželena za uporabo od Pearsonove statistike, še posebno, ko so logistični modeli vgrajeni z metodo maksimuma verjetja, saj ocene maksimuma verjetja za verjetnosti uspeha maksimirajo funkcijo verjetja za vgrajeni model in s temi ocenami je odklonskost minimizirana [3].

Obe statistiki sta uporabni, ko je pričakovano število, da se dogodek zgodi in pričakovano število, da se dogodek ne zgodi za vsak profil vsaj 5. Vendar pa večina sodobnih aplikacij logistične regresije uporablja podatke, ki ne dovolijo združevanja v profile, ker model vsebuje eno ali več zveznih napovednih spremenljivk. Ko imamo samo en primer na profil, imata odklonskost in Pearsonova $\chi^2$ statistika takšno porazdelitev, ki bistveno odstopa od prave hi-kvadrat porazdelitve, pri čemer dobimo P-vrednosti, ki so nepravilne. Ko imamo dejansko samo en primer na profil, odklonskost sploh ni več odvisna od opazovanih vrednosti, tako da postane povsem neuporabna kot mera ustreznosti prileganja modela [2].

Za rešitev prej omenjene težave imamo na voljo *Hosmer-Lemeshowov test ustreznosti prileganja*. Hosmer in Lemeshow sta predlagala združevanje primerov glede na njihove napovedane vrednosti iz modela logistične regresije. Napovedane vrednosti so razvrščene od najmanjše do največje in nato razdeljene v več skupin, ki so približno enakih velikosti. Običajno izberemo deset skupin in za vsako od skupin nato izračunamo opazovano ter pričakovano število, da se dogodek zgodi ali ne zgodi. Pričakovano število, da se je dogodek zgodil, je le seštevek napovedanih verjetnosti za vse posameznike v skupini. Pričakovano število, da se dogodek ne bo zgodil, je število subjektov skupine, od katerega odštejemo pričakovano število, da se je dogodek zgodil. Nato lahko apliciramo Pearsonovo $\chi^2$ statistiko, s katero primerjamo opazovana štetja s pričakovanimi. Kot pri klasičnih testih ustreznosti prileganja modela, nam nizka P-vrednost nakazuje, da moramo zavrniti model [2].

### 2.7.2 Mere napovedne moči modela

Mere napovedne moči modela merijo, kako dobro lahko napovemo odvisno spremenljivko na osnovi neodvisnih spremenljivk. Običajno se vrednosti teh mer nahajajo med 0 in 1, kjer 0 pomeni, da model nima napovedne moči in 1 pomeni, da model poda

popolno napoved. Očitno je, da višja vrednost pomeni boljšo napovedno moč, vendar je redko določena natančna meja, ki razlikuje sprejemljiv model od nesprejemljivega. Najpogosteje uporabljene mere napovedne moči so *R-kvadrat*, *klasifikacijske tabele* in *območje pod ROC krivuljo*.

#### 2.7.2.1 R-kvadrat statistika za logistično regresijo

Obstaja veliko različnih načinov za izračun statistike $R^2$ za logistično regresijo, vendar žal težko rečemo, katera je najboljša. Mittlbock in Schemper sta preiskovala lastnosti 12 različnih mer z uporabo naslednjih kriterijev [7]:

(1) interpretacija mere mora biti enostavna za razumevanje,

(2) kvadrat mere mora imeti spodnjo mejo 0 in zgornjo mejo 1,

(3) mera mora biti karakterno usklajena z logistično regresijo (to pomeni, da ni spremenjena z linearno transformacijo kovariat modela).

Po raziskovanju sta priporočala uporabo dveh mer za rutinsko uporabo, in sicer: kvadrat Pearsonovega korelacijskega koeficienta ter vsoto kvadratov $R^2$. Ostale mere, vključno z nekaterimi popularnimi $R^2$ statistikami, ki temeljijo na verjetju, so bile ocenjene kot neustrezne vsaj za enega izmed zgoraj naštetih kriterijev.

Ko imamo $n$ kovariatnih vzorcev, je kvadrat Pearsonovega korelacijskega koeficienta enak

$$r^2 = \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y})(\hat{\pi}_i - \bar{\pi})\right]^2}{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] \times \left[\sum_{i=1}^{n}(\hat{\pi}_i - \bar{\pi})^2\right]},$$

kjer je $\bar{y} = \bar{\pi} = n_1/n$. Vsoto kvadratov $R^2$ pa zapišemo kot

$$R^2_{ss} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Primer, ko imamo $J < n$ kovariatnih vzorcev ni bil predviden s strani Mittlbocka in Schemperja. Razširitev teh dveh mer bi v takem primeru zapisali kot

$$r^2_c = \frac{\left[\sum_{j=1}^{J}(y_j - m_j\bar{y})(m_j\hat{\pi}_j - m_j\bar{\pi})\right]^2}{\left[\sum_{j=1}^{J}(y_j - m_j\bar{y})^2\right] \times \left[\sum_{j=1}^{J}(m_j\hat{\pi}_j - m_j\bar{\pi})^2\right]},$$

in

$$R^2_{ssc} = 1 - \frac{\sum_{j=1}^{J}(y_j - m_j\hat{\pi}_j)^2}{\sum_{j=1}^{J}(y_j - m_j\bar{y})^2}.$$

Vedeti moramo, da so pri logistični regresiji nizke vrednosti $R^2$ nekaj običajnega. To velja tudi za ustrezne in dobre modele logistične regresije. Zato imamo lahko problem,

ko rezultate logistične regresije predstavljamo množici, ki je vajena rezultatov linearne regresije. Vseeno pa nam $R^2$ lahko pomaga v fazi gradnje modela, kot statistika za ocenjevanje in primerjanje konkurenčnih modelov [7].

#### 2.7.2.2 Klasifikacijska tabela

Klasifikacijska tabela navzkrižno razvršča binarni odziv z napovedjo ali je $y = 0$ ali 1. Napoved $\hat{y} = 1$ dobimo, ko je $\hat{\pi}_i > \pi_0$ in $\hat{y} = 0$, ko je $\hat{\pi}_i \leq \pi_0$, za neko mejo $\pi_0$. Večina klasifikacijskih tabel uporablja $\pi_0 = 0,5$ in povzamejo napovedno moč kot

$$\text{občutljivost} = P(\hat{y} = 1|y = 1) \quad \text{in} \quad \text{specifičnost} = P(\hat{y} = 0|y = 0).$$

Razlago pojmov *občutljivost* in *specifičnost* si bomo ogledali na naslednjem primeru. Recimo, da želimo postaviti diagnozo za raka na dojki. Z diagnostičnimi testi za to bolezen lahko dobimo dve pravilni diagnozi, in sicer pozitivni izid testa, kadar testiranec ima preiskovano bolezen ter negativni izid testa, kadar testiranec nima bolezni. V primeru, ko testiranec ima bolezen, pravimo pogojni verjetnosti, da je diagnostični test pozitiven, *občutljivost* (angl. sensitivity). V primeru, ko pa testiranec nima te bolezni, pa pravimo pogojni verjetnosti, da je diagnostični test negativen, *specifičnost* (angl. specificity). Želimo, da sta ti dve vrednosti čim bolj visoki.

Omejitve te tabele so, da pretvori zvezne napovedne vrednosti $\hat{\pi}$ v binarne in da je izbira $\pi_0$ poljubna [1].

#### 2.7.2.3 ROC krivulja

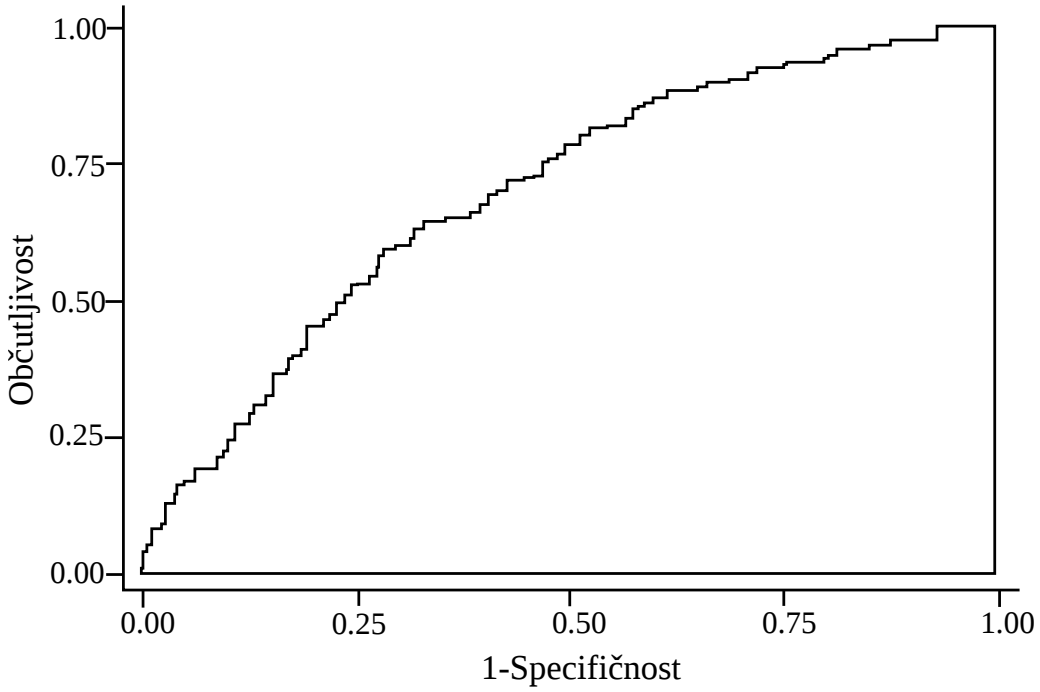
Občutljivost in specifičnost se zanašata le na eno določeno mejo za klasificiranje rezultatov testa, bolj celosten opis natančnosti razvrstitve pa dobimo z območjem pod *ROC* (angl. receiver operating characteristic) *krivuljo*. Krivulja nam prikazuje grafični prikaz med verjetnostjo zaznavanja pravilnega signala (občutljivost) in pa napačnega signala (1 - specifičnost) za celoten nabor možnih mejnih točk. Krivulja je običajno konkavne oblike in povezuje točke $(0, 0)$ in $(1, 1)$ kot lahko vidimo na sliki 2.

Območje pod ROC krivuljo predstavlja mero modelove sposobnosti razlikovanja med subjekti, ki imajo izid, kateri nam je v interesu v primerjavi s tistimi, ki nimajo tega izida. Če je naš cilj izbrati optimalno točko po kateri bomo subjekte razlikovali, je najboljša rešitev, da vzamemo točko, ki maksimira tako občutljivost kot specifičnost. Večje kot je območje pod ROC krivuljo, boljša je sposobnost razlikovanja. Splošno pravilo je [7]:

- Če $ROC = 0,5$: model nima sposobnosti razlikovanja.
- Če $0,7 \leq ROC < 0,8$: model ima sprejemljivo sposobnost razlikovanja.

- Če $0,8 \leq ROC < 0,9$: model ima odlično sposobnost razlikovanja.
- Če $ROC \geq 0,9$: model ima izjemno sposobnost razlikovanja.

V praksi je zelo neobičajno, da bi imeli območje pod ROC krivuljo večje od $0,9$. Se pa lahko zgodi, da imajo modeli s slabšim prileganjem dobro moč razlikovanja.




Slika 2: Primer ROC krivulje.

## 2.8 Podatkovno rudarjenje

Izgradnja modela iz zelo velikih baz podatkov predstavlja poseben izziv. Testi pomembnosti so običajno irelevantni, saj ima skoraj vsaka spremenljivka značilen efekt, če je $n$ dovolj velik. Strategije izgradnje modela vidijo določene modele za napovedovanje kot uporabne, kljub temu, da imajo kompleksno strukturo. Vseeno pa, če dodajamo preveliko število napovednih spremenljivk, dobimo zmanjšano učinkovitost, saj so po določeni točki novi kazalniki v veliki korelaciji z linearno kombinacijo ostalih kazalnikov, ki so prisotni v modelu in zaradi tega ne izboljšajo napovedne moči modela.

Z naraščanjem računalniške moči, se pojavljajo vedno večje zbirke podatkov. Na primer, finančne institucije, ki tržijo kreditne kartice, imajo na voljo odzivne podatke o tisočih oseb, katerim so poslali oglaševalno akcijo, o tem ali so naročili kartico ali ne. Za svoje komitente imajo mesečne podatke, če redno plačujejo svoje obveznosti in še mnogo drugih informacij, ki jih spremljajo o obnašanju svojih komitentov. Analizi ogromnih baz podatkov pravimo *podatkovno rudarjenje* (angl. data mining).

# 3 Modelski pristop k tveganju za neplačilo v poslovni banki

Banke se pri svojem poslovanju srečujejo z različnimi tveganji, ki jih lahko z učinkovitim upravljanjem precej obvladujejo. Eno izmed glavnih tveganj, s katerimi se banka srečuje, je *kreditno tveganje*. To je tveganje finančne izgube, ki je posledica komitentove nezmožnosti, da poravna obveznosti v celoti. Pri tem se gleda, kakšna je verjetnost, da dolžnik svojih obveznosti ne bo mogel ali ne bo hotel izpolnjevati. Cilj banke je oceniti verjetnost nastanka takega dogodka, z namenom, da si izračuna zadostne rezervacije in s tem minimizira stroške, ki nastanejo kot posledica odprave te škode, kar omogoči boljše upravljanje s kapitalskimi zahtevami.

Pristop IRB (angl. Internal Rating Based Approach) dovoljuje bankam, da uporabijo svoj statistični model za *ocenjevanje verjetnosti neplačila*, če banka upošteva minimalne zahteve s strani pristojnega nadzornika. Statistika nam ponuja različne metode za izdelavo modela za napovedovanje verjetnosti neplačila in najpogosteje je uporabljena ravno logistična regresija.

V nadaljevanju si bomo pogledali glavne komponente kreditnega tveganja in minimalne zahteve, ki jim mora model ustrezati, da lahko banka dobi odobritev s strani pristojnega nadzornika. V drugem delu tega poglavja bom predstavila postopek izdelave bonitetnega modela za segment velikih podjetij znotraj poslovne banke. Cilj je ugotoviti, katere spremenljivke oziroma kazalniki poslovanja podjetja v največji meri vplivajo na to, da bo komitent banke postal neplačnik in priti do formule, da bomo lahko za vsakega komitenta iz segmenta velikih podjetij izračunali verjetnost neplačila.

## 3.1 Merjenje kreditnega tveganja

Če želimo dobro upravljati s kreditnim tveganjem, moramo poznati njegove glavne komponente, ki predstavljajo ključne vhodne podatke za metode v okviru IRB pristopa in te so [4]:

- **Verjetnost neplačila** (angl. probability of default – PD)
  To je verjetnost, da posojilojemalec ne bo pravočasno ali v celoti poravnal svojih obveznosti. Časovni horizont za oceno verjetnosti neplačila je eno leto.

- **Izpostavljenost ob neplačilu** (angl. exposure at default – EAD)
  To je izpostavljenost banke do komitenta v trenutku, ko ta postane neplačnik. Izpostavljenost banke do posamezne osebe je vsota vseh postavk sredstev in zunajbilančnih postavk, ki izkazujejo terjatve in pogojne terjatve banke do te osebe in naložbe banke v finančne instrumente in kapitalske deleže te osebe [18].

- **Izguba v primeru neplačila** (angl. loss given default – LGD)
  To je izguba, ki se izraža v odstotku od izpostavljenosti ob upoštevanju zavarovanj za odobreno posojilo in prednostnih pravic.

- **Zapadlost** (angl. effective maturity – M) in **velikost dolžnika**.

Za merjenje kreditnega tveganja je najpomembnejša ravno verjetnost neplačila (PD) [9]. Da lahko banka meri verjetnost neplačila, potrebuje veliko količino informacij o komitentu, ki jih nato vključi v najrazličnejše modele izračunavanja verjetnosti, ki so lahko kvantitativni ali kvalitativni [16]. Končni rezultat je nato izračun verjetnosti neplačila dolga oziroma razvrstitev komitentov banke v določene razrede tveganosti, ki jim banke pripišejo vnaprej določene verjetnosti neplačila. Osnovni vhodni podatki, na katerih temeljijo ti modeli, se prilagajajo značilnostim posamezne skupine komitentov in so lahko povsem objektivni, torej ekonomski ali finančni dejavniki ocenjevanja, lahko pa tudi subjektivni podatki komitentovega položaja. Pri pravnih osebah lahko pridobimo podatke iz različnih analiz poslovanja in bilančnih izkazov, pri fizičnih osebah pa lahko podatkom finančnega značaja, kot sta na primer dohodek in rednost plačevanja, dodamo še podatke kot so delovna doba, starost komitenta in podobno. Na podlagi merjenja kreditnega tveganja se banka odloči, ali bo komitentu posojilo odobrila in pod kakšnimi pogoji.

Modeli kreditnega tveganja se običajno delijo v dve skupini, in sicer [16]:

- **Kvalitativni modeli**
  Pri teh modelih pridobivamo informacije s pomočjo notranjih virov banke in bonitetnih agencij. Te informacije proučujejo dejavnike, ki so značilni za vsakega posojilojemalca in dejavnike kreditnega tveganja, ki so vezani na vsa podjetja na trgu ter imajo vpliv na vse posojilojemalce. Te dejavnike nato subjektivno ovrednoti analitik in na ta način pride do končne bonitetne ocene komitenta.

- **Kvantitativni modeli**
  Tem modelom pravimo tudi *modeli kreditnega točkovanja* (angl. credit scoring models) in se uporabljajo za izračun verjetnosti neplačila ter za razvrščanje posojilojemalcev v skupine z enako verjetnostjo neplačila. Pri teh modelih se ukvarjamo z izbiro in povezovanjem različnih ekonomskih in finančnih značilnosti posojilojemalca in na ta način spoznamo, kateri dejavniki so pomembni za kreditno

tveganje. Takšno ocenjevanje omogoči lažjo izločitev posojilojemalca s previsokim kreditnim tveganjem, hkrati pa banki pomaga tudi pri izračunavanju potrebnih rezervacij za bodoče izgube. V to skupino modelov uvrščamo model linearne verjetnosti, model logistične regresije in modele diskriminantne analize.

Kvalitativni in kvantitativni modeli se med seboj ne izključujejo in banka lahko uporabi več modelov hkrati ter si s tem zagotovi boljšo oziroma bolj zanesljivo kreditno oceno. V slovenskem bančnem sistemu temelji ocenjevanje bonitete kreditojemalcev zaenkrat predvsem na kvalitativnih modelih. Kvantitativni modeli so se začeli uporabljati šele s prihodom kapitalskega sporazuma Basel II, ki ga bomo podrobneje spoznali v naslednjem poglavju.

## 3.2 Bonitetni sistem in Basel II

Bonitetni sistem, ki ga banke uporabljajo pri ocenjevanju tveganja svojih komitentov mora ustrezati minimalnim zahtevam za uvedbo IRB pristopa po kapitalskem sporazumu, ki se imenuje Basel II. Bonitetni sistem obsega metode, procese, kontrolo, zbiranje podatkov in informacijsko tehnologijo, ki služijo za ocenjevanje kreditnega tveganja, dodelitev internih bonitetnih ocen posameznim dolžnikom ali terjatvam ter določitev ocen neplačil in izgub.

Minimalne zahteve se nanašajo predvsem na strukturo bonitetnega sistema in vhodnih podatkov. Spodaj so navedene nekatere izmed zahtev, ki jih mora banka izpolnjevati za odobritev modela s strani nacionalnega nadzornika [4]:

- Banka mora oblikovati najmanj sedem bonitetnih razredov za porazdelitev tistih komitentov, ki izpolnjujejo svoje plačilne obveznosti in enega za tiste, ki so v statusu neplačila.
- Izbrati je potrebno primerno število razredov, da ne bi prihajalo do prevelike koncentracije v določenem razredu.
- Bonitetni razred komitenta je na podrobnih in jasnih kriterijih temelječa ocena tveganosti kreditojemalca, iz katere se nato dobi ocene verjetnosti neplačila.
- Definicija razreda mora vsebovati opis stopnje tveganja neplačila glede na profil dolžnikov, ki padejo v posamezen bonitetni razred in kriterije, ki so bili uporabljeni za identifikacijo te stopnje kreditnega tveganja.
- Pri dodeljevanju bonitetnih ocen mora banka uporabljati vse odločujoče in ažurne informacije, ki jih ima na voljo.

Za banke, ki uporabljajo napredni IRB pristop, ni predpisano minimalno število bonitetnih razredov za terjatve. Banka mora imeti dovolj razredov, da bo preprečila razvrščanje terjatev z zelo različnimi ocenami izgube v isti razred. Kriteriji za določitev teh bonitetnih razredov morajo biti utemeljeni na osnovi empiričnih dokazov.

Zahteve ne razkrijejo nobene preference glede izbire metode, saj je ena izmed glavnih idej IRB pristopa, da se banka sama odloči, kateri model bo izbrala. Seveda pa so nekatere metode bolj primerne od ostalih. Na primer, logit model je še posebej primeren za izdelavo stresnih testov, saj zajema tudi zgodovinske podatke. Metode pri katerih lahko testiramo posamezne vhodne kazalnike (logit in probit model), zagotavljajo lažji način za dokazovanje primernosti in verodostojnosti vhodnih kazalnikov. Ko je izid, ki nam ga da model, zvezna spremenljivka, lahko definiramo bonitetne razrede na bolj prilagodljiv način. Po drugi strani pa nobena od pomanjkljivosti obravnavanih modelov ne izključuje nobene od metod. Če je recimo banki všeč linearna regresijska analiza, pri kateri verodostojnosti vhodnih dejavnikov ni mogoče preveriti s statističnimi testi, ker nam linearna regresija ne vrne verjetnosti, bo banka morala iskati alternativne načine, da bo izpolnjevala minimalne zahteve.

## 3.3 Izdelava modela za ocenjevanja tveganja v banki

To poglavje opisuje primer izdelave modela za velika podjetja v poslovni banki. Pri razvoju in analizah takšnega modela sem tudi sama sodelovala, zato bom postopek opisala na podlagi svojih izkušenj. Podatki in model v tem poglavju so prirejeni in ne odražajo rezultatov dejanskega portfelja banke.

Glavna značilnost teh modelov je, da se osredotočajo na finančne kazalnike, zato si bomo podrobneje ogledali izgradnjo **finančnega modela za ocenjevanje tveganja**. Banka v končnem modelu vključuje tudi ocenjevanje tveganosti komitentov na podlagi njihovega poslovanja z banko (pravimo mu *vedenjski model*), kjer se upoštevajo podatki, ki jih dobimo iz notranjih bančnih sistemov ali iz centralnega bančnega sistema. Na koncu je dodana še subjektivna ocena izdelana s tako-imenovanimi mehkimi informacijami, ki jih pridobimo na podlagi vprašalnika, na katerega odgovori bančni skrbnik komitenta. Komitentu je na podlagi finančnega in vedenjskega modela dodeljena integrirana ocena, ki se preračuna v verjetnost neplačila in nato preslika v določeno bonitetno oceno. Z vprašalnikom lahko naknadno to oceno izboljšamo ali poslabšamo za toliko stopenj, kolikor se banka odloči. V primeru, da se skrbnik z oceno, ki mu jo vrne model, ne strinja, se lahko z utemeljenimi razlogi nanjo pritoži in predlaga drugačno oceno, seveda v okviru določenih omejitev.

V nadaljevanju bom predstavila metode, s katerimi se srečujemo pri razvoju notranjega bonitetnega sistema za poslovno banko. Pred začetkom izdelave modela pa se

moramo zavedati osnovnih vodil pri gradnji, ki so [8]:

- model mora biti **razumljiv**,
- model mora biti **učinkovit**,
- model mora biti **kalibriran na verjetnost neplačila** ter
- model mora biti **empirično preverjen**.

### 3.3.1 Izbira modela

V prejšnjem poglavju smo spoznali, da imamo različne metode, ki so primerne za ocenjevanje kreditnega tveganja, vendar je najbolj priljubljena logistična regresija oziroma logit model. Glavna dva razloga za to sta, da lahko izhodni podatek, ki nam ga da logit model, neposredno interpretiramo kot verjetnost neplačila in lahko enostavno preverimo, ali obstaja ekonomsko smiselna povezava med izbranimi kazalniki (napovednimi spremenljivkami) in tveganjem neplačila.

### 3.3.2 Opredelitev neplačila

Dogodek neplačila bo pri našem modelu odvisna spremenljivka, zato ga je najprej potrebno dobro definirati. V preteklosti so bonitetne modele razvijali na podlagi informacije, ali je podjetje v stečajnem postopku, saj so do tega podatka enostavno dostopali. Vendar pa imajo banke izgube tudi preden gre podjetje v stečaj, na primer, ko komitentu omogočijo odlog plačila brez nadomestila, v upanju, da bo posojilojemalec, ki je v težavah, v prihodnosti uspel odplačati svoj dolg. Zato je Baselski odbor za bančni nadzor leta 2001 opredelil definicijo neplačila, ki vključuje vse situacije, v katerih banka izgublja denar in določil, da bodo banke morale uporabljati to opredelitev neplačila za uporabo internega bonitetnega modela.

Da je prišlo do neplačila s strani dolžnika, se šteje, ko se zgodi eden ali oba od naslednjih dogodkov [17]:

- Banka meni, da obstaja majhna verjetnost, da bo dolžnik poravnal svoje kreditne obveznosti do banke v celoti, ne da bi bilo za poplačilo treba uporabiti ukrepe, kakor je unovčenje zavarovanja (v kolikor obstaja).
- Dolžnik več kakor 90 dni zamuja s plačilom katere koli pomembne kreditne obveznosti do banke, njene nadrejene družbe ali katere koli njej podrejene družbe.

Odvisna spremenljivka, ki bo opredeljena z izbranimi finančnimi kazalniki, bo pri tistih dolžnikih, ki v roku enega leta niso izpolnjevali nobenega od zgoraj opredeljenih

dogodkov, enaka 0. Vsi dolžniki, ki pa jim je bil pripisan kateri izmed teh dogodkov, bodo imeli vrednost 1.

Nazadnje je potrebno izbrati še časovni horizont za določitev ocene verjetnosti neplačila. Banke se običajno odločijo za obdobje enega leta, saj je to dovolj časa, da izvedejo določene ukrepe za preprečitev nastanka neplačila, po drugi strani pa je enoletni časovni zamik dovolj kratek, da zagotavlja ažurnost in primernost vhodnih podatkov za vnos v bonitetni model. Po minimalnih zahtevah se kljub temu pričakuje, da bo banka določila daljše časovno obdobje, na podlagi katerega bo dodeljevala bonitetne ocene, saj mora ta predstavljati oceno sposobnosti in pripravljenosti kreditojemalca izpolniti pogodbene obveznosti do banke, tudi v primeru, ko pride do poslabšanja gospodarskih razmer ali nepričakovanih dogodkov.

### 3.3.3 Populacija

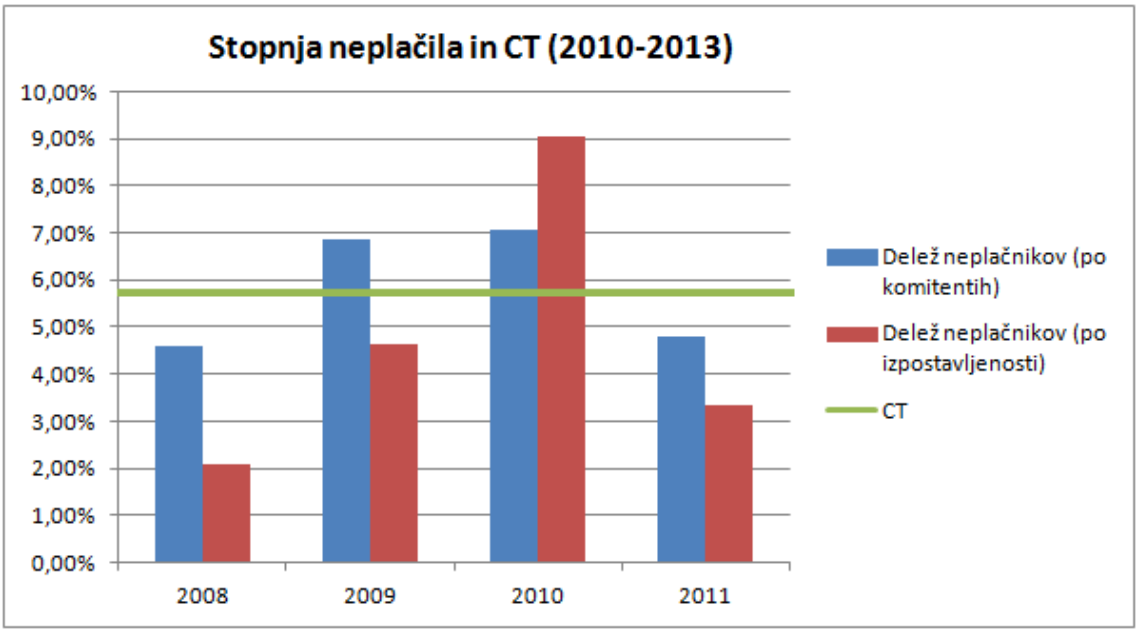
Da bi banka lahko dokazala, da so podatki, ki so jih uporabili za izdelavo modela, reprezentativni za dejanski portfelj dolžnikov in terjatev banke, mora najprej jasno definirati populacijo, za katero bo model veljaven. V našem primeru bodo to pravne osebe, ki so komitenti banke in spadajo po pravilih banke v segment *velikih podjetij*. Podjetje se smatra kot komitent banke v primeru, da je banka v trenutku obravnave do njega beležila izpostavljenost. Iz populacije izvzamemo podjetja z značilno drugačno strukturo bilance, in sicer finančne ustanove, podjetja javnega sektorja, socialne in zdravstvene ustanove ter tujce. S tem, ko izločimo določene kategorije podjetij, je naša populacija lahko bolj natančno opredeljena in lahko naredimo boljši model [8].

V tabeli 2 je predstavljen primer populacije velikih družb v poslovni banki, za podatke zbrane od leta 2010 pa do leta 2013, z izračunano *stopnjo neplačila* (angl. default rate - DR) glede na število komitentov in glede na izpostavljenost. Stopnja neplačila se izračuna kot delež komitentov, ki so znotraj opazovanega leta postali neplačniki.

Tabela 2: Populacija velikih podjetij (2010-2013)

| Leto | Št. podjetij | Izpostavljenost (v mln EUR) | Delež neplačnikov (po komitentih) | Delež neplačnikov (po izpostav.) |
|---|---|---|---|---|
| 2010 | 412 | 1.055 | 4,61% | 2,08% |
| 2011 | 478 | 1.024 | 6,87% | 4,63% |
| 2012 | 435 | 945 | 7,05% | 9,05% |
| 2013 | 533 | 855 | 4,78% | 3,32% |
| **Skupaj** | **1858** | **3879** | **5,83%** | **4,77%** |

Skupna stopnja neplačila, ki je glede na komitente enaka $5,83\%$, je izračunana kot povprečje stopenj neplačila za vsa štiri leta (v angleščini temu podatku pravimo *central tendency - CT* in je pomemben za kasnejšo kalibracijo modela). Na sliki 3 lahko vidimo, kako se giblje stopnja neplačila v obdobju od leta 2010 do leta 2013 in primerjava s CT.




Slika 3: Primerjava deleža neplačila in CT za velika podjetja (2010-2013)

### 3.3.4 Razvojni vzorec

Za potrebe izgradnje modela moramo najprej definirati *razvojni vzorec* (angl. development sample), na katerem bomo nato izbrali spremenljivke in ocenjevali parametre modela. Razvojni vzorec predstavlja del populacije, na podlagi katerega lahko izvedemo sklepanje o celotni populaciji.

V kolikor banka razpolaga s časovno vrsto, ki zajema daljše časovno obdobje od petih let, se lahko vzorči tako, da se najprej množico s podatki o komitentih razdelimo na dve podmnožici: na razvojni in testni vzorec. Razvojni vzorec, ki navadno vsebuje večino vseh zapisov, se uporabi za ocenjevanje bonitetnega sistema, medtem ko preostali podatki (angl. out-of-sample) služijo za vrednotenje modela. Pri ločevanju moramo biti pozorni na to, da so vsi podatki enega podjetja le v eni izmed podmnožic, ter da je razmerje med plačniki in neplačniki podobno v obeh podmnožicah. Navadno si za razvojno množico izberemo okrog $70\%$ vseh zapisov.

Ker je v našem primeru časovna vrsta krajša in imamo majhno število komitentov, bomo v razvojni vzorec za finančni model vzeli komitente iz celotne populacije, ki

zadostijo naslednjima dvema pogojema:

- komitenti, ki so v začetku opazovanega obdobja (leta) dobri (to pomeni, da niso neplačniki) in
- komitenti, ki imajo veljavne (ne starejše od treh let) in kvalitetne finančne izkaze za pripadajoče časovno obdobje.

V spodnji tabeli lahko vidimo predstavljen primer razvojnega vzorca finančnega modela, ki je razdeljen na tiste komitente, ki so znotraj obdobja ostali dobri (plačniki) in na tiste, ki so zašli v težave (neplačniki). Stopnjo neplačila dobimo tako, da izračunamo razmerje med številom neplačnikov in skupnim številom komitentov iz posameznega obdobja.

Tabela 3: Razvojni vzorec finančnega modela za velika podjetja (2010-2013)

| Leto | Plačniki | Neplačniki | Skupaj | Stopnja neplačila |
|---|---|---|---|---|
| 2010 | 340 | 13 | 353 | 3,68% |
| 2011 | 379 | 27 | 406 | 6,65% |
| 2012 | 332 | 28 | 360 | 7,78% |
| 2013 | 410 | 19 | 429 | 4,43% |
| **Skupaj** | **1461** | **87** | **1548** | **5,62%** |

### 3.3.5 Nabor podatkov

Podatki, s katerimi operiramo pri gradnji modela, so lahko interne narave, torej shranjeni v podatkovnem skladišču banke, ali pa jih banka pridobi od zunanjih institucij. Za naš primer je potrebna informacija ali je komitent postal neplačnik in ta podatek se hrani znotraj banke. Poleg tega so potrebni še finančni podatki iz bilanc podjetij, ki jih banke običajno pridobijo od Agencije Republike Slovenije za javnopravne evidence in storitve (AJPES).

### 3.3.6 Obdelava podatkov

To poglavje obravnava glavne dejavnosti obdelave podatkov, ki jih je potrebno izvesti preden se ocenjuje parametre modela. Te dejavnosti vključujejo čiščenje podatkov, izračun finančnih kazalnikov, univariatno analizo kazalnikov in na koncu še transformacijo ter normalizacijo kazalnikov.

#### 3.3.6.1 Čiščenje podatkov

Pomembno pri podatkih, ki jih bomo uporabili za izdelavo modela, je:

- da so brez očitnih napak,
- da nabor podatkov vključuje tiste spremenljivke, ki bodo uporabljene za izdelavo finančnih kazalnikov in imajo povezavo s stanjem neplačila,
- da je informacija o stanju neplačila na voljo za vse kreditojemalce.

Vse manjkajoče postavke finančnih podatkov morajo biti ustrezno obravnavane. Za nekatere kreditojemalce je običajno, da nekaterih finančnih podatkov nimajo. Če je število takih komitentov dokaj majhno, je najlažji način za odpravo tega problema, da jih izključimo iz obravnave. Če bi to pomenilo, da izgubimo preveliko število komitentov, se iz analize izključi vse spremenljivke, ki imajo velik delež manjkajočih vrednosti.

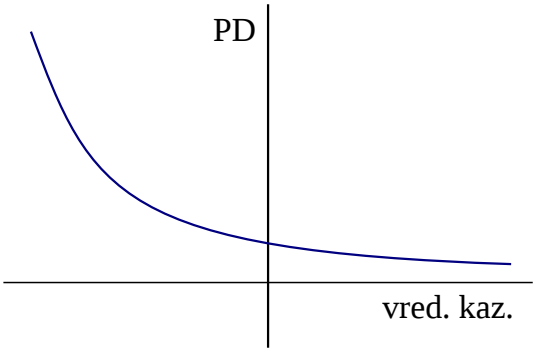
#### 3.3.6.2 Izračun finančnih kazalnikov

Ko je zagotovljena kakovost osnovnih finančnih podatkov, je potrebno izbrati potencialne pojasnjevalne spremenljivke. Najprej iz bilančnih postavk oblikujemo smiselne količnike, ki jim pravimo *kazalniki*. Količnike oblikujemo zato, da poenotimo razpoložljive informacije. Na primer, količnik »Dobiček / Sredstva« omogoča primerjavo donosnosti različno velikih podjetij. V splošnem bi morali izbrani vhodni bilančni kazalniki predstavljati najpomembnejše dejavnike kreditnega tveganja, kot so likvidnost, produktivnost, čiste prihodke od prodaje, sposobnost servisiranja dolga, dejavnost, donosnost, velikost podjetja, stopnjo rasti in finančni vzvod. Celotna množica lahko vsebuje tudi 200 in več kazalnikov (angl. long list) iz katerih je na koncu potrebno izbrati tiste, ki najbolje napovedujejo tveganje za neplačilo. V tabeli 4 lahko vidimo primer izdelanih kazalnikov, ki predstavljajo majhno množico tipičnih poslovnih količnikov, ki se uporabljajo v banki.

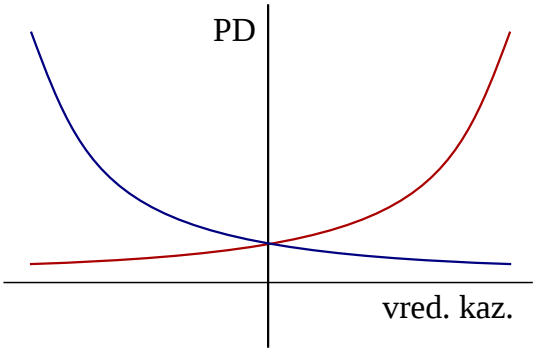
Tabela 4: Primer finančnih količnikov

| **Kazalnik** | **Kategorija** | **Opis** | **Hipoteza** | **Predznak** |
|---|---|---|---|---|
| DE2 | Dejavnost | Kratkoročne poslovne obveznosti / Čisti prihodki od prodaje | $-$ | $-1$ |
| FV3 | Finančni vzvod | (Kapital + Dolgoročne finančne obveznosti) / Dolgoročne finančne obveznosti | $+$ | 1 |
| L1 | Likvidnost | (Kratkoročne obveznosti + Kratkoročne pasivne časovne razmejitve) / Čisti prihodki od prodaje | $-$ | $-1$ |

Predzadnji stolpec v tabeli 4 prikazuje pričakovano odvisnost med finančnim kazalnikom in verjetnostjo neplačila, kjer znak + pomeni, da večja vrednost količnika vodi v izboljšanje stanja podjetja (večja verjetnosti plačila), znak − pa nakazuje nasprotno vedenje. Zadnji stolpec nam pove, s katero vrednostjo moramo pomnožiti celotni finančni količnik, da pridobimo enolično monotonost vseh kazalnikov, saj želimo, da je odvisnost med finančnim količnikom in verjetnostjo neplačila za vse kazalnike enaka. Na sliki 4 in 5 nam abscisna os predstavlja vrednost kazalnika, ordinata pa verjetnost neplačila (PD). Slika 4 nam prikazuje, kako naj bi se vsak količnik gibal. Na sliki 5 pa je z rdečo črto označeno napačno gibanje kazalnika, ki ga moramo pomnožiti z -1, da dobimo pravilno gibanje (modra črta).




Slika 4: Pravilno gibanje kazalnika




Slika 5: Napačno gibanje kazalnika

Ko sestavimo vhodne finančne količnike, je potrebno narediti še analizo napak pri izračunu. Za vsako kombinacijo števca in imenovalca (na primer števec pozitiven/imenovalec negativen, števec pozitiven/imenovalec enak 0, itd.) se je potrebno odločiti, kako jo bomo obravnavali. V določenih primerih je izračun kazalnika res lahko napaka in ga kot takega tudi pustimo, lahko pa imamo izjeme, ko nam kazalnik, ki se ni izračunal, v resnici pomeni nekaj dobrega ali pa nekaj zelo slabega. V takih primerih se lahko odločimo, da vrednost kazalnika nadomestimo z minimumom, maksimumom ali mediano tega kazalnika, izračunanimi na celotni populaciji za ustrezno časovno obdobje. Izbira je seveda odvisna od ekonomskega smisla. V tabeli 5 lahko vidimo primer kriterijev, ki veljajo za vsako kombinacijo števca in imenovalca kazalnikov iz tabele 4.

Tabela 5: Kriteriji za obravnavo kazalnikov

| Kazalnik | Den>0, Num>0 | Den>0, Num=0 | Den>0, Num<0 | Den =0, Num>0 | Den=0, Num=0 | Den=0, Num<0 | Den<0, Num>0 | Den<0, Num=0 | Den<0, Num<0 |
|---|---|---|---|---|---|---|---|---|---|
| DE2 | ok | 0 | ok | max | napaka | mediana | napaka | napaka | napaka |
| FV3 | ok | min | min | max | mediana | min | napaka | napaka | napaka |
| L1 | ok | napaka | napaka | napaka | napaka | napaka | napaka | napaka | napaka |

#### 3.3.6.3 Univariatna analiza in skrajšan seznam kazalnikov

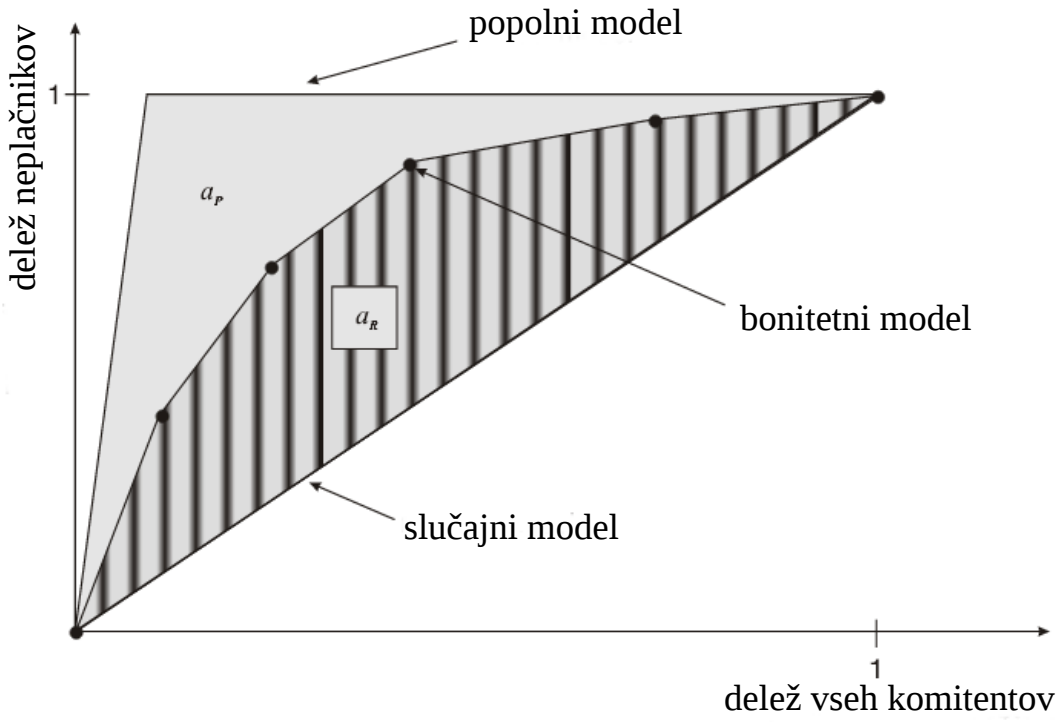
Pred začetkom univariatne analize je potrebno najti *osamelce* (angl. outliers) posameznega količnika, saj lahko močno izkrivijo ocenjene parametre modela. Odstopanja se pri količnikih lahko pojavijo tudi, če so osnovni podatki prečiščeni, na primer, ko je imenovalec količnika zelo majhna vrednost blizu ničle. Da se izognemo takšnim težavam, običajno zamenjamo ekstremne vrednosti s 1. oziroma 99. percentilom pripadajočega kazalnika. Ta korak je pomemben, saj s tem preprečimo prevelik vpliv na povprečne vrednosti, ki jih računamo posebej za plačnike in neplačnike. Bolj kot sta si povprečji različni, večjo sposobnost ima kazalnik, da razlikuje med dobrimi in slabimi komitenti.

Pri univariatni analizi se za vsak količnik izračuna razne statistike, kot na primer:

- koeficient natančnosti (angl. accuracy ratio - AR),
- mediano, povprečje itd.
- število manjkajočih in ničelnih vrednosti kazalnika,
- predznak naklona (v našem primeru mora biti negativen). Glede na monotonost kazalnikov (kazalniki z višjo vrednostjo morejo biti povezani z dobrimi komitenti in obratno) mora biti povprečje dobrih komitentov vedno višje od povprečja slabih komitentov. Če temu ni tako, potem naklon prejme pozitiven predznak.

*Koeficient natančnosti* (angl. accuracy ratio – AR) je mera, ki je uporabna za ocenjevanje sposobnosti modela (ali posameznega kazalnika) pri razločevanju med dobrimi in slabimi komitenti. Izračun koeficienta natančnosti izhaja iz *CAP krivulje* (angl. Cumulative accuracy profiles). CAP krivulja je zelo podobna ROC krivulji in prikazuje kumulativno frekvenco neplačnikov glede na kumulativno frekvenco vseh komitentov.

Na sliki 6 je prikazan koncept CAP krivulje. Strmejša kot je krivulja, točnejši je bonitetni model. *Popolni model* (angl. perfect model) bi vsem neplačnikom dodeljeval najnižje rezultate (oziroma najvišjo stopnjo tveganja). V takem primeru bi CAP krivulja na začetku linearno naraščala, nato pa bi postala horizontalna in bi ostala pri 100%. *Slučajni model* (angl. random model) bi dolžnike naključno razvrščal v bonitetne razrede, kar bi pomenilo, da nima napovedne moči. CAP krivulja bi v tem primeru potekala po diagonali. V realnosti pa se CAP krivulja, ki pripada zgrajenemu *bonitetnemu modelu* (angl. rating model) oblikuje med obema omenjenima skrajnostma [6].




Slika 6: CAP krivulja in koeficient natančnosti AR [6]

Mera AR povzema informacije, ki jih določa CAP krivulja in se izračuna kot

$$AR = \frac{a_R}{a_P},$$

kjer je $a_R$ površina med dobljeno CAP krivuljo in krivuljo slučajnega modela, $a_P$ pa površina med krivuljo popolnega in slučajnega modela.

Vrednost AR se nahaja v razponu med 0% (slučajni model) in 100% (popolni model), kjer vrednost, ki je bližje 100%, pomeni boljši bonitetni model. Praksa je pokazala, da se kot modele, ki dobro ločijo slabe od dobrih komitentov, upošteva tiste, ki imajo AR med $60\% - 70\%$ [8].

Da iz celotnega seznama kazalnikov dobimo *skrajšan seznam* (angl. short list), si določimo mejo izračunanih kriterijev, ki jo bomo upoštevali pri izločanju kazalnikov. V našem primeru je kazalnik prišel na skrajšan seznam, če je imel izpolnjene naslednje pogoje:

- AR kazalnika je višji od $0,05$,
- predznak naklona mora biti negativen in
- kazalnik nima več kot 50% ničelnih vrednosti.

V skladu z zgoraj navedenimi pogoji je iz nadaljnje obravnave izločenih 50 kazalnikov. Da dobimo končni skrajšan seznam, so preostali kazalniki vključeni še v *metodo razvrščanja v skupine* (angl. cluster analysis). To je matematična multivariatna metoda, ki omogoča združevanje enot (v našem primeru so to kazalniki) v homogene skupine na osnovi določenih kriterijev sorodnosti (spremenljivk), obenem pa pokaže tudi tipične predstavnike teh skupin. Da lahko enote razvrstimo v homogene skupine potrebujemo mero, s katero je mogoče presojati podobnost med enotami. Običajno se v ta namen uporablja evklidska razdalja. Kvadrat evklidske razdalje je vsota kvadriranih razlik med vrednostmi dveh spremenljivk za vse možne pare enot in jo zapišemo kot:

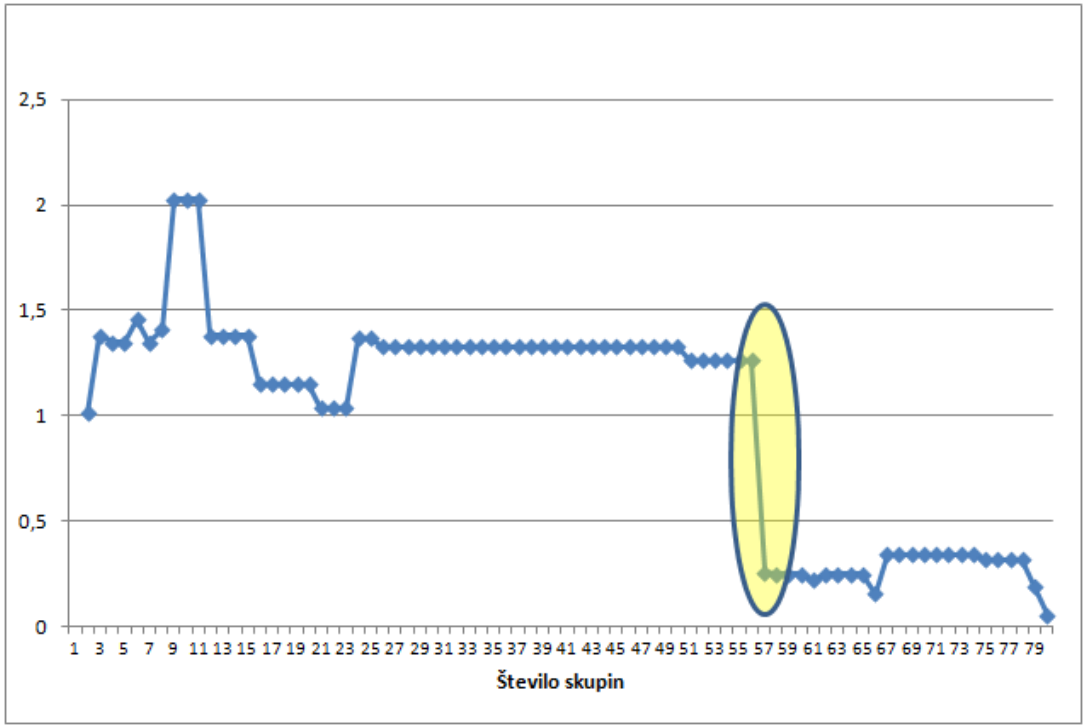
$$d_{rs}^2 = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2,$$

kjer je $d_{rs}^2$ kvadrirana evklidska razdalja med enotama $r$ in $s$, $x_{rj}$ vrednost $j$-te spremenljivke pri enoti $r$, $x_{sj}$ vrednost $j$-te spremenljivke pri enoti $s$ in $p$ število spremenljivk. Poznamo več metod, ki omogočajo združevanje enot v skupine. Ena izmed njih je *metoda variance* oziroma *Wardova metoda*, pri kateri se skupine tvorijo na osnovi minimiziranja variance znotraj skupin oziroma maksimiranja homogenosti znotraj skupin. Vsota kvadratov znotraj skupin služi kot merilo homogenosti. Na vsakem koraku se skupine oblikujejo tako, da je za oblikovane skupine vsota kvadratov znotraj skupin minimalna. Wardova metoda zahteva uporabo evklidske razdalje, na podlagi katerih se tvori matrika podobnosti, ki omogoča združevanje enot v skupine po različnih metodah, na primer *metoda hierarhičnega razvrščanja*. Ta metoda prične razvrščanje s številom skupin, ki je enako številu enot, nato pa se v vsakem koraku število skupin zmanjša za eno. Koliko skupin bomo na koncu izbrali, ni natančno določeno.

Postopek razvrščanja v skupine lahko naredimo z metodo VARCLUS v statističnem programu SAS. Ta metoda nam omogoča razvrščanje kazalnikov v skupine in izračune različnih statistik, kot sta korelacija med kazalniki znotraj skupine ter korelacija med skupinami. Korelacija znotraj skupine mora biti čim večja, med skupinami pa čim manjša. Kot rezultat te metode dobimo tudi količnik, ki nam na vsakem koraku konstrukcije skupin izračuna vrednost

$$\frac{1 - (\text{korelacija znotraj skupine})^2}{1 - (\text{korelacija med skupinami})^2},$$

kjer nam *korelacija znotraj skupine* da minimalno vrednost korelacije znotraj skupine in *korelacija med skupinami* najvišjo vrednost korelacije med danimi skupinami. Vrednost tega količnika je nizka, če so skupine med sabo dobro ločene [15]. Izračunane vrednosti količnika glede na število skupin, lahko vidimo na sliki 7.Vidimo, da je smiselno izbrati 57 skupin, saj se prej omenjeni količnik pri manjšem številu skupin poveča, kar pomeni slabše razločevanje med skupinami.




Slika 7: Vrednost statistike za ocenitev primernega števila skupin

Da dobimo končni skrajšan seznam, izberemo po enega izmed kazalnikov iz vsake skupine in to tistega, ki ima najvišji koeficient natančnosti AR. Na koncu smo torej dobili 57 kazalnikov, ki so razdeljeni po kategorijah kot je prikazano v tabeli 6.

Tabela 6: Skrajšan seznam finančnih kazalnikov

| Kategorija | Šifra kategorije | Št. kazalnikov |
|---|---|---|
| Dejavnost | DE | 11 |
| Sposobnost servisiranja dolga | SSD | 13 |
| Finančni vzvod | FV | 15 |
| Likvidnost | L | 8 |
| Donosnost | DO | 10 |
| **Skupno število kazalnikov** | | **57** |

#### 3.3.6.4 Transformacija in normalizacija kazalnikov

Namen transformacije in normalizacije količnikov je v prvi vrsti razrešiti nelinearnost v odvisnosti med danimi kazalniki in verjetnostjo neplačila. Transformacija nam omogoči, da izstopajoče vrednosti nimajo velikega vpliva na regresijo in naredi količnike monotone v primerjavi s stopnjo neplačila. Lahko bi se nam na primer zgodilo, da bi nam kateri od kazalnikov prikazoval nelinearno in nemonotono obnašanje in bi zaradi predpostavke o linearnosti, ki je del modela logit, razmerje med kazalnikom in dogod-

kom neplačila bilo napačno zajeto v model. Če kazalniki, ki so vključeni v regresijo niso normalizirani, potem ni možno neposredno interpretirati uteži, ki so jim dodeljene. S tem, ko normaliziramo vse kazalnike, omogočimo interpretacijo koeficientov kazalnikov kot uteži in jih na ta način lahko med seboj primerjamo.

Transformacijo in normalizacijo naredimo z naslednjima dvema korakoma:

(1) Najprej je kazalnik transformiran z logistično funkcijo:

$$\text{logistično-transformirana vrednost} = \frac{1}{1 + e^{-Sl \cdot (x - Md)}},$$

kjer je $x$ vrednost kazalnika in sta $Md$ (srednja vrednost) ter $Sl$ (naklon) izračunani kot:

$$Md = \frac{L + U}{2}; \quad Sl = \frac{2,95}{U - Md},$$

kjer je $L$ spodnja mejna vrednost in $U$ zgornja mejna vrednost kazalnika, izračunani kot 5. in 95. percentil na razvojnem vzorcu. Konstanta $2,95$ predstavlja 95. percentil logistične porazdelitve in omogoči, da dobimo transformirane vrednosti enake 0 za kazalnike z vrednostjo enako ali višjo od 95. percentila ter vrednost 1 za kazalnike z vrednostjo enako ali nižjo od 5. percentila.

(2) Sedaj pride na vrsto normalizacija, kjer dobimo končno vrednost kazalnika ($T(x)$) s povprečjem 0 in standardnim odklonom 50. Potrebovali bomo še dva dodatna parametra, in sicer povprečje transformiranega kazalnika ($M$) in njegov standardni odklon ($S$). Formula, ki jo uporabimo, je naslednja:

$$T(x) = \frac{\frac{1}{1+e^{S \cdot (M-x)}} - \mu}{\sigma} \cdot 50.$$

V nadaljnjih analizah transformirane in normalizirane vrednosti nadomestijo originalne vrednosti količnikov, kar lahko bistveno vpliva na kakovost modela.

### 3.3.7 Izdelava končnega bonitetnega modela

Kazalnike iz skrajšanega seznama lahko sedaj uporabimo za izdelavo končnega multivariatnega logističnega modela. Ocena okvirnega števila finančnih kazalnikov temelji na raziskavah, ki so pokazale, koliko faktorjev je moč razločiti znotraj nabora več sto finančnih kazalnikov. Za kvantitativne modele za ocenjevanje verjetnosti neplačila naj bi bilo primerno število okoli sedem finančnih kazalnikov, s katerimi lahko dokaj zanesljivo napovedujemo dogodek neplačila [12]. Seveda pa ta številka ni zanesljiva in nam je lahko le v pomoč, ko se odločamo, koliko kazalnikov bomo na koncu dejansko izbrali.

Model se razvija na vnaprej definiranem manjšem vzorcu. Za izbor končnih kazalnikov med multivariatno analizo uporabimo metoda postopne izbire (stepwise) skupaj z analizo statistične značilnosti kazalnikov in koeficienta natančnosti (AR).

Ko so izbrani končni kazalniki, se za namen stabilizacije uteži kazalnikov uporabi tehniko *samovzorčenja brez vračanja*, kar pomeni, da se v vzorcu, ki ga obravnavamo, posamezen komitent lahko pojavi le enkrat. Bistvo tehnike samovzorčenja je, da iz razvojnega vzorca ustvarimo $K$ podvzorcev, ki imajo $n$ enot. Če uporabimo metodo brez vračanja, to pomeni, da je vsaka enota lahko izbrana le enkrat. Iz vsakega podvzorca izračunamo vzorčne ocene parametrov, ki nas zanimajo in če je porazdelitev vzorčne ocene približno normalna, nato izračunamo oceno za povprečje ter za varianco vzorčne ocene. S to tehniko na koncu dosežemo, da uteži modela niso pristranske in odvisne od enega samega vzorca.

Ker je tveganost podjetja odvisna tudi od dejavnosti, s katero se ukvarja, se v model lahko doda še kategorična spremenljivka, ki opredeljuje sektor oziroma glavno dejavnost podjetja. Izdelavo takšne spremenljivke bom podrobneje opisala v naslednjem podpoglavju.

#### 3.3.7.1 Vpeljava kategorične spremenljivke

Zaradi predhodno omenjenega razloga, se v model lahko vključi še kategorična spremenljivka, ki predstavlja dejavnost podjetja (npr. gradbeništvo, industrija, trgovina, storitve ...). S to spremenljivko v modelu dosežemo to, da so na primer komitenti iz področja gradbeništva, ki je v zadnjih letih precej tvegana dejavnost, zaznani kot bolj tvegani glede na komitente iz drugih dejavnosti.

To dodatno spremenljivko obravnavamo z *metodo WOE* (angl. weight of evidence), ki je pogosto v uporabi pri kategoričnih spremenljivkah. WOE je tehnika s katero linearizirамo tveganje, ki je povezano z vsako kategorijo te spremenljivke. Statistiko WOE definiramo kot logaritemsko razmerje med deležem dobrih in deležem slabih komitentov znotraj posamezne kategorije $i$ glede na celotno število komitentov v vseh kategorijah:

$$\mathrm{WOE}_i = \ln\left(\frac{\mathrm{dobri}_i/\mathrm{dobri}}{\mathrm{slabi}_i/\mathrm{slabi}}\right).$$

Vrednosti, ki jih statistika WOE zavzame, interpretiramo na naslednji način:

- Če je vrednost WOE enaka 0, potem ima opazovana kategorija povprečno stopnjo neplačila.
- Če je vrednost WOE negativna, potem ima opazovana kategorija višjo stopnjo neplačila kot ostale kategorije (višje tveganje).
- Če je vrednost WOE pozitivna, potem ima opazovana kategorija nižjo stopnjo neplačila kot ostale kategorije (nižje tveganje).

Kakšno povezavo ima WOE statistika z logistično funkcijo, lahko vidimo v naslednji enačbi:

$$\text{logit}_i = \text{WOE}_i - \ln\left(\frac{1-PD_i}{PD_i}\right) = \text{WOE}_i - \ln\left(\frac{dobri}{slabi}\right). \tag{3.1}$$

Količina, ki jo odštejemo od vrednosti $\text{WOE}_i$, da dobimo $\text{logit}_i$, je logaritemsko razmerje med vsemi dobrimi in vsemi slabimi komitenti iz razvojnega vzorca.

Ravno zaradi linearne povezave z logistično funkcijo, je primerna uporaba kategoričnih spremenljivk transformiranih z WOE tehniko. V tabeli 7 lahko vidimo, kakšne vrednosti smo dobili za različne kategorije dejavnosti. Vidimo lahko, da je bil sektor gradbeništva najbolj tvegana kategorija v izbranem obdobju.

Tabela 7: Vrednosti WOE glede na kategorijo dejavnosti komitenta

| Sektor (kategorija) | Dobri | Slabi | WOE |
|---|---|---|---|
| Industrija | 429 | 23 | 0,08542 |
| Trgovina | 409 | 26 | -0,08492 |
| Gradbeništvo | 173 | 22 | -0,77829 |
| Storitve | 386 | 14 | 0,47624 |
| Ostalo | 64 | 2 | 0,62520 |

#### 3.3.7.2 Vzorčenje

Za potrebe logistične regresije, so končni kazalniki izbrani na manjšem vzorcu. V ta vzorec se vključi vse neplačnike iz razvojnega vzorca in delež dobrih komitentov, ki so bili izbrani naključno (brez vračanja), tako da je stopnja neplačila približno 25%. Na enak način se nato ustvari 500 različnih vzorcev, s katerimi se stabilizirali uteži modela in jih naredi nepristranske.

Tabela 8: Populacija za samovzorčenje in število komitentov v vsaki iteraciji

| Populacija za samovzorčenje | | | Delež izvzetih komitentov iz populacije za samovzorčenje | | Število komitentov v vzorcih za samovzorčenje | | |
|---|---|---|---|---|---|---|---|
| Dobri | Slabi | DR | Dobri | Slabi | Št. dobrih | Št. slabih | DR |
| 1461 | 87 | 5,62% | 17,86% | 100% | 261 | 87 | 0,25 |

Vzorčenje dobrih komitentov omogoča metoda *enostavnega slučajnega vzorčenja* (angl. simple random sampling - SRS) v programu SAS. Vsak vzorec ima $n$ različnih enot izmed $N$ možnih, ki imajo enako verjetnost izbora v vzorec. Verjetnost, da bo posamezna enota izbrana v vzorec je enaka $n/N$.

#### 3.3.7.3 Končni model

Proces izbora kazalnikov za končni model vključuje izbiro končne kombinacije kazalnikov, ki je najbolj učinkovita in ekonomsko smiselna med vsemi ostalimi kombinacijami. Z metodo postopne izbire, ki jo izvedemo s programom SAS, dobimo izbor kazalnikov, katerih kombinacija je najbolj učinkovita. Sedaj je potrebno preveriti ali ima ta kombinacija tudi ekonomski smisel.

Končno izbiro modela se običajno določi v sodelovanju z oddelkom, ki se ukvarja z analizo kreditnih predlogov, saj bodo oni kasneje tudi uporabniki izhodnih napovedi, ki jih bo ponujal model. Ključni elementi pri izbiri končnega modela so napovedna moč in pa ravnovesje med statistično analizo ter izkušnjami ekonomistov.

Ko so optimalni kazalniki izbrani za finančni model, se uporabi postopek samovzorčenja brez vračanja za namen stabilizacije uteži in neodvisnosti od vzorca. V procesu opravimo 500 iteracij, pri katerih je vsakič ustvarjen nov vzorec, ki je primerljiv z vzorcem preko katerega so bili izbrani končni kazalniki in je imel stopnjo neplačila enako 25%. Parametri, ki so kasneje implementirani v končni finančni model, so enaki povprečnim vrednostim teh parametrov, ki smo jih izračunali na 500-tih vzorcih.

V tabeli 9 lahko vidimo primer izbranih kazalnikov ter njihovi pripadajoči koeficienti ter uteži. Iz tabele je razvidno, da ima največjo utež kazalnik FV2, ki je tudi statistično najbolj značilen.

Tabela 9: Končni model - izbrani kazalniki in njihovi koeficienti

| **Kazalnik** | $\beta$ | **Utež** | **Kategorija dejavnosti** | **Opis kazalnika** |
|---|---|---|---|---|
| SSD1 | -0,002 | 6% | Sposobnost servisiranja dolga | (Dolgoročne finančne obveznosti + Kratkoročne finančne obveznosti - Denarna sredstva)/ (Čisti dobiček - Čista izguba + Odpisi vrednosti) |
| SSD3 | -0,005 | 13% | Sposobnost servisiranja dolga | Finančni odhodki iz finančnih obveznosti/(Dolgoročne finančne obveznosti + Kratkoročne finančne obveznosti) |
| FV2 | -0,018 | 45% | Finančni vzvod | (Obveznosti do virov sredstev - Kapital - Denarna sredstva)/ Sredstva |
| DO5 | -0,008 | 20% | Donosnost | (Čisti Dobiček - Čista Izguba)/(Kapital + Dolgoročne Obveznosti) |
| Sektor | -0,802 | 16% | Sektor | Sektor: Industrija, Trgovina, Gradbeništvo, Storitve, Ostalo |

Kot prikazano v formuli (2.4), nam linearna kombinacija kazalnikov in izračunanih koeficientov $\beta$, skupaj s koeficientom odziva $\alpha = -1,6067$, da *izid modela* (angl. score).

### 3.3.8 Ocenjevanje ustreznosti in učinkovitosti modela

To poglavje povzema najpomembnejše rezultate, ki nam jih vrne SAS-ova procedura za izbrani model in so pridobljeni z uporabo posebne procedure za logistično regresijo na izbranem vzorcu, ki smo ga opisali v poglavju 3.3.4. Opisi rezultatov procedure so povzeti po SAS-ovem uporabniškem priročniku [14].

Procedura logistične regresije v SAS-u nam ponudi mnogo informacij. Poleg tega, da omogoča oceno in primerjavo različnih modelov, nam pomaga tudi ugotoviti, katera od neodvisnih spremenljivk največ prispeva k učinkovitosti modela.

V kolikor želimo preveriti, kakšna je *skupna statistična značilnost izbranih neodvisnih spremenljivk*, preverimo rezultate v tabeli »Testing Global Null Hypothesis: BETA = 0«, zlasti vrednosti za $\chi^2$ (Chi-Square) in P-vrednost (P-value) pri testu razmerja verjetij (Likelihood Ratio). V tej tabeli dobimo rezultate treh testov, in sicer testa razmerja verjetij, testa ocene in Waldovega testa. Vsi trije testi preverjajo ničelno hipotezo, ki pravi, da so vsi koeficienti napovednih spremenljivk v modelu enaki 0. Alternativna hipoteza pa pravi, da je vsaj eden izmed koeficientov napovednih spremenljivk različen od 0. Če želimo hipotezo zavrniti, mora biti P-vrednost pri vseh treh testih manjša od 0,05. V tabeli 10 lahko vidimo rezultate našega modela.

Tabela 10: Skupna statistična značilnost izbranih kazalnikov

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | P-value |
| Likelihood Ratio | 92,5669 | 5 | <0,0001 |
| Score | 81,1920 | 5 | <0,0001 |
| Wald | 67,0474 | 5 | <0,0001 |

Vse P-vrednosti v tabeli 10 so pod mejo $\alpha = 0,05$, kar nas vodi do zaključka, da lahko ničelno hipotezo zavrnemo, in da je vsaj eden izmed regresijskih koeficientov različen od 0. Lahko rečemo, da so izbrani kazalniki, skupno gledano, statistično značilni.

Rezultati *značilnosti za posamezne kazalnike* in nekatere druge statistike so zbrani v izhodni tabeli z imenom »Analysis of Maximum Likelihood Estimates«. Poleg koeficientov dobimo izračunane še standardno napako, test značilnosti (Waldow hi-kvadrat test ter pripadajoča P-vrednost) in pa standardizirane koeficiente. S pomočjo te tabele izračunamo končne uteži modela, tako da iz predhodno omenjenih 500-tih izbranih vzorcev izračunamo povprečje standardiziranih koeficientov. Rezultati za predstavljen model so prikazani v tabeli 11.

Tabela 11: Statistična značilnost posameznih kazalnikov

| Analysis of Maximum Likelihood Estimate | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | P-value | Standardized Estimate |
| Odziv | 1 | -1,60670 | 0,14590 | 121,2850 | <0,0001 | |
| SSD1 | 1 | -0,00219 | 0,00172 | 1,6061 | 0,2050 | -0,0687 |
| SSD3 | 1 | -0,00498 | 0,00218 | 5,2224 | 0,0223 | -0,1367 |
| FV2 | 1 | -0,01840 | 0,00298 | 37,8476 | <0,0001 | -0,5029 |
| DO5 | 1 | -0,00816 | 0,00321 | 6,4815 | 0,0109 | -0,2134 |
| Sektor | 1 | -0,80230 | 0,26970 | 8,8522 | 0,0029 | -0,1723 |

Vidimo lahko, da so statistično značilni (P-vrednost pod 0,05) vsi kazalniki razen SSD1, ki ima P-vrednost višjo od 0,05. Kot sem že omenila, se pri izboru modela upošteva tudi ekonomska smiselnost modela in je lahko kateri izmed kazalnikov dodan po posvetu z bančnimi analitiki, četudi posamično ne izkazuje napovedne moči.

Če želimo videti *primerjavo med napovedanimi verjetnostmi in opazovanimi odgovori* ter sposobnost razlikovanja med dobrimi in slabimi komitenti, so rezultati prikazani v tabeli z naslovom »Association of Predicted Probabilities and Observed Responses«, ki vsebuje naslednje statistike:

- c $= (n_c + 0,5(t - n_c - n_d))/t$,
- Somer's D $= (n_c - n_d)/t$,
- Goodman-Kruskal Gamma $= (n_c - n_d)/(n_c + n_d)$,
- Kendall's Tau-a $= 2(n_c - n_d)/(N(N-1))$,

kjer je $t$ število vezanih parov z različnimi odgovori, $n_c$ število skladnih parov (angl. concordant), $n_d$ število neskladnih parov (angl. discordant), $t - n_c - n_d$ število parov z enakimi odgovori in $N$ je število opazovanih podatkov. Najprej definiramo vrednost odziva, ki bo pripadala opazovanemu odgovoru 1. Paru rečemo, da je skladen, če imata enoti v paru različen opazovani odgovor, vendar ima enota z nižjo vrednostjo odziva tudi nižjo napovedano verjetnost kot enota z višjo vrednostjo odziva. Par pa je neskladen, če imata enoti različen odgovor in hkrati ima enota z nižjo vrednostjo odziva, višjo napovedano verjetnost od enote z višjo vrednostjo odziva. Parom, ki niso ne skladni in niti neskladni, vseeno pa imajo različna opazovana odziva, pravimo vezani pari (angl. tied).

Parameter $c$ je ekvivalenten meri območja pod ROC krivuljo in njegove vrednosti se gibljejo med 0,5 in 1, kjer 0,5 pomeni, da model naključno napoveduje odgovore in

1 pomeni, da model popolnoma pravilno napoveduje status komitentov. Za modele z binarnim odzivom je vrednost $c$ povezana tudi z mero natančnosti AR in sicer s formulo: $AR = 2 \cdot c - 1$.

Somer's D se uporablja za določanje moči in smeri razmerja med pari spremenljivk. Vrednost te statistike se giblje med $-1$ (noben par se ne ujema) in 1 (vsi pari se ujemajo). Model, ki je bolj sposoben razlikovati med plačniki in neplačniki ima pri prej omenjenih statistikah višje vrednosti. Za modele z binarnim odzivom nam ta statistika predstavlja mero natančnosti AR.

Goodman-Kruskal Gamma statistika ignorira vezane pare in meri moč združevanja med pari. Vrednost te statistike se giblje med $-1$ (vsi pari so napačno združeni) in 1 (pari so popolno združeni). Vrednost 0 bi pomenila, da sta enoti v paru neodvisni.

Kendall's Tau-a statistika je pravzaprav modifikacija statistike Somer's D in je definirana kot razmerje med razliko števila skladnih in neskladnih parov ter številom vseh možnih parov. Običajno je precej manjša od statistike Somer's D, saj pričakujemo, da imamo mnogo parov z enakim odzivom.

Model, ki je boljši v razlikovanju med plačniki in neplačniki ima za zgoraj opisane statistike visoke pozitivne vrednosti. V tabeli 12 so rezultati predstavljenega modela.

Tabela 12: Sposobnost modela pri ločevanju plačnikov od neplačnikov

| Association of Predicted Probabilities and Observed Responses | | | |
|---|---|---|---|
| Percent Concordant | 77,0 | Somers' D | 0,543 |
| Percent Discordant | 22,7 | Gamma | 0,544 |
| Percent Tied | 0,3 | Tau-a | 0,198 |
| Pairs | 52761 | c | 0,771 |

Kot lahko vidimo so vse vrednosti statistik pozitivne in precej visoke, kar pomeni, da model dobro razlikuje med dobrimi in slabimi komitenti.

Med mere napovedne moči spada tudi *klasifikacijska tabela* oziroma *matrika zamenjav* (angl. confusion matrix), ki se uporablja za binarne klasifikacijske probleme. V SAS-u dobimo rezultate matrike v tabeli »Classification Table«. Vsak vhod v tabeli zabeleži število zapisov iz napovedanega izida za opazovani izid. V našem primeru testiramo, kako uspešno model prepozna slabo podjetje. Recimo, da imamo ničelno hipotezo, ki pravi: $H_0$ - *Podjetje je slabo.* Če bomo podjetje prepoznali kot slabo podjetje, bomo to smatrali kot sprejetje hipoteze oziroma kot pozitiven odgovor. Če bomo podjetje ocenili kot dobro, bomo to smatrali kot negativen odgovor (zavrnitev hipoteze). V nadaljevanju lahko vidimo, kako zgleda matrika zamenjav.

Tabela 13: Matrika zamenjav - splošen prikaz

| | | Dejanski status | |
|---|---|---|---|
| | | Slabo podjetje | Dobro podjetje |
| Napovedan status | Slabo podjetje | TP | FP |
| | Dobro podjetje | FN | TN |

Vrednosti v tabeli 13 imajo naslednji pomen:

- TP – true positive: število pravilnih napovedi, da je podjetje slabo.
- TN – true negative: število pravilnih napovedi, da je podjetje dobro.
- FP – false positive: število napačnih napovedi, da je podjetje slabo (napovemo, da je podjetje slabo, ko je v resnici dobro – temu pravimo tudi napaka 1. tipa)
- FN – false negative: število napačnih napovedi, da je podjetje dobro (napovemo, da je podjetje dobro, ko je v resnici slabo – temu pravimo tudi napaka 2. tipa)

Matrika zamenjav nam ponuja informacije potrebne za ocenitev kakovosti našega napovednega modela, ki jih lahko izrazimo z naslednjimi merami:

- Občutljivost (angl. Sensitivity - True Positive rate)= $\frac{TP}{(TP+FN)}$: meri odstotek slabih podjetij, ki smo jih pravilno prepoznali kot slabe.
- Specifičnost (angl. Specificity - True Negative rate)= $\frac{TN}{(FP+TN)}$: meri odstotek dobrih podjetij, ki smo jih pravilno prepoznali kot dobre.
- Stopnja izpadlih (angl. Fall Out - False Positive rate)= $\frac{FP}{(FP+TN)}$: meri odstotek dobrih podjetij, ki smo jih narobe prepoznali kot slabe.
- Stopnja zgrešenih (angl. Miss Rate - False Negative rate)= $\frac{FN}{(FN+TP)}$: meri odstotek slabih podjetij, ki smo jih narobe prepoznali kot dobre.

Želimo si, da sta meri občutljivosti in specifičnosti čim višji, stopnji izpadlih in zgrešenih pa čim nižji. V tabeli 14 je prikazana matrika zamenjav za izbrani model. V tabeli 15 pa lahko vidimo izpis iz SAS-a, kjer dobimo izračune za občutljivost, specifičnost, stopnjo izpadlih in stopnjo zgrešenih napovedi za izbrani model. Kot lahko vidimo sta vrednosti za občutljivost in specifičnost precej visoki, kar pomeni, da model dokaj dobro prepozna tako dobre kot slabe komitente. Tudi mera zgrešenih, ki je precej bolj kritična od mere izpadlih, je nizka, kar pomeni, da nismo prevelikemu številu slabih komitentov pripisali, da so dobri.

Tabela 14: Tabela zamenjav - vrednosti za model

| | | Dejanski status | |
|---|---|---|---|
| | | Slabo podjetje | Dobro podjetje |
| Napovedan status | Slabo podjetje | 68 | 98 |
| | Dobro podjetje | 25 | 193 |

Tabela 15: Izračunane mere uspešnosti iz tabele zamenjav

| Classification Table | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Correct | | Incorrect | | Percentages | | | | |
| Event | Non-Event | Event | Non-Event | Correct | Sensitivity | Specificity | False POS | False NEG |
| 68 | 193 | 98 | 25 | 68,0 | 73,1 | 66,3 | 33,7 | 26,9 |

V kolikor želimo preveriti, kako dobro se *model prilega podatkom*, pa si lahko pogledamo Hosmer-Lemeshowo statistiko, ki se nahaja v tabelah »Partition for the Hosmer and Lemeshow Test« in »Hosmer and Lemeshow Goodness-of-Fit Test«. Hosmer-Lemeshowa statistika meri, kako dobro logit model predstavlja dejanske verjetnosti neplačila za skupine podjetij z različno stopnjo tveganosti. Podatki so združeni v skupine na podlagi percentilov ocenjenih verjetnosti neplačila. Običajno se uporabi 10-procentne intervale, torej dobimo 10 skupin. Nato se za vsako izmed skupin izračuna povprečno ocenjeno verjetnost neplačila in izpelje pričakovano število neplačnikov v skupini. Nato se to število primerja s količino dejanskih neplačnikov v skupini. Hosmer-Lemeshowa statistika nato povzame te informacije za vse skupine. V tabeli 16 so prikazani rezultati tega testa za izbrani model.

Pri Hosmer-Lemeshow statistiki preverjamo ničelno hipotezo, ki predpostavlja, da ni razlike med opazovanimi in napovedanimi vrednostmi odzivne spremenljivke. V tabeli 17 lahko vidimo rezultate testiranja te hipoteze. Kot lahko vidimo je P-vrednost enaka 0,0149, kar pomeni, da hipotezo pri meji $\alpha = 0,05$ zavrnemo in to pomeni, da obstaja razlika med opazovanimi in napovedanimi vrednostmi. Zavedati pa se moramo, da je bil izbrani vzorec res majhen in je to lahko razlog slabših rezultatov tega testa.

Tabela 16: Podatki za Hosmer-Lemeshowo statistiko

| Partition for the Hosmer and Lemeshow Test | | | | | |
|---|---|---|---|---|---|
| Group | Total | status = 1 | | status = 0 | |
| | | Observed | Expected | Observed | Expected |
| 1 | 38 | 4 | 1,16 | 34 | 36,84 |
| 2 | 38 | 1 | 2,01 | 37 | 35,99 |
| 3 | 38 | 1 | 2,99 | 37 | 35,01 |
| 4 | 38 | 1 | 4,52 | 37 | 33,48 |
| 5 | 38 | 8 | 6,91 | 30 | 31,09 |
| 6 | 38 | 9 | 9,27 | 29 | 28,73 |
| 7 | 37 | 12 | 11,53 | 25 | 25,47 |
| 8 | 37 | 12 | 13,78 | 25 | 23,22 |
| 9 | 37 | 16 | 16,94 | 21 | 20,06 |
| 10 | 37 | 23 | 20,76 | 14 | 16,24 |

Tabela 17: Rezultati Hosmer-Lemeshowe statistike

| Hosmer and Lemeshow Goodness-of-Fit Test | | |
|---|---|---|
| Chi-Square | DF | P-value |
| 18,994 | 8 | 0,0149 |

### 3.3.9 Kalibracija modela

Kalibracija modela je potrebna zato, da se izid modela (angl. score) prilagodi stopnji neplačila, ki jo pričakujemo na populaciji, za katero se bo model uporabljal. Ta faza vključuje oceno povprečne stopnje neplačila na izbranem obdobju (angl. central tendency - CT) in izračun končne povezave med izidom modela ter verjetnostjo neplačila (PD).

Da bi razumeli razloge za kalibracijo, moramo upoštevati:

- Model mora biti prilagojen na stopnjo neplačila, ki je predvidena za populacijo, na kateri bomo model uporabljali v prihodnje.
- Stopnja neplačila, s katero bomo kalibrirali, je običajno različna od tiste, ki jo je imel razvojni vzorec, s katerim smo ocenjevali parametre.
- Stopnja neplačila, na kateri bomo kalibrirali, mora biti dodeljena glede na pretekle podatke.

Glede na to, da je bil izbor končnega modela narejen na vzorcu, kjer smo določili pretirano stopnjo neplačila (25%), je bilo potrebno kalibrirati formulo glede na dejansko

stopnjo neplačila.

Za kalibracijo potrebujemo zagotoviti naslednje tri korake:

- definicija povprečja stopnje neplačila (central tendency),
- definicija koeficientov ($\delta_0$ in $\delta_1$) kalibracijske krivulje razvite na podlagi povprečja stopnje neplačila in
- izračun verjetnosti neplačila (PD) po formuli:

$$PD = \frac{1}{1 + e^{\delta_0 + \delta_1 \cdot x}},$$

kjer je $x$ izid (score), ki nam ga ponudi model.

Najprej je potrebno izračunati povprečje stopnje neplačila (CT), ki je bilo že predhodno prikazano v tabeli 2, kjer vidimo, da je CT=$5,83\%$. Nato je potrebno poiskati povezavo med izidom linearne kombinacije transformiranih kazalnikov in verjetnostjo neplačila. Kalibracijski proces se torej osredotoča na konstrukcijo krivulje, ki povezuje verjetnost nastanka neplačila z izidom modela ob upoštevanju vrednosti CT.

Da to dosežemo, ustvarimo 10 razredov, ki vsebujejo enako število komitentov glede na vrednost izida, ki nam ga je dal model (od najnižjega do najvišjega). Komitenti z nižjimi vrednostmi so vključeni v prvi razred, komitenti z najvišjimi vrednostmi pa v deseti razred. Za vsak razred najprej, na podlagi števila plačnikov in neplačnikov, izračunamo stopnjo neplačila. Nato pa izračunamo še stopnjo neplačila s pomočjo Bayesove formule, ob upoštevanju stopnje neplačila na razvojnem vzorcu ($5,62\%$) in vrednosti CT ($5,83\%$). Povprečno pričakovano verjetnost neplačila (PD) za razred $i$ pa izračunamo z uporabo srednje vrednosti izida modela ($x_i$) za obravnavan razred in parametrov $\delta_0$ ter $\delta_1$, s formulo $PD = 1/(1 + e^{\delta_0 + \delta_1 \cdot x_i})$. Vrednosti $\delta_0$ in $\delta_1$ prilagodimo tako, da bo povprečje verjetnosti neplačila, izračunano na vseh komitentih iz razvojnega vzorca, enako CT in da minimiziramo razliko med opazovano stopnjo neplačila in izračunano pričakovano verjetnostjo neplačila za posamezni razred.

V tabeli 18 je predstavljen končni rezultat kalibracije, kjer dobimo izračunane parametre za izračun verjetnosti neplačila PD.

Končna formula s katero iz izida modela ($x$) dobimo verjetnosti neplačila je naslednja:

$$PD = \frac{1}{1 + e^{-0,0354 - 0,8792 \cdot x}}.$$

S tem smo tudi prišli do želenega cilja, in sicer do formule, ki omogoča izračun verjetnosti neplačila za posameznega komitenta banke. S tem korakom se razvoj samega modela zaključi. V nadaljevanju bo sledil še sklep na podlagi modela ter kratek opis uporabe v praksi.

Tabela 18: Rezultati kalibracije in parametri krivulje

| Parametri krivulje | |
|---|---|
| $\delta_0$ | -0,0354 |
| $\delta_1$ | -0,8792 |
| PD povpr. | 5,83% |
| AR | 68,30% |

## 3.4 Sklepi o modelu in njegova uporaba v praksi

V predhodnih poglavjih smo si podrobneje pogledali gradnjo finančnega modela v banki. Finančni model je le del celotnega bonitetnega modela za ocenjevanje verjetnosti dogodka neplačila. Banka poleg finančnih kazalnikov uporablja še vedenjske kazalnike, ki povzemajo dolžnikovo vedenje pri poslovanju z banko (gleda se disciplina odplačevanja, zamujanje z odplačili, višina neplačanih dolgov v primerjavi z višino kredita in podobno). Da ocena komitenta ne temelji le na statističnih rezultatih, pa banka lahko uporabi še vprašalnik, na katerega odgovarja skrbnik podjetja in s subjektivnimi informacijami vpliva na končno oceno tveganja (lahko ostane ista ali pa se zviša/zniža). Z združitvijo vseh omenjenih komponent se napovedna moč modela zviša, kar pomeni, da lahko boljše napovedujemo, kateri izmed komitentov bo postal neplačnik.

Za potrebe analiz in poročanja si banka ustvari svojo lestvico bonitetnih razredov. Vsak bonitetni razred ima določeno zgornjo mejo in spodnjo mejo verjetnosti neplačila ter srednjo vrednost verjetnosti neplačila, ki se jo za ta razred upošteva. Glede na verjetnost, ki nam jo vrne model, se nato določi v kateri bonitetni razred spada komitent.

Banka rezultate bonitetnega modela uporablja za pomoč pri kreditiranju komitentov, izračunu rezervacij, raznih analizah bančnega poslovanja, poročanju in podobnih procesih, s katerimi lahko uspešno obvladuje kreditna tveganja.

Pri opisanem modelu obstajajo določene pomanjkljivosti in omejitve, ki jih lahko banka v naslednjih verzijah odpravi. Ena izmed pomanjkljivosti, ki se jo bo relativno lahko odpravilo, je dolžina časovne vrste. Po zahtevah Basla II naj bi bila vsaj petletna, uporabljena pa je bila štiriletna, zaradi skromne razpoložljivosti baze podatkov. Prav tako v trenutni verziji modela nekaterih kazalnikov v model nismo mogli vključiti, zaradi prevelikega števila manjkajočih podatkov. Zato je za banko zelo pomembno, da ima zgrajene dobre baze podatkov, ki so čim bolj pravilno in v celoti zapolnjene. Da bo model skozi leta ohranjal zadovoljivo napovedno moč, je potrebna redna validacija in kalibracija modela.

# 4 Zaključek

V magistrski nalogi smo najprej preučili kategorične modele za binarne odzive, od katerih je bil podrobneje predstavljen model logistične regresije. Spoznali smo, kaj moramo upoštevati pri izbiri napovednih spremenljivk ter kako testiramo parametre modela. Podroben opis omenjenih metod nam je omogočal lažje razumevanje nadaljevanja magistrske naloge, ki povzema primer izgradnje modela za merjenje tveganja v banki.

Banka je ob izdaji kredita izpostavljena tveganju finančne izgube, zaradi česar je pomembno, da učinkovito upravlja s kreditnim tveganjem. Uvedba kapitalskega sporazuma Basel II, je bankam omogočila nove možnosti za učinkovitejše obvladovanje kreditnega tveganja. V tem delu magistrske naloge smo podrobneje spoznali osnovne pojme povezane z bančnim kreditiranjem in pa predstavitev IRB pristopa.

V zadnjem delu magistrske naloge je sledila predstavitev razvoja bonitetnega modela na portfelju velikih podjetij v banki. Prikazan je bil princip izgradnje modela v banki, ki poteka v naslednjih korakih:

- najprej si izberemo populacijo na kateri bomo model razvijali;
- opredelimo dogodek neplačila, ki ga bomo preučevali;
- opredelimo razvojni vzorec iz populacije, na katerem bomo model razvili;
- zberemo vse razpoložljive podatke, ki jih imamo na voljo za komitente banke;
- podatke obdelamo in prečistimo;
- iz danih podatkov sestavimo seznam kazalnikov, ki jih bomo testirali za model;
- na podatkih izvedemo univariatno analizo;
- na podlagi analize izberemo najučinkovitejše kazalnike, ki jih nato transformiramo in normaliziramo;
- s pomočjo metode postopne izbire poiščemo model z največjo napovedno močjo;
- ocenimo učinkovitost končnega modela;
- končni model kalibriramo, da dobimo formulo za izračun verjetnosti neplačila.

Čeprav celoten proces zbiranja in obdelave podatkov, gradnje in testiranja modela zahteva kar nekaj časa in truda, naša naloga ni zaključena, ko implementiramo dobljen bonitetni model. Napovedna moč vseh statističnih modelov sloni na predpostavki, da bo pretekla zveza med spremenljivkami modela in stanjem neplačila ostala nespremenjena tudi v prihodnosti. Glede na širok razpon možnih dogodkov, kot so spremembe računovodskih politik v podjetjih ali strukturne spremembe v nekaterih sektorjih, prej omenjene predpostavke ne veljajo za daljše časovno obdobje. Zato je potrebna redna validacija in kalibracija modela, da se njegova napovedna moč ne zmanjšuje.

# 5 Literatura

[1] A. Agresti, *Categorical Data Analysis, Second Edition*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2002.

[2] P. D. Allison, Measures of fit for logistic regression. *SAS Global Forum, Washington, DC*, Paper 1485-2014, 2014. http://support.sas.com/resources/papers/proceedings14/1485-2014.pdf (Datum ogleda: 24. 5. 2015).

[3] D. Collett, *Modelling binary data*, Chapman & Hall, London, 1991.

[4] A. Čargo in M. Štajner, *Minimalne zahteve za uvedbo IRB pristopa.* Banka Slovenije, 2004.

[5] N. Dhand, *Tutorial on Logistic Regression Model Building*, Faculty of Veterinary Science, The University of Sydney, Camden, 2010. http://sydney.edu.au/vetscience/biostat/macros/logistic_tut_intro.shtml (Datum ogleda: 24. 5. 2015).

[6] B. Engelmann, E. Hayden in D. Tasche, Measuring the discriminative power of rating systems. *Discussion paper No. 01/2003, Banking and Financial Supervision, Deutsche Bundesbank*, 2003.

[7] D. W. Hosmer in S. Lemeshow, *Applied Logistic Regression, Second Edition.* John Wiley & Sons, Inc., New York, 2000.

[8] M. Jovan in M. Šušteršič, *Statistično ocenjevanje verjetnosti neplačila za slovenska podjetja.* 10. strokovno posvetovanje o bančništvu: Novi bančni standardi in ERM 2, Zveza ekonomistov Slovenije, Ljubljana, 2004.

[9] P. Karpe, Klasična tveganja bančnega poslovanja - kreditno tveganje. *Bančni vestnik* 4 (1997) 36–38.

[10] B. Kerbler, Modeli diskretne izbire. *Urbani izziv* letnik 17, št. 1–2 (2006) 134–138.

[11] K. Košmelj, *Uporabna statistika, druga dopolnjena izdaja*, Biotehniška fakulteta, Ljubljana, 2007.

[12] E. G. Falkenstein, A. Boral, L. V. Carty, *RiskCalc for Private Companies: Moody's Default Model*, https://riskcalc.moodysrms.com/us/research/crm/56402.pdf. (Datum ogleda: 18. 8. 2015.)

[13] M. Raič, *Statistika - zapiski s predavanj*, http://valjhun.fmf.uni-lj.si/ raicm/Vaje/BPSt/Statistika.pdf. (Datum ogleda: 18. 8. 2015.)

[14] SAS Institute Inc., The LOGISTIC Procedure. *SAS/STAT User's Guide* Version 8, Chapter 39 (1999) 1903–2042.

[15] SAS Institute Inc., The VARCLUS Procedure. *SAS/STAT User's Guide* Version 8, Chapter 68 (1999) 3593–3620.

[16] A. Saunders in M. M. Cornett, *Financial institutions management: a risk management approach, Sixth Edition.* McGraw-Hill/Irwin, New York, 2008.

[17] *Sklep o izračunu kapitalske zahteve za kreditno tveganje po pristopu na podlagi notranjih bonitetnih sistemov za banke in hranilnice (Uradni list RS, št. 135/06 z dne 21. 12. 2006)*, http://www.pisrs.si/Pis.web/pregledPredpisa?id=SKLE6797. (Datum ogleda: 24. 5. 2015.)

[18] *Zakon o bančništvu – ZBan-1 (Uradni list RS, št. 131/06 z dne 14. 12. 2006)*, http://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO4300. (Datum ogleda: 24. 5. 2015.)