EKONOMICKÁ UNIVERZITA V BRATISLAVE

FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103003/B/2023/421000353285

Regresné modely v programovacom jazyku R

Bakalárska práca

Bratislava 2023 **Roman Kliman**

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Regresné modely v programovacom jazyku R

Bakalárska práca

Študijný program: Data science v ekonómii
Študijný odbor: Ekonómia a manažment
Školiace pracovisko: Katedra štatistiky
Vedúci záverečnej práce: Ing. Patrik Mihalech

Bratislava 2023 **Roman Kliman**

Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Roman Kliman
Študijný program:	data science v ekonómii (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor:	ekonómia a manažment
Typ záverečnej práce:	Bakalárska záverečná práca
Jazyk záverečnej práce:	slovenský
Sekundárny jazyk:	anglický

Názov: Regresné modely v programovacom jazyku R

Anotácia: Regresné modely patria medzi najčastejšie používané modely pri štatistickej analýze. Jedným z najsilnejších nástrojov na tvorbu regresných modelov je voľne dostupný programovací jazyk R. Cieľom bakalárskej práce bude tvorba regresného modelu zo zvoleného dátového súboru, overenie splnenia predpokladov regresného modelu a využitie regresných modelov pri prediktívnej analýze.

Vedúci:	Ing. Patrik Mihalech
Katedra:	KŠ FHI - Katedra štatistiky
Vedúci katedry:	doc. Ing. Mária Vojtková, PhD.
Dátum zadania:	28.03.2022

Dátum schválenia: 19.04.2023

prof. Mgr. Juraj Pekár, PhD.
osoba zodpovedná za realizáciu študijného programu

Poďakovanie

Rád by som sa touto cestou poďakoval Ing. Patrikovi Mihalechovi za jeho nenahraditeľné a cenné rady, ktoré mi poskytol počas vypracovania mojej bakalárskej práce a v neposlednom rade tiež aj za jeho ochotu a trpezlivosť.

Bratislava, 11.5.2023

Roman Kliman

Abstrakt

KLIMAN, Roman: *Regresné modely v programovacom jazyku R.* – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra štatistiky. – Vedúci záverečnej práce: Ing. Patrik Mihalech Bratislava: FHI EU, 2023, 56 s.

Bakalárska práca je vypracovaná na tému: *Regresné modely v programovacom jazyku R* a je rozdelená do troch kapitol. Regresné modely majú svoje nezastupiteľné miesto v štatistickej analýze a tiež sa vyznačujú tým, že majú široké spektrum využitia.

Cieľom našej bakalárskej práce bolo priblížiť využitie regresných modelov v programovacom jazyku R, či už z hľadiska historického, teda ich vývoja ale aj z hľadiska využitia v praxi. V prvej kapitole sa zameriavame na históriu jazyka R, jeho súčasný stav a balíky potrebné pre regresnú analýzu. Tiež sme si priblížili jeho výhody a nevýhody v porovnaní s inými štatistickými balíkmi. V druhej kapitole sme sa snažili podobnejšie popísať, čo je podstatou regresnej analýzy a bližšie sa venujeme regresnému modelu, ktorého vlastnosti vysvetľujeme prostredníctvom lineárneho regresného modelu. Podrobnejšie sme si vysvetlili niektoré štatistické techniky. V tretej kapitole sme sa snažili pretaviť teoretické poznatky do praktických príkladov venovali sme sa skúmaniu vlastnosti lineárneho regresného modelu na konkrétnych údajoch pomocou jazyka R v programovacom prostredí RStudio.

Kľúčové slová: jazyk R, RStudio, dáta, regresný model, metóda najmenších štvorcov

Abstract

KLIMAN Roman: Regression models in the programming language R. – University of Economics in Bratislava. Faculty of Economic Informatics; Department of Statistics. Supervisor of the final thesis: Ing. Patrik Mihalech Bratislava: FHI EU, 2023, 56 p.

The bachelor thesis is developed on the topic: Regression models in the R programming language and is divided into three chapters. Regression models have their irreplaceable place in statistical analysis and are also characterized by the fact that they have a wide range of uses.

The goal of our bachelor's thesis was to approach the use of regression models in the programming language R, both from the historical point of view, i.e. their development, but also from the point of view of use in practice. In the first chapter, we focus on the history of the R language, its current state, and the packages needed for regression analysis. We also zoomed in on its pros and cons compared to other stats packages. In the second chapter, we tried to describe the essence of regression analysis in a similar way, and we take a closer look at the regression model, the properties of which we explain through the linear regression model. We explained some statistical techniques in more detail. In the third chapter, we tried to transform theoretical knowledge into practical examples, we devoted ourselves to investigating the properties of the linear regression model on specific data using the R language in the RStudio programming environment.

Keywords: language R, RStudio, data, regression model, least squares method

Obsah

Zoznam obrázkov a tabuliek

Úvod

Regresná analýza sa zaoberá štatistickým modelovaním vzťahov medzi jednou alebo viacerými nezávislými premennými a jednou alebo viacerými závislými premennými. Cieľom regresnej analýzy je odhadnúť vzťah medzi premennými a použiť tento model na predpovedanie hodnôt závislej premennej na základe hodnôt nezávislých premenných. Existujú rôzne typy regresných analýz, ale najčastejšie používaná je lineárna regresia. Lineárna regresia sa používa, keď je závislá premenná spojitá a výsledný model je lineárny vzhľadom k parametrom. Tento model sa často používa v ekonómii, marketingu a v rôznych vedeckých disciplínach, kde sa snažíme odhadnúť vzťah medzi dvoma premennými. Jedným z najčastejšie používaných jazykov na vytváranie a skúmanie regresných modelov je voľne dostupný jazyk R. V jazyku R je k dispozícii množstvo balíkov a funkcií, ktoré nám umožňujú vytvárať a vizualizovať práve tieto modely. Na prácu s jazykom R sa používa programovacie prostredie RStudio, ktoré umožňuje používateľom prehľadne organizovať svoje projekty a súbory vrátane importu a exportu dát.

V prvej kapitole bakalárskej práce sa zaoberáme históriou jazyka R, jazykom S, ktorý je jeho predchodcom a filozofiou jazyka S, ktorá tvorí základ dnešného jazyka R. Pokračovaním kapitoly je súčasný stav jazyka R, jeho základné vlastnosti, výhody a nevýhody používania jazyka R. Na lepšiu prácu s jazykom R sa používa programovacie prostredie RStudio. Jazyk R a Rstudio sú voľne dostupné a nie je potrebné si kupovať licenciu na ich používanie. Na zjednodušenie práce s jazykom R sa v prostredí RStudio využíva množstvo dostupných balíkov (*packages*), ktoré si cez RStuido vieme nainštalovať. Záver prvej kapitoly tvorí načítanie samotných balíkov v prostredí RStudio a najčastejšie používaných balíkov R potrebných pre regresnú analýzu.

V druhej kapitole bakalárskej práce sa venujeme cieľom a metodike práce. Metodika práce je rozdelená na dve časti. V prvej časti sa zameriavame na objasnenie regresnej analýzy. V druhej časti sa zameriavame na regresný model a bližšie sa venujeme lineárnemu regresnému modelu a jeho aspektom.

Praktická časť bakalárskej práce je venovaná prepájaniu lineárneho regresného modelu s jazykom R v programovacom prostredí RStudio. V praktickej časti používame konkrétne dátové údaje na ktorých testujeme vlastnosti lineárneho regresného modelu.

1 Súčasný stav riešenej problematiky

1.1 História jazyka R

Z hľadiska historického vývoja je programovací jazyk R dialektom jazyka S. Ten neprišiel spomedzi tradičných programovacích jazykov. Cieľom jeho autorov bolo vymyslieť, ako uľahčiť analýzu údajov. Kľúčovým bodom tu bol prechod od používateľa ku vývojárovi. Jazyk R si zachovalo pôvodnú filozofiu jazyka S, teda poskytnúť jednak interaktivitu pri práci a jednak možnosť vývoja nových nástrojov.[1]

1.2 Jazyk S

S je jazyk, ktorý vyvinul John Chambers a ďalší v starých Bell Telephone Laboratories, pôvodne súčasťou AT&T Corp. S bol spustený v roku 1976 ako prostredie internej štatistickej analýzy – pôvodne implementované ako knižnice Fortran. Skoré verzie jazyka neobsahovali ani funkcie pre štatistické modelovanie.

V roku 1988 bol systém prepísaný do C a začal sa podobať systému, ktorý máme dnes (toto bola verzia jazyka 3). Kniha Statistical Models in S od Chambersa a Hastieho dokumentuje funkčnosť štatistickej analýzy. Verzia 4, jazyka S bola vydaná v roku 1998 a je to verzia, ktorú používame v súčasnosti. Kniha Programming with Data od Johna Chambersa dokumentuje túto verziu jazyka.

Od začiatku 90. rokov sa život jazyka S uberal dosť kľukatou cestou. V roku 1993 spoločnosť Bell Labs udelila spoločnosti StatSci (neskôr Insightful Corp.) exkluzívnu licenciu na vývoj a predaj jazyka S. V roku 2004 Insightful kúpil jazyk S od spoločnosti Lucent za 2 milióny dolárov. V roku 2006 Alcatel kúpil Lucent Technologies a teraz sa nazýva Alcatel-Lucent.

Insightful predal svoju implementáciu jazyka S pod produktovým názvom S-PLUS a na jeho vrchole postavil množstvo efektných funkcií (väčšinou GUI) – teda „PLUS“.

V roku 2008 získala spoločnosť Insightful spoločnosť TIBCO za 25 miliónov dolárov. Spoločnosť TIBCO je súčasným vlastníkom jazyka S a je jeho výhradným vývojárom.

Základy samotného jazyka S sa od vydania knihy Progamming with Data od Johna Chambersa v roku 1998 dramaticky nezmenili. V roku 1998 získal jazyk S cenu Asociácie

[1] Microsoft R Application Network. A (Brief) History of R. [online]. [cit. 2023-05-12]. Dostupné na: < https://mran.microsoft.com/documents/what-is-r#rhistory>

pre počítačové stroje za softvérový systém, čo je vysoko prestížne ocenenie v oblasti informatiky.[2]

1.2.1 Filozofia jazyka S

Všeobecnú filozofiu jazyka S je dôležité pochopiť pre používateľov S a R, pretože pripravuje zázemie pre návrh samotného jazyka, čo mnohí starší používatelia programu považujú za trochu zvláštne a mätúce. Predovšetkým je dôležité si uvedomiť, že jazyk S mal svoje korene v analýze údajov a nepochádzal z prostredia tradičného programovacieho jazyka. Jeho vynálezcovia sa zamerali na to, ako uľahčiť analýzu údajov najprv pre seba a v neposlednom rade aj pre ostatných.

V Stages in the Evolution of S John Chambers píše:

„*Chceli sme, aby používatelia mohli začať v interaktívnom prostredí, kde sa vedome nepovažujú za programovanie. Následne po vyjasnení ich potrieb sa ich sofistikovanosť sa zvýšila, mali by byť schopní postupne skĺznuť do programovania, keď sa jazykové a systémové aspekty stanú dôležitejšími.*“[3]

Kľúčovou časťou tu bolo prepojenie od používateľa k vývojárovi. Chceli vytvoriť jazyk, ktorý by mohol ľahko slúžiť obom „ľudom“. Z technického hľadiska potrebovali vytvoriť jazyk, ktorý by bol vhodný na interaktívnu analýzu údajov, ako aj na písanie dlhších programov.[4]

[2] PENG, D. *Roger*. *R Programming for Data Science*. 5.Vydanie. Vydavateľstvo LULU, 2016. 194s. ISBN 978-1365056826.
[3] Jasonheppler. BootcampR: An Introduction to R. [online]. [cit. 2023-04-18]. Dostupné na: https://jasonheppler.org/courses/bootcampr.2020/reading/02-reading/
[4] Jasonheppler. BootcampR: An Introduction to R. [online]. [cit. 2023-04-18]. Dostupné na: https://jasonheppler.org/courses/bootcampr.2020/reading/02-reading/

1.3 Jazyk R

Jazyk R sa začal používať pomerne dosť po vyvinutí jazyka S. Jedným z kľúčových obmedzení jazyka S bolo, že bol dostupný iba v komerčnom balíku S-PLUS. V roku 1991 vytvorili jazyk R Ross Ihaka a Robert Gentleman na Katedre štatistiky Univerzity v Aucklande. V roku 1993 bolo prvé predstavenie jazyka R verejnosti. Skúsenosti Rossa a Roberta s vývojom R sú zdokumentované v článku z roku 1996 v časopise Journal of Computational and Graphical Statistics:[5]

V roku 1995 Martin Mächler významne prispel tým, že presvedčil Rossa a Roberta, aby použili GNU General Public License na vytvorenie R slobodného softvéru. To bolo kritické, pretože to umožnilo, aby zdrojový kód celého systému R bol prístupný každému, kto si ho chcel vyskúšať (slobodnému softvéru sa budeme bližšie venovať v ďalšej časti práce).[6]

V roku 1996 bol vytvorený verejný mailing list (R-help a R-devel zoznamy) a v roku 1997 bola vytvorená R Core Group, ktorá obsahovala niektorých ľudí spojených s jazykom S a S-PLUS. V súčasnosti hlavná skupina riadi zdrojový kód pre R a je schopná kontrolovať zmeny v hlavnom zdrojovom strome R. Nakoniec bola v roku 2000 pre verejnosť uvoľnená verzia R 1.0.0.[7]

1.3.1 Slobodný softvér

Hlavnou výhodou, ktorú má R oproti mnohým iným štatistickým balíkom, je, že je zadarmo. Autorské práva na primárny zdrojový kód pre jazyk R vlastní R Foundation a sú publikované pod GNU General Public License verzie 2.0.

Podľa Free Software Foundation vám slobodný softvér poskytuje nasledujúce štyri slobody:

- Sloboda spustiť program na akýkoľvek účel.
- Sloboda študovať, ako program funguje a prispôsobiť ho svojim potrebám. Predpokladom je prístup k zdrojovému kódu.
- Sloboda redistribúcie kópií, aby ste mohli pomôcť svojmu blížnemu.

[5] Ross Ihaka a Robert Gentleman. R: Jazyk pre analýzu údajov a grafiku. Journal of Computational and Graphical Statistics, 5(3):299–314, 1996.
[6] Ross Ihaka a Robert Gentleman. R: Jazyk pre analýzu údajov a grafiku. Journal of Computational and Graphical Statistics, 5(3):299–314, 1996.
[7] PENG, D. *Roger*. *R Programming for Data Science*. 5.Vydanie. Vydavateľstvo LULU, 2016. 194s. ISBN 978-1365056826.

- Sloboda vylepšovať program a zverejňovať svoje vylepšenia tak, aby to robila celá komunita[8]

1.3.2 Základné vlastnosti R

Veľkou výhodou jazyka R je, že jeho syntax sa veľmi podobá jazyku S, čo používateľom S-PLUS uľahčilo prepínanie. Zatiaľ čo syntax R je takmer identická so syntaxou S, sémantika R, hoci je povrchne podobná S, je úplne odlišná. V skutočnosti je R technicky oveľa bližšie k jazyku Scheme ako k pôvodnému jazyku S, pokiaľ ide o to, ako R funguje pod kapotou.[9]

Dnes R beží na takmer každej štandardnej počítačovej platforme a operačnom systéme. Jeho open source povaha znamená, že ktokoľvek môže voľne prispôsobiť softvér akejkoľvek platforme, ktorú si vyberie. V skutočnosti sa uvádza, že R beží na moderných tabletoch, telefónoch, PDA a herných konzolách.[10]

Populárnou vlastnosťou, ktorú jazyk R má sú časté vydania, ktoré jazyk R zdieľa s mnohými populárnymi open source projektmi. V súčasnosti existuje veľké ročné vydanie, zvyčajne v októbri, kde sú začlenené a zverejnené hlavné nové funkcie. V priebehu roka sú vždy podľa potreby vydávané menšie opravy chýb. Časté vydania a pravidelný cyklus vydávania naznačujú aktívny vývoj softvéru a zaisťujú, že chyby budú včas odstraňované. Samozrejme, zatiaľ čo hlavní vývojári ovládajú primárny zdrojový strom pre R, mnoho ľudí na celom svete prispieva vo forme nových funkcií, opráv chýb alebo oboch.[11]

Ďalšou kľúčovou výhodou, ktorú má R oproti mnohým iným štatistickým balíkom, sú jeho grafické možnosti. Schopnosť R vytvárať grafiku „publikačnej kvality“ existuje od úplného začiatku a vo všeobecnosti bola lepšia ako konkurenčné balíčky. Dnes, keď je k dispozícii oveľa viac vizualizačných balíkov ako predtým, tento trend pokračuje. Základný grafický systém R umožňuje veľmi jemnú kontrolu nad v podstate každým aspektom grafu. Iné novšie grafické systémy ako napríklad *ggplot2* umožňujú komplexné a sofistikované vizualizácie viacrozmerných údajov.[12]

---

[8] PENG, D. *Roger. R Programming for Data Science*. 5.Vydanie. Vydavateľstvo LULU, 2016. 194s. ISBN 978-1365056826.
[9]Tutorialspoint. R-Overview. [online]. [cit. 2023-04-18]. Dostupné na: https://www.tutorialspoint.com/r/r_overview.htm
[10] Tutorialspoint. R-Overview. [online]. [cit. 2023-04-18]. Dostupné na: https://www.tutorialspoint.com/r/r_overview.htm
[11] TechVidvan. 15 Features of R Programming you can’t afford to overlook. [online]. [cit. 2023-04-18]. Dostupné na: https://techvidvan.com/tutorials/r-features/
[12] TechVidvan. 15 Features of R Programming you can’t afford to overlook. [online]. [cit. 2023-04-18]. Dostupné na: https://techvidvan.com/tutorials/r-features/

Jazyk R si zachoval pôvodnú filozofiu jazyka S, ktorá spočíva v tom, že poskytuje jazyk, ktorý je užitočný pre interaktívnu prácu, ale zároveň obsahuje výkonný programovací jazyk na vývoj nových nástrojov. To umožňuje používateľovi, ktorý používa existujúce nástroje a aplikuje ich na dáta, stať sa pomaly, ale isto vývojárom, ktorý vytvára nové nástroje.[13]

Napokon, jedna z výhod používania R nemá nič spoločné so samotným jazykom, ale skôr s aktívnou a živou používateľskou komunitou. V mnohých ohľadoch je jazyk úspešný, pretože vytvára platformu, s ktorou môže veľa ľudí vytvárať nové veci. R je platforma a tisíce ľudí na celom svete sa spojili, aby prispeli k jazyku R, vyvinuli balíčky (packages) a navzájom si pomáhali používať R pre všetky druhy aplikácií.[14]

1.3.3 Výhody R

Bezplatné a open source:

Programovací jazyk R je open source a je vydaný pod licenciou General Public License (GNU). To znamená, že môžete využívať všetky funkcie R zadarmo bez akýchkoľvek obmedzení alebo licenčných požiadaviek. Keďže R je open source, každý je vítaný, aby mohol prispieť k projektu, a keďže je voľne dostupný, komunita open source ľahko zistí a opraví chyby[15].

Popularita:

Programovací jazyk R sa najlepšie umiestnil na 8. mieste podľa rebríčku TIOBE Index z augusta 2020 a najnovšie zdroje z apríla 2023 uvádzajú 16. miesto.[16] Podľa edX je to druhý najpopulárnejší programovací jazyk pre dátovú vedu hneď za programovacím jazykom Python. Popularita R tiež znamená, že na platformách ako Stackoverflow existuje rozsiahla komunitná podpora. R má tiež podrobnú online dokumentáciu, do ktorej môžu používatelia R konzultovať pomoc.[17]

---

[13] TechVidvan. 15 Features of R Programming you can't afford to overlook. [online]. [cit. 2023-04-18]. Dostupné na: https://techvidvan.com/tutorials/r-features/

[14] TechVidvan. 15 Features of R Programming you can't afford to overlook. [online]. [cit. 2023-04-18]. Dostupné na: https://techvidvan.com/tutorials/r-features/

[15] DataFlair. Pros and Cons of R Programming Language. [online]. [cit. 2023-04-19]. Dostupné na:https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/

[16] TIOBE. Index for April 2023. [online]. [cit. 2023-04-18]. Dostupné na: https://www.tiobe.com/tiobe-index/

[17] DataFlair. Pros and Cons of R Programming Language. [online]. [cit. 2023-04-19]. Dostupné na:https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/

Vysokokvalitná vizualizácia:

Programovací jazyk R je známy kvalitnými vizualizáciami. R uľahčuje kvalitné vykresľovanie a vytváranie grafov. Populárne knižnice ako napríklad `ggplot2` obsahujú estetické a vizuálne dobre vyzerajúce grafy, ktoré odlišujú R od iných programovacích jazykov.[18]

Široké uplatnenie v akademickej aj priemyselnej sfére:

Programovací jazyk R je dôveryhodný a široko používaný v akademickej komunite na výskum. R je stále viac používaný vládnymi agentúrami, sociálnymi médiami, telekomunikáciami, finančnými spoločnosťami, elektronickými obchodmi, výrobnými a farmaceutickými spoločnosťami. Medzi najlepšie spoločnosti, ktoré používajú R, patria Amazon, Google, ANZ Bank, Twitter, LinkedIn, Facebook a mnohé ďalšie. Dobré zvládnutie programovacieho jazyka R otvára všetky druhy príležitostí v akademickej sfére a priemysle[19].

1.3.4 Nevýhody R

Žiadny programovací jazyk ani systém štatistickej analýzy nie sú dokonalé. R má určite množstvo nevýhod. Pre začiatok, R je v podstate založený na takmer 50 rokov starej technológii, ktorá sa vracia k pôvodnému systému S vyvinutému v Bell Labs. Podpora dynamickej alebo 3-D grafiky bola pôvodne málo zabudovaná (ale veci sa od „starých čias“ výrazne zlepšili).[20]

Ďalším bežne uvádzaným obmedzením R je, že objekty musia byť vo všeobecnosti uložené vo fyzickej pamäti. Je to čiastočne kvôli pravidlám určovania rozsahu jazyka. Na riešenie tohto problému však došlo k niekoľkým pokrokom, a to ako v jadre R, tak aj v množstve balíkov vyvinutých prispievateľmi[21].

[18] ROOT.CZ. Programovací jazyk R: úvodní informace. [online]. [cit. 2023-04-20]. Dostupné na: https://www.root.cz/clanky/programovaci-jazyk-r-uvodni-informace/
[19]ROOT.CZ. Programovací jazyk R: úvodní informace. [online]. [cit. 2023-04-20]. Dostupné na: https://www.root.cz/clanky/programovaci-jazyk-r-uvodni-informace/
[20] ROOT.CZ. Programovací jazyk R: úvodní informace. [online]. [cit. 2023-04-20]. Dostupné na: https://www.root.cz/clanky/programovaci-jazyk-r-uvodni-informace/
[21] https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/

1.4 R jazyk a R studio

R je programovací jazyk, ktorý spúšťa výpočty, zatiaľ čo RStudio je integrované vývojové prostredie (IDE), ktoré poskytuje rozhranie pridaním mnohých pohodlných funkcií a nástrojov. Takže tak, ako prístup k rýchlomeru, spätným zrkadlám a navigačnému systému značne uľahčuje jazdu, používanie rozhrania RStudio výrazne uľahčuje aj používanie R.[22]

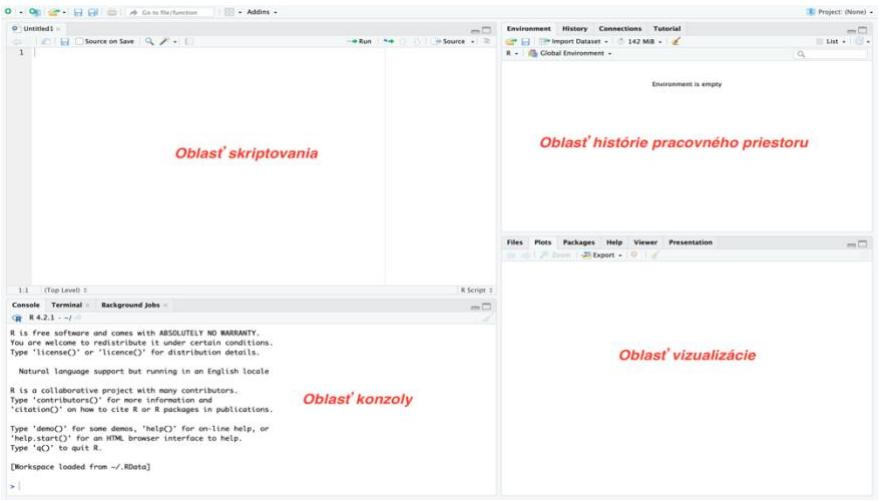
Základné vývojové prostredie R (v operačnom systéme Windows a macOS) tvorí textový editor, pomocou ktorého používateľ píše zdrojový kód v jazyku R, a príkazový riadok (*konzola*), v ktorom je odoslaný kód interpretovaný. Grafické výstupy sú presmerované do samostatných okien. Pre pohodlnú prácu odporúčame použitie integrovaného vývojového prostredia RStudio, ktoré farebne zvýrazňuje syntax, poskytuje nápovedu, sprístupňuje zoznam definovaných objektov, uľahčuje tvorbu dokumentácie a veľa ďalších užitočných nástrojov.[23]

Zdrojový kód sa na interpretáciu do príkazového riadku posiela typicky buď po riadkoch alebo vyznačením časti kódu, a stlačením kombinácie kláves (*Ctrl-R* v základnom prostredí, a *Ctrl-Enter* v prostredí RStudio).

Prostredie sa skladá zo štyroch rôznych oblastí:

- Oblasť skriptovania: V tejto oblasti môžeme otvárať, vytvárať a písať naše skripty
- Oblasť konzoly: Táto zóna je skutočná konzola R, kde sa vykonávajú príkazy
- Oblasť histórie pracovného priestoru: V tejto oblasti môžete nájsť zoznam všetkých objektov vytvorených v pracovnom priestore, kde pracujeme
- Oblasť vizualizácie: V tejto oblasti môžeme jednoducho načítať balíky a otvárať súbory pomocníka R, ale čo je dôležitejšie, môžeme si prezerať grafy[24]

[22] Mercury. Introduction to R and RStudio. [online]. [cit. 2023-04-20]. Dostupné na: http://mercury.webster.edu/aleshunas/R_learning_infrastructure/Introduction_to_R_and_RStudio.html
[23] Bookdown. Data Analysis and Processing with R based on IBIS data. . [online]. [cit. 2023-04-20]. Dostupné na: https://bookdown.org/kdonovan125/ibis_data_analysis_r4/#preface
[24] Mercury. Introduction to R and RStudio. [online]. [cit. 2023-04-20]. Dostupné na: http://mercury.webster.edu/aleshunas/R_learning_infrastructure/Introduction_to_R_and_RStudio.html



Obrázok 1 Prostredie Rstudio

Zdroj: Vlastné spracovanie

1.4.1 Inštalácia jazyka R a Rstudia

Prvá vec, ktorú musíme urobiť, aby sme mohli začať s jazykom R je nainštalovať ho do počítača. R funguje takmer na každej dostupnej platforme vrátane široko dostupných systémov Windows, Mac OS X a Linux.

Klikneme na nižšie uvedený odkaz, ktorý nás presmeruje na oficiálnu stránku pre stiahnutie jazyka R:

https://cran.r-project.org

Po inštalácii jazyka R si nainštalujeme programovacie prostredie Rstudio cez nižšie uvedený odkaz, ktorý nás opäť presmeruje na oficiálnu stránku na stiahnutie:

https://posit.co

1.4.2 Balíky (packages) v jazyku R

Balíky R rozširujú funkčnosť R poskytovaním ďalších funkcií, údajov a dokumentácie. Sú napísané celosvetovou komunitou používateľov R a dajú sa bezplatne stiahnuť z internetu. Sú uložené v adresári s názvom „library“ v prostredí R. Štandardne R nainštaluje sadu balíkov počas inštalácie. Ďalšie balíčky sa pridávajú neskôr, keď sú potrebné na konkrétny účel. Keď spustíme konzolu R, štandardne sú k dispozícii iba

predvolené balíčky. Ostatné balíky, ktoré sú už nainštalované, musia byť explicitne načítané, aby ich mohol použiť program R[25].

1.4.3 Inštalácia balíkov

Existujú dva spôsoby inštalácie balíka R: jednoduchý spôsob a pokročilejší spôsob. Ako príklad uvedieme jeden z najsťahovanejších balíkov *ggplot2*.

Na panely Súbory v RStudio:

1. Klikneme na kartu „Packages“.
2. Klikneme na „Install“ vedľa položky Aktualizovať.
3. Zadáme názov balíka v časti „Packages (násobok oddeľte medzerou alebo čiarkou):“ V tomto prípade zadáme *ggplot2*.
4. Klikneme na „Install“.

Alternatívnym, ale o niečo menej pohodlným spôsobom inštalácie balíka je napísanie `install.packages("ggplot2")` na panely konzoly RStudia a stlačením Return/Enter na klávesnici. Musíme si dať pozor však na to aby okolo názvu balíka boli uvedené úvodzovky[26].

1.4.4 Načítanie balíkov

Pripomeňme si, že po nainštalovaní balíka ho musíme „načítať“. Inými slovami, musíme ho „otvoriť“. Urobíme to pomocou príkazu `library()`. Ak chceme napríklad načítať balík `ggplot2`, spustíme nasledujúci kód na paneli konzoly. Pod pojmom „spustiť nasledujúci kód“ máme na mysli zadanie alebo skopírovanie a následné prilepenie kódu `library (ggplot2)` do panela konzoly a potom stlačenia klávesy Enter.

Jednou veľmi častou chybou, ktorú noví používatelia R robia, keď chcú použiť konkrétne balíky je, že ich zabudnú „načítať“ najskôr pomocou príkazu `library()`. Musíme si pamätať, že pri každom spustení RStudio musíme načítať každý balík, ktorý

[25] Rbasics. Getting used to R, Rstudio and R Markdown. [online]. [cit. 2023-04-20]. Dostupné na: https://rbasics.netlify.app
[26] ISMAY, Chester – KIM, Y. Albert. Statistical Inference via Data Science. 1.Vydanie. Vydavateľstvo CHAPMAN AND HALL/CRC, 2019. 430s. ISBN 978-0367409821.

chceme použiť. Ak balík najprv „nenačítame“, ale pokúsime sa použiť niektorú z jeho funkcií, zobrazí sa chybové hlásenie[27].

1.4.5 Balíky R potrebné pre regresnú analýzu

Balíček R „stats“

R „stats“ je balík, ktorý obsahuje mnoho užitočných funkcií na štatistické výpočty a generovanie náhodných čísel. V balíku sa nachádza nespočetne veľa funkcií a preto spomenieme len tie, ktoré sú najbližšie k regresnej analýze. Toto sú najužitočnejšie funkcie používané v regresnej analýze[28]:

lm: sa používa na prispôsobenie lineárnych modelov. Môže sa použiť na vykonanie lineárnej regresie, jednofaktorovej analýzy rozptylu a analýzy kovariancie.

Summary. lm: vráti súhrn pre lineárne prispôsobenie modelu.

coef: Pomocou tejto funkcie je možné extrahovať koeficienty z objektov vrátených modelovacími funkciami. Koeficienty sú pre to alias.

fitted: Predpokladané hodnoty vypočítané modelom.

formula: poskytuje spôsob extrahovania vzorcov, ktoré boli zahrnuté v iných objektoch.

predict: predpovedá hodnoty na základe objektov lineárneho modelu.

residuals: extrahuje rezíduá modelu z objektov vrátených funkciami modelovania.

confint: počíta intervaly spoľahlivosti pre jeden alebo viacero parametrov v prispôsobenom modeli. Base má metódu pre objekty dediace z triedy lm.

deviance: vráti odchýlku prispôsobeného objektu modelu.

[27] ISMAY, Chester – KIM, Y. Albert. Statistical Inference via Data Science. 1.Vydanie. Vydavateľstvo CHAPMAN AND HALL/CRC, 2019. 430s. ISBN 978-0367409821.
[28] Packt. R packages for regression. [online]. [cit. 2023-05-10]. Dostupné na: na:https://subscription.packtpub.com/book/data/9781788627306/1/ch01lvl1sec18/r-packages-for-regression

lm.influence: funkcia poskytuje základné veličiny používané pri vytváraní širokej škály diagnostiky na kontrolu kvality regresných preložení.

ls.diag: počíta základné štatistiky vrátane štandardných chýb, t-hodnôt a p-hodnôt pre regresné koeficienty.

glm: Funkcia na výpočet zovšeobecnených regresných modelov.[29]

To, čo sme spomenuli, sú len niektoré z mnohých funkcií obsiahnutých v balíku „stats“. Ako vidíme, so zdrojmi, ktoré ponúka tento balík, môžeme zostaviť lineárny regresný model, ako aj GLM (ako je viacnásobná lineárna regresia, polynomická regresia a logistická regresia). Budeme tiež schopní urobiť modelovú diagnostiku, aby sme overili vierohodnosť klasických hypotéz, ktoré sú základom regresného modelu, ale môžeme tiež riešiť lokálne regresné modely s neparametrickým prístupom.[30]

[29] Packt. R packages for regression. [online]. [cit. 2023-05-10]. Dostupné na: na:https://subscription.packtpub.com/book/data/9781788627306/1/ch01lvl1sec18/r-packages-for-regression
[30] Packt. R packages for regression. [online]. [cit. 2023-05-10]. Dostupné na: na:https://subscription.packtpub.com/book/data/9781788627306/1/ch01lvl1sec18/r-packages-for-regression

2 Metodika a ciele práce

2.1 Ciele práce

Hlavným cieľom bakalárskej práce je ukázať prepojenie regresnej analýzy s programovacím jazykom R. Ako ukážku využijeme lineárny regresný model, ktorý budeme programovať v programovacom prostredí RStudio. Pomocou rôznych balíkov a funkcií, ktoré jazyk R má sa budeme snažiť overiť vlastnosti modelu: štatistickú významnosť samotného modelu, overenie významnosti koeficienta β_0 a β_1, grafické znázornenie pomocou funkcií `plot` a `ggplot2`, výpočet rezíduí daného modelu, overenie, či je v modeli prítomná autokorelácia, či sú premenné v modeli lineárne závislé a následne celý model ako celok interpretovať.

2.2 Regresná analýza

Pojem regresia je známy z prác anglického polyhistora Francisa Galtona, ktorý sa v rokoch 1877 až 1885 zaoberal vzťahom medzi výškou otcov a ich synov. Pozorovaním tejto vlastnosti dospel k tomu, že aj keď vysokí otcovia majú nadpriemerne vysokých synov, ich výška sa približuje k priemeru a synovia nízkych otcov už nie sú až takí nízki, pretože ich výška sa približuje k priemernej výške v populácii mužov. Zákonitosť návratu smerom k priemernej výške Galton pôvodne pomenoval *reversion*, no neskôr to pomenoval ako *regression* (spätný krok)[31].

Aplikácie regresnej analýzy zaznamenali rapídny vzostup vďaka širokému využívaniu výpočtovej techniky, či už v psychológii, medicíne, marketingu a pod.
V dnešnej dobe je regresná analýza jednou z najčastejšie využívaných techník štatistiky v manažmente[32].

Regresná analýza predstavuje súhrn štatistických metód a postupov slúžiacich na štúdium vzájomných vzťahov medzi dvoma alebo viacerými premennými, prostredníctvom regresného modelu.[33]

[31] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.
[32] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.
[33] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

2.3 Regresný model

Regresný model je matematický predpis, ktorý zjednodušene charakterizuje vzťahy medzi premennými. Z hľadiska postavenia premenných v regresnom modeli rozlišujeme:

- Vysvetľovanú (závislú) premennú Y. Je to číselná premenná, ktorej závislosť od iných premenných skúmame. Jej pozorovanie skúmame y_i pre i = 1, 2,... n.
- Vysvetľujúce (nezávislé) premenné X_1, X_2,... X_j,... X_k. Sú to premenné u ktorých predpokladáme, že vyvolávajú zmeny závislej premennej a pomocou nich odhadujeme hodnoty závislej premennej.[34]

$$y_i = \underbrace{f(x_{i1}, x_{i2}, \dots, x_{ik};\ \beta_0, \beta_1, \dots, \beta_k)}_{\eta_i} + \varepsilon_i$$

kde: η_i je regresná funkcia,
ε_i je náhodná chyba.

Regresným modelom matematicky opisujeme voľnú (stochastickú) závislosť. Pri tejto závislosti pozorujeme na rôznych štatistických jednotkách pre rovnakú kombináciu hodnôt vysvetľujúcich premenných (X) rôzne hodnoty vysvetľovanej premennej (Y). Táto charakteristická menlivosť alebo variabilita je súčasťou náhodnej zložky (ε) regresného modelu. Regresný model sa teda skladá z dvoch zložiek:

- *Deterministická zložka* (η). Je to regresná funkcia $f(X, \beta)$ s premennými $X_1, X_2, \dots X_k$ a parametrami $\beta_0, \beta_1, \dots \beta_k$. Dá sa povedať, že hodnota regresnej funkcie je stredná hodnota vysvetľovanej premennej Y vzhľadom na hodnoty vysvetľujúcich premenných $X_1, X_2, \dots X_k$.
- *Náhodná zložka* (ε). Táto zložka odráža pôsobenie náhodných vplyvov, ale aj faktorov nezaradených do regresnej funkcie.[35]

[34] LABUDOVÁ, Viera a kol.:Štatistické metódy pre ekonómov a manažérov. Vydavateľstvo WOLTERS KLUWER, 2021. 392s. ISBN978-80-571-0401-8.
[35] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

2.3.1 Lineárny regresný model

Najjednoduchší model párovej regresie je model s lineárnou regresnou funkciou, ktorej grafom je priamka. Takýto model sa nazýva jednoduchý lineárny regresný model.

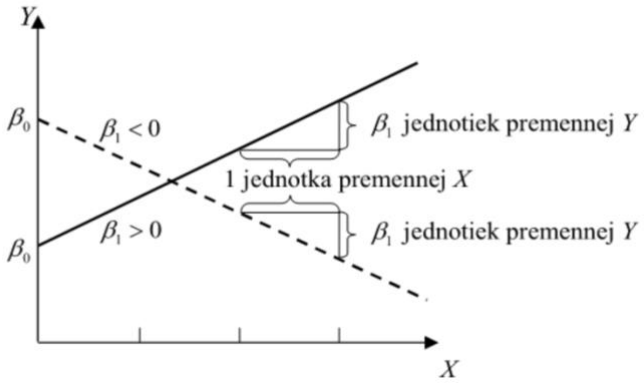
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ , pričom } i = 1, 2, \dots n$$

kde: y_i je i-tá pozorovaná hodnota vysvetľovanej premennej,
β_0, β_1 sú neznáme parametre regresného modelu,
x_i je i-tá hodnota vysvetľujúcej premennej,
ε_i je náhodná chyba i-tého pozorovania
n je počet pozorovaní

Parameter β_0 regresného modelu sa nazýva lokujúca konštanta. Interpretuje sa ako podmienená stredná hodnota závislej premennej Y za predpokladu, že hodnota vysvetľujúcej premennej X sa rovná nule. Avšak lokujúca konštanta niekedy nemá vhodnú interpretáciu a určuje len postavenie regresnej priamky v rovine (priesečník priamky s osou y).

Parameter β_1 sa nazýva regresný koeficient. Je smernicou regresnej priamky a určuje, aký prírastok ($\beta_1 > 0$) alebo úbytok ($\beta_1 < 0$) strednej hodnoty závislej premennej Y zodpovedá jednotkovému prírastku nezávislej premennej X, inak povedané o koľko sa nám zväčší alebo zmenší premenná Y ak sa premenná X zmení o jednu jednotku. Regresný koeficient je nositeľom informácie o priebehu štatistickej závislosti. Ak β_1 nadobúda kladné hodnoty, znamená to, že je medzi premennými X a Y priama lineárna závislosť. A zase naopak, ak regresný koeficient nadobúda záporné hodnoty, tak je prítomná nepriama lineárna závislosť.[36]

[36] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.



Obrázok 2 Regresná priamka $\boldsymbol{\eta_i = \beta_0 + \beta_1 x_i}$

Zdroj: Regresná a korelačná analýza s aplikáciami v softvéri SAS (Šoltés Erik, 2019)

2.3.2 Metóda najmenších štvorcov

Metóda najmenších štvorcov je štatistická technika používaná na nájdenie najlepšie vyhovujúcej čiary alebo krivky pre súbor údajových bodov minimalizovaním súčtu štvorcových rozdielov medzi pozorovanými hodnotami a predpokladanými hodnotami. Cieľom je nájsť takú čiaru alebo krivku, ktorá minimalizuje celkovú vzdialenosť medzi pozorovanými dátovými bodmi a predpovedanými hodnotami, čo sa dosiahne minimalizáciou súčtu štvorcov rezíduí.

Metódu najmenších štvorcov možno použiť na lineárne aj nelineárne modely a je široko používaná v mnohých oblastiach, ako je ekonómia, fyzika, inžinierstvo a financie. Metóda je užitočná najmä pri práci s neúplnými údajmi, pretože poskytuje spôsob, ako odhadnúť základné trendy v údajoch.

Ak predpokladáme lineárnu závislosť medzi premennými X a Y, môžeme použiť metódu lineárnej regresie na určenie rovnice vyrovnávajúcej regresnej priamky, ktorá bude odchýlok najlepšie vystihovať túto závislosť. Táto priamka sa snaží minimalizovať súčet štvorcov medzi pozorovaniami Y a príslušnými predpovedaniami na základe X. Ak je táto priamka správne určená, môžeme použiť jej rovnicu na predpovedanie hodnôt Y na základe zadaných hodnôt X. Je však dôležité mať na pamäti, že pred použitím tejto metódy by sme mali najskôr overiť, či skutočne existuje lineárna závislosť medzi premennými X a Y a že nie sú prítomné žiadne významné odchýlky alebo vplyvy tretích premenných[37].

[37] Zuzana Gibova: Metóda najmenších štvorcov. [online]. [cit. 2023-05-10]. Dostupné na: https://zuzana.gibova.website.tuke.sk/files/kap-3.2.pdf

2.3.3 Vyrovnávajúca regresná priamka

$$\hat{y}_i = b_0 + b_1 x_i$$

kde: $\hat{y}_i$ je i-tá vyrovnaná (očakávaná, teoretická) hodnota závislej premennej Y,
x_i je hodnota nezávislej premennej X pre i-té pozorovanie,
b_0 je bodový odhad parametra β_0,
b_1 je bodový odhad parametra β_1.

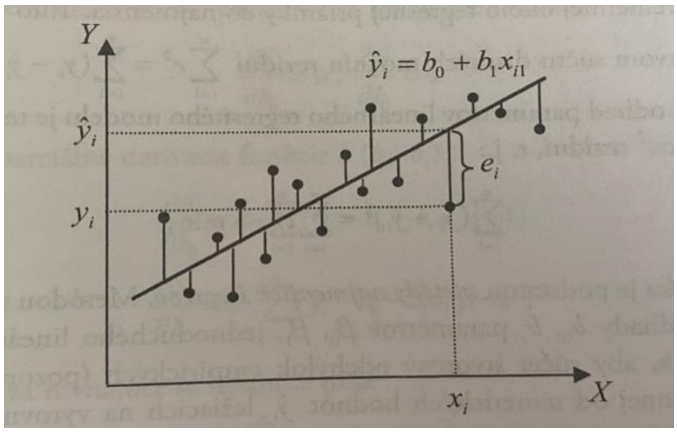
Vyrovnávajúca priamka, ktorú vytvárame pomocou metódy lineárnej regresie, je bodovým odhadom regresnej priamky, ktorá najlepšie vystihuje závislosť medzi premennými X a Y. Rezíduá alebo reziduálne odchýlky sú rozdiely medzi skutočne nameranými hodnotami Y a predpovedanými hodnotami Y na základe vyrovnávajúcej priamky. Tieto odchýlky označujeme ako e_i a sú bodovými odhadmi náhodných chýb $\mathcal{E}_i$ v regresnom modeli. Ich hodnoty môžu poskytnúť dôležité informácie o tom, ako dobre sa náš model prispôsobuje skutočným dátam a či sú nejaké významné odchýlky alebo vplyvy tretích premenných, ktoré by mohli ovplyvniť našu analýzu. Preto je dôležité venovať pozornosť týmto odchýlkam a vyhodnotiť ich, aby sme mohli zabezpečiť správnosť a spoľahlivosť nášho modelu.

Rezíduá v regresnom zapisujeme nasledovne[38]:

$$est\mathcal{E}_i = e_i = y_i - \hat{y}_i$$

Rezíduá pre regresnú priamku sú znázornené na nasledujúcom obrázku.

[38] FBIW.UNIZA. Lineárny regresný model. [online]. [cit. 2023-05-10]. Dostupné na: http://fbiw.uniza.sk/kkm/old/publikacie/ek/ek_kap_3.pdf



Obrázok 3 Vyrovnávajúca priamka s vyznačením rezíduí

Zdroj: Štatistické metódy pre ekonómov a manažérov (Labudová Viera, 2021)

Z geometrického hľadiska sa snažíme získať také odhady parametrov β_0 a β_1, aby sa skutočné hodnoty závislej premennej čo najmenej odchyľovali od vyrovnávajúcej priamky, inak povedané, aby rezíduá boli čo najmenšie. Logické by bolo zabezpečiť nulový súčet rezíduí. Avšak pri hľadaní vhodnej vyrovnávajúcej priamky nemôžeme jednoducho zabezpečiť nulový súčet rezíduí. Keďže niektoré pozorovania sú nad priamkou a iné pod ňou, znamená, že niektoré rezíduá sú kladné a iné sú záporné. Preto by nulový súčet rezíduí znamenal, že súčet týchto kladných a záporných hodnôt by bol nulový, čo nie je nutne optimálne. V skutočnosti sa snažíme minimalizovať súčet štvorcov rezíduí, ktorý je väčšinou bližší k nule, ale nie nutne nulový. Preto sa pri výbere vyrovnávajúcej priamky pomocou metódy najmenších štvorcov snažíme minimalizovať súčet štvorcov rezíduí a zabezpečiť, aby rozdelenie pozorovaní nad a pod priamkou bolo približne rovnomerné. To nám umožní získať regresnú priamku, ktorá čo najlepšie vystihuje závislosť medzi premennými X a Y.

Podmienka $\sum_{i=1}^{n} e_i = 0$ je iba nutnou podmienkou, ale nie je postačujúcou podmienkou na odhad parametrov lineárneho regresného modelu. Opísaný problém sa odstraňuje tak, že sa hľadá taká regresná priamka, pre ktorú je variabilita pozorovaných hodnôt vysvetľovanej premennej okolo regresnej priamky čo najmenšia.[39] Túto variabilitu meriame prostredníctvom súčtu druhých mocnín rezíduí podľa nasledovného vzorca:

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

[39] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

Postačujúcou podmienkou na odhad parametrov lineárneho regresného modelu je teda minimalizovanie súčtu štvorcov rezíduí:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow min.$$

Z matematického hľadiska je súčet štvorcov rezíduí v prípade párovej lineárnej regresie funkciou dvoch premenných b_0 a b_1. Túto funkciu zapisujeme:

$$S = S\left(b_{0,} b_1\right) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Minimum funkcie $S\left(b_{0,} b_1\right)$ nájdeme tak, že jej parciálne derivácie podľa obidvoch premenných b_0 a b_1 položíme za rovnajúce sa nule:

$$\frac{\partial S}{\partial b_0} = 0, \frac{\partial S}{\partial b_1} = 0$$

Podrobný výpočet nájdeme v knihe *Štatistické metódy pre ekonómov a manažérov*.

Po konečnom rozšírení zlomku výrazom $\frac{1}{n^2}$ získame pre regresný koeficient b_1 vyjadrenie:

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

Regresný koeficient b_0 vyjadríme zjednoduším prvej rovnice nasledovne:

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

Z daného vzťahu teda vieme zistiť, že vyrovnávajúca priamka vždy prechádza bodom $[\bar{x}, \bar{y}]$

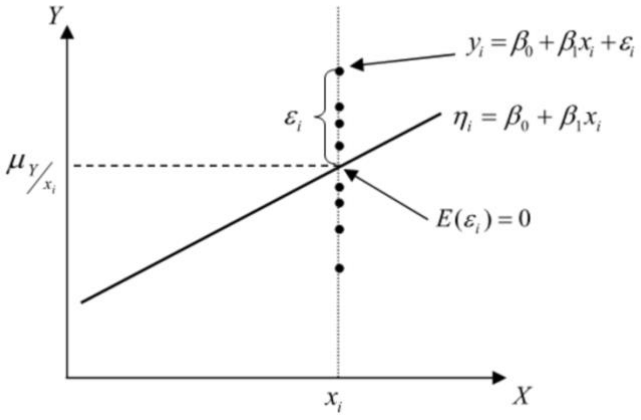
2.3.4 Predpoklady o náhodnej zložke regresného modelu

Metóda najmenších štvorcov, vysvetlená v predchádzajúcich kapitolách, poskytuje najlepšie neskreslené odhady lineárneho regresného modelu, ak model spĺňa nasledujúce predpoklady o náhodnej zložke.[40]

Predpoklad 1: Stredná hodnota náhodných chýb ε_i sa rovná nule, a to pre ľubovoľnú úroveň x_i vysvetľujúcej premennej X:

$$E(\varepsilon_i) = 0 \ \ pre \ \ i = 1, 2, \dots n$$

Tento predpoklad hovorí o tom, že pre každú hodnotu nezávislej premennej X musia byť stredné hodnoty závislej premennej Y (označené ako μ_{Y/x_i}), umiestnené na rovnakej vyrovnávajúcej priamke. Inými slovami, priemer odchýlok pozorovaných hodnôt Y od priamky, musí byť nulový pre každú hodnotu X. V praxi to znamená, že niektoré pozorované hodnoty Y budú ležať nad priamkou (kladná náhodná chyba $\varepsilon_i > 0$), zatiaľ čo iné budú ležať pod priamkou (záporná náhodná chyba $\varepsilon_i < 0$), avšak ich priemerná hodnota by mala byť na priamke. [41]



Obrázok 4 Grafické znázornenie predpokladu 1

Zdroj: Regresná a korelačná analýza s aplikáciami v softvéri SAS (Šoltés Erik, 2019)

[40] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.
[41] LABUDOVÁ, Viera a kol.: Štatistické metódy pre ekonómov a manažérov. Vydavateľstvo WOLTERS KLUWER, 2021. 392s. ISBN978-80-571-0401-8.

Predpoklad 2: Rozptýlenosť (variabilita) hodnôt závislej premennej Y pre ktorúkoľvek hodnotu vysvetľujúcej premennej X je rovnaká, t. j. rozptyl náhodnej zložky je konštantný:[42]

$$D(\varepsilon_i) = {\sigma_\varepsilon}^2 \ \ pre \ \ i = 1, 2, \dots n$$

Tento predpoklad sa týka homoskedasticity, čo znamená, že pri zmenách hodnôt nezávislej premennej sa rozptyl náhodných chýb ${\sigma_\varepsilon}^2$ nezmení. Ak je tento predpoklad splnený, hovoríme o homoskedastickom modeli. Ak sa však variabilita vysvetľovanej premennej mení so zmenami vysvetľujúcej premennej, potom hovoríme o heteroskedasticite náhodných chýb a o heteroskedastickom modeli. Tento predpoklad nie je v praxi vždy splnený, a to najmä pri skúmaní závislosti výdavkov domácností od príjmu domácností. Nižšie príjmové skupiny majú tendenciu mať relatívne stabilné výdavky, ktoré sa zameriavajú predovšetkým na zabezpečenie základných potrieb. Naopak, domácnosti s vyššími výdavkami môžu mať veľké rozdiely v ich výdavkoch, pretože sú viac zamerané na úspory a investície. Heteroskedasticita sa môže prejaviť aj v iných oblastiach, ako napríklad pri skúmaní efektov liečby na zdravie pacientov.[43]

Predpoklad 3: Náhodné chyby sú pre ľubovoľnú dvojicu hodnôt vysvetľujúcej premennej $x_i \neq x_j$ vzájomne nezávislé, čo znamená, že ich kovariancie (závislosti medzi dvoma náhodnými premennými) sa rovnajú nule:

$$cov(\varepsilon_i \varepsilon_j) = 0 \ \ pre \ všetky \ \ i \neq j$$

Základnými charakteristikami lineárnej závislosti dvoch premenných X a Y sú párový (jednoduchý) koeficient korelácie ρ_{xy} a párový (jednoduchý) koeficient determinácie p_{xy}^2. *Párový koeficient korelácie* meria obojstrannú lineárnu závislosť dvoch premenných a nadobúda hodnoty z intervalu $\langle -1, 1 \rangle$. Silu lineárnej závislosti vieme posúdiť z absolútnej hodnoty koeficienta korelácie.

[42] LABUDOVÁ, Viera a kol.:Štatistické metódy pre ekonómov a manažérov. Vydavateľstvo WOLTERS KLUWER, 2021. 392s. ISBN978-80-571-0401-8.
[43] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

Čím je absolútna hodnota koeficienta korelácie bližšia k hodnote 1, tým je závislosť silnejšia. Ak sa koeficient korelácie rovná 1 alebo -1, hovoríme o úplnej korelácii (priamej alebo nepriamej).

Bodovým odhadom párového koeficienta korelácie ρ_{xy} je výberový párový koeficient korelácie r_{xy}. Jeho výpočet vychádza z kovariancie $cov\ xy$, ktorá je daná vzťahmi:[44]

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

$$cov\ xy = \overline{xy} - \bar{x} \cdot \bar{y}$$

Kovariancia môže nadobúdať hodnoty z intervalu $(-\infty, \infty)$ a vyjadruje smer lineárnej závislosti medzi dvoma premennými. Pre kovarianciu platia nasledovné vzťahy:

$cov\ xy = 0 \Leftrightarrow$ premenné X a Y nie sú lineárne závislé
$cov\ xy > 0 \Leftrightarrow$ medzi premennými X a Y je priama lineárna závislosť
$cov\ xy < 0 \Leftrightarrow$ medzi premennými X a Y je nepriama lineárna závislosť

Avšak, keď hovoríme o koeficiente korelácie, nezískavame len informáciu o tom, či medzi premennými existuje priama alebo nepriama závislosť, alebo či neexistuje lineárna závislosť. Dôležitou súčasťou koeficientu korelácie je aj informácia o intenzite tejto závislosti. Preto sa kovariancia premenných X a Y delí smerodajnými odchýlkami oboch premenných, čím získavame výberový koeficient korelácie:

$$r_{xy} = \frac{cov\ xy}{s_x \cdot s_y}$$

Párový koeficient determinácie p_{xy}^2 slúži na určenie tesnosti lineárnej závislosti medzi premennými X a Y. Jeho hodnoty sa pohybujú v intervale $\langle -1, 1 \rangle$ a jeho stonásobok

[44] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

udáva, koľko percent celkovej variability závislej premennej je vysvetlených lineárnou regresnou funkciou s vysvetľujúcou premennou X.

Ak je párový koeficient determinácie rovný jednej, znamená to, že regresná priamka $\eta_i = \beta_0 + \beta_1 x_i$ úplne vystihuje variabilitu vysvetľovanej premennej Y, čo znamená, že medzi premennými X a Y existuje deterministická závislosť. [45]

Naopak, ak je párový koeficient determinácie nulový, znamená to, že premenné X a Y sú lineárne nezávislé. Bodovým odhadom párového koeficienta determinácie p_{xy}^2 je výberový párový koeficient determinácie r_{xy}^2. Vypočítame ho ako podiel variability premennej Y, ktorú vysvetľuje regresný model, a celkovej variability závislej premennej, podľa definície koeficienta p_{xy}^2. Tento odhad nám pomáha určiť, do akej miery sa vysvetľujúca premenná X podieľa na variabilite vysvetľovanej premennej Y.[46]

$$r_{xy}^2 = \frac{SSM}{SST} = 1 - \frac{SSR}{SST}$$

Ak však *predpoklad 3* nie je splnený a kovariancie dvoch premenných sa nerovnajú nule, hovoríme o autokorelácii chýb.

Autokorelácia je štatistický termín, ktorý sa používa na opisovanie vzájomnej závislosti medzi pozorovaniami v časovom rade. Jednoducho povedané, autokorelácia znamená, že hodnoty v časovom rade sú navzájom závislé. Táto závislosť môže byť buď pozitívna (keď vysoké hodnoty sú spojené s vysokými hodnotami a nízke hodnoty sú spojené s nízkymi hodnotami) alebo negatívna (keď vysoké hodnoty sú spojené s nízkymi hodnotami a naopak). Prítomnosť autokorelácie najčastejšie overujeme pomocou Durbin-Watsonovho testu, inak nazývaného aj d-štatistika.[47] Bližšie informácie o tomto teste nájdeme v praktickej časti práce.

[45] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.
[46] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.
[47] https://yolkki.ru/sk/tehnika/sushchnost-i-posledstviya-avtokorrelyacii-vremennye-ryady/

Predpoklad 4: Rozdelenie pravdepodobnosti náhodných chýb je normálne so strednou hodnotou 0 a rozptylom σ_ε^2:[48]

$$\varepsilon_i \sim N(0;\ \sigma_\varepsilon^2)\ \ pre\ \ i = 1, 2, \dots n$$

Lineárny regresný model, ktorého náhodná zložka spĺňa uvedené štyri predpoklady, sa nazýva klasický lineárny model (KLRM). Predstavuje určitý štandard, s ktorým sa porovnávajú modely, v ktorých aspoň jeden predpoklad nie je splnený.
V klasickom lineárnom modeli sa teda predpokladá, že náhodné chyby ε_i sú navzájom nezávislé normálne rozdelené náhodné premenné som strednou hodnotou nula a konštantným rozptylom σ_ε^2.

2.3.5 Overenie štatistickej významnosti regresného modelu

Na základe *analýzy rozptylu* vysvetľovanej premennej posúdime, do akej miery je vplyv vysvetľujúcej premennej na vysvetľovanú premennú relevantný a ako dobre lineárny regresný model, odhadnutý metódou najmenších štvorcov, vystihuje variabilitu závislej premennej. Vychádzame pritom z nasledovného vzorca:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Umocnením tejto rovnice, jej sumáciou cez $i = 1, 2, \dots n$ a po úprave dostaneme rovnosť:

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

Vzorec $(y_i - \bar{y})$ vyjadruje rozdiel medzi i-tým pozorovaním závislej premennej a jej priemerom. Súčet štvorcov týchto rozdielov, označovaný ako SST alebo SST_{otal}, ktorý je vyjadrený ľavou stranou rovnice $\sum_{i=1}^{n}(y_i - \bar{y})^2$. SST charakterizuje celkovú variabilitu závislej premennej. Ak sú hodnoty premennej Y blízko k jej priemeru $\bar{y}$ potom bude hodnota SST menšia, pretože je menšia variabilita.

[48] LABUDOVÁ, Viera a kol.:Štatistické metódy pre ekonómov a manažérov. Vydavateľstvo WOLTERS KLUWER, 2021. 392s. ISBN978-80-571-0401-8.

Vzorec $(\hat{y}_i - \bar{y})$ vyjadruje rozdiel medzi i-tou predpokladanou hodnotou závislej premennej a jej priemerom. Súčet štvorcov týchto rozdielov, označovaný ako SSM alebo SSM_{odel}, predstavuje variabilitu vyrovnaných hodnôt závislej premennej okolo jej priemeru $\bar{y}$, teda variabilitu, ktorá je vysvetlená regresným modelom.

Rozdiel skutočnej a vyrovnanej hodnoty závislej premennej $(y_i - \hat{y}_i)$ je rezíduum. Súčet druhých mocnín rezíduí, označovaný ako SSR^3 alebo $SSR_{esidual}$, vyjadruje variabilitu závislej premennej, ktorá nie je vysvetlená regresným modelom. SSR teda predstavuje nevysvetlenú (zvyškovú) variabilitu závislej premennej okolo regresnej funkcie $\hat{y}_i = \eta_i$.[49]

Predošlé rovnice sa teda dajú zapísať aj v tvare:

$$SST_{otal} = SSM_{odel} + SSR_{esidual}$$

kde: SST_{otal} je celková variabilita premennej Y
SSM_{odel} je variabilita Y vysvetlená modelom
$SSR_{esidual}$ je variabilita Y nevysvetlená modelom

Každému súčtu štvorcov je priradený počet stupňov voľnosti, ktorý určuje počet nezávislých sčítancov, z ktorých súčet štvorcov vznikol. Tento počet sa určuje ako počet prvkov v súbore mínus počet vypočítaných štatistík. SST obsahuje $(n-1)$ stupňov voľnosti, pretože obsahuje n nezávislých sčítancov (pozorovaní premennej Y), z ktorých jedno je určené priemerom $\bar{y}$. SSM má jeden stupeň voľnosti (dva parametre β_0, β_1 regresnej priamky mínus jeden priemer $\bar{y}$). SSR má $(n-2)$ stupňov voľnosti, pretože obsahuje n nezávislých sčítancov (pozorovaní premennej Y) a dva nezávislé sčítance (parametre regresnej priamky β_0 *a* β_1)[50].

[49] LABUDOVÁ, Viera a kol.:Štatistické metódy pre ekonómov a manažérov. Vydavateľstvo WOLTERS KLUWER, 2021. 392s. ISBN978-80-571-0401-8.
[50] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

Variabilita premennej Y	**Súčet štvorcov SS**	**Stupne voľnosti DF**	**Priemerný súčet štvorcov MS**	**Testovacia štatistika F**
Vysvetlená modelom - M_{odel}	$SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	$MSM = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{1}$	$\frac{MSM}{MSE}$
Nevysvetlená modelom - $R_{esidual}$	$SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$MSR = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1}$	
Celková - T_{otal}	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

Tabuľka 1 Analýza rozptylu pre lineárny regresný model

Zdroj: Vlastné spracovanie

„Priemerný štvorec“ (Mean Square - MS), ktorý sa uvádza v tabuľke, sa získava podielom súčtu štvorcov (Sum of Squares - SS) a počtu stupňov voľnosti (Degrees of Freedom - DF). V poslednom stĺpci tabuľky sa nachádza pomer priemerných štvorcov MSM a MSR. Výpočet tejto testovacej charakteristiky sa používa na overenie, či regresný model má štatistický význam[51].

Jeho výpočet závisí od toho, či je variabilita závislej premennej dostatočne veľká na to, aby regresný model skutočne vysvetľoval zmeny vysvetľovanej premennej. Podľa F-štatistiky je test známy pod názvom F-test štatistickej významnosti modelu alebo celkový F-test. Overuje sa ním nulová hypotéza:

$$H_0 : regresný\ model\ nie\ je\ štatisticky\ významný$$

Oproti alternatívnej hypotéze:

$$H_1 : regresný\ model\ je\ štatisticky\ významný$$

Za predpokladu platnosti nulovej hypotézy má testovacia charakteristika:

$$F = \frac{MSM}{MSR} = \frac{\frac{SSM}{1}}{\frac{SSR}{n-2}} = \frac{(n-2)\cdot\sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

F-rozdelenie je pravdepodobnostné rozdelenie, ktoré má dva stupne voľnosti - čitateľ a menovateľ. V regresnej analýze sa používa F-rozdelenie s 1 a (n-2) stupňami voľnosti, kde n predstavuje počet pozorovaní. Hladina významnosti α určuje pravdepodobnosť zamietnutia nulovej hypotézy, že regresný model nevysvetľuje variabilitu závislej

[51] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

premennej. Pokiaľ je výsledná hodnota F-výpočtu väčšia ako kritická hodnota F s danými stupňami voľnosti, zamietame nulovú hypotézu a považujeme regresný model za vhodný. V opačnom prípade ju prijmeme a model nepovažujeme za dostatočne vysvetľujúci variabilitu závislej premennej a tým pádom považujeme regresný model za nevhodný.[52]

2.3.6 Testy hypotéz a intervaly spoľahlivosti pre parametre klasického lineárneho modelu

Metóda najmenších štvorcov, ktorú sme použili na odhad parametrov regresnej priamky, minimalizuje súčet štvorcov rezíduí $S = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \sum_{i=1}^{n} e_i^2$.
Veľkosť rezíduí $e_i = y_i - \hat{y_i}$ a tým aj veľkosť súčtu štvorcov rezíduí závisí od vzdialenosti pozorovaných bodov $[x_i, y_i]$ od vyrovnávajúcej regresnej funkcie. Ak sú tieto vzdialenosti malé, aj rezíduá a súčet štvorcov rezíduí je malý. Vydelením tohto súčtu počtom stupňov voľnosti (v párovej lineárnej regresii $n - 2$) získame rozptyl rezíduí:[53]

$$s_{rez}^2 = MSR = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n-2}$$

a jeho odmocnením štandardnú odchýlku rezíduí:

$$s_{rez} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n-2}}$$

ktorá je mierou variability hodnôt y_i okolo vyrovnávajúcej priamky.

Rozptyl rezíduí je neskreslený odhad rozptylu náhodných chýb, štandardná odchýlka rezíduí je neskreslený odhad štandardnej odchýlky náhodných chýb, čo možno zapísať takto:

$$est\sigma_\varepsilon^2 = s_{rez}^2 \;\; a \;\; est\sigma_\varepsilon = s_{rez}$$

Následne vieme vypočítať, že neskreslený odhad rozptylu lokujúcej konštanty b_0 je:

[52] LABUDOVÁ, Viera a kol.:Štatistické metódy pre ekonómov a manažérov. Vydavateľstvo WOLTERS KLUWER, 2021. 392s. ISBN978-80-571-0401-8.
[53] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

$$s_{b_0}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

a najideálnejším odhadom rozptylu regresného koeficienta b_1 je:

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

Čím je rozptyl regresného koeficienta $s_{b_1}^2$ vyšší, tým je väčšia šanca, že sa odhad b_1 bude výrazne líšiť od skutočnej hodnoty parametra β_1. Naopak, ak sa rozptyl regresného koeficienta zmenšuje, tak sa pravdepodobnosť relatívne veľkej odchýlky bodového odhadu b_1 od skutočnej hodnoty parametra β_1 znižuje. Preto je žiadúce minimalizovať rozptyl regresného koeficienta b_1, aby sme získali čo najpresnejší odhad parametra β_1.

Testovanie hypotéz a výpočet intervalov spoľahlivosti pre regresné koeficienty β_1, β_{10} sú založené na predpoklade, že klasický lineárny regresný model je platný a že výberové charakteristiky majú Studentovo t-rozdelenie s (n-2) stupňami voľnosti:

$$T_1 = \frac{b_0 - \beta_0}{s_{b_0}} \quad a \quad T_2 = \frac{b_1 - \beta_1}{s_{b_1}}$$

2.3.7 Testy hypotéz pre regresný koeficient

a) Testy hypotéz o zhode regresného koeficienta β_1 so známou konštantou (v tomto prípade použijeme konštantu β_{10})

Na účely overenia nulovej hypotézy:

$$H_0: \beta_1 = \beta_{10}$$

oproti alternatívnej hypotéze:

$$H_1: \beta_1 \neq \beta_{10} \ \ alebo \ \ H_1: \beta_1 < \beta_{10}, resp. \ H_1: \beta_1 > \beta_{10}$$

dosadíme predpokladanú hodnotu β_{10} regresného koeficienta do predošlého vzťahu a dostaneme testovaciu charakteristiku:

$$t = \frac{b_1 - \beta_{10}}{s_{b_1}}$$

Ktorá má za predpokladu platnosti nulovej hypotézy Studentovho rozdelenie pravdepodobnosti so stupňami $(n-2)$.[54]

Pre obojstranný test s hladinou významnosti α je kritická oblasť definovaná nerovnicou $|t| > t_{1-\frac{\alpha}{2}}(n-2)$. Ak je prijatá nulová hypotéza, znamená to, že zväčšením vysvetľujúcej premennej o jednotku sa podmienená stredná hodnota vysvetľovanej premennej Y zmení v priemere o konštantu β_{10}. V prípade jednostranného testu s alternatívnou hypotézou $H_1: \beta_1 < \beta_{10}$ *alebo* $H_1: \beta_1 > \beta_{10}$ je kritická oblasť určená nerovnicou $t < -t_{1-\alpha}(n-2)$ alebo $t > t_{1-\alpha}(n-2)$[55].

b) Test štatistickej významnosti regresného koeficienta β_1

Test je špeciálny prípad predchádzajúceho testu a overujeme ním, či regresný koeficient v základnom súbore má hodnotu $\beta_{10} = 0$. Testujeme v ňom nulovú hypotézu:

$$H_0: \beta_1 = 0 \rightarrow regresný\ koeficient\ nie\ je\ štatisticky\ významný$$

oproti alternatívnej hypotéze:

$$H_1: \beta_1 \neq 0 \rightarrow regresný\ koeficient\ je\ štatisticky\ významný$$

[54] LABUDOVÁ, Viera a kol.: Štatistické metódy pre ekonómov a manažérov. Vydavateľstvo WOLTERS KLUWER, 2021. 392s. ISBN 978-80-571-0401-8.
[55] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

V tomto prípade bude testovacou charakteristikou:

$$t = \frac{b_1}{s_{b_1}}$$

Ak je hodnota t-štatistiky v oblasti zamietnutia nulovej hypotézy $|t| > t_{1-\frac{\alpha}{2}}(n-2)$, môžeme predpokladať, že zmeny nezávislej premennej X spôsobujú zmeny závislej premennej Y. To znamená, že existuje lineárna závislosť medzi premennými X a Y. Ak je prijatá nulová hypotéza na hladine významnosti α, nemôžeme predpokladať lineárnu závislosť medzi premennými X a Y. V takom prípade môžeme buď vylúčiť vysvetľujúcu premennú X z regresného modelu alebo uvažovať o inej funkcii premennej X. Tento test je jedným z najdôležitejších testov pre regresný koeficient a je súčasťou procedúry regresnej analýzy vo všetkých štatistických softvéroch[56].

2.3.8 Interval spoľahlivosti pre regresný koeficient

Na základe poznatkov zo štatistickej indukcie jednoducho odvodíme intervalový odhad regresného koeficienta β_1. Pre dvojstranný $100 \cdot (1-\alpha)\%$ interval spoľahlivosti využijeme symetrickosť Studentovho rozdelenia, vďaka ktorej platí:

$$P\left(-t_{1-\frac{\alpha}{2}} < T < t_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

odkiaľ po dosadení štatistiky T_2 z odvodeného vzťahu Studentovho rozdelenia dostaneme:

$$P\left(-t_{1-\frac{\alpha}{2}} < \frac{b_1-\beta_1}{s_{b_1}} < t_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Konečnou úpravou získame výsledný tvar $100 \cdot (1-\alpha)\%$ dvojstranného intervalu spoľahlivosti pre regresný koeficient β_1:

[56] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

$$P\left(b_1 - t_{1-\frac{\alpha}{2}} \cdot s_{b_1} < \beta_1 < b_1 + t_{1-\frac{\alpha}{2}} \cdot s_{b_1}\right) = 1 - \alpha$$

kde: $t_{1-\frac{\alpha}{2}}$ je kvantil Studentovho rozdelenia s $(n-2)$ stupňami voľnosti[57].

[57] ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

3 Výsledky práce

V tejto kapitole sa budeme venovať zostrojeniu lineárneho regresného modelu. Naše teoretické poznatky z predchádzajúcich kapitol sa v tejto časti pokúsime doplniť o praktické príklady, vďaka ktorým získame lepšiu predstavu o danej problematike. Potrebné výpočty budeme robiť pomocou programu R, ktorý obsahuje niekoľko balíčkov zameraných práve na štatistiku a analýzu dát.

3.1 Jednoduchá lineárna regresia v jazyku R

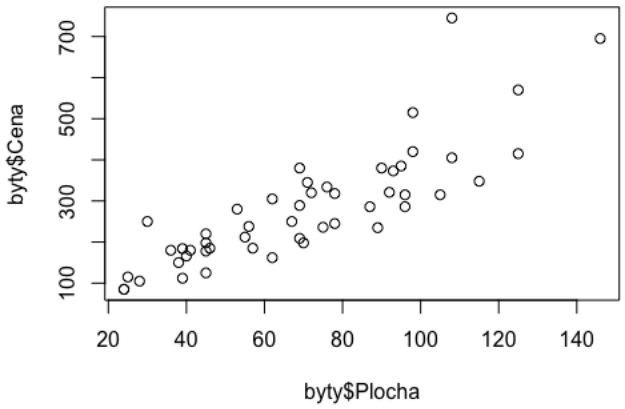
Zameriame sa na príklad týkajúci sa jednoduchého lineárneho modelu, aby sme, čo najlepšie a najprehľadnejšie dokázali vysvetliť základné funkcie a osvojili si syntax jazyka. Ako ukážku použijeme nasimulované dáta cien bytov (v tis. €) v závislosti od plochy bytu (v m^2). Súbor obsahuje 50 pozorovaní s 2 premennými. Budeme sa zameriavať na závislú premennú `Cena`, ktorá predstavuje cenu jednotlivých bytov a nezávislú premennú `Plocha`, ktorá predstavuje plochu daných bytov.

Na začiatok si ukážeme, ako by vyzeral jednoduchý lineárny model pre rovnicu .

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

kde y_i predstavuje hodnotu ceny bytu (`Cena`), x_i udáva plochu (`Plocha`). Už z teórie vieme, že β_i sú neznáme parametre, ktoré sa snažíme odhadnúť a platí, že náhodné veličiny $\varepsilon_i = N(0, \sigma^2)$. Ako prvé si necháme vykresliť dáta, aby sme mali lepšiu predstavu, s čím vlastne pracujeme. Na to nám v jazyku R slúži príkaz `plot()`.

```
> plot(byty$Plocha, byty$Cena)
```



Obrázok 5 Párový graf: `byty$Plocha, byty$Cena`

Zdroj: Vlastné spracovanie

Na obrázku 5 vidíme párový graf jednotlivých premenných. I keď sa nám z obrázku môže javiť určitá závislosť medzi premennými, je vhodné previesť presné výpočty. Výpočet budeme prevádzať pomocou metódy najmenších štvorcov. V jazyku R si pomocou funkcie `lm` vytvoríme jednoduchý lineárny model.

```
model <- lm(Cena ~ Plocha, data=byty)
```

Argumenty funkcie špecifikujú tvar modelu. Závislá premenná je v tomto prípade `Cena` a nezávislá premenná `Plocha`. Nasledovným dotázaním sa na model (`model`) sa nám zobrazí formula modelu a hľadané hodnoty koeficientov β_i:

```
> model

Call:

lm(formula = Cena ~ Plocha, data = byty)

Coefficients:

(Intercept)       Plocha

      2.754        3.976
```

Vidíme, že odhady $\beta_0 = 2{,}754$, $\beta_1 = 3{,}976$. Interpretácia je potom jednoduchá. Ak sa zväčší plocha bytu (`Plocha`) o jeden m^2, potom vzrastie očakávaná cena bytu o 3,976 tisíc € a teda medzi týmito dvoma veličinami je kladný vzťah. Potom tu máme ešte hodnotu `Intercept`, čo je tzv. úrovňová konštanta. Určuje hodnotu závislej premennej pri nulových hodnotách nezávislých premenných.

Rozšírené informácie o modeli spolu s analýzou rezíduí získame použitím príkazu `summary()`:

```
> summary(model)

Call:

lm(formula = Cena ~ Plocha, data = byty)

Residuals:

    Min      1Q  Median      3Q     Max

-121.60  -54.38   -0.07   28.72  312.86

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)   2.7536    27.7746   0.099    0.921

Plocha        3.9758     0.3662  10.857  1.6e-14 ***

---

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

Residual standard error: 76.22 on 48 degrees of freedom

Multiple R-squared:  0.7106,  Adjusted R-squared:  0.7046
```

`Call` hovorí o tvare modelu, ktorým sa zaoberáme. Rezíduá (`Residuals`) nás upozorňujú na niekoľko informácií o normalite, respektíve jej porušení. Ak rezíduá spĺňajú podmienku normality, medián by sa nemal príliš vzďaľovať od nuly. Taktiež priveľká vzdialenosť medzi prvým a tretím kvartilom môže byť dôsledkom šikmosti množiny rezíduí. Samotnú normalitu modelu budeme overovať v ďalšej časti našej práce.

Koeficienty (`Coefficients`) nám neukazujú len ich jednotlivé hodnoty, ale rovno celú tabuľku, ktorá nám dovolí si utvoriť lepší obraz. Prvý stĺpec ukazuje jednotlivé odhady koeficientov. Druhý stĺpec obsahuje odhady ich smerodajných odchýlok $\hat{\sigma}$. Tretí stĺpec je tvorený hodnotami T-štatistík, kde nulová hypotéza je tvaru: H_0: $\beta_1 = 0$.

Posledný stĺpec Pr(>|t|) je tzv. *p-hodnota* testu nulovosti jednotlivých koeficientov, čo je vlastne pravdepodobnosť, že hodnota náhodnej veličiny bude aspoň taká veľká ako absolútna hodnota sledovanej štatistiky T . Vysoká p-hodnota indikuje, že hypotéza je konzistentná s dátami. Posledný stĺpec nám ukazuje významnosť jednotlivých koeficientov. Ak je niektorý koeficient štatisticky významný, jeho p-hodnota testu nulovosti je malá, pričom intervaly hladiny významnosti ukazuje ďalší riadok (`Signif. codes`). V našom prípade, na hladine významnosti 0,05, nie je lokujúca konštanta (`Intercept`) štatisticky významná nakoľko jej p-hodnota je väčšia než 0,05. Parameter `Plocha` je štatisticky významný nakoľko jeho p-hodota je nižšia ako hladina významnosti 0,05 .

Hodnota `Multiple R-squared` (R^2) poskytuje informácie o tom, ako dobre model vysvetľuje variabilitu dát. Hodnota R^2 sa pohybuje v intervale od 0 do 1, pričom vyššia hodnota znamená, že model lepšie vysvetľuje variabilitu dát. Ako môžeme vidieť z nášeho výstupu, hodnota `Multiple R-squared` je `0.7106`, čo znamená, že 71,06% variability ceny bytov vysvetľuje premenná `plocha`, zatiaľ čo zvyšných 28,94% variability je spôsobených náhodnými vplyvmi alebo premennými nezahrnutými do regresného modelu.

A nakoniec v našom výstupe môžeme vidieť hodnotu `F-statistic` (F-test). F-test slúži na overenie štatistickej významnosti modelu ako celku. Významnosť modelu vieme jednoducho overiť tak, že porovnáme p-hodnotu s hladinou významnosti 0,05. Ak je p-hodnota menšia ako 0,05, tak zamietneme nulovú hypotézu a náš model bude štatisticky

významný. V našom prípade je p-hodnota menšia ako 0,05, čo znamená, že náš model je štatisticky významný.

3.1.1 Predpokladané hodnoty a rezíduá

Predpokladané (alebo predikované) hodnoty sú y-hodnoty, ktoré by ste očakávali pre dané x-hodnoty v súlade s postaveným regresným modelom (alebo vizuálne, s najlepšie sa hodiacim priamkovým regresným riešením). V R môžeme jednoducho rozšíriť naše dáta a pridať predpokladané hodnoty a reziduá pomocou funkcie `augment()`. Nazvyme výstup `model.diag.metrics`, pretože obsahuje niekoľko výstupov, ktoré sú užitočné pre diagnostiku regresie.

```
library(broom)

model.diag.metrics <- augment(model)

head(model.diag.metrics)
```

Po zadaní funkcie `head()` dostaneme nasledujúci výstup:

```
> head(model.diag.metrics)

# A tibble: 6 × 8

   Cena Plocha .fitted .resid   .hat .sigma  .cooksd .std.resid

  <dbl>  <dbl>   <dbl>  <dbl>  <dbl>  <dbl>    <dbl>      <dbl>

1    85     24    98.2 -13.2  0.0686   77.0 0.00118      -0.179

2    85     24    98.2 -13.2  0.0686   77.0 0.00118      -0.179

3   105     28   114.   -9.08 0.0605   77.0 0.000486     -0.123

4   112     39   158.  -45.8  0.0420   76.7 0.00827      -0.614

5   115     25   102.   12.9  0.0665   77.0 0.00109       0.175

6   125     45   182.  -56.7  0.0343   76.6 0.0102       -0.756
```

1. **Prispôsobené hodnoty (.fitted):** sú predpokladané hodnoty závislej premennej (y), získané z regresného modelu na základe daných nezávislých premenných (x). Jednoducho povedané, predstavujú hodnoty, ktoré model odhaduje ako výstupnú premennú na základe vstupných údajov.
2. **Reziduálne chyby (.resid):** rezíduá sú rozdiely medzi pozorovanými hodnotami závislej premennej (skutočné hodnoty y) a prislúchajúcimi prispôsobenými hodnotami z regresného modelu. Dokazujú, ako veľmi sa odlišujú predpovede modelu od skutočných hodnôt.
3. **Hat hodnoty (.hat):** Hat hodnoty sa používajú pri diagnostike regresie na identifikáciu bodov s vysokým vplyvom. Bod s vysokým vplyvom je dátový bod s extrémnymi hodnotami nezávislých premenných (x), ktorý môže výrazne ovplyvniť kvalitu regresného modelu. Hat hodnoty sú diagonálne prvky tzv. "hat matice", ktorá sa používa na výpočet predpovedaných hodnôt v regresii.
4. **Štandardizované rezíduálne chyby (.std.resid):** Štandardizované rezíduálne chyby sú rezíduá, ktoré sa delia ich odhadnutými štandardnými chybami. Pomáhajú identifikovať odľahlé hodnoty alebo extrémne hodnoty závislej premennej (y) po zohľadnení variability vysvetlenej modelom.
5. **Cookova vzdialenosť (.cooksd):** Cookova vzdialenosť predstavuje mieru vplyvu každého dátového bodu na regresný model. Kombinuje informácie o rezíduálnych chybách a vplyve (hat hodnoty) každého dátového bodu. Vysoké hodnoty Cookovej vzdialenosti naznačujú významné body, ktoré môžu výrazne ovplyvniť regresné koeficienty.

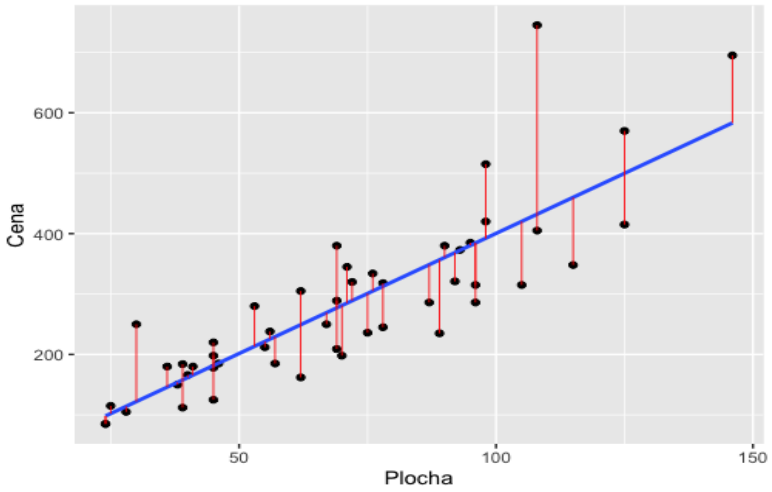
Vyššie uvedené pojmy majú významnú úlohu pri diagnostike regresných modelov, umožňujú analyzovať kvalitu modelu, identifikovať významné body, detekovať odľahlé hodnoty a posúdiť celkový výkon modelu.

Nasledujúci R kód vykresľuje chyby reziduálov (v červenej farbe) medzi pozorovanými hodnotami a regresnou priamkou. Každý vertikálny červený segment reprezentuje reziduálnu chybu medzi pozorovanou hodnotou ceny bytu a prislúchajúcou predpokladanou (tj. predikovanou) hodnotou.

```
ggplot(model.diag.metrics, aes(Plocha, Cena)) +
  geom_point() +
```

```
  stat_smooth(method = lm, se = FALSE) +

  geom_segment(aes(xend = Plocha, yend = .fitted), color =
"red", size = 0.3)
```

Z daného kódu dostaneme nasledovný graf:



Obrázok 6 Graf reziduálnych chýb

Zdroj: Vlastné spracovanie

Z grafu môžeme vidieť, že nie všetky hodnoty dát padajú presne na predpokladanú regresnú priamku. To znamená, že pre danú plochu bytov sa pozorované (alebo namerané) hodnoty cien môžu líšiť od predikovaných hodnôt cien. Táto odchýlka sa nazýva reziduálne chyby a je znázornená pomocou vertikálnych červených čiar.

3.1.2 Predpoklady lineárnej regresie

Jedným z prvých krokov po zostavení lineárneho regresného modelu je skontrolovať, či náš model spĺňa predpoklady lineárnej regresie. Tieto predpoklady sú dôležitou súčasťou posudzovania, či je model správne špecifikovaný. V tejto kapitole sa budeme venovať tomu, aké sú predpoklady lineárnej regresie a ako ich otestovať pomocou jazyku R.

Predpoklady, ktorými sa budeme bližšie zaoberať:

1. Lineárnosť: Predpokladá sa, že vzťah medzi závislou premennou (Y) a každou predpokladanou premennou (X) je lineárny. To znamená, že zmena v Y je priamo úmerná zmenám v X.

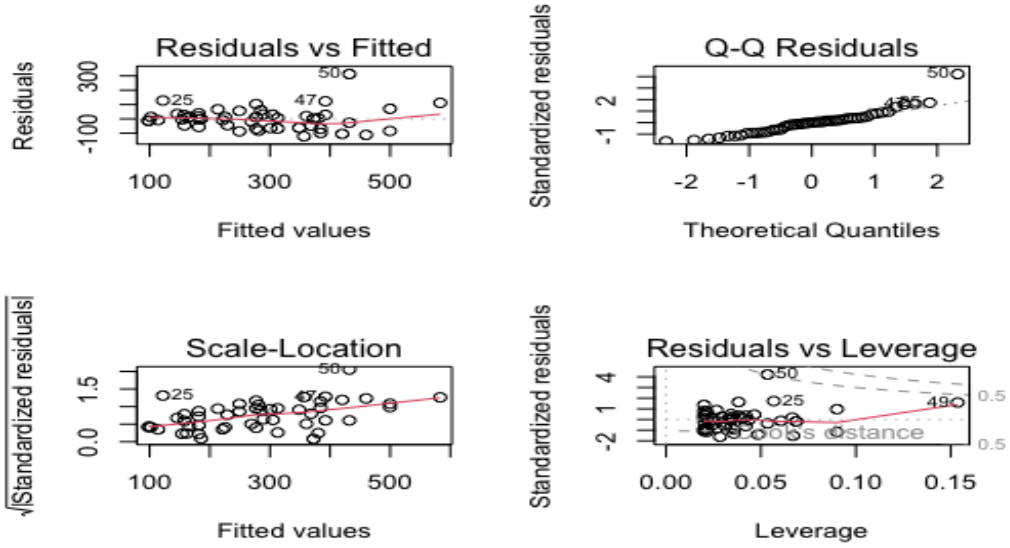
2. Nezávislosť: Pozorovania v dátovej sade by mali byť nezávislé. Každý dátový bod by nemal ovplyvňovať ani súvisieť s iným dátovým bodom.
3. Homoskedasticita: Tento predpoklad vyžaduje, aby variabilita rezíduí (rozdiely medzi skutočnými a predpovedanými hodnotami) bola konštantná pre všetky hodnoty prediktorov. Inak povedané, rozptyl rezíduí by nemal systematicky narastať alebo klesať s hodnotami prediktorov.
4. Normalita: Predpokladá sa, že rezíduá sú normálne distribuované. To znamená, že chyby (rezíduá) by mali nasledovať normálnu distribúciu s priemerom nula.

Ak tieto predpoklady nie sú splnené, môže to viesť k skresleným odhadom, nesprávnemu testovaniu hypotéz a nespoľahlivým predpovediam. Je dôležité overiť platnosť týchto predpokladov pred interpretáciou výsledkov lineárnej regresnej analýzy. Rôzne diagnostické nástroje a štatistické testy sa používajú na posúdenie, do akej miery sú tieto predpoklady splnené a na identifikáciu prípadných problémov v dátach. Ak sú predpoklady značne porušené, môže byť potrebné použiť alternatívne regresné modely alebo transformácie dát pre získanie spoľahlivejších výsledkov.

3.1.2.1 Diagnostické grafy

Diagnostické grafy (Diagnostic plots) sú vizuálne nástroje, ktoré sa používajú na overenie predpokladov lineárnej regresie a identifikáciu prípadných problémov v regresnom modeli. Tieto grafy nám pomáhajú posúdiť, do akej miery sú splnené predpoklady modelu a či je model vhodný pre naše dáta. Diagnostické grafy regresie môžu byť vytvorené pomocou základnej funkcie `plot()` alebo funkcie `autoplot()`:

```
par(mfrow = c(2, 2))

plot(model)
```



Obrázok 7 Diagnostické grafy

Zdroj: Vlastné spracovanie

Diagnostické grafy ukazujú rezíduá štyrmi rôznymi spôsobmi:

1. Residuals vs Fitted: Tento graf sa používa na overenie predpokladov linearity. Ak sú rezíduá rovnomerne rozmiestnené okolo horizontálnej čiary bez zreteľných vzorov (červená čiara je približne horizontálna na nule), naznačuje to lineárny vzťah.
2. Normal Q-Q: Tento graf sa používa na overenie predpokladu normality rezíduí. Ak väčšina rezíduí sleduje priamku so zlomenou čiarou, predpoklad je splnený.
3. Scale-Location: Tento graf slúži na overenie predpokladu homoskedasticity rezíduí (rovnakého rozptylu rezíduí). Ak sú rezíduá rozmiestnené náhodne a vidíme horizontálnu čiaru s rovnomerne (náhodne) rozptýlenými bodmi, predpoklad je splnený.
4. Residuals vs Leverage: Tento graf sa používa na identifikáciu vplyvných hodnôt v dátovej sade. Vplyvné hodnoty sú extrémne hodnoty, ktoré môžu ovplyvniť výsledky regresie, keď sú zaradené alebo vylúčené z analýzy.

Tieto diagnostické grafy nám poskytujú dôležité informácie o správnosti nášho regresného modelu a o tom, či sú splnené predpoklady lineárnej regresie. Ak diagnostické grafy odhalia odchýlky od predpokladov, môže byť potrebné upraviť model alebo dáta pre dosiahnutie spoľahlivých výsledkov.

3.1.2.2 Testovanie predpokladov

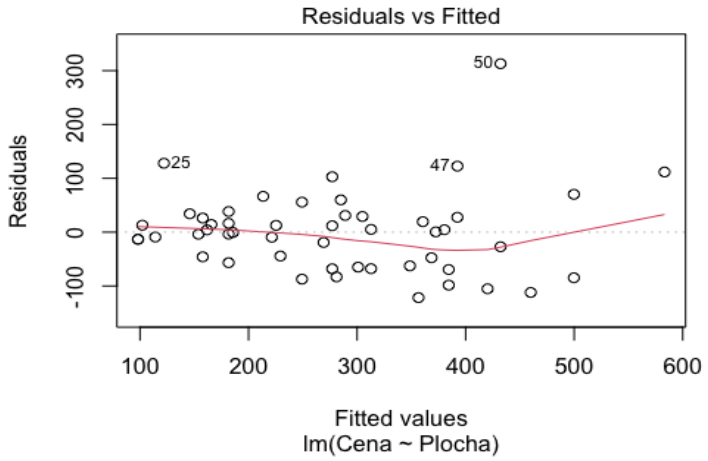
V tejto časti kapitoly sa budeme venovať testovaniu predpokladov pomocou diagnostických grafov a testovacích štatistík. Následne si ukážeme výsledky a interpretujeme si ich.

3.1.2.2.1 Lineárnosť

Linearitu dát môžeme overiť pozorovaním grafu Residuals vs Fitted (1. graf). Ideálne by tento graf nemal mať žiadny vzor a červená čiara by mala byť približne horizontálna pri nule.

Pri testovaní si načítame graf Residuals vs Fitted:

```
> plot(model, 1)
```



Obrázok 8 Graf Residuals vs Fitted

Zdroj: Vlastné spracovanie

Ideálne by rezíduá nemali ukazovať žiadny výrazný vzor. To znamená, že červená čiara by mala byť približne horizontálna pri nule. Prítomnosť vzoru môže naznačovať problém s niektorým aspektom lineárneho modelu. V našom príklade môžeme vidieť jemný pokles v grafe rezíduí. To naznačuje, že nemôžeme s istotou predpokladať lineárny vzťah medzi prediktorom a závislými premennými.

3.1.2.2.2 Nezávislosť

Najjednoduchší spôsob, ako overiť predpoklad nezávislosti, je pomocou Durbin-Watson testu. Tento test môžeme vykonať pomocou zabudovanej funkcie v R s názvom "durbinWatsonTest" alebo "dwtest" na náš model.
Po spustení tohto testu získame výstup s p-hodnotou, ktorá nám pomôže určiť, či je predpoklad splnený alebo nie.

```
dwtest(model)
```

Po spustení Durbin-Watson testu získavame:

```
Durbin-Watson test
data:  model
DW = 1.3298, p-value = 0.004959
alternative hypothesis: true autocorrelation is greater than 0
```

Nulová hypotéza tvrdí, že chyby nie sú autokorelované (sú nezávislé). Preto, ak získame p-hodnotu väčšiu ako 0,05, zlyháme v zamietnutí nulovej hypotézy. To nám poskytne dostatok dôkazov na to, aby sme tvrdili, že predpoklad nezávislosti je splnený. V našom prípade sme však dostali p-hodnotu `0.004959`, čo je menšie ako 0,05 a preto môžeme povedať, že nulovú hypotézu prijímame a v modeli je prítomná autokorelácia. Na jej odstránenie sa používajú rôzne metódy. Pre nás bude najideálnejšie ak zlogaritmujeme daný model pomocou funkcie `log ()`:

```
model2 <- lm(log(Cena) ~ Plocha, data = byty)
```

A následne otestujeme náš nový model pomocou Durbin-Watson testu:

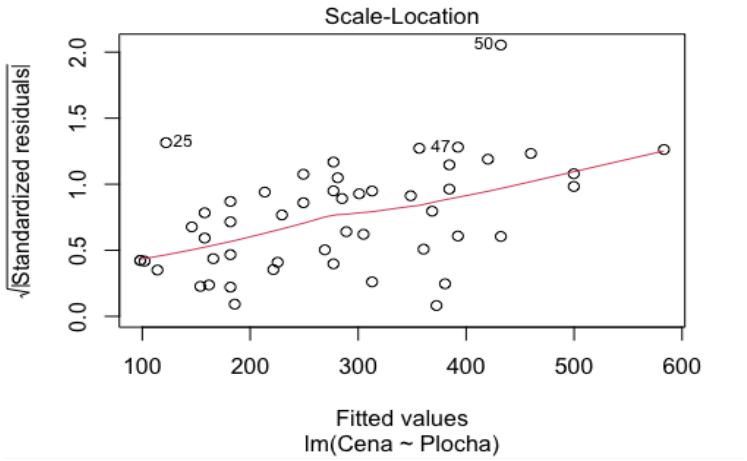
```
> dwtest(model2)

Durbin-Watson test

data:  model2
DW = 1.5683, p-value = 0.05324
alternative hypothesis: true autocorrelation is greater than 0
```

Teraz môžeme vidieť, že p-hodnota je väčšia ako 0,05, čo znamená, že zamietame nulovú hypotézu a podarilo sa nám z modelu odstrániť autokoreláciu.

3.1.2.2.3 Homoskedasticita

Tento predpoklad môžeme overiť pomocou grafu Scale-Location. V tomto grafe zobrazujeme predikované hodnoty oproti odmocnenej štandardizovanej reziduálnej hodnote. Ideálne by sme chceli vidieť body rezíduí rovnomerne rozptýlené okolo červenej čiary, čo by naznačovalo konštantný rozptyl. Graf si jednoducho zobrazíme pomocou funkcie `plot ()`:

```
> plot(model, 3)
```



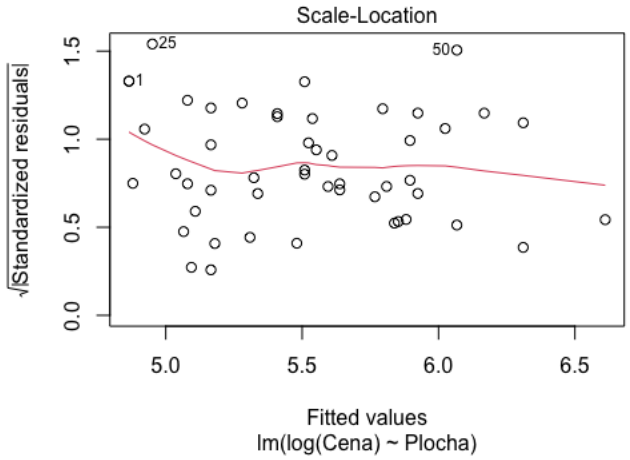
Obrázok 9 Graf Scale-Location

Zdroj: Vlastné spracovanie

V uvedenom grafe vidíme, že body rezíduí nie sú rovnomerne rozptýlené, a tým pádom tento predpoklad nie je splnený. Jedno bežné riešenie tohto problému je zlogaritmovať model pomocou funkcie `log ()`. Týmto spôsobom sa môže dosiahnuť, že rozptyl reziduálnych hodnôt bude konštantný, čo je požadovaný predpoklad pre homoskedasticitu.

```
model2 <- lm(log(Cena) ~ Plocha, data = byty)

plot(model2, 3)
```

Po zlogaritmovaní našeho modelu by sme mali dostať konštantnejší rozptyl reziduálnych hodnôt.



Obrázok 10 Zlogaritmovaný graf Scale-Location

Zdroj: Vlastné spracovanie

Ďalším spôsobom akým môžeme testovať homoskedasticitu je pomocou Breusch-Pagan testu.

```
bptest(model)
```

Pre ukážku použijeme znova začiatočný model bez logaritmu aby sme si boli istí, že naše tvrdenia sú správne.

```
> bptest(model)

studentized Breusch-Pagan test

data:  model

BP = 4.9959, df = 1, p-value = 0.02541
```

Z daného výstupu môžeme vidieť, že naša p-hodnota sa rovná `0.02541`, čo je menšie ako 0,05 a tým pádom boli naše tvrdenia správne a v tomto modeli nie je prítomná homoskedasticita a je prítomná heteroskedasticita, ktorá hovorí o tom, že naše reziduálne chyby v modeli nie sú konštantné. Práve preto sme daný model zlogaritmovali a tým aj

dostali konštantnejšie vyobrazenie modelu. Teraz si skúsme pomocou Breusch-Pagan testu overiť pravdivosť zlogaritmovaného modelu.

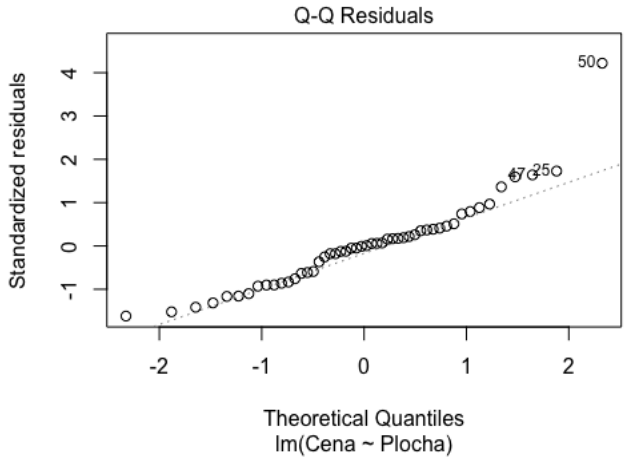
```
> bptest(model2)

studentized Breusch-Pagan test

data:  model2

BP = 1.0374, df = 1, p-value = 0.3084
```

Z tochto výstupu môžeme vidieť, že naša p-hodnota sa rovná `0.3084`, čo znamená, že nemôžeme zamietnuť nulovú hypotézu, pretože p-hodnota je väčšia než 0,05 a tým pádom je v modeli prítomná homoskedasticita. Predpoklad o homoskedasticite je splnený.

3.1.2.2.4 Normalita

QQ graf reziduálov môžeme vizuálne použiť na overenie predpokladu o normálnosti. Snažíme sa dosiahnuť taký graf reziduálov, kde by hodnoty približne nasledovali priamku.

```
> plot(model, 2)
```



Obrázok 11 Graf Q-Q Residuals

Zdroj: Vlastné spracovanie

Väčšina bodov sa približne nachádza pozdĺž priamky, takže môžeme predpokladať normalitu a tým môžeme potvrdiť náš posledný sledovaný predpoklad.

Záver

Cieľom našej bakalárskej práce bolo bližšie priblížiť matematický popis regresného modelu v programovacom jazyku R ako z hľadiska teoretického ale aj praktického. Ukázali sme prepojenie regresnej analýzy s programovacím jazykom R. Postupy regresnej analýzy spadajú medzi jedny z najčastejšie využívané štatistické techniky v rozličných vedných oblastiach. Ich použitie umožňuje popísanie komplikovaných vzťahov v rámci rôznych premenných, majú široké možnosti uplatnenia nielen v ekonomických, ale aj sociálnych analýzach nevynímajúc prírodné vedy.

V teoretickej časti bakalárskej práce sme sa opierali o dostupnú domácu ale aj zahraničnú literatúru, pomocou ktorej sme si bližšie predstavili históriu vzniku programovacieho jazyka R, jeho výhody a nevýhody. Priblížili sme si aj jeho uplatnenie v praxi respektíve v ktorých oblastiach sa najviac využíva. Tiež sme sa venovali aj balíkom R, ktoré sú potrebné pre regresnú analýzu.

Kapitola Metodika a cieľ práce, je rozdelená do dvoch častí. V prvej časti sa zameriavame na objasnenie regresnej analýzy a druhá časť sa zameriava na regresný model, bližšie sa venujeme lineárnemu regresnému modelu a jeho aspektom.

Kapitola Výsledky práce je venovaná prepájaniu lineárneho regresného modelu s jazykom R v programovacom prostredí RStudio. V praktickej časti sme použili konkrétne dátové údaje na ktorých sme testovali vlastnosti lineárneho regresného modelu. Zamerali sme sa na príklad, ktorý sa týkal jednoduchého lineárneho modelu, aby sme mohli , čo najlepšie a najprehľadnejšie vysvetliť základné funkcie a osvojili si tak syntax jazyka.

Zoznam použitej literatúry

Knižné zdroje:

1. ISMAY, Chester – KIM, Y. Albert. Statistical Inference via Data Science. 1.Vydanie. Vydavateľstvo CHAPMAN AND HALL/CRC, 2019. 430s. ISBN 978-0367409821.

2. LABUDOVÁ, Viera a kol.: Štatistické metódy pre ekonómov a manažérov. Vydavateľstvo W FBIW.UNIZA. WOLTERS KLUWER, 2021. 392s. ISBN 978-80-571-0401-8.

3. PENG, D. *Roger. R Programming for Data Science*. 5.Vydanie. Vydavateľstvo LULU, 2016. 194s. ISBN 978-1365056826.

4. Ross Ihaka a Robert Gentleman. R: Jazyk pre analýzu údajov a grafiku. Journal of Computational and Graphical Statistics, 5(3):299–314, 1996.

5. ŠOLTÉS Erik. Regresná a korelačná analýza s aplikáciami v softvéri SAS. Vydavateľstvo LETRA EDU, 2019. 235s. ISBN 978-80-89962-38-9.

Internetové zdroje:

6. Bookdown. Data Analysis and Processing with R based on IBIS data. . [online]. [cit. 2023-04-20]. Dostupné na: https://bookdown.org/kdonovan125/ibis_data_analysis_r4/#preface

7. DataFlair. Pros and Cons of R Programming Language. [online]. [cit. 2023-04-19]. Dostupné na:https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/

8. Jasonheppler. BootcampR: An Introduction to R. [online]. [cit. 2023-04-18]. Dostupné na: https://jasonheppler.org/courses/bootcampr.2020/reading/02-reading/

9. Lineárny regresný model. [online]. [cit. 2023-05-10]. Dostupné na: http://fbiw.uniza.sk/kkm/old/publikacie/ek/ek_kap_3.pdf

10. Microsoft R Application Network. A (Brief) History of R. [online]. [cit. 2023-05-12]. Dostupné na: < https://mran.microsoft.com/documents/what-is-r#rhistory>

11. Mercury. Introduction to R and RStudio. [online]. [cit. 2023-04-20]. Dostupné na: http://mercury.webster.edu/aleshunas/R_learning_infrastructure/Introduction_to_R_and_RStudio.html

12. Packt. R packages for regression. [online]. [cit. 2023-05-10]. Dostupné na: na:https://subscription.packtpub.com/book/data/9781788627306/1/ch01lvl1sec18/r-packages-for-regression

13. Podstata a dôsledky autorkorelácie. Časové rady, Technika.[online] 2019. Dostupné na : https://yolkki.ru/sk/tehnika/sushchnost-i-posledstviya-avtokorrelyacii-vremennye-ryady/

14. Rbasics. Getting used to R, Rstudio and R Markdown. [online]. [cit. 2023-04-20]. Dostupné na: https://rbasics.netlify.app

15. ROOT.CZ. Programovací jazyk R: úvodní informace. [online]. [cit. 2023-04-20]. Dostupné na: https://www.root.cz/clanky/programovaci-jazyk-r-uvodni-informace/

16. TIOBE. Index for April 2023. [online]. [cit. 2023-04-18]. Dostupné na: https://www.tiobe.com/tiobe-index/

17. TechVidvan. 15 Features of R Programming you can't afford to overlook. [online]. [cit. 2023-04-18]. Dostupné na: https://techvidvan.com/tutorials/r-features/

18. Tutorialspoint. R-Overview. [online]. [cit. 2023-04-18]. Dostupné na: https://www.tutorialspoint.com/r/r_overview.htm

19. Zuzana Gibova: Metóda najmenších štvorcov. [online]. [cit. 2023-05-10]. Dostupné na: https://zuzana.gibova.website.tuke.sk/files/kap-3.2.pdf