**VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE**

FAKULTA INFORMATIKY A STATISTIKY


Název bakalářské práce:

# Odhad pravěpodobnosti defaultu pomocí logistické regrese

Autor: **Pavel Jiřičko**

Katedra: **Katedra ekonometrie**

Obor: **Matematické metody v ekonomii**

Vedoucí práce: **Ing. Zuzana Dlouhá, Ph.D.**

**Prohlášení**:

Prohlašuji, že jsem bakalářskou práci na téma „Odhad pravděpodobnosti defaultu pomocí logistické regrese“ zpracoval samostatně. Veškerou použitou literaturu a další podkladové materiály uvádím v seznamu použité literatury.

V Praze dne 14. května 2017 .....

Pavel Jiřičko

**Poděkování**:

Rád bych na tomto místě poděkoval Ing. Zuzaně Dlouhé, Ph.D. za vedení mé bakalářské práce a za podnětné návrhy, které ji obohatily. Také bych chtěl poděkovat za dodání dat nezbytných k praktické části této práce.

## Abstrakt

*Název práce*: Odhad pravděpodobnosti defaultu pomocí logistické regrese
*Autor*: Pavel Jiřičko
*Katedra*: Katedra ekonometrie
*Vedoucí práce*: Ing. Zuzana Dlouhá, Ph.D.

Cílem práce je sestavit model pro odhad pravděpodobnosti nesplacení úvěru klientem. Nejprve byly vybrány vhodné vysvětlující proměnné, které byly v práci podrobně popsány. U kvalitativních proměnných byly v jednotlivých kategoriích zkoumány veličiny jako míra selhání nebo skóre. U kvantitativních proměnných byly spočítány veličiny jako průměr nebo směrodatná odchylka. Samotný odhad modelu byl proveden metodou logistické regrese. U odhadnutých koeficientů jednotlivých vysvětlujících proměnných byly zkoumány jejich statistické významnosti. V práci byl také popsán vliv jednotlivých vysvětlujících proměnných na pravděpodobnost defaultu, a ta byla také spočítána pro dva klienty s konkrétními hodnotami vysvětlujících proměnných. Podařilo se tedy sestavit model, který je možné použít k tvorbě predikcí o pravděpodobnosti defaultu klienta.



**Klíčová slova:** pravděpodobnost defaultu, skóre, logistická regrese



## Abstract

*Title*: The Estimation of Probability of Default Using Logistic Regression
*Author*: Pavel Jiřičko
*Department*: Department of Econometrics
*Supervisor*: Ing. Zuzana Dlouhá, Ph.D.

The aim of the thesis was to build a probability prediction model for client loan repayment. First, the author selected suitable explanatory variables and explained these in detail in the thesis. The author examined various quantities both for qualitative variables (e.g. failure rates and scores) and quantitative variables (e.g. means and standard deviations) describing individual categories. The predictions were done by the use of logistic regression. The estimated coefficients of individual explanatory variables were examined in terms of their statistical significance. The author also described the impact of individual explanatory variables on the probability of a client's default. The probabilities were calculated for two clients with specific values of explanatory variables. The author managed to build a model that can be used to predict probability of client loan repayment defaults.



**Keywords**: propability of default, score, logistic regression

# Obsah

# 1. Úvod

Poskytování úvěrů je v dnešní době zcela nedílnou součástí dobře fungující ekonomiky. S nějakou formou úvěru se během života setká téměř každý. Může se jednat například i o situaci, kdy Vám kolega v práci půjčí peníze na svačinu, protože jste si zrovna zapomněli doma peněženku. V takovém případě mu nejspíše bude stačit Vaše slovo, že mu peníze zítra vrátíte. Pravděpodobně se nebude snažit uzavřít s Vámi nějakou smlouvu, nebo dokonce se zabývat tím, s jakou pravděpodobností mu zapůjčené peníze druhý den vrátíte. Přesto celý tento proces ovlivní jeho budoucí rozhodování v podobné situaci. Pokud mu peníze druhý den bez problémů vrátíte, nebude mít žádný problém Vám je půjčit znovu, jestliže je budete potřebovat. Dojde-li však k tomu, že mu peníze dlouhou dobu vracet nebudete, nebo je dokonce vůbec nevrátíte, pak se dá předpokládat, že Vám již nebude chtít žádné půjčit.

V našem případě se Váš kolega rozhodoval na základě předchozích zkušeností. Protože se však jednalo o částku nepatrné hodnoty, neměl potřebu podrobněji zkoumat Vaši finanční situaci. Odlišný případ nastane, pokud budete potřebovat půjčit větší finanční obnos. Rozhodnete-li se oslovit svého kolegu s tím, jestli by Vám nepůjčil například 100 000 Kč, bude již požadovat písemnou smlouvu a nejspíše si také nechá nějaký čas na rozmyšlení. Bude se také podrobněji zajímat o Vaši finanční situaci. Bude se snažit určit, jaká bude Vaše schopnost dostát svým závazkům a dluh ve stanoveném termínu splatit.

V této situaci si již Váš kolega bude zjišťovat o Vás některé základní informace, aby si mohl udělat obrázek o Vaší finanční situaci. Pokud však nenajdete ve svém okolí žádného kolegu nebo kamaráda, který by Vám byl ochotný půjčit 100 000 Kč, nezbývá Vám nic jiného než o úvěr požádat v bance, nebo v některé z nebankovních finančních institucí. Také zde se budou snažit zjistit, jaká je Vaše schopnost dostát svým závazkům. Na rozdíl od Vašeho kolegy nebo kamaráda se ji však budou snažit kvantifikovat. Budou po Vás požadovat spoustu informací, na základě kterých se budou snažit vyčíslit s jakou pravděpodobností svůj dluh splatíte, nebo naopak nesplatíte. Případu, kdy klient svůj dluh nesplatí, budeme říkat že zdefaultuje.

A přesně tato kvantifikace platební schopnosti klienta je součástí této práce. Budeme se zabývat odhadem pravděpodobnosti, že klient zdefaultuje na základě informací, které o klientovi máme. Nástrojem, který budeme pro odhad této pravděpodobnosti používat, bude logistická regrese.

# 2. Logistická regrese

## 2.1 Logistická křivka

Pro získání tvaru logistické křivky vyjdeme z diferenciální rovnice

$$y' = ay(s - y)$$

Kdybychom měli pouze model $y' = ay$ jednalo by se o rovnici exponenciálního růstu. V tomto případě zde však máme ještě člen v závorce $s - y$, který představuje jakousi „brzdu“. Samotnému členu $s$ říkáme hladina saturace, nebo také hladina nasycení. Už nyní si tedy můžeme udělat představu o tom, jak nejspíše bude logistická křivka vypadat. Do určitého bodu bude konvexní a bude připomínat exponenciální funkci. Člen $s - y$ však způsobí, že se křivka v jednom konkrétním bodě změní na konkávní a bude se limitně blížit k hodnotě $s$, tedy k hladině nasycení. Takovému bodu říkáme inflexní bod.

Abychom získali rovnici pro tvar logistické křivky, budeme potřebovat tuto diferenciální rovnici vyřešit. Ještě předtím však zavedeme předpoklad o hodnotách $y$ a $s$, že $0 < y < s$.

Pro výpočet použijeme metodu separace proměnných. Dostaneme tedy

$$\frac{dy}{dt} = ay(s - y)$$

Po úpravách dostaneme

$$\int \frac{1}{y(s-y)}\, dy = \int (a) dt$$

Vidíme, že na pravé straně rovnice dostaneme po integraci výraz $at + k$, kde k představuje integrační konstantu.

Integrál na pravé straně je o něco složitější. Pro jeho řešení použijeme několik „triků“. Nejprve vnitřek integrálu vynásobíme výrazem $\frac{s}{s}$ , tedy vlastně jedničkou, a část $\frac{1}{s}$ vytkneme před integrál. Dostáváme tedy

$$\frac{1}{s} \int \frac{s}{y(s-y)}\, dy$$

Nyní v čitateli odečteme a přičteme $y$, čímž hodnotu čitatele nijak nezměníme.

$$\frac{1}{s} \int \frac{s-y+y}{y(s-y)}\, dy$$

Tento integrál můžeme rozložit na součet dvou integrálů

$$\frac{1}{s} \left[ \int \frac{s-y}{y(s-y)}\, dy + \int \frac{y}{y(s-y)}\, dy \right]$$

Po zkrácení dostaneme

$$\frac{1}{s}\left[\int \frac{1}{y}dy + \int \frac{1}{s-y}dy\right]$$

Tyto výrazy již můžeme snadno zintegrovat

$$\frac{1}{s}(ln|y| - ln|s-y|)$$

Vyjdeme-li nyní z předpokladu, který jsme stanovili na začátku $0 < y < s$, můžeme snadnými úpravami dostat tvar pro levou stranu naší původní rovnice

$$\frac{1}{s}\ln(\frac{y}{s-y})$$

Naší původní rovnici, tak nyní máme ve tvaru

$$\frac{1}{s}\ln\left(\frac{y}{s-y}\right) = at + k$$

Tuto rovnici vynásobíme výrazem $s$ a následně na celou rovnici použijeme exponenciálu. Dostaneme tedy

$$\frac{y}{s-y} = e^{sat} + sk$$

V tuto chvíli celou rovnici vynásobíme výrazem $s - y$ a členy obsahující $y$ převedeme na levou stranu, kde $y$ následně vytkneme

$$y(1 + e^{sat}) = se^{sat} + sk$$

Celou rovnici teď vydělíme výrazem $1 + e^{sat}$ a vzniklý zlomek na pravé straně usměrníme výrazem $\frac{e^{-(sat+sk)}}{e^{-(sat+sk)}}$ a dostaneme rovnici ve tvaru

$$y = \frac{s}{1+e^{-sat-sk}}$$

Tento tvar již velmi připomíná tvar logistické křivky. Nyní již pouze reparametrizujeme tento model. Zavedeme nové parametry $b$ a $c$, které definujeme následovně

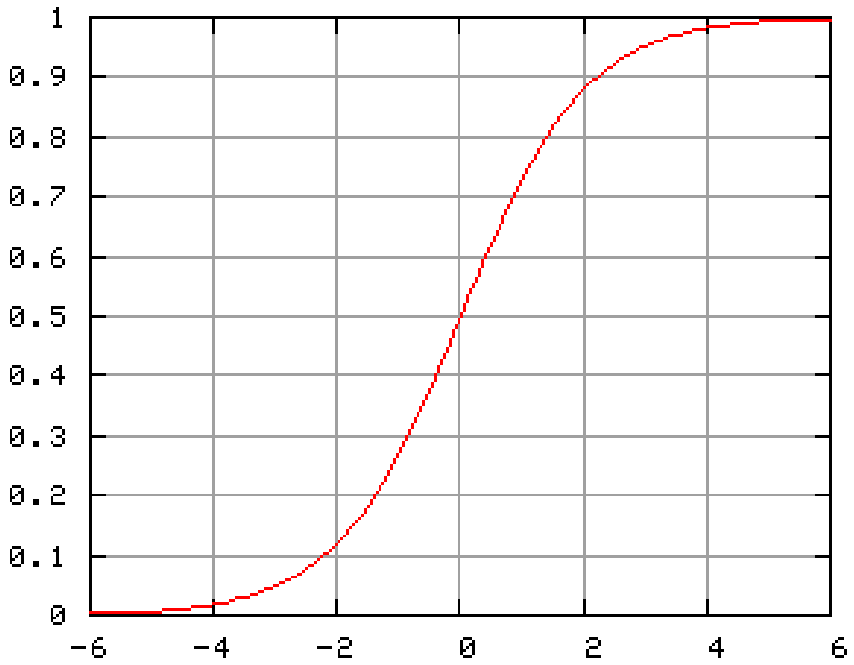
$$b = sa$$

$$c = -\frac{k}{a}$$

Po reparametrizaci již dostáváme rovnici pro logistickou křivku

$$\boxed{y(t) = \frac{s}{1 + e^{-b(t-c)}}}$$

Kde $t$ v závorce nám říká, že $y$ je funkcí času.[1]




OBRÁZEK 2.1 LOGISTICKÁ KŘIVKA[2]

## 2.2 Model logistické regrese

V předchozí části jsme si odvodili tvar logistické křivky. Nyní se budeme zabývat již samotnou logistickou regresí. Podobně jako v lineární regresi, tak i v logistické regresi budeme mít nějakou vysvětlovanou proměnnou a několik obecně $k$ vysvětlujících proměnných, kterým budeme říkat regresory. Data však nebudeme prokládat přímkou jako u lineární regrese, nýbrž logistickou funkcí ve tvaru

$$Y_i = \frac{s}{1+e^{-(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}$$

kde $Y_i$ je vysvětlovaná proměnná, vektor $[b_0\ b_1 \dots\ b_k]^T$ je vektorem neznámých parametrů, které se budeme snažit odhadnout a vektor $[1\ x_1\ \ x_2 \dots\ x_k\ ]$ představuje vektor regresorů. První složka tohoto vektoru je rovna jedné, protože parametr $b_0$ není násoben žádným regresorem.

Logistickou regresi můžeme použít v různých situacích, kdy se snažíme něco odhadovat. Například nás může zajímat pravděpodobnost výskytu nějaké nemoci v závislosti na různých faktorech. V nejjednodušším případě si můžeme představit, že zkoumáme pravděpodobnost výskytu dané nemoci u muže a u ženy. V takovém případě jednu kategorii zvolíme za referenční a použijeme dummy proměnnou. Reálně budeme samozřejmě mít vysvětlujících proměnných mnohem více.

My se budeme zabývat problémem odhadu pravděpodobnosti, že klient který si u nás chce půjčit peníze zdefaultuje. Jak jsme si již řekli v úvodu, pokud si půjdete půjčit peníze od nějaké finanční instituce, bude po Vás požadovat mnoho informací na základě kterých se

bude snažit kvantifikovat Vaši platební schopnost a rozhodnout se tak, zda Vám peníze půjčí nebo ne. My jako nástroj použijeme logistickou regresi. Logistickou funkci jsme si již ukázali. Jelikož budeme odhadovat pravděpodobnost, kterou budeme značit písmenem $P$, tak je zřejmé, že naší horní hranicí bude hodnota 1. Logistickou funkci tak budeme používat ve tvaru

$$Y_i = \frac{1}{1+e^{-(b_0+b_1x_1+\cdots+b_kx_k)}}$$

Pravděpodobnostní model má pak tvar

$$P\big(Y_i = 0|x_{i,1}, x_{i,2}, \dots, x_{i,k}\big) = F(b_0 + x_{i,1}b_0 + \cdots + x_{i,k}b_k)$$

Nebo ekvivalentně

$$P\big(Y_i = 1|x_{i,1}, x_{i,2}, \dots, x_{i,k}\big) = 1 - F(b_0 + x_{i,1}b_0 + \cdots + x_{i,k}b_k)$$

Kde $F$ představuje vhodnou pravděpodobnostní distribuční funkci. Nejčastěji se za tuto funkci volí funkce logistického nebo normálního rozdělení. [6]

## 2.3 Odhady parametrů

Stejně jako u lineární regrese je našim cílem odhadnout hodnoty parametrů modelu $b_{0,}, b_1, \dots, b_k$. K tomuto odhadu se nejčastěji používají 2 metody. Metoda maximální věrohodnosti a metoda nejmenších čtverců.

Jednou z možných metod je metoda nejmenších čtverců, která spočívá v minimalizaci čtverců odchylek skutečných hodnot (získaných z dat) od vyrovnaných hodnot (odhadnutých modelem). Chceme tedy minimalizovat funkci

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Vzhledem k tomu, že naše vysvětlovaná proměnná je binární, tak není metoda nejmenších čtverců úplně vhodná, jelikož může poskytovat zkreslené odhady. Proto použijeme metodu maximální věrohodnosti, která spočívá v nalezení maxima věrohodnostní funkce, jejíž tvar pro *i*-té pozorování je

$$P(Y_i = y_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$

Kde $\pi(x_i)$ značí pravděpodobnost, že vysvětlovaná proměnná $Y_i$ nabyde hodnoty 1. V našem případě tedy značí pravděpodobnost, že klient nezdefaultuje.

Budeme-li předpokládat nezávislost jednotlivých pozorování, můžeme věrohodnostní funkci přepsat ve tvaru

$$l(b) = \prod_{i=1}^{n}\pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$

Jak již bylo zmíněno, metoda maximální věrohodnosti spočívá v nalezení maxima věrohodnostní funkce. Maximum funkce hledáme tak, že všechny parciální derivace (v našem případě podle parametrů modelu $b_i$ $(i = 1 \dots . k)$) položíme rovny 0. Věrohodnostní funkce ve

tvaru, ve kterém jsme si ji uvedli, však není pro derivování vhodná, proto využijeme toho, že maximum zlogaritmované funkce je shodné s maximem původní funkce a naši věrohodnostní funkci zlogaritmujeme. Dostaneme tedy tvar

$$\log\big(l(b)\big) = \log(\prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i})$$

A tento tvar můžeme dále přepsat

$$\log\big(l(b)\big) = \sum_{i=1}^{n}(y_i \log(\pi(x_i)) + (1-y_i)\log(1-\pi(x_i)))$$

Funkci v tomto tvaru již je možné derivovat podle všech parametrů modelu $b_i$. Dostaneme soustavu rovnic, které říkáme soustava věrohodnostních rovnic.

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

$$\sum_{i=1}^{n}\Big(x_{ij}\big(y_i - \pi(x_i)\big)\Big) = 0$$

Vyřešit tuto soustavu ručně není lehké, proto se k nalezení jejího řešení používá některý ze statistických softwarů. Řešením získáme maximálně věrohodný odhad parametrů $b$. [3]

Dále nás také může zajímat intervalový odhad parametrů $b$. Ten nám říká, v jakém intervalu se bude nacházet konkrétní parametr $b_i$ s určitou námi zvolenou pravděpodobností. Na námi zvolené hladině významnosti α pak dostaneme intervalový odhad pro parametr $b_i$ následovně

$$P\left(\beta_i - u_{1-\frac{\alpha}{2}}Se(b_i) < b_i < \beta_i + u_{1-\frac{\alpha}{2}}Se(b_i)\right) = 1-\alpha$$

kde $\beta_i$ představuje bodový odhad parametru $b_i$ a $u_{1-\frac{\alpha}{2}}$ představuje příslušný kvantil normálního rozdělení. [4]

## 2.4 Testování parametrů modelu

Z lineární regrese známe 2 významné testy pro testování parametrů modelu. Jedním z nich je t-test, pomocí něhož zkoumáme statistickou významnost jednotlivých parametrů a druhým je celkový F-test, pomocí něhož zkoumáme významnost modelu jako celku. Podobně tomu bude i u logistické regrese. K testování významnosti jednotlivých parametrů použijeme Waldův test, který odpovídá t-testu, jak ho známe z lineární regrese. Testové kritérium které budeme značit $W$ vypočítáme jako

$$W = \frac{\widehat{b_i}}{\widehat{Se(b_i)}}$$

Nulová hypotéza $b_i = 0$ nám říká, že parametr $b_i$ je nevýznamný a vysvětlovanou proměnnou v podstatě neovlivňuje. Hodnotu testového kritéria $W$ porovnáváme s $1-\frac{\alpha}{2}$ kvantilem normálního rozdělení. Hladinu významnosti α si sami zvolíme podle potřeby. [4]

K testování významnosti modelu jako celku jsme u lineární regrese používali celkový F-test o modelu. U logistické regrese vyjdeme ze zlogaritmované věrohodnostní funkce a dále budeme definovat takzvaný saturovaný model. Tento model obsahuje takový počet parametrů, aby s pravděpodobností rovné jedné vystihoval pozorovaná data. Následně definujeme ještě funkci $D$ jako

$$D = -2\log\left(\frac{l(b)}{l_s(b_s)}\right)$$

Kde $l(b)$ je původní věrohodnostní funkce a $l_s(b_s)$ představuje věrohodnostní funkci saturovaného modelu. Ta je v našem případě rovna jedné a proto můžeme funkci $D$ přepsat ve tvaru

$$D = -2\log(\prod_{i=1}^{n} \hat{\pi}(x_i)^{y_i}(1-\hat{\pi}(x_i))^{1-y_i})$$

a tento tvar můžeme dále ještě upravit

$$D = -2\sum_{i=1}^{n}\left(y_i \log\left(\hat{\pi}(x_i)\right) + (1-y_i)\log\left(1-\hat{\pi}(x_i)\right)\right)$$

Nyní když položíme všech $k$ parametrů rovno nule, budeme porovnávat hodnoty původního neomezeného modelu s omezeným modelem, který má všechny parametry nulové. K tomu použijeme charakteristiku $G$ jako

$$G = -2\log(\frac{k_r(b_r)}{k_u(b_u)})$$

Nulová hypotéza nám tedy nyní říká, že všechny parametry modelu jsou nevýznamné (rovny 0), což můžeme zjednodušeně interpretovat tak, že námi odhadnutý model je velmi nevhodný pro tvorbu jakýchkoliv předpovědí. $G$ se řídí chí kvadrát rozdělením s $k$ stupni volnosti. Je však nutné podotknout, že nulovou hypotézu budeme zamítat v případě, že alespoň jeden parametr $b$ je nenulový. Což však rozhodně nemusí nutně znamenat, že zvolený model je vhodný pro tvorbu předpovědí. Vždy je dobré ještě provést i Waldův test o významnosti jednotlivých parametrů.

Tento test je možné použít také v případě, kdy chceme testovat významnost jen určité části modelu. Obecně testujeme významnost $l$ vysvětlujících proměnných. Pak bude omezený model mít nulových jen právě $l$ parametrů, které testujeme. [3]

## 2.5 Vysvětlující proměnné modelu

U logistické regrese stejně jako u lineární regrese se snažíme vysvětlovanou proměnnou vysvětlit pomocí takzvaných vysvětlujících proměnných. Vysvětlující proměnná může být například značka vozu a můžeme se pomocí této proměnné snažit odhadnout jeho cenu. Zkoumáme tedy závislost ceny na značce vozu.

Obecně je možné vysvětlující proměnné rozdělit do dvou základních skupin. Kvantitativní a kvalitativní vysvětlující proměnné. Kvantitativní vysvětlující proměnné vyjadřují množství, velikost, míru a podobně. Příkladem kvantitativní proměnné mohou být pracovní zkušenosti člověka vyjádřené v letech a můžeme zkoumat, jak tato proměnná ovlivňuje mzdu. Nebo se můžeme snažit vyjádřit cenu vozu v závislosti na jeho stáří v letech. Stáří vozu bude opět kvantitativní proměnná. Příkladů bychom jistě našli celou řadu.

Naproti tomu kvalitativní proměnná značí nejčastěji nějakou kategorii. Příkladem kvalitativní proměnné jsou medaile, kde máme kategorie zlatá, stříbrná a bronzová. Dalším příkladem může být typ vlastnictví bytu nebo místo trvalého pobytu. Kvalitativní proměnné se pak dále dělí na nominální a ordinální. Liší se v tom, zda je možné jednotlivé kategorie logicky uspořádat. Náš příklad s medailemi by byla proměnná ordinální, jelikož medaile je možné logicky uspořádat od bronzové po zlatou. Naopak nominální proměnnou není možné logicky uspořádat. Jako příklad můžeme uvést barvu očí s kategoriemi modrá, hnědá a zelená nebo výše zmíněný typ vlastnictví bytu. Opět bychom našli celou řadu příkladů jak pro ordinální, tak pro nominální kvalitativní proměnné. [5]

Kvalitativní proměnné zařazujeme do modelu nejčastěji v podobě takzvaných dummy proměnných. To jsou proměnné, které nabývají pouze hodnot 0 nebo 1. Pokud bychom například odhadovali cenu vozu v závislosti na značce a měli bychom kategorie Octavia, Fabia a Superb, tak by náš model v lineárním tvaru vypadal následovně

$$y(cena) = b_0 + b_1 x_1 + b_2 x_2$$

kde $x_1$ představuje model Octavia a $x_2$ model Fabia. Model Superb by v modelu vůbec nebyl. Takové proměnné pak říkáme referenční kategorie a víme, že se jedná o model Superb v případě, kdy proměnné $x_1$ i $x_2$ jsou nulové.

U kvantitativních proměnných to bývá jednodušší, jelikož většinou dosazujeme do modelu přesně ty hodnoty, pro které se snažíme odhadnout hodnotu vysvětlované proměnné. Takže jako úplně nejjednodušší příklad si můžeme uvést situaci, kdy se snažíme cenu vozu odhadnout v závislosti na jeho stáří.

$$y(cena) = b_0 + b_1 x_1$$

Kde $x_1$ představuje stáří vozu.

I takovouto evidentně kvantitativní proměnnou je však možné rozdělit do kategorií. V našem případě například do kategorií stáří vozu 0-5 let, 5-10 let, 10-15 a 15 a více. Takto jednoduše můžeme z kvantitativní proměnné vytvořit proměnnou kvalitativní. A opět bychom takový model odhadovali pomocí dummy proměnných.

V reálných modelech je samozřejmě proměnných celá řada a jsou jak kvalitativní, tak kvantitativní. Opět si můžeme uvést jako příklad cenu vozu. Tentokrát ji však budeme odhadovat jak pomocí stáří vozu, tak pomocí modelu.

$$y(cena) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Kde $x_1$ představuje stáří vozu, $x_2$ představuje model Octavia a $x_3$ představuje model Fabia. Model Superb jsme zvolili za referenční kategorii a bude se jednat o Superb v případě, kdy proměnné $x_1$ i $x_2$ budou nulové.

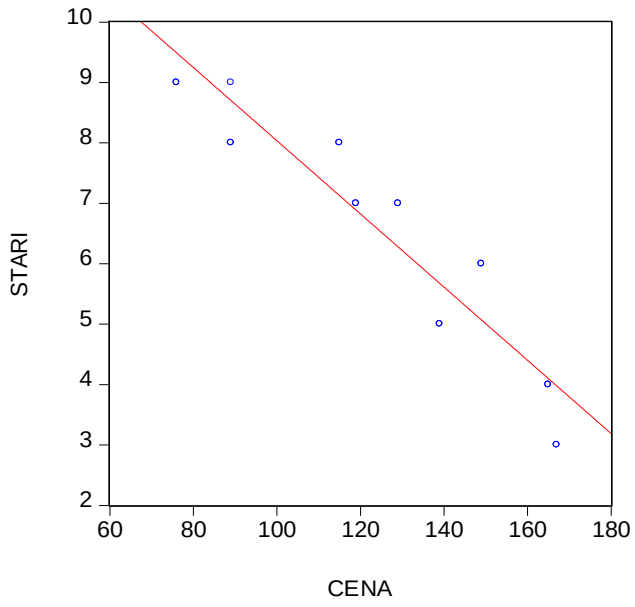
# 3. Diverzifikační schopnost modelu

## 3.1 Diverzifikační schopnost modelu

Našim cílem při odhadování pravděpodobnosti defaultu je oddělení klientů u kterých předpokládáme, že poskytnutý úvěr bez větších potíží splatí od těch, u kterých naopak předpokládáme, že budou mít se splácením problémy. Narážíme zde však na otázku, zda je skutečně možné sestavit model, který spolehlivě tyto skupiny klientů oddělí. Ptáme se tedy, zda je možné najít takovou hodnotu skóre $s_0$, pro kterou bude platit, že všichni klienti kteří splatili své dluhy, budou mít hodnotu skóre větší než $s_0$ a naopak všichni klienti kteří své dluhy nesplatili, nebo s jejich splácením měli problémy, budou mít hodnotu skóre menší než $s_0$.

V praxi se nám takovýto model samozřejmě sestavit nepodaří. Vždy budou existovat klienti, jejichž hodnota skóre bude velmi nízká, a přesto své dluhy splatili a stejně tak klienti, kteří sice budou mít hodnotu skóre vysokou a přesto své dluhy z nějakého důvodu splatit nezvládli, nebo minimálně měli s jejich splácením problémy. [3]

Pro lepší představu, proč není možné sestavit model, který by měl dokonalou diverzifikační schopnost, si stačí vzpomenout na lineární regresi. Pomocí lineární regrese jsme se snažili odhadnout hodnotu nějaké vysvětlované proměnné pomocí jedné nebo více vysvětlujících proměnných. Odhad jsme prováděli tak, že jsme naměřená data prokládali přímkou s určitými vlastnostmi. A víme, že jen zcela ojediněle nastala situace, kdy všechny pozorované hodnoty leželi přesně na konkrétní regresní přímce. Taková situace nastává pouze tehdy, pokud je mezi vysvětlovanou a vysvětlující proměnnou přímá nebo nepřímá lineární závislost, což v praxi není příliš časté. Na obrázku níže můžeme jako příklad vidět lineární regresní přímku, pomocí které se snažíme vysvětlit proměnnou cena vozu pomocí vysvětlující proměnné stáří vozu.




OBRÁZEK 3.1 LINEÁRNÍ REGRESE [VLASTNÍ]

Vidíme, že kdybychom chtěli odhadnout cenu auta starého 5 let pomocí této regresní přímky, tak bychom dostali jinou hodnotu, než kterou jsme získali z naměřených dat. A přesně toto je důvod, proč se nám nepodaří nikdy získat model, který by byl zcela perfektní. Můžeme se pouze snažit vyjádřit, jak dobrý daný model je. V případě lineární regrese používáme například koeficient determinace, který nám říká, jak velkou část variability vysvětlované proměnné se nám daným modelem podařilo vysvětlit. V našem případě, kdy se snažíme oddělit od sebe dvě skupiny klientů, budeme hovořit o schopnosti diverzifikace daného modelu.

## 3.2 Lorenzova křivka

S Lorenzovou křivkou se setkáme nejčastěji v ekonomii, kde se používá pro znázornění nerovnosti rozdělení bohatství ve společnosti. Toto však není její jediné praktické využití. Dá se použít mimo jiné také pro hodnocení skóringových modelů. Pro její konstrukci použijeme distribuční funkce skóre klientů, kteří splatili své dluhy a distribuční funkce skóre klientů, kteří měli problém se splacením svých dluhů.

Tyto distribuční funkce definujeme pro každé $s \epsilon S$, kde $S$ představuje obor hodnot skóringové funkce jako pravděpodobnost, že náhodně vybraný klient bude mít skóre menší než $s$ pro distribuční funkci skóre klientů kteří splatili své dluhy (budeme ji značit $F^G(s)$) a jako pravděpodobnost, že náhodně vybraný klient bude mít skóre větší než $s$ pro distribuční funkci skóre klientů kteří měli problémy se splacením svých dluhů (budeme značit $F^B(s)$).

V praxi tyto distribuční funkce většinou neznáme, proto je nahrazujeme jejich konzistentními dohady, které definujeme následovně

$$F^G(s) = \frac{G(s_i<s)}{G}$$

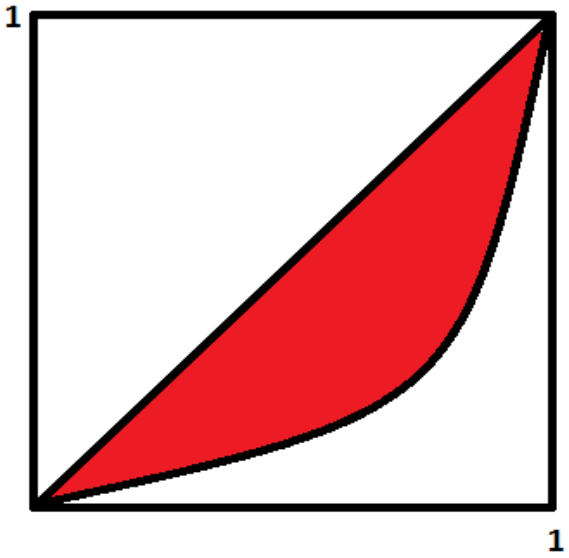
$$F^B(s) = \frac{B(s_i<s)}{B}$$

kde $G(s_i < s)$ představuje počet klientů, kteří splatili své dluhy a jejichž skóre bylo menší než $s$, $G$ představuje počet všech klientů, kteří své dluhy bez problémů splatili, $B(s_i < s)$ symbolizuje počet klientů, kteří měli problémy se splacením svých dluhů a jejich skóre bylo menší než $s$ a $B$ představuje počet všech klientů, kteří měli se splácením dluhů problémy.

Lorenzovu křivku pak definujeme jako množinu bodů

$$L = \{[F^B(s), F^G(s)] \epsilon R^2 : s \epsilon S\}$$

Takto definovaná Lorenzova křivka se bude nacházet uvnitř čtverce s hranou délky 1. Čím větší je obsah mezi úhlopříčkou čtverce a Lorenzovou křivkou, tím větší diverzifikační schopnost má námi zvolený model. Příklad Lorenzovy křivky definované touto množinou bodů můžeme vidět na následujícím obrázku

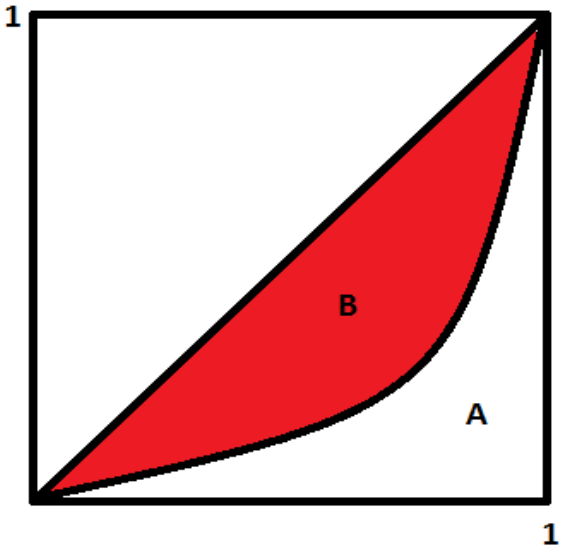



OBRÁZEK 3.2 LORENZOVA KŘIVKA [VLASTNÍ]

Červená plocha nám vyjadřuje diverzifikační schopnost daného modelu. V ideálním případě by obsah této plochy byl 0,5. V takovém případě by totiž Lorenzova křivka kopírovala strany trojúhelníku a červeně vybarvená by byla celá plocha pod úhlopříčkou. Znamenalo by to, že náš model má perfektní schopnost diverzifikace. [3]

## 3.3 Giniho koeficient

V předchozí části jsme si řekli, že kvalitu skóringového modelu hodnotíme jeho diverzifikační schopností. Nyní nám půjde o to, tuto schopnost diverzifikace kvantifikovat. K tomu použijeme právě Giniho koeficient. Vezmeme si předchozí graf Lorenzovy křivky a zavedeme symboly pro červenou plochu $B$ a pro zbytek plochy pod diagonálou $A$, jak to máme znázorněno na následujícím obrázku.




OBRÁZEK 3.3 GINIHO KOEFICIENT [VLASTNÍ]

Potom hodnotu Giniho koeficientu $G$ získáme následovně

$$G = \frac{B}{A+B}$$

Protože celkový obsah plochy pod diagonálou $A + B$ je roven polovině obsahu jednotkového čtverce (který nabývá hodnoty 0,5), můžeme výraz upravit

$$G = \frac{A}{\frac{1}{2}}$$

$$G = 2A$$

Tento výraz můžeme ještě přepsat do tvaru

$$G = 1 - 2B$$

A odtud již dostaneme tvar

$$G = 1 - 2\int_s F^G(s)dF^B(s)$$

Giniho koeficient je tedy poměr červeně vyznačené plochy a celkové plochy pod diagonálou. Je tedy zřejmé, že nabývá hodnot v intervalu $\langle -1,1 \rangle$. V případě, že Giniho koeficient nabývá hodnoty 1, pak máme model, který má perfektní schopnost diverzifikace a v případě, že by nabýval hodnoty 0, tak je náš model zcela nevhodný pro oddělování klientů. Pokud by hodnota Giniho koeficientu byla záporná, znamená to, že máme skóringovou funkci s opačnou klasifikací. Lorenzova křivka by ležela nad diagonálou. Snažíme se tedy najít model s co nejvyšší hodnotou Giniho koeficientu. [3]

## 3.4 Odhad Giniho koeficientu

Pro odhad Giniho koeficientu použijeme Somerovu $d$ statistiku definovanou následujícím vztahem

$$d = \frac{a-b}{a+b+c}$$

kde - $a$ představuje počet všech dvojic klientů u kterých byl klient, jenž bez problémů splatil své závazky, ohodnocen vyšším skóre než klient, který zdefaultoval.

- $b$ představuje počet všech dvojic klientů u kterých byl klient, jenž bez problémů splatil své závazky, ohodnocen nižším skóre než klient, který zdefaultoval.

- $c$ představuje počet všech dvojic klientů u kterých byl klient, jenž bez problémů splatil své závazky, ohodnocen stejným skóre jako klient, který zdefaultoval. [3]

# 4. Skóringové modely

## 4.1 Skóringové modely

Jak jsme si již řekli v úvodu, každá půjčka s sebou přináší pro věřitele riziko, že klient, kterému peníze půjčujeme, nebude schopen své závazky splatit. Pro každou instituci, která se poskytováním úvěrů zabývá, je tedy důležité nějakým způsobem toto riziko kvantifikovat. A k tomu nám mohou sloužit právě skóringové modely. Každému žadateli o úvěr daný model přiřadí takzvané skóre. Čím vyšší skóre daný klient má, tím je u něj vyšší pravděpodobnost, že své závazky splatí bez problémů a nezdefaultuje.

Skóringový model vychází z databáze klientů, kterým instituce poskytla úvěr v minulosti a u kterých má vedeny určité informace. A na základě těchto informací a skutečnosti, zda klient s danými vlastnostmi zdefaultoval či nezdefaultoval, je pak daným skóringovým modelem klientovi přiřazeno skóre. Na základě tohoto skóre se pak instituce rozhodne, zda žadateli úvěr poskytne nebo neposkytne. [3]

Skóringových modelů je více, my se však zaměříme na dva konkrétní.

- Independence model
- WOE model

## 4.2 Independence model

Skóringovou funkci u Independence modelu budeme značit $S^{IM}(x)$ a definujeme ji následovně

$$S^{IM}(x) = odds \prod_{(i,j)\in Z} (OR_j^i)^{x_j^i}$$

kde $odds$ představuje takzvanou šanci, tedy poměr klientů, kteří bez problémů své závazky splatili ku počtu klientů, kteří zdefaultovali; $x$ jsou nezávislé proměnné charakterizující daného klienta a $OR_j^i$ je definována takto

$$OR_j^i = \frac{odds_j^i}{odds}$$

kde $odds_j^i$ vyjadřuje poměr klientů, kteří splatili své závazku ku počtu klientů, kteří zdefaultovali v dané kategorii (například v kategorii muži).

Vidíme, že jednou ze zásadních nevýhod independence modelu je skutečnost, že každé kategorii je přiřazena stejná váha. To však může být zkreslující. Například skutečnost, zda se jedná o muže či ženu, bude mít menší vliv na schopnost splácet než to, jestli je daný klient zaměstnaný nebo nezaměstnaný. Další nevýhodou je předpoklad o nezávislosti regresorů. [3]

## 4.3 WOE model

Jak jsme si již řekli, nevýhodou Independence modelu je fakt, že každé kategorii je přiřazena stejná váha. Proto je možné použít jako skóringový model WOE model, kde zkratka WOE pochází z anglického weight of evidence. Je tedy zřejmé, že tento model bude každé kategorii přidělovat jinou váhu. WOE model má tvar

$$S^{WOE}(x, \lambda) = odds \prod_{(i,j)\in Z} (OR_j^i)^{\lambda^i x_j^i}$$

kde λ je vektorem vah jednotlivých kategorií a $x$ jsou stejně jako u Independence modelu nezávisle proměnné.

Logaritmus této funkce nazýváme $logit(x)$ a má následující tvar

$$logit(x) = \ln(odds) + \sum_{(i,j)\in Z} \lambda^i x_j^i (OR_j^i)$$

Vektor parametrů λ je odtud možné odhadovat pomocí logistické regrese. [3]

# 5. Logistická regrese na reálných datech

V této části budeme logistickou regresi aplikovat na reálná data. Sestavíme model a odhadneme koeficienty. Oboje následně otestujeme. Jako nástroj použijeme program EViews.

## 5.1 Data

Datový soubor je poměrně veliký. Celkem máme údaje o 30 732 klientech. Zároveň máme k dispozici skutečně velké množství různých vlastností (dále proměnné), které u daných klientů sledujeme. Do našeho modelu však zařadíme jen některé. Jednu proměnnou vybereme jako vysvětlovanou a některé další jako vysvětlující proměnné. Ukázku datového souboru můžeme vidět na následujícím obrázku. Celý datový soubor je pak možné nalézt v příloze.

| SEX | AGE | maturity | Region | INSURANCE | CHILD_NUM | EDUCATION | FAMILY_STATE |
|---|---|---|---|---|---|---|---|
| Z | 32 | 11 | HC | N | 2 | High school | Married |
| Z | 39 | 11 | KG | Y | 2 | Secondary school | Married |
| M | 26 | 8 | HC | Y | 0 | High school | Single |
| M | 35 | 10 | VL | Y | 2 | Secondary school | Married |
| M | 46 | 22 | LA | N | 2 | Secondary school | Married |
| M | 43 | 7 | Tv | Y | 2 | Secondary school | Widowed |
| M | 35 | 16 | HC | N | 1 | Secondary school | Married |
| Z | 30 | 10 | KG | Y | 0 | Secondary school | Single |
| Z | 43 | 15 | HN | N | 2 | Bachelor | Married |
| M | 25 | 7 | LD | Y | 0 | High school | Married |
| M | 25 | 11 | BL | Y | 0 | Secondary school | Single |
| Z | 22 | 11 | QN | N | 1 | Secondary school | Married |
| Z | 48 | 14 | HU | Y | 3 | High school | Married |
| M | 48 | 15 | HC | Y | 2 | High school | Married |
| Z | 56 | 12 | HC | N | 2 | Secondary school | Married |
| Z | 44 | 11 | HC | Y | 2 | Secondary school | Married |
| Z | 44 | 11 | HU | Y | 2 | Secondary school | Married |

TABULKA 5.1 UKÁZKA DATOVÉHO SOUBORU

Na této ukázce můžeme vidět, že o klientech máme údaje jako pohlaví, věk, dosažené vzdělání nebo rodinný stav. Můžeme také již na základě této ukázky říci, že máme v souboru mnoho proměnných, které bude potřeba kódovat jako 0,1 proměnné. Tedy tzv. dummy proměnné. To je příklad právě rodinného stavu nebo dosaženého vzdělání.

## 5.2 Model

Nyní vybereme některé proměnné, které zahrneme do modelu. Jako první zvolíme vysvětlovanou proměnnou. V našem datovém souboru je hned několik možných ¨kandidátů¨. Popíšeme si je tedy trochu podrobněji.

FPD10 – První splátka 10 dní po splatnosti

FPD30 – První splátka 30 dní po splatnosti

FPD90 – První splátka 90 dní po splatnosti

SPD30 – Druhá splátka 30 dní po splatnosti

SPD90 – Druhá splátka 90 dní po splatnosti

My si jako vysvětlovanou proměnnou vybereme FPD10. Tedy první splátku 10 dní po splatnosti. Důvod je především ten, že ostatní z těchto proměnných neobsahují buď žádný, nebo zcela zanedbatelný počet jedniček. Proměnná FPD10, stejně jako všechny ostatní, je 0,1 proměnná, která nabývá hodnoty 0 v případě, že klient nezdefaultoval a hodnoty 1 v případě, že klient zdeafaultoval. Celkem máme v datovém souboru 302 klientů, kteří zdefaultovali.

Dále vybereme vhodné vysvětlující proměnné. Chceme vybrat především takové proměnné, které budou vycházet statisticky významné. Do modelu tedy zahrneme následující proměnné.

CREDIT – Proměnná má dvě kategorie CASH a POS. Kategorie CASH znamená, že spotřebitelský úvěr byl poskytnut bez účelu. Kategorie POS udává úvěr, který byl poskytnut na konkrétní zboží. Proměnná nabývá hodnoty 1 v případě, že úvěr byl poskytnut bez účelu a hodnoty 0 v opačném případě.

INSURANCE – Proměnná, která nabývá hodnoty 0 v případě, že klient nemá pojištění a hodnoty 1 v případě, že klient má pojištění.

POHLAVI – Tato proměnná nabývá hodnoty 0 v případě, že se jedná o ženu a hodnoty 1 v případě, že se jedná o muže.

SALES_TYPE – Udává nám typ prodejního místa. Opět se jedná o 0,1 proměnnou. Hodnota 0 značí NON-ALDI (mimo pobočku), hodnota 1 ALDI (na pobočce).

SPLATNOST – Jedná se o jedinou kvantitativní proměnnou, kterou v našem modelu máme. Značí dobu splatnosti daného úvěru.

NAD15 – Tato proměnná je opět 0,1 proměnná, která nabývá hodnoty 1 v případě, že výše půjčky je 15 001-25 000 a hodnoty 0, pokud výše půjčky v tomto intervalu není.

NAD25 – Platí to samé, co u předchozí proměnné, pouze pro výši půjčky v intervalu 25 001 – 35 000.

NAD35 – Nabývá hodnoty 1 pro půjčku vyšší než 35 000 a hodnoty 0 pro půjčku ve výši 35 000 a nižší.

NAD5 – Stejný popis jako u proměnných NAD15 a NAD5, ovšem pro výši půjčky v intervalu 5 001 – 15 000.

NAD400 – Opět platí to samé, pouze pro půjčku jejíž výše je v intervalu 401 – 5 000.

FAMILY – Poslední proměnná vstupující do modelu. Nabývá hodnoty 0 v případě, že klient nemá rodinu v době zřizování půjčky a hodnoty 1 v případě, že klient rodinu v této době má.

Ukázku vybraných proměnných můžeme vidět v následující tabulce. Proměnné NAD15, NAD25, NAD35, NAD5 a NAD400 jsou ve sloupci výše půjčky. Pro potřeby modelu byly překódovány do 0,1 proměnných.

| FPD10 | Credit | INSURANCE | POHLAVI | Sales_Type | Výše půjcky | FAMILY | maturity |
|---|---|---|---|---|---|---|---|
| 0 | CASH | N | z | NON-ALDI | 5 001 - 15 000 | Married | 11 |
| 0 | CASH | Y | z | NON-ALDI | 25 001 - 35 000 | Married | 11 |
| 0 | POS | Y | m | NON-ALDI | 5 001 - 15 000 | Single | 8 |
| 0 | POS | Y | m | NON-ALDI | 25 001 - 35 000 | Married | 10 |
| 0 | CASH | N | m | NON-ALDI | 15 001 - 25 000 | Married | 22 |
| 1 | POS | Y | m | NON-ALDI | 5 001 - 15 000 | Widowed | 7 |
| 0 | CASH | N | m | NON-ALDI | 15 001 - 25 000 | Married | 16 |
| 0 | POS | Y | z | NON-ALDI | 15 001 - 25 000 | Single | 10 |
| 0 | CASH | N | z | NON-ALDI | 25 001 - 35 000 | Married | 15 |
| 0 | POS | Y | m | NON-ALDI | 15 001 - 25 000 | Married | 7 |
| 0 | CASH | Y | m | NON-ALDI | 15 001 - 25 000 | Single | 11 |
| 0 | POS | N | z | NON-ALDI | 5 001 - 15 000 | Married | 11 |
| 0 | CASH | Y | z | NON-ALDI | 25 001 - 35 000 | Married | 14 |
| 0 | CASH | Y | m | NON-ALDI | 25 001 - 35 000 | Married | 15 |
| 0 | CASH | N | z | NON-ALDI | 5 001 - 15 000 | Married | 12 |

TABULKA 5.2 – VYBRANÉ PROMĚNNÉ

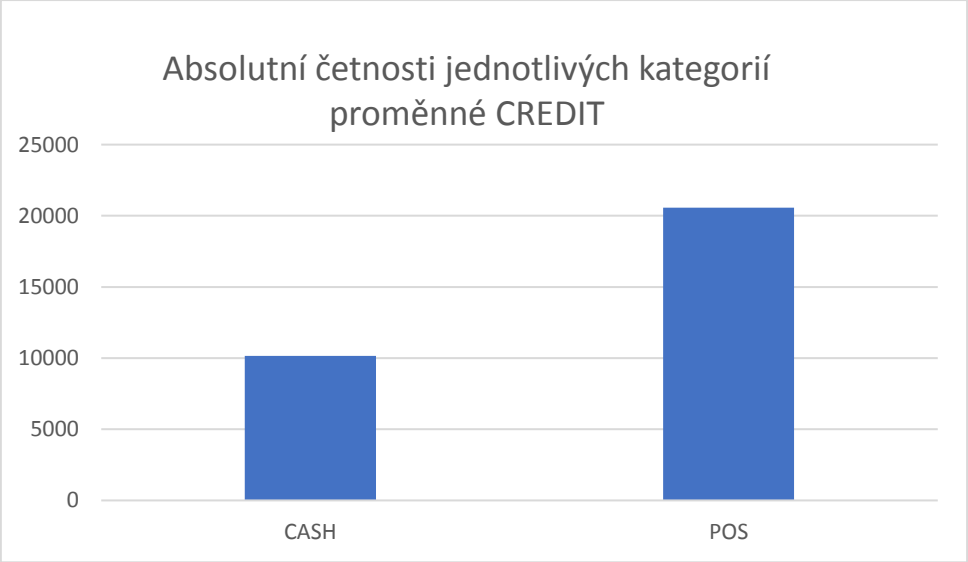
Proměnná výše půjčky je kvalitativní proměnná, kterou jsme rozdělili do kategorií NAD15, NAD25, NAD35, NAD5 a NAD400. Z těchto pěti kategorií však do modelu zahrneme pouze čtyři. Jednu z nich vybereme jako takzvanou referenční kategorii. Interpretace odhadnutých koeficientů u ostatních čtyř se tak bude vztahovat právě k referenční kategorii. Jako referenční kategorii zvolíme proměnnou NAD400.

## 5.3 Popisná statistika proměnných

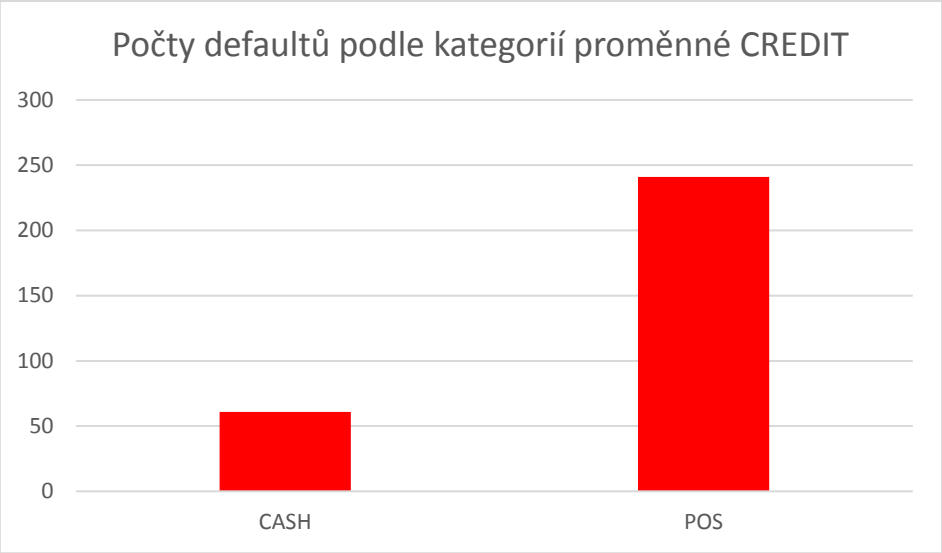
V předchozí části jsme si vysvětlili, co nám jednotlivé proměnné vybrané do modelu říkají. Nyní se podíváme na základní statistiky týkající se těchto proměnných. Jelikož většina našich proměnných jsou kvalitativní proměnné, tak nás budou zajímat především absolutní a relativní četnosti jednotlivých kategorií, míry selhání v těchto kategoriích a také nás pro každou kategorii bude zajímat hodnota skóre. Tedy to, co jsme v teoretické části značili $OR_j^i$.

1. CREDIT

První proměnnou, u které budeme zjišťovat výše zmíněné vlastnosti, je proměnná CREDIT. Ta obsahuje dvě kategorie (CASH a POS). Počty klientů v jednotlivých kategoriích a počty defaultů v těchto kategoriích vystihují následující grafy.




GRAF 5.1




GRAF 5.2

Pro obě kategorie vypočítáme míry selhání, které budeme značit $d$ a skóre ($OR$)

$$d_{CASH} = \frac{61}{10152} * 100 = 0{,}601\%$$

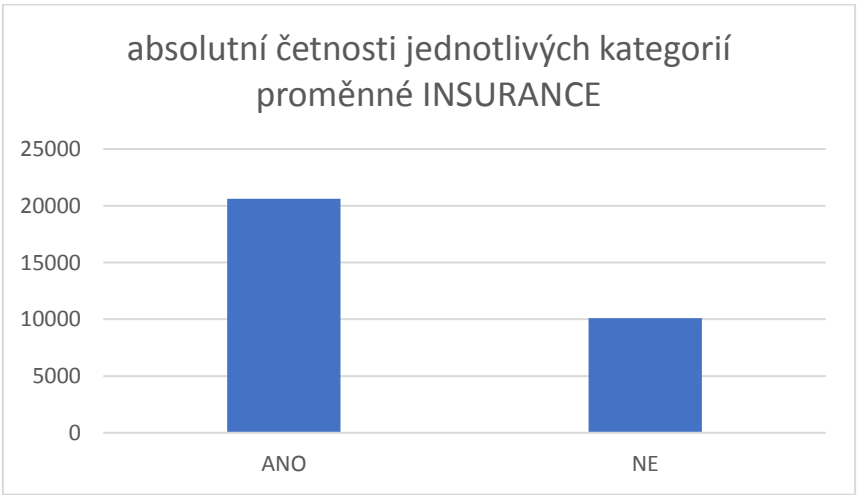
$$d_{POS} = \frac{241}{20580} * 100 = 1{,}171\%$$

$$OR_{CASH} = \frac{30732 - 302}{302} : \frac{10152 - 61}{61} = 1{,}642$$

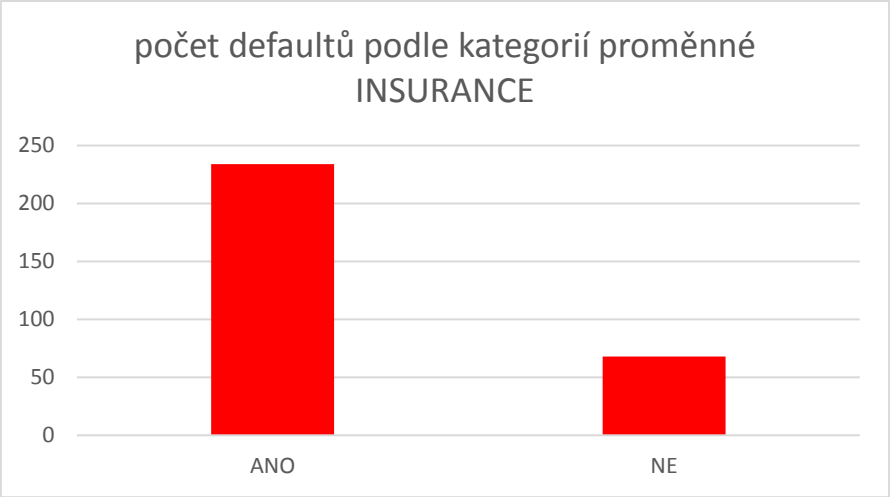
$$OR_{POS} = \frac{30732 - 302}{302} : \frac{20580 - 241}{241} = 0{,}838$$

Vidíme, že míra selhání je vyšší u kategorie POS, což také znamená, že má tato kategorie vyšší skóre. Klient v této kategorii má tedy nižší pravděpodobnost defaultu.

2. INSURANCE




GRAF 5.3




GRAF 5.4

Opět spočítáme pro každou kategorii míry selhání a skóre

$$d_{ANO} = \frac{234}{20626} * 100 = 1{,}134\%$$

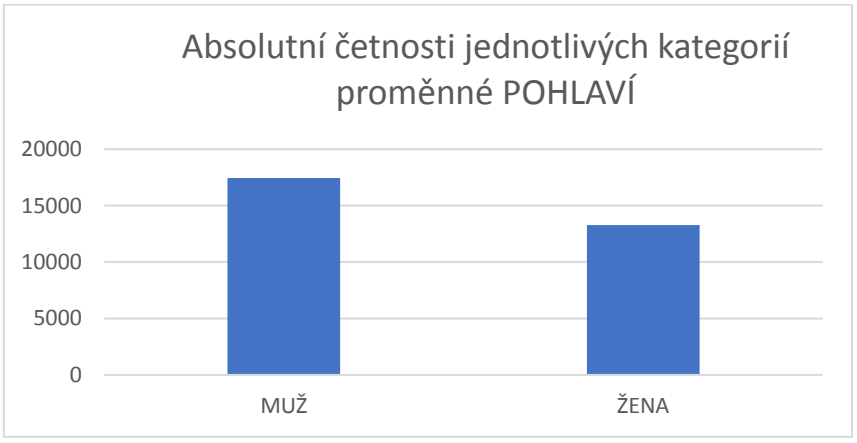
$$d_{NE} = \frac{68}{10106} * 100 = 0{,}673\%$$

$$OR_{ANO} = \frac{30732 - 302}{302} : \frac{20626 - 234}{234} = 0{,}865$$

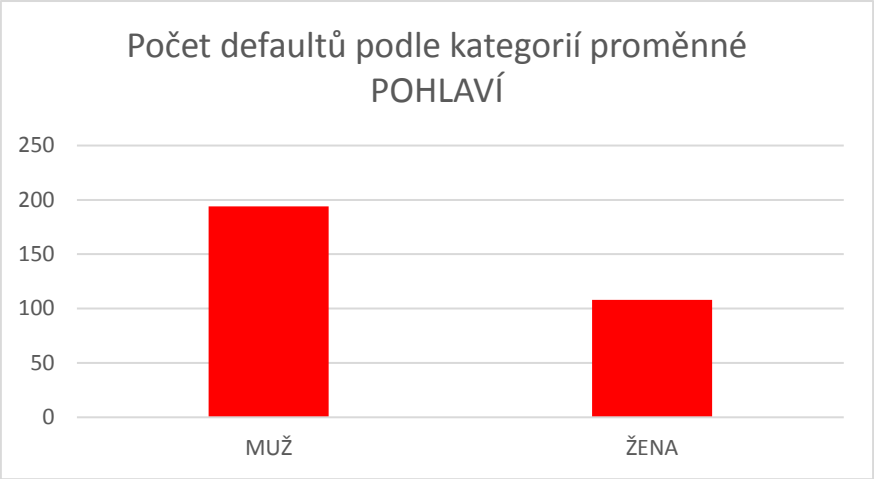
$$OR_{NE} = \frac{30732 - 302}{302} : \frac{10106 - 68}{68} = 1{,}465$$

Klienti s pojištěním mají vyšší míru selhání, tím i nižší skóre a mají tak vyšší pravděpodobnost defaultu.

3. POHLAVI




GRAF 5.5




GRAF 5.6

Míry selhání a skóre pro každou kategorii spočítáme zde

$$d_{MUZ} = \frac{194}{17456} * 100 = 1{,}111\%$$

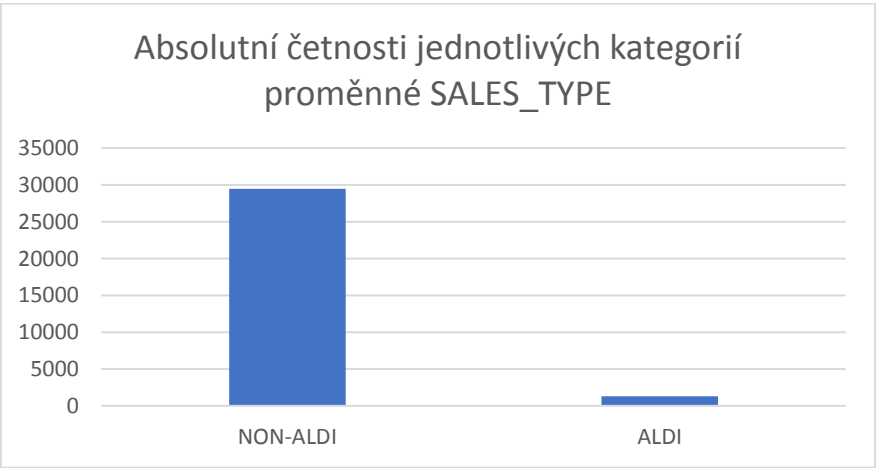
$$d_{ZENA} = \frac{108}{13276} * 100 = 0{,}813\%$$

$$OR_{MUZ} = \frac{30732 - 302}{302} : \frac{17456 - 194}{194} = 0{,}883$$

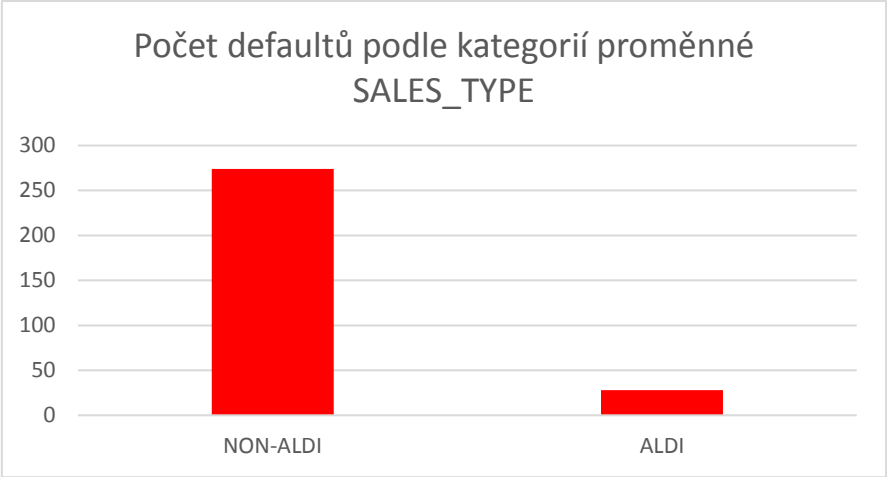
$$OR_{ZENA} = \frac{30732 - 302}{302} : \frac{13276 - 108}{108} = 1{,}210$$

V tomto případě mají vyšší míru selhání a nižší skóre muži než ženy. Tuto skutečnost se pokusíme vysvětlit po odhadu modelu.

4. SALES_TYPE




GRAF 5.7




GRAF 5.8

Stejně jako u všech předchozích kvalitativních proměnných spočítáme míry selhání a skóre pro jednotlivé kategorie

$$d_{NON-ALDI} = \frac{274}{29444} * 100 = 0{,}931\%$$

$$d_{ALDI} = \frac{28}{1288} * 100 = 2{,}174\%$$

$$OR_{NON-ALDI} = \frac{30732 - 302}{302} : \frac{29444 - 274}{274} = 1{,}057$$

$$OR_{ALDI} = \frac{30732 - 302}{302} : \frac{1288 - 28}{28} = 0{,}447$$

Zde vidíme že klienti, kteří sjednávají půjčku na pobočce, mají výrazně vyšší míru selhání. Vysvětlení této skutečnosti není zcela jednoznačné.

5. SPLATNOST

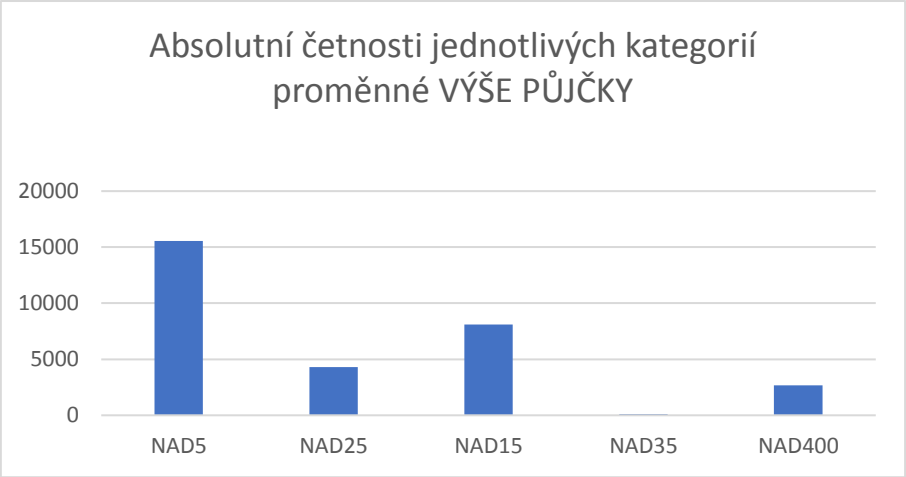
SPLATNOST je jediná kvantitativní proměnná, kterou jsme zařadili do modelu. U této proměnné nás budou zajímat (mimo jiné) i takové statistické údaje jako je průměr, minimální a maximální hodnota, nebo medián. Všechny tyto údaje můžeme vidět v následující tabulce

| | průměr | Medián | maximum | minimum | sm. odchylka | rozptyl |
|---|---|---|---|---|---|---|
| SPLATNOST | 11,221 | 11 | 36 | 4 | 4,833 | 23,361 |

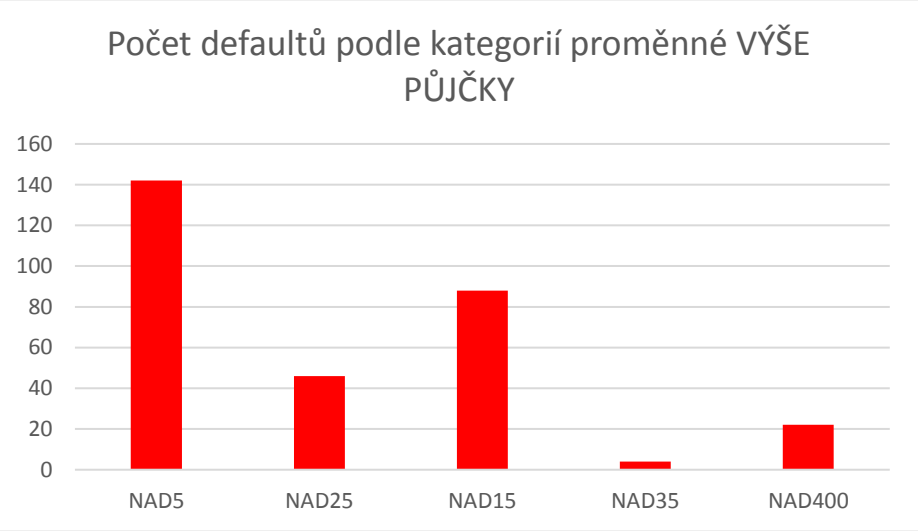
TABULKA 5.3 – POPISNÁ STATISTIKA PROMĚNNÉ SPLATNOST

6. VÝŠE PŮJČKY

Jedná se o jedinou proměnnou, která má více než 2 kategorie. Celkem obsahuje 5 kategorií (NAD15, NAD25, NAD35, NAD5 a NAD400).




GRAF 5.9




GRAF 5.10

Také u této kvalitativní proměnné určíme míry selhání a skóre jednotlivých kategorií. Tentokrát však máme kategorií více než jen dvě

$$d_{NAD5} = \frac{142}{15559} * 100 = 0{,}913\%$$

$$d_{NAD25} = \frac{46}{4289} * 100 = 1{,}073\%$$

$$d_{NAD15} = \frac{88}{8110} * 100 = 1{,}085\%$$

$$d_{NAD35} = \frac{4}{91} * 100 = 4{,}396\%$$

$$d_{NAD400} = \frac{22}{2680} * 100 = 0{,}821\%$$

$$OR_{NAD5} = \frac{30732 - 302}{302} : \frac{15559 - 142}{142} = 1{,}077$$

$$OR_{NAD25} = \frac{30732 - 302}{302} : \frac{4289 - 46}{46} = 0{,}915$$

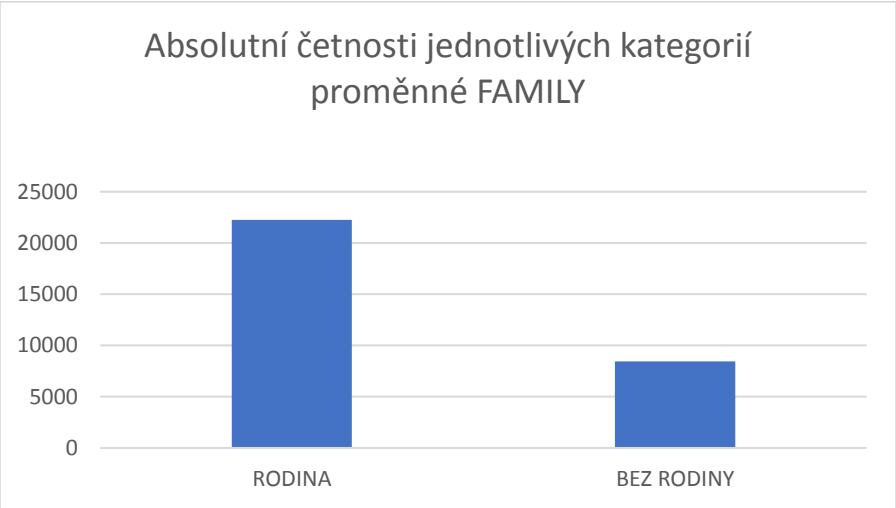
$$OR_{NAD15} = \frac{30732 - 302}{302} : \frac{8110 - 88}{88} = 0{,}905$$

$$OR_{NAD35} = \frac{30732 - 302}{302} : \frac{91 - 4}{4} = 0{,}216$$

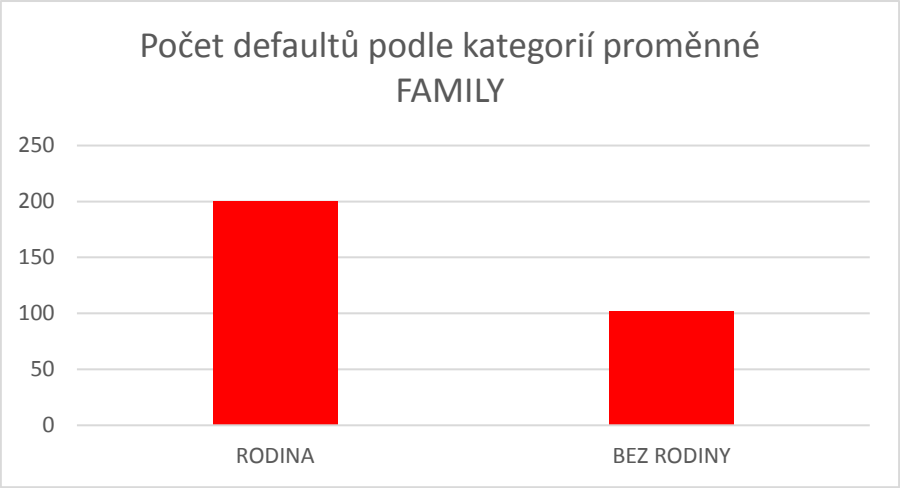
$$OR_{NAD400} = \frac{30732 - 302}{302} : \frac{2680 - 22}{22} = 1{,}199$$

Vidíme, že míra selhání roste s výší půjčky. Analogicky tak s výší půjčky klesá skóre a roste pravděpodobnost defaultu.

7. FAMILY




GRAF 5.11




GRAF 5.12

$$d_{RODINA} = \frac{200}{22269} * 100 = 0{,}898\%$$

$$d_{BEZ\ RODINY} = \frac{102}{8463} * 100 = 1{,}205\%$$

$$OR_{RODINA} = \frac{30732 - 302}{302} : \frac{22269 - 200}{200} = 1{,}095$$

$$OR_{BEZ\ RODINY} = \frac{30732 - 302}{302} : \frac{8463 - 102}{102} = 0{,}814$$

Nežije-li klient sám, je u něj nižší míra selhaní, a samozřejmě také vyšší skóre a nižší pravděpodobnost defaultu. Opět se tuto skutečnost pokusíme vysvětlit po odhadu modelu.

## 5.4 Odhad modelu

Nyní si tedy celý model odhadneme. Jak již bylo řečeno, k odhadu použijeme program EViews. Výstup z tohoto programu můžeme vidět na následujícím obrázku.

Dependent Variable: FPD
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 05/02/17 Time: 15:46
Sample: 1 30732
Included observations: 30729
Convergence achieved after 9 iterations
Coefficient covariance computed using observed Hessian

| Variable | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| C | -5.873229 | 0.310624 | -18.90782 | 0.0000 |
| CREDIT | -0.876396 | 0.173603 | -5.048263 | 0.0000 |
| INSURANCE | 0.483955 | 0.152880 | 3.165598 | 0.0015 |
| POHLAVI | 0.292848 | 0.121413 | 2.411996 | 0.0159 |
| SALES_TYPE | 1.027435 | 0.221579 | 4.636890 | 0.0000 |
| SPLATNOST | 0.045068 | 0.012940 | 3.482733 | 0.0005 |
| NAD15 | 0.718532 | 0.257686 | 2.788398 | 0.0053 |
| NAD25 | 0.816724 | 0.289399 | 2.822136 | 0.0048 |
| NAD35 | 2.036356 | 0.568995 | 3.578866 | 0.0003 |
| NAD5 | 0.415181 | 0.235715 | 1.761369 | 0.0782 |
| FAMILY | -0.378955 | 0.126922 | -2.985739 | 0.0028 |

OBRÁZEK 5.1 – VÝSTUP Z PROGRAMU EVIEWS

Pokud bychom hladinu významnosti zvolili 5%, tak vidíme že nám všechny proměnné (kromě proměnné NAD5) vyšly statisticky významné. Když se podíváme na koeficienty u jednotlivých proměnných, tak vidíme, že většina z nich má pozitivní vliv na pravděpodobnost defaultu. Jinými slovy zvyšují tuto pravděpodobnost. U proměnných NAD15, NAD25, NAD35 a NAD5 je nutné si uvědomit, že koeficienty všech těchto proměnných se vztahují k jedné referenční kategorii, kterou je proměnná NAD400. Z čehož při pohledu na hodnoty koeficientů můžeme soudit, že právě klienti spadající do kategorie NAD400 mají nejnižší pravděpodobnost defaultu, pokud bychom neuvažovali žádné další faktory. Obecně můžeme z hodnot koeficientů usoudit, že s výší půjčky, roste i pravděpodobnost defaultu.

Naproti tomu proměnná, která negativně ovlivňuje pravděpodobnost defaultu (snižuje ji) je proměnná FAMILY. Pokud tedy někdo nežije sám, je u něj nižší pravděpodobnost, že nebude schopen splácet své závazky. To je nejspíše dáno tím, že dva lidé obvykle znamenají dva příjmy. Pokud tedy jeden z nich přijde o své příjmy, například byl propuštěn ze zaměstnání, stále mají ještě jeden příjem, ze kterého mohou splácet své závazky.

Zajímavý je také vliv pohlaví. Vidíme, že u mužů je pravděpodobnost defaultu vyšší než u žen. V tomto případě bychom mohli říci, že to může mít souvislost s proměnnou FAMILY. Muži totiž žijí častěji (a hlavně déle) sami než ženy.

Překvapující může být poměrně velký vliv typu prodejního místa. Je-li totiž půjčka sjednána na pobočce, zvyšuje to pravděpodobnost defaultu. Není však zcela jednoznačné,

z jakého důvodu tomu tak je. Vzhledem k mírám selhání a k hodnotám skóre, které jsme pro každou kategorii každé kvalitativní proměnné vypočítali, by nás však hodnoty koeficientů v odhadnutém modelu neměli nijak překvapit.

Nyní se tedy na základě tohoto modelu pokusíme predikovat pravděpodobnost defaultu dvou klientů s určitými vlastnostmi.

První klient bude svobodný muž, který si půjčil částku 10000 se splatností 10 let, prodejní místo bude mimo pobočku, úvěr nebyl poskytnut na konkrétní zboží a nebude mít pojištění. Logitová transformace tedy bude

$$Logit(\pi) = -5{,}87 - 0{,}88 * 1 + 0{,}48 * 0 + 0{,}29 * 1 + 1{,}03 * 0 + 0{,}05 * 10 + 0{,}72 * 0 + 0{,}82 * 0 + 2{,}04 * 0 + 0{,}42 * 1 - 0{,}38 * 0 = -5{,}54$$

Pravděpodobnost defaultu pak bude

$$p_i = \frac{e^{-5{,}54}}{1 + e^{-5{,}54}} = 0{,}00391$$

Pravděpodobnost defaultu klienta s těmito vlastnostmi je tedy 0,391%.

Nyní si ještě určíme, kolikrát je pravděpodobnost, že tento klient zdefaultuje větší, než pravděpodobnost, že nezdefaultuje.

$$\frac{0{,}00391}{1 - 0{,}00391} = 0{,}0039$$

Vidíme, že pravděpodobnost že tento klient zdefaultuje je 0,0039krát větší, než že nezdefaultuje.

Teď si vezměme druhého klienta, kterým bude vdaná žena, která si půjčila částku 30 000 se splatností 5 let, prodejní místo bude na pobočce, úvěr byl poskytnut na konkrétní zboží a bude mít pojištění. Logitová transformace tedy bude

$$Logit(\pi) = -5{,}87 - 0{,}88 * 0 + 0{,}48 * 1 + 0{,}29 * 0 + 1{,}03 * 1 + 0{,}05 * 5 + 0{,}72 * 0 + 0{,}82 * 1 + 2{,}04 * 0 + 0{,}42 * 0 - 0{,}38 * 1 = -3{,}67$$

Pravděpodobnost defaultu pak bude

$$p_i = \frac{e^{-3{,}67}}{1 + e^{-3{,}67}} = 0{,}0248$$

Pravděpodobnost defaultu klienta s těmito vlastnostmi bude tedy 2,48%.

A opět si určíme kolikrát je větší pravděpodobnost, že tento klient zdefalutuje, než že nezdefaultuje

$$\frac{0{,}0248}{1 - 0{,}0248} = 0{,}0254$$

Pravděpodobnost, že klient s těmito vlastnostmi zdefaultuje je 0,0254krát větší, než pravděpodobnost, že tento klient nezdefaultuje.

Již jsme si tedy popsali odhadnutý model, zjistili jsme statistickou významnost jednotlivých proměnných a zkusili jsme si predikovat pravděpodobnost defaultu pro konkrétního klienta s konkrétními vlastnostmi. Ještě se tedy podíváme na další výstup z programu Eviews týkající se tohoto modelu. Zajímat nás bude především celkový test o modelu.

| | | | |
|---|---|---|---|
| McFadden R-squared | 0.028136 | Mean dependent var | 0.009828 |
| S.D. dependent var | 0.098649 | S.E. of regression | 0.098477 |
| Akaike info criterion | 0.108027 | Sum squared resid | 297.8920 |
| Schwarz criterion | 0.111010 | Log likelihood | -1648.783 |
| Hannan-Quinn criter. | 0.108983 | Deviance | 3297.566 |
| Restr. deviance | 3393.033 | Restr. log likelihood | -1696.517 |
| LR statistic | 95.46710 | Avg. log likelihood | -0.053656 |
| Prob(LR statistic) | 0.000000 | | |
| Obs with Dep=0 | 30427 | Total obs | 30729 |
| Obs with Dep=1 | 302 | | |

OBRÁZEK 5.2 – VÝSTUP Z PROGRAMU EVIEWS

Vidíme, že p-hodnota u celkového testu o modelu vyšla velmi malá. Zamítáme tedy nulovou hypotézu, že by všechny koeficienty u vysvětlujících proměnných byly nulové ve prospěch alternativní hypotézy, že alespoň jeden z nich je nenulový.

# 6. Závěr

V této práci byla tedy nejprve teoreticky popsána logistická regrese, odvozena logistická křivka, popsány metody odhadu parametrů a jejich testování. Dále byly v práci popsány typy proměnných, pojem diverzifikace modelu a vybrané skóringové modely. V praktické části poté byla logistická regrese aplikována na konkrétních datech klientů banky. Byla odhadována pravděpodobnost defaultu (platebního selhání) u klientů této banky.

Do modelu bylo zahrnuto celkem 11 vysvětlujících proměnných, z nichž pouze jedna vyšla na hladině významnosti 5% statisticky nevýznamná. Proměnné byly statisticky popsány, u kvalitativních proměnných byly zjištěny míry selhání pro každou kategorii, a také hodnota skóre. Na základě tohoto odhadnutého modelu pak byly predikovány pravděpodobnosti defaultu, pro klienty s konkrétními vlastnostmi. Bylo tedy zjištěno, které proměnné mají na pravděpodobnost defaultu pozitivní, a které naopak negativní vliv. Byly také popsány možné příčiny toho vlivu.

Bylo zjištěno, že pravděpodobnost defaultu s výší půjčky, s dobou splatnosti nebo v kategorii mužů roste. Naopak je nižší u lidí, kteří nežijí sami. Zajímavé bylo zjištění, že pravděpodobnost defaultu je ovlivněna také typem prodejního místa. U klientů, kteří si půjčku sjednají na pobočce je pravděpodobnost defaultu vyšší než u těch, kteří si ji sjednají jinde. Model tedy slouží pro predikci pravděpodobnosti defaultu pro každého klienta, u kterého budou známy informace o hodnotách všech vysvětlujících proměnných. Případně je možné model upravit přidáním dalších vysvětlujících proměnných.

# Literatura

[1]http://www.maths.nuigalway.ie/~emil/handouts/ma101/lec20.pdf

[2] https://cs.wikipedia.org/wiki/Logistick%C3%A1_funkce

[3] RYCHNOVSKÝ, Michal. *Postupná výstavba modelu ohodnocení kreditního rizika.* Praha: Bakalářská práce, 2008

[4] KESELY, Ladislav. *Srovnání logistické regrese a rozhodovacích stromů při tvorbě skóringových modelů.* Praha: Bakalářská práce, 2014

[5] HINDLS, R. A KOL.: *STATISTIKA PRO EKONOMY*, 8. vydání, Professional Publishing, 2007, ISBN 978-80-869-4643-6

[6] HOLÝ, Vladimír a ČERNÝ, Michal. *Kvantitativní ekonomie.* Praha, 2016