Univerzita Hradec Králové

Přírodovědecká fakulta

BAKALÁŘSKÁ PRÁCE

2020 Andrea Široká

Univerzita Hradec Králové

Přírodovědecká fakulta

Katedra matematiky

Zobecněný lineární model a aplikace logistické regrese ve financích

Bakalářská práce

Autor: Andrea Široká
Studijní program: Aplikovaná matematika
Studijní obor: Finanční a pojistná matematika

Vedoucí práce: Mgr. Jitka Kühnová, Ph.D

Hradec Králové duben 2020

Univerzita Hradec Králové
Přírodovědecká fakulta

Zadání bakalářské práce

Autor:	**Andrea Široká**
Studium:	S17AM016BP
Studijní program:	B1103 Aplikovaná matematika
Studijní obor:	Finanční a pojistná matematika
Název bakalářské práce:	**Zobecněný lineární model a aplikace logistické regrese ve financích**
Název bakalářské práce AJ:	Generalized linear model and application of logistic regression in finance

Cíl, metody, literatura, předpoklady:

Tato práce bude obsahovat popis Zobecněného lineárního modelu, především z matematického hlediska, ale i z hlediska použití. Součástí práce bude podrobný popis jednoho z modelů, tedy konkrétně klasické Logistické regrese. Logistická regrese se využívá i v oblasti financí, a proto cílem této práce bude aplikace modelu na reálných datech z oblasti investování na finančním trhu podle vhodně zvolených aspektů.

Anděl, J. Základy matematické statistiky, 2011, ISBN 978-80-7378-162-0

Garantující pracoviště:	Katedra matematiky, Přírodovědecká fakulta
Vedoucí práce:	Mgr. Jitka Kühnová, Ph.D.
Oponent:	Mgr. Tomáš Zuščák, Ph.D.

Datum zadání závěrečné práce: 23.1.2019

Prohlášení:

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, a že jsem v seznamu použité literatury uvedla všechny prameny, ze kterých jsem vycházela.

V Hradeci Králové dne Andrea Široká

Poděkování:

Velmi ráda bych na tomto místě poděkovala paní Mgr. Jitce Kühnové, Ph.D, za cenné rady, ochotu a především čas, který mi věnovala při vedení této bakalářské práce. Zároveň bych ráda poděkovala za odborné rady investičnímu specialitovi panu Janu Portyšovi.

Anotace

Tato práce bude obsahovat popis Zobecněného lineárního modelu, především z matematického hlediska, ale i z hlediska použití. Součástí práce bude podrobný popis jednoho z modelů, tedy konkrétně klasické logistické regrese. Logistická regrese se využívá i v oblasti financí, a proto cílem této práce bude aplikace modelu na reálných datech z oblasti investování na finančním trhu podle vhodně zvolených aspektů.

Klíčová slova

zobecněný lineární model, logistická regrese, akciové investice

Annotation

This work will contain a description of the Generalized Linear Model, especially from the mathematical viewpoint but also viewpoint of it use. Part of the work will be a detailed description of one of the models, namely the classic logistic regression. Logistic regression is also used in the part of finance and therefore the aim of this work will be the application of the model on real data from the part of investment in the financial market according to appropriately chosen aspects.

Keywords

generalized linear model, logistic regression, equity investments

Obsah

Úvod

V této práci se budeme zabývat *Zobecněnými lineárními modely* a to hlavně z matematického hlediska. Tyto modely popisují závislost středních hodnot náhodných veličin na veličinách nenáhodných. Termín „zobecněný“ značí, že náhodné veličiny se řídí rozdělením z rodiny exponenciálního typu, do které patří například – alternativní, binomické, Poissonovo, Gamma rozdělení a další.

V první části teoretického textu se budeme věnovat speciálnímu případu zobecněného lineárního modelu a to tedy *klasickému lineárnímu modelu*, v němž se závislá proměnná řídí normálním rozdělením. Detailně si popíšeme jak tento model funguje, v čem spočívá a jaké se v něm uplatňují metody. Tento jednoduchý model nám bude sloužit pro názornost, abychom následně mohli odvodit vztahy pro jeho zobecnění.

Zobecněný lineární model tedy představuje rozšíření lineárních regresních modelů pro data, která nesplňují všechny předpoklady modelu lineárního. Zobecněný model byl formulován jako způsob, jak sjednotit různé jiné statistické modely, včetně logistické regrese.

Dále se zaměříme konkrétně na **logistickou regresi**, což je model s binární závisle proměnnou. Uvedeme její matematickou definici – základní vlastnosti, odhad parametrů, devianci a celkovou diagnostiku modelu. Podíváme se i na její historii a především na její využití v praxi.

Protože se v praktické části budeme zabývat objekty z *kapitálového trhu*, bude mu věnována i jedna kapitola. Kapitálový trh je jedna z částí finančního trhu, na které dochází k pohybu cenných papírů. Předmětem obchodování je střednědobý a dlouhodobý kapitál, tedy kapitál s nízkou likviditou. Jedná se o cenné papíry s dobou splatnosti více než jeden rok, což jsou například námi zvolené **akcie**, nebo také podílové listy, dluhopisy, a podobně.

V poslední kapitole této práce se pokusíme aplikovat model logistické regrese na reálných datech z oblasti akciových investic. Na těchto datech ukážeme jak tento model funguje a jaký nám přináší užitek. Na základě námi zvolených aspektů, budeme pozorovat vývoj vybrané akcie a pomocí logistické regrese odhadneme její vývoj do budoucna. Podle výsledku se rozhodneme, zda danou akcii „nakoupíme“. Cílem pro nás bude dosáhnout zisku.

Kapitola 1

Zobecněný lineární model

Zobecněný lineární model je překladem z anglického výrazu *Generalized linear model*, a proto je pro něj velmi často používána zkratka **GLM**. Jak už plyne z jeho názvu, je tento model zobecněním klasického lineárního modelu. Umožňuje nám jeho využití i v případě, kdy nejsou splněny předpoklady, které jsou kladeny pro model lineární. Tím nám vzniká celá řada jeho zobecnění a stává se tak prakticky použitelnějším. Budeme se tedy nejdříve věnovat lineárnímu regresnímu modelu a poté přejdeme k definici jeho zobecnění.

1.1 Lineární regresní model

Často se nám v praxi stává, že chceme prozkoumat vztah mezi veličinami, kde na jedné straně stojí tzv. **nezávisle proměnné** X_j kde $j = 1, \ldots, k$, které mají ovlivňovat tzv. **závislou proměnnou** Y, která stojí na straně druhé. Pro naše účely budeme předpokládat, že jsou všechny veličiny spojité.

Prvním krokem pro jejich zkoumání je zakreslení dat do bodového grafu, tzv. **korelačního pole** a ověření toho, zda mezi nimi skutečně existuje předpokládaná závislost, neboli **regrese**. Nejjednodušší formou je *jednoduchá lineární regrese*, která předpokládá lineární závislost pouze mezi dvěma veličinami.

Když zjistíme mezi veličinami nějakou závislost, zajímá nás často také typ a tvar této závislosti. Pro hledání typu, tvaru a konkrétních koeficientů závislosti náhodných veličin používáme regresní metody – příkladem takové metody může být *Metoda maximální věrohodnosti* (viz. kapitola 1.2.1), nebo právě níže popsaná *Metoda nejmenších čtverců* (viz. kapitola 1.1.2).

1.1.1 Zavedení lineárního regresního modelu

Máme naměřenené hodnoty, které zapisujeme do tzv. *matice plánu* $\mathbf{X}$, která má rozměr $n \times k$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ x_{21} & \dots & x_{2k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix}$$

jinak ji také můžeme nazývat *regresní matice* nebo *matice modelu*. Dále uvažujeme vektor hodnot $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^T$, které jsou závislé a hledáme vektor parametrů $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)^T$ takový, že je splněna lineární závislost:

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

kde $\boldsymbol{\varepsilon}$ je vektor náhodných chyb, který má normální rozdělení $N_n(\mathbf{0}, \mathbf{V})$.

Rovnici (1.1) můžeme přepsat ve tvaru:

$$\begin{aligned} Y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1, \\ Y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2, \\ &\vdots \\ Y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n. \end{aligned}$$

Hledáme $\widehat{\mathbf{Y}} = \mathbf{Y} \cdot \mathbf{b}$, což je odhad rovnice (1.1), kde $\mathbf{b} = \widehat{\boldsymbol{\beta}}$ je odhad vektoru parametrů. V odhadu zanedbáváme náhodné chyby. Kvalitu odhadu určuje **reziduální rozptyl** $E[(\mathbf{Y} - \widehat{\mathbf{Y}})^2]$.

Definice 1.1. n-rozměrný vektor $\mathbf{Y}$ se řídí **lineárním regresním modelem** (LRM) s $n \times k$ maticí plánu $\mathbf{X}$, vektorem chyb $\boldsymbol{\varepsilon}$ a vektorem parametrů $\boldsymbol{\beta}$, pokud splňuje (1.1). Přičemž musí být splněny podmínky:

1. $E(\varepsilon) = \mathbf{0}$;
2. $\mathrm{cov}(\mathbf{Y}) = \mathrm{cov}(\boldsymbol{\varepsilon}) = \mathbf{V}$.

Zapisujeme ve tvaru:

$$(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

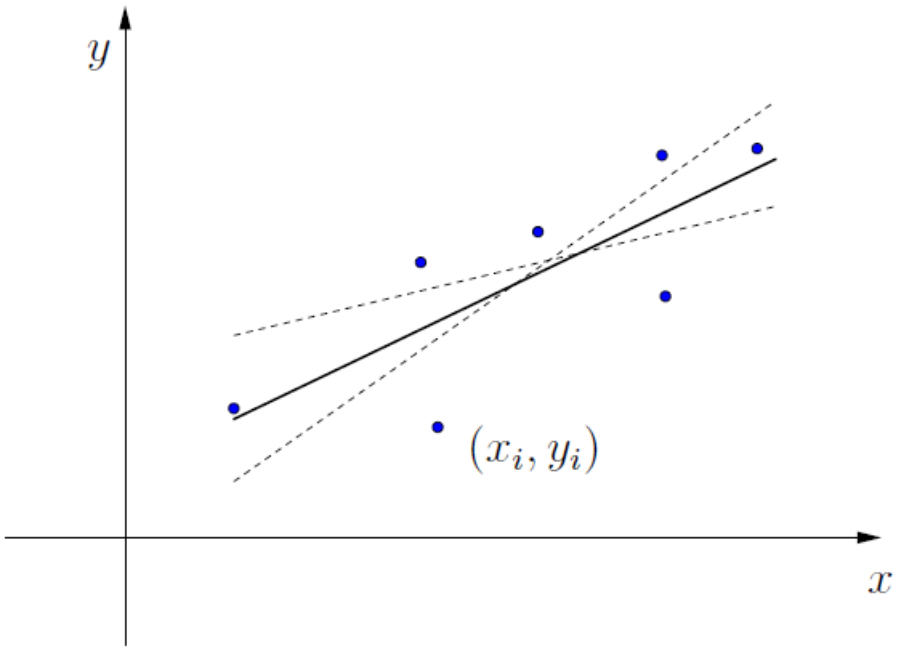
Poznámka 1.1. Střední hodnota vektoru $\mathbf{Y}$ je rovna $\mathbf{X}\boldsymbol{\beta}$, jak můžeme vidět v odvození:

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}$$

Poznámka 1.2. Pokud je $h(\mathbf{X}) = k \leq n$ a matice typu $\mathrm{cov}(Y) = \mathbf{V}$ je regulární, pak se jedná o **model plné hodnosti**. [3] Nejčastěji se setkáváme s případem, kdy $\mathbf{V} = \sigma^2 \boldsymbol{I}$. Pokud existuje pouze jedna nezávisle proměnná X mluvíme o jednoduché lineární regresi – jedná se o klasickou regresní přímku s dvěma parametry $\rightarrow k = 2$. Pokud existuje více nezávislých proměnných X_j mluvíme o mnohonásobné lineární regresi.

1.1.2 Metoda nejmenších čtverců

Příklad 1.1. Mějme naměřeny hodnoty nezávisle (x_i) a závisle proměnné (y_j) a zakreslíme si je do bodového grafu. Následně se je pokusíme aproximovat přímkou. Hledáme takovou přímku, která bude minimalizovat „vzdálenost“ naměřených hodnot tak, aby co nejlépe popisovala sled bodů v grafu (1.1):



Obrázek 1.1: Aproximace bodů přímkou [3]

Snaha aproximovat proměnné co nejjednoduší funkcí je z důvodu toho, abychom mohli lépe popsat jak se data chovají a určit jejich průběh, vlastnosti a vzájemné vztahy. K tomu nám poslouží nejlépe lineární funkce, jejímž grafem je přímka.

Metoda nejmenších čtverců (MNČ) anglicky *Least squares method* je nejpoužívanější metodou při odhadu parametrů lineárních regresních modelů.

Základem této metody je snaha o proložení vhodné funkce mezi naměřenými hodnotami tak, aby vzdálenost od všech naměřených hodnot byla minimální. Z charakteru hodnot předem stanovíme jakého typu bude funkce, kterou chceme najít. Například:

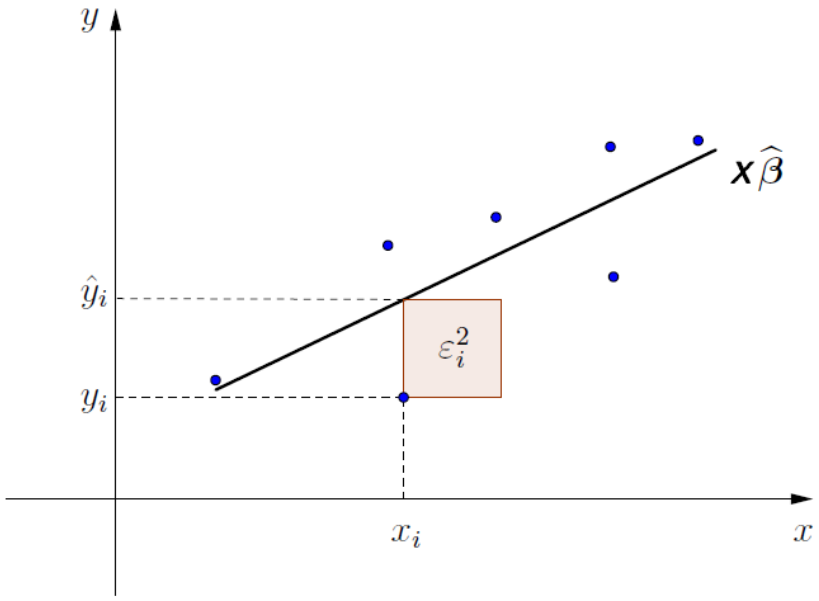
- Lineární aproximace – hledáme funkci typu $y = \beta_1 x_1 + \beta_2$
- Kvadratická aproximace – hledáme funkci typu $y = \beta_1 x_1^2 + \beta_2 x_2 + \beta_3$
- apod.

V našem případě budeme uvažovat pouze **lineární aproximaci**, tedy hledaná funkce bude lineární. Následně MNČ definujeme pomocí geometrické interpretace (viz. Obrázek 1.2).

Definice 1.2. Jsou dány body $[x_0, y_0], [x_1, y_1], \ldots, [x_n, y_n]$. Hledáme funkci typu $y = \beta_1 x_1 + \beta_2$, která aproximuje tyto body tak, že reziduální součet čtverců

$$S_e = \sum_{i=1}^{n} \varepsilon_i^2$$

je minimální.



Obrázek 1.2: Součet čtverců odchylek [3]

Věta 1.1 (Zobecněná metoda nejmenších čtverců)**.** *Mějme lineární model* $(\boldsymbol{Y}, \boldsymbol{X\beta}, \boldsymbol{V})$, *zavedeme funkci*

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{i=1}^{k} x_{ij}\beta_j \right)^2 = \| \boldsymbol{Y} - \boldsymbol{X\beta} \|^2 .$$

Pak tato funkce nabývá svého minima v bodě $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(Y)$, *tedy*

$$\| \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \|^2 = \min_{\beta \in \mathbb{R}^2} \| \boldsymbol{Y} - \boldsymbol{X\beta} \|^2 .$$

Věta 1.2. *Nechť* $(\boldsymbol{Y}, \boldsymbol{X\beta}, \boldsymbol{V})$ *je lineární regresní model plné hodnosti. Odhad* $\widehat{\boldsymbol{\beta}}$ *parametrů* $\boldsymbol{\beta}$ *získaný metodou nejmenších čtverců je ekvivalentní řešení tzv.* ***normálních rovnic***

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$$

tedy

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Důkaz. viz.([3], str.29) □

1.1.3 Koeficient determinace

Koeficient determinace R^2 představuje v matematické statistice **míru kvality regresního modelu**. Tato míra ve své základní podobě vyjadřuje, jaký podíl variability mezi nezávislými proměnnými X_j a závislou proměnnou Y model vysvětluje. Přesněji řečeno udává, kolik procent rozptylu závisle proměnné je vysvětleno modelem a kolik zůstalo nevysvětleno.

Tento koeficient může nabývat hodnoty maximálně 1 (resp. 100 %), což znamená dokonalou predikci hodnot závisle proměnné. Naopak hodnota 0 (resp. 0 %) znamená, že model nepřináší žádnou informaci a je tedy zcela neužitečný.

$$R^2 \in \langle 0; 1 \rangle$$

Koeficient determinace LRM se definuje jako jedna minus podíl reziduálního a celkového součtu čtverců, což můžeme vidět v následujícím vzorci (1.2) [1]:

$$R^2 = 1 - \frac{S_e}{S_t}, \tag{1.2}$$

kde S_e je reziduální součet čtverců a S_t je celkový součet čtverců

$$S_e = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2,$$
$$S_t = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Čím více se hodnota koeficientu blíží k jedničce, tím je regresní model úspěšnější.

1.2 Zobecněný lineární model

Ve statistice je tento model zobecněním klasického lineárního regresního modelu, který jsme definovali v předchozí kapitole (1.1). Pro rekapitulaci si následně uvedeme hlavní rozdíly mezi klasickým lineárním modelem a jeho zobecněním, a dále budeme pokračovat v definici zobecněného lineárního modelu.

LRM

- Závislá proměnná Y má normální rozdělení $\rightarrow Y \sim N(\mu, \sigma^2)$
- $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \rightarrow$ střední hodnota závisle proměnné je lineární kombinací X_j
- Odhad parametrů $\boldsymbol{\beta}$ se provádějí pomocí *metody nejmenších čtverců*
- Kvalita modelu se určuje pomocí *koeficientu determinace* R^2 (resp. reziduálního součtu čtverců)

GLM

- Závislá proměnná Y má rozdělení z rodiny exponenciálního typu, kde θ je přirozený a ϕ je disperzní parametr $\rightarrow$ alternativní, multinomické, binomické, Poissonovo, Gamma, apod. – normální rozdělelní je speciálním případem
- $\mathrm{E}(\mathbf{Y}) = g^{-1}(\mathbf{X}\boldsymbol{\beta}) \rightarrow$ střední hodnota závisle proměnné je funkcí lineární kombinace X_j
- Odhad parametrů $\boldsymbol{\beta}$ se provádí pomocí *metody maximální věrohodnosti*
- Kvalitu modelu určujeme pomocí *deviance*

GLM byl formulován dvěma statistiky *Johnem Nelderem* a *Robertem Wedderburnem* jako způsob, jak sjednotit různé jiné statistické modely, včetně lineární regrese a logistické regrese.

V zobecněných lineárních modelech vystupují náhodné veličiny $Y_1, \ldots, Y_n$ a daná matice plánu $\mathbf{X} = (x_{ij})$ typu $n \times k$. Rozdělení těchto náhodných veličin závisí na daných x_{ij}, které budeme nazývat nezávisle proměnné, neboli **prediktory**. Náhodné veličiny Y_i budeme nazývat závisle proměnné.

První komponentu modelu tvoří vzájemně nezávislé náhodné veličiny $Y_1, \ldots, Y_n$, o kterých předpokládáme, že jsou stejně rozdělené a toto rozdělení patří do rodiny exponenciálního typu s disperzním parametrem ϕ. To znamená, že vyhovují hustotě: [4]

$$f(y_i, \theta_i, \phi) = \exp\left\{ \frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi) \right\}, \tag{1.3}$$

kde θ_i je přirozený parametr. Po funkci $b(\theta_i)$ chceme, aby byla ryze monotónní, dvakrát spojitě diferencovatelná a druhá derivace kladná.

Pro střední hodnotu a rozptyl platí:

$$\mathrm{E}(Y_i) = b'(\theta_i) = \mu_i = \mu_i(\theta_i), \tag{1.4}$$
$$\mathrm{D}(Y_i) = b''(\theta_i)a(\phi). \tag{1.5}$$

Ze vztahu (1.5) je vidět, že rozptyl je funkcí střední hodnoty. Pokud známe vztah mezi střední hodnotou a rozptylem, můžeme určit o jaké rozdělení z rodiny exponenciálního typu se jedná. [4]

Vlastnosti hustoty

Přirozenou třídou hustoty, se kterou pracujeme je třída hustot exponenciálního typu. Řekli jsme, že u lineární regrese, což je speciální případ zobecněného lineárního modelu, má hustota normální rozdělení. Hustota, se kterou budeme pracovat v logistické regresi má rozdělení binomické, a proto si je následně blíže popíšeme:

1. **Normální rozdělení** – Mějme $Y \sim N(\mu, \sigma^2)$, kde platí $\mu \in \mathbb{R}, \sigma^2 > 0$. Pak má hustota tvar:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\Big(\frac{y-\mu}{\sigma}\Big)^2\right\} = \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} + \left(-\frac{1}{2}\frac{y^2}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right)\right\}$$

   kde

$$b(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2 \qquad \Rightarrow b'(\theta) = \theta = \mu$$
$$\Rightarrow b''(\theta) = 1$$
$$a(\phi) = \sigma^2 \qquad \Rightarrow \phi = \sigma^2.$$

   Skutečně platí:

$$\mathrm{E}(Y) = b'(\theta) = \mu,$$
$$\mathrm{D}(Y) = b''(\theta)a(\phi) = \sigma^2.$$

   Tedy pro přirozený paramer platí $\theta = \mu$, a pro disperzní parametr $\phi = \sigma^2$.

2. **Binomické rozdělení** – Mějme $Z \sim Bi(n, \pi)$, kde platí $n \in \mathbb{N}, \pi \in (0, 1)$. Pak má hustota tvar:

$$f_Z(z) = \exp\left\{z\theta - n\ln(1 + \mathrm{e}^\theta) + \ln\binom{n}{z}\right\}$$

kde pro $\pi = \frac{e^\theta}{1+e^\theta}$ platí

$$b(\theta) = n\ln(1+e^\theta) \qquad \Rightarrow b'(\theta) = n\frac{e^\theta}{(1+e^\theta)} = n\pi = \mu$$
$$\Rightarrow b''(\theta) = n\frac{e^\theta}{(1+e^\theta)^2} = n\pi(1-\pi) = \mu\Big(1-\frac{\mu}{n}\Big)$$
$$a(\phi) = 1$$

Skutečně tedy platí:

$$\mathrm{E}(Z) = b'(\theta) = \mu,$$
$$\mathrm{D}(Z) = b''(\theta)a(\phi) = n\pi(1-\pi).$$

Z výše uvedeného vidíme, že pro přirozený parametr platí $\theta = \ln(\frac{\mu}{n-\mu})$. [6]

Druhou komponentu GLM tvoří vektor $(\eta_1, \ldots, \eta_n)$, který je lineární funkcí nezávislých proměnných $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Závislost je ve tvaru[4]:

$$\eta_i = \sum_{j=1}^{k} \beta_j x_{ij}, \qquad i = 1, \ldots, n.$$

Zobecněné lineární modely tedy představují rozšíření lineárních regresních modelů pro data, která nesplňují všechny předpoklady modelu lineárního. Tyto modely mají dva důležité parametry:

Transformační funkce, neboli *link function*, která převádí hodnoty prediktoru na smysluplné hodnoty závislé proměnné;

Typ rozložení dané tak, aby postihlo vztah mezi rozptylem a očekávanou hodnotou y (binomické, Poissonovo, Gamma, exponenciální, apod.).

Tyto dvě výše popsané komponenty jsou propojeny transformační (linkovací) funkcí, která je diferencovatelná a ryze monotónní a platí pro ni vztah:

$$g(\mathrm{E}(Y_i)) = \eta_i. \tag{1.6}$$

Na tomto místě je důležité poznamenat, že v zobecněných lineárních modelech se nevyskytuje chybový člen jako u lineární regrese. Důvodem je, že levá strana rovnice (1.6) je funkcí střední hodnoty veličiny Y_i, a ne pouze samotné veličiny Y_i.

Z výše uvedeného je vidět, že linkovací funkci g i komponenty η_i můžeme navíc chápat jako funkci parametrů β_j nebo θ_i:

$$\begin{aligned} g(E(Y_i)) = g(\mu_i) = g(b'(\theta)) = g(\theta_i) = \eta_i = \sum_{j=1}^{k} \beta_j x_{ij} = \eta_i(\boldsymbol{\beta}), \\ \eta_i = g(EY_i) = \eta_i(\mu_i) = \eta_i(\boldsymbol{\beta}) = \eta_i(\theta_i). \end{aligned} \tag{1.7}$$

Pokud platí, že $g(\mathrm{E}(Y_i)) = g(\theta_i) = \theta_i$, mluvíme o tzv. *kanonické transformační funkci.* Pro přirozený parametr pak dostáváme přímou závislost $\theta_i = \sum_{j=1}^{k} \beta_j x_{ij}$. Z toto vyplývá, že nalezení vhodné spojovací funkce je při aplikaci modelu velice zásadní. [4]

1.2.1 Metoda maximální věrohodnosti

K bodovému odhadu parametrů u lineární regrese se používá metoda nejmenších čtverců, která spočívá v nalezení takových hodnot koeficientů modelu které minimalizují tzv. *součet čtverců*, viz (1.1.2). U složitějších modelů této metody nelze využít kvůli charakteru závisle proměnné, a proto se k odhadu parametrů využívá tzv. *metoda maximální věrohodnosti*, která má širší využití.

Principy metody

Metoda maximální věrohodnosti je nástroj pro jednoduché odhady parametrů, ale i pro netriviální odhady v nelineárních modelech s daty z jiného než normálního rozdělení. Principem metody je najít odhad parametru θ (případně vektoru parametrů $\boldsymbol{\theta}$) – jinými slovy se snažíme najít takovou hodnotu θ pro niž je pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení, **maximální**. Odhad se tedy snaží maximálně přizpůsobit pozorovaným datům.

Uvažujme náhodný výběr $Y_1, \ldots, Y_n$, tedy n nezávislých náhodných veličin se stejným rozdělením pravděpodobnosti s hustotou $f(\mathbf{y}, \boldsymbol{\theta})$, kde $\boldsymbol{\theta}$ představuje vektor neznámých parametrů. Sdružená hustota, případně pravděpodobnostní funkce, odpovídá n realizacím náhodné veličiny Y.

Za předpokladu, že známe $\boldsymbol{\theta}$, vyjadřuje větší hodnota sdružené hustoty větší shodu pozorovaných hodnot s předpokládaným rozdělením s hustotou $f(y, \boldsymbol{\theta})$. Hlavní myšlenkou metody maximální věrohodnosti je dívat se na sdruženou hustotu jako na funkci vektoru $\boldsymbol{\theta}$.

Definice 1.3. Nechť náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ má sdruženou hustotu $f(\mathbf{y}, \boldsymbol{\theta})$, kde $\boldsymbol{\theta} \in \Omega$. Při pevné hodnotě $\mathbf{y}$ se funkce $f(\mathbf{y}, \boldsymbol{\theta})$ jakožto funkce $\boldsymbol{\theta}$ nazývá *věrohodnostní funkce.*
Hodnota $\widehat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}$ která maximalizuje věrohodnostní funkci $f(\mathbf{y}, \boldsymbol{\theta})$ pro dané $\mathbf{Y} = \mathbf{y}$, se nazývá *maximálně věrohodný odhad* parametru $\boldsymbol{\theta}$. [2]

Metoda maximální věrohodnosti má široké využití v matematické statistice, například:

1. při testování hypotéz
2. ve faktorové analýze

Navíc se tato metoda často využívá i v jiných oborech, například:

3. při rozpoznávání objektů v obrazových datech,

4. v ekonometrii a modelování finančních trhů,

5. při lokalizaci (pomocí GPS apod.).

1.2.2 Odhad parametrů v GLM

Předpokládejme, že je dáno n nezávislých náhodných veličin $Y_1, \ldots, Y_N$, které se řídí zobecněným lineárním modelem s transformační funkcí g a hustotou exponeniálního typu (1.3). Hustota veličiny Y_i závisí na prametru θ_i, kde $i = 1, \ldots, n$. Předpokládejme, že disperzní parametr ϕ je známý pro všechna pozorování $Y_1, \ldots, Y_N$.

Pro střední hodnotu Y_i dostaneme

$$\mu_i = \mathrm{E}(Y_i) = b'(\theta_i), \qquad i = 1, \ldots, n$$

Pomocí tranformační funkce g lze lineární prediktor

$$\eta_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$

vyjádřit jako funkci střední hodnoty μ_i ve tvaru

$$\eta_i = g(\mu_i), \qquad i = 1, \ldots, n.$$

Uvedené vztahy využijeme při odvozování věrohodnostních rovnic pro výpočet odhadů neznámých parametrů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$.

Věrohodnostní rovnice

Mějme $\ell(\boldsymbol{\theta}, \phi) = \prod_{i=1}^{N} f(Y_i; \theta_i, \phi)$ *věrohodnostní funkci* a její logaritmus $L(\boldsymbol{\theta}) = \log(\ell(\boldsymbol{\theta}, \phi))$. Jelikož víme, že $\mu_i = b'(\theta_i)$ a funkce b je ryze monotónní (existuje k ní i její derivace inverzí funkce), pak můžeme chápat θ_i jako $\theta_i = \theta_i(\mu_i)$. Obdobně pak dle (1.7) můžeme psát $\mu_i = \mu_i(\eta_i)$ a $\eta_i = \eta_i(\boldsymbol{\beta})$. Proto můžeme uvažovat $L(\boldsymbol{\theta}) = L(\boldsymbol{\beta})$.

Maximálně věrohodným odhadem parametru $\boldsymbol{\beta}$ (kde $\boldsymbol{\beta}$ je vektor parametrů modelu) je hodnota, při které je maximalizována *věrohodnostní funkce* a tedy i její logaritmus

$$L(\boldsymbol{\beta}) = \log(\ell(\boldsymbol{\beta})) = \sum_{i=1}^{N} \left\{ \frac{[Y_i \theta_i - b(\theta_i)]}{a(\phi)} + c(Y_i, \phi) \right\}. \tag{1.8}$$

Funkce je maximalizována pro hodnotu, kdy $\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = 0$ pro $j = 1, \ldots, k$. Použitím řetízkového pravidla dostaneme

$$\frac{\partial L}{\partial \beta_j} = \frac{\partial L}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} \cdot \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} \cdot \frac{\partial \boldsymbol{\eta}}{\partial \beta_j} \qquad j = 1, \ldots, k. \tag{1.9}$$

Užitím vztahů (1.4) a (1.5) máme $\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = \frac{\mathrm{D}(Y_i)}{a(\phi)}$. Podle věty o derivaci inverzní funkce víme, že $\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}$. Dosazením do výrazu (1.9) získáme soustavu rovnic

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_i} = \sum_{i=1}^{N} \left[\frac{Y_i - \mathrm{E}(Y_i)}{\mathrm{D}(Y_i)} \frac{1}{g'(\mu_i)} x_{ij} \right] = 0 \qquad j = 1, \ldots, k. \quad [4] \tag{1.10}$$

Výsledná soustava rovnic (1.10) lze interpretovat také tak, že odhad parametru $\boldsymbol{\beta}$ metodou maximální věrohodnosti závisí na středních hodnotách a rozptylech náhodných veličin $Y_1, \ldots, Y_N$, nikoliv na jejich rozdělení. Navíc z předchozí kapitoly víme, že v případě daného hustotou (1.3) je rozptyl funkcí střední hodnoty. Za použití kanonické transformační funkce se soustava (1.10) zjednoduší na tvar

$$\sum_{i=1}^{N} \frac{Y_i - b'(\theta_i)}{a(\phi)} x_{ij} \qquad j = 1, \ldots, k.$$

Pokud je $a(\phi)$ konstantní, pak hledaný odhad vyhovuje maticovému zápisu

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\widehat{\boldsymbol{\mu}},$$

kde $\mathbf{X}$ je matice $N \times k$ s hodnotami x_{ij} a $\widehat{\boldsymbol{\mu}}$ značí odhad střední hodnoty. [4] Pro zobecněný lineární model s kanonickou transformační funkcí vyhovují odhady *soustavě normálních rovnic,* neboli maximalizací věrohodnostní funkce docházíme ke stejným výsledkům, k jakým bychom se dostali aplikací *metody nejmenších čtverců.*

1.2.3 Testování podmodelu

K testování podmodelu lze použít test poměrem věrohodnosti založený na odhadech v modelu a v podmodelu. Test se zpravidla provádí prostřednictvím tzv. *deviancí*, které si nyní zavedeme.

Deviance

Deviance v zobecněných lineárních modelech je obdobou rozptylu u klasických lineárních regresních modelů. Je tedy kritériem vhodnosti GLM.

Označme si devianci proměnnou D. Čím je náš model méně přiléhavý, tím je hodnota D větší, podobně, jako je větší reziduální součet čtverců v lineárním modelu. [1]

Pro ověření, zda je model vhodně použit, se zavádí tzv. *nejbohatší model*, který se nazývá *saturovaný*, nebo také *nasycený*. Takový model obsahuje odlišný parametr pro každé pozorování. Sám o sobě není použitelný, ale může sloužit jako dobrá výchozí pozice pro testování správnosti ostatních modelů.

Uvažujeme tedy saturovaný model, který obsahuje odlišný parametr pro každé pozorování. Označme $\widetilde{\theta}_i$ jako odhad parametru θ_i pro saturovaný model. Z definice nasyceného modelu dostáváme, že $\widetilde{\mu}_i = Y_i$ pro každé $i = 1, \ldots, N$. $\widehat{\theta}_i$ a $\widehat{\mu}_i$ jsou příslušné odhady pro testovaný model.

Dále $L_{\max}$ značí maximální hodnotu logaritmické věrohodnostní funkce pro saturovaný model a L_1 pro testovaný. Dle (1.8) a možným zjednodušením $a(\phi) = \frac{\phi}{\omega_i}$, kde $\omega_i > 0$

jsou známé apriorní váhy a $\phi > 0$ je neznámý parametr, který se též nazývá škálovým či rozptylovým parametrem.

Při testování vhodnosti modelu hraje velmi důležitou roli tzv. *škálová deviance*, kterou můžeme vyjádřit takto:

$$\frac{D(\mathbf{Y}, \widehat{\boldsymbol{\mu}})}{\phi} = -2(L_1 - L_{\max}) = 2\sum_{i=1}^{N} \omega_i \frac{[Y_i(\widetilde{\theta_i} - \widehat{\theta_i}) - b(\widetilde{\theta_i}) + b(\widehat{\theta_i})]}{\phi}.$$

Nazvěme dále $D(\mathbf{Y}, \widehat{\boldsymbol{\mu}})$ *deviance.* Aplikaci deviance lze zobecnit na srovnávání dvou nesaturovaných do sebe vnořených modelů, a tím získáme test poměrem věrohodností

$$-2(L_1 - L_2) = -2(L_1 - L_{\max}) - [-2(L_2 - L_{\max})] = D(\mathbf{Y}; \widehat{\boldsymbol{\mu}}_{\mathbf{1}}) - D(\mathbf{Y}; \widehat{\boldsymbol{\mu}}_{\mathbf{2}}).$$

Tato statistika má za platnosti testovaného podmodelu asymptoticky rozdělení χ^2_{n-k}, kde $n - k$ je rozdíl nezávislých parametrů v testovaných modelech. To znamená, že test poměrem pravděpodobností pro dva do sebe vnořené modely je jednoduše rozdílem jejich škálovaných deviancí.[4]

1.3 Logistická regrese

Logistická regrese, též také nazývána *logistický regresní model* nebo *model logistické regrese* je jeden z případů zobecněných lineárních modelů s binární závisle proměnnou.

Byla navržena v 60. letech 20. století jako alternativa k *metodě nejmenších čtverců*, kterou jsme definovali v kapitole (1.1.2). Dříve se logistická regrese používala v několika oblastech:

Lékařství a farmacie, ve kterých se logistická regrese používala a dodnes používá nejčastěji. Závislá proměnná Y představuje např. přítomnost nebo nepřítomnost choroby. Tato regrese pak umožňuje modelovat např. riziko vzniku srdeční choroby jako funkci různých parametrů (pohlaví, věk, BMI, krevní tlak, hladina cholesterolu, kouření, apod.)

Průmysl je jednou z oblastí, kde můžeme sledovat úspěšnost nebo neúspěšnost nějakého výrobku a logistickou regresí lze určit, které veličiny se na úspěšnosti výrazně podílejí.

Bankovnictví – zde se používá logistická regrese k vytvoření modelů, které dokáží odhadnout na základě řady parametrů o klientovi banky (např. věk, pohlaví, nejvyšší dosažené vzdělání) žádajícím o úvěr, jestli bude tento úvěr splácet řádně či nikoliv.

V této práci využijeme logistickou regresi na poli kapitálového trhu, pomocí níž budeme rozhodovat o koupi či nekoupi akcie na základě zvolených parametrů tak, abychom dosáhli zisku.

Logistická regrese se od lineární liší v tom, že predikuje pravděpodobnost toho, zdali se nějaká událost stane či nestane. Hlavním rozdílem ale je, že logistická regrese používá kategorickou závislou proměnnou Y, kdežto u linerání je závislá proměnná spojitá.

Podle typu závislé proměnné Y rozlišujeme:

Binární logistickou regeresi – týká se binární závisle proměnné, která nabývá pouze dvou možných hodnot, např. přítomnosti či nepřítomnost jevu

Ordinální logistickou regresi – závislou proměnnou je veličina ordinálního typu, nabývající více možných stavů, mezi nimiž existuje přirozené uspořádání, např. stadium závažnosti nějakého onemocnění

(Multi)nominální logistickou regresi – týká se nominální závisle proměnné o více než dvou úrovních stavů, mezi nimiž existuje pouze odlišnost, např. barva očí, rasa, apod.

Obdobně jako u lineární regrese, vektor nezávislých proměnných u všech třech druhů logistické regrese může obsahovat více proměnných, a to jak spojitých – **prediktory**, tak kategoriálních – **faktory**.[5] V dalším textu se pod pojmem *logistická regrese* bude rozumět *binární logistická regrese.*

Proč tedy nepoužijeme jednodušší model, když se jedná o binární nezávislou proměnnou – lineární regresi? Jedním z důvodů je, že se mnohem častěji vyskytuje nelineární vztah mezi podmíněnou pravděpodobností náhodné nezávislé veličiny a nenáhodnými prediktory. Dalším důvodem je, že daná funkce není omezena v intervalu $(0; 1)$, což pro odhad parametrů binomického rozdělení je potřeba.

Logistická regrese se používá při modelování pravděpodobnosti nějakého jevu v závislosti na hodnotě spojité proměnné. Předpokládá se, že náhodná proměnná Y má alternativní rozdělení s parametrem π, který odpovídá pravděpodobnosti výsledku 1 a mění se monotónně s hodnotou nezávisle proměnné. Výsledný model je právě odhadem tohoto parametru v závislosti na $\mathbf{x} \to P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$.

1.3.1 Základní vlastnosti

Model logistické regerese umožňuje analýzu dat, kdy náhodná veličina Y nabývá hodnot 0 a 1, případně poměry v itervalu $\langle 0; 1\rangle$. Předpokládejme, že závisle proměnná Y je binární proměnná, která nabývá hodnoty jedna, pokud sledovaný jev nastal, v opačném případě je rovna nule. Jelikož se jedná o regresní model, bude nás zajímat vztah pravděpodobností úspěchu či neúspěchu k hodnotám regresorů $\mathbf{x} = (x_1, \ldots, x_k)^T$ a budeme tedy zkoumat pravděpodobnost [6]:

$$P(Y = 1|x_1, \ldots, x_k) = \pi(\mathbf{x}),$$
$$P(Y = 0|x_1, \ldots, x_k) = 1 - \pi(\mathbf{x}).$$

Předpokládejme, že lineární prediktor je roven

$$\eta(\mathbf{x}) = \alpha + \boldsymbol{\beta}\mathbf{x}.$$

Jelikož platí:

$$\mu = \mathrm{E}(Y|\mathbf{x}) = 1 \cdot P(Y = 1|\mathbf{x}) + 0 \cdot P(Y = 0|\mathbf{x}) = P(Y = 1|\mathbf{x}) = \pi(\mathbf{x}),$$

pak dle (1.6) dostáváme

$$\pi(\mathbf{x}) = \mu = g^{-1}(\eta(\mathbf{x})) = g^{-1}(\alpha + \boldsymbol{\beta}\mathbf{x}).$$

Hledáme tedy takovou linkovací funkci g, jejíž inverze zobrazí hodnoty prediktorů na interval $\langle 0; 1\rangle$. Případně je možné se na daný vztah dívat i obráceně, tedy hledáme takovou linkovací funkci $g(\pi(\mathbf{x}))$, která transformuje hodnotu podmíněné pravděpodobnosti na celá reálná čísla.

Pro tento účel zavedeme podíl pravděpodobností:

$$\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})},$$

který se nazývá *poměr šancí* (anglicky odds ratio). Tento poměr nám již umožní trasformovat hodnoty pravděpodobností na interval $\langle 0; \infty)$. Pro úpravu na celou reálnou osu ještě využijeme funkci $\log(x)$, čímž dostáváme vztah:

$$\log \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \alpha + \boldsymbol{\beta}\mathbf{x}. \tag{1.11}$$

Jelikož náhodná veličina Y má aleternativní rozdělení, což je speciálním případem rozdělení binomického s parametrem $n = 1$, pak transformace prametru π v tomto tvaru vychází přímo ze vztahu popsaného v 2 (popis binomického rozdělení u funkce hustoty, str. 17). Po úpravě získáme

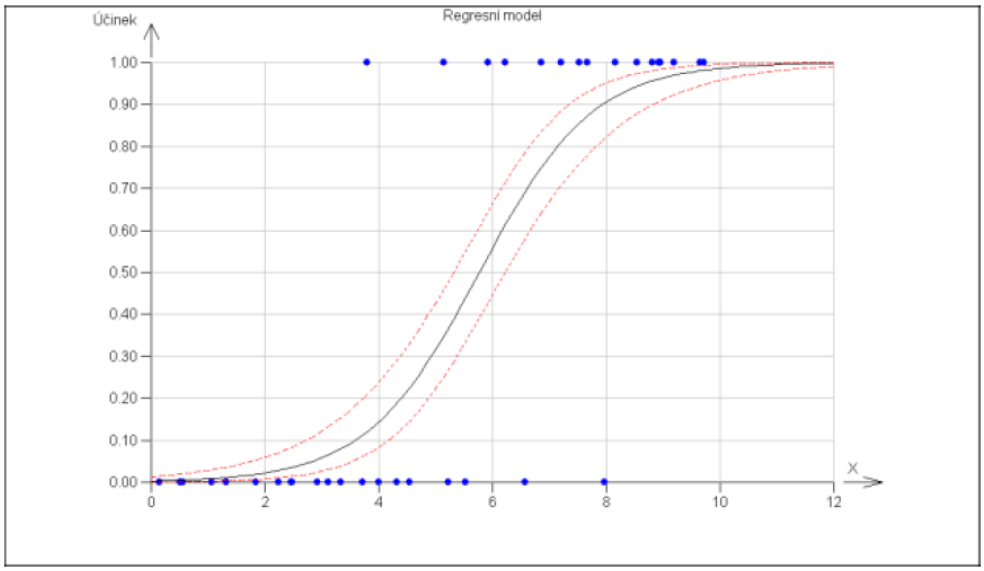
$$\pi(\mathbf{x}) = \frac{1}{1 + \exp(-(\alpha + \boldsymbol{\beta}\mathbf{x}))} = \frac{1}{1 + \exp(-\eta(\mathbf{x}))} \tag{1.12}$$

$$1 - \pi(\mathbf{x}) = \frac{\exp(-(\alpha + \boldsymbol{\beta}\mathbf{x}))}{1 + \exp(-(\alpha + \boldsymbol{\beta}\mathbf{x}))} = \frac{\exp(-\eta(\mathbf{x}))}{1 + \exp(-\eta(\mathbf{x}))} \tag{1.13}$$

Výraz na levé straně rovnice (1.11) se nazývá **logit**, neboli *logistická funkce*. V kontextu s zobecněným lineráním modelem je logit speciálním případem transformační funkce (*link function*) a vidíme, že jde dokonce o kanonickou transformační funkci. Hodnoty α a $\boldsymbol{\beta}$ jsou regresní koeficienty. Z omezení $\pi(\mathbf{x}) \in (0, 1)$ dostáváme, že logistická funkce je dobře definována.[5]

Data a parametry

Data musí zahrnovat jeden nebo více sloupců nezávisle proměnné x_i a jeden sloupec závisle proměnné y. Nezávisle proměnné mohou nabývat číselných hodnot. Binární závisle proměnná musí mít hodnoty 0 nebo 1, které odpovídají výskytu či nevýskytu sledovaného jevu, což můžeme v grafické podobě vidět na obrázku (1.3).



Obrázek 1.3: Logistický model pro binární data[7]

Na volbě zda 1 bude výskyt a 0 nevýskyt či naopak nezáleží, je však důležité označení zachovávat. Binární proměnná odpovídá případu, kdy pro danou hodnotu prediktoru máme pouze jediný výsledek typu *ano/ne*.

Proměnná, neboli frekvenční závisle proměnná odpovídá případu, kdy pro danou hodnotu prediktoru X provedeme s testů a máme tedy s výsledků z nichž t je pozitivních a $(s-t)$ negativních. Do sloupce závisle proměnné pak můžeme zapsat poměr $\frac{t}{s}$, popř. $\frac{(s-t)}{s}$.

Interpretace parametru β

Pro lepší představu odvodíme pro jednu nezávislou proměnnou x.

1. **Analytický pohled**
   Kladné (resp. záporné) znaménko nám říká, zda je pravděpodobnost $\pi(x)$ rostoucí (resp. klesající). Sklon $\pi(x)$ je dán velikostí $|\beta|$. Pokud zderivujeme $\pi(x)$ dle x

$$\begin{aligned}\frac{\mathrm{d}\pi}{\mathrm{d}x}(x) &= \frac{\beta\{\exp(\alpha+\beta x)[1+\exp(\alpha+\beta x)] - [\exp(\alpha+\beta x)]^2\}}{[\exp(\alpha+\beta x)]^2} \\ &= \beta\pi(x)[1-\pi(x)],\end{aligned}$$

   snadno nahlédneme symetrii, křivka $\pi(x)$ se blíží k jedné pod stejným úhlem jako k nule a dosahuje maximálního sklonu pro $\pi(x) = \frac{1}{2}$. Pro tuto hodnotu je argument v exponenciále roven nule, které se nabývá pro $x = \frac{-\alpha}{\beta}$. [4]

2. **Statistický pohled**
Hodnota e^{β} vyjadřuje **poměr šancí** v modelu pro prediktory $x+1$ a x. Platí[4]

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

Šance vyjadřuje, kolikrát se zvýší pravděpodobnost toho, že dojdeme k úspěchu (hodnoty v modelu nabudou 1), když se x zvýší o jedna. Je třeba si uvědomit, že šance není pravděpodobnost, ale jedná se pouze o poměr pravděpodobností – pravděpodobnosti výskytu sledovaného jevu P_1, a jejím doplňku $1–P_1$. Šanci (*odds*), která odpovídá výskytu sledovaného jevu v dané skupině, můžeme vyjádřit pomocí následujícího vztahu:

$$oods = \frac{P_1}{1-P_1},$$

Poměr šancí (*Oods Ratio*), jak již vyplývá z názvu, je pak dán poměrem dvou šancí, které odpovídají srovnávaným skupinám, experimentální a kontrolní. Poté ho můžeme obecně definovat vztahem:

$$OR = \frac{\frac{P_1}{1-P1}}{\frac{P_0}{1-P_0}}.$$

1.3.2 Odhady parametrů pro model logistické regrese

Tak jako jsme si určili odhad parametrů pro zobecněný lineární model, určíme si následně odhad parametrů konkrétně pro model logistické regrese. Odhad parametru $\boldsymbol{\beta}$ určíme *metodou maximální věrohodnosti* (1.2.1).

Věrohodnostní rovnice

Mějme N nezávislých náhodných veličin s binomickým rozdělením $Y_i \sim Bi(n_i, \pi(\mathbf{x}_i))$, které odpovídají pozorováním v bodech $\mathbf{x}_i$, kde $i = 1, \ldots, N$. Pak platí:

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{N} \binom{n_i}{y_i} (\pi(\mathbf{x}_i))^{yi} (1-\pi(\mathbf{x}_i))^{n_i - y_i}. \tag{1.14}$$

Opět budeme pracovat s logaritmem funkce $\ell(\boldsymbol{\beta})$. Označme $\mathbf{b}$ hodnotu, při které je maximalizována $L(\boldsymbol{\beta}) = \log \ell(\boldsymbol{\beta})$. Před dalším odvozováním je vhodné si celý výraz upravit a použít vztahů

$$\log \frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} = \sum_{j=1}^{k} \beta_j x_{ij}, \tag{1.15}$$

$$1 - \pi(\mathbf{x}_i) = \frac{1}{1 + \exp(\sum_{j=1}^{k} \beta_j x_{ij})}. \tag{1.16}$$

Užitím vztahů (1.15) a (1.16) se logaritmus výrazu (1.14) zjednoduší na tvar

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ \log \binom{n_i}{y_i} + y_i \sum_{j=1}^{k} \beta_j x_{ij} + n_i \log \left[\frac{1}{1 + \exp(\sum_{j=1}^{k} \beta_j x_{ij})} \right] \right\}.$$

Nutná podmínka pro hledání extrémů funkce $L(\boldsymbol{\beta})$ dává rovnici

$$0 = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N} \left[y_i x_{ij} - n_i x_{ij} \frac{\exp(\sum_{j=1}^{N} \beta_j x_{ij})}{1 + \exp(\sum_{j=1}^{N} \beta_j x_{ij})} \right], \quad j = 0, \ldots, k.$$

Dostáváme nelineární rovnici pro odhad parametru $\boldsymbol{\beta}$

$$\sum_{i=1}^{N} [y_i x_{ij} - n_i x_{ij} \widehat{p}(\mathbf{x}_i)] = 0, \quad j = 0, \ldots, p.$$

Výsledkem je, že maximálně věrohodný odhad $\mathbf{b}$ splňuje nelineární rovnici: [4]

$$\mathbf{x}'\mathbf{y} = \mathbf{x}' \frac{\exp(\mathbf{b}'\mathbf{x})}{1 + \exp(\mathbf{b}'\mathbf{x})}.$$

1.3.3 Deviance v logistické regresi

Stejně tak jako jsme definovali pro zobecněný lineární model devianci, zaměříme se na ní i u logistické regrese. Definujeme si ji z hlediska použití v modelu.

Deviance se zpravidla využívá jako nástroj pro testování podmodelu, proto si ho teď zavedeme. K testování podmodelu lze použít test *poměrem věrohodnosti* založený na odhadech $\mathbf{b}$ v modelu a $\widetilde{\mathbf{b}}$ v podmodelu. Provedeme ho právě pomocí deviancí. Uvažujme tedy model, který má právě tolik parametrů, kolik je různých hodnot vektorů $\mathbf{x}_i$. Tento model je saturovaný a jeho maximální hodnotu věrohodnostní funkce si označíme $L_{\max}$. Přiléhavost modelu můžeme posoudit pomocí

$$D(\mathbf{b}) = 2(L_{\max} - L(\mathbf{b})).$$

Předpokládejme, že všechny vektory $\mathbf{x}_i$ jsou různé. Saturovaný model má pak n parametrů $\mu_1, \ldots, \mu_n$. Odhadem střední hodnoty μ_i je přímo Y_i a bez důkazu uvedeme, že pro $L_{\max}$ platí:

$$L_{\max} = \sum_{i=1}^{N} (Y_i \log Y_i + (1 - Y_i) \log(1 - Y_i)) = 0. \tag{1.17}$$

Důkaz. viz.([1], str. 179) □

Devianci v modelu logistické regerese vyjádříme pomocí (1.17) jako:

$$D(\mathbf{b}) = -2L(\mathbf{b}) = -2 \sum_{i=1}^{n} (Y_i \log \widehat{\mu}_i + (1 - Y_i) \log(1 - \widehat{\mu}_i)).$$

Chceme–li porovnat nějaký obecný model M a jeho podmodel $\widetilde{M}$, použijeme test poměrem věrohodností, a to pomocí deviancí modelu a podmodelu. Testovou statistikou je

$$\begin{aligned} 2\big(L(\mathbf{b}) - L(\widetilde{\boldsymbol{b}})\big) &= \big(2(L_{\max} - L(\widetilde{\boldsymbol{b}}))\big) - \big(2\big(\ell_{\max} - L(\mathbf{b})\big)\big) \\ &= D(\widetilde{\boldsymbol{b}}) - D(\mathbf{b}). \end{aligned} \tag{1.18}$$

Tato testová statistika (1.18) má za platnosti testovaného podmodelu rodělení χ^2_f, kde f je rovno rozdílu počtu nezávislých parametrů v pozorovaných modelech. [1]

Nulovou hypotézu H_0: *$\widetilde{M}$ je podmodelem modelu M* zamítáme na hladině významnosti $\boldsymbol{\alpha}$, když $D(\widetilde{\mathbf{b}}) - D(\mathbf{b}) \geq \chi^2_{1-\alpha}(f)$.

„*Testování podmodelu se prakticky používá:*

1. *pro testování významnosti celého modelu porovnáním daného modelu s tzv. nulovým modelem $P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1-e^{-b_0}}$, tj. testuje se hypotéza*

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \quad \times \quad H_1 : \text{alespoň jedno } \beta_i \neq 0,$$

2. *pro ověření, zda nějaký podsoubor regresorů $\beta_{i1}, \ldots, \beta_{ik}$ (s výjimkou absolutního členu β_0) významně přispívá k vysvětlení variability závisle binární proměnné testováním hypotézy*

$$H_0 : \beta_{i1} = \cdots = \beta_{ik} = 0 \quad \times \quad H_1 : \text{alespoň jedno } \beta_{ik} \neq 0.$$

Hypotézu $H_0 : \beta_i = 0$ o nulovosti jednotlivých koeficientů modelu lze tedy testovat pomocí testů poměru věrohodnosti.“ [5]

1.3.4 Diagnostika

Pokud již máme vytvořený logistický regresní model, potřebujeme také posoudit, jak kvalitně nám popisuje data. K tomu se využivá řada koeficientů a grafických pomůcek, které si následně stručně popíšeme.

Skóre a prahový bod

Nechť tedy máme logistický regresní model M se závislou proměnnou Y a nazávisle proměnné $X_1, \ldots, X_k$. Tento model pak každému i-tému případu s realizací y_i přiřadí na základě jemu příslušných naměřených hodnot nezávislých proměnných $x_{i1}, \ldots, x_{ik}$ pravděpodobnost, že tato realizace nabude hodnoty 1. Tato pravděpodobnost se nazývá *skóre* nebo také *logitové skóre*

$$s_i = P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i).$$

Skóre tedy ohodnocuje každé měření zvlášť a tak umožňuje posoudit kvalitu modelu. Za ideální lze považovat takový model, který naměřeným hodnotám $y_i = 1$ přiřadí $s_i = 1$

a naopak hodnotám $y_i = 0$ přiřadí $s_i = 0$. S tak ideálním případem modelu, se ale v praxi nelze setkat. Cílem je vytvořit model, který se alespoň co nejvíce k ideálnímu blíží. [5]

Někdy nastává situace, kdy je skóre modelu např. $s_i \approx 0,5$. Pak vzniká dilema, do jaké skupiny model přiřadit. Z toho důvodu se zavádí tzv. *prahový bod*, označme ho P_C, který zařadí případ do správné skupiny:

$$\begin{cases} \text{pokud} & s_i > P_C & \text{pak přiřadí model do} & y_i = 1 \\ \text{pokud} & s_i \leq P_C & \text{pak přiřadí model do} & y_i = 0 \end{cases}$$

Koeficienty determinace

Koeficient determinace jsme si již definovali pro lineární model pomocí reziduálního součtu čtverců a vzhledem k určité podobnosti mezi LRM a logistickým modelem, byly snahy rozšířit koeficient determinace i na tento model.

Uvažujme tzv. *nulový model*, tj. model, který predikuje vždy stejnou pravděpodobnost $P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{1+\mathrm{e}^{-b_0}}$. Označme devianci tohoto modelu D_0. Porovnejme nulový model s nějakým jiným logistickým modelem, např. $P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{1+\mathrm{e}^{-x'b}}$, s deviancí $D(\mathbf{b})$. Potom lze zavést (důkaz: [1], str. 180):

$$\textbf{McFaddenův koeficient determinace}: \quad R_L^2 = 1 - \frac{D(\mathbf{b})}{D_0},$$

$$\textbf{Coxův-Snellův koeficient determinace}: \quad R_{CS}^2 = 1 - \mathrm{e}^{\frac{D(\mathbf{b})-D_0}{n}},$$

$$\textbf{Nagelkerkův koeficient determinace}: \quad R_N^2 = \frac{1 - \mathrm{e}^{\frac{D(\mathbf{b}-D_0)}{n}}}{1 - \mathrm{e}^{\frac{-D_0}{n}}}.$$

Čím je model M více vzdálen od nulového modelu tím jsou koeficienty determinace R_L^2, R_{CS}^2 a R_N^2 dále od nuly.

AIC

Akaike information criterion v překladu Akaikeho informační kritérium slouží k posouzení schopnosti různých modelů vysvětlit variabilitu v pozorovaných datech.

Předpokládejme, že máme statistický model, počet odhadovaných parametrů k a odhad maximální hodnoty pravděpodobnostní funkce $\widehat{\ell}$, pak pro toto informační kritérium můžeme zavést vztah:

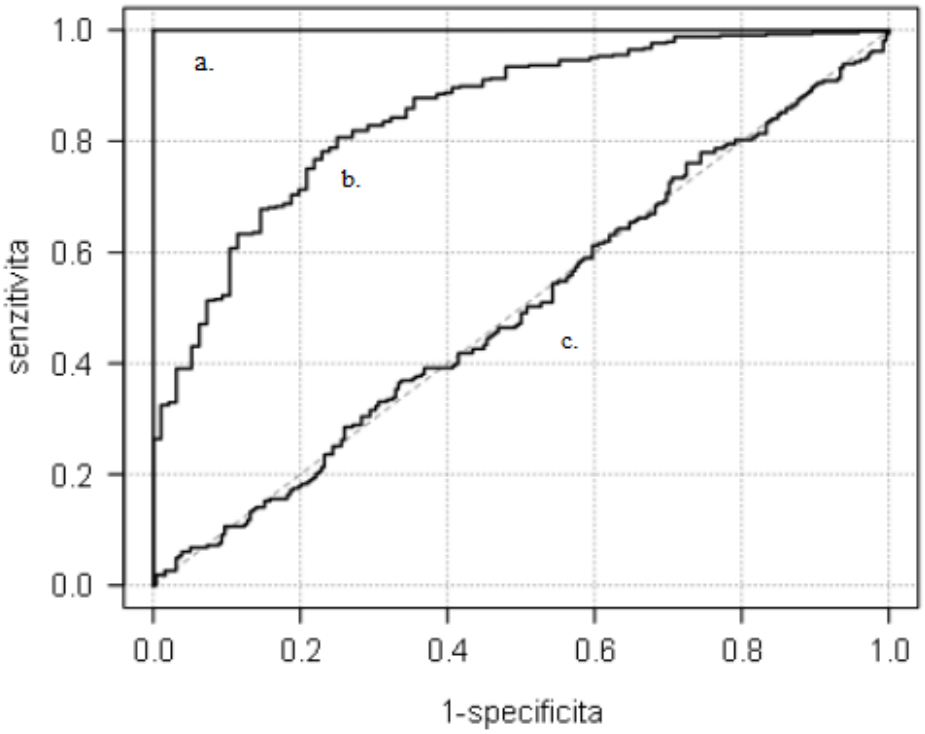
$$AIC = 2k - 2\ln(\widehat{\ell}).$$

Preferovány jsou modely s nižšími hodnotami, protože indikují lepší schopnost modelu „sedět“ na pozorovaná data.

ROC křivka

ROC je zkratka pro *Receiver Operating Characteristic*, tedy ROC křivka znamená v překladu *křivka operační prahové chrakteristiky*. Tato křivka je nástroj, který se v logistické re-

gresi používá pro hodnocení kvality vytvořeného modelu pomocí grafického znázornění, jak můžeme vidět na obrázku (1.4).



Obrázek 1.4: ROC křivky [5]

a. Ideální případ

b. Reálný případ

c. Náhodný případ

ROC křivku tedy můžeme definovat jako **graf**, který popisuje kvalitu binárního klasifikátoru, který se zabývá roztříděním dané množiny objektů na základě přítomnosti či nepřítomnosti určité vlastnosti v závislosti na klasifikačním prahu. Zjednodušeně lze říci, že nastavováním různých prahových hodnot hledáme na ROC křivce kompromis mezi množstvím falešně pozitivních (FP) a falešně negativních (FN) pozorování.

Křivka je úzce svázána s pojmy *senzitivita* a *specificita*, protože ukazuje vztah mezi nimy. Test je tím užitečnější, čím je jejich senzitivita a specificita vyšší.

Senzitivita je poměr správně pozitivních pozorování ku všem pozitivním případům, neboli ukazuje jaký podíl skutečných případů test zachytí.

Specificita je poměr správně negativních pozorování ku všem negativním případům, neboli ukazuje jaký podíl negativních případů test správně vyloučí.

Za ideální lze považovat takový případ, pro který budou senzitivita a specificita rovny 1.

Kapitola 2

Akciové investice

V této kapitole, budou nastíněny základní informace o kapitálových trzích, akciích a především o akciových investicích.

2.1 Kapitálový trh

Kapitálový trh je jedna z částí finančního trhu, na které dochází k pohybu cenných papírů, a kde předmětem obchodování je střednědobý a dlouhodobý kapitál, tedy kapitál s nízkou likviditou. Jedná se o cenné papíry s dobou splatnosti více než jeden rok, což jsou např. akcie, podílové listy, dluhopisy, apod.

Můžeme zde najít dvě strany. Na jedné straně stojí emitenti a na straně druhé investoři.

Emitent je společnočnost, či jiný oprávněný subjekt, který vydává cenné papíry a získává tak na kapitálovém trhu finanční zdroje ke svému podnikání. V případě akcií je emitentem akciová společnost.

Investor je fyzická či právnická osoba, která na kapitálovém trhu nakupuje cenné papíry. Je tedy jejich vlastníkem a klade za cíl zhodnotit své volné finanční zdroje.

Z hlediska emitenta se může jednat o vlastní nebo cizí zdroje, podle typu cenného papíru.[9]

Výhoda takového obchodování spočívá v tom, že investoři mohou cenné papíry směňovat navzájem a k tomu obvykle slouží burzy. Díky tomu lze kapitálový trh rozdělit na 2 části:

Akciový trh (Equity securities) – je označení pro všechny burzy, trhy a subjekty, na kterých se obchoduje s akciemi, nebo kde se vydávají.

Trh dluhopisů (Debt securities) – je označení pro finanční trh, který nabízí a prodává dluhopisy. Zahrnuje všechny subjekty, které vydávají, prodávají nebo jinak obchodují s dluhopisy.

Regulaci nad kapitálovým trhem má v rukou ministerstvo financí České republiky a jeho pravomoc je zakotvena v zákoně č.2/1969 Sb. Jeho hlavními úlohami jsou:

- vytvářet koncepci politiky,
- provádět jeho analýzu,
- vytvářet zákony,
- zajišťovat členství v mezinárodních finančních institucích a orgánech Evropské unie.[9]

Působnost Ministerstva financí v oblasti právní úpravy kapitálového trhu lze rozdělit do těchto oblastí:

- Podnikání na kapitálovém trhu
- Investiční fondy
- Cenné papíry a finanční zajištění

2.2 Akcie

Akcie, je dlouhodobý majetkový cenný papír, který na trh emituje akciová společnost. Investor, který akcii nakoupí, se stává akcionářem společnosti a tím získává určitý podíl na základním kapitálu emitenta.

Podíl na akciové společnosti je dán nominální hodnotou akcie, což je hodnota, za kterou akcii koupil. Pro akcionáře je však důležitá a také ho více zajímá tzv. *tržní hodnota*, tj. skutečná hodnota akcie, neboli také kurz akcie. Ta se v čase mění a odvíjí se od aktuální finanční situace společnosti. Tržní hodnota akcie se odvíjí od:

- situace na trhu,
- situace na burze cenných papírů,
- skutečné finanční situace,
- oceněné hodnoty podniku.

Z držení akcií vyplývá také výplata dividendy, což je výnos, který plyne akcionáři z vlasnictví akcie. Je to jeho podíl na zisku akciové společnosti a o její výši rozhoduje valná hromada společnosti. Výplata dividendy a tržní hodnota akcie spolu úzce souvisí. Změna tržní hodnoty, tedy kurzu akcie, má klíčový vliv na podíl majetku akcionáře. Finanční investice akcionáře je nevratná, ale může ji na kapitálovém trhu směnit. [10]

Akcie se dělí do několika skupin, podle několika různých hledisek.

1. Druhy akcií podle práv držitelů:

   Kmenové akcie jsou jedním z nejběžnějších typů akcií, a proto se také označují jako základní či běžné. Akcionářům poskytují právo na dividendu a možnost hlasování na valné hromadě, ale neposkytují jim žádná zvláštní práva.

   Přednostní akcie jinak také nazývány prioritní, umožňují přednostní výplatu dividendy a podílu na likvidačním zůstatku před akcionáři držícími kmenové akcie. Práva držitelů však mohou být stanovami akciové společnosti omezeny – např. nemožností hlasovat na valné hromadě.

   Zaměstnanecké akcie jsou vydávané společnostmi pro potřeby svých zaměstnanců. Jejich podíl nesmí překročit 5 % základního jmění společnosti. Tyto akcie mohou mít podle vnitřních stanov společnosti určité výhody.

2. Akcie podle formy své existence:

   Listinné akcie existují ve fyzické podobě a akcionář je má u sebe uschovány.

   Zaknihované akcie existují v elektronické podobě a jsou zapsané v registru cenných papírů.

3. Akcie dle formy vlastnictví a možností převodu:

   Akcie na jméno jsou vázané přímo na jméno konkrétní fyzické či právnické osoby.

   Akcie na doručitele nejsou vázané na jméno konkrétního akcionáře a výkon práv s nimi spojenými může provádět ten, kdo je jejich majitelem. Jejich převod je možný pouhým předáním cenného papíru.

2.3 Ekonomická data a trh

Zprávy hýbou trhy. Načasování většiny zpráv je nepředpověditelné – jako v případě válek, politických změn a přírodních katastrof. Oproti tomu, zprávy založené na datech z ekonomiky přicházejí v předem ohlášených časech, které jsou nastaveny rok nebo více dopředu. Prakticky všechna prohlášení mají co dělat s ekonomikou, zejména hospodářský růst a inflace mají potenciál výrazně pohnout trhy. Ekonomická data nejenže vytvářejí rámec pro způsob, jakým obchodníci nahlížejí na ekonomiku, ale také mají dopad na očekávání obchodníků, jak bude centrální banka provádět svou měnovou politiku. Silnější hospodářský růst nebo vyšší inflace zvyšuje pravděpodobnost, že centrální banka buď zpřísní, nebo přestane uvolňovat měnovou politiku. Všechna tato data ovlivňují očekávání obchodníků ohledně budoucího směru úrokových sazeb, ekonomiky a hlavně cen akcií. [8]

„*Trhy nereagují přímo na to, co je ohlášeno. Reagují spíše na rozdíl mezi tím, co obchodníci očekávali, že se stane a co se skutečně děje. Důvodem, proč trhy reagují pouze na rozdíl mezi očekáváním a tím co se skutečně nastane, je, že ceny cenných papírů již zahrnují všechny očekávané informace.* “ [8]

2.4 Měření rizika a výnosů

Vztah mezi rizikem a výnosy je základním ukazatelem při investování. Jakmile je stanoveno riziko a očekávaná výnosnot každého aktiva, můžeme sestavit pro investora nejlepší investiční portfolio.

Riziko a výnosy akcií jsou veličiny, které se nedají snadno měřit. Abychom mohli určit velikost rizika, nebo její výnosnost, musíme provést analýzu minulosti vývoje akcie abychom tak pochopili její budoucnost.

Výnos podléhá určitým změnám a je ovlivňván nahodilými faktory. Právě tato nahodilost vnáší na finanční trhy riziko. Pokud se na výnos z investiční možnosti díváme jako na náhodou veličinu X, vhodným nástrojem pro posouzení výnosu je její střední hodnota $E(X)$ a k posouzení rizika je to její rozptyl $D(X)$.

2.5 Měřítka pro ocenění akcií

V této kapitole budou popsána měřítka, která se používají k ocenění akcií a některá z nich nebo jejich poměry budeme posléze používat v rozhodování o koupi či prodeji akcií v praktické části této práce.

2.5.1 Poměr ceny a zisku

Poměr ceny a zisku, neboli P/E ratio je nejzákladnější měřítko pro ocenění akcie. Tento poměr je zobrazován prostým podílem ceny akcie a ročních zisků na akcii, což znázorňuje následující vztah:

$$\text{P/E ratio} = \frac{\text{cena akcie}}{\text{roční zisk na akcii}}.$$

Tento ukazatel měří, kolik je investor ochoten zaplatit za dolarovou hodnotu aktuálních zisků.

Jednoznačně nejdůležitější proměnnou určující P/E ratio pro jednotlivou akcii je očekávání budoucího růstu zisků. Růst zisků však není jediným faktorem ovlivňující tento poměr. Poměry P/E jsou také ovlivněny dalšími faktory jako jsou úrokové sazby, postoj investorů k riziku, daně a likvidita. [8]

V souvislosti s poměrem ceny a zisku je definován také tzv. **PEG poměr**, neboli PEG ratio, který představuje P/E ratio v závislosti na růstu (*Growth*). Tento poměr popisuje relativní cenu mezi cenou akcie, výnosem generovaným na akcii (EPS viz. 2.5.5) a očekávaným růstem dané společnosti.

$$\text{PEG ratio} = \frac{\text{P/E ratio}}{\text{EPS}}.$$

2.5.2 Účetní hodnota akcie

Účetní hodnota neboli tzv. *book value* je dalším častým ukazatelem pro ocenění akcií. Často je pro ni používána zkratka BV. Tento ukazatel označuje dosahování zisku dané společnosti v minulém a aktuálním období, který je rozdělován mezi stát (daně), vlastníky (dividendy) a podnik (reinvestice). Zobrazuje tedy podíl mezi vlastním kapitálem a množstvím emitovaných kmenových akcií, jak můžeme vidět v následujícím vztahu:

$$\text{Book value} = \frac{\text{vlastní kapitál}}{\text{počet kmenových emitovaných akcií}}.$$

Tento poměr využívají výhradně investoři, při hodnocení činnosti společnosti.

2.5.3 Sharpeho poměr

Problém mezi vztahem výnosnosti a rizikovosti, nám řeší *Sharpe ratio* neboli Sharpeho poměr. Tento poměr je jedním z nejznámějších a také nejčastěji používaných koeficientů k hodnocení fondů. Podle něj je výhodnější ta investice, jejíž poměr výnosu nad bezrizikovou mírou a rizikem je vyšší. Sharpeho poměr bere v potaz celkové riziko, a proto se hodí také pro porovnávání fondů napříč všemi kategoriemi.

Vzorec pro výpočet Sharpeho poměru se dá popsat jako:

$$\text{Sharpe ratio} = \frac{\text{výnos sledované investice} - \text{bezrizikový výnos}}{\text{riziko}}.$$

Riziko v tomto případě chápeme jako směrodatnou odchylku výnosu nad bezrizikový výnos. Čím větší je číslo, které po výpočtu dostaneme, tím je vyšší výnos, který investice dosáhla na jednotku rizika. [8]

V praxi Shrarpeho poměr dosahuje u stadratních akciových portfolií hodnoty kolem 0,5.

2.5.4 Míra volatility

Volatilita jako taková označuje v jaké míře kolísá hodnota daného aktviva nebo jeho míry výnosu. V našem případě je dané aktivum akcie. Obvykle je označována jako směrodatná odchylka změn za dané časové období.

Jedná se o nástroj, pomocí kterého lze s určitou pravděpodobností zjistit jaký bude potenciální výnos nebo ztráta na dané akcii na základě vývoje změn hodnot v minulosti.

2.5.5 Zisk na akcii

Zkratka EPS, neboli *Earnings Per Share* je v doslovném překladu do češtiny zisk na akcii. Je definována jako zisk připadající na jednu kmenovou akcii společnosti.

Zisk na akcii je obecně považován za jeden z nejdůležitějších faktorů při stanovení ceny. Je také hlavní složkou používanou k výpočtu poměru „P/E ratio“ (viz 2.5.1).

Na tento pojem lze nahlížet jako na čistý zisk na akcii, protože je po zdanění. Můžeme ho označit i jako tzv. *rentabilitu* na jednu akcii. Základní vztah pro výpočet čistého zisku na akcii popisuje následující vzorec: [11]

$$\text{Zisk na akcii} = \frac{\text{Čistý zisk společnosti}}{\text{počet emitovaných akcií společnosti}}.$$

2.6 Investiční filozofie

Každý investor, předtím než začně s investováním, si musí ujasnit svou tzv. *investiční filozofii*. Investiční filozofie představuje souhrn základních principů, které formulují investiční proces investora. Popisuje přímo jeho strategii v investování, také to jakou chová averzi vůči riziku apod. Dalo by se říci, že je tolik investičních filozofií, kolik je investorů.

Podle přístupu k trhu lze strategie rozlišit na:

Aktivní strategii, kdy se investor snaží „porazit trh“ a dosáhnout tak vyšší výnosnosti, než jaká se na trhu nachází.

Pasivní strategii, kdy se investor snaží dosáhnout přibližně stejné výnosnosti jakou trh v danou chvíli nabízí.

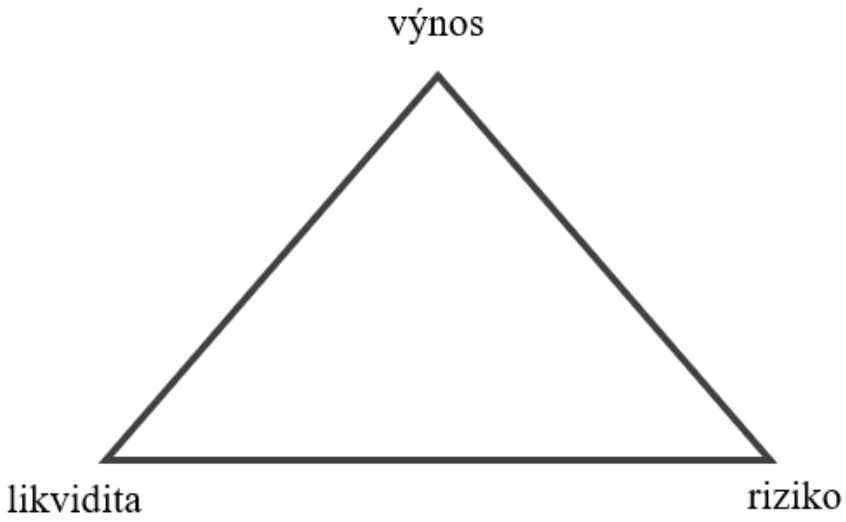
Investor posuzuje každou investici přinejmenším z hlediska tří základních faktorů, tzv. *magický trojúhelník investování*, který můžeme vidět níže na obrázku (2.1):

- očekávaný výnos investice,
- očekávané riziko investice,
- očekávaný důsledek na likviditu investora.

Investor se vždy pohybuje uvnitř tohoto trojúhelníku a nikdy nemůže dospět k ideální investici, která by současně maximalizovala výnos, byla by zcela bezpečná a bylo by možné ji okamžitě proměnit v hotové peníze.

Podle cíle investora a také podle toho jaké faktory upřednostňuje, jsou celé řady označení pro investiční filozofie, jako např.

- hodnotová
- růstová
- technická
- apod.



Obrázek 2.1: Trojúhelník investování

Jelikož žádná z nich nemá objektivně platnou definici a popis, v podstatě není důležité jak ji nazveme. Duležité však je, abychom ji měli ujasněnou. [12]

Kapitola 3

Aplikace modelu logistické regrese

V této části práce se budeme zabývat aplikací modelu logistické regrese v praxi. Jak už jsme zmínili v kapitole dříve, logistická regerse pracuje na principu 0 a 1, kdy na základně zvolených aspektů rozhoduje o „úspěchu“ a „neúspěchu“. Pokusíme se tedy v této části ukázat, jak tento model pracuje.

3.1 Popis datového souboru

Logistická regrese se využivá v různých oblastech, a jednou z nich je právě i finanční trh. Aplikujeme tedy model na reálná data z oblasti kapitálového trhu – konktrétně z části akciových investic.

Informace o vývoji subjektů (akcií), se kterými budeme následně pracovat, jsme získali z online investičního portálu finance.yahoo.com [14], kde je možné obchodovat s cennými papíry. Všechny akciové sploločnosti byly vybrány ze seznamu společností v akciovém indexu **S&P 500** (viz. [13]), což je 500 „největších“ veřejně obchodovaných firem na americké burze. Abychom byli schopni odhadnout, zda daný subjekt bude v budoucnosti dosahovat nějakého zisku, musíme ho nějaký čas pozorovat, proto o každém subjektu zjišťujeme vývoj zpětně.

Model budeme vytvářet s 25 subjekty – o každé akcii známe kvartální vývoj následujích pěti vybraných aspektů (viz. 3.1), s jejichž pomocí budeme určovat, zda danou akcii „nakoupíme“ či ne. Hodnoty aspektů jsou brány vždy k poslednímu dni kvartálu roku 2019.

P/E ratio:	poměr ceny a zisku, podle něhož očekáváme budoucí růst zisků
PEG ratio:	poměr P/E ratio a čistého zisku na akcii
Price/Sales:	poměr mezi cenou a prodejem
Price/Book:	účetní hodnota akcie
EV/Revenue:	poměr hodnoty a výnosu společnosti (EV = *enterprice value*)

Tabulka 3.1: Vysvětlivky aspektů

Od každého aspektu známe jeho průměrnou hodnotu za daný kvartál. Model budeme konstruovat ještě podle šestého aspektu – Koupim – což je tzv. *dichotomická veličina* nabývají hodnot 0 a 1, kde 0 značí logickou hodnotu FALSE a 1 logickou hodnotu TRUE. Tomuto aspektu jsem hodnotu přiřadila podle toho, zda hodnota akcie v dalším období stoupala nebo klesala, na základně prvních pěti aspektů.

Hodnota akcie	*Koupíme*	*Přiřazená hodnota*
Stoupla	ano	1
Klesla	ne	0

Tabulka 3.2: Postup přiřazení hodnoty

Akcie byly pro rozmanitost a určitou univerzálnost modelu vybrány z několika různých odvětví:

- energetika
- moderní technologie
- potravinářství
- komunikační služby
- platební služby
- zdravotnictví
- atd.

Sledované subjekty

Abychom si více přiblížili do jakých odvětví se chystáme investovat, zjistíme si základní informace o vybraných akciových společnostech.

Abbot Laboratories je akciová společnost z oblasti zdravotní péče, která celosvětově prodává zdravotnické potřeby, léky a výživové výrobky.

AES Corporation je jednou z předních světových energetických společností, která vyrábí a distribuuje elektrickou energii v 15 státech.

Amazon, Inc. je společnost, která provozuje internetový obchod Amazon.com, což je jeden z největších obchodních řetězců světa.

Apple, Inc. je akciová společnost, která rozvíjí moderní technologie a specializuje se na hardware a software. Jejich produkty jsou např. stolní počítače, notebooky, chytré telefony, hodinky, apod. s operačním systémem iOS.

eBay, Inc. je společnost, která spravuje internetové stránky eBay.com, což je online aukční a nákupní webová stránka, na které lidé a firmy nakupují a prodávají širokou škálu zboží a služeb po celém světě.

Equinix se specializuje na připojení k internetu a datová centra. Má vedoucí postavení na globálním trhu datových center ve 25 zemích.

Facebook je akciová společnost, která spravuje rozsáhlý společenský webový systém – umožňuje komunikaci mezi uživateli napříč světem, sdílení multimediálních dat, udržování vztahů a zábavu.

Gartner, Inc. je akciová společnost z oblasti výzkumu a poskytuje poradenské služby – poskytující informace, rady a nástroje pro vedoucí pracovníky v oblasti IT, financí, lidských zdrojů, zákaznického servisu a podpory, komunikace, právní předpisy a dodržování předpisů, marketing a prodej.

Hilton Worldwide Holdings, Inc. je nadnárodní pohostinná společenost, která spravuje široké portfolio hotelů a letovisek.

Intuit, Inc. je americká obchodní a finanční softwarová společnost, která vyvíjí a prodává finanční, účetní a daňový software.

Jacobs Engineering Group, Inc. je mezinárodní společnost z oblasti stavebního inženýrství.

Johson & Johson je farmaceutická společnost, která vyrábí léky, zdravotnické prostředky, a také toaletní a hygienické zboží.

Juniper Networks, Inc. je americká akciová společnost, vyrábějící síťová zařízení.

Las Vegas Sands Corporation je společnost, která provozuje kasino a letovisko – zprostředkovávají ubytování, hry a zábavu, kongresové a výstavní prostory, restaurace a kluby.

Monster Beverage Corporation je společnost z oblasti potravinářství, která vyrábí energetické nápoje.

Netflix, Inc. je společnost, která je poskytovatelem online filmů.

PepsiCo je společnost, zabývající se výrobou a prodejem nápojů a potravin.

Prologis, Inc. je realitní společnost, která investuje do logistických zařízení.

ResMed je společnost, která se pohybuje na trhu zdravotního průmyslu – poskytuje primárně zdravotnické prostředky – masky, respirátory, apod.

Teleflex Incorporated je poskytovatelem speciálních zdravotnických prostředků primárně do oblasti chirurgie.

The Walt Disney Company se stala lídrem v animačním průmyslu – zprostředkovává filmové produkce, televize a zábavní parky.

Tyson Foods působí v potravinářském průmyslu. Je druhým největším výrobcem a prodejcem kuřecího, hovězího a vepřového masa na světě.

Visa, Inc. provozuje největší světovou síť elektronických plateb, správu plateb mezi finančními institucemi, obchodníky, spotřebiteli a orgány státní správy.

Wal-mart Stores, Inc. je obchodní společnost, která provozuje řetězec prodejen Wal-mart.

Zoetis je největším světovým výrobcem léčiv a očkování pro domácí a hospodářská zvířata.

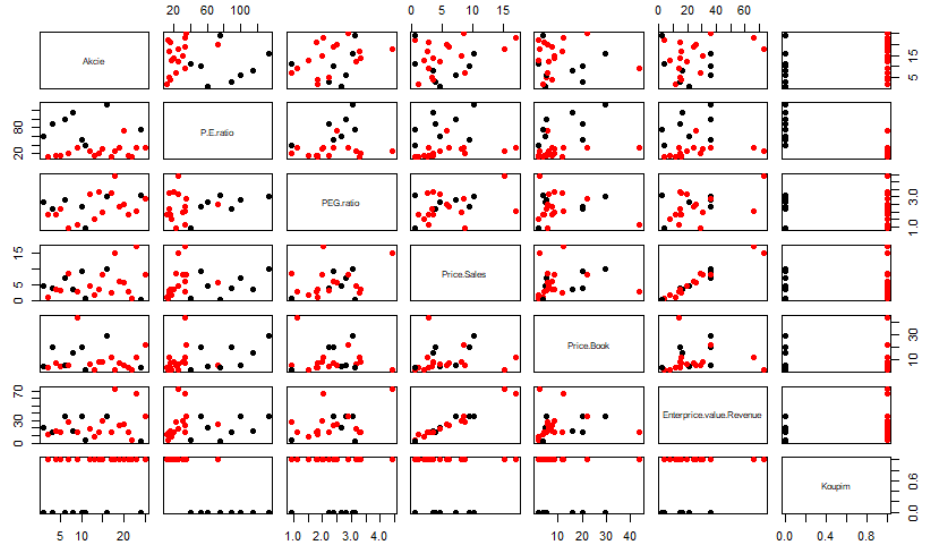
3.2 Konstrukce modelu

K vytvoření modelu budeme používat aplikaci **RStudio**, což je vývojové prostředí umožňující práci v programovacím jazyce **R**.

V této aplikaci budeme vytvářet logistický model, abychom mohli analyzovat naše data. Nejprve ho budeme vytvářet pro každý kvartál zvlášť a poté se ho pokusíme spojit, abychom došli k predikci vývoje pro další období. Vývoj akcie budeme sledovat za období prvních třech kvartálů roku 2019, na jehož základě bychom měli dojít k rohodnutí, které akcie „nakoupíme“, a které ne. Poslední kvartál daného roku budeme pozorovat z důvodu ověření, zda naše predikce byla správná a zda jsme dosáhli zisku.

3.2.1 I. kvartál

Jako první si do prostředí nahrajeme data z prvního kvartálu, které si vhodně pojmenujeme, a vykreslíme si je pomocí příkazu `View(data1)`. Podle vykresleného grafu (3.1) se budeme snažit posoudit, podle které proměnné by mohl být model průkazný.



Obrázek 3.1: Vykreslení dat

Podle obrázku (3.1) vidíme, že hodnota u proměnné *Koupim* by nám mohla ukazovat, že čím menší je hodnota u *P.E.ratio*, tím spíše danou akcii nakoupíme. Proto si vykreslíme konkrétní graf závisloti *P.E.ratio* na *Koupim*, abychom viděli jak jsou hodnoty rozložené.

Poté si vytvoříme logistický model, pomocí příkazu `fit_logistic<-glm(Koupim~P.E.ratio,data=data1,family="binomial")` a získáme tak výsledek, který si zobrazíme pomocí příkazu `summary(fit_logistic)`:

```
Call:
glm(formula = Koupim ~ P.E.ratio, family = "binomial", data = data1)

Deviance Residuals:
        Min       1Q       Median       3Q        Max
-1.79225  -0.06143   0.20035   0.33018   2.03670

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.46112    1.98235   2.755  0.00587 **
P.E.ratio   -0.10288    0.04154  -2.477  0.01326 *
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 31.343  on 24  degrees of freedom
Residual deviance: 11.645  on 23  degrees of freedom
AIC: 15.645

Number of Fisher Scoring iterations: 6
```

Když se podrobněji podíváme na to, co nám říká výsledek modelu, tak se jako první zastavíme u části *Coefficients*. První řádek nám ukazuje jaké má vlastnosti regresní koeficient α a druhý řádek obdobně popisuje koeficent β – v prvním sloupci (*Estimate*) vidíme přímo hodnotu koeficientu a poslední sloupec, neboli tzv. *test na parametr* nám říká, jak je daný koeficient statisticky významný.

Můžeme tedy dosadit konkrétní hodnoty parametrů do vztahu pro logistický model (1.12):

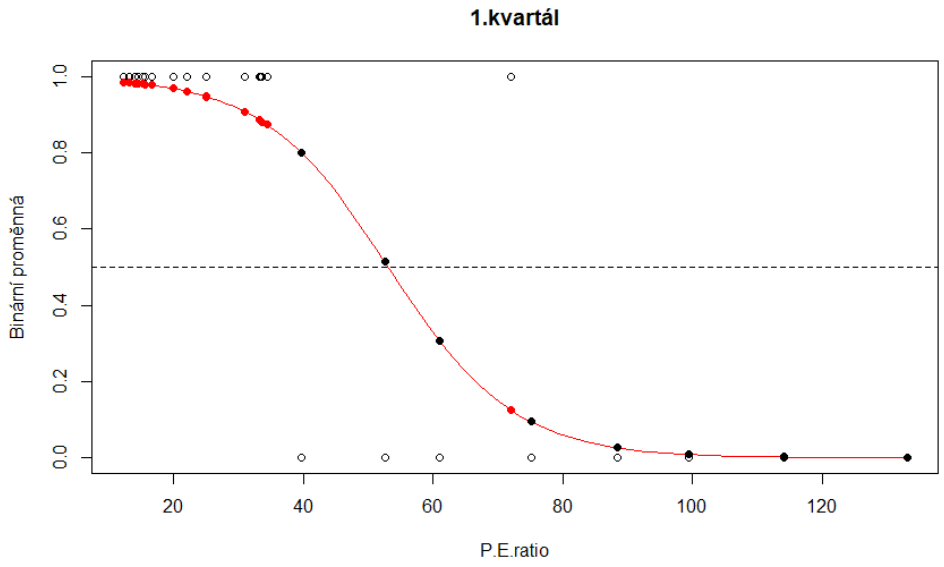
$$\widehat{\pi}(x) = \frac{1}{1 + \exp(-(5.46112 + (-0.10288x))}.$$

Hodnota 0.01326 v posledním sloupci nám říká, že daný koeficient na hladině významnosti 5% zamítáme – neboli jinak řečeno, že daný koeficient je pro náš model vysoce statisticky významný.

Poté se zastavíme u hodnoty *AIC*, která nám slouží k porovnání modelu a submodelu. Čím je hodnota nižší, tím model lépe sedí na pozorovaná data. Naše hodnota 15.645 je v pořádku.

Nakonec si vykreslíme graf modelu, který můžeme vidět na obrázku (3.2). Na tomtu grafu vidíme regresní křivku modelu, na které jsou umístěné hodnoty akcií. Pokud se hodnoty nacházejí pod dělící křivkou, přiřadíme jim hodnotu 0, tedy danou akcii nekoupíme. Naopak pokud se hodnoty nachází nad dělící křivkou, přidělíme jim hodnotu 1, tedy danou akcii nakoupíme.

Červené body zobrazují akcie, které bychom měli nakoupit, protože oproti minulému období vzrostly a zvyšuje se tak pravděpodobnost toho, že budou růst dále. Černé body zobrazují opačný případ, tedy akcie, které oproti minulému období klesly. Můžeme tedy v grafu vidět, že dvě akcie označené černě se nacházejí nad dělící čarou – to jsou hodnoty tzv. *False positive*, neboli falešně pozitivní (FP), což znamená, že bychom dané akcie nakoupili, ikdyž bychom neměli. Obdobný případ můžeme vidět i pod dělící čarou, kde se nachází jedna červeně označená akcie – to je hodnota tzv. *False negative*, neboli falešně negativní(FN), protože bychom danou akcii nekoupili, ikdyž bychom měli.



Obrázek 3.2: Vykreslení modelu za 1. kvartál

3.2.2 II. kvartál

Pro analýzu dat totoho kvartálu budeme postupovat obdobně, jako jsme postupovali v předchozí části.

Jako první si vykreslíme data pro druhý kvartál, která jsme si nahráli do prostředí aplikace. Poté si vykreslíme konktrétní graf závislosti *P.E.ratio* na *Koupim* a vytvoříme si logistický model. Následně si zobrazíme jeho výsledek, abychom mohli popsat jeho chování.

```
Call:
glm(formula = Koupim ~ P.E.ratio, family = "binomial", data = data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7598  -0.2410  -0.0009   0.3636   1.5311

Coefficients:
             Estimate   Std. Error z value Pr(>|z|)
(Intercept)  5.22898    2.33409   2.240   0.0251 *
P.E.ratio   -0.15198    0.07083  -2.146   0.0319 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 34.617  on 24  degrees of freedom
```

```
Residual deviance: 17.011  on 23  degrees of freedom
AIC: 21.011

Number of Fisher Scoring iterations: 7
```

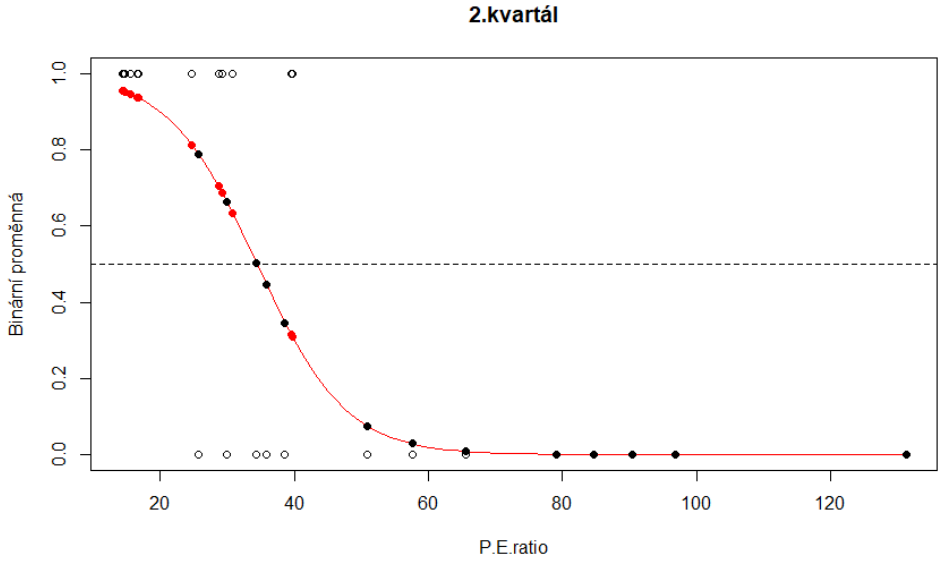
Opět nás budou zajímat hodnoty pro regresní koeficienty α a β, které najdeme v části *Coefficients* v prvním sloupci. Můžeme si je tedy dosadit do vztahu (1.12), z čehož dostaneme průběh funkce modelu:

$$\widehat{\pi}(x) = \frac{1}{1 + \exp(-(5.22898 + (-0.15198x))}.$$

Hodnota 0.0251 v posledním sloupci nám říká, že daný koeficient na hladině významnosti 5% zamítáme – neboli jinak řečeno, že daný koeficient je pro náš model vysoce statisticky významný.

Co se týče hodnoty AIC – 21.011 tak nám oproti minulému kvartálu nepatrně stoupla, ale stále o ní můžeme říci, že je v normě.

Následně si vykreslíme graf logistického modelu pro 2. kvartál, který můžeme vidět na obrázku (3.3).



Obrázek 3.3: Vykreslení modelu pro 2.kvartál

Na daný graf nahlížeme stejně jako na graf (3.2). V grafu můžeme vidět regresní křivku, na které jsou umístěné hodnoty akcií. Pokud se hodnoty nacházejí pod dělící křivkou, přiřadíme jim hodnotu 0, tedy danou akcii nekoupíme. Naopak pokud se hodnoty nachází nad dělící křivkou, přidělíme jim hodnotu 1, tedy danou akcii nakoupíme. Opět se nám v grafu vyskytují hodnoty FP a FN.

3.2.3 III. kvartál

Posouváme se tedy do druhého pololetí roku 2019, a na data ze třetího kvartálu budeme nahlížet obdobně jako na data z prvního pololetí.

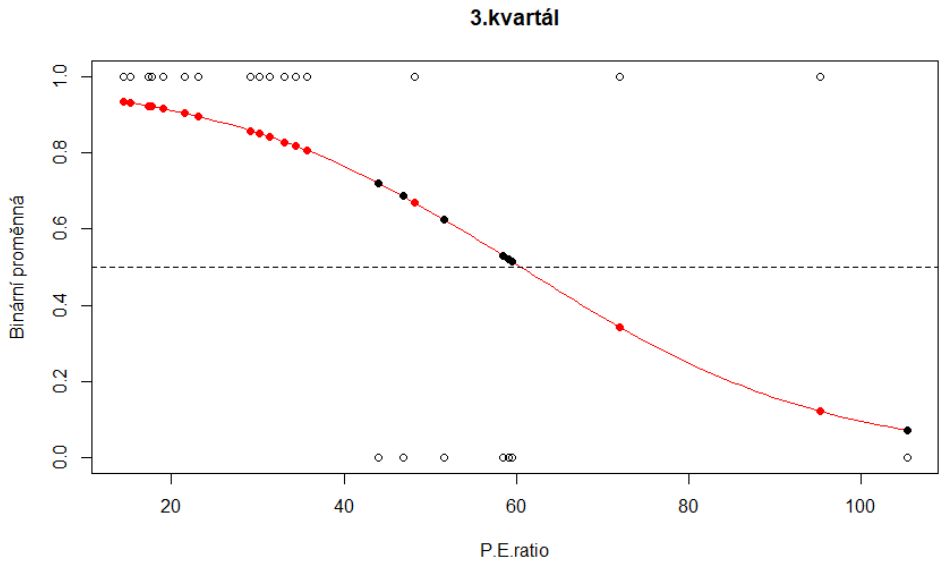
Opět si nahrajeme data do vývojového prostředí Rstudia, vhodně si je pojmenujeme a vykreslíme do grafů. Podle něj si vybereme konkrétní graf, který nám ukáže, podle které proměnné bude model průkazný. Opět je to graf závislosti *P.E.ratio* na proměnné *Koupim*. Na základě této závislosti vytvoříme logistický model a zobrazíme si jeho výsledek:

```
Call:
glm(formula = Koupim ~ P.E.ratio, family = "binomial", data = data3)

Deviance Residuals:
     Min      1Q      Median    3Q      Max
-1.5962  -0.3869   0.4222   0.5874   2.0552

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.45788    1.30372   2.652  0.00799 **
P.E.ratio   -0.05707    0.02644  -2.159  0.03088 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 29.648  on 24  degrees of freedom
Residual deviance: 22.457  on 23  degrees of freedom
AIC: 26.457

Number of Fisher Scoring iterations: 4
```

Opětovně nás z výsledku budou zajímat především hodnoty regresních koeficientů α a β abychom mohli odhadnout průběh funkce modelu. Hodnoty si dosadíme do předpisu (1.12):

$$\widehat{\pi}(x) = \frac{1}{1 + \exp(-(3.45788 + (-0.05707x))}.$$

Tímto předpisem získáme regresní křivku, kterou můžeme vidět na obrázku (3.4).



Obrázek 3.4: Vykreslení modelu pro 3. kvartál

Na regresní křivce jsou umístěny hodnoty akcií. Těm co se nacházejí pod dělící křivkou přiřadíme hodnotu 0, tedy danou akcii nekoupíme. Naopak pokud se hodnoty nachází nad dělící křivkou, přidělíme jim hodnotu 1, tedy danou akcii nakoupíme.

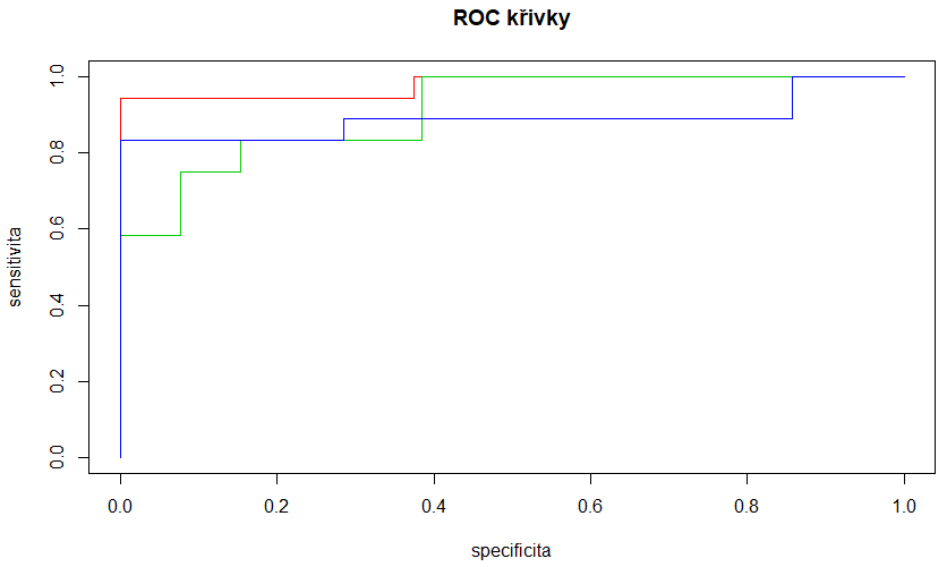
Podle barevného označení můžeme vidět, že se nám v modelu opět nacházejí hodnoty FP a FN.

3.2.4 ROC křivky modelu

ROC křivka modelu nám měří kvalitu testu – zjednodušeně lze říci, že hledáme na ROC křivce kompromis mezi množstvím falešně pozitivních a falešně negativních pozorování, nebo jinými slovy rozdíl mezi senzitivitou a specificitou.

Použijeme k tomu *diagnostický test*, který bude měřit obsah plochy pod křivkou (Area Under Curve – AUC). Je-li plocha rovna 1, je test ideální a má 100% senzitivitu i specificitu.

Na obrázku (3.5) můžeme vidět ROC křivky pro jednotlivé pozorované kvartály. Bude nás tedy zajímat, kolik procent plochy pod křivkou nám model popisuje.



Obrázek 3.5: ROC křivky modelu

- červená – I. kvartál – pro zjištění plochy pod křivkou opět využijeme prostředí Rstudia a pomocí funkce `roc` plochu jednoduše zobrazíme:

```
Call:
roc.default(response = data1$Koupim, predictor =
 = data1_P.E.ratio$pred, plot = T)

Data: data1_P.E.ratio$pred in 8 controls (data1$Koupim 0) < 17
cases (data1$Koupim 1).
Area under the curve: 0.9779
```

  jak můžeme vidět AUC je 97,79 %, což nám ukazuje, kolik procent dat nám model popsal. Tato křivka by se dala považovat skoro za ideální.

- zelená – II. kvartál - opět si zobrazíme výsledek testu:

```
Call:
roc.default(response = data2$Koupim, predictor =
= data2_P.E.ratio$pred, plot = T)

Data: data2_P.E.ratio$pred in 13 controls (data2$Koupim 0) < 12
cases (data2$Koupim 1).
Area under the curve: 0.9103
```

  plocha pod křivkou je 91,03 %. Tato křivka nám popisuje model velmi dobře.

- modrá – III. kvartál

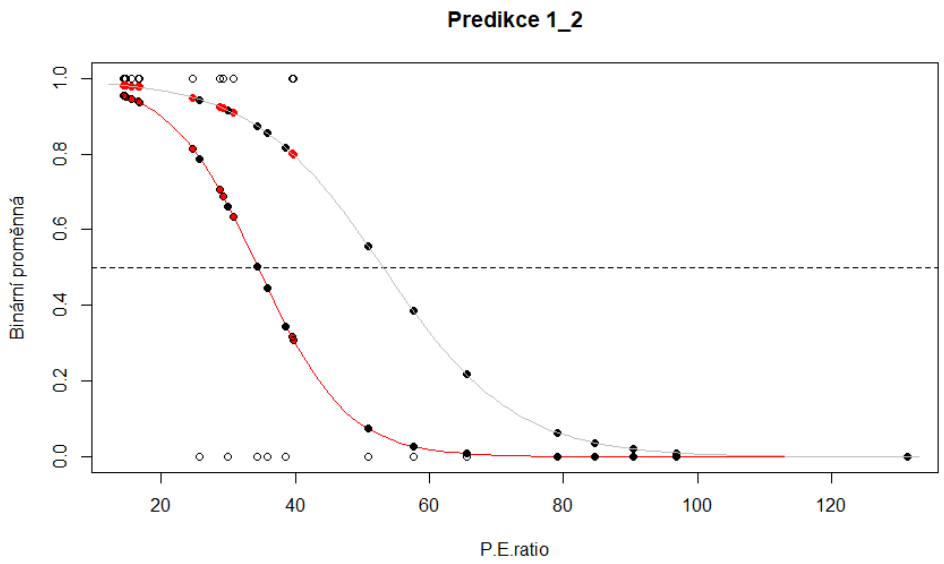
```
Call:
roc.default(response = data3$Koupim, predictor =
= data3_P.E.ratio$pred, plot = T)

Data: data3_P.E.ratio$pred in 7 controls (data3$Koupim 0) < 18
cases (data3$Koupim 1).
Area under the curve: 0.8889
```

plocha pod křivkou totoho kvartálu je 88,89 %, což je stále velice dobrý výsledek.

Čím více se blíží plocha pod křivkou 1, resp. 100 % tím je model kvalitnější. Na základě našich hodnot můžeme říci, že je model velmi dobrý.

3.3 Predikce do budoucích obchobí

V této části se pokusíme vytvořit predikci pro další období na základě pozorování těch minulých.Využijeme k tomu prostředí Rstudia, kam si nahrajeme data z prvního a druhého kvartálu. Použijeme k tomu již nadefinované logistické modely včetně predikcí a vzájemně je zkombinujeme. Vezmeme tedy model pro druhý kvartál a do jeho grafu si vykreslíme regresní křivku s predikcí z kvartálu prvního. Poté na ní překleslíme data druhého kvartálu a obarvíme je podle proměnné *Koupim*. Celý tento proces můžeme vidět na obrázku (3.6).



Obrázek 3.6: Predikce do 2. kvartálu

Děláme to proto, abychom se mohli podívat jestli akcie, které jsme se rozhodli „nakoupit“ v prvním kvartálu, bychom „nakoupili“ i v kvartálu druhém. Rozhoduje nám o tom *dělící křivka*. Akcie, které se nachází nad dělící křivkou nakupujeme, a ty které se nacházejí pod ní ne. Snažíme se tedy mezi křivkami porovnat, zda bychom se rozhodli pro dané akcie stejně, bez ohledu na barevné označení.

Pokud bychom posuzovali koupi podle dat z 2. kvartálu, přičemž bychom použili predikci z kvartálu prvního, tak bychom nakoupili 18 z 25 akcií, což můžeme vidět na obrázku (3.6) na šedé křivce. Pokud bychom nakupovali striktně podle modelu z 2. kvartálu, nakoupili bychom jich pouze 13 z 25. Podle tabulky (3.9) v přílohách, provedeme porovnání výsledků.

Pokud bychom se rozhodli podle predikce, tak bychom nakoupili všechny akcie, které oproti minulému období vzrostly, ale s nimi i 6 akcií, které klesly. Predikce se oproti našemu rozhodnutí v druhém kvartálu liší pro 8 z 25 akcií, které jsou zaznamenány v následujcí tabulce (3.3).

Akcie	**Koupim**	**Pred2**	**Pred1**	**Porovnání**
Hilton Worldwide Holdings Inc.	0	0	1	Koupili bychom ji, kdybychom se rozhodovali podle predikce, ale její hodnota klesla
Intuit Inc.	0	0	1	Koupili bychom ji, kdybychom se rozhodovali podle predikce, ale její hodnota klesla
Johson & Johson	0	1	1	Podle modelu i predikce bychom ji koupili, ikdyž klesla
Las Vegas Sands	0	1	1	Podle modelu i predikce bychom ji koupili, ikdyž klesla
Monster Beverage Corporation	0	1	1	Podle modelu i predikce bychom ji koupili, ikdyž klesla
ResMed Inc.	1	0	1	Podle modelu bychom ji nekoupili ikdyž vzrostly, ale podle predikce ano
Visa Inc.	0	0	1	Koupili bychom ji, kdybychom se rozhodovali podle predikce, ale její hodnota klesla
Zoetis	1	0	1	Podle modelu bychom ji nekoupili ikdyž vzrostla, ale podle predikce ano

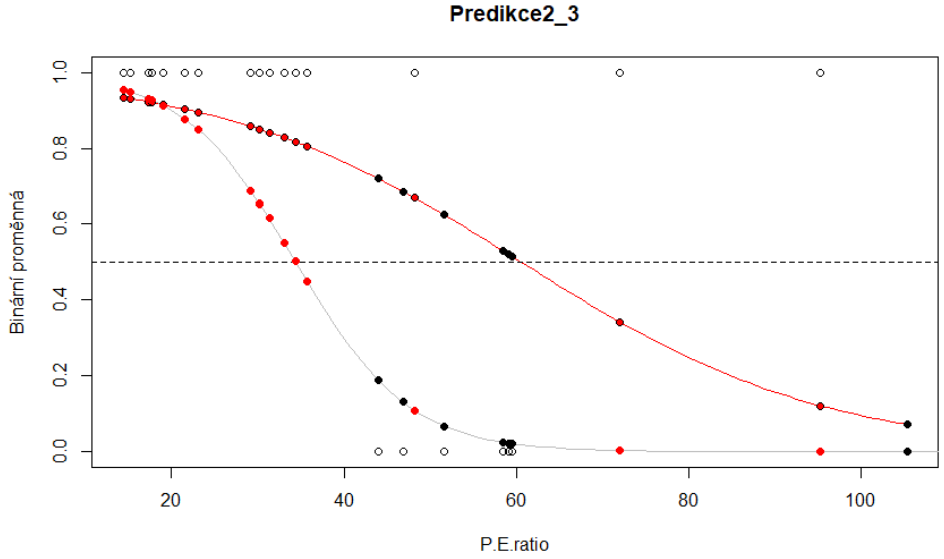
Tabulka 3.3: Změny v predikci oproti modelu

V 17 případech se predikce s naším rozhodnutím shodovala:

- 10 akcií oproti minulému období vrostlo, tedy jsme jim přiřadili hodnotu 1, a proto bychom se přiklonili k jejich koupi.
- 7 akcií oproti minulému období kleslo, proto jsme jim přiřadili hodnotu 0.

Zbývá otázka, jestli je výhodné se řídit predikcí z prvního kvartálu – pro náš případ určitě ano, protože nakoupíme všechny akcie, které vzrostly, což značí zisk. Sice s nimi nakoupíme i 6 akcií, které klesly, ale jsme schopni nést určité riziko a lehké ztráty.

Obdobným způsobem provedeme tento proces mezi druhým a třetím kvartálem. Na obrázku (3.7) můžeme vidět obě regresní křivky pro daná období. Stejně tak můžeme vidět, že pokud bychom posuzovali koupi podle predikce z 2. kvartálu, nakoupili bychom pouze 13 z 25 akcií, ale pokud bychom se řídili podle modelu ze 3. kvartálu, tak bychom vyloučili pouze 3 akcie. Podle tabulky (3.10) provedeme opět porovnání výsledků toho, jak



Obrázek 3.7: Predikce do 3. kvartálu

se predikce liší od našeho rozhodnutí, což nám zobrazuje tabulka (3.4). Oproti predikci se naše rozhodnutí liší pro 11 z 25 případů.

Akcie	**Koupim**	**Pred3**	**Pred2**	**Porovnání**
Abbot Laboratiories	0	1	0	Podle modelu bychom ji koupili, ale její hodnota klesla
Amazon Inc.	1	0	0	Nekoupili bychom ji ani podle modelu, ani podle predikce, ikdyž vzrotla
Equinix	1	0	0	Nekoupili bychom ji ani podle modelu, ani podle predikce, ikdyž vzrotla
Garther Inc.	0	1	0	Podle modelu bychom je koupili, ale její hodnota klesla
Hilton Worldwide Holdings Inc.	1	1	0	Akcie vzrostla, do predikce se nevejde, ale model ji zahrnuje
Intuit Inc.	0	1	0	Podle modelu bychom j koupili, ale její hodnota klesla
ResMed	1	1	0	Akcie vzrostla, do predikce se nevejde, ale model ji zahrnuje
Teleflex	0	1	0	Podle modelu bychom ji koupili, ale její hodnota klesla
Visa	1	1	0	Akcie vzrostla, do predikce se nevejde, ale model ji zahrnuje
Walmart	0	1	0	Podle modelu bychom ji koupili, ale její hodnota klesla
Zoetis	0	1	0	Podle modelu bychom ji koupili, ale její hodnota klesla

Tabulka 3.4: Změny v predikci oproti modelu

Pokud se rozhodneme nakoupit podle predikce, tak nakoupíme pouze akcie, které oproti minulému období vzrostly a tak je pro ně větší předpoklad, že budou růst dále. Tím bychom měli dosáhnout zisku. Proto tedy rohodneme podle predikce z 2. kvartálu a na konci 3. kvartálu nakoupíme akcie následujících společností:

- AES Corporation
- Apple Inc.
- eBay
- Facebook
- Jacobs Engineering Group Inc.
- Johson & Johson
- Juniper Networks
- Las Vegas Sands
- Monster Beverage Corporation
- PepsiCo
- Prologis
- The Walt Disney Company
- Tyson Foods

Na konci IV. kvartálu zhodnotíme, zda nakoupené akcie opět vzrostly a zda jsme dosáhli zisku.

3.4 IV. kvartál

Dostáváme se do posledního kvartálu roku 2019, ve kterém budeme sledovat, zda akcie, které jsme „nakoupili“ na základě pozorování během roku vzrostly a zda jsme dosáhli nějakého zisku.

I pro tento kvartál si vytvoříme logistický model a zobrazíme si jeho výsledek:

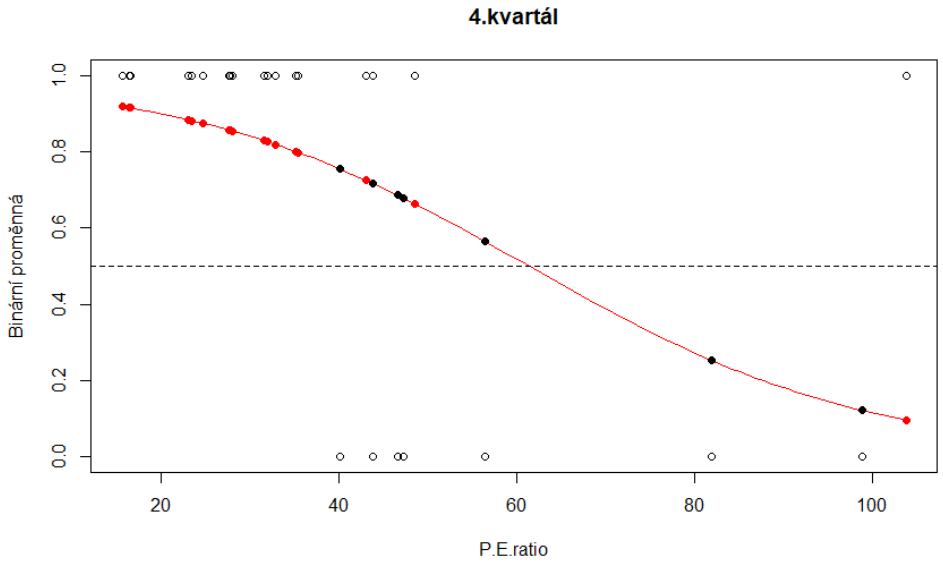
```
Call:
glm(formula = Koupim ~ P.E.ratio, family = "binomial", data = data4)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.6773  -0.5087   0.5188   0.6334   2.1644
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.24396    1.24182   2.612  0.00899 **
P.E.ratio   -0.05289    0.02638  -2.005  0.04494 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 29.648  on 24  degrees of freedom
Residual deviance: 23.536  on 23  degrees of freedom
AIC: 27.536

Number of Fisher Scoring iterations: 4
```

Jasně vidíme, že regresní funkce modelu (1.12) má předpis:

$$\widehat{\pi}(x) = \frac{1}{1 + \exp(-(3.24369 + (-0.05289x))}.$$

Hodnota v posledním sloupci u regresního koeficientu β nám říká, že je koeficient pro náš model velice statisticky významný. Vykreslíme si tedy i graf modelu (3.8).



Obrázek 3.8: Vykreslení modelu pro 4.kvartál

Z obrázku můžeme vidět, že bychom mohli „nakoupit“ až 22 z 25 vybraných akcií, protože v tomto kvartálu bychom se přiklonili k jejich koupi.

My jsme ale podle predikce nakoupili vybraných 13 akcií a podle obrázku (3.8) můžeme vidět, že se všechny nachází nad dělící křivkou. Podle tabulky (3.8) v přílohách vidíme, že u každé z nich je u proměnné *Koupim* hodnota 1, což znamená, že daná akcie oproti minulému období vzrostla a tím jsme dosáhli požadovaného zisku.

Závěr

Tato práce popisuje Zobecněný lineární model, především z matematického hlediska, ale i z hlediska jeho použití. Tento model má velmi široký rozsah, proto jsme se zaměřili pouze na jednu jeho část – logistickou regresi. Termín „zobecněný“ u lineárního modelu značí, že náhodné veličiny se řídí rozdělením z rodiny exponenciálního typu. Logistická regrese je model s binární závisle proměnnou, tedy závislá proměnná, kterou modeluje se řídí binomickým rozdělením.

Teoretická část obsahuje především matematickou teorii, která byla zapotřebí pro definování modelu logistické regrese. Proto jsme postupně nadefinovali jednoduší případy GLM, abychom měli prostor pro vysvětlení jak logistická regrese funguje na poli matematiky. Uvedli jsme i její historii a především její využití v praxi.

Tato práce si tedy kladla za cíl ukázat, jak logistická regrese funguje v obou výše zmíněných směrech. Využívá se v různých oblastech - lékařství, farmacie, průmysl a finance. Proto jsme v praktické části, rozhodli použít model logistické regrese na reálná data z oblasti akciových investic. Tedy poukázat, jak logistická regrese pracuje na poli kapitálového trhu. Na základě několika zvolených aspektů, jsme pomocí modelu logistické regrese rozhodovali o koupi, či nekoupi dané akcie tak, abychom dosáhli zisku.

Každého investora, který obchoduje s cennými papíry ovlivňuje také jeho cit. Chtěli jsme tedy vytvořit model, který by nám pomohl v rozhodnutí, zda dané akcie nakoupit bez této citové složky. Je to model, který mechanicky rozhoduje o koupi či nekoupi na základě pozorování vývoje do minulosti.

Poté jsme pomocí zvolených aspektů pozorovnávali vývoj vybraných akcií v jednotlivých kvartálech zvoleného roku. Pomocí modelu logistické regrese jsme vytvořili predikci do dalších období, a podle ní jsme poté rozhodli o „koupi“ 13 vybraných akcií, jejichž hodnoty v následujícím kvartálu vzrostly. Kdybychom tedy „nakoupené“ akcie na konci dalšího kvartálu prodali, dosáhli bychom zisku. I tento model sebou nese určitá rizika a nezahnuje vývoj celkové ekonomické situace ve světě. K naprogramování takového modelu, jsme využívali prostředí aplikace **Rstudio**.

Přílohy

Akcie	P/E ratio	PEG ratio	Price/Sales	Price/Book	Enterprice value/Revenue	Koupim
Abbot Laboratories	61,02	2,65	4,63	4,62	20,69	0
AES Corporation	12,22	1,83	1,12	3,74	11,24	1
Amazon	88,42	2,2	3,82	20,12	14,65	0
Apple Inc	15,57	1,82	3,56	7,42	15,93	1
eBay	14,56	2,23	3,43	5,18	14,49	1
Equinix	99,38	2,82	7,17	5,28	35,74	0
Facebook	22,02	0,93	8,72	5,66	28,89	1
Gartner Inc	114,05	-	3,51	16,06	16,25	0
Hilton Worldwide Holdings Inc.	33,24	1,11	2,85	44	14,14	1
Intuit Inc.	52,52	2,36	9,35	20,1	36,34	0
Jacobs Engineering Group Inc.	39,65	0,91	0,66	1,78	3,9	0
Johson & Johson	24,92	3,15	4,68	6,21	19,13	1
Juniper Networks	16,54	1,5	2,02	1,93	7,8	1
Las Vegas Sands	19,86	3,3	3,49	8,29	14,93	1
Monster Beverage Corporation	31,01	1,99	8,09	8,22	30,32	1
Netflix	133,04	3,05	10,19	29,76	35,89	0
Pepsi Co	13,96	3,29	2,7	11,85	15,17	1
Prologis	25,07	4,41	15,14	2,04	72,78	1
ResMed	33,32	2,36	6,11	7,62	24,08	1
Teleflex	71,94	2,49	5,78	5,63	25,62	1
The Walt Disney Company	15,19	1,8	2,81	3,93	14,46	1
Tyson Foods	13,12	-	0,64	1,93	3,54	1
Visa	33,59	2,03	17	12,34	66,27	1
Walmart	75,07	3,11	0,55	3,86	2,38	0
Zoetis	34,36	2,89	8,41	22,07	36,42	1

Tabulka 3.5: Data – I. kvartál

Akcie	P/E ratio	PEG ratio	Price/Sales	Price/Book	Enterprice value/Revenue	Koupim
Abbot Laboratories	57,6	2,78	4,85	4,81	20,57	0
AES Corporation	24,73	1,69	1,05	3,44	11,69	1
Amazon	79,03	1,32	3,93	19,32	15,01	0
Apple Inc	16,58	1,45	3,68	8,47	17,53	1
eBay	14,52	2,16	3,51	6,39	14,81	1
Equinix	96,79	3,24	7,81	5,08	38,37	0
Facebook	28,68	1,3	9,5	6,37	30,37	1
Gartner Inc	90,42	-	3,72	16,93	16,27	0
Hilton Worldwide Holdings Inc.	38,63	1,37	3,23	38	14,71	0
Intuit Inc.	50,88	2,52	10,59	23,29	19,48	0
Jacobs Engineering Group Inc.	30,8	1,16	0,73	2,1	4,3	1
Johson & Johson	25,79	3,1	4,64	6,24	18,67	0
Juniper Networks	16,75	1,52	2,06	1,89	7,04	1
Las Vegas Sands	29,99	3,56	3,35	8,24	16,06	0
Monster Beverage Corporation	34,32	2,22	9,12	9,4	30,63	0
Netflix	131,19	3,87	9,98	28,2	34,03	0
Pepsi Co	14,83	3,84	2,87	12,92	12,76	1
Prologis	29,23	4,7	17,1	2,27	77,74	1
ResMed	39,49	4,6	6,97	8,67	26,51	1
Teleflex	84,69	2,74	6,27	5,96	26,47	0
The Walt Disney Company	15,65	3,19	3,52	2,77	14,73	1
Tyson Foods	14,39	3,81	0,73	2,19	3,81	1
Visa	35,86	1,93	18,41	13,51	68,76	0
Walmart	65,5	3,3	0,59	4,06	2,78	0
Zoetis	39,68	3,75	9,31	23,4	38,21	1

Tabulka 3.6: Data – II. kvartál

Akcie	P/E ratio	PEG ratio	Price/Sales	Price/Book	Enterprice value/Revenue	Koupim
Abbot Laboratories	51,65	2,56	4,8	4,67	20,28	0
AES Corporation	29,18	1,42	1,03	3,38	11,06	1
Amazon	72	1,02	3,46	16,19	12,52	1
Apple Inc	19,01	2,04	4,09	10,32	15,76	1
eBay	15,23	1,26	3,33	7,78	14,07	1
Equinix	95,34	2,46	8,86	5,62	42,83	1
Facebook	30,18	0,94	8,22	5,73	26,48	1
Gartner Inc	59,58	-	3,23	13,27	15,52	0
Hilton Worldwide Holdings Inc.	34,36	1,07	2,99	-	14,45	1
Intuit Inc.	46,86	2,53	11,1	17,73	20,16	0
Jacobs Engineering Group Inc.	35,74	1,11	0,77	2,01	3,66	1
Johson & Johson	21,46	2,89	4,32	5,6	17,1	1
Juniper Networks	17,68	1,27	1,95	1,84	6,7	1
Las Vegas Sands	23,1	3,43	3,25	7,98	16,11	1
Monster Beverage Corporation	30,24	1,89	8,05	7,74	26,59	1
Netflix	105,36	1,76	6,86	19,21	23,81	0
Pepsi Co	15,22	3,73	2,97	13,7	12,73	1
Prologis	31,33	4,74	17,89	2,42	68,86	1
ResMed	48,25	4,06	7,49	9,37	30,17	1
Teleflex	59,09	2,39	6,33	5,9	27,22	0
The Walt Disney Company	17,31	3,19	3,19	2,57	14,85	1
Tyson Foods	14,4	2,41	0,76	2,26	4,02	1
Visa	33,08	1,9	17,64	12,97	64,21	1
Walmart	58,46	4,7	0,63	4,61	2,92	0
Zoetis	44,02	2,73	9,97	24,64	40,55	0

Tabulka 3.7: Data – III. kvartál

Akcie	P/E ratio	PEG ratio	Price/Sales	Price/Book	Enterprice value/Revenue	Koupim
Abbot Laboratories	47,21	2,65	4,93	4,81	20,19	0
AES Corporation	27,64	1,55	1,28	4,21	12,96	1
Amazon	81,87	1,32	3,5	16,28	10,71	0
Apple Inc	24,7	2,03	5,25	14,23	14,11	1
eBay	16,34	1,39	2,96	8,79	12,08	1
Equinix	98,76	2,55	8,92	5,68	42,79	0
Facebook	32,89	1,2	8,88	6,23	25,72	1
Gartner Inc	56,45	-	3,4	14,94	13,66	0
Hilton Worldwide Holdings Inc.	35,1	1,31	3,48	-	16,48	1
Intuit Inc.	43,72	2,78	10,02	17,88	28,42	0
Jacobs Engineering Group Inc.	42,98	1,3	0,98	2,09	3,78	1
Johson & Johson	27,78	3,43	4,82	6,6	19,05	1
Juniper Networks	23,46	1,29	1,95	1,78	6,52	1
Las Vegas Sands	27,95	4,31	3,9	9,84	17,35	1
Monster Beverage Corporation	31,93	2,24	8,52	8,3	32,31	1
Netflix	103,71	2,23	7,74	20,69	27,43	1
Pepsi Co	15,62	3,45	2,92	13,46	10,51	1
Prologis	31,5	3,77	17,63	2,51	81,42	1
ResMed	48,44	4,19	8,31	10,58	31,98	1
Teleflex	40,05	-	6,72	6,12	28,3	0
The Walt Disney Company	23,07	3,34	3,46	2,9	14,34	1
Tyson Foods	16,49	2,61	0,79	2,36	4,13	1
Visa	35,32	2,06	18,58	14,26	70,52	1
Walmart	46,53	4,11	0,66	4,73	3,11	0
Zoetis	43,82	2,62	10,39	23,54	40,48	1

Tabulka 3.8: Data – IV. kvartál

Akcie	**Koupim2**	**pred2**	**pred1**
Abbot Laboratories	0	0	0
AES Corporation	1	1	1
Amazon	0	0	0
Apple Inc	1	1	1
eBay	1	1	1
Equinix	0	0	0
Facebook	1	1	1
Gartner Inc	0	0	0
Hilton Worldwide Holdings Inc.	0	0	1
Intuit Inc.	0	0	1
Jacobs Engineering Group Inc.	1	1	1
Johson & Johson	0	1	1
Juniper Networks	1	1	1
Las Vegas Sands	0	1	1
Monster Beverage Corporation	0	1	1
Netflix	0	0	0
Pepsi Co	1	1	1
Prologis	1	1	1
ResMed	1	0	1
Teleflex	0	0	0
The Walt Disney Company	1	1	1
Tyson Foods	1	1	1
Visa	0	0	1
Walmart	0	0	0
Zoetis	1	0	1

Tabulka 3.9: Porovnání dat 1. a 2. kvartálu

Akcie	**Koupim3**	**pred3**	**pred2**
Abbot Laboratories	0	1	0
AES Corporation	1	1	1
Amazon	1	0	0
Apple Inc	1	1	1
eBay	1	1	1
Equinix	1	0	0
Facebook	1	1	1
Gartner Inc	0	1	0
Hilton Worldwide Holdings Inc.	1	1	0
Intuit Inc.	0	1	0
Jacobs Engineering Group Inc.	1	1	1
Johson & Johson	1	1	1
Juniper Networks	1	1	1
Las Vegas Sands	1	1	1
Monster Beverage Corporation	1	1	1
Netflix	0	0	0
Pepsi Co	1	1	1
Prologis	1	1	1
ResMed	1	1	0
Teleflex	0	1	0
The Walt Disney Company	1	1	1
Tyson Foods	1	1	1
Visa	1	1	0
Walmart	0	1	0
Zoetis	0	1	0

Tabulka 3.10: Porovnání dat 2. a 3. kvartálu

Seznam obrázků

Seznam tabulek

Literatura

[1] Karel ZVÁRA, *Regrese*, ISBN 978-80-7378-041-8, MATFYZPRESS, Praha, 2008

[2] Jiří ANDĚL, *Základy matematické statistiky*, ISBN 80-7378-001-1, MATFYZPRESS, vydavatelství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze, 2007

[3] Jitka KÜHNOVÁ, *Matematická statistika 2*, skriptum, přírodovědecká fakulta Univerzity Hradec Králové

[4] Jakub PETRÁSEK, *Aplikace zobecněného modelu lineární regrese v bankovnictví*, bakalářská práce, Univerzita Karlova v Praze, 2006

[5] Filip ZLÁMAL, *Logistická regrese v R*, bakalářská práce, Masarykova univerzita, Brno, 2013

[6] *Logistická regrese*, [online], dostupé z: `https://portal.matematickabiologie.cz/`, citováno [17.4.2020]

[7] *Logistický model pro binární data*, obrázek, [online], dostupné z: `https://www.trilobyte.cz/downloadfree/qcemanual/logreg.pdf`, citováno [30.3.2020]

[8] Jeremy SIEGEL, *Investice do akcií: běh na dlouhou trať*, ISBN 978-80-247-3860-4, GRANADA Publishing a.s., 2011

[9] *Ministerstvo financí ČR*, [online], dostupné z:`https://www.mfcr.cz/cs/soukromy-sektor/kapitalovy-trh/zakladni-informace`, citováno [27.11.2019]

[10] *Akcie*, [online], dostupné z: `https://managementmania.com/cs/akcie`, citováno [27.11.2019]

[11] *EPS*, [online], dostupné z: `https://managementmania.com/cs/cisty-zisk-na-akcii`, citováno [30.1.2020]

[12] Daniel GLADIŠ, *Akciové investice*, ISBN 978-80-247-5375-1, GRANADA Publishing a.s., 2015

[13] *Seznam akcií*, [online], dostupné z: `https://markets.businessinsider.com/index/components/s&p_500/a`, citováno [5.12.2019]

[14] *Informace o vývoji akcií*, [online], dostupné z: `finance.yahoo.com`