TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Regresní analýza v data miningových úlohách

Diplomová práce

Studijní program: N2612 – Elektrotechnika a informatika
Studijní obor: 1802T007 – Informační technologie

Autor práce: **Bc. Tomáš Kadleček**
Vedoucí práce: RNDr. Klára Císařová, Ph.D.

Liberec 2017

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

Regression analysis in data mining tasks

Master thesis

Study programme: N2612 – Electrical Engineering and Informatics
Study branch: 1802T007 – Information Technology

Author: **Bc. Tomáš Kadleček**
Supervisor: RNDr. Klára Císařová, Ph.D.

Liberec 2017

Technická univerzita v Liberci
Fakulta mechatroniky, informatiky a mezioborových studií
Akademický rok: **2016/2017**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Tomáš Kadleček**
Osobní číslo:	**M15000170**
Studijní program:	**N2612 Elektrotechnika a informatika**
Studijní obor:	**Informační technologie**
Název tématu:	**Regresní analýza v data miningových úlohách**
Zadávající katedra:	**Ústav mechatroniky a technické informatiky**

Zásady pro vypracování:

1. Zabývejte se regresní analýzou, lineární i logistickou a jejími uplatněními v data miningových modelech. Jako jeden ze studijních zdrojů použijte MOOC kurz Machine Learning.
2. Prostudujte principy ekonometrického modelování a uveďte příklady nasazení lineárních regresních modelů jako výkladovou studii do kurzu Data mining na ALS portále.
3. Vytipujte případy využití regrese v konkrétních data miningových úlohách a problém řešte pomoci Modeleru případně KNIMU až do postavení případové studie.
4. Regresní model naprogramujte tak, aby bylo možné problém řešit bez podpůrných dataminigových nástrojů a porovnejte obě řešení.

Rozsah grafických prací: **dle potřeby dokumentace**

Rozsah pracovní zprávy: **40–50 stran**

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

[1] **BERKA, Petr. Dobývání znalostí z databází. Praha: Academia, 2003, s.18. ISBN 80-200-1062-9.**

[2] **RUD, Olivia Parr. Data mining. Vyd. 1, Praha: Computer Press, 2006, XVII, 329 s. ISBN 80-722-6577-6.**

[3] **Yong Yin, Ikou Kaku, Jiafu Tang: Data Mining, Springer London Ltd , 2011.**

[4] **HANČLOVÁ, Jana. Ekonomerické modelování,. Vyd. 1, Praha: Professional Publishing, 2012, 2012s. ISBN 978-80-7431-088-1.**

Vedoucí diplomové práce: **RNDr. Klára Císařová, Ph.D.**
Ústav mechatroniky a technické informatiky

Datum zadání diplomové práce: **10. října 2016**

Termín odevzdání diplomové práce: **15. května 2017**

L.S.

prof. Ing. Zdeněk Plíva, Ph.D.
děkan

doc. Ing. Milan Kolář, CSc.
vedoucí ústavu

V Liberci dne 10. října 2016

Prohlášení

Byl jsem seznámen s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu TUL.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Diplomovou práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé diplomové práce a konzultantem.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum: 15. 5. 2017

Podpis: Kudlička

Abstrakt

Tato diplomová práce se zabývá problematikou regresní analýzy v data miningových úlohách a ekonometrickým modelováním. Jejím cílem je seznámit se z problematikou regresní analýzy a její aplikací v případové studii. Následně vysvětlit pojem ekonometrické modelování a uvést příklady její aplikace. V diplomové práci byla vypracována případová studie v programu IBM SPSS Modeler. Zabývá se odhadem ceny nemovitosti. Výsledkem mé práce je vypracovaná případová studie s nasazením lineárního regresního modelu. Dále byla naprogramovaná aplikace v prostředí Octave, která umožňuje vytvořit vlastní lineární regresní model. V závěru práce porovnávám výsledky regresní statistiky provedené pomocí programu Modeler a mé aplikace v prostředí Octave. Porovnání je provedeno na několika různých výběrových souborech.

Klíčová slova: regresní analýza, lineární regresní model, ekonometrie, ekonometrické modelování, případová studie, Octave

Abstract

This diploma thesis deals with problems of regression analysis in data mining tasks and econometric modeling. Its aim is to get acquainted with problems of regression analysis and its application in the case study. Further explain the concept of econometric modeling and give examples of its application. A case study in IBM SPSS Modeler was developed in the thesis. It deals with an estimate of the property price. The result of my work is a case study with the implementation of a linear regression model. In addition, an Octave application was programmed to create a custom linear regression model. At the end I compare the results of regression statistics made using Modeler and my application in Octave. The comparison is made on several different sample files.

Keywords: regression analysis, linear regression model, econometrics, econometric modeling, data analysis, case study, Octave

Poděkování

Chtěl bych poděkovat vedoucí mé práce RNDr. Kláře Císařové, Ph.D. za její trpělivost a rady, které mi poskytla při zpracovávání diplomové práce. Dále pak mé přítelkyni, která mi byla velkou oporou a mé rodině za jejich podporu při studiu a při zpracovávání diplomové práce.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zdrojových kódů

Seznam zkratek

MOOC	Hromadný otevřený online kurz
TUL	Technická univerzita v Liberci
MNČ	Metoda nejmenších čtverců
PRF	Populační regresní funkce
VRF	Výběrová regresní funkce
ML	Metoda maximální věrohodnosti
TSS	Celkový součet čtverců
RSS	Reziduální součet čtverců
ESS	Regresní (vysvětlený) součet čtverců
GPL	General Public License
ANOVA	Anylysis of variance
CPU	Centrální procesorová jednotka

1 Úvod

S nástupem internetu a sběru enormního množství elektronických dat se velice rozvíjejí technologie, pomocí kterých je máme možnost studovat a hledat v nich závislosti, vztahy a další informace. A právě Data mining je obor, který se touto problematikou zabývá.

Ve své diplomové práci se budu zabývat regresní analýzou a jejím uplatněním v data miningových modelech. V teoretické části budu rozebírat metody lineární a logistické regrese. Jako jeden z materiálů, ze kterých čerpám, jsem použil data minigový online kurz, konkrétně MOOC kurz Machine Learning pořádaný Univerzitou ve Stanfordu. Tímto kurzem jsem prošel v letním semestru roku 2015 v rámci přípravy na diplomovou práci a úspěšně jsem ho dokončil. V příloze B přikládám diplom udělený za absolvování kurzu. V následující části mé diplomové práce se budu zabývat ekonometrickým modelováním, které s regresní analýzou úzce souvisí. Ekonometrické modelování spočívá v analýze určitého ekonomického problému a snaží se nalézt souvislosti mezi ekonomickou teorií a reálnými daty. Tyto závislosti mohou být poté použity k predikci ekonomických proměnných. Téma ekonometrické modelování zpracuji jako výkladovou studii na ALS portále.

V praktické části budu uplatňovat teoretické poznatky z prvních dvou kapitol vytvořením případové studie. Studii budu vytvářet ve statistickém programu IBM SPSS Modeler. Dále naprogramuji lineární regresní model tak, aby nebylo třeba používat podpůrné nástroje, jako Modeler, Knime či jiné. K tomuto účelu použiji prostředí Octave. Tento program jsem zvolil z důvodů jeho nezávislosti a svobodné licence. Bude schopen načíst libovolná data v textovém formátu odpovídající struktury a na nich vytvořit vlastní lineární regresní model včetně jeho celkového otestování a testu jednotlivých atributů. V závěru práce porovnám oba uplatněné přístupy z hlediska jejich přesnosti a času výpočtu.

2 Regresní modely

Tato kapitola se zabývá regresními modely a jejich modifikacemi. Tato analýza patří k jednomu z nástrojů ekonometrického modelování.

2.1 Základní pojmy

Před použitím samotných regresních modelů, je třeba vysvětlit, některé ze základních pojmů v oboru statistky a představit základní nástroje pro analýzu dat.

2.1.1 Data

Data se rozdělují do několika skupin podle rozsahu, obsahu a typu. Ve statistice rozlišujeme statistický soubor, jednotku a znak.

- Statistický soubor je konečná neprázdná množina prvků M, které mají společné vlastnosti.
- Statistická jednotka je jeden prvek ze statistického souboru (jeden prvek množiny).
- Statistické znaky jsou vlastnosti statistických jednotek.

Pojmem rozsah souboru n představuje mohutnost množiny M (viz vztah 2.1).

$$n = |M| \tag{2.1}$$

Rozlišují se dva přístupy ke statistickému souboru:

1. Základní soubor
   Množina všech teoreticky možných prvků zkoumaného problému. Problémem u takové množiny dat je její přílišná obsáhlost, a proto ji obvykle v praxi není možné použít. Z tohoto důvodu se používá výběrový soubor.
2. Výběrový soubor
   Vzorek dat ze základního souboru. Následně se podle tohoto výběru provádí úsudek o základním souboru. Vlivem provedení výběru ze základního souboru dochází k určité výběrové chybě.

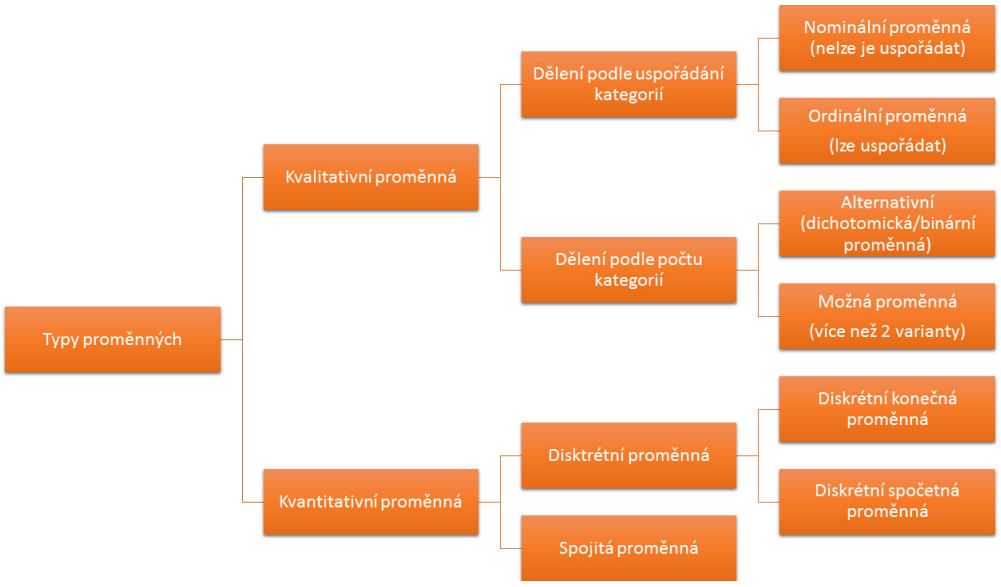
Každá statistická jednotka vykazuje vlastnosti, které se nazývají také atributy, nebo proměnné. Dělí se kvantitativní a kvalitativní. Základní rozdíl mezi těmito typy je, že kvantitativní jsou číselné charakteristiky, pomocí kterých definujeme nebo měříme různé jevy. Běžně to jsou číselné proměnné, které charakterizují určitou vlastnost objektu například výšku osob, naměřená data z přístroje atd. Obyčejně dává smysl provádět nad těmito hodnotami aritmetické operace.
Kvantitativní typy se dále dělí na:

1. Nespojité (Diskrétní) proměnné nabývají vždy určitých hodnot (např četnosti).
2. Spojité proměnné mohou nabývat teoreticky nekonečného počtu hodnot mezi určitým intervalem (např. výška osob).

Naproti tomu kvalitativní jsou jakákoli data výčtového typu, pomocí kterého popisujeme určité vlastnosti objektu. Ne vždy můžeme tyto hodnoty porovnávat (nominální). Podle uspořádání kategorií proměnných můžeme rozdělit kvalitativní data na:

- Nominální – lze říci, že se dvě hodnoty liší, ale nelze je porovnávat (např. výrobce, typ, národnost),
- Ordinální – stejné jako nominální, ale navíc lze hodnoty mezi sebou porovnávat (např. hodnocení, typ vzdělání).

Dále se dělí podle počtu kategorií na alternativní (dichotomické) a více-kategoriální proměnné. Shrnutí typů proměnných zobrazuje obrázek 2.1.



Obrázek 2.1: Typy proměnných

2.1.2 Základní nástroje pro analýzu dat

Tato kapitola se věnuje základním nástrojům pro analýzu dat, pomocí kterých můžeme reprezentovat daný výběrový soubor.

Rozdělení četností

Četnost je veličina, která udává kolikrát se daná hodnota statistického znaku vyskytuje ve statistické souboru. Uvažujeme-li statistický znak ve tvaru $x_1, x_2, ..., x_n$, kde n je rozsah statistického souboru a celkový počet různých hodnot znaku x je $k \leq n$[10].

Absolutní četnost hodnoty znaku x_j je počet statistických jednotek, které mají stejnou hodnotu znaku x_j pro $j = 1, 2..., k$.

$$\sum_{i=1}^{k} n_j = n \tag{2.2}$$

Relativní četnost hodnoty znaku x_j je podíl absolutní četnosti a rozsahu souboru, nejčastěji se vyjadřuje v procentech, označuje se jako v_j a jejich součet je jedna (v případě procent 100)(viz rovnice 2.3). Výhodu relativní četnosti je, že pomocí ní můžeme porovnávat dva výběrového soubory s rozdílnými rozsahy[10].

$$\sum_{i=1}^{k} v_j = 1. \tag{2.3}$$

Kumulativní absolutní četnost vyjadřuje součet všech předcházejících absolutních četností. Umožňuje zjistit kolik hodnot je menších než zadané číslo.

Kumulativní relativní četnost umožňuje zjistit procento hodnot menších než zadané číslo. Vypočteme jej vydělením příslušné absolutní kumulativní četnosti s rozsahem souboru, nebo sečtením relativních četností v intervalech, jejichž horní hranice je menší než zadané číslo.

Intervalové rozdělení četností - kategorizace

Toto rozdělení rozděluje statistický soubor na intervaly, kterým říkáme třídy. Používá se zejména v případech, kdy máme příliš mnoho variant znaků, například u spojitých hodnot, jako je výška osob nebo příjem. Použitím tohoto rozdělení zvýšíme přehlednost statistického souboru. Při vytváření intervalů je třeba dodržovat určité pravidla[10]. Počet tříd rozdělení – k odpovídá:

- Odmocninovému pravidlu $k = \sqrt{n}$,
- Sturgesovu pravidlu $k = 1 + 3,3 \log n$.

Pro určení šířky (počtu prvků) intervalu existuje několik metod. Jednou z nich je podíl rozdílu maximální a minimální hodnoty výběrové souboru a počtu tříd.

$$i = \frac{MAX - MIN}{k} \tag{2.4}$$

Mezi další metody řadíme určení šířky intervalu pomocí Kvantilů. Rozdělují statistický soubor na části, v závislosti na tom, kolika procentní kvantil je použit. Značí se x_p, kde p jsou procenta v intervalu $< 0, 100 >$.
Nejpoužívanější kvantily jsou:

- Medián - x_{50},
- Kvartily - x_{25}, x_{50}, x_{75} ,
- Decily - $x_{10}, x_{20} \ldots, x_{90}$,
- Percentily - $x_1, x_2, \ldots, x_{99}$.

Kategorizaci číselné proměnné zejména v data miningových řešeních lze provést mnoha dalšími postupy. Například algoritmy pro kategorizaci s respektem k cílové predikované hodnotě.

Charakteristiky statistického souboru

Při statistické analýze je často třeba porovnávat několik statistických souborů. Z tohoto důvodu se používají charakteristiky. Charakterizují základní rysy zkoumaného statistického souboru[10].
Existuje několik základních charakteristik:

- Polohy
- Variability
- Tvaru
- Kovariance

Charakteristika polohy

Představuje různé druhy středních hodnot výběrového souboru. Obecně označujeme střední hodnotu jako $E[X] = \overline{x} = \mu$.

Základní mírou polohy je Aritmetický průměr. Rozlišujeme průměr pro základní a výběrový soubor[10].

- Aritmetický průměr pro základní soubor:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}. \tag{2.5}$$

- Aritmetický průměr pro výběrový soubor:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{2.6}$$

- Vážený aritmetický průměr - je zvláštním případem výše zmíněného, kde n_i jsou váhy (četnosti) jednotlivých hodnot x_i. Nejčastěji to jsou počty výskytů hodnoty x_i ve výběrovém souboru.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i n_i}{\sum_{i=1}^{n} n_i} \tag{2.7}$$

K aritmetickému průměru se váže několik vlastností:

- Aritmetický průměr konstanty je konstanta.
- Přičtením, odečtením, vynásobením nebo vydělením všech hodnot znaku nenulovou konstantou se odpovídajícím způsobem změní také aritmetický průměr.
- Vynásobím-li všechny váhy nenulovou konstantou, tak se průměr nezmění.

Kromě aritmetického průměru existují další, které se používají ve speciálních případech (viz tabulka 2.1).

Název	**Vzorec**	**Použití**
Geometrický průměr	$\overline{x_G} = \sqrt[n]{\prod_{i=1}^{n} x_i}$	K výpočtu koeficientů růstu nebo řetězových indexů.
Vážený geometrický průměr	$\overline{x_G} = \sqrt[n]{\prod_{i=1}^{n} x_i^{n_i}}$	Je použit v případě, že jsou udaje zatříděny dle četností, nebo mají různé hodnoty různou váhu.
Harmonický průměr	$\overline{x_H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	pro měření úrovně poměrných čísel (rychlost, výkon, produktivita práce).
Vážený harmonický průměr	$\overline{x_H} = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} \frac{n_i}{x_i}}$	Je použit v případě, že jsou udaje zatříděny dle četností, nebo mají různé hodnoty různou váhu.
Kvadratický průměr	$\overline{x_K} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$	Při výpočtu střední kvadratické odchylky.
Vážený kvadratický průměr	$\overline{x_G K} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 n_i}{\sum_{i=1}^{n} n_i}}$	Je použit v případě, že jsou udaje zatříděny dle četností, nebo mají různé hodnoty různou váhu.

Tabulka 2.1: Seznam typů průměrů

Platí, že $\overline{x} \leq \overline{x_G} \leq \overline{x_H} \leq \overline{x_K}$
. Mezi další střední hodnoty se řadí také:

- Medián – hodnota, která je ve středu statistického souboru za předpokladu, že je seřazený.
- Modus – hodnota z nejvyšší četností znaku.

Charakteristika variability

Charakteristika variability udává, jak se liší hodnoty znaků prvků od zvolené charakteristiky polohy (průměru). Řadíme mezi ně variační rozpětí, průměrnou absolutní odchylky, rozptyl, směrodatnou odchylku a variační koeficient. Platí, že čím vyšší je variabilita hodnot znaku, tím nižší je vypovídací schopnost charakteristiky polohy (průměru atd.)[10].

Název	**Vzorec**
Variační rozpětí	$R = x_{max} - x_{min}$
Kvartilové rozpětí	$KR = \widetilde{x_{75}} - \widetilde{x_{25}}$
Kvartilová odchylka	$Q = \frac{(\widetilde{x_{75}} - \overline{x}) + (\overline{x} - \widetilde{x_{25}})}{2}$
Průměrná odchylka	$\overline{d_x} = \frac{\sum_{i=1}^{n} \lvert x_i - \overline{x} \rvert}{n}$
Relativní průměrná odchylka	$\overline{D} = \frac{\overline{d_x}}{\overline{x}}$
Střední diference	$\Delta = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \lvert x_i - x_j \rvert}{n(n-1)}$

Tabulka 2.2: Charakteristiky variability

Nejpoužívanější charakteristikou variability je rozptyl. Značí se také jako $var(X) = D(X) = E(X - E(X))^2 = \sigma^2$. Je definován jako průměr kvadrátů odchylek jednotlivých znaků x_i od jejich aritmetického průměru $\overline{x}$ [10].
Stejně jako u průměru rozlišujeme rozptyl základního souboru a výběrového souboru. Rozptyl základní souboru:

$$\sigma^2 = var(X) = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}. \tag{2.8}$$

Rozptyl výběrového souboru:

$$S_x^2 = var(X) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}, \tag{2.9}$$

kde ve jmenovateli výraz $n-1$, označuje počet stupňů volnosti výběrového souboru. Použitím výrazu $n-1$ místo velikosti souboru n docílíme přesnějšího odhadu skutečné hodnoty populačního rozptylu, zejména při výpočtu na základě malých výběrových souborů[10].
K rozptylu se váže několik vlastností:

- Rozptyl konstanty je nula.
- Přičteme-li ke všem hodnotám znaku stejnou konstantu $\Longrightarrow$ rozptyl se nezmění.
- Vynásobíme-li každou hodnotu znaku stejnou konstantou $\Longrightarrow$ rozptyl bude její násobek .
- Rozptyl součtu nebo rozdílu dvou znaků je roven součtu rozptylů obou znaků zvětšeném/zmenšeném o dvojnásobek kovariance: $S_z^2 = S_x^2 + S_y^2 \pm S_{xy}$.

Vzhledem k užití kvadrát zavádíme směrodatnou odchylku, která je definována jako:

$$S_x = \sqrt{S_x^2}. \tag{2.10}$$

Často je třeba porovnávat statistické soubory a může se stát, že znaky nejsou ve stejných jednotkách nebo mají nestejnou velikost. V takových případech využíváme charakteristiku relativní variability. Mezi ni řadíme variační koeficient. Označujeme ho V_x. A vypočítá se jako podíl směrodatné odchylky a průměru výběrového souboru[10]:

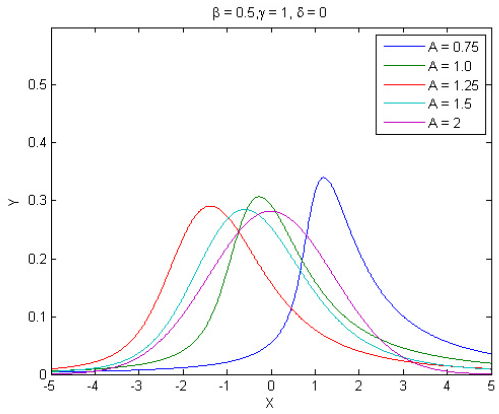
$$V_x = \frac{S_x}{\overline{x}}. \tag{2.11}$$

Charakteristika tvaru

Měří odchylku v rozložení četností hodnot znaků oproti danému referenčnímu rozdělení četností (obvykle normálnímu)[10]. Skládá se ze dvou složek:

- Asymetrie (Šikmosti) – udává symetrické/asymetrické rozložení hodnot kolem průměr.

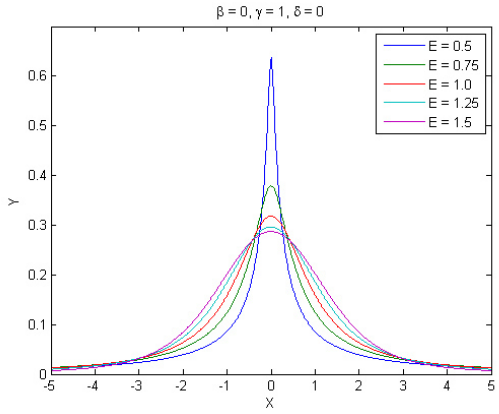
$$A = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{nS^3} \tag{2.12}$$



Obrázek 2.2: Charakteristiky asymetrie

- Špičatosti – porovnává četnost hodnoty znaků kolem průměru.

$$E = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{nS^4} \tag{2.13}$$



Obrázek 2.3: Charakteristiky špičatosti

Kovariance

Charakterizuje, jak se dva znaky x a y statistického souboru vzájemně ovlivňují. Značí se jako $cov(X;Y)$, nebo Sxy[10] a vypočítáme ji jako:.

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \overline{xy} - \overline{x}\,\overline{y}.$$

Pokud $S_{xy} > 0$ znak x roste (klesá), tak roste (klesá) y, např. vztah mezi výškou a váhou člověka.
Pokud $S_{xy} < 0$ znak x roste (klesá), tak y klesá (roste), např. vztah mezi hloubkou dezénu pneumatiky a brzdnou dráhou automobilu.

Platí, že čím vyšší je kovariance, tím více se znaky navzájem mění. Naopak ale nulová kovariance $S_{xy} = 0$ nemusí nutně znamenat, že mezi znaky neexistuje závislost. Jen se nemusí jednat o lineární závislost, ale například o kvadratickou.

Korelace

Korelace označuje míru závislosti dvou znaků x a y. Řekneme, že dvě proměnné jsou korelované jestliže hodnoty jedné proměnné mají tendenci vyskytovat se společně s hodnotami druhé proměnné[10]. Pro změření míry korelace je navržena řada koeficientů, které se liší podle typů proměnných a vlastnostmi. Při zkoumání vztahů korelace je důležitý kvalitativní rozbor dat. Jinak řečeno, nemá smysl hledat závislost tam, kde na základě logické úvahy nemůže existovat.
Jedním z nejpoužívanějších koeficientů je Pearsonův korelační koeficient. Označuje se jako r_{xy} a spočítá se jako podíl kovariance S_{xy} a násobku směrodatných odchylek S_x a S_y.

$$r_{xy} = \frac{S_{xy}}{S_x S_y}. \tag{2.14}$$

2.2 Jednoduchý lineární regresní model

Tato kapitola se bude zabývat jednoduchým lineárním regresním modelem. Tedy, kdy závislá (vysvětlovaná) proměnná Y je lineárním vztahem pouze jedné nezávislé (vysvětlující) proměnné X. Pomocí regresního modelu hledáme lineární vztah mezi proměnnou Y a X[9].

První dva pojmy, kterými se budeme zabývat jsou deterministická a stochastická populační regresní funkce, dále jen PRF. Deterministická PRF spojuje očekávané hodnoty vysvětlované proměnné Y_i pro daná X_i a je dána vztahem:

$$E(Y_i|X_i) = \beta_1 + \beta_2 X_i, \quad i = 1, 2, \ldots, n, \tag{2.15}$$

kde parametry β_1 je absolutní člen a β_2 definuje sklon regresní křivky.
Tyto modely nejsou příliš časté, protože existují další vlivy na vysvětlovanou proměnnou Y_i resp. náhodné složky, které do regresního modelu vnáší určitou chybu. Zanesením této chyby do modelu zadefinujeme stochastickou PRF. Je definovaná jako:

$$E(Y_i|X_i) = \beta_1 + \beta_2 X_i + u_i, \quad i = 1, 2, \ldots, n, \tag{2.16}$$

kde u_i je náhodná složka, tj. chyba, zanesená zanedbáním některých vlivů a dalších chyb, například z měření.

Jak již bylo výše zmíněno, obvykle se nestává, že bychom měli k dispozici data za celou populaci (základní soubor), tudíž jej nahrazujeme výběrovými soubory. Nazýváme ji výběrová regresní funkce, dále VRF. Je snaha o to, aby VRF konvergovala k PRF. Následující tabulka 2.3 shrnuje zmiňované funkce.

	Deterministická forma	**Stochastická forma**
Populační regresní funkce	$E(Y_i\|X_i) = Y_i = \beta_1 + \beta_2 X_i$	$E(Y_i\|X_i) = Y_i = \beta_1 + \beta_2 X_i + u_i$
Výběrová regresní funkce	$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} X_i$	$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} X_i + \widehat{u_i}$

Tabulka 2.3: Forma zápisu populační a výběrové regresní funkce

Symboly „$\widehat{\ }$“ nad proměnnými a parametry vyjadřují odhad pro výběrový soubor. To znamená, že $\widehat{Y_i}$ je odhad pro Y_i, $\widehat{\beta_1}$ a $\widehat{\beta_2}$ jsou odhady regresních parametrů, a $\widehat{u_i}$ představuje reziduální složku, což je odhad stochastické náhodné složky u_i[9]. Existuje několik metod pro odhad parametrů regresního modelu:

- metoda nejmenších čtverců (MNČ),
- metodu maximální věrohodnosti (ML),
- metoda momentů,
- zobecněná metoda momentů.

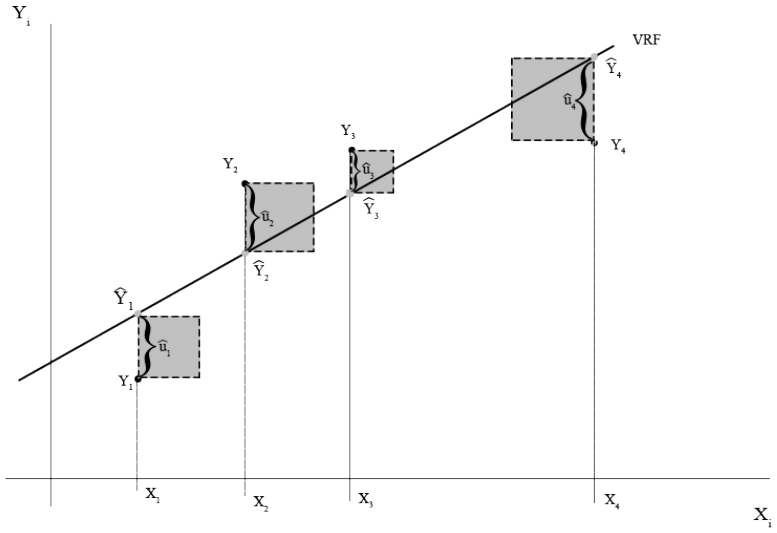
Tato práce se zaměřuje na první z uvedených metod, o druhé metodě se zmiňuje v souvislosti s logistickou regresní analýzou.

2.2.1 Metoda nejmenších čtverců

Tato metoda byla zavedena, německým matematike, Carlem Friedrichem Gaussem. Jedná se metodu zjištění parametrů $\widehat{\beta_1}$ a $\widehat{\beta_2}$ výběrové regresní funkce:

$$Y_i = \widehat{\beta_1} + \widehat{\beta_2} X_i + \widehat{u_i} = \widehat{Y_i} + \widehat{u_i}, \quad i = 1, 2, \ldots, n, \tag{2.17}$$

kde vývoj proměnné Y_i je determinován změnami X_i a tvar křivky je určen regresními parametry β_1 aβ_2[9]. Metoda proloží přímku jednotlivými hodnotami znaků, jak zobrazuje obrázek 2.4.



Obrázek 2.4: Princip metody nejmenších čtverců

Dále vyjádříme z rovnice 2.17 reziduální složku $\widehat{u_i}$:

$$\widehat{u_i} = Y_i - \widehat{Y_i} = Y_i - \widehat{\beta_1} + \widehat{\beta_2} X_i = f(\widehat{\beta_1}, \widehat{\beta_2}). \tag{2.18}$$

Z rovnice 2.18 je zřejmé, že reziduální složka je funkcí regresních parametrů. Obrázek 2.4 ukazuje, že reziduální složka $\widehat{u_i}$ může být kladná i záporná. Z těchto důvodů je třeba použít součet čtverců reziduálních odchylek. A tedy základem metody nejmenších čtverců je minimalizace tohoto součtu[9]:

$$\sum_{i=1}^{n} \widehat{u_i}^2 = f(\widehat{\beta_1}, \widehat{\beta_2}). \tag{2.19}$$

Pro nalezení minima funkce se použije metoda z matematické analýzy – hledání extrému funkce. Funkce 2.19 se parciálně zderivuje podle parametrů β_1 a β_2 a jednotlivé derivace položíme rovny nule:

$$\begin{aligned} \frac{\delta(\sum \widehat{u_i}^2)}{\delta\widehat{\beta_1}} &= 2\sum(-1)(Y_i - \widehat{\beta_1} - \widehat{\beta_2} X_i) = 0, \\ \frac{\delta(\sum \widehat{u_i}^2)}{\delta\widehat{\beta_2}} &= 2\sum(-X_i)(Y_i - \widehat{\beta_1} - \widehat{\beta_2} X_i) = 0. \end{aligned} \tag{2.20}$$

Úpravou obou těchto rovnic získáme 2 rovnice o dvou neznámých parametrech:

$$\begin{aligned} \sum Y_i &= n\widehat{\beta_1} + \widehat{\beta_2} \sum X_i, \\ \sum Y_i X_i &= \widehat{\beta_1} \sum X_i + \widehat{\beta_2} \sum X_i^2. \end{aligned} \tag{2.21}$$

Jejím vyřešením obdržíme odhady obou regresních parametrů:

$$\widehat{\beta_2} = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2},$$
$$\widehat{\beta_1} = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_iY_i}{n\sum X_i^2 - (\sum X_i)^2} = \overline{Y} - \widehat{\beta_2}\overline{X}, \tag{2.22}$$

kde $\overline{X}$ a $\overline{Y}$ jsou výběrové průměry pro X a Y.

2.2.2 Vlastnosti odhadové funkce nejmenších čtverců

Pomocí metody nejmenších čtverců byl proveden bodový odhad [1] parametrů $\widehat{\beta_1}$ a $\widehat{\beta_2}$ daného výběrového souboru. Za předpokladu dalších nezávislých výběrových souborů se získá výběrové rozdělení hodnot odhadů parametrů, a poté na jejím základě dochází k odhadu parametrů β_1 a β_2 základního souboru[9].
Odhadová funkce má tyto vlastnosti:

- nestrannost,
- vydatnost (eficience),
- konzistence.

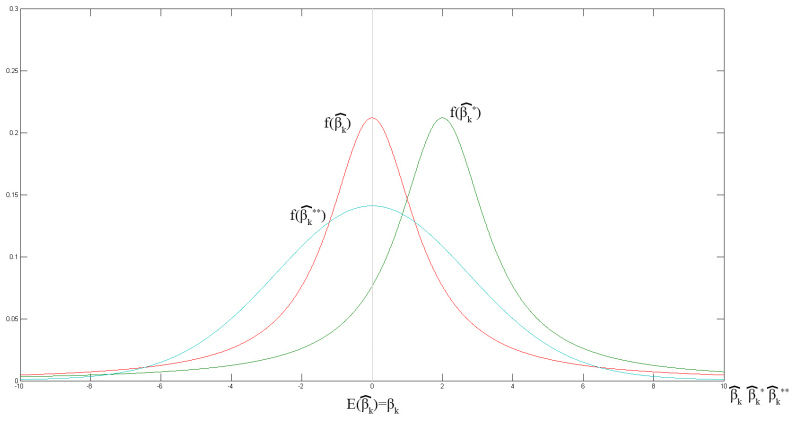
Nestrannost je vlastnost odhadové funkce $\widehat{\beta_k}$, která říká, že střední hodnota bodového regresního parametru je rovna populačnímu regresnímu parametru:

$$E(\widehat{\beta_k}) = \beta_k. \tag{2.23}$$

Tuto vlastnost zobrazuje obrázku 2.5, kde odhadová funkce $\widehat{\beta_k^*}$ (zelená) je vychýlená vůči odhadu $\widehat{\beta_k}$.

Další z vlastností je vydatnost (eficience). Naše odhadová funkce $\widehat{\beta_k}$ je eficientní vůči jiné téhož $\widehat{\beta_k^{**}}$ (modrá), jestliže nemá větší rozptyl. Vlastnost zobrazuje obrázek 2.5. Z něho vyplývá, že odhadová funkce $\widehat{\beta_k}$ je z dané třídy odhadových funkcí s nejmenším rozptylem. Obě tyto vlastnosti zkoumáme zejména na menších výběrových souborech[9].

[1] neznámý parametr základního souboru odhadujeme pomocí jediného čísla

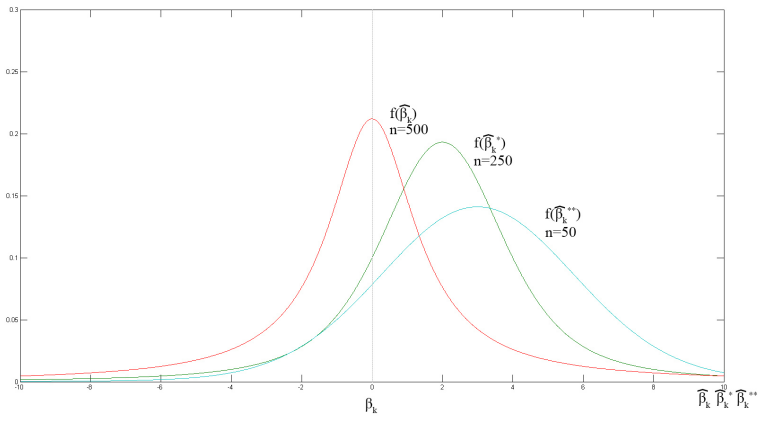


Obrázek 2.5: Nevychýlené a eficientní rozdělení parametrů $\widehat{\beta_k}$

Pro rozsáhlé soubory testujeme vlastnosti konzistence. Odhadová funkce $\widehat{\beta_k}$ je konzistentní s odhadovou funkcí β_k pro n limitně rostoucí do nekonečna, kde se n rovná rozsahu výběrového souboru, jestliže je:

- asymptoticky nestranná:
$$\lim_{n\to\infty} E(\widehat{\beta_k}) = \beta_k, \tag{2.24}$$
- s rostoucí hodnotou n parametr $\widehat{\beta_k}$ konverguje ke skutečné hodnotě odhadnutého parametru β_k:
$$\lim_{n\to\infty} \widehat{\beta_k} = \beta_k. \tag{2.25}$$

Obrázek 2.6 zobrazuje tři odhadové funkce $\beta_k, \beta_k^*, \beta_k^{**}$, kde s rostoucím rozsahem výběrového souboru n roste konzistence.



Obrázek 2.6: Konzistentní rozdělení parametrů $\widehat{\beta_k}$

2.2.3 Předpoklady pro použití metody nejmenších čtverců

Vlastnosti odhadové funkce zmiňované v kapitole 2.2.2 jsou splněny za několika předpokladů. Tato kapitola zkoumá tyto předpoklady a v tomto případě se zaměřuje pouze na jednoduchý lineární regresní model[9]. Tyto předpoklady jsou dále zobecněny pro vícerozměrný lineární regresní model v kapitole 2.4.2.

- **P1**: Lineární regresní model $Y_i = \beta_1 + \beta_2 X_i + u_i$ je lineární v parametrech.
- **P2**: Hodnoty X_i jsou fixní.
- **P3**: Střední hodnota náhodné složky je nulová $E(u_i|X_i) = 0$
- **P4**: Pro každou i-tou skupinu bude platit, že variabilita náhodné složky bude rovna σ^2. Tento předpoklad se také nazývá homoskedasticita $\Longrightarrow$ nemění se rozptyl náhodné složky v jednotlivých skupinách. Opakem je heteroskedasticita $\Longrightarrow$ rozptyl se mění, např. zvyšuje se s rostoucími hodnotami X_i.

$$var(u_i|X_i) = D(u_i|X_i) = E(u_i - E(u_i|X_i))^2 = E(u_i^2|X_i) = \sigma^2 \tag{2.26}$$

- **P5**: Náhodná složka z různých skupin není sériově závislá (korelovaná). V případě opaku mluvíme o sériové korelaci (autokorelaci) náhodné složky, která pak je pozitivní nebo negativní.

$$\begin{aligned} &cov(u_i; u_j|X_i; X_j) = E\{[u_i - E(u_i)]|X_i\}\{[u_j - E(u_j)]|X_j\} = \\ &= E\{u_i|X_i\}\{u_j|X_j\} = 0 \quad pro\ i \neq j \end{aligned} \tag{2.27}$$

- **P6**: Dalším předpokladem je nulová kovariance mezi náhodnou složkou u_i a X_i. Tento předpoklad zároveň vyjadřuje, že PRF můžeme rozdělit na dvě aditivní části tzn. na část deterministické regrese a stochastické regrese s náhodnou složkou.

$$\begin{aligned} &cov(u_i; X_i) = E[u_i - E(u_i)(X_i - E(X_i))] = E[u_i(X_i - E(X_i))] = \\ &= E(u_i X_i) - E(X_i)E(u_i) = E(u_i, X_i) = 0 \end{aligned} \tag{2.28}$$

- **P7**: Počet pozorování $|X| = n$ musí být větší, jak počet parametrů regresního modelu. U jednoduchého regresního modelu platí $n > 2$.
- **P8**: Náhodná složka má normální rozdělení $u_i \sim N(0; \sigma^2)$.

2.2.4 Koeficient determinace

Koeficient determinace je jedna z veličin pro hodnocení regresní analýzy[9]. Pro jeho vymezení je třeba definovat některé základní pojmy. Úplný součet čtverců (TSS) je součet kvadrátů rozdílů pozorované hodnoty vysvětlované proměnné a průměrné hodnoty:

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{2.29}$$

Úplný součet čtverců je možné rozložit na dvě složky:

- reziduální součet čtverců (RSS):

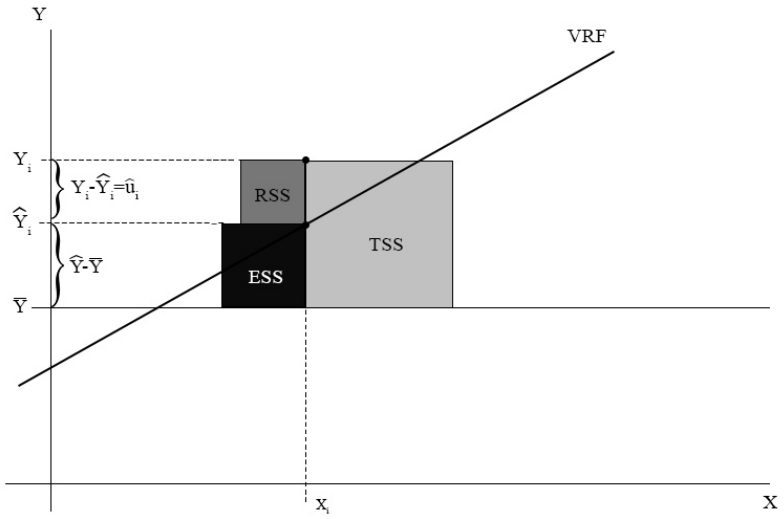
$$RSS = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2, \tag{2.30}$$

- vysvětlený (regresní) součet čtverců (ESS):

$$ESS = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2. \tag{2.31}$$

Použitím Pythagorovy věty platí (viz obrázek 2.7):

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2 = RSS + ESS. \tag{2.32}$$



Obrázek 2.7: Rozklad součtu čtverců TSS

Koeficient determinace R^2 je poté definován, jako podíl vysvětlovaného součtu čtverců a celkové součtu čtverců:

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}. \tag{2.33}$$

Udává stupeň vysvětlení závislé proměnné Y našeho regresního modelu[9].
Hodnota R^2 má několik vlastností:

- Nabývá hodnoty v intervalu $< 0, 1 >$.
- Pokud $R^2 = 1$, všechna výběrová pozorování leží přímo na vyrovnané regresní přímce. Nejlepší možná možnost.

- Pokud $R^2 = 0$, tak ani jedno pozorování neleží na regresní přímce a nepodařilo se nám vysvětlit žádnou část vysvětlované proměnné. Regresní model nemá smysl.

Z koeficientu determinace lze odvodit koeficient korelace R vztahem 2.34:

$$R = \pm\sqrt{R^2}. \tag{2.34}$$

Vzhledem k tomu, že s koeficientem determinace je spojeno několik problémů, které spočívají v tom, že adekvátně nereaguje na změny v počtu pozorování a nezohledňuje rozšíření počtu vysvětlujících proměnných, tak se z těchto důvodu používá korigovaný koeficient determinace[9].

2.2.5 Testování hypotéz o odhadnutých regresních parametrech

Po vytvoření jednoduchého regresního modelu metodou nejmenších čtverců začíná fáze statistické verifikace a dalšího testování hypotéz o odhadnutých parametrech i celého modelu. Základní principy testování hypotéz lze shrnout do tří základních fází[9]:

- formulace nulové a alternativní hypotézy(H_0, H_A),
- výpočet testovací statistiky,
- aplikace nebo použití rozhodovacího pravidla o přijetí, nebo zamítnutí nulové hypotézy pro stanovenou hladinu významnosti.

Toto testování může probíhat prostřednictvím oboustranného resp. jednostranného testu. Vzhledem k rozsahu práce se touto problematikou dále nezabývám a podrobnější informace ke statistické teorii testování hypotéz lze najít v publikacích [10] a [14].

2.3 Logistický regresní model

Dále se budeme věnovat logistickému regresnímu modelu. Základním rozdílem mezi lineárním a logistickým regresním modelem spočívá typech proměnných. Logistický, na rozdíl od lineárního pracuje s kategoriální závislou proměnnou. Například přítomnost/nepřítomnost choroby, existence atd.. Odhaduje míru pravděpodobnosti, že dané nezávislé proměnné $x_1, \ldots, x_n$ budou zařazené do určité kategorie. Z hlediska data miningu patří logistická regrese ke klasifikačním metodám[6].
Podle závislé proměnné se rozlišuje logistická regrese na:

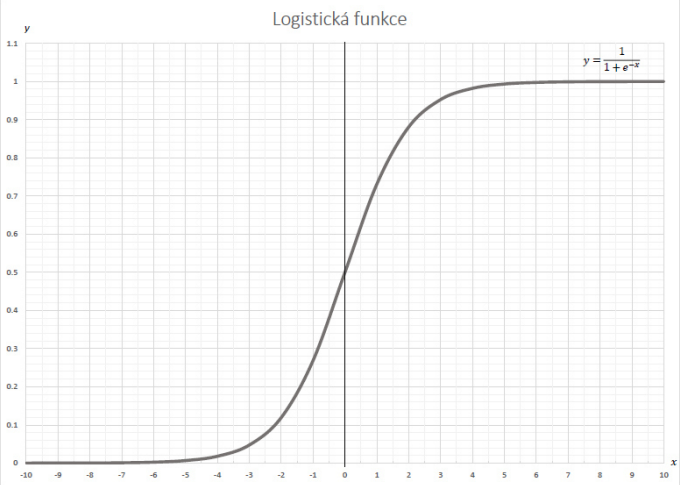
- binární (dichotomická) - nebývá pouze dvou hodnot, např. ano/ne, 1/0,
- ordinální - závislá proměnná nabývá více hodnot, mezi kterými existuje přirozené uspořádaní,
- (multi) nominální - závislá proměnná nabývá více než dvou hodnot, mezi kterými existuje pouze odlišnost, to znamená, že je nelze řadit, např. rasy, náboženství atd..

V logistickém regresním modelu je třeba určit, z jakou pravděpodobností nastane jev Y, jestliže nabývá hodnot 0 $\implies$ jev nenastal a 1 $\implies$ jev nastal. Lineární regresní model nelze použít z důvodu, že cílová proměnná je kategoriálního typu. Z rovnice (2.35) je patrné, že na levé straně jsou pouze dvě hodnoty 0 a 1 (může být i více kategorií), zatímco pravá strana rovnice nabývá libovolných hodnot.

$$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_i \tag{2.35}$$

Z těchto důvodů využijme logistickou funkci:

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{2.36}$$



Obrázek 2.8: Logistická funkce

Obrázek funkce (viz obr. 2.8) zobrazuje, že nabývá hodnot pouze v intervalu $(0, 1)$. Nyní tedy definujeme logistickou regresní funkci jako:

$$P(\widehat{Y}_i = 1 | X_i = x_i) = \frac{1}{1 + e^{-(\widehat{\beta}_1 + \widehat{\beta}_2 x_i)}}. \tag{2.37}$$

Pro odhady koeficientů $\widehat{\beta}_1$ a $\widehat{\beta}_2$ použijeme metodu maximální věrohodnosti.

2.3.1 Metoda maximální věrohodnosti

Tato metoda patří ke skupině základních metod bodových odhadů. Jedním z prvních pojmů, které je třeba definovat je tzv. věrohodnostní funkce.

Nechť $X = (X_1, \ldots, X_n)$ je náhodný výběr a $x = (x_i, \ldots, x_n)$ je jeho realizace. Dále nechť je populace (náhodný výběr) popsána pomocí určitého rozdělení $f(x, \Theta)$, kde θ je neznámý parametr. Potom funkci 2.38 nazveme věrohodnostní funkcí[15].

$$L(x, \theta) = L(x_1, \ldots, x_n, \theta) = f(x_1, \theta) f(x_2, \theta), \ldots f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \Theta). \tag{2.38}$$

Metoda maximální věrohodnosti spočívá v tom, že za odhad neznámého parametru (neznámých parametrů) zvolí hodnota $\widehat{\theta}$, která při daných hodnotách maximalizuje funkci věrohodnosti. Za předpokladu, že existuje bod $\widehat{\theta}$ z parametrického prostoru, takový, že pro všechny hodnoty parametru $\widehat{\theta}$ z parametrického prostoru platí: $L(X,\theta) \leq L(X,\widehat{\theta})$, potom nazveme tento bod maximálně věrohodným odhadem neznámého parametru $\widehat{\theta}$[6]. Dále pro jednoduchost budeme psát pouze tvar $L(\theta)$. Často je výhodnější použít místo věrohodnostní funkce její logaritmický tvar:

$$l(\theta) = lnL(\theta). \tag{2.39}$$

Tuto rovnici zapíšeme jako:

$$l(\Theta) = ln(\prod_{i=1}^{n} f(x_i,\Theta)) = \sum_{i=1}^{n} lnf(x_i,\Theta). \tag{2.40}$$

Tuto úpravu můžeme použít z důvodu, že logaritmická funkce je monotónní, tj. má-li funkce $L(\theta)$ maximum v bodě $\widehat{\theta}_{ML}$ má v tomtéž bodě maximum i funkce $lnL(\theta)$[6].

Pro nalezení maxima $\widehat{\theta}_{ML}$ použijeme metodu z matematické analýzy a to hledání extrémů funkce $l(\theta)$. Provedeme parciální derivaci podle parametru θ. Tím získáme systém věrohodnostních rovnic:

$$\frac{\delta L(\theta)}{\delta\theta_j} = 0, \quad j = 1,\ldots,m, \tag{2.41}$$

s řešením $\theta = \widehat{\theta}$. Musíme ověřit, zda v bodě $\widehat{\theta}$ nabývá funkce $L(\theta)$ svého maxima, musí tedy platit:

$$H(\widehat{\theta}) = \left(\frac{\delta^2 L(\theta)}{\delta\theta_i\delta\theta_j}\right)_{i,j=1}^{m}\Bigg|_{\theta=\widehat{\theta}} < 0 \tag{2.42}$$

tedy, že Hessova matice $H(\widehat{\theta})$ je negativně definitní[6].

2.3.2 Odhad koeficientů u logistického regresního modelu

Pro určení koeficientů budeme postupovat podle výše uvedené metody maximální věrohodnosti. Mějme náhodný výběr $Y_1,\ldots,Y_n$ alternativního rozdělení $A(\vartheta), 0 < \vartheta < 1$,s realizacemi $y_1,\ldots,y_n$[6].

$$P(Y_i = y_i) = \vartheta^{y_i}(1-\vartheta)^{1-y_i} \tag{2.43}$$

Pro střední hodnotu platí $E(Y_i) = \vartheta$ a pro rozptyl $D(Y_i) = \vartheta(1-\vartheta)$. Každému y_i přísluší realizace $x_{i1},\ldots,x_{in}$ veličin $X_{i1}\ldots,X_{in}$. Potom podle 2.37 modelujeme pravděpodobnost jako[6]:

$$\begin{aligned} P(Y_i = y_i|X_i = x_i) &= \left(\frac{1}{1+e^{-(\beta_1+\beta_2 x_i)}}\right)^{y_i}\left(1-\frac{1}{1+e^{-(\beta_1+\beta_2 x_i)}}\right)^{1-y_i} = \\ &= \frac{\left(e^{-(\beta_1+\beta_2 x_i)}\right)^{y_i-1}}{1+e^{-(\beta_1+\beta_2 x_i)}} \end{aligned} \tag{2.44}$$

Věrohodnostní funkce je poté ve tvaru:

$$L(\beta) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \frac{(e^{-(\beta_1+\beta_2 x_i)})^{1-y_i}}{1+e^{-(\beta_1+\beta_2 x_i)}}. \tag{2.45}$$

Použijeme logaritmickou věrohodnostní funkci (2.40), pomocí které z násobení dostaneme sčítání:

$$l(\beta) = ln(L(\beta)) = ln\left(\prod_{i=1}^{n} \frac{(e^{-(\beta_1+\beta_2 x_i)})^{1-y_i}}{1+e^{-(\beta_1+\beta_2 x_i)}}\right) = \sum_{i=1}^{n} ln\left(\frac{(e^{-(\beta_1+\beta_2 x_i)})^{1-y_i}}{1+e^{-(\beta_1+\beta_2 x_i)}}\right) =$$
$$= \sum_{i=1}^{n} [(y_i - 1)(\beta_1 + \beta_2 x_i) - ln(1 + e^{-(\beta_1+\beta_2 x_i)})]. \tag{2.46}$$

Nyní provedeme parciální derivace:

$$\frac{\delta(l(\beta))}{\delta\beta_1} = \sum_{i=1}^{n}(y_i - 1) + \frac{e^{-}(\beta_1 + \beta_2 x_i)}{1 + e^{-}(\beta_1 + \beta_2 x_i)} = 0,$$
$$\frac{\delta(l(\beta))}{\delta\beta_2} = \sum_{i=1}^{n}(y_i - 1)x_i + \frac{e^{-}(\beta_1 + \beta_2 x_i)}{1 + e^{-}(\beta_1 + \beta_2 x_i)} x_i = 0. \tag{2.47}$$

Rovnice dále upravíme:

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{1}{1+e^{\beta_1+\beta_2 x_i}} = 0,$$
$$\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{1}{1+e^{\beta_1+\beta_2 x_i}} x_i = 0. \tag{2.48}$$

Jedná se o soustavu nelineárních rovnic o dvou neznámých. Řešením těchto rovnic jsou koeficienty β_1 a β_2. Toto řešení nelze nalézt v algebraickém tvaru, proto se hledá numericky například pomocí Newtonovy-Raphsonovy metody. Více k numerickým metodám viz publikace[12].

2.4 Vícerozměrný lineární regresní model

V této kapitole se budeme věnovat rozšíření lineárního regresního modelu pro n vysvětlujíc proměnných, tedy $X_1 \ldots, X_n$. V praxi se budeme s tímto typem regresního modelu setkávat mnohem častěji, než s jednoduchou lineární regresí, protože vysvětlovaná proměnná Y je ovlivněna celou řadou dalších příčinných faktorů $X_1, \ldots, X_n$. Zařazení těchto faktorů do modelu přispěje k vyšší míře vysvětlení závislé proměnné Y[9].

Stejně jako u jednoduchého lineárního regresního modelu formulujeme deterministickou populační regresní funkci jako:

$$E(Y_i|X_{i2},\ldots,X_{ij}) = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_j X_{ij} \quad i = 1,2,\ldots,n. \tag{2.49}$$

Zahrnutím náhodné složky definujeme stochastickou PRF:

$$E(Y_i|X_{i2},\ldots,X_{ij}) = Y_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_j X_{ij} + u_i \quad i = 1,2,\ldots,n. \tag{2.50}$$

Tvar (2.50) můžeme pro jednotlivé hodnoty $i = 1,2\ldots,n$ rozepsat jako:

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{12} + \beta_3 X_{13} + \cdots + \beta_j X_{1j} + u_1 \\ Y_2 &= \beta_2 + \beta_2 X_{22} + \beta_3 X_{23} + \cdots + \beta_j X_{2j} + u_2 \\ &\vdots \\ Y_n &= \beta_n + \beta_2 X_{n2} + \beta_3 X_{n3} + \cdots + \beta_j X_{nj} + u_n. \end{aligned} \tag{2.51}$$

Soustavu rovnic 2.51 můžeme zapsat pomocí maticového zápisu ve tvaru:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{12} & X_{13} & \cdots & X_{1j} \\ 1 & X_{22} & X_{23} & \cdots & X_{2j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & X_{n3} & \cdots & X_{nj} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_j \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}. \tag{2.52}$$

Tento tvar vyjádříme prostřednictvím matice a vektorů:

$$\vec{y} = X \times \vec{\beta} + \vec{u}. \tag{2.53}$$

Kde:

$\vec{y}$: je vektor $(n \times 1)$ vysvětlované proměnné Y_i,

X: je matice $(n \times j)$ vysvětlujících proměnných X_i, kde první sloupec je jednotkový vektor, který odpovídá úrovňové konstantě,

$\vec{\beta}$: je vektor $(j \times 1)$ regresních koeficientů,

$\vec{u}$: je vektor $(n \times 1)$ náhodné složky.

Odhad pro výběrovou regresní funkci zapíšeme obdobně jako u jednoduchého lineárního regresního modelu ve tvaru[9]:

$$\vec{y} = X \times \vec{\widehat{\beta}} + \vec{\widehat{u}}. \tag{2.54}$$

Pro odhad neznámých parametrů $\widehat{\beta}_1,\ldots,\widehat{\beta}_j$ můžeme použít metodu nejmenších čtverců, metodu maximální věrohodnosti nebo zobecněnou metodu momentů. Následující kapitola se zaměřuje na první z uvedených metod.

2.4.1 Metoda nejmenších čtverců pro vícerozměrný lineární regresní model

Princip metody nejmenších čtverců pro vícerozměrný lineární regresní model je stejný jako pro jednoduchý. Hledáme odhady pro neznáme parametry $\vec{\widehat{\beta}} = \widehat{\beta_1} \ldots, \widehat{\beta_n}$ regresní funkce 2.54. Z této funkce vyjádříme reziduální složku:

$$\vec{\widehat{u}} = \vec{y} - X \times \vec{\widehat{\beta}}. \tag{2.55}$$

Tedy opět hledáme minimalizace součtu kvadrátu reziduálních složek $\vec{\widehat{u}}$. Kvadrát zapíšeme jako násobek transponovaného vektoru $\vec{\widehat{u}}^T$ a vektoru $\vec{\widehat{u}}$ (2.60) [9].

$$\vec{\widehat{u}}^T \vec{\widehat{u}} = (\vec{y} - X\widehat{\beta})^T(\vec{y} - X\widehat{\beta}) = \vec{y}^T\vec{y} - \vec{\widehat{\beta}}^T X^T\vec{y} - \vec{y}^T X\vec{\widehat{\beta}} + \vec{\widehat{\beta}}^T X^T X\vec{\widehat{\beta}} \tag{2.56}$$

Rovnici popsanou výše můžeme upravit do tvaru 2.57, protože platí, že $(\vec{y}^T X\vec{\widehat{\beta}})^T = \vec{\widehat{\beta}}^T X^T\vec{y}$, neboli transponovaný skalár je roven skaláru[11].

$$\vec{\widehat{u}}^T \vec{\widehat{u}} = (\vec{y} - X\widehat{\beta})^T(\vec{y} - X\widehat{\beta}) = \vec{y}^T\vec{y} - 2\vec{\widehat{\beta}}^T X^T\vec{y} + \vec{\widehat{\beta}}^T X^T X\vec{\widehat{\beta}} \tag{2.57}$$

Pro nalezení minima funkce 2.57 použijeme metodu matematické analýzy hledání extrému. Funkci parciálně zderivujeme podle $\vec{\widehat{\beta}}$ a položíme rovnu nule:

$$\frac{\delta(\vec{\widehat{u}}^T\vec{\widehat{u}})}{\delta\widehat{\beta}} = -2X^T\vec{y} + 2X^T X\vec{\widehat{\beta}} = 0. \tag{2.58}$$

Vyjádřením $\vec{\widehat{\beta}}$ získáme řešení ve tvaru:

$$\vec{\widehat{\beta}} = (X^T X)^{-1} X^T\vec{y}. \tag{2.59}$$

Jednou z metod výpočtu koeficientů jsou například numerické metody. Mezi ně patří také gradientní metoda, která je dále popsána v praktické části. Ostatními metodami se práce vzhledem rozsahu dále nezabývá a jejich popis je k dispozici v publikaci [12].

2.4.2 Rozšířené předpoklady pro metodu nejmenších čtverců

Následující kapitola rozšiřuje předpoklady pro vícerozměrný lineární regresní model[9].

- **P1**: Lineární regresní model $\vec{y} = X \times \vec{\beta} + \vec{u}$ je lineární v parametrech.
- **P2**: Matice X není stochastická tzn., že výběrový soubor má pevně dané proměnné $X_2, X_3, \ldots, X_n$
- **P3**: Střední hodnota náhodné složky je nulová $E(\vec{u}) = 0$

- **P4** a **P5**: Další dva předpoklady homoskedasticity a sériové nezávislosti náhodné složky můžeme můžeme vyjádřit současně prostřednictvím variačně-kovarianční maticí náhodných složek. Také na tomto předpokladu objasníme, příčinu násobení vektorů $\vec{u}^T\vec{u}$ (2.60).
  Vynásobením těchto dvou vektorů dostaneme následující tvar:

$$\vec{u}^T\vec{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \times \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix} = \begin{bmatrix} u_1u_1 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2u_2 & \cdots & u_2u_n \\ \vdots & \vdots & \ddots & \vdots \\ u_nu_1 & u_nu_2 & \cdots & u_nu_n \end{bmatrix} \tag{2.60}$$

  Podle předpokladů pro jednoduchý lineární regresní model víme, že podle 2.26 a 2.27 platí:

$$var(u_i|X_i) = E(u_i^2|X_i) = \sigma^2 \tag{2.61}$$

$$cov(u_i; u_j|X_i; X_j) = 0 \quad pro\, i \neq j \tag{2.62}$$

  Můžeme tedy matici 2.60 přepsat do tvaru

$$\begin{bmatrix} u_1u_1 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2u_2 & \cdots & u_2u_n \\ \vdots & \vdots & \ddots & \vdots \\ u_nu_1 & u_nu_2 & \cdots & u_nu_n \end{bmatrix} = \begin{bmatrix} var(u_1) & cov(u_1; u_2) & \cdots & cov(u_1; u_n) \\ cov(u_2; u_1) & var(u_2) & \cdots & cov(u_2; u_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(u_nu_1) & cov(u_nu_2) & \cdots & var(u_n) \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \cdot \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma^2 \cdot I_n. \tag{2.63}$$

  Předpoklad homoskedasticity náhodné složky vyjadřuje, že pro každé $x_{ij}\epsilon I_n$, kde $i = j$ bude $x_{ij} = 1$ a druhý předpoklad sériové nezávislosti náhodné složky (nepřítomnost autokorelace) vyjadřují prvky $x_{ij}\epsilon I_n$, kde $i \neq j$ takové, že $x_{ij} = 0$.

- **P6**: Tento předpoklad vyjadřuje nekoleraci sloupců matice X s vektorem náhodné složky $\vec{u}$.

$$E(X^T\vec{u}) = 0 \tag{2.64}$$

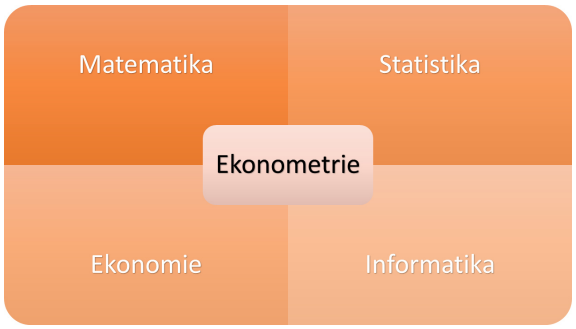
- **P7**: Počet nezávislých řádků se rovná součtu sloupců a ten je menší nebo roven počtu řádků této matice (počet pozorování).

$$h(X) = k \leq n \tag{2.65}$$

- **P8**: Náhodná složka má normální rozdělení $\vec{u} \approx N(0; \sigma^2 \cdot I_n)$

3 Ekonometrické modelování

Tato kapitola se věnuje ekonometrickému modelování. První pojem, který vymezuje je ekonometrie samotná. Jde o vědní disciplínu, jejímž cílem je pomocí kvantitativní a kvalitativní analýzy ověřovat závěry ekonomických teorií s využitím matematických nástrojů a statistické dedukce [9]. Tedy využívá matematiky, statistiky a informatiky pro hledání, měření a inferenci[1] vzájemných funkčních vztahů mezi ekonomickými veličinami v modelu. Počátky ekonometrie směřují ke 30 letům 20. století, kdy byla založena ekonometrická společnost (Econometric Society[2]). Hlavními důvody vzniku byla velká hospodářská krize, kritika ekonomického výzkumu a myšlenka využití matematiky a statistiky v ekonomii.

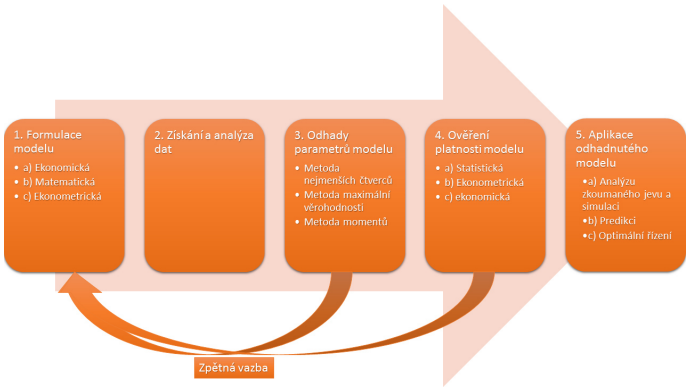


Obrázek 3.1: Ekonometrie

3.1 Proces ekonometrického modelování

Při ekonometrickém modelování je třeba dodržovat určitý metodologický postup pro vytvoření modelu, který je shrnut na obrázku 3.2. Dodržováním tohoto procesu se vyvarujeme chyb, které by jinak mohli nastat. První část proces zahrnuje určitou formulaci modelu a získání testovacích dat. Poté odhad regresních parametrů a jejich otestování. Poslední fází je aplikace vytvořené modelu. Následující kapitoly se věnují jednotlivým částem procesu podrobně.

[1] odvozování určitých výroků
[2] Mezinárodní společnost akademických ekonomů



Obrázek 3.2: Proces tvorby ekonometrického modelu

3.2 Formulace modelu

Počáteční fází procesu je formulace modelu. V této fázi určíme, jaký předmět chceme zkoumat, jaké proměnné obsahuje a matematicky ho definujeme[9]. Formulace modelu se skládá z několika fází.

První fází je vytvoření **ekonomického modelu**. Výsledkem bývá model, který podává informaci o možném vztahu mezi veličinami. Jedním z příkladů může být například spotřební funkce. Jedná se o závislost reálné spotřeby na reálném důchodu. Do této fáze můžeme zařadit:

- výběr zkoumaného předmětu,
- posuzování ekonomických veličin,
- popis vazeb a vztahů mezi veličinami,
- formulace základního tvrzení o chování ekonomických veličin.

Ekonomický model je z hlediska dalších fází velice důležitý a jeho zanedbání, či chybná interpretace, vede k závažným chybám.

Po této fázi následuje formulace **matematického modelu**.
Zde:

- vymezíme typy a rozsahy proměnných,
- ekonomický model transformujeme do analytického předpisu např. lineární, nelineární,
- nalezneme další očekávané vztahy, ať negativní nebo pozitivní a jiná omezení pro parametry.

Výsledkem této fáze může být například stanovení jednoduchého lineárního regresního modelu (viz rovnice 2.15).

Závěrečnou fází je vytvoření **ekonometrického modelu**. Spočívá zavedení náhodné složky u_i do matematického modelu, kterou budou stanoveny hypotézy o charakteru rozložení této chyby. Tímto z deterministického modelu získáme stochastický model. Výsledkem této fáze je například rovnice 2.16.

3.3 Získání a analýza dat

Sběr dat a jejich analýza je velice náročnou částí ekonometrického modelování. V prvním řadě je třeba najít vhodné informační zdroje a databáze. V těchto fázích jsou velice potřebné znalosti a dovednosti z ekonomické statistiky. Mezi vhodné informační zdroje patří:

- databáze Eurostatu,
- webový server Kaggle [8],
- Český statistický úřad [5],
- a další databáze.

Při analýze dat pracujeme obvykle s výběrovým souborem, nikoli ze základním, protože jeho získání není v mnoha případech možné. Prostřednictvím tohoto vzorku dat se snažíme zjistit informace o základním souboru. Vybraný výběrový datový soubor poté analyzujeme a upravujeme do podoby vhodné pro použití v dalších částí ekonometrického procesu. Mezi ně patří například časové, prostorové a obsahové vymezení, odstranění chybějících dat a eliminace odlehlých pozorování[9].

3.4 Odhady parametrů modelu

Tato část procesu se zabývá výběrem vhodné a dostupné metody odhadování parametrů. Selekce se provádí podle vlastností dat, složitosti modelovaného systému, dostupnosti technického a softwarového vybavení a v neposlední řadě také podle znalostí a zkušeností výzkumného pracovníka resp. výzkumného týmu[9].
Metody odhadu můžeme rozdělit do několika základních skupin:

- metoda nejmenších čtverců,
- metoda momentů,
- metoda maximální věrohodnosti.

3.5 Ověření platnosti modelu

V této fázi ověřujeme správnost a validitu vytvořeného modelu. Toto ověření probíhá na třech úrovních:

- statistické ověření,
- ekonometrické ověření,
- ekonomické ověření.

V případě, že v jakékoli z těchto tří úrovní zjistíme chybu, tak se vracíme zpět do předchozích částí modelování a podle příčiny chyby provádíme korekci modelu.

3.6 Aplikace odhadnutého modelu

Za předpokladu bezchybného ověření modelu následuje jeho aplikace.
Můžeme ji rozdělit do tří skupin:

- predikce budoucího vývoje,
- analýza vývoje nebo chování,
- využití odhadnutého modelu k optimálnímu řízení hospodářské politiky (simulace scénářů a jejich dopadů).

3.7 Příklady aplikací ekonometrického modelování

Mezi případy ekonometrického modelování se řadí nesčetně typových úloh.

Stručným příkladem ekonometrické modelování může být například jednoduchá spotřební funkce keynesiánského typu[3] pro české domácnosti, kde se predikuje vývoj spotřeby domácnosti s určitým měsíčním příjmem. První částí procesu podle kapitoly 3.2 má být formulace a popis ekonomického modelu spotřební funkce. Tedy stanovení předmětu zkoumání, tím jednoduchá spotřební funkce. Klasifikování ekonomických veličin - C_i (reálná spotřeba i-té domácnosti) a Y_i (příjem i-té domácnosti). Dále popis vztahu mezi veličinami. Zde se jedná o přímou závislost spotřeby na příjmu. A v poslední řadě formulace základního tvrzení o chování ekonomických veličin - tj. spotřeba roste pomaleji než důchod. Následuje formulace matematické modelu. Zde vymezíme klíčové proměnné použité v modelu - C_i spotřeba i-té domácnosti (v Kč/rok), Y_i - reálný příjem (v Kč/rok) a provedeme transformaci ekonomického modelu do analytické formy funkčního předpisu[9]:

$$C_i = \beta_1 + \beta_2 Y_i, \quad i = 1, 2, \ldots, n. \tag{3.1}$$

Poslední fází je formulace ekonometrického modelu spotřební funkce, která předpokládá zavedení náhodné složky u_i do rovnice 3.1:

$$C_i = \beta_1 + \beta_2 Y_i + u_i, \quad i = 1, 2, \ldots, n. \tag{3.2}$$

Následující krok by spočíval v získání a analýze dat, se kterými se bude pracovat. K tomuto účelu se mohou použít zdroje uvedené v 3.3. Další částí procesu by byl výběr metody odhadu regresních parametrů modelu a jejich vypočtení. Poté by následovalo otestování modelu. Za předpokladu úspěšného otestování by následovalo využití odhadnutého a verifikovaného modelu pro predikci, nebo bližší analýzu zkoumaného problému.

[3] Keynes, J. M.: The general theory of employment, interest and money. Kissimmee,USA:Singnalman Publishing, 2009, ISBN 978-0-9840614-0-2, 264 s.

4 Praktická část

4.1 Zkušenosti s MOOC kurzem na portále Coursera

Minulý rok (2016) jsem v rámci předmětu Dataming absolvoval online kurz Machine Learning (*Strojové učení*) na MOOC portále Coursera[4]. Jednalo se o mou první zkušenost s MOOC kurzem. Tento kurz trval přesně jedenáct týdnů a po jeho složení sliboval certifikát, který dokazuje znalost probíraných okruhů. Výuka byla předkládána formou přednášek a cvičení. Přednášky byly ve formě komentovaných videí s pauzami pro testy zaměřené na aktuální probírané téma. Každá video trvalo průměrně 10 až 15 minut. Cvičení byla vždy na konci probraného tématu a spočívaly v naprogramování zadaných témat v programu Octave. Pro všechny úlohy byli vytvořeny unit testy, které provedly kontrolu výsledků. Předmět byl vyučován v angličtině a k dispozici byli další čtyři jazyky (čeština mezi nimi nebyla). Časová náročnost každé lekce byla okolo 5-7 hodin týdně.

Mezi probrané okruhy patřily:

- regresní modely (lineární, logistický a vícerozměrný),
- neuronové sítě,
- support Vector Machines,
- učení bez učitele (clustering),
- práce s obsáhlými daty,
- praktické příklady nasazení.

Po úspěšném absolvovaní všech testů a cvičení my byl nabídnut certifikát. Jeho hodnota byla 49 $. V době platby tato částka odpovídala 1 201 Kč. Po zaplacení částky jsem obdržel nabízený certifikát (viz příloha B).

Forma předkládání informací byla na velmi dobré úrovni a z těchto důvodů mě kurz velice bavil. Nemohu jinak než tento kurz doporučit. Navíc informace poskytnuté v kurzu jsou k dispozici i po jeho absolvování, a tedy jsou výborným informačním zdrojem, za kterého čerpám i v této práci.

4.2 Kurz na ALS portále

V rámci této kapitoly jsem vytvořil výkladovou studii do kurzu Datamaning na ALS portále. Zpracoval jsem kapitoly jednoduchého, logistického, vícerozměrného regresního modelování a téma ekonometrie do několika výukových materiálů a vložil do kurzu. Tyto materiály budou rámci kurzu ke stažení. Dále jsem do kurzu přiložil také moji zpracovanou případu studii o odhadu ceny nemovitostí včetně mého proudu vytvořeného v programu Modeler, použitých dat a jejich popisu. Výslednou strukturu kurzu zobrazuje obrázek 4.1.

Obrázek 4.1: Snímek z vytvořeného kurzu Datamining

4.3 Případová studie - Odhad ceny nemovitosti

V této případové studii se budu zabývat odhadem ceny nemovitosti v závislosti na informacích o její kvalitě, rozloze a několika dalších údajích. Na základě ekonomické teorie předpokládám, že s rostoucí plochou nemovitosti a její kvalitou bude cena růst. Tato úloha se řadí mezi modelové příklady pro použití vícerozměrného lineárního regresního modelu.

4.3.1 Popis dat

K dispozici mám data, které udávají několik údajů o nemovitostech. Získal jsem je ze serveru kaggle.com, kde byla předána komunitě různých týmům nebo jednotlivců, aby je analyzovali a vytvářeli nad nimi své modely. Na serveru kaggle.com jsou k datům často vypisované soutěže, to ale nebyl tento případ. Datový soubor zahrnuje informace o 1460 amerických nemovitostí, konkrétně z města Ames v Iowě. Všechny plošné jednotky jsou ve čtverečních stopách a použitá měna pro cenu nemovitosti jsou dolary. Každý řádek výběrového souboru obsahuje jednu nemovitost a informace o ní, konkrétně 81 parametrů. Pro analýzu a následné použití v modelu jsem vybral 26 atributů. Jsou uvedeny v tabulkách 4.1 a 4.2. Kompletní seznam všech atributů je uveden v příloze D. Ostatní atributy jsem nezahrnul do případové studie z důvodu nízké variability hodnot a příliš nízké souvislosti s konečnou cenou nemovitosti.

Atribut	**Popis**
YearBuilt	Rok výstavby
YearRemodAdd	Rok rekonstrukce
OverallQual	Celkové ohodnoceni materiálu
OverallCond	Ohodnoceni celkového stavu nemovitosti
FullBath	Celkový počet koupelen
BedroomAbvGr	Počet ložnic
KitchenAbvGr	Počet kuchyní
TotRmsAbvGrd	Celkový počet místností bez koupelen
Fireplaces	Počet únikových východů
GarageYrBlt	Rok výstavby garáže

Tabulka 4.1: Vybrané kvalitativní atributy nemovitostí

Atribut	Popis
LotFrontage	Vzdálenost vzdušnou čarou od cesty k nemovitosti
LotArea	Celková rozloha nemovitosti
MasVnrArea	Plocha zdí
TotalBsmtSF	Celková plocha sklepa
1stFlrSF	Plocha prvního patra
2ndFlrSF	Plocha druhého patra
GrLivArea	Plocha přízemí
GarageArea	Celková plocha garáže ve čtverečních stopách
WoodDeckSF	Plocha dřevěné terasy
OpenPorchSF	Plocha otevřené verandy
EnclosedPorch	Plocha uzavřené verandy
3SsnPorch	Plocha zimní zahrady
ScreenPorch	Plocha prosklené verandy
PoolArea	Plocha bazénu
MiscVal	Hodnota nadstandardního vybavení
SalePrice	Prodejní cena

Tabulka 4.2: Vybrané kvantitativní atributy nemovitostí

4.3.2 Načtení dat do Modeleru

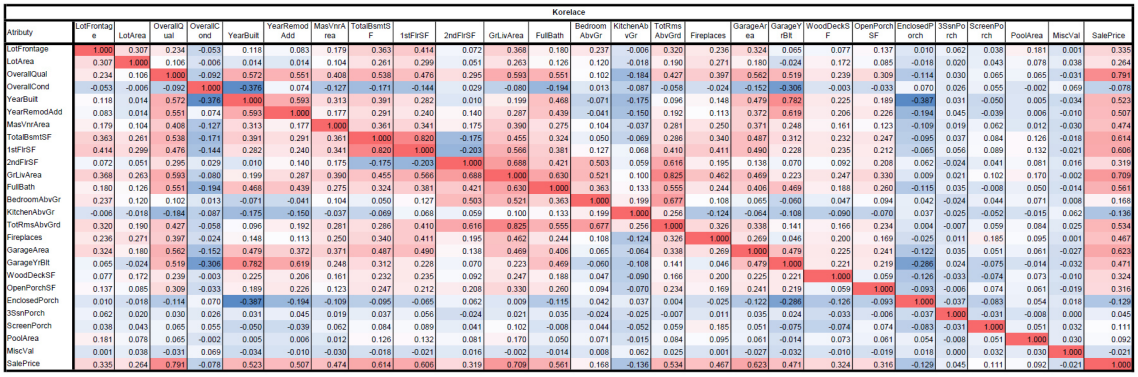
Pro načtení výběrového souboru s daty použiji blok *var. file*. Po jeho otevření jsem změnil oddělovač jednotlivých záznamů na čárku a potvrdil možnost, že data obsahují na prvním řádku popis atributů. Kontrolu, zda data byla načtena korektně provedu zařazením bloku *Table*, který zobrazí tabulku se všemi záznamy.

4.3.3 Analýza dat

Nejdříve jsem data testoval na chybějící a chybné hodnoty. Použil jsem blok *Data-Audit*, který podá základní statistické informace o všech atributech. Dále v záložce *Quality* zjistím úplnost dat vyjádřenou procenty. Tento blok mi podá informaci, že výběrový soubor je bez chybějících hodnot a všechny hodnoty jsou ve správných intervalech.

Následně určuji, jaké jsou závislosti jednotlivých atributů mezi sebou, a jaký mají vliv na cílovou proměnnou. Pro každou dvojici určím Pearsonův korelační koeficient a na základě těchto hodnot vytvořím korelační mapu, kterou zobrazuje obrázek 4.2. Otestoval jsem také Kandelův a Spearmanův koeficient, ale vzhledem k tomu, že jejich hodnoty se velmi podobaly Pearsnovu koeficientu a hledáme lineární závislosti, tak je nepoužiji. Korelační mapa mi podala informaci o tom, že mezi atributy je vysoká závislost a cena nemovitosti (`SalePrice`) je silně korelována téměř všemi vybranými atributy. Některé z vybraných atributů však na cenu nemovitosti nemají téměř žádný vliv, a některé opačný, než by se předpokládalo. Proto je do výsledného modelu nezařadím. Vyloučím tyto atributy `MiscVal, PoolArea, EnclosedPorch, 3SsnPorch, OverallCond, KitchenAbvGr,ScreenPorch a BedroomAbvGr`. Jejich

nízký vliv je způsoben nízkou variabilitou hodnot a pro velké množství nemovitostí nabývají nulových hodnoty.

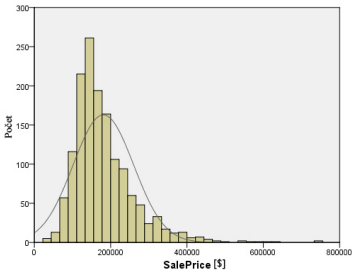


Obrázek 4.2: Korelační mapa atributů

Nyní se zaměřím na rozbor kvantitativních atributů. Prvním je cílový atribut `SalePrice`, který udává výslednou cenu nemovitosti v dolarech. Statistické informace udává tabulka 4.3 a obrázek 4.3 zobrazuje histogram rozložení hodnot. Na jeho základě usuzuji, že ceny nemovitostí jsou ve výběrovém soubor rozloženy rovnoměrně.

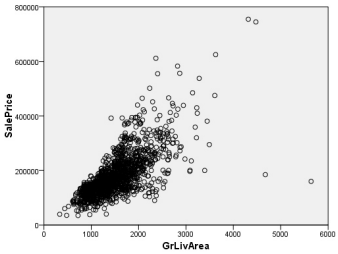
Název	**Hodnota**
Validní hodnoty	1460
Chybějící hodnoty	0
Průměr	180921.195891
Směrodatná odchylka	79442.502883
Kvartil 25%	129950
Medián	163000
Kvartil 75%	214000

Tabulka 4.3: Statistické údaje SalePrice

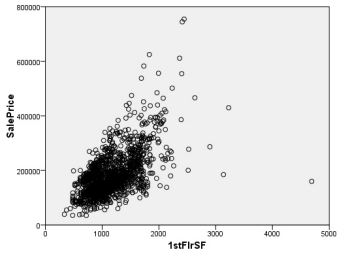


Obrázek 4.3: Histogram proměnné SalePrice

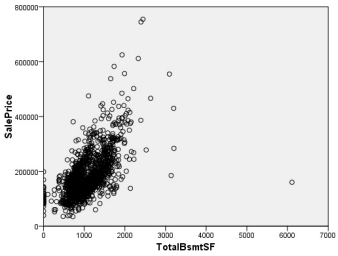
Z důvodu velkého počtu atributů v modelu zde provedu rozbor těch, které mají na cílovou proměnnou největší vliv. Patří mezi ně `GrLivArea`, `1stFlrSF`, `TotalBsmtSF` a `GarageArea`. Závislost těchto atributů na cílové proměnné zobrazuje soubor bodových grafů na obrázku 4.4. Ukazuje, že všechny mají pozitivní vliv na cenu nemovitosti a odpovídají vypočteným korelacím.



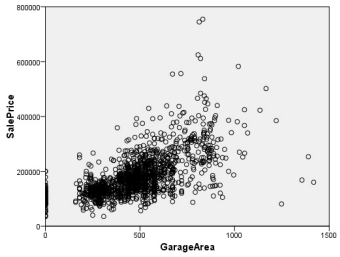
(a) GrLivArea



(b) 1stFlrSF



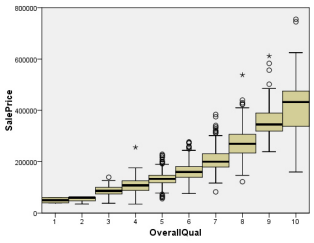
(c) TotalBsmtSF



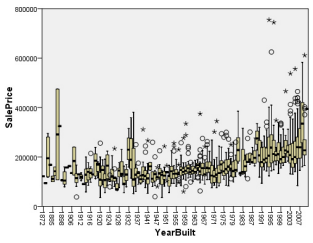
(d) GarageArea

Obrázek 4.4: Bodový graf závislostí

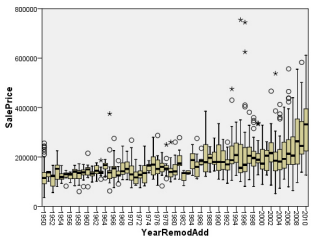
Nyní zpracuji kvalitativní atributy. Největší vliv na výslednou cenu má atribut `OverallQual a YearBuilt`. Parametr `OverallQual` popisuje celkovou kvalitu nemovitosti. Nabývá hodnoty $1 - 10$. Nejnižší hodnota odpovídá nejhorší kvalitě. Závislost těchto atributů na výsledné ceně zobrazují krabicové grafy na obrázku 4.5. Z krabicových grafů 5.6a vidím, že ve výběrovém souboru je většina nemovitostí s vyšší kvalitou. Dále, že menšina nemovitostí z nižší kvalitou má značně nižší cenu.



(a) OverallQual



(b) YearBuilt



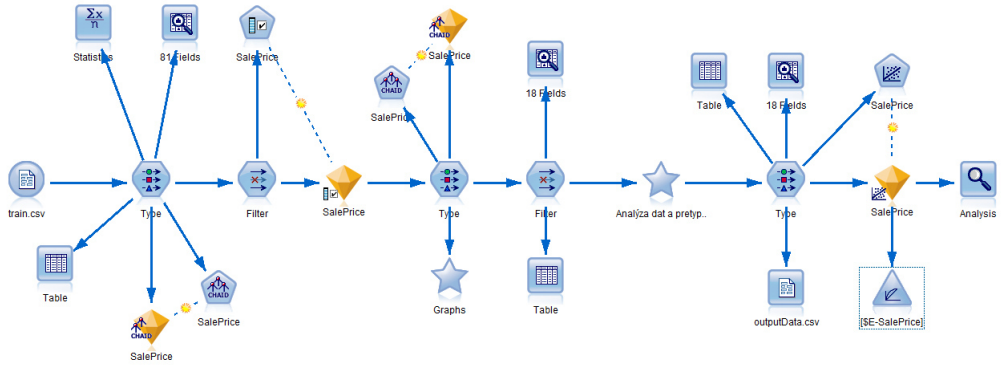
(c) YearBuiltAdd

Obrázek 4.5: Krabicové grafy

Výběrový soubor dále obsahuje tři atributy podávající časovou informaci o roku výstavby nemovitosti, její renovaci a výstavby garáže. Provedu kontrolu, zda jsou tyto hodnoty dostatečně unikátní. Například by se mohlo stát, že všechny nemovitosti budou mít tyto časové údaje stejné. Při jejich srovnání jsem zjistil, že pro 572 záznamů nabývají stejných hodnoty. Tento počet není příliš vysoký a z toho důvodu v regresním modelu použiji všechny tři atributy.

4.3.4 Tvorba modelu

Nyní přistoupím k odhadu lineárního regresního modelu pro výběrový soubor. V programu Modeler jsem použil blok *Regression*, který zařadím za blok *Type*. V tomto bloku jsem definoval typy použitých proměnných a nastavil jsem cílovou proměnnou, kterou je `SalePrice`. Celý proud spustím. Vytvořený model se zobrazí na pracovní ploše jako ikona žlutého diamantu. Po jeho otevření vidíme výsledný regresní model. Výstup obsahuje několik tabulek, ve kterých najdeme informace o hodnotách vypočtených koeficientů a o celkovém testování modelu. Výsledný proud vytvořený v Modeleru zobrazuje obrázek 4.6. Dále v kapitole testování modelu rozeberu výstupní informace, které nám podal vytvořený model.



Obrázek 4.6: Proud v aplikace Modeler

4.3.5 Testování modelu

Výstupní tabulky a grafy lze rozdělit do tří základních bloků:

- shrnutí základních regresních statistik o vytvořeném modelu (tabulka *Model Summary* - Tab. 4.4),
- výsledky analýzy rozptylu ANOVA (tabulka *ANOVA* - Tab. 4.5),
- odhady regresních parametrů (tabulka *Coefficients* - Tab. 4.6).

První tabulka *Model Summary* shrnuje pro každý odhadnutý model informace o koeficientu vícenásobné korelace, determinace a korigovaný koeficient determinace. V posledním sloupci je dále standardní chyba odhadu regrese.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.888[a]	.788	.785	36802.995

a. Predictors: (Constant), OpenPorchSF, WoodDeckSF, ...

Tabulka 4.4: Regresní statistika

Výsledky v tabulce 4.4 dokumentují, že odhadnutá cena nemovitosti je vyrovnána výběrovým souborem dat velmi dobře, tj. podle koeficientu determinace ($RSquare = 0,785$) je podíl vysvětlené regrese na celkovém součtu 78.5%. V případě porovnávání více modelů s různými rozsahy výběrových souborů použiji vyrovnaný koeficient determinace. Jeho výhody jsou popsány v kapitole 2.2.4.

Následuje tabulka 4.5 analýza rozptylu (**AN**ylysis **O**f **VA**riance). V této tabulce jsou ve druhém sloupci zachyceny jednotlivé součty čtverců, které jsou popsány v 2.2.4, další sloupec je počet stupňů volnosti, čtvrtý je podílem součtu čtverců a stupňů volnosti. Další sloupce slouží k testování statistické významnosti celého regresního modelu.

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.255E+12	17	4.268E+11	315.071	.000b
	Residual	1.953E+12	1442	1.35E+09		
	Total	9.208E+12	1459			

a. Dependent Variable: SalePrice
b. Predictors: (Constant), OpenPorchSF, WoodDeckSF, ...

Tabulka 4.5: Výsledky analýzy ANOVA

F-test je testem významnosti koeficientu determinace R^2. Pro její testování přistoupím k formulaci nulové a alternativní hypotézy:

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0, \qquad (4.1)$$
$$H_A : \beta_2 \neq 0 \vee \beta_3 \neq 0 \vee \cdots \vee \beta_k \neq 0.$$

Nulová hypotéza H_0 uvádí, že všechny regresní parametry spojené s cílovou proměnnou $\beta_2, \beta_3, \ldots, \beta_k$ jsou současně rovny nule s výjimkou úrovňové konstanty β_1. Alternativní hypotéza H_A znamená, že bude alespoň jedna vysvětlující proměnná statisticky vyznaným nenulovým regresním koeficientem. Pro rozhodnutí o zamítnutí nebo přijetí nulové hypotézy nejdříve stanovíme hladinu významnosti $\alpha = 0.05$ tj 5%. Pro tuto hladinu významnosti testuji platnost nulové hypotézy. Podle vypočtené hodnoty F-testu (porovnám s tabulkovou kritickou hodnotou) zamítám nulovou hypotézu a odhadnutý regresní model je statisticky významný na zvolené hladině významnosti α. Totéž by platilo na 1% hladině významnosti.

Poslední částí výstupu je tabulka *Coefficient* obsahující výsledky a statistiky pro odhadnuté regresní parametry (viz. Tab 4.6). Pro náš model tato tabulka zahrnuje několik řádků, které odpovídají odhadnutým regresním parametrům. První dva sloupce zaznamenávají nestandardizované odhady regresních parametrů β a jejich směrodatné odchylky. Tyto odhady jsou obtížně porovnatelné, proto jsou ve čtvrtém sloupci s označením *Stand. Coef. Beta* uloženy standardizované regresní parametry. Poslední dva sloupce v tabulce jsou opět určeny ke statistické významnosti jednotlivých regresních koeficientů (viz kapitola 2.2.5).

Coefficients						
Model	Atribute	Unstand. Coef. Beta	Std. Error	Stand. Coef. Beta	t statistika	Sig.
1	(Constant)	-1245621.962	134183.623		-9.283	.000
	LotFrontage	13.953	50.940	.004	.274	.784
	LotArea	.528	.107	.066	4.938	.000
	OverallQual	19220.507	1178.162	.335	16.314	.000
	YearBuilt	225.045	59.716	.086	3.769	.000
	YearRemodAdd	357.732	64.767	.093	5.523	.000
	MasVnrArea	32.400	6.197	.074	5.228	.000
	TotalBsmtSF	16.379	4.209	.090	3.891	.000
	1stFlrSF	31.432	21.373	.153	1.471	.142
	2ndFlrSF	22.992	21.023	.126	1.094	.274
	GrLivArea	16.421	20.865	.109	.787	.431
	FullBath	-3726.450	2638.392	-.026	-1.412	.158
	TotRmsAbvGrd	1084.673	1091.050	.022	.994	.320
	Fireplaces	8113.577	1814.274	.066	4.472	.000
	GarageArea	38.015	6.196	.102	6.136	.000
	GarageYrBlt	15.940	72.058	.005	.221	.825
	WoodDeckSF	32.132	8.253	.051	3.893	.000
	OpenPorchSF	5.820	15.830	.005	.368	.713

Tabulka 4.6: Odhady regresních parametrů

Výsledky dokumentují, že za předpokladu nulovosti všech koeficientů beta je cena nemovitosti -1245621.962 $. Podle standardizovaných koeficientů se potvrdilo, že největší vliv na výslednou cenu nemovitosti mají atributy `OverallQual`, celkové obyvatelná plocha a plošné výměry prvního a druhého patra. Naopak velice nízkého vlivu dosahuje atribut `FullBath`. Dále v tabulce 4.6 je v posledním sloupci označeném Sig. vypočtená p-hodnota, tj. sanovaná hladina významnosti, která odpovídá vypočtené t-statistice. Podrobnější informace ke statistické teorii testování hypotéz lze najít v publikaci [11]. U většiny byla tato vypočtená hodnota téměř nulová.

4.3.6 Zhodnocení případové studie

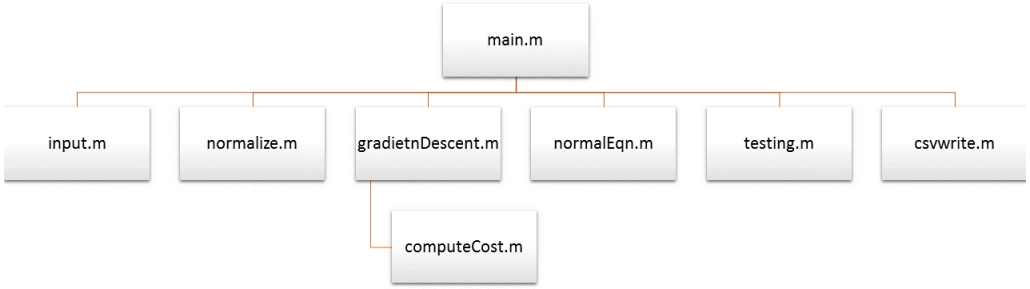
V případové studii jsem se zabýval odhadem ceny nemovitosti na základě několik známých skuteční o ní. Pomocí několika technik jsem provedl analýzu dat o nemovitostech a na jeho základě jsem vybral několik atributů, které jsem následně použil pro tvorbu regresního modelu. Vytvořený regresní model je na základě jeho statistik přesný a může být použit k další analýze nebo k predikci.

5 Vytvoření nezávislého modelu

V této části práce se budu zabývat tvorbou vlastního regresního modelu prostřednictvím nezávislé platformy. Pro jeho tvorbu použiji prostředí Octave. Řadí se mezi svobodný software a je šířený pod licencí GPL. Toto jsou hlavní důvody, kvůli kterým jsem ho zvolil jako nejvhodnější variantu, protože vytvořený program bude použit pro studijní účely a proto by prostředí, kde půjde spustit, mělo být snadno dostupné a bez potřeby pořízení licence.

5.1 Struktura programu

Program je rozdělen na několik částí, podle toho jakou funkci provádějí. Vstupem do programu je soubor `main.m`. Strukturu rozložení programu zobrazuje obrázek 5.1. V následujících kapitolách popíši jeho jednotlivé části.



Obrázek 5.1: Diagram funkcí

5.2 Čtení dat

Program přečte datový soubor, kde na každém řádku bude jednotlivý záznam a hodnoty budou odděleny libovolným oddělovačem. Je třeba, aby na prvním řádku byl uveden seznam všech atributů a další řádky odpovídaly pořadím tomuto seznamu. Jméno souboru společně s oddělovačem zadá uživatel při vstupu do programu. V případě neuvedení souboru se použije defaultní soubor s nemovitostmi a jako oddělovač čárka. V tomto případě budou data načítány z datového souboru `finalData.csv`. První řádek obsahuje textový popis atributů a na dalších jsou jednotlivé záznamy o nemovitostech. Hodnoty na řádku jsou odděleny čárkami. Vzorek souboru zobrazuje

zdrojový kód 5.1. Na těchto datech budu prezentovat funkčnost programu.

```
65,8450,7,2003,2003,196,856,856,854,1710,2,8,0,548,2003,0,61,208500
80,9600,6,1976,1976,0,1262,1262,0,1262,2,6,1,460,1976,298,0,181500
68,11250,7,2001,2002,162,920,920,866,1786,2,6,1,608,2001,0,42,223500
60,9550,7,1915,1970,0,756,961,756,1717,1,7,1,642,1998,0,35,140000
```

Zdrojový kód 5.1: Vzorek dat

Pro načtení souboru použiji funkci *input()*. Jako vstupní parametry přebírá jeho jméno a oddělovač hodnot. V této funkci načtu datový soubor a každý řádek rozděluji podle zadaného oddělovače. Na výstupu funkce vrací dvourozměrné pole `data` s jednotlivými hodnotami o velikosti počet záznamů × počet atributů a oddělenou hlavičku `h` z názvy atributů.

```
function [data,h] = input(fileName,delimeter)
```

Zdrojový kód 5.2: Funkce přo čtení

Kromě tohoto způsobu jsem otestoval také funkci *csvRead()*, která je součástí standardní instalace Octave. Tato funkce umožňuje načíst soubor typu *.csv*, ale čte pouze číselné hodnoty, tudíž jsem od jejího použití upustil i za cenu značného zvýšení doby čtení a parsování souboru. Čas výpočtu jednotlivých funkcí zobrazuje tabulka 5.1. K nárůstu došlo zejména z důvodu parsování a konverze hodnot do vhodné datové struktury.

funkce	**čas[s]**
input()	2.443
csvRead()	0.249

Tabulka 5.1: Porovnání času výpočtu funkcí pro čtení

Následně rozdělím atributy mezi závislé a nezávislé proměnné podle výběru uživatele. Závislou proměnnou zde představuje cena nemovitosti, kterou uložím do proměnné `y` a ostatní nezávislé atributy do dvourozměrného pole `X`.

5.3 Normalizace

Před výpočtem odhadů parametrů regresního modelu se může provést normalizace dat. Důvodem normalizace atributů je, že odhadnuté koeficienty regresního modelu budou standardizované. Důvody popisuji v kapitole testování 4.3.5. Normalizace provádím ve funkci *normalize()*. Jako vstup funkce přebírá datovou matici `X` s nezávislými proměnnými. Ná výstupu vrací pole hodnot, kde první parametr je normalizovaná datová matice, druhý a třetí parametr jsou střední hodnoty a směrodatné odchylky pro jednotlivé atributy.

```
function [X_norm, mu, sigma] = normalize(X)
```

Ve funkci nejprve určím střední hodnoty `mu` a směrodatné odchylky `sigma` pro jednotlivé atributy. Poté provedu normalizace hodnot. Od hodnot matice `X` na indexech i,j, kde j odpovídá j-tému atributy a i i-té hodnotě v atributu odečtu střední hodnotu a tento rozdíl dělím směrodatnou odchylkou. Algoritmus výpočtu zobrazuje zdrojový kód 5.3. Celková doba výpočtu funkce *normalize()* na datovém souboru s nemovitostmi odpovídá přibližně 1.038*s*.

```
for j = 1:size(X, 2)
 if (sigma(j) ~= 0)
  for i = 1:size(X, 1)
   X_norm(i, j) = (X(i, j)-mu(j))/sigma(j);
  end
 else
  % V tomto pripade hodnoty jsou nula
  % (stredni hodnota 0, odchylka sigma 0)
  X_norm(:, j) = zeros(size(X, 1), 1);
 end
end
```

Zdrojový kód 5.3: Normalizace hodnot

5.4 Odhad parametrů regresního modelu

Odhady koeficientů regresního modelu určuji pomocí metody nejmenších čtverců. Pro výpočet parametrů užívám dvě metody. První je analytické řešení pomocí standardní rovnice popsané v kapitole 2.4.1. Nevýhodu této metody může být použitelnost pouze do určitého počtu atributů (přibližně 1000). Jak je vidět v rovnici 2.59, výpočet inverzní matice může být při větším počtu atributů náročnější. Proto jako druhou užívám gradientní metodu nazývanou také gradient descent. Výpočet touto metodou nám umožňuje zadat stupeň učení a počet iterací. V následujících kapitolách provedu rozbor obou metod a doby čau výpočtu.

5.4.1 Standardní rovnice

Analytický výpočet regresních koeficientů metodou nejmenších čtverců provede funkce *normalEqn()*. Na vstupu funkce přebírá dva parametry, dvourozměrné pole nezávislých proměnných X a vybraný cílový atribut y. Před výpočtem je třeba vložit jednotkový vektor na počátek matice X. Formát matice X zobrazuje rovnice 2.52. Výpočet regresních parametrů provedu zadáním parametrů do rovnice 2.59. Algoritmus výpočtu zobrazuje zdrojový kód 5.4.

```
function [beta] = normalEqn(X, y)
%NORMALEQN Vypocet pomoci standartni rovnice
%NORMALEQN(X,y) X - nezavisle promenne,y - cilova promenna
%inicializace promenne beta
 beta = zeros(size(X, 2), 1);
%vypocet pomoci standartni rovnice
 beta = pinv(X'*X)*X'*y;
end
```

Zdrojový kód 5.4: Standartní rovnice

Funkce nejprve inicializuje proměnnou beta, která odpovídá počtu nezávislých atributů plus jednotkový vektor pro konstantní koeficient. Poté vektorově vypočítám jednotlivé koeficienty `beta`, které vracím jako výstup. Doba průběhu funkce nad testovacími daty je v průměru $0.002s$.

5.4.2 Gradientní metoda

Dalším příkladem výpočtu koeficientů regresní modelu pomocí nejmenších čtverců je gradientní metoda. Řadíme ji mezi iterační metody, které definujeme jako proces opakovaného použití funkce s cílem přiblížit se s určitou délkou kroku a počtu iterací co nejvíce k výsledku. Použiji funkci viz rovnice 2.58. Při výpočtu touto metodou nemusíme dosáhnout výsledku, protože se může stát, že metoda bude divergovat. Je tedy třeba ověřit její konvergenci. Postačující podmínku konvergence pro gradientní metodu je symetričnost matice $X^T X$[12]. Stačí ověřit, že matice je pozitivně definitní, tedy že všechny její hlavní minory jsou kladné. Pro násobek matic $(X^T \times X)$ tato podmínka je splněna a lze tedy konstatovat, že gradientní metodu lze pro danou soustavu použít.

Výpočet gradientní metodou provedu zavoláním funkce *gradientDescent()*. Jako vstupní parametry jsou cílová proměnná y, nezávislé atributy X, stupeň učení a počet iterací. Na počátku inicializujeme proměnnou `m`, která odpovídá rozsahu výběrového souboru, dále pole koeficientů `beta`. V proměnné `J_history` ukládám hodnoty funkce po každé iteraci a později mi umožní vykreslit křivku průběhu učení. V každé kroku cyklu vypočtu nové hodnoty funkce `gradJ` s aktuálními hodnotami koeficientů beta. Dále podle nastaveného stupně učení určím jejich nové hodnoty. Do proměnné `J_history` zapíšu hodnotu funkce *computCost()*. Přesnost a doba výpočtu metody bude záviset na daném stupni učení, tj. délka kroku a na počtu iterací. V případě volby příliš vysoké délky kroku se může stát, že metoda bude divergovat. Touto vlastností se budu v dalších částech této kapitoly zabývat. Zdrojový kód funkce zobrazuje 5.5.

```
function [beta, J_history] = gradientDescent(X, y, alpha, num_iters)
%GRADIENTDESCENT Vypocet koeficientu regresniho modelu gradientni metodou
%dochazi k postupnemu zpresnovani koeficientu beta
% inicializace promennych
m = length(y);
J_history = zeros(num_iters, 1);
beta = zeros(size(X, 2), 1);
%vypocet, num_iter - pocet itareci
for iter = 1:num_iters
  gradJ = 1/(2*m) * 2 * (X'*X*beta - X'*y);
  beta = beta - alpha * gradJ;
  J_history(iter) = computeCost(X, y, beta);
end
end
```

Zdrojový kód 5.5: Gradientní metoda

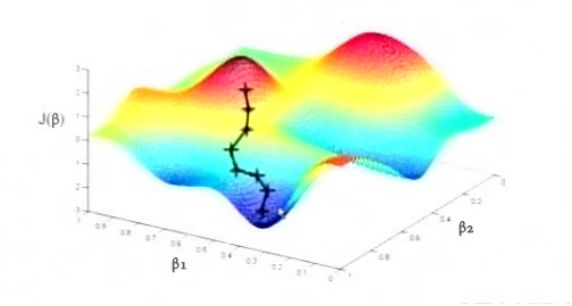
Provedu porovnání několik nastavení gradientní metody. Jako referenční hodnotu přebírám hodnotu koeficientu determinace modelu určené pomocí analytické formy

výpočtu, který odpovídá $R \cong 0.787885$. Tabulka 5.2 udává závislost počtu kroků na době průběhu výpočtu a přesnosti určených koeficientů. Vidíme, že čas roste lineárně s počtem provedených kroků a ze zvyšujícím se počtem iterací hodnota koeficientu determinace konverguje k referenční hodnotě.

Počet kroků	**Stupeň učení**	**čas[s]**	**Koeficient determinace**
50	0.1	≅.022	≅.780244
100	0.1	≅.041	≅.786300
1000	0.1	≅.368	≅.787869
10000	0.1	≅3.632	≅.787884
100000	0.1	≅36.297	≅.787885

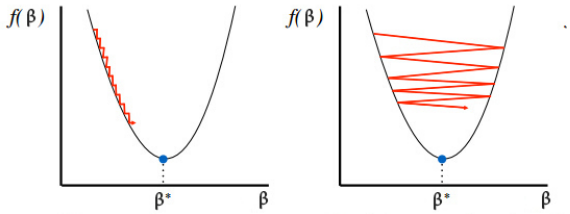
Tabulka 5.2: Závislost počtu kroků na době průběhu a přesnosti modelu

Stupeň učení bude ovlivňovat rychlost konvergence k referenční hodnotě. Průběh odhadu gradientní metodou pro koeficienty β_1 a β_2 (jednoduchý lineární regresní model) zobrazuje obrázek 5.2.



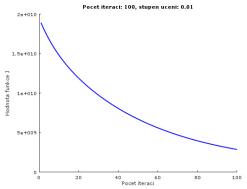
Obrázek 5.2: Gradientní metoda

Za předpokladu volby příliš velké délky kroku metoda nezachytí globální minimum a začne divergovat, naopak pokud je krok příliš malý, vykonávání metody skončí před tím, než nalezne globální minimum. Oba tyto případy pro odhad jednoho koeficientu β zobrazuje obrázek 5.3, nalevo je nastaven stupeň příliš nízký a naopak na pravém vysoký.

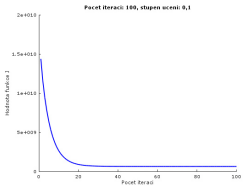


Obrázek 5.3: Nastavení stupně učení (délky kroku)

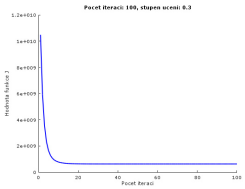
Výstupem provedené funkce jsou dvě pole, odhadnuté regresní koeficienty a průběh učení. Jeho hodnoty jsem zanesl do grafů a zobrazuje je soubor obrázků 5.4. Obsahuje průběhy křivky učení pro různá nastavení gradientní metody. Nejoptimálnější jsou na obrázcích 5.4b a 5.4c. Na tomto základě usuzuji, že nejlepšími nastaveními metody pro testovací data bude 50 až 100 iterací a stupeň učení $0,1$ nebo $0,3$.



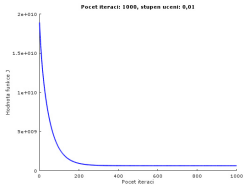
(a) 100 iterací, stupeň učení 0,01



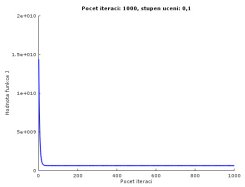
(b) 100 iterací, stupeň učení 0,1



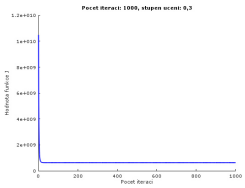
(c) 100 iterací, stupeň učení 0,3



(d) 1000 iterací, stupeň učení 0,01



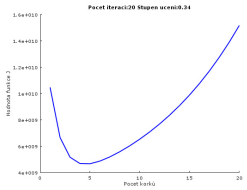
(e) 1000 iterací, stupeň učení 0,1



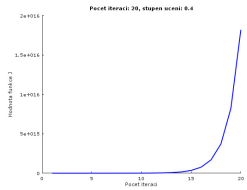
(f) 1000 iterací, stupeň učení 0,3

Obrázek 5.4: Křivky učení

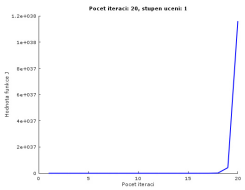
V případě nastavení kroku většího, jak 0.33 došlo k divergování metody. Abych mohl vykreslit požadovaný graf, nastavil jsem počet iterací na 20 a jeho průběh zobrazuje obrázek 5.5. Z grafu vidíme, že hodnota funkce roste exponenciálně do nekonečna.



(a) 20 iterací, stupeň učení 0,34



(b) 20 iterací, stupeň učení 0,4



(c) 20 iterací, stupeň učení 1

Obrázek 5.5: Divergování gradientní metody

5.5 Testování modelu

Nyní přistoupím k testování sestaveného regresního modelu. Provedu několik testů, které budou vypovídat o kvalitě vytvořeného regresního modelu a dále otestuji všechny

atributy. Jednotlivé testy provádím ve funkci *testing()*. Jako vstupní parametry přebírá hodnoty nezávislých atributů, cílovou proměnnou, odhad cílové proměnné na základě sestaveného modelu a jeho koeficienty. Celková doba výpočtu testování pro výběrový soubor s nemovitostmi se pohybovala okolo $0.005912s$. V následujících kapitolách se budu zabývat výpočtem jednotlivých testů regresního modelu.

```
function [F,PvalF,R2,R2adj,RCor,t,Pval,Sy,Se,St] = testing(y,yOdhad,beta,X)
```

5.5.1 Koeficient determinace

Začnu shrnutím základních regresních statistik o odhadnutém regresním modelu, mezi než patří koeficient vícenásobné korelace, determinace a jeho korigovaná varianta. Pro tyto výpočty užiji poznatky získané v kapitole 2.2.4. Nejdříve určím celkový, reziduální a vysvětlený (regresní) součet čtverců. Dále vypočítám koeficient determinace jako podíl vysvětlovaného a celkového součtu čtverců. Vícenásobnou korelaci určím jako jeho odmocninu. Vzhledem k tomu, že s koeficientem determinace je spojeno několik problémů (viz 2.2.4) spočítám také jeho korigovanou variantu. Pro tuto variantu je třeba zjistit počet stupňů volnosti pro jednotlivé části modelu. Jejich hodnoty určím následujícím způsobem:

- regresní část: $DFM = p - 1$,
- reziduální část: $DFE = n - p$,
- celkový součet: $DFT = n - 1$,

kde n odpovídá rozsahu výběrového souboru a p je počet odhadnutých koeficientů β. Korigovaný koeficient determinace odpovídá:

$$R^2adjust = 1 - \frac{(1 - R^2) * DFT}{DFE}, \tag{5.1}$$

kde R^2 je rovno jeho nekorigované variantě. Část zdrojového kódu viz 5.6.

```
%soucty ctvercu
Sy = sum((y-meanY).^2); %Celkovy
Se = sum((y-yOdhad).^2); %Rezidualni
St = sum((yOdhad-meanY).^2); %Vysvetleny
%koeficient determinace
R2 = St/Sy;
%korelace
RCor = sqrt(R);
%stupne volnosti
DFM = p-1;DFE = n-p;DFT = n-1;
%korigovany koeficient determinace
R2adj = 1-(((1-R2)*DFT)/DFE);
```

Zdrojový kód 5.6: Koeficient determinace

Jednotlivé vypočtené koeficienty pro model vytvořený nad testovacími daty z nemovitosti jsou shrnuty v tabulce 5.3.

R	R^2	R^2adjust
.88763	.78789	.78538

Tabulka 5.3: Regresní statistiky pro odhad ceny nemovitosti

5.5.2 Analýza rozptylu - ANOVA

Nyní následuje analýza rozptylu (**An**alysis **O**f **Va**riance). Pro výpočet použiji součty čtverců jednotlivých částí regresního modelu a jejich stupně volnosti určených při výpočtu korigovaného koeficientu determinace. Určím střední hodnotu jednotlivých součtu čtverců vydělením odpovídajícím stupněm volnosti. F-statistiku určím jako podíl středních hodnot regresní a reziduální části. Výsledky analýzy rozptylu pro model vytvořený nad testovacími daty jsou uvedeny v tabulce 5.4.

Model	Součet čtverců	Stupeň volnosti	F statistika
Regresní	$7.2548e+012$	17	
Reziduální	$1.9531e+012$	1442	315.071380
Celková	$9.2079e+012$	1459	

Tabulka 5.4: Analýza rozptylu

5.5.3 T-testy atributů

Další částí je testování odhadnutých regresních parametrů. Pro tento výpočet určím inverzní matici k $(X' * X)$ a dále rozptyl rozdílu původní a odhadnuté cílové proměnné. Dále v cyklu pro každý odhadnutý regresní parametr určuji hodnotu t-testu. Konkrétní výpočet udávám v rovnici 5.2. Nejčastější způsob formulace hodnoty t-testu je prostřednictvím p-hodnoty. Definujeme ji jako nejmenší hladinu významnosti testu, při níž na daných datech ještě zamítneme nulovou hypotézu [14]. Postup výpočtu obou hodnot zobrazuje zdrojový kód 5.7.

$$t(i) = \frac{\beta(i)}{\sqrt{rozpyl^2 * V[i,i]}} \tag{5.2}$$

```
%jednostranne t-testy atributu
V = pinv(X'*X);
t = ones(p,1);
rozdil = y-yOdhad;
rozptyl = std(rozdil);
for i = 1:p
 t(i) = beta(i)/sqrt(rozptyl^2*V(i,i));
end
%vypocet p-value
Pval=2*(1-tcdf(abs(t),DFT));
```

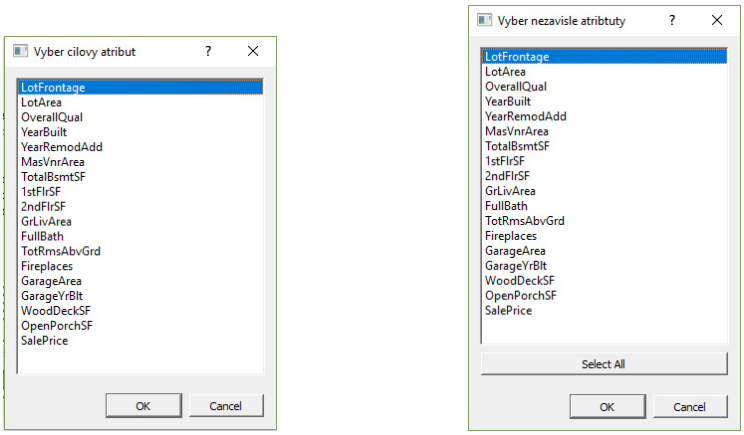
Zdrojový kód 5.7: T-test a p-hodnota

Tabulku regresních koeficientů s jejich statistikami pro jednotlivé atributy uvádím vzhledem k jejich obsáhlosti příloze D.

5.6 Ovládání programu

Program s uživatelem komunikuje formou dialogových oken. Prostředí Octave umožňuje použít několik grafických oken [7]. Ve svém programu užívám informační okna, dialogy s tlačítky a možností výběru více položek ze seznamu a další.

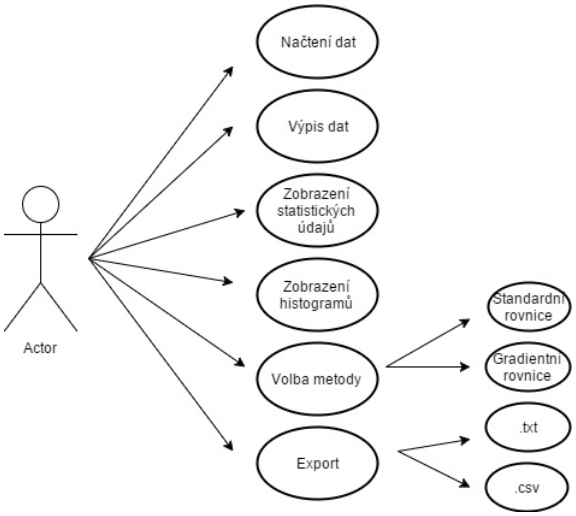
Vstupem do programu je soubor `main.m`. Po jeho spuštění je uživatel vyzván k zadání vstupního souboru. Také má možnost použít již před-připravená data z nemovitostmi, na kterých je prezentována funkčnost programu. Po načtení souboru je uživatel vyzván k ověření jejich korektnosti. V případě chyby program uživateli umožní opětovné načtení. Po korektním načtením dat je uživatel vyzván k výběru cílového atributu a nezávislých proměnných. Výběrové dialogy zobrazuje obrázek 5.6.



(a) Výběr cílového atributu (b) Výběr nezávislých atributů

Obrázek 5.6: Výběr atributů

Po výběru je uživateli umožněn náhled na statistické údaje ke všem vybraných atributů, zobrazení jejich histogramů a korelační závislost na cílovém atributu. V případě, že uživatel již nepotřebuje žádné další informace o použitých atributech následuje v programu krok odhadu parametrů regresního modelu a jeho celkové testování. Program umožňuje výpočet odhadu parametrů dvěma metodami. Volbu metody si uživatel zvolí v dialogovém okně. V případě volby *standardní rovnice* dojde k rovnou vytvoření regresního modelu včetně jejího testování. Pokud uživatel zvolí *gradientní metodu* je mu dána možnost volby jejího nastavení. Uživatel nastaví počet iterací a stupeň učení gradientní metody. Po zadání hodnot dojde k vypočtení odhadu a k vytvoření křivky průběhu učení. Ta je zobrazena uživateli a v případě, že uživatel potvrdí její korektnost dojde opět k otestování celého modelu. Pokud je průběh učení chybný (viz. kapitola 5.4.2) je uživateli umožněno opakovat výpočet odhadu parametrů s novými nastaveními. Celkové výsledky regresního modelu, které zahrnují odhady parametrů a statistiky o provedeném modelu, jsou uživateli vypsány do konzole programu Octave. Je umožněno také exportování výsledků do souboru typu *.txt* nebo *.csv*. Obrázek 5.7 zobrazuje grafický popis ovládání programu.



Obrázek 5.7: Grafický popis ovládání programu

5.7 Zhodnocení

Výsledný program umožňuje vytvoření vlastního lineárního regresního modelu. Může pracovat s libovolnou datovou sadou, kde bude sada atributů číselného typu a jeden z nich bude vhodnou cílovou proměnnou. Umožňuje uživateli z načtených atributů vybrat cílovou proměnnou a nezávislé atributy. Dále zobrazí statistické údaje o načtených datech. Z důvodu pohodlnosti ovládání program komunikuje s uživatelem formou dialogových oken. K odhadu parametrů je užita metoda nejmenších čtverců. Program ji umožňuje spočítat dvěma způsoby, standardní rovnice popsanou v kapitole 2.4.1 a gradientní metodou viz 5.4.2. Výstupem programu jsou odhadnuté regresní parametry a celkové testování modelu včetně testu nezávislých atributů. Výstup programu je možné exportovat do souboru typu *.txt* nebo *.csv*. Tento program je vhodný pro výuku lineární regresní analýzy a jejího celkového testování společně s testováním odhadnutých parametrů.

6 Srovnání

V této kapitole provedu srovnání lineárního regresního modelu vytvořené prostřednictvím aplikace Modeler s výstupem programu v Octave. Následně porovnávám časové rozdíly výpočtu. Srovnávání provedu na několika výběrovými soubory.

6.1 Srovnání výsledků modelů

6.1.1 Výběrový soubor - Nemovitosti

První srovnání provedu s výsledky mé případové studie, kde jsem odhadoval cenu nemovitosti na základě dalších známých skutečností. Výsledky prezentuji v několika výstupních tabulkách. V první tabulce 6.1 porovnávám výsledky regresní statistiky. Na řádku s označením Model č. 1 jsou výsledky podané programem Modeler, na řádku č. 2 jsou výsledky mého regresního modelu vytvořeného v Octave. Z uvedené tabulky vidíme, že oba modely podávají srovnatelné výsledky. Veškeré výsledky zaokrouhluji na 5 desetinných míst z důvodu prezentace přesnosti.

Model Summary			
Model	R	R Square	Adjusted R Square
1	.88763	.78789	.78538
2	.88763	.78789	.78538

Tabulka 6.1: Regresní statistika test č. 1

Další výstupní tabulkou 6.2 je analýza rozptylu (ANOVA - test). V tabulce je opět na první řádku výstup z aplikace Modeler a druhém řádku výsledky mého modelu. Určené hodnoty se opět srovnatelné.

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	$7.255e+12$	17	$4.268e+11$	315.071	.000
	Residual	$1.953e+12$	1442	$1.35e+09$		
	Total	$9.208e+12$	1459			
2	Regression	$7.2548e+12$	17	$4.2675e+11$	315.0713	.000
	Residual	$1.9531e+12$	1442	$1.3545e+09$		
	Total	$9.2079e+12$	1459			

Tabulka 6.2: Výsledky analýzy ANOVA test č. 1

V poslední tabulce 6.3 srovnávám jednotlivé odhady parametrů a jejich statistiky. V levé části tabulky vidíme výsledky podané programem Modeler a v pravé části výsledky mého provedené modelu. Opět vidíme, že se výsledky téměř neliší, až na některé výjimky, kde u druhého a třetího desetinného místa došlo k drobné odchylce. Tato odchylka mohla být způsobena zaokrouhlováním.

Coefficients								
Model	1				2			
Atribut	Unstand.	Stand.	t	Sig.	Unstand.	Stand.	t	Sig.
(Constant)	-1245621.962		-9.283	.000	-1245621.955		-9.338	.000
LotFrontage	13.953	.004	.274	.784	13.953	.004	.276	.783
LotArea	.528	.066	4.938	.000	.528	.066	4.967	.000
OverallQual	19220.507	.335	16.314	.000	19220.508	.335	16.410	.000
YearBuilt	225.045	.086	3.769	.000	225.045	.086	3.791	.000
YearRemodAdd	357.732	.093	5.523	.000	357.732	.093	5.556	.000
MasVnrArea	32.400	.074	5.228	.000	32.400	.074	5.259	.000
TotalBsmtSF	16.379	.090	3.891	.000	16.379	.091	3.914	.000
1stFlrSF	31.432	.153	1.471	.142	31.432	.153	1.479	.139
2ndFlrSF	22.992	.126	1.094	.274	22.992	.126	1.100	.272
GrLivArea	16.421	.109	.787	.431	16.421	.109	.792	.429
FullBath	-3726.450	-.026	-1.412	.158	-3726.450	-.026	-1.421	.156
TotRmsAbvGrd	1084.673	.022	.994	.320	1084.673	.022	1.000	.318
Fireplaces	8113.577	.066	4.472	.000	8113.577	.066	4.498	.000
GarageArea	38.015	.102	6.136	.000	38.015	.102	6.172	.000
GarageYrBlt	15.940	.005	.221	.825	15.940	.005	.223	.824
WoodDeckSF	32.132	.051	3.893	.000	32.132	.051	3.916	.000
OpenPorchSF	5.820	.005	.368	.713	5.820	.005	.370	.712

Tabulka 6.3: Odhady regresních parametrů test č. 1

6.1.2 Výběrový soubor - Výkon CPU

Další srovnání provedu na výběrovém souboru, kde odhaduji hodnotu relativního výkonu v závislosti na době cyklu, velikost paměti atd. Více informací o stanovení

relativní hodnoty výkonu podává publikace [2]. Data jsem získal z portálu *UCI*, který je zaměřen na výuku strojového učení[1]. Popis atributů uvádím v příloze E. Výběrový soubor obsahoval 209 záznamů. Tato data již byla předzpracovaná, a tudíž jsem na ně aplikoval pouze regresní analýzu. Z uvedených tabulek vidíme, že oba modely dosáhly opět srovnatelných výsledků. Jejich rozložení odpovídá testu č. 1.

Model Summary			
Model	R	R Square	Adjusted R Square
1	.930	.865	.861
2	.929995	.864891	.860878

Predictors: (Constant), CHMAX, MYCT, MMIN, CACH, CHMIN, MMAX

Tabulka 6.4: Regresní statistika test č. 2

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4653317.024	6	775552.837	215.514	.000
	Residual	726920.115	202	3598.614		
	Total	5380237.139	208			
2	Regression	4653317.023564	6	775552.837	215.514291	.000
	Residual	726920.115204	202	3598.614		
	Total	5380237.138756	208			

Dependent Variable: PRP
Predictors: (Constant), CHMAX, MYCT, MMIN, CACH, CHMIN, MMAX

Tabulka 6.5: Výsledky analýzy ANOVA test č. 2

Coefficients								
Model	1				2			
Atribut	Unstand.	Stand.	t	Sig.	Unstand.	Stand.	t	Sig.
(Constant)	-55.894		-6.948	.000	-55.8939		-7.0501	.000
MYCT	.049	.079	2.789	.006	.048855	.079059	2.829779	0.005114
MMIN	.015	.369	8.371	.000	.015293	.368810	8.494595	.000
MMAX	.006	.406	8.681	.000	.005571	.406224	8.808762	.000
CACH	.641	.162	4.596	.000	.641401	.162029	4.663581	.000006
CHMIN	-.270	-.011	-.316	.752	-.270358	-.011458	-.320618	.748822
CHMAX	1.482	.240	6.737	.000	1.482472	.239633	6.836666	.000

Dependent Variable: PRP

Tabulka 6.6: Odhady regresních parametrů test č. 2

6.2 Časové srovnání

V této části provedu časové srovnání budování regresního modelu mezi Modeler a program v Octave. Porovnávám jak čas dílčích částí procesu, tak i celkovou dobu. V Modeleru lze při spuštění celého proudu a libovolných jeho částí zaznamenávat čas jejich vykonání, ale sofistikovanější údaje o dobách výkonu jednotlivých částí případě spuštění celého proudu v něm nejsou k dispozici. Naproti tomu Octave poskytuje několik nástrojů pro časování a zjištění doby průběhu dílčích funkcí a celého procesu. Jedním z nich je nástroj *profile*, který lze aktivovat a deaktivovat v libovolném úseku programu. Jeho výstup lze zobrazit zapsáním příkazu *profshow* do příkazové konzole v prostředí Octave. Jeho výstup zobrazuje tabulka 6.7. Další možností časování je použití funkcí *tic()* a *toc()*[7].

Funkce	**Celkový čas(s)**	**Procesorový čas (s)**	**Počet provedení**
questdlg	8.18	.002	7
listdlg	3.751	.001	2
input	2.443	1.963	1
normalize	1.15	1.108	2
testing	.005	.003	1
normalEqn	.003	.002	2
fprintf	.003	.003	82

Tabulka 6.7: Čas výpočtu

První dva řádky tabulky 6.7 s označením *questdlg* a *listdlg* jsou dialogová okna, která čekají na vstup od uživatele. Dále funkce *input* reprezentuje dobu čtení. Modeler v tomto případě dosahoval přibližně času 0.12 sekundy. Tedy podobného času jako funkce *csvread()*, kterou jsem se rozhodl nepoužít z důvodů, které udávám v kapitole 5.2. Vlastní výpočet parametrů regresního modelu a jeho otestování dosáhlo v programu Octave procesorového času 0.008 (součet doby funkcí *normalEqn() a testing()*). Podle zaznamenaných údajů z Modeleru byl procesorový čas (*CPU time*) tvorby regresního modelu 0.008. Tedy oba dosáhly srovnatelného času při odhadu regresních parametrů a celkového testování modelu.

7 Závěr

Zpracováním diplomové práce jsem si prohloubil znalosti regresní analýzy a strojového učení, ale přiměřeně rozsahu diplomové práce, které díky tomu můžu aplikovat v data miningových úlohách. Podrobně jsem rozebral metody lineární a logistické regrese. Následně jsem vysvětlil proces ekonometrického modelování a na praktickém příkladě předvedl její uplatnění společně s aplikací lineárního regresního modelu. Téma ekonometrie a ekonometrické modelování jsem zpracoval jako výkladovou studii do kurzu Datamining na ALS portále.

V rámci přípravy na diplomovou práci a v předmětu Datamining jsem se seznámil z kurzy typy MOOC a sám jsem jeden absolvoval. Konkrétně kurz *Machine Learning* zaměřený na strojové učení. Šlo o kurz na serveru Coursera, který provozuje univerzita Stanford. Byla to má první zkušenost s MOOC kurzem takového typu. Kurz mě velice bavil a byl cenným informačním zdrojem nejen pro tuto práci, ale i pro další aplikace. Nemohu jinak než tuto formu výukových kurzů doporučit.

V praktické části jsem zpracoval případovou studii v programu IBM SPSS Modeler. Konkrétně odhad ceny nemovitosti na základě známých skutečností a aplikoval na ni lineární regresní model. Podle podaných výsledků model dosáhl dobré kvality. Dále jsem naprogramoval aplikaci v prostředí Octave, pomocí které je možno vytvořit vlastní lineární regresní model na libovolných datech společně s jeho otestováním a testováním jednotlivých atributů. Výstupy obou těchto řešení jsem porovnal z hlediska přesnosti výsledků a časové náročnosti. Porovnání proběhlo nad několika výběrovými soubory a bylo dosaženo srovnatelné přesnosti, jako v případě regresního modelu sestaveného v programu Modeler. Podle zaznamenaných časů stavby regresního modelu byla také obě řešení porovnatelná.

Z celkového hlediska jsou programy diametrálně odlišné, a tedy neporovnatelné. Program IBM SPSS Modeler je komplexní nástroj pro provedení celkové analýzy dat a umožňuje použití mnoho rozličných nástrojů pro práci s nimi. Naproti tomu můj program se zaměřuje na lineární regresní model a je vhodný pro použití tam, kde nejsou třeba tak rozsáhlé programy typu Modeler. Dalším důvodem může být potřeba licence, kterou program Modeler vyžaduje a s tím spojené i licenční poplatky.

Vytvořený program je vhodný pro výuku lineární regresní analýzy, protože celkový proces výpočtu lineární regrese včetně jeho testování je názorně popsán a zdokumentován. Program může pracovat s libovolnou datovou sadou, kde bude

sada atributů číselného typu a jeden z nich bude vhodnou cílovou proměnnou. Jediné omezení je ve formátu souboru, který je pro analytické a data miningové zadání typický. Bude použit pro výuku regresní analýzy v předmětu Datamining. Další výhodou vytvořeného programu je jeho funkčnost ve svobodném prostředí Octave, tudíž je vhodný pro studijní účely. V případě dalších rozšíření vytvořeného programu bych se zaměřil na logistický regresní model a jeho možnosti řešení a testování.

Literatura

[1] Aha, D.; Asuncion, A.; Newman, D.: UCI Machine Learning Repository[online]. Dostupné z: `http://archive.ics.uci.edu/ml/`.

[2] Aha, D. W.; Kibler, D. F.; Albert, M. K.: *Instance-based prediction of real-valued attributes*. 1989, 51 s.

[3] Berka, P.: *Dobývání znalostí z databází*. Praha: Academia, 2003, ISBN 80-200-1062-9, 336 s.

[4] Coursera: Free Online Courses From Top Universities [online]. Dostupné z: `https://www.coursera.org/`.

[5] CSU: Český statistický úřad [online]. Dostupné z: `https://www.czso.cz/`.

[6] Duda, R. O.; Hart, P. E.; Stork, D. G.: *Pattern classification,, 2 nd edition*. 2001 [cit. 2017-02-18], ISBN 111858600X, 680 s., dostupné z: `https://books.google.cz/books?isbn=111858600X`.

[7] Eaton, J. W.: *GNU Octave documentation [online]*. Dostupné z: `https://www.gnu.org/software/octave/doc/v4.0.3`.

[8] Goldbloom, A.: Kaggle - Your Home for Data Science[online]. Dostupné z: `https://www.kaggle.com/`.

[9] Hančlová, J.: *Ekonomerické modelování*. Praha: Professional Publishing, 2012, ISBN 978-80-7431-088-1, 214 s.

[10] Hendl, J.: *Přehled statistických metod*. Praha: 4 rozš. Vyd. Portál s.r.o., 2012 [cit. 2016-12-25], ISBN 978-80-262-0200-4, 736 s.

[11] Hindls, R.; Hronová, S.; Seger, J.: *Statistika pro ekonomy*. Praha: Professional Publishing, vyd. 7, 2004 [cit. 2017-03-15], ISBN 80-86419-59-2, 417 s.

[12] Horová, I.; Zelinka, J.: *Numerické metody [online]*. Brno: Masarykova univerzity, 2008 [cit. 2017-02-28], dostupné z: `https://www.math.muni.cz/~zelinka/dokumenty/numerika.pdf`.

[13] Keynes, J. M.: *The general theory of employment, interest and money*. Kissimmee,USA: Singnalman Publishing, 2009, ISBN 978-0-9840614-0-2, 264 s.

[14] Kohout, V.: *Teorie odhadu Kapitola 10*. ZČU Plzeň: Skriptum ZCU [online], 22.04.2004 [cit. 2017-02-14], dostupné z: http://www.kmt.zcu.cz/person/Kohout/info_soubory/letnisem/zs/stat10.pdf.

[15] Myung, I. J.: Tutorial on maximum likelihood estimation [online]. 2003-02 [cit. 2017-01-15], dostupné z: http://www.sciencedirect.com/science/article/pii/S0022249602000287.

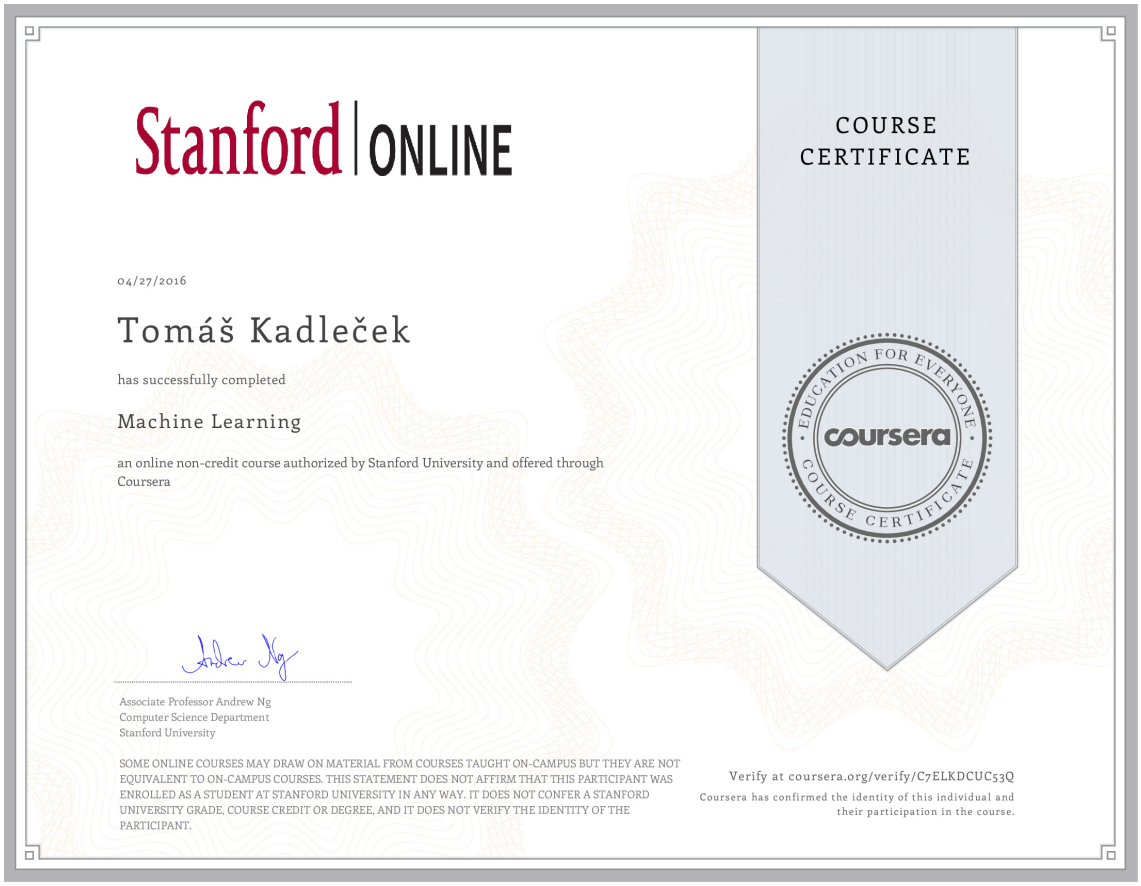
[16] Rud, O. P.: *Datamining*. Praha: Computer Press, 2006, ISBN 80-722-6577-6, 416 s.

[17] Yin, Y.; Kaku, I.; Tang, J.: *Data Mining,*. London Ltd: Springer, 2011.

A Obsah přiloženého CD

- diplomová_práce_2017_tomáš_kadleček.pdf
- Případová studie
  - Případová studie.pdf
  - PopisDat.pdf
  - PripadovaStudieModeler.str
  - train.csv
- Program v Octave
- MOOC kurz
  - Diplom.pdf
- Kurz Datamining
  - Jednoduchý lineární regresní model.pdf
  - Logistický regresní model.pdf
  - Vícerozměrný lineární regresní model.pdf
  - Ekonometrie.pdf
  - Případová studie.pdf
- Data test č. 2
  - data.csv
  - popis dat.pdf
- Obrázky použité v práci

B Certifikát o absolvování kurzu Machine learning



Obrázek B.1: Certifikát o absolvování kurzu Machine Learning

C Nemovitosti

Atribut	Popis
Id	Identifikátor nemovitosti
MSSubClass	Identifikuje třídu nemovitosti
MSZoning	Zóna nemovitosti
LotFrontage	Vzdálenost vzdušnou čarou od cesty k nemovitosti
LotArea	Celková rozloha nemovitosti ve čtverečních stopách
Street	Ulice ve které se nemovitost nalézá
Alley	Alej ve které se nemovitost nalézá
LotShape	Tvar nemovitosti
LandContour	Rovnost pozemku
Utilities	Připojena voda,elektřina,plyn
LotConfig	Umístěni nemovitosti
LandSlope	Sklon pozemku
Neighborhood	Sousedství
Condition1	Vzdálenost k objektům v okolí
Condition2	Vzdálenost k objektům v okolí 2
BldgType	Typ budovy - rodinný, bytovka atd.
HouseStyle	Typ budovy - počet pater
OverallQual	Celkové ohodnoceni materiálu
OverallCond	Ohodnoceni celkového stavu nemovitosti
YearBuilt	Rok výstavby
YearRemodAdd	Rok rekonstrukce
RoofStyle	Typ zastřešení
RoofMatl	Materiál použitý pro zastřešení
Exterior1st	Externí zastřešení nemovitosti
Exterior2nd	Externí zastřešení nemovitosti 2
MasVnrType	Typ použitého zdiva
MasVnrArea	Plocha zdi ve čtverečních stopách
ExterQual	Kvalita externího materiálů
ExterCond	Aktuální kvalita externích materiálů
Foundation	Materiál použitý pro základy
BsmtQual	Přibližná výška sklepa
BsmtCond	Hodnoceni kvality sklepa
BsmtExposure	Stav kvality sklepa
BsmtFinType1	Hodnoceni kvality dokončeného sklepa 1

Tabulka C.1: Popis všech atributů 1

Atribut	Popis
BsmtFinSF1	Plocha dokončeného sklepa 1 ve čtverečních stopách
BsmtFinType2	Hodnoceni kvality dokončeného sklepa 2
BsmtFinSF2	Plocha dokončeného sklepa 2 ve čtverečních stopách
BsmtUnfSF	Plocha nedokončeného sklepa ve čtverečních stopách
TotalBsmtSF	Celková plocha sklepa ve čtverečních stopách
Heating	Typ zatepleni
HeatingQC	Kvalita zatepleni
CentralAir	Klimatizace
Electrical	Elektrické vedení
1stFlrSF	Plocha prvního patra ve čtverečních stopách
2ndFlrSF	Plocha druhého patra ve čtverečních stopách
LowQualFinSF	Plocha oblasti špatného stavu nemovitosti ve čtverečních stopách
GrLivArea	Plocha přízemí ve čtverečních stopách
BsmtFullBath	Počet malých koupelen v suterénu
BsmtHalfBath	Počet koupelen v suterénu
FullBath	Celkový počet malých koupelen
HalfBath	Celkový počet koupelen
BedroomAbvGr	Počet ložnic
KitchenAbvGr	Počet kuchyní
KitchenQual	Kvalita kuchyně
TotRmsAbvGrd	Celkový počet místností bez koupelen
Functional	Funkcionalita nemovitostí
Fireplaces	Počet únikových východů
FireplaceQu	Kvalita únikových východů
GarageType	Typ garáže
GarageYrBlt	Rok výstavby garáže
GarageFinish	Dokončenost garáže
GarageCars	Počet automobilů, který se vejde do garáže
GarageArea	Celková plocha garáže ve čtverečních stopách
GarageQual	Kvalita garáže
GarageCond	Aktuální stav garáže
PavedDrive	Příjezdová cesta
WoodDeckSF	Plocha dřevěné terasy
OpenPorchSF	Plocha otevřené verandy
EnclosedPorch	Plocha uzavřené verandy
3SsnPorch	Plocha zimní zahrady
ScreenPorch	Plocha balkonu
PoolArea	Plocha bazénu
PoolQC	Kvalita bazénu
Fence	Kvalita oplocení
MiscFeature	Nadstandardní vybavení
MiscVal	Hodnota nadstandardního vybavení
MoSold	Měsíc prodeje
YrSold	Rok prodeje
SaleType	Typ prodeje
SaleCondition	Podmínky prodeje
SalePrice	Prodejní cena

Tabulka C.2: Popis všech atributů 2

D Vypočtené koeficienty a jejich statistiky

Atribut	Koeficient	T statistika	Sign. t
Konstanta	-1245621.9551	-9.3375	.0000
LotFrontage	13.9533	0.2755	.7830
LotArea	.5284	4.9670	.0000
OverallQual	19220.5075	16.4099	.0000
YearBuilt	225.0445	3.7908	.0002
YearRemodAdd	357.7320	5.5558	.0000
MasVnrArea	32.4003	5.2588	.0000
TotalBsmtSF	16.3792	3.9140	.0001
1stFlrSF	31.4315	1.4793	.1393
2ndFlrSF	22.9916	1.1001	.2715
GrLivArea	16.4208	.7916	.4287
FullBath	-3726.4499	-1.4207	.1556
TotRmsAbvGrd	1084.6726	1.0000	.3175
Fireplaces	8113.5768	4.4984	.0000
GarageArea	38.0152	6.1719	.0000
GarageYrBlt	15.9399	.2225	.8239
WoodDeckSF	32.1318	3.9162	.0001
OpenPorchSF	5.8197	.3698	.7116

Tabulka D.1: Odhady regresních parametrů

E Popis atributů testu č. 2

1. Název prodejce: 30 (adviser, amdahl,apollo, basf, bti, burroughs, c.r.d, cambex, cdc, dec, dg, formation, four-phase, gould, honeywell, hp, ibm, ipl, magnuson, microdata, nas, ncr, nixdorf, perkin-elmer, prime, siemens, sperry, sratus, wang)
2. Název modelu: mnoho jedinečných symbolů
3. MYCT: takt stroje v nanosekund (celé číslo)
4. Mmin: minimální hlavní paměť v kB (celé číslo)
5. Mmax: maximální hlavní paměť v KB (celé číslo)
6. CACH: vyrovnávací paměť v kB (celé číslo)
7. CHMIN: minimální kanály (celé číslo)
8. CHMAX: maximální kanály (celé číslo)
9. PRP: publikovaná relativní výkonnost (celé číslo)
10. ERP: odhadnutý relativní výkon viz publikace [2] (celé číslo)