Univerzita Hradec Králové

Přírodovědecká fakulta

Katedra matematiky

Užití regresní analýzy ve financích

Bakalářská práce

Autor: Denisa Sehnoutková

Studijní obor: Finanční a pojistná matematika

Vedoucí práce: RNDr. Michal Čihák, Ph.D.

Odborný konzultant: Mgr. Jitka Kühnová, Ph.D.

Hradec Králové květen 2016

Prohlášení:

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně a s použitím uvedené literatury.

V Hradci Králové dne Denisa Sehnoutková

Poděkování

Chtěla bych poděkovat RNDr. Michalu Čihákovi, Ph.D. za odborné vedení bakalářské práce, rady a čas, které mi věnoval.

Anotace

SEHNOUTKOVÁ, D. *Užití regresní analýzy ve financích*. Hradec Králové, 2016. Bakalářská práce na Přírodovědecké fakultě Univerzity Hradec Králové. Vedoucí bakalářské práce RNDr. Michal Čihák, Ph.D.

Regresní analýza patří mezi základní a nejčastěji používané statistické metody. Cílem nabízené práce je ukázat některé její konkrétní aplikace z oblasti financí. Součástí práce by měl být i podrobný popis použitých metod regresní analýzy (neměla by chybět některá vícerozměrná metoda), a to jak z hlediska jejich matematické podstaty, tak i z hlediska metodologie jejich použití. V neposlední řadě by měl být zvolen vhodný statistický software pro provedení regresní analýzy a otestovány jeho možnosti v této oblasti.

Klíčová slova: lineární regrese, rezidua, koeficient determinace, finance

Annotation

SEHNOUTKOVÁ, D. *Using regression analysis in finance.* Hradec Králové, 2016. Bachelor Thesis at Faculty of Science University of Hradec Králové. Thesis Supervisor RNDr. Michal Čihák, Ph.D.

Regression analysis is one of the basic and most frequently used statistical methods. The aim of the thesis is to show its concrete applications in finance. Detailed description of regression methods will be included in this work (multivariate methods should not be missed) both in terms of their mathematical nature and in terms of their usage. Finally appropriate statistical software for regression analysis should be chosen and tested.

Keywords: linear regression, residuals, coefficient of determination, finance

Obsah

Úvod

Regresní analýza je jednou z nejpoužívanějších a nejznámějších statistických metod, uvádí se, že dokonce naprostá většina aplikací je nějakou formou regresních metod. Nalézá široké uplatnění v mnoha oblastech běžného života, nejčastěji ve společenských vědách. Počátky regresní analýzy jsou dokumentovány již z let 1877 až 1885, kdy se antropolog a meteorolog Francis Galton zabýval vztahem mezi výškou otců a jejich prvorozených synů.

Hlavním cílem této práce je seznámení s lineární regresní analýzou a jejími aplikacemi na vybraná data z oblasti financí. Tato práce je rozdělena do čtyř kapitol.

V první kapitole se seznámíme s cíli regresní analýzy a obecným postupem při regresní analýze. Dále se budeme věnovat jednoduché lineární regresi, kde si vysvětlíme princip metody nejmenších čtverců a také definujeme jednoduchý lineární model. Budeme se také zabývat mnohonásobnou lineární regresí a určením mnohonásobného regresního modelu.

Jakmile určíme regresní model, velice důležitá součást regresní analýzy je zhodnocení kvality modelu. Tímto se budeme zabývat ve druhé kapitole, kde se seznámíme s různými statistickými nástroji, které určitým způsobem dovolují hodnotit využitelnost regresního modelu.

Ve třetí kapitole se zabývám regresní diagnostikou, kde si připomeneme předpoklady regresního modelu a podíváme se blíže na rezidua, která slouží k hodnocení kvality modelu, ale i k posuzování splnění předpokladů zvoleného regresního modelu.

V poslední kapitole tyto teoretické dovednosti využijeme pro aplikaci na konkrétní data z oblasti financí. Ke zpracování těchto dat použijeme R software, základní informace o R jsou také uvedeny v této kapitole.

1 Regresní analýza

Tato kapitola zabývající se regresní analýzou je zpracována pomocí známých knižních zdrojů, jsou zde využity zdroje [1], [4], [6] a internetový zdroj [9].

Pojem regrese se používá ve statistice ke zkoumání vztahu mezi jednou náhodnou veličinou na jedné straně (závisle proměnná) a jednou nebo více náhodnými veličinami na straně druhé (nezávisle proměnné, regresory), tedy zkoumání závislosti vysvětlované proměnné na proměnných vysvětlujících. Hlavním úkolem je najít „idealizující" matematickou funkci tak, aby co nejlépe vyjadřovala charakter závislosti. Tato matematická funkce se nazývá regresní funkce.

Úkolem regresní analýzy je poznání vztahů mezi statistickými znaky. Východiskem k popisu statistických závislostí jsou statistické údaje. Statistický soubor n pozorování sledovaných statistických znaků můžeme získat různými způsoby:

- údaje jsou získány pozorováním n statistických jednotek, přičemž statistický soubor byl prostorově, časově i věcně vymezen,
- údaje jsou získány pozorováním určité statistické jednotky v n časových okamžicích či intervalech,
- pozorování vznikla n-násobným opakováním určitého pokusu prováděného za stejných nebo přibližně stejných podmínek.

S uvedeným hlavním úkolem regresní analýzy souvisí řada dílčích úkolů. Některé z nich jsou:

- shromáždit a matematicky formulovat apriorní představy o charakteru regresní funkce,
- formulovat naše představy (předpoklady) o souhrnném působení neuvažovaných statistických znaků,
- odhadnout empirickou regresní funkci na základě statistických pozorování,
- posoudit kvalitu empirické regresní funkce z hlediska důvodů a cílů statistického zjišťování.

1.1 Postup

Regresní analýza má několik kroků

1. Jako první krok je nutno určit, jakou úlohu máme řešit.
2. Na této úvaze závisí další důležité kroky, jako jsou především výběr proměnných a specifikace modelu.
3. Výběr potenciálně relevantních proměnných a sběr dat

V této fázi je třeba vybrat nezávislé proměnné, které by mohly mít vliv na závislou proměnnou. Do výběru je lepší zahrnout i proměnné, o jejichž relevanci si nejsme jisti. Regresní analýza nám umožňuje tuto relevanci testovat a popřípadě je tak můžeme z modelu vyřadit. Pro každou z n statistických jednotek v naší úloze získáme $k + 1$ hodnot pro k nezávisle proměnných a jednu závisle proměnnou. Uspořádáním dat pro nezávisle

proměnné po řádcích (co řádek, to jedno pozorování) do matice o n řádcích a k sloupcích získáme tzv. matici plánu, která se značí $\boldsymbol{X}$.

4. Specifikace modelu

Tato část je velmi obtížná a často rozhodující, neboť nevhodně zvolený model může vést k zavádějícím výsledkům. Zde je často potřeba spolupráce statistika s odborníkem v oboru, jenž má znalosti o vztazích mezi veličinami. Modely je možné zpětně validovat a lze i porovnávat dva různé modely. Tím dospějeme k takovému modelu, jenž nejlépe popisuje situaci.

5. Odhad regresních parametrů

Následuje samotný výpočet regresních koeficientů $\beta_1, \beta_2, \ldots, \beta_n$. Nejčastěji se užívá tzv. *metoda nejmenších čtverců.* Za jistých předpokladů, jež je nutné znát a ověřit, tato metoda dává spolehlivé odhady regresních koeficientů.

6. Zhodnocení kvality modelu a užití modelu k řešení úlohy

Po úspěšném zjištění přesné podoby regresní rovnice lze pro každé pozorování vyjádřit takzvané *reziduum,* tedy rozdíl skutečné hodnoty závislé proměnné a výsledku vypočteného na základě modelu. Za pomoci reziduí i za pomoci *reziduálního součtu čtverců (SSE)* můžeme odhadnout, zda je model správně sestaven, a porovnat ho s jinými. A samozřejmě nejlepší kontrolou jsou reálná data, která (ne)korespondují s výsledky modelu. Je-li model shledán dostatečně kvalitním, můžeme jej užít k řešení úlohy.

1.2 Jednoduchá lineární regrese

Jednoduchá lineární regrese je statistická metoda, která nám umožňuje shrnout a studovat vztahy mezi dvěma souvislými (kvantitativními) proměnnými. Jedna proměnná, označujeme ji x, je považována za nezávisle proměnnou neboli vysvětlující proměnnou. Druhá proměnná, označená y, je výsledek reakce (závisle proměnná). Jednoduchá lineární regrese, se nazývá jednoduchá, protože studium se týká pouze jedné nezávisle proměnné x.

Nejjednodušší a nejčastěji používaná je přímková regrese, měli bychom charakterizovat její rovnici, která má tvar

$$y_i = \beta_0 + \beta_1 x_i$$

Máme tedy data a vzájemnou závislost chceme vystihnout pomocí přímky. Nastává tedy otázka, jak nejlépe přímku daty proložit. Tedy jak určit parametry β_0 a β_1. Přímka nemůže procházet všemi body zároveň, chtěli bychom však takové proložení, které nějakým způsobem nejlépe vystihuje průběh dat. Je třeba určit souhrnnou míru, která bude charakterizovat, jak moc je přímka blízko bodům. Nejpoužívanější je metoda nejmenších čtverců.

1.2.1 Metoda nejmenších čtverců

Cílem metody je najít vyrovnávací (odhadovanou) přímku, jejíž rovnice má tvar

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

kde x_i označuje předpokládanou hodnotu pozorovaného objektu a $\hat{y}_i$ je předpokládaný výsledek pozorovaného objektu. Parametry β_0 a β_1 musíme zvolit tak, abychom získali co nejméně rozptýlený soubor odchylek e_i, tzv. reziduálních chyb, metoda spočívá v minimalizaci součtů kvadrátů reziduí.

$$e_i = y_i - \hat{y}_i$$

Součet čtverců reziduí:

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_1))^2$$

Součet čtverců Q je funkcí neznámých parametrů. Pro určení jeho minima je nutné vypočítat první parciální derivace podle β_0 a β_1 a ty potom položit rovny nule. Nahradíme-li β_j jejich odhady b_j, dostaneme

$$\frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial Q}{\partial b_1} = -2 \sum_{i=1}^{n} [(y_i - b_0 - b_1 x_i) \cdot x_i] = 0$$

Danou soustavu upravíme na tvar

$$\sum_{i=1}^{n} y_i - n b_0 - b_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

Pro odhady parametrů b_0 a b_1 platí

$$b_0 = \frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Tatáž úloha, ale v maticovém zápisu. Hledáme přímku $y = b_0 + b_1 x$. Až ji najdeme, bude v bodech x_i platit

$$\hat{y}_0 = b_0 + b_1 x_0$$

$$\hat{y}_1 = b_0 + b_1 x_1$$

$$\vdots$$

$$\hat{y}_n = b_0 + b_1 x_n$$

Kde $\hat{y}_i$ jsou funkční hodnoty vypočtené dosazením x_i do rovnice přímky. Maticově můžeme psát

$$\begin{pmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

Což označením vektorů a matice je

$$\hat{\boldsymbol{y}} = \boldsymbol{X} \cdot \boldsymbol{b}$$

Je možné ukázat, že normálním rovnicím pro určení koeficientů b_0 a b_1 odpovídá maticový zápis

$$\boldsymbol{X}^T \cdot \boldsymbol{y} = \boldsymbol{X}^T \cdot \boldsymbol{X} \cdot \boldsymbol{b}$$

kde

$$\boldsymbol{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

jsou původní naměřené *y*-ové souřadnice zadávaných bodů. Odtud lze maticově psát přímo vzorec pro výpočet hledaných koeficientů b_0 a b_1:

$$\hat{\boldsymbol{b}} = (\hat{b}_0, \hat{b}_1)^T = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

1.2.2 Jednoduchý lineární regresní model

Model má podobu $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, kde $i = 1, \ldots, n$

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

v maticovém zápisu

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

kde

- y_i jsou hodnoty závisle proměnná,
- x_i jsou hodnoty nezávisle proměnné (regresoru),
- β_0, β_1 jsou neznámé parametry (které se snažíme odhadnout),
- ε_i je náhodná složka (rezidua).

Stanovme si předpoklady ohledně rozdělení y:

1. Linearita: Pro každou náhodnou veličinu y_i platí, že její střední hodnota leží na skutečné regresní přímce.
2. Homogenní rozptyl: Všechna y_i mají stejný rozptyl.
3. Nezávislost: Náhodné veličiny y_i jsou navzájem nezávislé.
4. Normalita: Náhodné veličiny y_i mají pro $i = 1, 2, \ldots, n$ normální rozdělení.

V případě regresního modelu ve tvaru:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

by měly být navíc splněny následující podmínky:

- $E(\varepsilon_i) = 0$ pro každé $i = 1, 2, \ldots, n$. To znamená, že náhodné složky mají ve všech výběrech nulovou střední hodnotou.
- $D(\varepsilon_i) = \sigma^2$ pro každé $i = 1, 2, \ldots, n$. Rozptyl náhodné složky je konstantní.
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ pro každé $i \neq j$, kde $i, j = 1, 2, \ldots, n$. Kovariance náhodné složky je nulová. Tedy hodnoty náhodné složky jsou nekorelované.
- Normalita, náhodné složky ε_i mají pro $i = 1, 2, \ldots, n$ normální rozdělení.
- Regresní parametry β_i mohou nabývat libovolných hodnot.
- Regresní model je lineární v parametrech.

1.3 Vícenásobná lineární regrese

Posuneme se z jednoduché lineární regrese s jednou nezávisle proměnnou na mnohonásobnou lineární regresi s dvěma a více nezávisle proměnnými. V mnohonásobné regresi může být potencionálně velké množství nezávisle proměnných, je tedy efektivní použít matice k definování regresního modelu.

Volba vhodného typu vícenásobné regresní funkce je obtížná. Možnost zachycení grafického průběhu závislosti i logického posouzení vhodného typu regresní funkce zde odpadá. Při hledání tohoto vhodného typu se opíráme o matematicko-statistická kritéria (různé testy, směrodatné chyby regresních koeficientů apod.), která nám obvykle umožní najít ten nejvhodnější.

Jestliže je závisle proměnná y lineárně závislá na každé z vysvětlujících proměnných, které jsou vzájemně nezávislé, používáme pro určení vývoje závisle proměnné mnohonásobnou lineární funkci proměnných $x_1, x_2, \ldots, x_k$.

1.3.1 Vícenásobný lineární regresní model

Pro každou zvolenou kombinaci hodnot nenáhodných vysvětlujících proměnných $x_1, x_2, \ldots, x_k$ je y náhodnou veličinou s určitým pravděpodobnostním rozdělením, konečnou

střední hodnotou $E(y)$ a s konečným rozptylem $D(y)$. Pro n kombinací hodnot nenáhodných vysvětlujících proměnných se předpoklad zcela lineárního modelu zapisuje jako n lineárních rovnic ve tvaru

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n$$

tedy ve tvaru

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

kde $i = 1, 2, \ldots, n$.

Rovnice můžeme zapsat maticově

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \text{ a } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

kde

- $\boldsymbol{y}$ je n-členný náhodný vektor napozorovaných hodnot vysvětlované proměnné y,
- $\boldsymbol{X}$ – nenáhodná matice typu n x (k+1) zvolených n kombinací hodnot vysvětlujících proměnných,
- $\boldsymbol{\beta}$ – je (k+1)-členný vektor neznámých parametrů modelu,
- $\boldsymbol{\varepsilon}$ – n-členný vektor nepozorovatelné rušivé (náhodné) složky.

Z uvedeného zápisu je vidět, že v n lineárních rovnicích je $p = k+1$ neznámých regresních parametrů a n hodnot náhodné složky.

Z předchozí rovnice dostaneme pro vektor podmíněných středních hodnot

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X\beta} + E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{X\beta}$$

a pro kovarianční matici

$$cov(\boldsymbol{y}|\boldsymbol{X}) = E\{[\boldsymbol{y} - E(\boldsymbol{y}|\boldsymbol{X})][\boldsymbol{y} - E(\boldsymbol{y}|\boldsymbol{X})]^T|\boldsymbol{X}\} = E\{[\boldsymbol{y} - \boldsymbol{X\beta}][\boldsymbol{y} - \boldsymbol{X\beta}]^T|\boldsymbol{X}\}$$
$$= E\{(\boldsymbol{\varepsilon}|\boldsymbol{X})(\boldsymbol{\varepsilon}|\boldsymbol{X})^T|\boldsymbol{X}\} = \sigma^2\boldsymbol{I}$$

Neznámé parametry β lze odhadnou metodou nejmenších čtverců, tj. nalezení takového vektoru $\boldsymbol{b}$, pro který je nejmenší reziduální součet čtverců (SSE) odchylek pozorovaných hodnot od jejich odhadů z modelu

$$SSE = (\boldsymbol{y} - \boldsymbol{Xb})^T(\boldsymbol{y} - \boldsymbol{Xb})$$

Položíme-li derivaci tohoto výrazu podle vektoru $\boldsymbol{b}$, tj.

$$\frac{\partial SSE}{\partial \boldsymbol{b}} = \frac{\partial}{\partial \boldsymbol{b}}(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}) = \frac{\partial}{\partial \boldsymbol{b}}(-2\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b})$$
$$= -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}$$

rovnu nulovému vektoru, dostaneme soustavu normálních rovnic

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}$$

Za předpokladu, že k matici $\boldsymbol{X}^T\boldsymbol{X}$ existuje matice inverzní, dostaneme vektor odhadovaných parametrů podle vztahu

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Takto určené odhady se nazývají OLS-odhady (Ordinary Least Squares). Tyto odhady jsou nestranné, neboť

$$E(\boldsymbol{b}) = E(\boldsymbol{b}|\boldsymbol{X}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{y}|\boldsymbol{X}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

Vektor $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}$ je lineární kombinací vektorů regresorů, tj. leží v prostoru (přímce, rovině, nadrovině), jehož dimenze je rovna počtu regresorů. Dosadíme-li za **b**, dostaneme

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

Matice $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ je matice projekce vektoru $\boldsymbol{y}$ do prostoru určeného vektory regresorů. Požadavek formulovaný v metodě nejmenších čtverců vlastně znamená, že tato projekce je ortogonální. Pak tedy vektory $\hat{\boldsymbol{y}}$ a $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ jsou ortogonální vektory, tzn. $\hat{\boldsymbol{y}}^T\boldsymbol{e} = 0$, o čemž se můžeme přesvědčit:

$$(\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = (\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}) = \boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}) = 0$$

neboť výraz v závorce je nulový vektor.

Příklad 1.1 Kvadratická regrese

Model má podobu $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon$, v maticovém zápisu

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

I v tomto případě jde o lineární regresní model. Rozhodující je linearita vzhledem k regresním parametrům.

Příklad 1.2 Vícenásobná regrese

Model může mít následující podobu $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$,

kde $x_{i3} = x_{i1} \,.\, x_{i2}$. To lze zapsat maticově následujícím způsobem

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Příklad 1.3 Dva nezávislé výběry

Výsledky měření u první skupiny lze vyjádřit rovnicí $y_{1i} = \mu_1 + \varepsilon_{1i}$ a u druhé skupiny $y_{2j} = \mu_2 + \varepsilon_{2j}$. Měření pro obě skupiny můžeme popsat souhrnným modelem

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \end{pmatrix}$$

2 Hodnocení kvality modelu

Ke zpracování této kapitoly jsem využila zdroje [4], [5], [6] a [13]. Pro zhodnocení kvality modelu je vždy rozhodujícím kritériem cíl analýzy, a tím i použitelnost výsledků. Vážné důsledky má volba špatného typu regresního modelu, nedostatky použitých statistických dat, ale i výběr nevyhovující metody odhadu parametrů a neoprávněnost některých předpokladů a podmínek. V dalších částech této kapitoly bude pozornost zaměřena na některé statistické nástroje, které určitým způsobem dovolují hodnotit využitelnost regresního modelu.

2.1 Součty čtverců

K hodnocení míry variability náhodné veličiny y slouží následující charakteristiky. Celkovou variabilitu náhodné veličiny y charakterizuje celkový součet čtverců. Celkový součet čtverců $SSTO$ je možné vyjádřit jako součet objasněné části rozptylu SSR (vysvětleného, nebo také regresního součtu čtverců) a neobjasněné části rozptylu SSE (reziduálního součtu čtverců).

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Maticově

$$SSTO = (\boldsymbol{y} - \overline{\boldsymbol{y}})^T (\boldsymbol{y} - \overline{\boldsymbol{y}})$$

$$SSR = (\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}})^T (\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}})$$

$$SSE = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$$

Platí:

$$SSTO = SSR + SSE$$

Z rozkladu celkového součtu čtverců vychází i analýza rozptylu, která je obvyklou součástí statistických softwarů. Tabulka analýzy rozptylu neboli ANOVA tabulka má většinou tento formát:

Zdroj rozptýlenosti	Součet čtverců	Stupně volnosti	Průměrný čtverec	Testová statistika F
Model	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$f_R = k - 1$	$MSR = \frac{SSR}{k}$	
Náhodná složka	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$f_E = n - k$	$MSE = \frac{SSE}{n-k-1}$	$F = \frac{MSR}{MSE}$
Celkový	$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$F_{TO} = n - 1$		

Tabulka č. 1: Tabulka ANOVA, vlastní zpracování [13]

2.2 Koeficient determinace

O kvalitě použitého modelu jako celku (tedy jaká část rozptylu vysvětlované proměnné je vysvětlena modelem) vypovídá koeficient determinace R^2, který má tvar:

$$R^2 = \frac{SSR}{SSTO} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Suma ve jmenovateli představuje celkový součet čtverců, který získáme součtem reziduálního a teoretického součtu čtverců. Koeficient determinace pak můžeme interpretovat jako poměr mezi regresním a celkovým součtem čtverců. Tento koeficient ukazuje, jakou část variability závisle proměnné se pomocí uvažované závislosti podařilo vysvětlit variabilitou nezávisle proměnných. Tento koeficient se často vyjadřuje v procentech.

Hodnota koeficientu determinace se pohybuje v intervalu $0 \leq R^2 \leq 1$. Nízká hodnota R^2, nemusí znamenat nízký stupeň závislosti mezi jednotlivými proměnnými, ale většinou signalizuje chybný typ regresní funkce.

Přesnost regresního modelu popisuje také upravený koeficient determinace (*adjusted* R^2), který je dán vzorcem

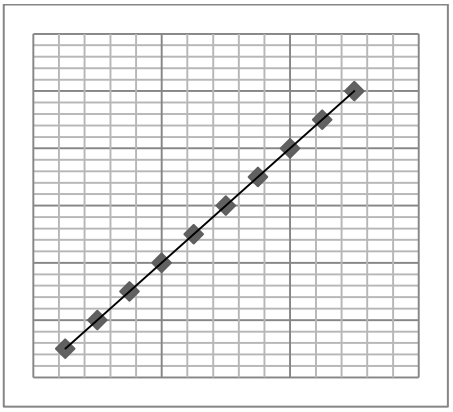
$$R^2_{adj} = 1 - (1 - R^2) \frac{n-1}{n-k}$$

Tento koeficient zohledňuje počet odhadovaných parametrů modelu vzhledem k počtu měření. Upravený koeficient determinace je vždy menší nebo roven R^2 a může být i záporný.

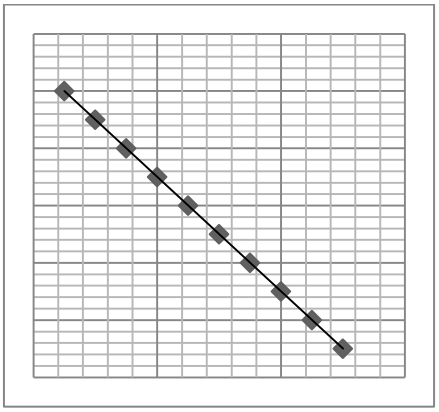
2.3 Pearsonův korelační koeficient

Pomocí korelačního koeficientu zjišťujeme míru lineárního vztahu dvou nenáhodných spojitých proměnných *x* a *y*. Lineární závislost dvou statistických jednotek lze zjistit vynesením proměnných do grafu. V případě korelace si můžeme přímku představit jako vyjádření lineárního vztahu a z odchylek bodů od přímky pak odhadnout míru tohoto vztahu.

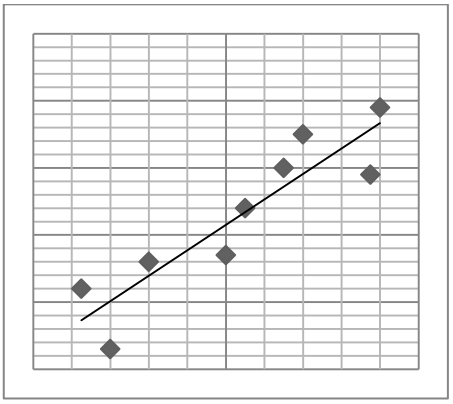
Korelační koeficient R je přímo spojený s koeficientem determinace R^2. Nabývá hodnot od $-1 \leq R \leq 1$, krajní hodnoty značí perfektní lineární vztah (záporný nebo kladný). Pokud se korelační koeficient rovná nule, jedná se o nelineární vztah.



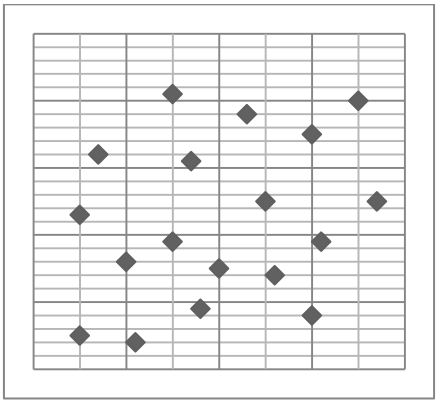
Graf č. 1: R = 1, vlastní zpracování [12]



Graf č. 3: R = −1, vlastní zpracování [12]



Graf č. 2: R = 0,8, vlastní zpracování [12]



Graf č. 4: R = 0, vlastní zpracování [12]

Výpočet korelačního koeficientu:

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

Součty čtverců ve jmenovateli jsou $n - 1$ násobkem výběrových rozptylů. Proto se také setkáváme s jednodušším vyjádřením:

$$R = \frac{s_{xy}}{s_x s_y}$$

Kde s_x je výběrová směrodatná odchylka proměnné x, s_y směrodatná odchylka proměnné y a s_{xy} kovariance proměnných x a y.

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

2.4 Dílčí t-testy

Dílčí t-testy jsou testy o hodnotách jednotlivých parametrů regresní funkce a umožňují nám testovat oprávněnost setrvání příslušné funkce vysvětlující proměnné v regresním modelu. Používají se tedy pro zjištění statistické významnosti jednotlivých regresních koeficientů.

Na hladině významnosti α testujeme nulovou hypotézu $H_0: \beta_j = 0$, proti alternativní hypotéze $H_1: \beta_j \neq 0$. To znamená, že se ptáme, zda j-tý prediktor v modelu nemá významný vliv na hodnotu závisle proměnné. Testová statistika

$$t = \frac{b_j}{s(b_j)}$$

kde $s(b_j)$ je odhad směrodatné chyby odhadu parametru β_j, se řídí Studentovým rozložením $t = (n - k - 1)$, je-li H_0 pravdivá. Nulovou hypotézu zamítáme na hladině významnosti α, platí-li $|t| > t_{1-\alpha/2}$.

2.5 Celkový F-test

Hodnota statistiky *F*, tedy společná statistická významnost všech koeficientů dohromady, je určující pro významnost jednotlivých koeficientů. Nejprve bychom se měli zajímat o hodnotu *F*-testu, a pokud naznačuje významnost regresních koeficientů jako sady, teprve pak kontrolovat významnost jednotlivých koeficientů.

Celkovým F-testem testujeme, zda vysvětlovaná proměnná je lineární kombinací vybraných funkcí vysvětlujících proměnných. Na hladině významnosti α testujeme nulovou hypotézu H_0, která tvrdí, že všechny parametry β_j (bez β_0) modelu jsou rovny nule, tj. $\beta_j = 0, j = 1, 2, \ldots, k$, oproti alternativní hypotéze H_1, která tvrdí, že aspoň pro jedno *j* je $\beta_j \neq 0$. Statistika má tvar

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k}}{\frac{\sum_{i=1}^{n}(y - \hat{y}_i)^2}{n-k-1}}$$

kde n je počet pozorování, k je počet prediktorů v modelu, řídí se Fisherovým-Snedecorovým rozložením $F = (k, n - k - 1)$, je-li H_0 pravdivá. Nulovou hypotézu zamítáme na hladině významnosti α, platí-li $F > F_{1-\alpha}$.

3 Regresní diagnostika

Tato kapitola byla zpracována pomocí zdrojů [4], [5], [9] a [11]. Cílem regresní diagnostiky je ověřit, zda navržený model koresponduje s reálnými daty a zda použití metody nejmenších čtverců bylo vhodné. Metoda nejmenších čtverců poskytuje optimální výsledky jen při splnění předpokladů o datech a regresním modelu. Regresní diagnostika obsahuje postupy k posouzení předpokladů metody nejmenších čtverců, kvality modelu a k posouzení kvality dat pro navržený model. Rozdíl mezi regresní diagnostikou a klasickými testy spočívá v tom, že u regresní diagnostiky není potřeba formulovat alternativní hypotézu a jsou odhaleny typy odchylek od ideální situace.

Regresní diagnostika vyšetřuje regresní triplet, který se zaměřujeme na následující oblasti:

- kritika a analýza vstupních dat (matice $\boldsymbol{X}$ a vektoru $\boldsymbol{y}$),
- kritika a analýza modelu jako celku,
- kritika a analýza metody odhadů, resp. zhodnocení splnění předpokladů použité metod nejmenších čtverců.

Připomeňme předpoklady lineárního regresního modelu:

- jednotlivé složky vektoru reziduí jsou nekorelované,
- reziduální složka má normální rozdělení,
- střední hodnota reziduální složky se rovná nule,
- reziduální složka má konstantní rozptyl (pro každé pozorování má příslušná složka vektoru ε_i stejný rozptyl),
- správně specifikovaný model, tedy rovnice modelu je vybrána správně.

3.1 Rezidua a jejich vlastnosti

Rezidua jsou základním diagnostickým nástrojem, a to nejen při hodnocení kvality regresní funkce a dat, ale i obecněji při posuzování předpokladů zvoleného regresního modelu. Lze říci, že jakákoli systematičnost (nenáhodnost) zjištěná u reziduí značí nějaký nedostatek. Vektor reziduí vyjádříme pomocí projekční matice $\boldsymbol{H}$:

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{I}\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

Pak kovarianční matice reziduí je

$$\begin{aligned} cov(\boldsymbol{\varepsilon}) = cov[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}] = (\boldsymbol{I} - \boldsymbol{H})cov(\boldsymbol{y})(\boldsymbol{I} - \boldsymbol{H})^T = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\sigma^2}\boldsymbol{I}(\boldsymbol{I} - \boldsymbol{H})^T \\ = \boldsymbol{\sigma^2}(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^T = \boldsymbol{\sigma}^2(\boldsymbol{I} - \boldsymbol{H} - \boldsymbol{H}^T + \boldsymbol{H}\boldsymbol{H}^T) = \boldsymbol{\sigma^2}(\boldsymbol{I} - \boldsymbol{H}) \end{aligned}$$

Neboť projekční matice $\boldsymbol{H}$ je symetrická ($\boldsymbol{H}^T = \boldsymbol{H}$) a idempotentní ($\boldsymbol{H}^2 = \boldsymbol{H}$)

$$\boldsymbol{H}\boldsymbol{H}^T = \boldsymbol{H}^2 = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}$$

Vektor reziduí je funkcí náhodných vektorů $\boldsymbol{y}$ a $\boldsymbol{b}$. Matice $\boldsymbol{H}$ s prvky h_{ij}, $i, j = 1, 2, \ldots, n$ je symetrická, ale nemusí být diagonální.

3.1.1 Typy reziduí

1. Klasická rezidua

Jsou to rezidua, která jsme doposud používali $\varepsilon_i = y_i - \hat{y}_i$, i = 1, 2, …, n, což jsou rozdíly mezi skutečnými a odhadnutými hodnotami vysvětlované proměnné y.

Jejich rozptyly $var(\varepsilon_i) = s_e^2(1 - h_{ii})$ nejsou konstantní, i když $var(\varepsilon_i) = \sigma^2$ konstantní je.

2. Predikovaná rezidua

Jsou to rezidua počítaná bez i-tého pozorování. Označme $\hat{y}_{i(-i)}$ vyrovnanou hodnotu, kterou jsme získali na základě $n - 1$ pozorování při vypuštění i-tého pozorování. Predikované reziduum je vypočteno jako rozdíl skutečné a odhadnuté hodnoty

$$\varepsilon_{i(-i)} = y_i - \hat{y}_{i(-i)}$$

3. Standardizovaná rezidua

Také se jim říká vnitřně studentizovaná residua a jsou definovány takto

$$\varepsilon_{Si} = \frac{\varepsilon_i}{s\sqrt{1 - h_{ii}}}$$

4. Plně studentizovaná rezidua

Jsou alternativou k vnitřně studentizovaným reziduím, nazývají se Jackknife rezidua, která jsou konstruována podobně jako predikovaná rezidua na vypuštění i-tého pozorování. Jackknife rezidua jsou definovaná jako

$$\varepsilon_{Ji} = \frac{\varepsilon_i}{s_{(-i)}\sqrt{1 - h_{ii}}}$$

5. Leverage

Tyto charakteristiky ohodnocují vliv i-tého bodu na hodnoty odhadů parametrů. Jsou to diagonální prvky projekční matice. Platí, že

$$0 < h_{ii} < 1 \text{ a } \sum_{i=1}^{n} h_{ii} = k + 1$$

kde k je počet regresorů.

6. Cookova vzdálenost

Je vlastně euklidovská vzdálenost mezi vektorem predikce závisle proměnné získaném metodou nejmenších čtverců a tímtéž vektorem při vynechání i-tého bodu. Tato charakteristika slouží také k posouzení vlivu i-tého pozorování na odhad vektorů regresních parametrů $\boldsymbol{b}$.

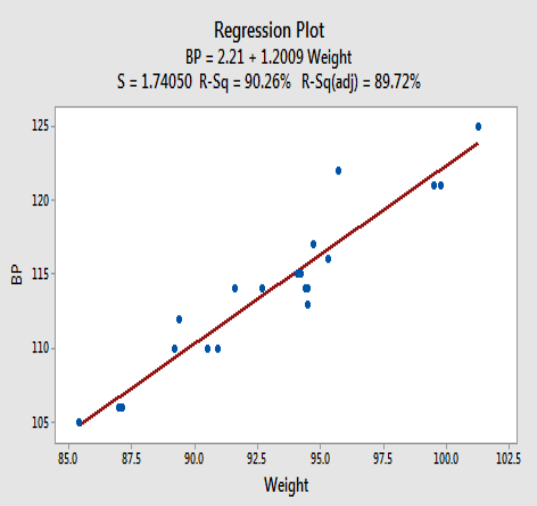
3.1.2 Grafická analýza reziduí

Regresní přímka jednoduchým způsobem vystihuje vztah mezi proměnnými, ale i odchylky od tohoto vztahu jsou důležité. Například se chceme přesvědčit, zda je vztah

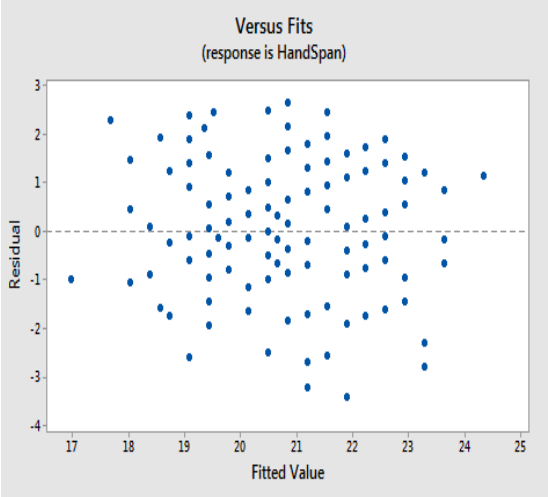
opravdu lineární nebo se zajímáme o hodnoty, které jsou v konfiguraci dat neobvyklé. K tomu slouží především analýza reziduálních hodnot. Sestrojujeme dvojrozměrný bodový graf, jenž zachycuje vztah reziduí k hodnotám nezávisle proměnné *x*.

Bude následovat skupina obrázků, kde jsou vyznačeny různé konfigurace reziduálních hodnot ε_i. Pro všechny následující situace je obrázek vlevo vizualizace dat s červenou předpovězenou regresní přímkou. Pravý obrázek je pak obrázek grafů předpovědí vůči reziduím.

Model splňující předpoklady lineárního vztahu. V následujícím grafu data splňují všechny předpoklady. Je vidět, že regresní přímka je velmi blízko skutečné závislosti. Rezidua jsou náhodně rozmístěna kolem nuly, systematicky se nezvyšují ani se systematicky nesnižují spolu s předpovídanými hodnotami. Takto by měl vypadat graf v případě správně zvoleného modelu a při modelování bychom se měli snažit tohoto výsledku dosáhnout.

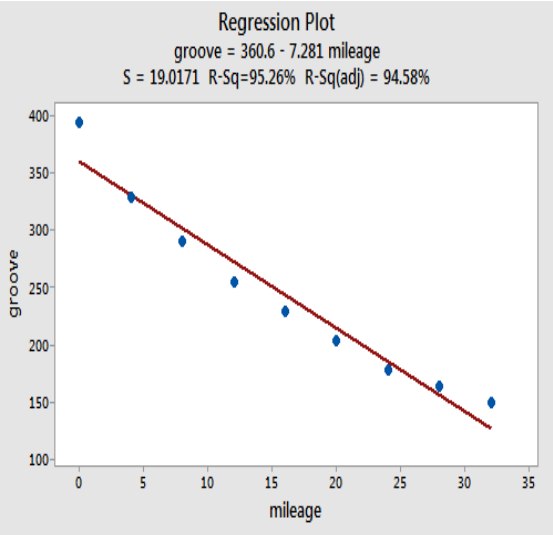


Obrázek č. 1: Správný model [9]

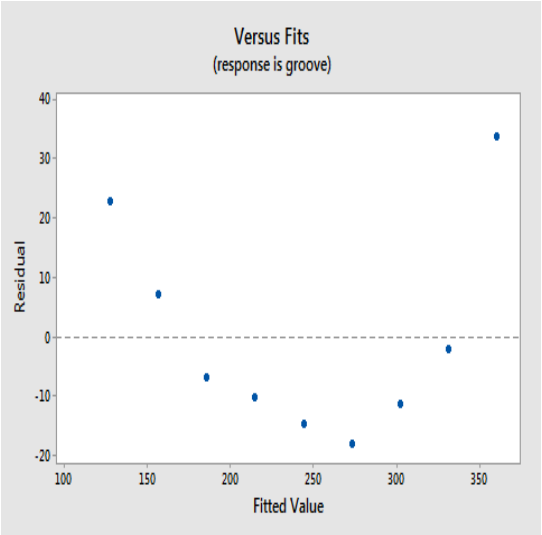


Obrázek č. 2: Správný model [9]

Nesprávně zvolený model (nelinearita). Reziduální odchylky jsou uspořádané nenáhodně kolem nuly, jejich tvar rozmístění naznačuje, jaký typ regresní funkce je třeba uvažovat. Na tomto obrázku je jasně vidět, že body naznačují parabolu, měli bychom tedy použít kvadratickou regresi.

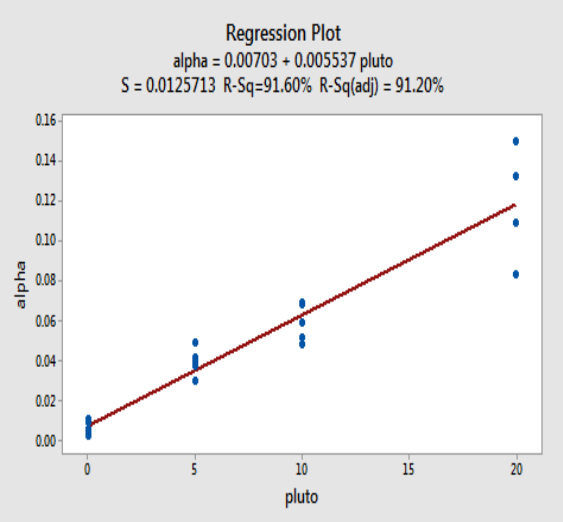


Obrázek č. 3: Nesprávný model [9]

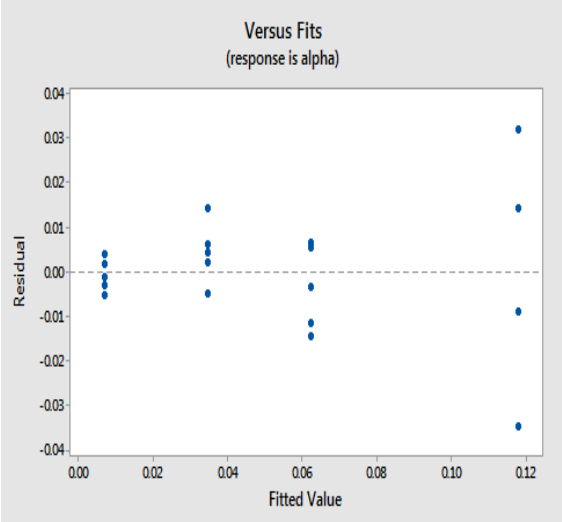


Obrázek č. 4: Nesprávný model [9]

Modely s nekonstantním rozptylem (tzv. heteroskedasticita reziduí). Případ nekonstantnosti rozptylu může typicky vypadat tak, že rozptyl bodů kolem regresní přímky roste spolu s rostoucím *x*. Není tedy splněn předpoklad konstantního rozptylu.

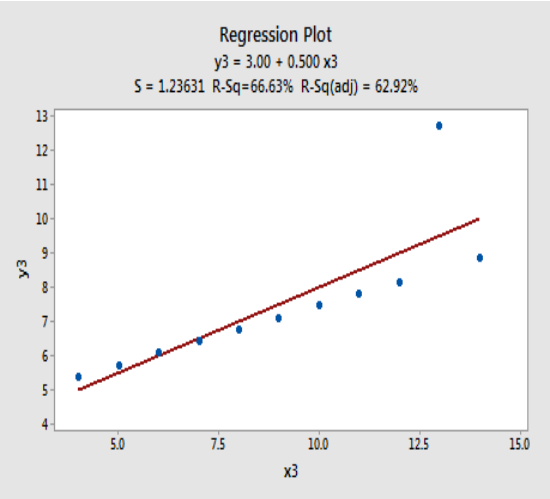


Obrázek č. 5: Nekonstantní rozptyl [9]

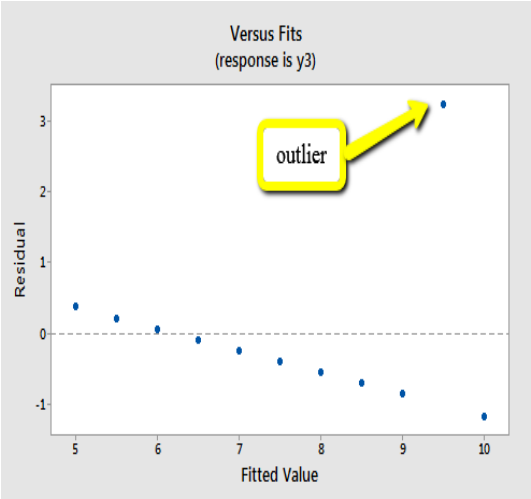


Obrázek č. 6: Nekonstantní rozptyl [9]

Model s extrémními hodnotami (odlehlé hodnoty, anglicky outliers). Tato situace znamená porušení předpokladů o rozdělení chybového členu. Odlehlé pozorování je pozorování, jehož hodnoty jsou velmi vzdáleny od hodnot většiny dat. Klasickým důsledkem přítomnosti odlehlých dat je to, že už i jen málo odlehlých pozorování dokáže velmi vychýlit odhad. Odlehlý bod je takový, který leží mimo základní konfiguraci bodů v grafu. Bod může být odlehlý ve směru y, ve směru x nebo v obou směrech. Odlehlý údaj ve směru nezávisle proměnné se nazývá vybočující. Odlehlé hodnoty při regresi jsou nápadně velké reziduální hodnoty, upozorňující na špatnou predikci závisle proměnné. Bod nazýváme vlivný, pokud se po jeho odstranění podstatně změní poloha regresní přímky, tedy podstatně ovlivňuje odhady regresních koeficientů. Vybočující pozorování jsou nezvyklé konfigurace hodnot týkající se společného rozdělení nezávislých proměnných.



Obrázek č. 7: Extrémní hodnoty [9]



Obrázek č. 8: Extrémní hodnoty [9]

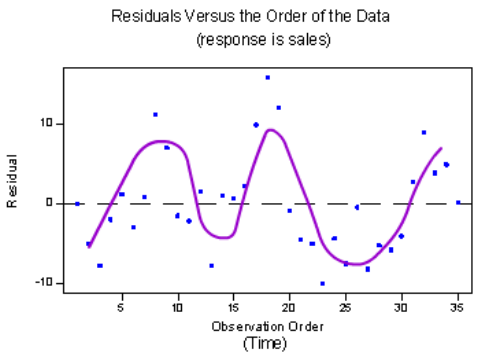
Závislost mezi složkami reziduálního členu (tzv. autokorelace reziduí). Mezi náhodnými složkami existuje regresní vztah. Tím je porušen požadavek na kovarianční matici pro odhad regresních parametrů metodou nejmenších čtverců. Nejvíce se vyskytuje u jednorovnicového modelu, jehož pozorování tvoří časové řady. Může vzniknout chybnou specifikací modelu, užitím zpožděných vysvětlujících proměnných nebo také chybami v měření. O přítomnosti autokorelace náhodných složek se lze přesvědčit jen nepřímo, vyšetřením reziduálních hodnot. Nejčastějším případem je autokorelace prvního řádu, což zapisujeme:

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i$$

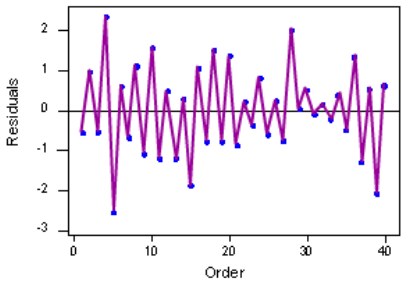
ve které ρ je neznámý parametr. Analogicky bychom sestrojili autokorelační strukturu druhého a třetího řádu. Velmi názorný obrázek o míře autokorelovanosti náhodných složek podává Durbin-Watsonův test určený výrazem:

$$D = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})}{\sum_{i=1}^{n}\varepsilon_i^2}$$

Rozsah přípustných hodnot se pohybuje v rozmezí $< 0{,}4 >$, přičemž obě krajní hodnoty signalizují maximální možnou korelovanost následujících reziduálních hodnot. Pokud D = 0 jde o kladnou autokorelaci prvního řádu, v případě D ≅ 4 jde o zápornou autokorelaci, zatímco prostřední hodnota D ≅ 2 znamená nepřítomnost autokorelace.



Obrázek č. 9: Pozitivní autokorelace [9]



Obrázek č. 10: Negativní autokorelace [9]

3.1.3 Testování hypotéz

F-testem určujeme, zda pokusný zásah má vliv na rozptyl (proměnlivost) zkoumané náhodné veličiny. Výpočet vychází z dat dvou výběrových souborů, které srovnáváme. Předpokládejme, že rozdělíme pozorovaná rezidua do dvou skupin, první budou rezidua s nízkými hodnotami prediktoru a druhá s vysokými hodnotami. Testujeme

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Použijeme F-statistiku $F^* = s_1^2/s_2^2$. Tato testová statistika je rozdělena podle F_{n_1-1,n_2-1} rozdělení, pokud $F^* \geq F_{n_1-1,n_2-1;1-\alpha}$, potom zamítáme nulovou hypotézu a závěrem je, že rozptyl není konstantní.

Dále můžeme provést test, v němž nulová hypotéza tvrdí, že rezidua mají normální rozdělení oproti alternativní hypotéze, že rezidua nemají normální rozdělení. K tomu slouží několik testů, např. Shapiro-Wilkův test, Ryan-Joiner test anebo Kolmogorovův-Smirnovův test dobré shody. Více o těchto testech nalezneme v [13, str. 128].

3.2 Multikolinearita

Považuje se za jedno z nejzávažnějších narušení vypovídající schopnosti hodnot regresních koeficientů a nepřímo se dotýká i kvality regresních odhadů. Silná vzájemná závislost vysvětlujících proměnných výrazně zhoršuje interpretaci regresních koeficientů. Pojem je velmi úzce svázán se silnou vzájemnou lineární závislostí vysvětlujících proměnných, jejímž důsledkem je špatně podmíněná matice $\boldsymbol{X}$. Přesnou multikolinearitou je případ, kdy jednotlivé sloupce $\boldsymbol{x}_j$, $j = 1, 2, \ldots, K$, matice $\boldsymbol{X}$ jsou lineárně závislé, takže pro aspoň jednu nenulovou konstantu c_j platí

$$c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2 + \cdots + c_K\boldsymbol{x}_K = \boldsymbol{o}_n$$

Vektory hodnot vysvětlujících proměnných lze vyjádřit jako lineární kombinace vektorů hodnot jiných vysvětlujících proměnných. Tato situace může vzniknout nadbytečností některých vysvětlujících proměnných. Za to může špatná volba kombinací hodnot vysvětlujících proměnných, ale i náhoda při malém rozsahu výběru.

Multikolinearitou se míní případ, kdy sloupce matice $\boldsymbol{X}$ (jednotlivé regresory) jsou téměř lineárně závislé. Souvisí s předpokladem o pozitivní definitivnosti matice $\boldsymbol{X}^T\boldsymbol{X}$, a tím jednoznačnosti rovnice

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

To má za následek že:

- determinant matice $\boldsymbol{X}^T\boldsymbol{X}$ je číslo blízké nule,
- některá vlastní čísla matice $\boldsymbol{X}^T\boldsymbol{X}$ jsou blízká nule.

Multikolinearita má za následek nadhodnocení součtu čtverců regresních koeficientů, to znamená, že některé proměnné se mohou zdát důležitější, než ve skutečnosti jsou. Také zvyšuje rozptyly odhadů, což má za následek nízké hodnoty testové statistiky pro

individuální t-testy, při kterých se některé regresní koeficienty ukazují jako statisticky nevýznamně odlišné od nuly i v případě kvalitního regresního modelu.

Je možné ji odstranit v případě přeurčeného regresního modelu, neboli v případě výskytu zbytečných vysvětlujících proměnných, jejich identifikací a vypuštěním z regresní rovnice. Je-li způsobená nevhodnou volbou kombinací hodnot vysvětlujících proměnných, je možné nedostatky napravit a pořídit si kvalitnější data. Nejčastějším případem je věcně zdůvodněná závislost vzájemně propojených veličin. V takovém případě vypuštění proměnných povede k systematickým chybám a pořízení nových dat také nepomůže. Jedinou možností je maximálně využít všechny věcné a empirické informace o regresním modelu, což vede ke zvýšení kvality.

4 Aplikace regresní analýzy

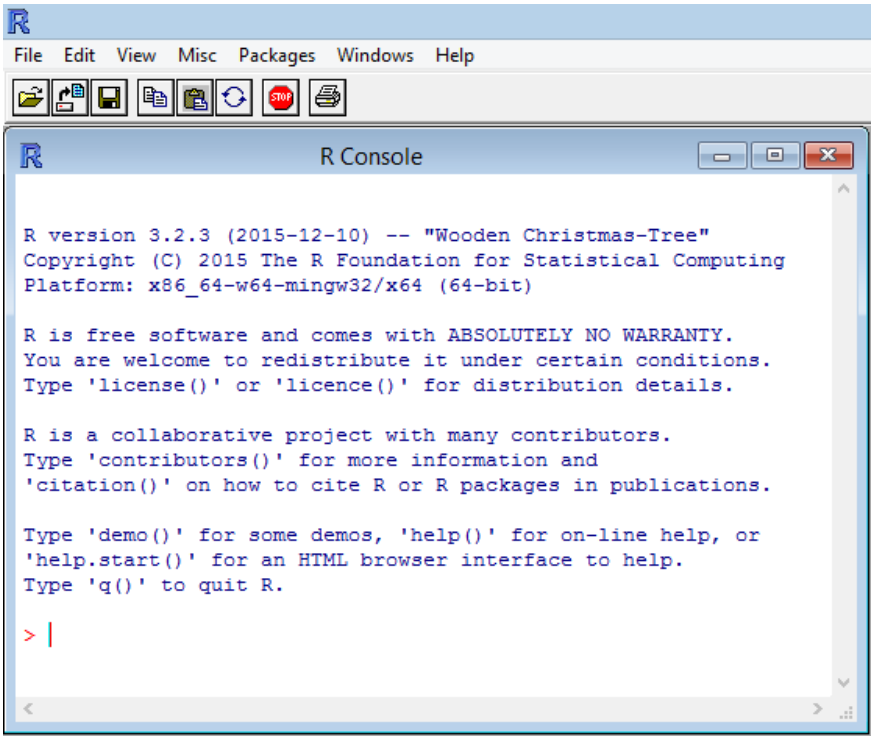
4.1 R software

Ke zpracování dat si můžeme vybrat z velké škály statistických softwarů. Zvolila jsem si R software, se kterým mám už zkušenosti z předchozího semestru. R je matematický software, který se specializuje na statistiku a grafické výstupy. R funguje na široké škále UNIX platforem a jim podobných systémech (FreeBSD a Linux) a dále v operačních systémech Windows a MacOS. Je možné analyzovat data z veškerých lidských oblastí (astrofyzika, ekonomika, geografie, zemědělství a mnoho dalších).

R je jazyk a prostředí pro výpočet a grafické znázornění statistických úloh. Může být považováno za jiné provedení programovacího jazyka S vyvinutého v Bell Laboratories (nyní Lucent Technologies). Mezi R a S jazyky lze nalézt rozdíly, ale mnoho kódů napsaných pro S běží i pod jazykem R. Jazyk R nabízí mnoho statistických technik od samostatného vytváření lineárních a nelineárních modelů po klasifikaci, clustering či jen můžeme provádět klasické lineární testování a to vše lze zakončit grafickými výstupy.

Ke stažení je zdarma pod licencí GNU/GPL na stránkách samotné organizace http://www.r-project.org/ , zde jsou také návody, informace o R, volně stažitelné balíky, které pomáhají zlepšit práci a ostatní dokumentace.

Pokud chceme rozšířit R, lze stáhnout tzv. packages (balíčky). Základní jsou spuštěny po instalaci R a mnoho dalších lze najít a získat z databáze CRAN. V ní udržované balíčky tematicky pokrývají prakticky všechny oblasti současné statistiky. Při zapnutí R softwaru je zobrazeno okno s informacemi o verzi, licenci a možnosti vyvolání nápovědy (help). Zde také začneme s psaním kódu, jak naznačuje červený kurzor.



Obrázek č. 11: Rozhraní R softwaru

4.2 Příjem obyvatel

Nyní si na konkrétním datovém souboru ukážeme postup při regresní analýze v prostředí R. Soubor se týká hodnocení kvality života ve vybraných městech na Floridě a pochází z [2]. Data jsou k dispozici v podobě textového souboru, jehož obsah si můžete prohlédnout v tabulce č. 2. Konkrétní názvy měst v souboru můžeme najít v [3].

Income	Commute	JobGrowth	Physicians	MurderRate	RapeRate	Golf	Restaurants	Housing	MedianAge	Literacy	HouseholdIncome	Recreation
26000	49.2	10.8	1987	5.3	51.3	925	5582	109400	35.3	5.15	68000	2620
29300	45.3	9.5	517	6.6	50.8	364	9988	97000	43.2	5.97	70400	3066
24800	39.8	8.2	592	8.2	77.7	1627	20511	114700	29.5	9.41	60500	1297
27900	46.8	7.6	3310	6.7	51.2	956	8946	99100	40.5	4.61	65900	2902
37500	39.9	12.2	975	5.1	40.1	426	4000	122200	47.1	5.64	84700	2214
31900	49.5	7.7	2238	6.9	38.0	1459	8970	145300	39.3	4.80	75800	1402
25300	44.4	5.4	611	4.5	38.8	1063	9570	99500	38.6	6.84	62600	2900
22000	44.8	6.2	272	7.5	65.7	951	19101	76400	41.6	2.79	54800	2448
29400	44.9	7.8	381	8.4	48.7	349	12099	112500	41.8	4.48	72900	2756
42400	44.7	8.0	1812	8.1	45.4	397	10953	143500	41.2	5.16	100000	2508
40500	40.0	10.9	294	8.0	69.6	191	2655	173600	41.7	6.41	102000	3000
24700	38.7	9.0	196	2.8	19.0	449	15796	129200	33.4	1.66	65300	1570
24400	41.1	8.7	404	7.3	77.2	1590	16001	126500	30.6	5.60	62200	1713
22400	42.8	8.3	534	5.7	57.9	1160	16712	102700	34.5	2.16	59200	2190
22200	37.8	8.4	166	5.6	50.9	815	11856	110300	35.4	2.72	57100	2142
27500	48.4	8.1	1553	14.0	83.6	1195	12348	107400	34.3	4.03	72000	2657
23100	44.5	4.7	502	7.9	42.7	556	65804	116000	38.5	2.07	59400	2066
25000	41.4	13.9	172	4.0	17.8	459	36151	120000	52.7	3.61	57300	1467
25800	53.5	5.3	4143	16.8	57.4	3054	14310	132800	36.2	5.03	71900	3520
22600	45.0	6.5	526	5.5	52.2	861	8878	86500	41.5	5.29	54000	2977

Tabulka č. 2: Data vybraná pro regresní analýzu [2]

- **Income:** průměrný příjem na obyvatele
- **Commute**: průměrná denní doba dojíždění v minutách
- **Job Growth**: předpokládané procento nárůstu počtu pracovních míst v příštích pěti letech
- **Physicians:** počet lékařů na 100 000 obyvatel
- **Murder Rate:** průměrný počet vražd za 8 let na 100 000 obyvatel
- **Rape Rate:** průměrný počet znásilnění za 8 let na 100 000 obyvatel
- **Golf:** počet obyvatel připadající na jednu golfovou jamku
- **Restaurants:** index kvality restaurací (rezidenti na jakostní bod)
- **Housing:** medián ceny nemovitostí
- **Median Age:** medián věku obyvatel
- **Literacy:** počet návštěv veřejných knihoven připadající na jednoho obyvatele
- **Household Income:** průměrný příjem domácnosti
- **Recreation**: hodnocení příležitostí pro trávení volného času podle publikace Places Rated Almanac's rating (vyšší číslo indikuje více rekreace)

Data budeme zpracovávat ve výše představeném R softwaru. Výstupy uvádíme bez větších editačních úprav v surovém stavu. Po zapnutí R musíme načíst data, důležitým krokem je změnit složku, ze které data načítáme (v menu *File* zvolíme položku *Change directory*).

V textovém souboru (který můžeme v MS Windows editovat například pomocí programu Poznámkový blok) máme naše data s názvem „*city*“, která budeme vkládat do R pomocí příkazu

```
city<-read.table("city.txt",header=T)
```

Zpřístupníme sloupce tabulky, abychom na ně mohli odkazovat přímo jejich názvy, to vše pomocí příkazu `atttach(city)`, a můžeme názvy sloupců vypsat příkazem `names(city)`.

Ve snaze předejít multikolinearitě (kapitola 3.2) bychom se měli podívat, jak jsou jednotlivé regresory lineárně závislé, tedy jaké jsou jejich korelační koeficienty. Tato symetrická tabulka ukazuje, jak je silná závislost mezi proměnnými.

	Income	Commute	JobGrowth	Physicians	MurderRate	RapeRate	Golf
Income	1.0000000	0.16842105	0.16842105	0.29473684	0.15789474	-0.16842105	-0.25263158
Commute	0.1684211	1.00000000	-0.28421053	0.51578947	0.25263158	0.09473684	0.17894737
JobGrowth	0.1684211	-0.28421053	1.00000000	-0.26315789	-0.29473684	-0.07368421	-0.26315789
Physicians	0.2947368	0.51578947	-0.26315789	1.00000000	0.16842105	0.09473684	0.32631579
MurderRate	0.1578947	0.25263158	-0.29473684	0.16842105	1.00000000	0.44210526	0.16842105
RapeRate	-0.1684211	0.09473684	-0.07368421	0.09473684	0.44210526	1.00000000	0.28421053
Golf	-0.2526316	0.17894737	-0.26315789	0.32631579	0.16842105	0.28421053	1.00000000
Restaurants	-0.4421053	-0.13684211	-0.13684211	-0.24210526	0.18947368	0.05263158	0.22105263
Housing	0.3157895	-0.13684211	0.17894737	0.05263158	0.08421053	-0.20000000	-0.09473684
MedianAge	0.2526316	0.07368421	0.11578947	-0.09473684	-0.10526316	-0.30526316	-0.47368421
Literacy	0.3052632	0.00000000	0.08421053	0.16842105	0.05263158	0.16842105	0.04210526
HouseholdIncome	0.8105263	0.16842105	0.14736842	0.31578947	0.22105263	-0.06315789	-0.21052632
Recreation	0.1473684	0.36842105	-0.20000000	0.22105263	0.16842105	0.20000000	-0.17894737

	Restaurants	Housing	MedianAge	Literacy	HouseholdIncome	Recreation
Income	-0.44210526	0.31578947	0.25263158	0.30526316	0.81052632	0.1473684
Commute	-0.13684211	-0.13684211	0.07368421	0.00000000	0.16842105	0.3684211
JobGrowth	-0.13684211	0.17894737	0.11578947	0.08421053	0.14736842	-0.2000000
Physicians	-0.24210526	0.05263158	-0.09473684	0.16842105	0.31578947	0.2210526
MurderRate	0.18947368	0.08421053	-0.10526316	0.05263158	0.22105263	0.1684211
RapeRate	0.05263158	-0.20000000	-0.30526316	0.16842105	-0.06315789	0.2000000
Golf	0.22105263	-0.09473684	-0.47368421	0.04210526	-0.21052632	-0.1789474
Restaurants	1.00000000	-0.03157895	-0.28421053	-0.37894737	-0.42105263	-0.3684211
Housing	-0.03157895	1.00000000	-0.05263158	0.04210526	0.42105263	-0.2842105
MedianAge	-0.28421053	-0.05263158	1.00000000	0.10526316	0.16842105	0.2631579
Literacy	-0.37894737	0.04210526	0.10526316	1.00000000	0.22105263	0.2526316
HouseholdIncome	-0.42105263	0.42105263	0.16842105	0.22105263	1.00000000	0.1684211
Recreation	-0.36842105	-0.28421053	0.26315789	0.25263158	0.16842105	1.0000000

Názorným případem multikolinearity je vysoká hodnota korelace mezi proměnnými *Income* a *HouseholdIncome*. To znamená, že proměnnou *HouseholdIncome* můžeme z modelu vyloučit. Je logické, že průměrný příjem domácností (*HouseholdIncome*) bude úzce souviset s průměrným příjmem osob (*Income*).

K analýze budeme používat vícenásobný regresní model (tak, jak je popsaný v kapitole 1.3.1). Po zahrnutí sloupce jedniček je $k = 12$, $n = 20$. Provedeme lineární regresi proměnné průměrný příjem na obyvatele (*Income*) vzhledem ke všem ostatním nezávisle proměnným. Metoda začíná s modelem, ve kterém jsou zahrnuty všechny potencionálně relevantní nezávislé proměnné. Jsou-li všechny proměnné pomocí testů vyhodnoceny jako statisticky důležité, není potřeba pokračovat. Pokud ne, následně se vypustí ta proměnná, jejíž přínos ke zmenšení hodnoty SSE je nejmenší, tedy ta, jejíž absolutní hodnota statistiky T je nejmenší. Poté sestavíme nový model a proces opakujeme. Konec nastane tehdy, jsou-li všechny proměnné statisticky významné. Použijeme následující příkaz a statistický software vypočítá tyto výsledky

```
Call:
lm(formula = Income ~ Commute + JobGrowth + Physicians + MurderRate +
    RapeRate + Golf + Restaurants + Housing + MedianAge + Literacy +
    Recreation)

Coefficients:
(Intercept)      Commute    JobGrowth   Physicians   MurderRate     RapeRate
 10063.7865    -133.9266    -597.6988       2.1946     501.9910      42.0713
       Golf  Restaurants      Housing    MedianAge     Literacy   Recreation
    -6.7428      -0.1434       0.1424     355.0373     690.0484      -1.9207
```

Z první části se dozvíme pouze hodnoty regresních koeficientů, pro podrobnější výpis výsledků lineární regrese použijeme příkaz `summary`, kde právě uvidíme hodnotu statistiky T a zjistíme, které proměnné můžeme považovat za relevantní. Pro zjednodušení si ale nejdříve pojmenujeme předchozí příkaz pro lineární model jako „*model*“

```
model<-
lm(Income~Commute+JobGrowth+Physicians+MurderRate+RapeRate+Golf
+Restaurants+Housing+MedianAge+Literacy+Recreation)

summary(model)
```

```
Call:
lm(formula = Income ~ Commute + JobGrowth + Physicians + MurderRate +
    RapeRate + Golf + Restaurants + Housing + MedianAge + Literacy +
    Recreation)

Residuals:
    Min      1Q  Median      3Q     Max
-2506.1 -1548.6  -156.5   795.7  3590.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.006e+04  1.645e+04   0.612  0.55762
Commute     -1.339e+02  2.988e+02  -0.448  0.66593
JobGrowth   -5.977e+02  4.229e+02  -1.413  0.19529
Physicians   2.195e+00  1.121e+00   1.957  0.08604 .
MurderRate   5.020e+02  4.710e+02   1.066  0.31758
RapeRate     4.207e+01  7.112e+01   0.592  0.57050
Golf        -6.743e+00  1.798e+00  -3.750  0.00562 **
Restaurants -1.434e-01  6.508e-02  -2.203  0.05870 .
Housing      1.424e-01  3.876e-02   3.674  0.00627 **
MedianAge    3.550e+02  1.959e+02   1.812  0.10758
Literacy     6.900e+02  4.175e+02   1.653  0.13697
Recreation  -1.921e+00  1.646e+00  -1.167  0.27690
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2668 on 8 degrees of freedom
Multiple R-squared:  0.9169,    Adjusted R-squared:  0.8026
F-statistic: 8.022 on 11 and 8 DF,  p-value: 0.003332
```

V této části si všimneme hlavně řádku *Multiple R-squared*, který udává hodnotu koeficientu determinace R^2. Lze konstatovat, že 91,69% variability závisle proměnné se modelem podařilo vysvětlit. Ke srovnání modelu s dalšími modely použijeme hodnotu *adjusted R-squared*, zde 80,26%. Z výsledků lze dále vidět, že některé proměnné nejsou v modelu statisticky významné, protože p-hodnota Pr(>|t|) je větší než 0,05. Tyto proměnné tedy lze

z modelu vyloučit. Provedeme zpracování *modelu2* bez proměnných, které nejsou statisticky významné pomocí

```
model2<-lm(Income~Physicians+Golf+Restaurants+Housing)

summary(model2)
```

```
Call:
lm(formula = Income ~ Physicians + Golf + Restaurants + Housing)

Residuals:
    Min      1Q  Median      3Q     Max
-5897.5 -1053.5    35.6  1302.8  5258.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.424e+04  4.221e+03   3.375  0.00417 **
Physicians   2.120e+00  8.557e-01   2.478  0.02561 *
Golf        -5.451e+00  1.403e+00  -3.885  0.00147 **
Restaurants -1.213e-01  5.381e-02  -2.254  0.03961 *
Housing      1.571e-01  3.337e-02   4.709  0.00028 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3135 on 15 degrees of freedom
Multiple R-squared:  0.7848,    Adjusted R-squared:  0.7274
F-statistic: 13.67 on 4 and 15 DF,  p-value: 6.832e-05
```

Po vyloučení proměnných, které nejsou statisticky významné, jsme dostali *model2*, kde koeficient determinace už sice není tak vysoký R^2 = 78,48%, ale model je stále statisticky významný. Z tohoto modelu plyne zjištění, že hodnoty veličiny y (průměrný příjem na obyvatele) je možné dobře odhadnout na základě mediánu ceny nemovitostí, počtu lékařů, indexu kvality restaurací a počtu obyvatel připadajících na jednu golfovou jamku. Výpočtem dojdeme k vektoru $\boldsymbol{b} = \begin{pmatrix} 14244{,}93 \\ 2{,}12 \\ -5{,}45 \\ -0{,}12 \\ 0{,}16 \end{pmatrix}$, kde koeficienty proměnných následují po koeficientu absolutního členu. Hledaný vztah má tedy podobu rovnice: $Income = 14244{,}93 + 2{,}12 \cdot Physicians - 5{,}45 \cdot Golf - 0{,}12 \cdot Restaurants + 0{,}16 \cdot Housing + \varepsilon$

Další možnosti, které nám software nabízí, jsou například vypsání regresních koeficientů modelu, tabulka ANOVA, 95% - ní intervaly spolehlivosti pro regresní koeficienty, vypsání hodnot reziduí a také grafy reziduí.

```
                    2.5 %        97.5 %
(Intercept) 5247.52949526  2.324232e+04
Physicians     0.29628769  3.943918e+00
Golf          -8.44161120 -2.460509e+00
Restaurants   -0.23594970 -6.572233e-03
Housing        0.08599925  2.282416e-01
```

Obrázek č. 12: Intervaly spolehlivosti pro regresní koeficienty

```
Analysis of Variance Table

Response: Income
             Df    Sum Sq   Mean Sq F value    Pr(>F)
Physicians    1  23022506  23022506  2.3426 0.1466958
Golf          1 234942511 234942511 23.9057 0.0001964 ***
Restaurants   1  61672539  61672539  6.2753 0.0242628 *
Housing       1 217909882 217909882 22.1726 0.0002798 ***
Residuals    15 147418062   9827871
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obrázek č. 13: ANOVA tabulka

1	2	3	4	5	6
-3927.43545	1913.63995	2634.31856	-2637.08611	4795.05320	-878.50640
7	8	9	10	11	12
1081.15306	2674.56930	40.82329	5258.90893	-281.24196	-5897.46107
13	14	15	16	17	18
30.30064	-763.58576	-3846.96217	1099.16785	575.05901	-1578.29309
19	20				
288.67757	-581.09936				

Obrázek č. 14: Hodnoty reziduí

Jak už jsem zmínila dalším výstupem, který nám statistický software dovoluje, je grafická analýza reziduí. Pomocí příkazů

```
layout(matrix(c(1,2,3,4),2,2)), plot(model2)
```

vytvoříme grafické okno znázorňující čtyři grafy v jednom.

První graf se nazývá *Residuals versus Fitted* (nahoře vlevo) tento graf ukazuje hodnoty reziduí proti předpovězeným hodnotám. Aby byl splněn předpoklad normality a linearity, rezidua by měla mít normální rozložení se střední hodnotou nula. To znamená, že by měla být nesystematicky rozložena okolo nuly. Pokud graf ukazuje nějaký důkaz zakřiveného vztahu, navrhuje se použít jinou než lineární regresi, například kvadratickou.

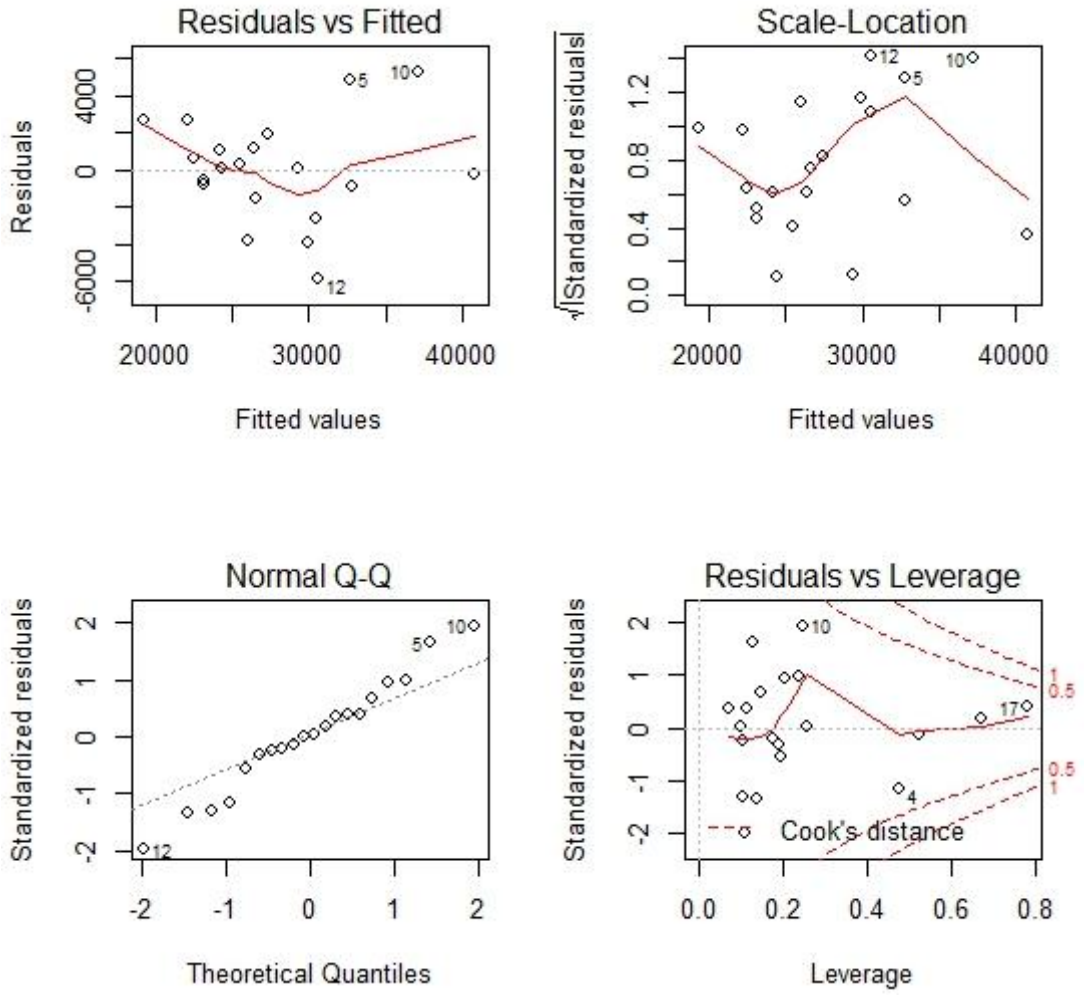
Druhý graf *Scale Location* (nahoře vpravo) znázorňuje odmocniny standardizovaných reziduí oproti předpovězeným hodnotám. Splňujeme-li předpoklad konstantního rozptylu, body v tomto grafu by měly být umístěny okolo horizontální přímky. Pokud by tady byl problém, jako je například zvětšující se rozptyl reziduí při zvětšujících se předpovězených hodnotách, body budou rozděleny do podoby trojúhelníkového tvaru.

Další graf je normální *Q-Q-graf* (dole vlevo) standardizovaných reziduí sloužící k prověření normality reziduí. Čím blíže leží body Q-Q-grafu k přímce, tím mají bližší rozdělení normálnímu rozdělení. Rezidua mají mít normální rozdělení, proto je naším cílem mít body v grafu co nejvíce na přímce.

Poslední graf, který software nabízí, se nazývá *Residuals versus Leverage* (dole vpravo) graf standardizovaných reziduí proti leverage, který přidává pásma odpovídající Cookově vzdálenosti 0,5 a 1. Smyslem tohoto grafu je upozornit na hodnoty, které mají největší vliv na odhady parametrů, to znamená, že body jsou blízko Cookovy vzdálenosti. Tento graf identifikuje odlehlé hodnoty (outliers), body s vysokou hodnotou leverage a vlivná

pozorování (influential observations). Přečtení tohoto grafu shledávám poměrně obtížným a ne až tak použitelným.

Výsledky jsou zobrazeny na obrázku č. 15, z grafů můžeme vidět, že předpoklady modelu se zdají splněny.

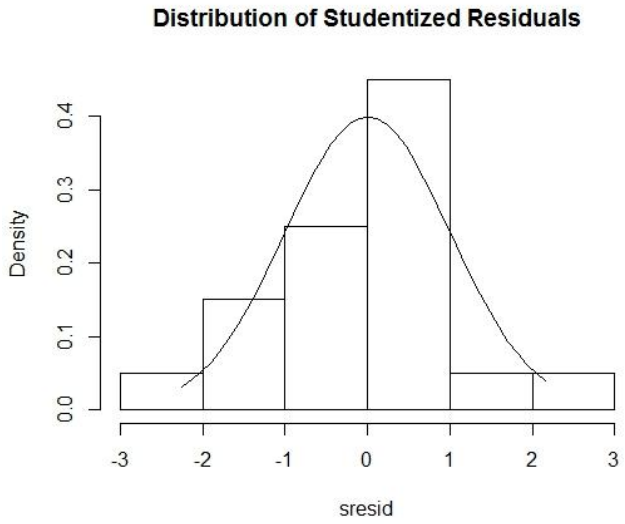


Obrázek č. 15: Grafické znázornění reziduí, vlastní zpracování v R

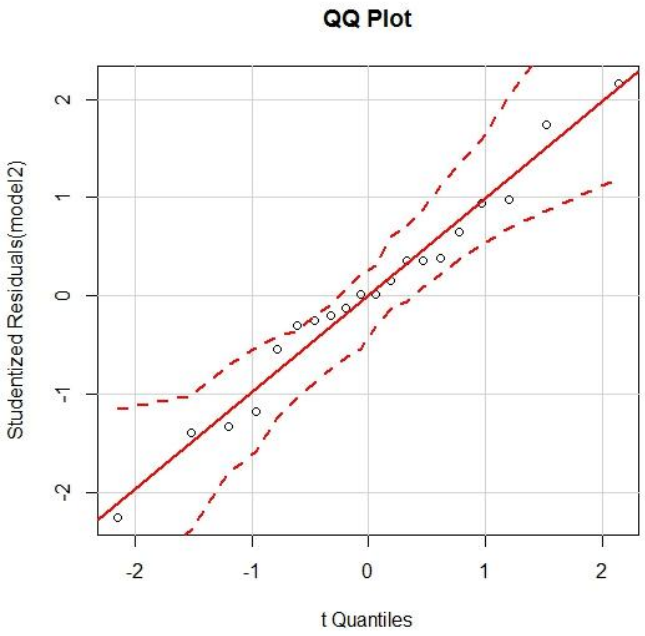
Nejjednodušším prostředkem pro hodnocení tvaru rozdělení je histogram rozdělení standardizovaných reziduí, který je vyobrazen na obrázku č. 16, ale je zásadní vhodné určení počtu dělení sloupců. Pro lepší porovnání můžeme grafem proložit teoretickou Gaussovu křivku. Skutečnost, že vektor reziduí má normální rozdělení se střední hodnotou nula, ověřuje právě histogram, i když malý počet pozorování způsobil jistou deformaci rozdělení.

Jelikož Q-Q grafy jsou významným a hojně využívaným pomocníkem pro vizuální kontrolu předpokladů, podíváme se podrobněji na tento graf. Q-Q graf je založený na porovnání kvantilů teoretického rozdělení a naměřených kvantilů a tedy je vhodnějším nástrojem pro určení rozdělení náhodné veličiny. Tento způsob kontroly je náročný na zkušenost uživatele. Je třeba odhadnout, co lze ještě považovat za přijatelnou přímku a co už nelze. Následující obrázek ukazuje Q-Q graf, z grafu vidíme, že rozdělení reziduí můžeme

považovat za normální, odchylky od přímky nejsou velké. Tedy data i model vyhovují předpokladu a výsledky můžeme považovat za nezavádějící.



Obrázek č. 16: Histogram standardizovaných reziduí, vlastní zpracování v R



Obrázek č. 17: Normální Q-Q graf, vlastní zpracování v R

Mezi předpoklady modelu patří i nekorelovanost reziduí, o míře autokorelovanosti náhodných složek podává informaci Durbin-Watsonův test, který je popsán v kapitole 3.1.2. D-W statistika je rovna 2,248872, tak se nemusíme znepokojovat autokorelací reziduí.

4.3 Prodejní cena domu

Tento datový soubor se týká informací o prodejcích cenách domů, data pochází z [10].

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	4.9176	1.0	3.4720	0.998	1.0	7	4	42	3	1	0	25.9
2	5.0208	1.0	3.5310	1.500	2.0	7	4	62	1	1	0	27.9
3	4.5429	1.0	2.2750	1.175	1.0	6	3	40	2	1	0	27.9
4	4.5573	1.0	4.0500	1.232	1.0	6	3	54	4	1	0	25.9
5	5.0597	1.0	4.4550	1.121	1.0	6	3	42	3	1	0	29.9
6	3.8910	1.0	4.4550	0.988	1.0	6	3	56	2	1	0	29.9
7	5.8980	1.0	5.8500	1.240	1.0	7	3	51	2	1	1	30.9
8	5.6039	1.0	9.5200	1.501	0.0	6	3	32	1	1	0	28.9
9	16.4202	2.5	9.8000	3.420	2.0	10	5	42	2	1	1	84.9
10	14.4598	2.5	12.8000	3.000	2.0	9	5	14	4	1	1	82.9
11	5.8282	1.0	6.4350	1.225	2.0	6	3	32	1	1	0	35.9
12	5.3003	1.0	4.9883	1.552	1.0	6	3	30	1	2	0	31.5
13	6.2712	1.0	5.5200	0.975	1.0	5	2	30	1	2	0	31.0
14	5.9592	1.0	6.6660	1.121	2.0	6	3	32	2	1	0	30.9
15	5.0500	1.0	5.0000	1.020	0.0	5	2	46	4	1	1	30.0
16	5.6039	1.0	9.5200	1.501	0.0	6	3	32	1	1	0	28.9
17	8.2464	1.5	5.1500	1.664	2.0	8	4	50	4	1	0	36.9
18	6.6969	1.5	6.9020	1.488	1.5	7	3	22	1	1	1	41.9
19	7.7841	1.5	7.1020	1.376	1.0	6	3	17	2	1	0	40.5
20	9.0384	1.0	7.8000	1.500	1.5	7	3	23	3	3	0	43.9
21	5.9894	1.0	5.5200	1.256	2.0	6	3	40	4	1	1	37.5
22	7.5422	1.5	4.0000	1.690	1.0	6	3	22	1	1	0	37.9
23	8.7951	1.5	9.8900	1.820	2.0	8	4	50	1	1	1	44.5
24	6.0931	1.5	6.7265	1.652	1.0	6	3	44	4	1	0	37.9
25	8.3607	1.5	9.1500	1.777	2.0	8	4	48	1	1	1	38.9
26	8.1400	1.0	8.0000	1.504	2.0	7	3	3	1	3	0	36.9
27	9.1416	1.5	7.3262	1.831	1.5	8	4	31	4	1	0	45.8
28	12.0000	1.5	5.0000	1.200	2.0	6	3	30	3	1	1	41.0

Tabulka č. 3: Data vybraná pro regresní analýzu [10]

- V2: místní prodejní ceny ve stovkách dolarů
- V3: počet koupelen
- V4: plocha pozemku (v tisících stop čtverečných)
- V5: velikost obytného prostoru (v tis. stop čtverečných)
- V6: počet garáží
- V7: počet místností
- V8: počet ložnic
- V9: stáří domu v letech
- V10: konstrukční typ 1 = cihla, 2 = cihla/dřevo, 3 = hliník/dřevo, 4 = dřevo.
- V11: architektonický typ 1 = dvoupatrový, 2 = mezonetový, 3 = ranč
- V12: počet krbů
- V13: prodejní cena

V těchto datech budeme zkoumat závislost prodejní ceny na ostatních proměnných, tedy závisle proměnná je prodejní cena (*V13*) a nezávisle proměnné jsou všechny zbývající. Budeme postupovat stejně jako v předchozím příkladu. Načteme data pomocí příkazu

```
house<-read.table("house.txt", header=F,row.names=1)
```

Pro analýzu použijeme vícenásobný lineární regresní model, kde $k = 13$, $n = 28$. Nejprve tedy spočteme korelační koeficienty.

```
             V2         V3          V4          V5          V6          V7
V2   1.00000000  0.6510749  0.51664860  0.52393802  0.44545795  0.46507578
V3   0.65107492  1.0000000  0.36837270  0.58067524  0.36676793  0.51485078
V4   0.51664860  0.3683727  1.00000000  0.44859853  0.22362049  0.35258533
V5   0.52393802  0.5806752  0.44859853  1.00000000  0.28074782  0.58051376
V6   0.44545795  0.3667679  0.22362049  0.28074782  1.00000000  0.53890816
V7   0.46507578  0.5148508  0.35258533  0.58051376  0.53890816  1.00000000
V8   0.34282216  0.5423872  0.23663363  0.55498903  0.50975870  0.85187405
V9  -0.34871106 -0.1796246 -0.22254853 -0.12738812 -0.03891144  0.03546635
V10  0.05223061  0.1333275 -0.11104872 -0.07085313  0.01105649  0.03664577
V11  0.12360629 -0.3073093  0.07756315 -0.02581989  0.01235604 -0.07371541
V12  0.35052684  0.4219414  0.29261586  0.17770466  0.32598708  0.31004372
V13  0.75836136  0.6807842  0.48793741  0.49264119  0.43242586  0.42678622
             V8          V9         V10         V11         V12         V13
V2   0.34282216 -0.34871106  0.05223061  0.12360629  0.35052684  0.75836136
V3   0.54238719 -0.17962461  0.13332755 -0.30730933  0.42194137  0.68078417
V4   0.23663363 -0.22254853 -0.11104872  0.07756315  0.29261586  0.48793741
V5   0.55498903 -0.12738812 -0.07085313 -0.02581989  0.17770466  0.49264119
V6   0.50975870 -0.03891144  0.01105649  0.01235604  0.32598708  0.43242586
V7   0.85187405  0.03546635  0.03664577 -0.07371541  0.31004372  0.42678622
V8   1.00000000  0.15648888  0.09099926 -0.29121760  0.20149017  0.30916855
V9   0.15648888  1.00000000  0.15655377 -0.38839927  0.07224722 -0.30478475
V10  0.09099926  0.15655377  1.00000000 -0.23265449  0.10036250  0.05258045
V11 -0.29121760 -0.38839927 -0.23265449  1.00000000 -0.27529888  0.08814089
V12  0.20149017  0.07224722  0.10036250 -0.27529888  1.00000000  0.40441816
V13  0.30916855 -0.30478475  0.05258045  0.08814089  0.40441816  1.00000000
```

Z této symetrické tabulky vidíme vyšší závislost mezi proměnnou prodejní cena (*V13*) a místní prodejní cena (*V2*). Je celkem jasné, že tyto proměnné spolu budou úzce souviset. Z tohoto důvodu bychom proměnnou *V2* mohli z modelu vyloučit, ale v tomto případě jsem se rozhodla ji zatím v modelu ponechat. Provedeme tedy lineární regresi závisle proměnné *V13* na zbývajících proměnných. Pro výpis výsledků rovnou použijeme příkaz `summary`, jenž nám dá podrobnější popis výsledků, který pro analýzu potřebujeme.

```
summary(lm(V13~V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12))
```

```
Call:
lm(formula = V13 ~ V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 +
    V11 + V12)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4860 -2.4901  0.5456  1.7226  7.0276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.54246    8.40025   0.303   0.7660
V2           0.84203    0.79835   1.055   0.3072
V3           9.13727    7.20045   1.269   0.2226
V4           0.18055    0.52129   0.346   0.7336
V5          13.31512    4.96493   2.682   0.0164 *
V6           1.93053    1.77362   1.088   0.2925
V7          -1.07030    2.63897  -0.406   0.6904
V8          -0.30201    4.14240  -0.073   0.9428
V9          -0.07199    0.09188  -0.783   0.4448
V10          1.02264    0.70397   1.453   0.1656
V11          1.33991    2.71736   0.493   0.6286
V12          2.78686    2.49657   1.116   0.2808
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.038 on 16 degrees of freedom
Multiple R-squared:  0.9518,    Adjusted R-squared:  0.9186
F-statistic: 28.72 on 11 and 16 DF,  p-value: 1.942e-08
```

Z výsledků dostaneme pouze jednu statisticky významnou proměnnou velikost obytného prostoru (*V5*), tedy její p-hodnota vychází menší než 0,05. To znamená, že všechny zbývající proměnné můžeme z modelu vyloučit a bude nám stačit pouze jednoduchý lineární regresní model tak, jak je popsaný v kapitole 1.2.2, ve kterém bude závisle proměnná prodejní cena (*V13*) na nezávisle proměnná velikost obytného prostoru (*V5*). Provedeme tedy jednoduchou lineární regresi následujícím příkazem

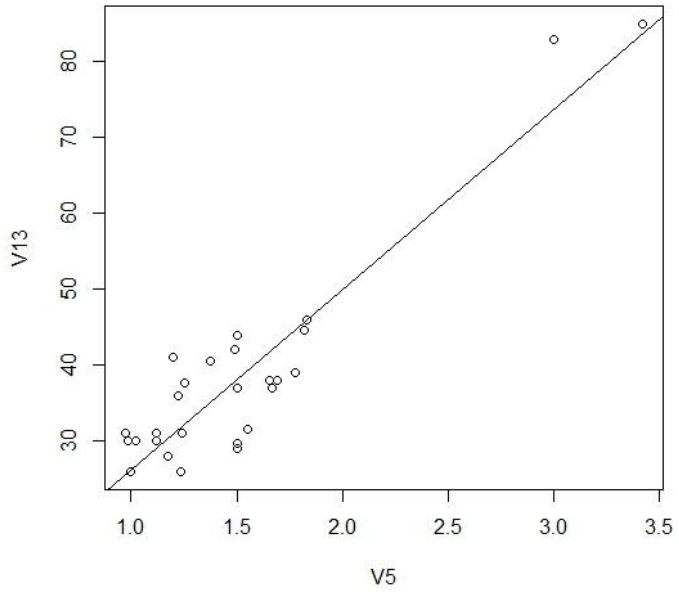
```
summary(lm(V13~V5))
```

```
Call:
lm(formula = V13 ~ V5)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0028 -4.5994  0.0073  4.3728 10.2660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.154      3.155   0.683    0.501
V5            23.817      1.966  12.112 3.41e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.598 on 26 degrees of freedom
Multiple R-squared:  0.8494,    Adjusted R-squared:  0.8436
F-statistic: 146.7 on 1 and 26 DF,  p-value: 3.414e-12
```

Z této části výsledků nás zajímá hlavně koeficient determinace R^2 = 84,94 % (*Multiple R-squared*). Koeficient determinace je v tomto případě statisticky velmi významný. Model tedy lze považovat za dostatečně kvalitní. Dále je vidět, že proměnná *V5* je statisticky významná, p-hodnota je rovna $3{,}41 \cdot 10^{-12}$. Matici plánu sestavíme tak, že první sloupec zaplní jednotkový vektor a druhý sloupec jednotlivá pozorování. V programu můžeme vypočíst hodnotu vektoru $\boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$ podle odvozených vzorců. Výsledkem je vektor $\begin{pmatrix} 2{,}154 \\ 23{,}817 \end{pmatrix}$. Hledaná rovnice nejlepšího modelu je tedy: $V13 = 2{,}154 + 23{,}817 \cdot V5 + \varepsilon$

V případě jednoduché regrese můžeme znázornit graf s proloženou regresní přímkou příkazem

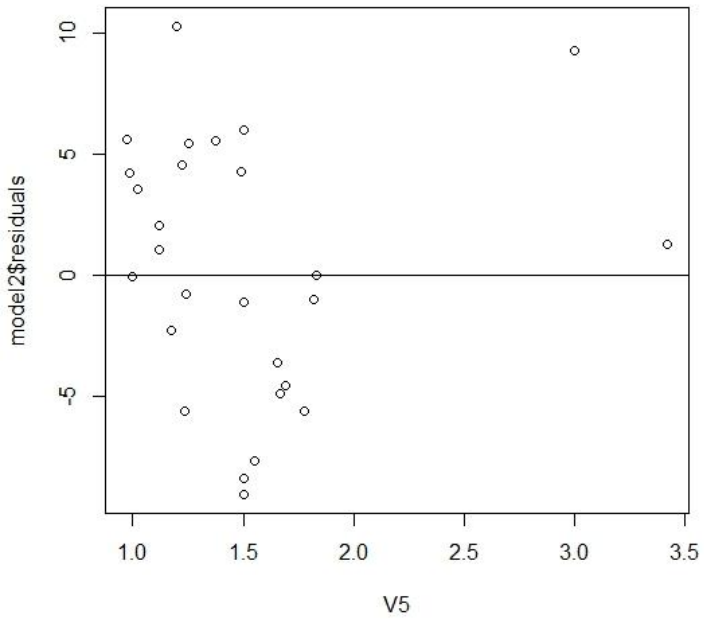
```
plot(V5,V13), abline(model2)
```



Obrázek č. 18: Prodejní cena domu v závislosti na velikosti obytného prostoru s proloženou regresní přímkou

Vykreslením dat na obrázku č. 18 můžeme vidět, že v datech se nachází dvě odlehlá pozorování. Pozorováním čtyř diagnostických grafů se toho pravděpodobně potvrdí. Statistický software nám také umožňuje znázornit graf standardizovaných reziduí proti nezávisle proměnné. Tento graf vytvoříme příkazem

```
plot(V5, model2$residuals), abline(h=0)
```



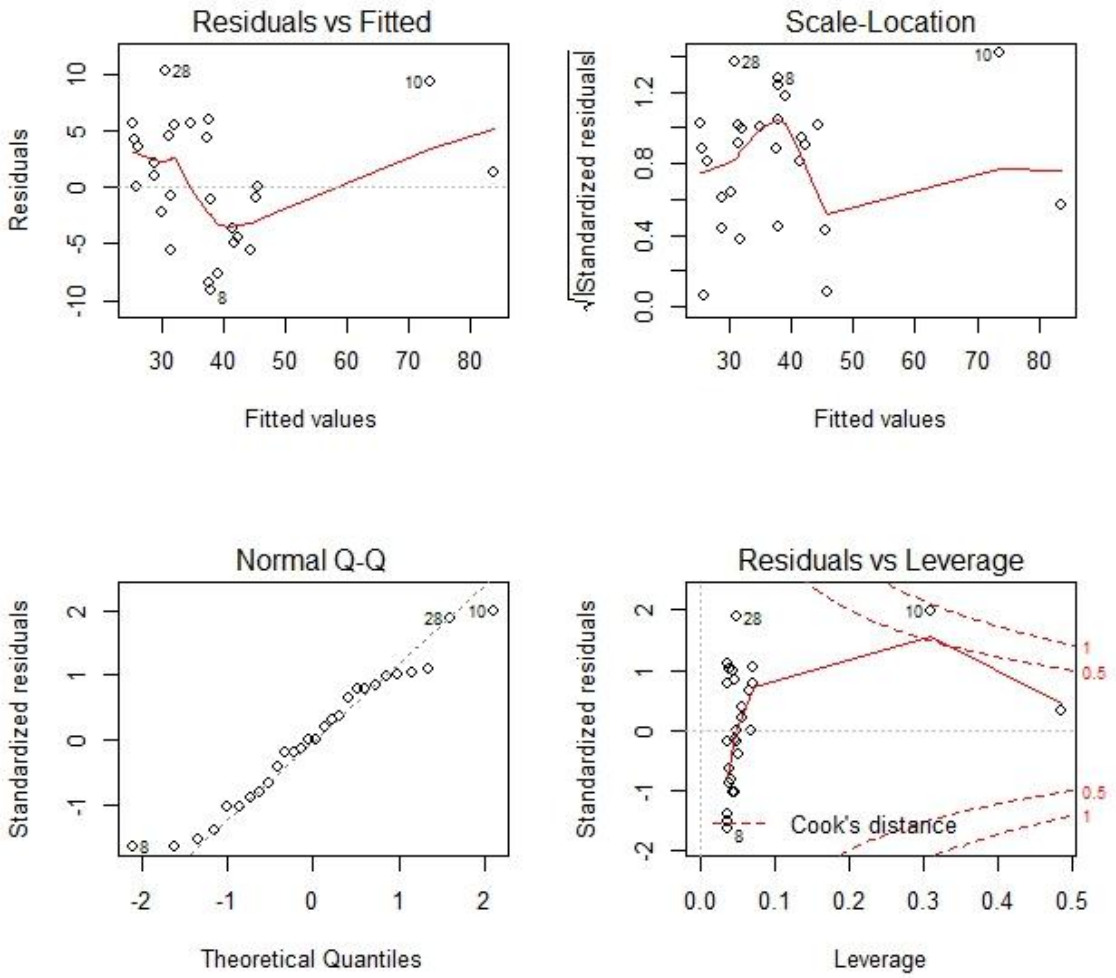
Obrázek č. 19: Graf standardizovaných reziduí proti nezávislé proměnné, vlastní zpracování v R

Z grafu na obrázku č. 19 vidíme, že rezidua jsou náhodně rozmístěna kolem nuly, systematicky se nezvyšují ani nesnižují.

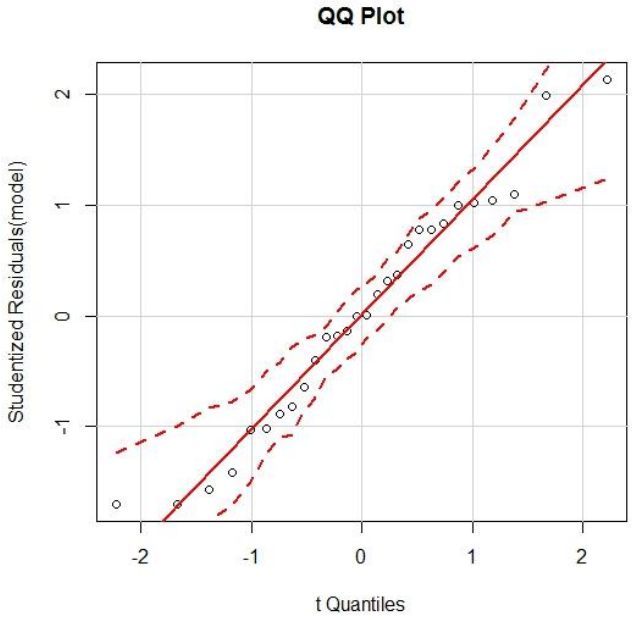
Dalším výstupem, který provedeme, je grafická analýza reziduí. Následující grafické okno se čtyřmi grafy získáme pomocí příkazů

```
layout(matrix(c(1,2,3,4),2,2)), plot (model2)
```

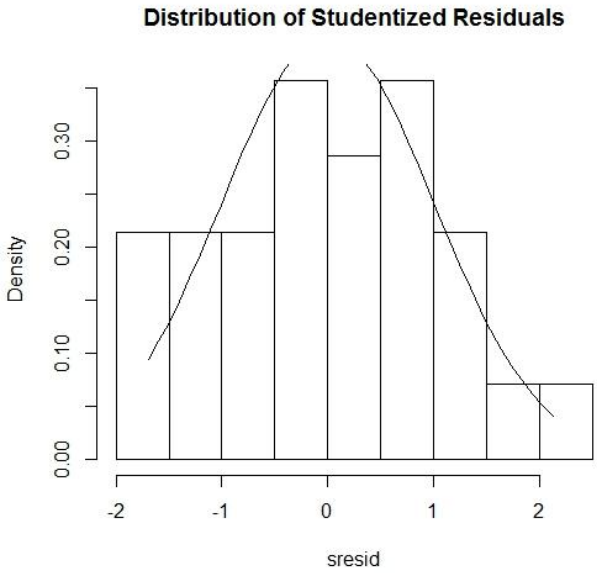
Na prvním grafu vidíme mírné předpokládané zkreslení, v grafu je lehký náznak vzoru, ve správně zvoleném modelu by neměly být vidět zjevné vzory. To lze vysvětlit tím, že v našem příkladě máme celkem malý počet pozorování a to může zkreslit výsledky. Normální Q-Q graf navazuje na diagonálu celkem pěkně, dá se říct, že rezidua mají normální rozdělení. O tom se můžeme přesvědčit na obrázku č. 21, kde je vykreslen samostatně Q-Q graf s pásmem spolehlivosti a také na histogramu standardizovaných reziduí, který je vyobrazen na obrázku č. 22. Podle mého názoru jsou tam dvě odlehlá pozorování (outliers).



Obrázek č. 20: Grafické znázornění reziduí, vlastní zpracování v R



Obrázek č. 21: Q-Q graf, vlastní zpracování v R



Obrázek č. 22: Histogram standardizovaných reziduí, vlastní zpracování v R

Závěr

Regresní analýza je termín, který se vztahuje na širokou škálu metod ve statistice. Ve své práci jsem se zabývala především regresními modely, hodnocením kvality těchto modelů pomocí různých statistických testů a jako poslední také regresní diagnostikou, kde jsem sledovala splnění předpokladů regresního modelu. Nejvíce jsem se věnovala tomu, jak funguje analýza reziduí. V praktické části jsem aplikovala regresní analýzu na dva různé soubory dat z oblasti financí pomocí R softwaru. První soubor dat byl zaměřen na hledání závislosti příjmů obyvatel na ostatních proměnných, ve druhém bylo ukázáno s čím souvisí prodejní cena domů.

Cílem práce bylo podat základní informace o regresní analýze, vyložit přístup k řešení problému a seznámit s nezbytným teoretickým základem. Na první pohled se regresní analýza zdá jako velmi jednoduchá statistická metoda, která s příslušným softwarem dává snadno a lehce cenné poznatky. Praktická aplikace regresní analýzy přináší mnohé problémy. V předchozích příkladech lze najít problém jiného než normálního rozdělení reziduí, problém multikolinearity, problém volby množiny regresorů nebo problém korelace reziduí. Při podrobnějším zkoumání dat jsem zjistila, že je náročné interpretovat výsledky a je potřeba konzultovat vybraná data s odborníkem v oboru, z něhož data pocházejí.

Pro uspokojivé výsledky je tedy nutno znát a tedy i použít podstatně více metod, než jen ty, které jsou v práci uvedeny. I na příklady, které jsou v práci uvedeny, by bylo možno aplikovat složitější postupy, které by možná přinesly přesnější výsledky. Regresní analýza přesto zůstává nejvyužívanější statistickou metodou aplikovanou v mnoha různých oborech.

Literatura

[1] ANDĚL, Jiří. *Statistické metody*. Druhé vydání. Praha 2: MATFYZPRESS, 1998. ISBN 80-85863-27-8.

[2] *Dr. John Rasp's Statistics Website: Business Statistics* [online]. [cit. 2016-04-12]. Dostupné z: http://www2.stetson.edu/~jrasp/stat201/resources/bestcity/best_cities_1.htm

[3] *Dr. John Rasp's Statistics Website: Business Statistics: The Best Cities to Live in Florida* [online]. [cit. 2016-04-12]. Dostupné z: http://www2.stetson.edu/~jrasp/stat201/resources/bestcity/best_cities_4.htm

[4] HEBÁK, Petr, Jiří HUSTOPECKÝ a Iva MALÁ. *Vícerozměrné statistické metody [2]*. Praha 4: INFORMATORIUM, spol. s. r. o., 2005. ISBN 80-7333-036-9.

[5] HENDL, Jan. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Druhé vydání. Praha: Portál, 2006. ISBN 80-7367-123-9

[6] HINDLS, Richard. *Statistika pro ekonomy*. Osmé vydání. Praha: Professional Publishing, 2007. ISBN 8086946436.

[7] *Quick-R accessing the power of R: Multiple (Linear) Regression* [online]. 2016 [cit. 2016-03-20]. Dostupné z: http://www.statmethods.net/stats/regression.html

[8] *R project* [online]. [cit. 2016-02-25]. Dostupné z: https://www.r-project.org/about.html

[9] *Regression Methods* [online]. [cit. 2016-02-25]. Dostupné z: https://onlinecourses.science.psu.edu/stat501/

[10] *REGRESSION: Linear Regression Datasets* [online]. [cit. 2016-04-12]. Dostupné z: http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html

[11] ŘEZANKOVA, Hana; MAREK, Luboš; VRABEC, Michal. *IASTAT*[online]. 2001 [cit. 2016-03-16]. *Interaktivní učebnice statistiky*. Dostupné z: http://iastat.vse.cz/

[12] SVATOŠOVÁ, Libuše a Bohumil KÁBA. *Statistické metody I*. 1. vydání. Praha: Česká zemědělská univerzita v Praze, 2013. ISBN 978-80-213-1672-0.

[13] ZVÁRA, Karel. *Regrese*. První vydání. Praha: MATFYZPRESS, 2008. ISBN 978-80-7378-041-8.