VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY
INSTITUTE OF MATHEMATICS

VÍCEROZMĚRNÉ REGRESNÍ MODELY
MULTIDIMENSIONAL REGRESSION MODELS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Gabriela Hrubešová**
AUTHOR

VEDOUCÍ PRÁCE — **doc. RNDr. Libor Žák, Ph.D.**
SUPERVISOR

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání diplomové práce

Ústav:	Ústav matematiky
Studentka:	**Bc. Gabriela Hrubešová**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**doc. RNDr. Libor Žák, Ph.D.**
Akademický rok:	2017/18

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Vícerozměrné regresní modely

Stručná charakteristika problematiky úkolu:

1. Popis technického nebo ekonomického problému
2. Tvorba regresního modelu
3. Aplikace na reálná data

Cíle diplomové práce:

– teoretické základy pro regresní analýzu
– tvorba vícerozměrného regresního modelu
– nelineární vícerozměrné modely
– popis procesu, na který se bude aplikovat regresní analýza
– provedení regresní analýzy
– vyhodnocení výsledků

Seznam doporučené literatury:

ANDĚL, J. Základy matematické statistiky. 3. opr. vyd. Praha: Matfyzpress, 2011. ISBN 978-80-73-8-001-2.

ZVÁRA, K. Regresní analýza. Praha: Academia, 1989. ISBN: 80-200-0125-5 .

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2017/18

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

Abstrakt

Předmětem této diplomové práce je využití poznatků vícerozměrných regresních modelů v praxi. V první části jsou popsány teoretické základy pro regresní analýzu. Dále se věnujeme teorii tvorby modelu a nelineárním vícerozměrným modelům. V další části práce je popsán reálný problém. Jde o šířku mezery řezu slitin titanu po použití metody elektojiskrové řezání drátovou elektrodou (WEDM). Na tomto problému aplikujeme většinu teoretických poznatků a pomocí softwaru Minitab provedeme regresní analýzu. Data budeme analyzovat z různých pohledů a vytvoříme několik vícerozměrných regresních modelů. V poslední části shrneme výsledky získané regresní analýzou.

Abstract

The subject of this diploma thesis is the use of knowledge of multidimensional regression models in practice. The first part describes the theoretical basis for regression analysis. Then further we focus on modeling theory and nonlinear multidimensional models. The next part describes a real problem. This is the width of kerf of titanium alloy using wire electrical discharge machining (WEDM). We apply most of the theoretical knowledge to this problem, and we use regression analysis using Minitab software. We will analyze the data from different perspectives and create several multidimensional regression models. In the last part we summarize the results obtained by the regression analysis.

klíčová slova

vícerozměrná regresní analýza, lineární model, šířka mezery řezu, WEDM, titan

key words

multidimensional regression analysis, linear model, width of kerf, WEDM, titanium

HRUBEŠOVÁ, G. *Vícerozměrné regresní modely*. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2018. 83 s. Vedoucí diplomové práce doc. RNDr. Libor Žák, Ph.D..

Čestné prohlášení

Prohlašuji, že předložená bakalářská práce je původní a zpracovala jsem ji samostatně. Prohlašuji, že citace použitých pramenů je úplná, že jsem ve své práci neporušila autorská práva (ve smyslu Zákona č. 121/2000 Sb., o právu autorském a o právech souvisejících s právem autorským).

V Brně dne 24. května 2018

.....

podpis studenta

Děkuji svému školiteli panu doc. RNDr. Liboru Žákovi, Ph.D. za četné rady a připomínky při vedení mé diplomové práce. Také bych chtěla poděkovat oponentovi práce panu Ing. Josefu Bednářovi, Ph.D. za odbornou oponenturu.

Bc. Gabriela Hrubešová

Obsah

ÚVOD **15**

1 TEORETICKÁ ČÁST **17**

1.1 Základy regresní analýzy 17

1.2 Lineární regresní model 18

1.2.1 Metoda nejmenších čtverců 18

1.2.2 Intervaly spolehlivosti 20

1.2.3 Testování základních hypotéz 21

1.3 Obecný lineární regresní model 22

1.4 Model s neúplnou hodností 23

1.5 Regrese se dvěma nezávisle proměnnými 24

1.6 Submodely a jejich testování 24

1.7 Ověřování předpokladů modelu 26

1.7.1 Stálost rozptylu 26

1.7.2 Nezávislost pozorování 28

1.7.3 Normalita rozdělení odchylek 30

1.8 Multikolinearita 31

1.8.1 Odhalování multikolinearity 31

1.8.2 Vychýlené odhady 33

1.9 Volba vícerozměrného regresního modelu 34

1.9.1 Porovnávání modelů 34

1.9.2 Výběr podmnožin regresorů 38

1.9.3 Transformace závisle proměnné 39

1.10 Nelineární regresní modely 41

1.10.1 Exponenciální regresní funkce 41

1.10.2 Modifikovaný exponenciální trend 42

1.10.3 Logistický trend 42

1.10.4 Gompertzova křivka 42

1.10.5 Kompartmentový model 42

1.10.6 Michaelisův - Mentenův model 42

2 POPIS PROBLÉMU **43**

2.1 Titan 43

2.2 Slitiny titanu 44

2.2.1 Slitina Ti-6Al-4V 45

2.3 Elektroerozivní metody obrábění 45

2.3.1 Princip elektroerozivního obrábění 46
2.4 Elektrojiskrové řezání drátovou elektrodou 46
2.5 Průběh měření dat 47

3 ANALÝZA DAT A TVORBA MODELŮ 49
3.1 Analýza a tvorba modelu pro datový soubor MEAN DATA 50
3.2 Analýza a tvorba modelu pro datový soubor TOTAL DATA 63
3.3 Analýza a tvorba modelu pro datový soubor FIRST PART DATA 66
3.4 Analýza a tvorba modelu pro datový soubor SECOND PART DATA . . . 69

4 ZÁVĚREČNÉ ZHODNOCENÍ MODELŮ 72
4.1 Porovnání modelů pro FIRST a SECOND DATA 73
4.2 Optimální nastavení parametrů stroje 74

ZÁVĚR 77

SEZNAM POUŽITÝCH ZDROJŮ 78

SEZNAM POUŽITÝCH ZKRATEK 80

SEZNAM GRAFŮ 81

SEZNAM OBRÁZKŮ 82

SEZNAM TABULEK 83

ÚVOD

Při výrobě součástek je důležitá přesnost. Metoda elektrojiskrového řezání drátovou elektrodou neboli WEDM tuto přesnost umožňuje. Díky této metodě jsme schopni vyrábět složité tvary z materiálů, které se klasickými metodami dají jen těžko obrobit. To je velmi užitečné v letectví a medicíně. V těchto odvětvích vyžadujeme rychlé a efektivní obrábění titanových slitin, které vyžadují přesnost a preciznost. Tak jako v každé výrobě však vyžadujeme spolu s přesností, produktivitou také úspornost.

Při obrábění metodou WEDM vzniká při odpařování materiálu šířka mezery řezu. Ta ovlivní konečný rozměr obráběné součástky, což zasahuje jak do přesnosti obrábění, tak to úspornosti. Slitiny titanu bývají velice drahé a my chceme výrábět ekonomicky. Při větší šířce mezery řezu vzniká větší odpad materiálu a dochází tak k neefektivní výrobě. Šířku mezery řezu nejsme prozatím schopni zcela odstranit. Můžeme se ji však pokusit minimalizovat a to správným nastavením parametrů stroje, který provádí obráběcí proces.

Při procesu WEDM pomocí CNC stroje nastavujeme parametry gap voltage (V), pulse on time (μs), pulse off time (μs), wire speed (m/min) a dicharge current (A). Pokud bychom nalezli závislost mezi těmito parametry a šířkou mezery řezu a byli bychom schopni tuto závislost popsat, dokázali bychom pochopit chování šířky mezery řezu při obrábění titanových slitin.

Pro nalezení a popsání závislosti nastavení CNC stroje a šířky mezery řezu je vhodné využít vícerozměrné regresní modely. Díky těmto modelům jsme schopni s určitou přesností popsat chování tohoto jevu. Když se podaří nalézt závislost, můžeme určit optimální nastavení, které zajistí minimální šířku mezery řezu. Tak dosáhneme ekonomické, efektivní a přesné výroby. Cílem této práce bude nalézt závislost mezi parametry stroje a šířkou mezery řezu a popsat ji vícerozměrným regresním modelem. Pokud se závislost podaří nalézt, zjistíme, zda se šířka mezery řezu chová jinak na začátku řezu. Pokusíme se také o zjistění optimálního nastavení stroje.

V první části se budeme věnovat teoretickým poznatkům. Uvedeme jakýsi průřez regresní analýzou. Položíme základy regresní analýzy, budeme se věnovat lineárnímu regresnímu modelu, jeho tvorbě a testování. Pro zúplnění přehledu si nastíníme i nelineární regresní modely.

Další část věnujeme popisu zkoumaného problému. Abychom pochopili celou podstatu, rozebereme stručně materiál a metodu obrábění, které se v problému vyskytují. Dozvíme se něco o titanu a o titanových slitinách, dále o elektroerozivních metodách obrábění a technologii WEDM. Poté přejdeme k popisu měření dat, která budeme zkoumat.

Nasleduje část práce, která řeší analýzu dat a tvorbu vícerozměrných regresních modelů. Provedeme regresní analýzu z několika pohledů na data s pomocí softwaru Minitab.

Vytvořené modely využijeme k odpovědím na otázky, zda existuje a jaká je závislost mezi parametry stroje a šířkou mezery řezu a zda je tato šířka mezery řezu jiná při počáteční fázi řezu a ve zbylé fázi.

V poslední části zhodnotíme modely. Zjistíme, zda se šířka mezery řezu chová jinak na začátku řezu a nalezneme optimální nastavení CNC stroje.

1 TEORETICKÁ ČÁST

V této části se budeme věnovat teorii regresní analýzy. Položíme si její základy. Dále si popíšeme lineární regresní model. Mimo jiné budeme studovat tvorbu modelu a nelineární regresní modely. Soustředíme se však převážně na lineární regresní modely.

1.1 Základy regresní analýzy

Definice 1.1. Mějme náhodné veličiny $Y, X_1, X_2, ..., X_n$. Naším cílem je nalézt co nejlepší lineární aproximaci veličiny Y pomocí veličin $X_1, ..., X_n$. Předpokládejme, že náhodné veličiny $Y, X_1, X_2, ..., X_n$ mají konečné druhé momenty. Lineární aproximace bude tvaru

$$\alpha + \beta_1 X_1 + ... + \beta_n X_n = \alpha + \boldsymbol{\beta}'\boldsymbol{X}, \tag{1.1}$$

kde $\boldsymbol{\beta} = (\beta_1, ...\beta_n)'$ a $\boldsymbol{X} = (X_1, ..., X_n)'$.

Zda je námi zvolená aproximace kvalitní zjistíme pomocí střední kvadratické chyby $\mathrm{E}(Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X})^2$ [1].

Věta 1.2. Mějme regulární matici $\boldsymbol{V} = \mathrm{var}\ \boldsymbol{X}$. Pak podle [1] platí nerovnost

$$\mathrm{E}(Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X})^2 \geq \mathrm{var}\ Y - \mathrm{cov}\ (Y, \boldsymbol{X})\boldsymbol{V}^{-1}\mathrm{cov}\ (\boldsymbol{X}, Y).$$

Rovnost bude platit právě tehdy, když

$$\boldsymbol{\beta} = \boldsymbol{V}^{-1}\mathrm{cov}\ (\boldsymbol{X}, Y), \qquad \alpha = \mathrm{E}Y - \boldsymbol{\beta}'\mathrm{E}\boldsymbol{X}.$$

Důkaz: Mějme libovolnou náhodnou veličinu Z s rozptylem, který je konečný. Pak platí, že $\mathrm{E}Z^2 = \mathrm{var}\ Z + (\mathrm{E}Z)^2$. Pokud položíme $Z = Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X}$, pak dostaneme

$$\mathrm{E}(Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X})^2 \geq \mathrm{var}\ (Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X}).$$

Rovnost je dosažena, jestliže $\mathrm{E}(Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X}) = 0$, to znamená, že $\alpha = \mathrm{E}Y - \boldsymbol{\beta}'\mathrm{E}\boldsymbol{X}$. Dále mějme

$$\mathrm{var}\ (Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X}) = \mathrm{var}\ (Y - \boldsymbol{\beta}'\boldsymbol{X})$$

$$= \mathrm{var}\ (Y) - \boldsymbol{\beta}'\mathrm{cov}\ (\boldsymbol{X}, Y) - \mathrm{cov}\ (Y, \boldsymbol{X})\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}$$

$$= \left[\boldsymbol{\beta} - \boldsymbol{V}^{-1}\mathrm{cov}\ (\boldsymbol{X}, Y)\right]'\boldsymbol{V}\left[\boldsymbol{\beta} - \boldsymbol{V}^{-1}\mathrm{cov}\ (\boldsymbol{X}, Y)\right] + \mathrm{var}Y - \mathrm{cov}(Y, \boldsymbol{X})\boldsymbol{V}^{-1}\mathrm{cov}(\boldsymbol{X}, Y)$$

$$\geq \mathrm{var}\ Y - \mathrm{cov}\ (Y, \boldsymbol{X})\boldsymbol{V}^{-1}\mathrm{cov}\ (\boldsymbol{X}, Y)$$

a rovnosti je dosaženo, právě tehdy, když $\boldsymbol{\beta} = \boldsymbol{V}^{-1}\mathrm{cov}\ (\boldsymbol{X}, Y)$ [1].

Nejlepší lineární aproximací náhodné veličiny Y je tedy veličina $\hat{Y} = \alpha + \boldsymbol{\beta}'\boldsymbol{X}$, kde $\boldsymbol{\beta} = \boldsymbol{V}^{-1}\mathrm{cov}\ (\boldsymbol{X}, Y)$ a $\alpha = \mathrm{E}Y - \boldsymbol{\beta}'\mathrm{E}\boldsymbol{X}$ [1].

Definice 1.3. Vzorec

$$\mathrm{E}(Y - \hat{Y})^2 = \sigma_{Y.\boldsymbol{X}} = \mathrm{var}\, Y - \mathrm{cov}\,(Y, \boldsymbol{X})\boldsymbol{V}^{-1}\mathrm{cov}\,(\boldsymbol{X}, Y) \tag{1.2}$$

udává reziduální rozptyl. Po úpravě dostaneme vzorec

$$\sigma_{Y.\boldsymbol{X}} = \mathrm{var}\, Y - \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}, \tag{1.3}$$

kde $\boldsymbol{\beta} = \boldsymbol{V}^{-1}\mathrm{cov}\,(\boldsymbol{X}, Y)$ [1].

1.2 Lineární regresní model

Definice 1.4. Nechť $Y_1, ..., Y_n$ jsou náhodné veličiny a $\boldsymbol{X} = (x_{ij})$ je matice daných čísel typu $n \times k$, kde $k < n$. Předpokládejme, že pro náhodný vektor $\boldsymbol{Y} = (Y_1, ..., Y_n)'$ platí

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.4}$$

kde $\boldsymbol{\beta} = (\beta_1, ...\beta_k)'$ je vektor neznámých parametrů a $\boldsymbol{e} = (e_1, ..., e_n)'$ je náhodný vektor, pro který platí $\mathrm{E}\boldsymbol{e} = \boldsymbol{0},\ \mathrm{var}\ \boldsymbol{e} = \sigma^2\boldsymbol{I}$. Neznámým parametrem je také $\sigma^2 > 0$. Model uvedený v rovnici 1.4 se nazývá lineární regresní model. Model je lineární, protože $\boldsymbol{Y}$ závisí na $\boldsymbol{\beta}$ lineárně [1].

V každém regresním modelu zahrnujeme vysvětlující proměnné x_{ij}, které jsou pro model významné. Matice $\boldsymbol{X}$ by měla mít lineárně nezávislé sloupce [1]. V případě, že by matice $\boldsymbol{X}$ měla lineárně závislé sloupce, lineární odhad vektoru $\boldsymbol{\beta}$ neexistuje [2]. Dříve jsme předpokládali $k < n$, takže $h(\boldsymbol{X}) = k$ z toho také vyplývá, že matice $\boldsymbol{X}'\boldsymbol{X}$ je regulární [1].

Dále platí $\mathrm{E}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$ a $\mathrm{var}\ \boldsymbol{Y} = \sigma^2\boldsymbol{I}$, protože vektor $\boldsymbol{X}\boldsymbol{\beta}$ je nenáhodný [1].

1.2.1 Metoda nejmenších čtverců

Abychom mohli celý model popsat, potřebujeme odhadnout parametry $\beta_1, ..., \beta_k$. Tyto odhady získáme metodou nejmenších čtverců, jejímž jádrem je podmínka minimálnosti výrazu $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. Odhady těchto parametrů značíme $\boldsymbol{b} = (b_1, ..., b_k)'$ [1].

Věta 1.5. Odhady parametrů $\beta_1, ..., \beta_k$ metodou nejmenších čtverců jsou dle [1]:

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{1.5}$$

Důkaz: Jelikož platí $\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = 0$, dostáváme

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = [(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + (\boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta})]'\,[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + (\boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta})]$$

$$= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + (\boldsymbol{b} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{\beta})$$

$$\geq (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})$$

Z konstrukce matice $\boldsymbol{X}'\boldsymbol{X}$ plyne její pozitivní definitnost. Proto musí rovnost nastat právě tehdy, když $\boldsymbol{\beta} = \boldsymbol{b}$ [1].

Definice 1.6. Soustavu lineárních rovnic $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$ s neznámou $\boldsymbol{b}$ nazveme soustavou normálních rovnic. Nejlepší aproximací vektoru $\boldsymbol{Y}$, vytvořenou lineární kombinací sloupců matice $\boldsymbol{X}$, je vektor $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ [1].

Věta 1.7. Jelikož je na obou stranách soustavy nějaká lineární kombinace řádků matice $\boldsymbol{X}$, je tato soustava lineárních rovnic vždy řešitelná. To však neznamená, že řešení bude jednoznačné [2].

Označme $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ a $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{H}$. Z tohoto plyne, že $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$. Matice $\boldsymbol{H}$ je projekční matice, je symetrická a idempotentní, stejně tak matice $\boldsymbol{M}$ je symetrická a idempotentní [1]. Obě tyto matice jsou dány jednoznačně [2].

Věta 1.8. Hodnost matice $\boldsymbol{H}$ je k a hodnost matice $\boldsymbol{M}$ je $n - k$ [1].

Důkaz: $h(\boldsymbol{H}) = \mathrm{Tr}\ \boldsymbol{H} = \mathrm{Tr}\ \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \mathrm{Tr}\ \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \mathrm{Tr}\ \boldsymbol{I_k} = k$ a $h(\boldsymbol{M}) = n - k$ [1].

Definice 1.9. Rozdíl $Y_i - \hat{Y}_i$ pro $i = 1, ..., k$ se nazývá reziduum veličiny Y_i [3].

Definice 1.10. Veličinu $S_e = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}})$ nazveme reziduální součet čtverců [1]. Pro tuto veličinu platí $S_e = \boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y}$ a $S_e = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y}$ [1].

Důkaz:

$$S_e = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = (\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y})'(\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}) = (\boldsymbol{M}\boldsymbol{Y})'(\boldsymbol{M}\boldsymbol{Y}) = \boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y} =$$
$$= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y}$$

Věta 1.11. Odhad $\boldsymbol{b}$ metodou nejmenších čtverců je nejlepším nestranným odhadem vektoru $\boldsymbol{\beta}$, tzn. $\mathrm{E}\boldsymbol{b} = \boldsymbol{\beta}$ a $\mathrm{var}\ \boldsymbol{b} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ [1, 2].

Důkaz:

$$\mathrm{E}\boldsymbol{b} = \mathrm{E}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{E}\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$
$$\mathrm{var}\ \boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\mathrm{var}\ \boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$
$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Věta 1.12. Mějme náhodnou veličinu $s^2 = \frac{S_e}{(n-k)}$. Tato náhodná veličina je nestranným odhadem parametru σ^2 [1].

Důkaz:

Abychom mohli určit těsnost regresní závislosti modelu na reálných datech, musíme zavést koeficient determinace R^2 a upravený koeficient determinace $\bar{R}^2$. Hodnoty blíže jedné znamenají těsnější regresní závislost [1].

$$R^2 = 1 - \frac{S_e}{S_T}, \qquad \bar{R}^2 = 1 - \frac{n-1}{n-(k-1)-1}\frac{S_e}{S_T}, \tag{1.6}$$

kde

$$S_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2.$$

Veškeré pojmy zavedené po tuto část nevyžadují předpoklad normality. V dalším textu budeme předpokládat normalitu reziduí, tedy že $\boldsymbol{e} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Z tohoto vyplývá, že $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ [1].

Věta 1.13. Jelikož $\boldsymbol{b}$ vzniká lineární transformací z $\boldsymbol{Y}$, pak vektor $\boldsymbol{b}$ musí mít také normální rozdělení, tedy $\boldsymbol{b} \sim \mathrm{N}[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X'X})^{-1}]$ [1].

1.2.2 Intervaly spolehlivosti

V této části budeme předpokládat, že složky náhodného vektoru $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ vyhovují vztahu

$$Ey_i = \sum_{j=1}^{k} \beta_j x_j(w_i) = \boldsymbol{\beta'x}(w_i),$$

$i = 1, ..., n$, $x_i(\boldsymbol{w}), ..., x_k(\boldsymbol{w})$ jsou známé funkce, w_i jsou známá pevná čísla a funkce $\boldsymbol{x}(\boldsymbol{w}) = (x_1(\boldsymbol{w}), ..., x_k(\boldsymbol{w}))'$ je vektorová funkce argumentu $\boldsymbol{w}$ [4].

Nejlepším nestranným lineárním odhadem parametrické funkce $\boldsymbol{\beta'x}(w)$ je statistika $\boldsymbol{b'x}(w)$. Tato statistika má rozdělení

$$N(\boldsymbol{\beta'x}(w), \sigma^2 d^2(w)),$$

kde $d(w) = [\boldsymbol{x'}(w)(\boldsymbol{X'X})^{-}\boldsymbol{x}(w)]^{1/2}$ [4].

Pro pevné w^* je interval

$$\left\langle \boldsymbol{b'x}(w^*) - sd(w^*)t_{n-p}\left(1 - \frac{\alpha}{2}\right), \boldsymbol{b'x}(w^*) + sd(w^*)t_{n-p}\left(1 - \frac{\alpha}{2}\right)\right\rangle \tag{1.7}$$

intervalem spolehlivosti pro $\boldsymbol{\beta'x}(w^*)$ s koeficientem spolehlivosti $1 - \alpha$. Pro dané w^* překrývá tento interval neznámou hodnotu $\boldsymbol{\beta'x}(w^*)$ s předem danou pravděpodobností $1 - \alpha$ [4].

Máme budoucí pozorování

$$Y^* = Y(w^*) = \boldsymbol{\beta'x}(w^*) + e^*,$$

kde $e^* \sim N(0, \sigma^2)$ je náhodná veličina nezávislá na $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X\beta}$. Pro pevné w^* dostaneme interval

$$\left\langle \boldsymbol{b'x}(w^*) - s(1 + d^2(w^*))^{1/2}t_{n-p}\left(1 - \frac{\alpha}{2}\right), \boldsymbol{b'x}(w^*) + s(1 + d^2(w^*))^{1/2}t_{n-p}\left(1 - \frac{\alpha}{2}\right)\right\rangle,$$

který překryje jednu realizaci náhodné veličiny $Y^* = Y(w^*)$ s pravděpodobností $1 - \alpha$ [4].

Platí, že $L(w) \leq \boldsymbol{\beta}'\boldsymbol{x}(w) \leq U(w)$. Dále platí vztah

$$L(w) = \boldsymbol{b}'\boldsymbol{x}(w) - sd(w)\left[pF_{p,n-p}(1-\alpha)\right]^{1/2} \tag{1.8}$$

a obdobně

$$U(w) = \boldsymbol{b}'\boldsymbol{x}(w) + sd(w)\left[pF_{p,n-p}(1-\alpha)\right]^{1/2} \tag{1.9}$$

což nám dává konfidenční pás, tedy pás spolehlivosti, který překryje neznámou regresní funkci s danou pravděpodobností. Přesné odvození je uvedeno v [4].

1.2.3 Testování základních hypotéz

Věta 1.14. Prvky matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ označíme v_{ij}. Dále mějme $T_i = \frac{(b_i-\beta_i)}{\sqrt{s^2 v_{ii}}}$. Pak platí, že $T_i \sim t_{n-k}$ pro všechna $i = 1, ..., k$ [1].

Hypotéza $H_0 : \beta_i = \beta_i^0$, kde i je pevně zvolené číslo. Alternativní hypotéza bude tedy $H_1 : \beta_i \neq \beta_i^0$ [1].

Pokud platí

$$T = \frac{|b_i - \beta_i^0|}{\sqrt{s^2 v_{ii}}} \geq t_{n-k}(\alpha),$$

pak hypotézu H_0 zamítneme na hladině významnosti α [1].

Hypotéza $H_0 : \beta_i^0 = 0$, kde i je pevně zvolené číslo. Alternativní hypotéza bude tedy $H_1 : \beta_i^0 \neq 0$. Tato hypotéza se využívá nejčastěji k otestování, zda $\boldsymbol{Y}$ závisí na i-tém sloupci matice $\boldsymbol{X}$. Pokud bychom potvrdili hypotézu H_0, pak můžeme i-tý sloupec matice $\boldsymbol{X}$ vypustit a získat tak jednodušší model [1].

Pokud potřebujeme testovat více parametrů najednou. Nechť

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix}, \qquad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} \boldsymbol{V} & \boldsymbol{U} \\ \boldsymbol{U}' & \boldsymbol{W} \end{pmatrix},$$

kde $\boldsymbol{\beta}_1$ a $\boldsymbol{b}_1$ mají p složek, $\boldsymbol{\beta}_2$ a $\boldsymbol{b}_2$ mají q složek. Platí, že $p + q = k$. $\boldsymbol{V}$ je matice typu $p \times p$ a $\boldsymbol{W}$ je typu $q \times q$ [1].

Věta 1.15. Zaveďme statistiku Z, pro kterou platí

$$Z = \frac{1}{qs^2}(\boldsymbol{b}_2 - \boldsymbol{\beta}_2)'\boldsymbol{W}^{-1}(\boldsymbol{b}_2 - \boldsymbol{\beta}_2) \sim F_{q,n-k}.$$

Tuto statistiku využijeme k dalšímu testování [1].

Hypotéza $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$ a alternativní hypotéza bude $H_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_2^0$ [1].

Pokud platí

$$Z = \frac{1}{qs^2}(\boldsymbol{b}_2 - \boldsymbol{\beta}_2^0)'\boldsymbol{W}^{-1}(\boldsymbol{b}_2 - \boldsymbol{\beta}_2^0) \geq F_{q,n-k}(\alpha),$$

pak hypotézu H_0 zamítneme na hladině významnosti α [1].

Hypotéza $H_0 : \boldsymbol{\beta}_2^0 = \mathbf{0}$ a alternativní hypotéza bude $H_1 : \boldsymbol{\beta}_2^0 \neq \mathbf{0}$. Touto hypotézou testujeme, zda $\boldsymbol{Y}$ závisí na posledních q sloupcích matice $\boldsymbol{X}$. Pokud bychom potvrdili hypotézu H_0, pak můžeme posledních q sloupců matice $\boldsymbol{X}$ vypustit [1].

Dalším testem může být případ, kdy chceme zjisti, zda všechny parametry jsou nulové.

Hypotéza $H_0 : \boldsymbol{\beta} = \mathbf{0}$ a alternativní hypotéza bude $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$ [1].

Pokud platí

$$Z = \frac{\boldsymbol{Y}'\boldsymbol{Y} - S_e}{ks^2} \geq F_{k,n-k}(\alpha),$$

pak hypotézu H_0 zamítneme na hladině významnosti α [1].

V případě, že potřebujeme testovat konkrétní lineární kombinaci složek vektoru $\boldsymbol{\beta}$ využíváme následující postup [1].

Věta 1.16. Nechť $\boldsymbol{c} = (c_1, ..., c_k)'$ je daný nenulový vektor. Dále $\mathrm{E}\boldsymbol{c}'\boldsymbol{b} = \boldsymbol{c}'\boldsymbol{\beta}$ [1]. Pak platí

$$T = \frac{\boldsymbol{c}'\boldsymbol{b} - \boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim t_{n-k}.$$

Pomocí statistiky T můžeme určit oboustranný interval spolehlivosti pro $\boldsymbol{c}'\boldsymbol{\beta}$

$$\left\langle \boldsymbol{c}'\boldsymbol{b} - t_{n-k}\left(1 - \frac{\alpha}{2}\right)\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}},\ \boldsymbol{c}'\boldsymbol{b} + t_{n-k}\left(1 - \frac{\alpha}{2}\right)\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}} \right\rangle,$$

s koeficientem spolehlivosti $1 - \alpha$ [1].

1.3 Obecný lineární regresní model

Definice 1.17. Obecný lineární regresní model je ve tvaru

$$y_i = \sum_{j=1}^{k} \beta_j f_j(x_{1i}, ..., x_{li}) + e_i, \tag{1.10}$$

kde $f_1, ..., f_k$ jsou funkce známého tvaru proměnných $x_1, ..., x_l$ a neobsahují žádné neznámé parametry. Čísla $x_{1i}, ..., x_{li}$ jsou konkrétní hodnoty proměnných $x_1, ..., x_l$ [3].

Odhady parametrů $\beta_1, ..., \beta_k$ metodou nejmenších čtverců splňují podmínku

$$\sum_{i=1}^{n}[y_i - \sum_{j=1}^{k} b_j f_j(x_{1i}, ..., x_{li})]^2 = min. \tag{1.11}$$

Tato podmínka znamená minimálnost součtu kvadrátů čtverců reziduí [3].

1.4 Model s neúplnou hodností

Důkazy pro tuto část jsou uvedeny v [1].

Pokud má matice $\boldsymbol{X}$, která je tvaru $n \times k$, hodnost menší než k, mluvíme o modelu s neúplnou hodností [1]. Nechť platí $\mathrm{var}\, \boldsymbol{Y} = \sigma^2 \boldsymbol{I}$, kde $\sigma^2 > 0$ je neznámý parametr [1].

Věta 1.18. Pokud je $\boldsymbol{b}$ řešením soustavy rovnic

$$\boldsymbol{X'Xb} = \boldsymbol{X'Y}, \tag{1.12}$$

pak výraz

$$(\boldsymbol{Y} - \boldsymbol{Xb})'(\boldsymbol{Y} - \boldsymbol{Xb})$$

nabývá vzhledem k $\boldsymbol{b}$ nejmenší hodnoty. Tato hodnota je stejná pro všechna $\boldsymbol{b}$, která jsou řešením soustavy rovnic výše [1].

Věta 1.19. Soustavu 1.12 ($\boldsymbol{X'Xb} = \boldsymbol{X'Y}$) můžeme zapsat ve tvaru

$$\frac{\partial}{\partial b_j}(\boldsymbol{Y} - \boldsymbol{Xb})'(\boldsymbol{Y} - \boldsymbol{Xb}) = 0, \quad j = 1, ..., k,$$

což je ekvivalentní zápis.

Věta 1.20. Parametr $\theta = \boldsymbol{c'\beta}$ jsme schopni odhadnout právě tehdy, když je $\boldsymbol{c}$ nějaká lineární kombinace řádků matice $\boldsymbol{X}$ [1].

Věta 1.21. Parametr $\theta = \boldsymbol{c'\beta}$ jsme schopni odhadnout právě tehdy, když je θ nějaká lineární kombinace složek vektoru $\mathbf{E}\boldsymbol{Y}$ [1].

Věta 1.22. Řekněme, že parametr $\theta = \boldsymbol{c'\beta}$ můžeme odhadnout. Pak nejlepší nestranný lineární odhad takovéhoto parametru je $\hat{\theta} = \boldsymbol{c'b}$, kde $\boldsymbol{b}$ je libovolné řešení soustavy $\boldsymbol{X'Xb} = \boldsymbol{X'Y}$. Hodnota $\boldsymbol{c'b}$ je stejná pro všechna řešení dané soustavy $\boldsymbol{X'Xb} = \boldsymbol{X'Y}$ [1].

Věta 1.23. Vektor $\boldsymbol{\theta} = \mathbf{E}\boldsymbol{Y}$ jsme schopni odhadnout vždy a nejlepší nestranným lineárním odhadem pro $\boldsymbol{\theta}$ je podle [1]

$$\hat{\boldsymbol{\theta}} = \boldsymbol{X}(\boldsymbol{X'X})^{-}\boldsymbol{X'Y}.$$

Věta 1.24. Nyní pro reziduální součet čtverců $S_e = (\boldsymbol{Y} - \boldsymbol{Yb})'(\boldsymbol{Y} - \boldsymbol{Yb})$ platí dle [1]

$$S_e = \boldsymbol{Y}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-}\boldsymbol{X'}]\boldsymbol{Y}. \tag{1.13}$$

Věta 1.25. Mějme $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ a hodnost matice $\boldsymbol{X}$ je r. Pak $\frac{S_e}{\sigma^2} \sim \chi^2_{n-r}$ a $s^2 = \frac{S_e}{(n-r)}$ je nestranným odhadem parametru σ^2 [1].

Věta 1.26. Pokud má vektor $\boldsymbol{Y}$ normální rozdělení, pak jsou vektor $\boldsymbol{X'Y}$ a veličina S_e nezávislé a také vektor $\boldsymbol{b}$ a veličina s^2 jsou nezávislé [1].

1.5 Regrese se dvěma nezávisle proměnnými

Mějme $n \geq 4$ a $Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i$, kde $i = 1, ..., n$ a x_i a z_i jsou daná čísla získaná měřením. Pak

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ & \vdots & \\ 1 & x_n & z_n \end{pmatrix}, \qquad \boldsymbol{X'X} = \begin{pmatrix} n & \sum x_i & \sum z_i \\ \sum x_i & \sum x_i^2 & \sum x_i z_i \\ \sum z_i & \sum x_i z_i & \sum z_i^2 \end{pmatrix},$$

$$\boldsymbol{X'Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum z_i Y_i \end{pmatrix}.$$

Odhad parametrů β_i tedy b_i vypočteme podle [1] takto $\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$.

1.6 Submodely a jejich testování

Tato kapitola bude názorně vysvětlena na lineárním modelu se dvěma submodely. Pokud bychom měli větší počet submodelů, postup bude analogický [1]. Důkazy některých tvrzení nalezneme v [1].

Nechť

$$\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\alpha}, \sigma^2\boldsymbol{I})$$

je lineární model, kde $\boldsymbol{Y} = (Y_1, ..., Y_n)'$ je náhodný vektor, $\boldsymbol{X}$ je typu $n \times k$, $\boldsymbol{\alpha}$ je k-rozměrný vektor neznámých parametrů, σ^2 je neznámý parametr a $\boldsymbol{I}$ je typu $n \times n$. Výše uvedený model označíme jako model M [1].

Berme v úvahu také modely M_1 a M_2 [1].

$$M_1 : \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{U}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \tag{1.14}$$

$$M_2 : \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{T}\boldsymbol{\gamma}, \sigma^2\boldsymbol{I}), \tag{1.15}$$

kde $\boldsymbol{U}$ je typu $n \times k_1$, vektor $\boldsymbol{\beta}$ má k_1 složek, $\boldsymbol{T}$ je typu $n \times k_2$ a $\boldsymbol{\gamma}$ má k_2 složek. Dále předpokládáme, že hodnost matice $\boldsymbol{X}$ je r, hodnost matice $\boldsymbol{U}$ je r_1 a hodnost matice $\boldsymbol{T}$ je r_2 [1].

Model M_1 je submodelem modelu M, pokud každý sloupec matice $\boldsymbol{U}$ patří do lineárního obalu sloupců matice $\boldsymbol{X}$ a zároveň $r > r_1$ [1].

Pokud bude existovat matice $\boldsymbol{K}$, která bude typu $k \times k_1$ a bude platit, že

$$\boldsymbol{U} = \boldsymbol{X}\boldsymbol{K}$$

pak právě tehdy bude každý sloupec matice $\boldsymbol{U}$ patřit do lineárního obalu sloupců matice $\boldsymbol{X}$ [1].

Model M_2 je submodelem modelu M_1, pokud každý sloupec matice $\boldsymbol{T}$ patří do lineárního obalu sloupců matice $\boldsymbol{U}$ a zároveň $r_1 > r_2$ [1].

Pokud bude existovat matice $\boldsymbol{L}$, která bude typu $k_1 \times k_2$ a bude platit, že

$$\boldsymbol{T} = \boldsymbol{U}\boldsymbol{L}$$

pak právě tehdy bude každý sloupec matice $\boldsymbol{T}$ patřit do lineárního obalu sloupců matice $\boldsymbol{U}$ [1].

Dostáváme tedy $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{L} = \boldsymbol{X}\boldsymbol{K}\boldsymbol{L}$. Submodely a modely téměř vždy budeme volit tak, aby platilo $k > k_1 > k_2 \geq 1$ [1].

Pokud je M_1 submodelem modelu M, pak pokud platí pro $\boldsymbol{Y}$ model M_1, pak platí pro $\boldsymbol{Y}$ také model M. Pokud platí pro $\boldsymbol{Y}$ model M_2, pak platí pro $\boldsymbol{Y}$ také model M i model M_1 [1].

Nechť

$$\hat{\boldsymbol{\mu}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \quad \hat{\boldsymbol{\upsilon}} = \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{Y} \quad \hat{\boldsymbol{\tau}} = \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y}.$$

Pak nejlepším nestranným lineárním odhadem $\mathrm{E}\boldsymbol{Y}$ je $\hat{\boldsymbol{\mu}}$ u modelu M, $\hat{\boldsymbol{\upsilon}}$ u modelu M_1, $\hat{\boldsymbol{\tau}}$ u modelu M_2 [1].

Model M se vybírá tak, aby s jistotou popisoval dobře chování $\boldsymbol{Y}$. Chceme však zjistit, zda nemůžeme použít jednodušší model M_1 nebo dokonce až model M_2 [1]. Tento výběr budeme provádět na základě následující tabulky (Tabulka 1.).

Tabulka 1: Výběr modelu a submodelu. [vlastní zpracování]

Vztah mezi vektory	**Rozšíření nebo redukce**
$\hat{\boldsymbol{\upsilon}}$ a $\hat{\boldsymbol{\mu}}$ se znatelně neliší	M_1 se nevyplatí rozšiřovat na M
$\hat{\boldsymbol{\upsilon}}$ a $\hat{\boldsymbol{\mu}}$ jsou znatelně odlišné	M nemohu redukovat na M_1
$\hat{\boldsymbol{\tau}}$ a $\hat{\boldsymbol{\upsilon}}$ se znatelně neliší	M_2 se nevyplatí rozšiřovat na M_1
$\hat{\boldsymbol{\tau}}$ a $\hat{\boldsymbol{\upsilon}}$ jsou znatelně odlišné	M_1 nemohu redukovat na M_2

Mějme $S_e = (\boldsymbol{Y} - \hat{\boldsymbol{\mu}})'(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$ a $s^2 = \frac{S_e}{(n-r)}$ [1].

Věta 1.27. Pokud pro $\boldsymbol{Y}$ platí model M_1, pak má náhodná veličina

$$F_1 = \frac{(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{v}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{v}})}{r - r_1} \frac{1}{s^2} \tag{1.16}$$

rozdělení $F_{r-r_1,n-r}$ [1].

Pokud platí model M_2, pak má náhodná veličina

$$F_2 = \frac{(\hat{\boldsymbol{v}} - \hat{\boldsymbol{\tau}})'(\hat{\boldsymbol{v}} - \hat{\boldsymbol{\tau}})}{r_1 - r_2} \frac{1}{s^2} \tag{1.17}$$

rozdělení $F_{r1-r_2,n-r}$ [1].

Věta 1.28. Pro veličiny S_e, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{\tau}}$ podle [1] platí

$$S_e = \boldsymbol{Y'Y} - \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}}$$

$$(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{v}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{v}}) = \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{v}}'\hat{\boldsymbol{v}}$$

$$(\hat{\boldsymbol{v}} - \hat{\boldsymbol{\tau}})'(\hat{\boldsymbol{v}} - \hat{\boldsymbol{\tau}}) = \hat{\boldsymbol{v}}'\hat{\boldsymbol{v}} - \hat{\boldsymbol{\tau}}'\hat{\boldsymbol{\tau}}$$

Věta 1.29. Pro veličiny S_e, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{\tau}}$ dle [1] platí dále

$$S_e + (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{v}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{v}}) + (\hat{\boldsymbol{v}} - \hat{\boldsymbol{\tau}})'(\hat{\boldsymbol{v}} - \hat{\boldsymbol{\tau}}) = \boldsymbol{Y'Y} - \hat{\boldsymbol{\tau}}'\hat{\boldsymbol{\tau}}$$

1.7 Ověřování předpokladů modelu

V této části popíšeme způsoby, kterými můžeme ověřovat předpoklady modelů. Předpoklady modelu jsou stálost rozptylu, nezávislost pozorování a normalita rozdělení odchylek.

1.7.1 Stálost rozptylu

Pokud máme k dispozici nezávislé odhady $s_1^2, ..., s_r^2$ takové, že $\frac{f_i s_i^2}{\sigma_i^2}$ má rozdělení $\chi^2_{f_i}$, kde $i = 1, ..., r$, pak můžeme využít klasické testy k ověření hypotézy $\sigma_1^2 = ... = \sigma_r^2$ [4].

Následující tři testy můžeme využít v případě, že alespoň pro některé hodnoty nezávisle proměnné máme k dispozici opakovaná pozorování [4].

Věta 1.30. Bartlettův test využívá statistiku

$$B = \frac{(\sum_i f_i) \ln s^2 - \sum_i f_i \ln s_i^2}{C}, \tag{1.18}$$

kde

$$s^2 = \frac{\sum_i f_i s_i^2}{\sum_i f_i},$$

$$C = 1 + \frac{\sum_i \frac{1}{f_i} - \frac{1}{\sum_i f_i}}{3(r-1)}.$$

Pokud platí hypotéza, statistika má rozdělení χ^2_{r-1}, ale pouze pokud $f_i \geq 6$, $i = 1, ..., r$. Tuto statistiku velice ovlivní porušení předpokladu o normálním rozdělení náhodných veličin, z nichž jsou určeny odhady s_i^2 [4].

Věta 1.31. Cochranův test je omezen případem, kdy $f_1 = ... = f_r$. Využívá statistiku

$$\max_i \frac{s_i^2}{\sum_{i=1}^r s_i^2}. \tag{1.19}$$

Tento test je silný, pokud je jeden z rozptylů $\sigma_1^2 = ... = \sigma_r^2$ výrazně větší než ostatní [4].

Věta 1.32. Hartleyův test využívá statistiku

$$\frac{\max_i s_i^2}{\min_i s_i^2}. \tag{1.20}$$

Pro tento test musí opět platit $f_1 = ... = f_r$ [4].

Předchozí testy mají různá omezení. Univerzálním testem může být Goldfeld-Quandtův test. Testujeme hypotézu stálého rozptylu. Alternativní hypotézou je monotonní závislost na zvoleném regresoru $\boldsymbol{x}_{.l}$ [4]. Uspořádáme pozorování tak, aby platilo $x_{1l} \leq ... \leq x_{nl}$. V dalším kroku vypočítáme reziduální rozptyl s_1^2 z prvních q pozorování a nezávisle rozptyl s_2^2 z posledních q pozorování. Musí platit $q > p$ a zároveň $2q < n$. Obvykle bývá voleno $q \doteq \frac{n}{3}$ [4]. Tento test pracuje se statistikou

$$F = \frac{s_1^2}{s_2^2}, \tag{1.21}$$

která má v případě platnosti hypotézy rozdělení $F_{q-p,q-p}$ [4].

Další možností jsou rekurzivní rezidua. Opět musí platit $x_{1l} \leq ... \leq x_{nl}$. Pokud platí hypotéza, tvoří rezidua $r_1^*, ...r_{n-p}^*$ náhodný výběr z $N(0, \sigma^2)$. Volíme q tak, aby platilo $0 < 2q < n - p$. Nechť $\boldsymbol{r_1^*}$ obsahuje prvních q a $\boldsymbol{r_2^*}$ posledních q složek vektoru $\boldsymbol{r^*}$. Použijeme statistiku

$$F = \frac{\left\|\boldsymbol{r_1^*}\right\|}{\left\|\boldsymbol{r_2^*}\right\|}. \tag{1.22}$$

Pokud hypotéza platí, statistika má rozdělení $F_{q,q}$ [4].

Pokud díky testování zjistíme nestálost rozptylu, je třeba model přeformulovat tak, abychom dosáhli opaku. Nejčastěji se využívá vážená regrese nebo vhodně zvolená transformace [4].

1.7.2 Nezávislost pozorování

Tento problém se velice často objevuje u dat, která vlastně tvoří časovou řadu. Předpokládáme, že v regresní matici je sloupec jedniček, tedy, že regresní model obsahuje absolutní člen. Máme náhodné veličiny

$$Y_i = (\boldsymbol{x_{i.}})'\boldsymbol{\beta} + e_i,$$

kde $e_i \sim N(0, \sigma^2)$ a připustíme závislost náhodných veličin $e_1, ..., e_n$. Tyto veličiny budou tvořit autoregresní proces prvního řádu $e_i = \rho e_{i-1} + \epsilon_i$ a teprve ϵ_i jsou nezávislé veličiny. Pokud $\rho = 0$, dostaneme klasický normální lineární model [2].

Zda jsou po sobě jdoucí pozorování nezávislá ověřujeme pomocí Durbin-Watsonova testu. Ten používá statistiku

$$d = \frac{\sum_{i=1}^{n-1}(u_{i+1} - u_i)^2}{\sum_{i=1}^{n} u_i^2} = \frac{\boldsymbol{u'Au}}{\boldsymbol{u'u}} \tag{1.23}$$

kde $\boldsymbol{A}$ je matice

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix},$$

která je symetrická a pozitivně semidefinitní. Dále $\boldsymbol{u} = \boldsymbol{Me}$ a matici $\boldsymbol{M}$ můžeme zapsat pomocí ortonormální báze jako $\boldsymbol{M} = \boldsymbol{NN'}$[2].

Mějme hypotézu $H_0 : \rho = 0$ a pokusme se zjistit, jako rozdělení má statistika d při platnosti této hypotézy [2]. Zavedeme náhodný vektor

$$\boldsymbol{U} = \frac{1}{\sigma}\boldsymbol{N'e} \sim N(\boldsymbol{0}, \boldsymbol{I_{n-r}}).$$

Poté můžeme statistiku d zapsat jako

$$d = \frac{\boldsymbol{U'N'ANU}}{\boldsymbol{U'U}}.$$

Dále nalezneme spektrální rozklad $\boldsymbol{Q\Lambda Q'}$ k matici $\boldsymbol{N'AN}$. Matice $\boldsymbol{Q}$ je ortonormální matice řádu $n - r$ a $\boldsymbol{\Lambda}$ je diagonální matice

$$\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & & \cdots & 0 & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda_{n-r} \end{pmatrix},$$

kde pro prvky platí $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{n-r} \geq 0$ [2].

Zavedeme náhodný vektor $\boldsymbol{Z} = \boldsymbol{Q'U}$, přičemž je jasné, že $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I_{n-r}})$. Statistika

$$d = \frac{\boldsymbol{Z'\Lambda Z}}{\boldsymbol{Z'Z}} = \frac{\sum_{i=1}^{n-r} \lambda_i Z_i^2}{\sum_{i=1}^{n-r} Z_i^2} \tag{1.24}$$

kde čitatel obsahuje lineární kombinaci náhodných veličin s rozdělením χ_1^2 a jmenovatel náhodné veličiny s rozdělením χ_{n-r}^2. Konstanty λ_i závisí na původní regresní matici $\boldsymbol{X}$ [4]. Označíme vlastní čísla matice $\boldsymbol{A}$ jako $\alpha_1 > ... > \alpha_n$. Platí

$$\alpha_i = 2\left(1 - \cos\frac{\pi(n-i)}{n}\right). \tag{1.25}$$

Nechť $\boldsymbol{p_i, ..., p_n}$ jsou příslušné ortonormální vektory. Platí $\alpha_n = 0$ a $\boldsymbol{p_n} = n^{-\frac{1}{2}}\boldsymbol{1}$. Jestli je matice $\boldsymbol{X}_{n\times k}$ zcela obecná a hodnosti p, pak dle [4] platí nerovnosti

$$\lambda_i \leq \alpha_i, \quad i = 1, ..., n-p$$

$$\lambda_i \geq \alpha_{i+p}, \quad i = 1, ..., n-p.$$

Jelikož jsme však předpokládali, že regresní funkce obsahuje absolutní člen, můžeme nerovnosti zapsat pouze do jedné

$$\lambda_i \geq \alpha_{i+p-1}, \quad i = 1, ..., n-p.$$

Z těchto nerovností podle plyne

$$d_L = \frac{\sum_{i=1}^{n-p} \alpha_{i+p-1} Z_i^2}{\sum_{i=1}^{n-p} Z_i^2} \leq d \leq \frac{\sum_{i=1}^{n-p} \alpha_i Z_i^2}{\sum_{i=1}^{n-p} Z_i^2} = d_U. \tag{1.26}$$

Statistiky d_L a d_U mají rozdělení nezávislé na matici $\boldsymbol{X}$ [4].

Momenty rozdělení statistiky d jsou dle [4]

$$\mathrm{E}d = \frac{\sum_{i=1}^{n-p} \lambda_i}{(n-p)} = \bar{\lambda},$$

$$\mathrm{var}d = \frac{2\sum_{i=1}^{n-p}(\lambda_i - \bar{\lambda})^2}{(n-p)(n-p+2)}.$$

Hypotézu nezávislosti zamítneme v případě, že $d \leq d_L(\alpha)$ a nezamítneme v případě, že $d \geq d_U(\alpha)$. Pokud však hodnota statistiky bude přímo v intervalu $(d_L(\alpha), d_U(\alpha))$, je třeba alespoň přibližně odhadnout kvantil $d(\alpha)$ [4].

Pokud tedy zjistíme závislost neboli významnou hodnotu statistiky d. Můžeme tuto závislost odstranit dle následujících úvah v případě, že odchylky e_i nejsou nezávislé, ale platí $e_i = \rho e_{i-1} + \epsilon_i$ s nenulovým ρ [4].

Zavedeme veličiny

$$y_i^* = y_{i+1} - \rho y_i, \;\; x_{ij}^* = x_{i+1,j} - \rho x_{ij}, \qquad i = 1, ..., n-1, \;\; j = 1, ..., k.$$

Platí vztahy

$$y_{i+1} = \sum_{j=1}^{k} x_{i+1,j}\beta_j + (\rho e_i + \epsilon_{i+1}),$$

$$y_i = \sum_{j=1}^{k} x_{i,j}\beta_j + e_i.$$

Lineární kombinací těchto vztahů dostaneme

$$y_i^* = \sum_{j=1}^{k} x_{i,j}^*\beta_j + \epsilon_{i+1}, \;\; i = 1, ..., n-1,$$

kde jsou členy ϵ_i nezávislé [4]. Neznámý parametr ρ můžeme nahradit odhadem

$$\tilde{\rho} = \frac{\sum_{i=1}^{n-1} u_{i+1}u_i}{\sum_{i=1}^{n-1} u_i^2},$$

který je zhruba roven hodnotě $1 - \frac{d}{2}$ [4].

Statisticky významná hodnota d může také znamenat, že jsme zvolili nesprávný tvar regresní funkce [4].

1.7.3 Normalita rozdělení odchylek

Normalitu můžeme testovat například pomocí χ^2 testem dobré shody. K tomuto testu je však potřeba velké množství pozorování [4].

Klasické testy využívají výběrovou šikmost a špičatost jako statistiky pro testování normality [4]:

$$c_1 = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^3 \Big/ \left(\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2\right)^{\frac{3}{2}}, \tag{1.27}$$

$$c_2 = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^4 \Big/ \left(\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2\right)^{2} - 3. \tag{1.28}$$

Mějme uspořádané hodnoty

$$e_{(1)} \leq ... \leq e_{(n)}.$$

Dalším známým testem je Shapiro-Wilkův test. V tomto testu bereme v úvahu statistiku

$$W = \left[\sum_{i=1}^{(n/2)} a_{i,n}(e_{(n-i+1)} - e_{(i)}) \right]^2 / \sum_{i=1}^{n} (e_{(i)} - \bar{e})^2. \tag{1.29}$$

Takto vlastně porovnáváme odhad rozptylu pomocí směrnice přímky z diagramu normality s běžným odhadem, který je založen na součtu čtverců odchylek od průměru [4].

Hodnoty $a_{i,n}$ jsou uvedeny v [5]. Tyto konstanty byly odvozeny ze středních hodnot a varianční matice uspořádaného výběru rozsahu n z rozdělení $N(0,1)$ [4].

Andersonova-Darlingova statistika také vychází z uspořádaného výběru. Tato statistika

$$A^2 = -n^{-1} \sum_{i=1}^{n} (2i-1) \left[\ln(\Phi(\tilde{e}_{(i)})) + \ln(1 - \Phi(\tilde{e}_{(n-i+1)})) \right] - n \tag{1.30}$$

používá normované protějšky místo veličin $e_{(i)}$ a Φ je distribuční funkce normálního rozdělení $N(0,1)$ [4]. Hodnoty $\tilde{e}_{(i)}$ vypočítáme

$$\tilde{e}_{(i)} = \frac{(e_{(i)} - \bar{e})}{\hat{\sigma}_e} \;\; i = 1, ..., n,$$

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_{(i)}, \;\; \hat{\sigma}_e^2 = \frac{1}{n} \sum_{i=1}^{n} (e_{(i)} - \bar{e})^2.$$

1.8 Multikolinearita

Pro tuto část budeme předpokládat, že náš model je model s plnou hodností [4].

Multikolinearita znamená, že sloupce matice $\boldsymbol{X}$ jsou téměř lineárně závislé. Zaznamenáme ji při numerickém řešení normální rovnice, řešení bude nestabilní. Můžeme ji také odhalit pomocí velkých hodnot výběrových korelačních koeficientů mezi regresory. Multikolinearita způsobí malé hodnoty t statistik u jednotlivých odhadů b_j [4].

1.8.1 Odhalování multikolinearity

Níže uvedený postup nám pomůže nalézt multikolinearitu v modelu. V postupu se využívá čísla podmíněnosti, které přiblížíme ještě před postupem nalezení multikolinearity [6].

Věta 1.33. Nechť

$$\boldsymbol{X} = \boldsymbol{P}\boldsymbol{T}\boldsymbol{Q}' \tag{1.31}$$

$$\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{I}_k \tag{1.32}$$

je rozklad matice $\boldsymbol{X}$ podle singulárních hodnot. Víme, že diagonální prvky matice $\boldsymbol{T}$ splňují nerovnosti

$$t_1 \geq ... \geq t_k > 0,$$

což lze předpokládat díky úplné hodnosti matice $\boldsymbol{X}$ [4].

Mějme sloupce matic $\boldsymbol{P} = (\boldsymbol{p}_1, ..., \boldsymbol{p}_k)$ a $\boldsymbol{Q} = (\boldsymbol{q}_1, ..., \boldsymbol{q}_k)$ [4].

Definice 1.34. Indexem podmíněnosti budeme označovat

$$\eta_j = \frac{t_1}{t_j}, \tag{1.33}$$

kde $j = 1, ..., k$ [4]. A předpis

$$\kappa(\boldsymbol{X}) = \eta_k \tag{1.34}$$

budeme označovat za číslo podmíněnosti matice $\boldsymbol{X}$ [4].

V prvním kroku postupu zjišťování multikolinearity je potřeba normovat sloupce regresní matice a to i se sloupcem $\mathbf{1}$. Pro matici $\boldsymbol{X}$, kde platí $\|\boldsymbol{x}_{.j}\|$, $j = 1, ..., k$ se číslo podmíněnosti skoro neliší od nejmenšího možného čísla podmíněnosti, když měníme jednotlivé $\|\boldsymbol{x}_{.j}\|$ [6].

V dalším kroku zjistíme singulární hodnoty $t_1, ..., t_k$ matice $\boldsymbol{X}$ a vypočítáme indexy podmíněnosti $\eta_1, ..., \eta_k$. Zvolíme konstantu η^*, která se volí v rozmezí $10 - 30$. Určíme, které indexy podmíněnosti jsou vyšší než zvolená konstanta η^*. Tím jsme zjistili počet téměř lineárních vztahů mezi sloupci matice $\boldsymbol{X}$. Pokud by některý z indexů podmíněnosti překročil dokonce hodnotu 100, jedná se o silný vztah [6].

Nyní určíme vliv singulárních hodnot t_j (indexů podmíněnosti η_j) na rozptyl statistiky b_s, $s = 1, ..., k$ [6]. Nechť

$$\mathrm{var}b_s = \sigma^2 \sum_{j=1}^{k} \left(\frac{q_{sj}}{t_j}\right)^2.$$

Zavedeme veličiny

$$\pi_{js} = \frac{\left(\frac{q_{sj}}{t_j}\right)^2}{\sum_{l=1}^{k} \left(\frac{q_{sl}}{t_l}\right)^2}, \tag{1.35}$$

kde $j, s = 1, ..., k$. Tyto veličiny nám říkají, jaký je podíl vlivu jednotlivých indexů podmíněnosti na rozptyl b_s. Závislé regresory zjistíme podle hodnot πjs, které odpovídají indexům podmíněnosti větším než η^* dle následujícího postupu [6].

a) Pokud alespoň dvě hodnoty π_{ks}, $s = 1, ..., k$, jsou větší než π^*, pak se u největšího indexu podmíněnosti projevila multikolinearita. Konstanta π^* se volí obvykle $0,5$ [6].

b) V případě, že jsou dva indexy podmíněnosti velmi blízké, pak hodnoty π_{js} v obou řádcích posuzujeme společně. Jsou-li to indexy η_j, η_{j+1}, pak namísto π_{js} budeme pracovat s hodnotami $\pi_{js} + \pi_{j+1}, s$, $s = 1, ..., k$ [6].

c) Je-li některý index podstatně větší než η_j, pak některé hodnoty π_{js}, $s = 1, ..., k$ mohou být potlačeny dominujícím řádkem většího indexu podmíněnosti [6].

V posledním kroku odhalování mulitkolinearity máme tedy určené regresory $\boldsymbol{x}_{.s}$, pro které je $\eta_j > \eta^*$ a $\pi_{js} > \pi^*$. Pro každý index podmíněnosti (každé j) vybereme jediný regresor a pokusíme se jej vyjádřit pomocí ostatních nevybraných. Při tom můžeme použít pomocný lineární model, ve kterém do matice $\boldsymbol{X}$ zahrneme regresory nevybrané v žádném řádku s $\eta_j > \eta^*$. Závisle proměnná bude pro každé $\eta_j > \eta^*$ vybraný regresor. Tímto postupem dokážeme určit, jak závisí vynechané regresory na ostatních [6].

1.8.2 Vychýlené odhady

Abychom se vyhnuli velkým rozptylům statistik b_j, které jsou způsobeny multikolinearitou, můžeme využít vychýlených odhadů. Nejčastěji se používají metody hřebenové regrese a regrese na hlavních komponentách matice $\boldsymbol{X}$ [4]. Důkazy tvrzení nalezneme v [4].

Hřebenová regrese využívá hřebenového odhadu vektoru $\boldsymbol{\beta}$ tvaru

$$\boldsymbol{b}_\delta = (\boldsymbol{X}'\boldsymbol{X} + \delta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{1.36}$$

kde $\delta \geq 0$ [4]. Tento odhad můžeme také vyjádřit pomocí vektoru $\boldsymbol{b}$

$$\begin{aligned} \boldsymbol{b}_\delta &= (\boldsymbol{X}'\boldsymbol{X} + \delta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X} + \delta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} \\ &= \left[(\boldsymbol{I} + \delta(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]^{-1}\boldsymbol{b}. \end{aligned} \tag{1.37}$$

Odhad $\boldsymbol{b}_\delta$ má menší střední čtvercovou chybu než $\boldsymbol{b}$ pro dostatečně malé nenulové δ [4].

Metoda regrese na hlavních komponentách matice $\boldsymbol{X}$ používá podle [4] vychýlený odhad

$$\boldsymbol{b}_r = (\boldsymbol{X}_r'\boldsymbol{X}_r + \boldsymbol{K}'\boldsymbol{K})^{-1}\boldsymbol{X}_r'\boldsymbol{y} = (\boldsymbol{Q}_r\boldsymbol{T}_r^2\boldsymbol{Q}_r' + \boldsymbol{Q}_{k-r}\boldsymbol{Q}_{k-r}')^{-1}\boldsymbol{Q}_r\boldsymbol{T}_r\boldsymbol{P}_r'\boldsymbol{y}$$

$$= \left[(\boldsymbol{Q}_r, \boldsymbol{Q}_{k-r})\begin{pmatrix} \boldsymbol{T}_r^2 & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_{k-r} \end{pmatrix}\begin{pmatrix} \boldsymbol{Q}_r' \\ \boldsymbol{Q}_{k-r}' \end{pmatrix}\right]^{-1}\boldsymbol{Q}_r\boldsymbol{T}_r\boldsymbol{P}_r'\boldsymbol{y} \tag{1.38}$$

$$= \boldsymbol{Q} \begin{pmatrix} \boldsymbol{T}_r^{-2} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_{k-r} \end{pmatrix} \boldsymbol{Q}'\boldsymbol{Q}_r\boldsymbol{T}_r\boldsymbol{P}'_r\boldsymbol{y} = (\boldsymbol{Q}_r\boldsymbol{T}_r^{-2}\boldsymbol{Q}'_r + \boldsymbol{Q}_{k-r}\boldsymbol{Q}'_{k-r})\boldsymbol{Q}_r\boldsymbol{T}_r\boldsymbol{P}'_r\boldsymbol{y}$$

$$= \boldsymbol{Q}_r\boldsymbol{T}_r^{-1}\boldsymbol{P}'_r\boldsymbol{y} = \sum_{j=1}^{k} t_j^{-1}(\boldsymbol{p}'_j\boldsymbol{y})\boldsymbol{q}_j,$$

kde $h(\boldsymbol{K}) = k - r$ a $h(\boldsymbol{X}) = r$

$\boldsymbol{K}' = (\boldsymbol{q}_{r+1}, ..., \boldsymbol{q}_k)$,
$\boldsymbol{X}_r = \sum_{i=1}^{k} \boldsymbol{t}_j\boldsymbol{p}_j\boldsymbol{q}'_j$
$\boldsymbol{P} = (\boldsymbol{P}_r, \boldsymbol{P}_{k-r})$, $\boldsymbol{Q} = (\boldsymbol{Q}_r, \boldsymbol{Q}_{k-r})$,
$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{T}_{k-r} \end{pmatrix}$.

1.9 Volba vícerozměrného regresního modelu

Tato část se zabývá postupy, které nám pomohou zvolit vhodný regresní model, který bude dostatečně vystihovat náš problém.

1.9.1 Porovnávání modelů

Cílem regrese je popsat variabilitu závisle proměnné co možná nejlépe, avšak s použitím co možná nejmenšího počtu regresorů a tak, aby interpretace byla co nejjednodušší. Pro porovnávání modelů se používají různá kritéria, ty nejpoužívanější popíšeme níže [4].

Reziduální součet čtverců

Vypočítejme RSS v modelu $\boldsymbol{y} \sim (\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ a RSS_g v modelu $\boldsymbol{y} \sim (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}, \sigma^2\boldsymbol{I})$, ve kterém platí $h(\boldsymbol{X}, \boldsymbol{Z}) = p + m$, kde m je počet sloupců matice $\boldsymbol{Z}$ [4]. Dále předpokládáme, že

$$RSS - RSS_g = \|\boldsymbol{d}\|^2 = \boldsymbol{y}'\boldsymbol{M}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{y}.$$

Jelikož je $h(\boldsymbol{X}, \boldsymbol{Z}) > h(\boldsymbol{X})$, je $\boldsymbol{M}\boldsymbol{Z} \neq 0$ a rozdíl $RSS - RSS_g$ je tedy skoro jistě kladný. Z toho vyplývá, že reziduální součet čtverců je nejmenší pro model obsahující všechny dostupné regresory. S každým ubráním regresoru se hodnota RSS skoro jistě zvýší [4].

Ta část variability závisle proměnné, která není modelem vysvětlena, je vyjádřena reziduálním součtem čtverců. Hodnota RSS je závislá na měřítku použitém pro závisle proměnnou. Matici $\boldsymbol{M}$ můžeme zapsat pomocí ortonormální báze jako $\boldsymbol{M} = \boldsymbol{N}\boldsymbol{N}'$. Blíže je tato matice popsána v kapitole 1.7.2 [4].

Koeficient determinace R^2

Tohle kritérium lze použít pouze v modelech, kde regresní funkce obsahuje absolutní člen. Koeficient determinace můžeme vyjádřit

$$R^2 = 1 - \frac{RSS}{\|\boldsymbol{y} - \bar{y}\mathbf{1}\|^2} = 1 - \frac{RSS}{TSS}, \tag{1.39}$$

z čehož můžeme odvodit, že koeficient determinace je klesající funkcí reziduálního součtu čtverců. Koeficient determinace je však bezrozměrnou veličinou nabývající hodnoty z intervalu a $\langle 0, 1\rangle$ a vyjadřuje míru variability závisle proměnné vyjádřené modelem [4].

Reziduální rozptyl

V modelu $\boldsymbol{y} \sim N(\boldsymbol{X\beta} + \boldsymbol{Z\gamma}, \sigma^2\boldsymbol{I})$ platí

$$\mathrm{E}\, s^2 = \sigma^2 + \frac{\|\boldsymbol{MZ\gamma}\|^2}{n-p},$$

z čehož můžeme soudit, že pokud není model $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ správný, statistika s^2 dá odhad rozptylu σ^2 vychýlený směrem nahoru. Pokud však platí alespoň nějaký podmodel modelu $\boldsymbol{y} \sim (\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, pak je statistika s^2 nestranným odhadem rozptylu [4].

Díky těmto vlastnostem můžeme zavést odhad řádu modelu. Odhad řádu modelu chápeme jako odhad nejmenšího počtu regresorů, které dají lineární kombinací vektor $\mathrm{E}\boldsymbol{y}$ [4].

Mějme regresní matici s hodností j. Nejmenším z reziduálních rozptylů modelu je s_j^2. Chceme nalézt index, od kterého bude posloupnost statistik s_j^2, $j = 1, ...,$ kolísat kolem hodnoty σ^2. Jelikož mají statistiky s_j^2 náhodný charakter, je takovéto určení řádu velice složité [4].

K určení řádu modelu se vyjádříme ještě později.

Nerovnosti

$$F = \frac{\|d\|^2}{RSS_g}\frac{n-p-m}{m} \leq \frac{1}{m}$$

$$F \leq 1,$$

kde statistika F má rozdělení $F_{m,n-p-m}$, naznačují, kdy je vhodnější použít podmodel $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ oproti správnému modelu $\boldsymbol{y} \sim N(\boldsymbol{X\beta} + \boldsymbol{Z\gamma}, \sigma^2\boldsymbol{I})$ vzhledem ke zvolenému kritériu [4].

Statistiku F můžeme zapsat

$$F = \frac{RSS - RSS_g}{RSS_g}\frac{n-p-m}{m} = \frac{(n-p)s^2 - (n-p-p)s_g^2}{ms_g^2} = \frac{n-p}{ms_g^2}(s^2 - s_g^2) + 1,$$

kde $s_g = \frac{RSS_g}{n-p-m}$. Proto jsou nerovnosti $F \leq 1$ a $s^2 \leq s_g^2$ ekvivalentní, pokud se jedná o úlohu interpolace. V případě extrapolace je omezení shora v nerovnostech ekvivalentní s $F \leq \frac{1}{m}$ [4]. Platí tedy $s^2 \leq s_g^2$ a

$$s_g^2 \frac{n-p-m}{n-p} \leq s^2 \leq s_g^2 \frac{n-p-m+1}{n-p}.$$

Upravený koeficient determinace

V případě normálního lineárního modelu můžeme vyjádřit koeficient determinace R^2 takto:

$$R^2 = 1 - \frac{\hat{\sigma}_{Y.X}^2}{\hat{\sigma}_Y^2} \tag{1.40}$$

kde $\hat{\sigma}_{Y.X}^2$ je odhad rozptylu σ^2 metodou maximální věrohodnosti v modelu $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ a $\hat{\sigma}_Y^2 = \frac{\|\boldsymbol{y}-\bar{y}\boldsymbol{1}\|^2}{n}$ je podobný odhad rozptylu v modelu $\boldsymbol{y} \sim N(\gamma\boldsymbol{1}, \sigma^2\boldsymbol{I})$. Pokud místo těchto odhadů použijeme jejich nestranné protějšky, získáme upravený koeficient determinace [4]

$$\bar{R} = 1 - \frac{RSS}{\|\boldsymbol{y}-\bar{y}\boldsymbol{1}\|^2}\frac{n-1}{n-p} = 1 - \frac{RSS}{TSS}\frac{n-1}{n-p}. \tag{1.41}$$

Upravený koeficient můžeme vyjádřit s použitím reziduálního rozptylu takto:

$$\bar{R}^2 = 1 - \frac{n-1}{\|\boldsymbol{y}-\bar{y}\boldsymbol{1}\|^2}s^2 = 1 - \frac{n-1}{TSS}s^2, \tag{1.42}$$

z čehož je vidět, že jde o klesající funkci reziduálního rozptylu s^2 [4].

Mallowsovo C_p

Mějme celkovou čtvercovou chybu statistik $\hat{y}_i$, které jsou brány jako odhady středních hodnot $\mathrm{E}y_i$ pro $i = 1, ..., n$ v modelu $\boldsymbol{y} \sim (\boldsymbol{X\beta} + \boldsymbol{Z\gamma}, \sigma^2\boldsymbol{I})$. Tuto chybu můžeme vyjádřit

$$p\sigma^2 + \|\boldsymbol{MZ\gamma}\|^2 = \mathrm{E}RSS + (2p-n)\sigma^2,$$

kde $\mathrm{E}RSS = (n-p)\sigma^2 + \|\boldsymbol{MZ\gamma}\|^2$ [4].

Chybu můžeme vztáhnout k σ^2

$$\Gamma_p = \frac{\mathrm{E}RSS}{\sigma^2} + 2p - n,$$

dále

$$C_p = \frac{RSS}{s_g^2} + 2p - n$$

můžeme brát jako odhad Γ_p [4].

Pokud je podmodel $\boldsymbol{y} \sim (\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ modelu $\boldsymbol{y} \sim (\boldsymbol{X\beta} + \boldsymbol{Z\gamma}, \sigma^2\boldsymbol{I})$ platný, pak je $\mathrm{E}RSS = (n-p)\sigma^2$ a $\Gamma_p = p$[4].

Jsme schopni dle [4] vyjádřit také statistiku F pomocí C_p takto

$$m(F-1) = \frac{(n-p-m)RSS - (n-p-m)RSS_g - mRSS_g}{RSS_g}$$
$$= \frac{RSS}{s_g^2} - (n-p) = C_p - p.$$

Nerovnost $F \leq 1$ je ekvivalentní s nerovností $C_p \leq p$ a nerovnost $F \leq \frac{1}{m}$ platí, když $C_p \leq p - m + 1$ [4].

Průměrný rozptyl předpovědi

Mějme nová pozorování $Y(\boldsymbol{x}_{i.})$ odpovídající vektorům regresorů $\boldsymbol{x}_{i.}$, $i = 1, ..., n$. Tato nová pozorování budeme předpovídat pomocí statistik $\hat{y}_i$, které mají rozptyl $\sigma^2 h_{ii}$. Každá z náhodných veličin $Y(\boldsymbol{x}_i)$ je nezávislá na $\boldsymbol{y}$ z modelu $\boldsymbol{y} \sim (\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. Odtud plyne, že rozptyl chyby předpovědi $Y(\boldsymbol{x}_{i.}) - \hat{y}_i$ je

$$\sigma^2(1 + h_{ii}),$$

což značí variabilitu $\hat{y}_i$ jako předpovědi pro $Y(\boldsymbol{x}_i)$. Průměrná hodnota rozptylu chyby předpovědi pro $i = 1, ..., n$ je

$$\sigma^2 \sum_{i=1}^{n} (1 + h_{ii})/n = \sigma^2(1 + p/n).$$

Rozptyl σ^2 však neznáme. Mějme tedy statistiku

$$J_p = s^2(1 + p/n),$$

kde je neznámý rozptyl nahrazen jeho nestranným odhadem s^2. Odhad řádu modelu založený na minimalizaci J_p přes p není konzistentní[4].

Konzistentní odhady řádu modelu získáme pomocí těchto statistik

$$A(p) = s^2(1 + cpn^{-\alpha}),\ \alpha \in \left(0, \frac{1}{2}\right),\ c > 0, \tag{1.43}$$

$$HQ(p) = \ln s^2 + 2\, cp\, (\ln \ln n)/n,\ c > 0, \tag{1.44}$$

$$SR(p) = \ln s^2 + p \ln n/n, \tag{1.45}$$

které minimalizujeme [4].

1.9.2 Výběr podmnožin regresorů

Mějme mimo hodnoty závisle proměnné také hodnoty K potencionálních regresorů. My z nich chceme vybrat k, aby co nejlépe vystihovaly variabilitu závisle proměnné. Předem však neznáme hodnotu k. K porovnávání různých podmnožin využijeme kritéria popsaná dříve. Pokud je hodnota K relativně malá, můžeme kritéria spočítat pro všechny možné modely, kterých je 2^K. Častěji se však využívají různé metody. Ty jednodušší a často využívané si popíšeme níže [4].

Vzestupný výběr regresorů

Začínáme s prázdnou množinou regresorů popřípadě s konstantním regresorem. V každém kroku do modelu vložíme regresor, který sníží hodnotu reziduálního součtu čtverců nejvíce. Pro každý doposud nezařazený regresor $\boldsymbol{z}_{.j}$ potřebujeme určit hodnotu

$$RSS_{gj} = RSS - \|\boldsymbol{d}_j\|^2,$$

kde je

$$\|\boldsymbol{d}_j\|^2 = \boldsymbol{u}'\boldsymbol{z}_{.j}(\boldsymbol{z}'_{.j}\boldsymbol{M}\boldsymbol{z}_{.j}) - \boldsymbol{z}'_{.j}\boldsymbol{u} = \boldsymbol{y}'\boldsymbol{M}\boldsymbol{z}_{.j}(\boldsymbol{z}'_{.j}\boldsymbol{M}\boldsymbol{z}_{.j}) - \boldsymbol{z}'_{.j}\boldsymbol{M}\boldsymbol{y}.$$

Tyto hodnoty potřebujeme určit, abychom mohli zjistit, který regresor bude snižovat hodnotu reziduálního součtu čtverců nejvíce [4].

Víme, že platí $\boldsymbol{M}\boldsymbol{z}_{.j} \neq \boldsymbol{0}$, $j = 1, ..., m$, takže můžeme $\|\boldsymbol{d}_j\|^2$ vyjádřit také jako

$$\|\boldsymbol{d}_j\|^2 = \frac{(\boldsymbol{y}'\boldsymbol{M}\boldsymbol{z}_{.j})^2}{(\boldsymbol{z}'_{.j}\boldsymbol{M}\boldsymbol{z}_{.j})} = \frac{\|\boldsymbol{y} - \bar{y}\boldsymbol{1}\|^2 (\boldsymbol{y}'\boldsymbol{M}\boldsymbol{z}_{.j})^2}{(\boldsymbol{z}'_{.j}\boldsymbol{M}\boldsymbol{z}_{.j}) \|\boldsymbol{y} - \bar{y}\boldsymbol{1}\|^2}.$$

Pokud model $\boldsymbol{y} \sim (\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ obsahuje absolutní člen, tzn. platí $\boldsymbol{M}\boldsymbol{1} = \boldsymbol{0}$, pak platí, že $\boldsymbol{1}'\boldsymbol{M}\boldsymbol{z}_{.j} = 0$. Označíme si výběrový koeficient parciální korelace vektorů $\boldsymbol{y}$ jako $\boldsymbol{r}_{y,z_j.x}$, pak můžeme psát

$$RSS_{gj} = RSS - \|\boldsymbol{y} - \bar{y}\boldsymbol{1}\|^2 \boldsymbol{r}_{y,z_j.x}.$$

Odtud je vidět, že výběr zařazení jednotlivých regresorů podle reziduálních součtů je ekvivalentní s rozhodováním podle výběrového koeficientu parciální korelace [4].

Nejčastějším kritériem pro výběr regresoru pro zařazení do modelu je F test. Mějme normální model $\boldsymbol{y} \sim (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\gamma}\boldsymbol{z}_{.j}, \sigma^2\boldsymbol{I})$. Předpokládáme stále $h(\boldsymbol{X}, \boldsymbol{Z}) = p+m$. Pokud platí hypotéza $\gamma = 0$, pak má statistika

$$F_j = \frac{\|\boldsymbol{d}_j\|^2}{RSS_{gj}} \frac{n-p-1}{1} = \frac{\|\boldsymbol{d}_j\|^2}{RSS - \|\boldsymbol{d}_j\|^2} \frac{n-p-1}{1}$$

rozdělení $F_{1,n-p-1}$. Takové rozhodování je ekvivalentní s rozhodováním podle rezidálního součtu čtverců nebo podle koeficientu parciální korelace. Do modelu zahrneme takový nový regresor, jehož hodnota $F_j$je maximální. V okamžiku získání hodnoty maximálního

F_j, která bude nižší než hodnota F^*, proces přidávání regresorů ukončíme. Za F^* volíme obvykle konstanty, $F^* = 2$ nebo $F^* = 4$ nebo můžeme hodnotu F^* vypočítat pomocí hodnoty kvantilové funkce $F_{1,n-p-1}(1-\alpha^*)$ [4].

Sestupný výběr regresorů

Metoda sestupného výběru vychází naopak z modelu obsahující všechny možné regresory, které postupně vylučujeme. V každém kroku vyloučíme takový regresor, který napomáhá k vysvětlení závisle proměnné nejméně. Nejčastěji vyloučíme regresor na základě F testu s hypotézou, že daný regresor je nulový. Regresor, jehož hodnota F testu je nejmenší, z modelu vyloučíme. V okamžiku, kdy je hodnota některého z nejmenších hodnot statistiky F větší než konstanta F^{**}, proces končí. Za F^{**} volíme obvykle konstanty, $F^{**} = 2$ nebo $F^{**} = 4$ nebo můžeme hodnotu F^* vypočítat pomocí hodnoty kvantilové funkce $F_{1,n-p}(1-\alpha^{**})$ [4].

Kroková regrese

Metoda krokové regrese kombinuje obě předchozí metody. Začínáme obvykle s konstantním regresorem. V každém kroku provedeme nejprve sestupný výběr a vyloučíme regresory, které lze na základě tohoto výběru vyloučit. Poté použijeme vzestupný výběr a pokusíme se zahrnout jeden regresor, který lze přidat. Dále opět postupujeme sestupnou metodou. Pokud však nelze přidat ani vyloučit nějaký regresor, proces ukončíme [4].

Nevýhodou všech tří metod popsaných výše je, že nám dají jediný model. Může ale existovat řešení se stejnými hodnotami statistik, avšak s jednodušší interpretací. Z tohoto důvodu se využívá upravená kroková regrese, kdy do modelu zahrneme přednostně některé regresory [4].

1.9.3 Transformace závisle proměnné

Někdy potřebujeme dosáhnout lineární závislosti, stabilního rozptylu nebo přibližně normálního rozdělení, k čemuž nám může pomoci vhodná transformace závisle proměnné nebo regresorů [4].

Velké množství transformací kladné závisle proměnné y můžeme zapsat ve tvaru

$$y^{(\lambda)} = \frac{y^{(\lambda)} - 1}{\lambda},\ \lambda \neq 0, \tag{1.46}$$

$$y^{(0)} = \ln y,$$

které upravují Tukeyho transformace [7]. Tento tvar je komplikovaný, avšak výhodou je, že funkce $y^{(\lambda)}$ je spojitá i v bodě 0 [4]:

$$\lim_{\lambda \to 0} y^{(\lambda)} = \ln y.$$

Nechť pro nějaké λ přibližně platí model

$$\boldsymbol{y}^{(\lambda)} \sim N(\boldsymbol{X\beta}, \boldsymbol{\sigma^2 I})$$

kde $\boldsymbol{y}^{(\lambda)}$ je vektor se složkami $y_i^{(\lambda)}$. Tento model platí přesně pouze pro $\lambda = 0$ [4].

Hodnota λ je neznámým parametrem, který odhadujeme stejně jako $\boldsymbol{\beta}$ a σ^2 pomocí metody maximální věrohodnosti [4].

Náhodný vektor $\boldsymbol{y}^{\lambda}$ má sdruženou hustotu

$$(2\pi\sigma^2)^{-n/2} exp\left[\frac{-1}{(2\sigma^2)}\sum_{i=1}^{n}(y_i^{(\lambda)} - \boldsymbol{x}'_{i.}\boldsymbol{\beta})^2\right].$$

Dále má náhodný vektor $\boldsymbol{y}$ hustotu

$$f(\boldsymbol{y}) = (2\pi\sigma^2)^{-n/2} exp\left(\frac{-1}{(2\sigma^2)}\sum_{i=1}^{n}(y_i^{(\lambda)} - \boldsymbol{x}'_{i.}\boldsymbol{\beta})^2\right)\boldsymbol{J},$$

kde $\boldsymbol{J}$ je absolutní hodnota jakobiánu transformace $\boldsymbol{y} \mapsto \boldsymbol{y}^{(\lambda)}$. Pro $\lambda \neq 0$ platí

$$J = \prod_{i=1}^{n}\frac{\partial y_i^{\lambda-1}}{\lambda} = \prod_{i=1}^{n} y_i^{\lambda-1} = \left(\prod_{i=1}^{n} y_i\right)^{\lambda-1} = \dot{y}^{n(\lambda-1},$$

kde $\dot{y}$ je geometrický průměr z $y_1, ..., y_n$. Jakobián vyjde stejně i pro $\lambda = 0$ [4].

Dále nalezneme odhady metodou maximální věrohodnosti. Mějme hustotu $f(\boldsymbol{y})$ jako funkci neznámých parametrů $\boldsymbol{\beta}$, σ^2, λ. Označíme

$$l(\boldsymbol{\beta}, \sigma^2, \lambda) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^{(\lambda)} - \boldsymbol{x}'_{i.}\boldsymbol{\beta})^2 + n(\lambda - 1)\ln\dot{y}$$

logaritmem hustoty $f(\boldsymbol{y})$. Pokud je výraz

$$\sum_{i=1}^{n}(y_i^{(\lambda)} - \boldsymbol{x}'_{i.}\boldsymbol{\beta})^2$$

minimální a hodnoty σ^2 a λ jsou pevné, funkce $l(\boldsymbol{\beta}, \sigma^2, \lambda)$ je maximální. To znamená, že regresní koeficienty budeme odhadovat metodou nejmenších čtverců. Odhad koeficientů $\boldsymbol{\beta}$ označíme $\boldsymbol{b}^{(\lambda)}$. Nejdříve však potřebujeme odhadnout λ [4].

Odhad $\hat{\lambda}$ parametru λ získáme minimalizací funkce

$$RSS_z(\lambda) = \dot{y}^{2(1-\lambda)} RSS_y(\lambda),$$

kde RSS_z se vztahuje k modelu $^{(\lambda)} \sim N(\boldsymbol{X\beta_z}, \boldsymbol{\sigma^2 I})$ a $\boldsymbol{z}^{(\lambda)}$. Konfidenční interval pro λ je takový, kde λ splňují

$$RSS_z(\lambda) \leq RSS_z(\hat{\lambda}) exp\left[\chi_1^2(1-\alpha)/n\right]$$

Podrobné odvození je uvedeno v [4].

Když hledáme vhodný tvar závislosti, můžeme použít i neformálnější postup. Předpokládejme, že máme k dispozici jedinou nezávisle proměnnou s kladnými hodnotami, k dané závisle proměnné s kladnými hodnotami [4]. Vybíráme z transformací typu

$$..., \frac{-1}{x^2}, \frac{-1}{x}, \ln x, \frac{-1}{\sqrt{n}}, \sqrt{x}, x, x^2....$$

Této řadě transformací se říká transformační žebřík nebo žebřík transformací [2]. Snažíme se závislost linearizovat tak, že postupujeme po těchto transformacích vzhůru nebo dolů v jedné nebo druhé proměnné. Jakmile se nám to podaří, snažíme se dále stabilizovat rozptyl současným postupem v obou proměnných. Jaký směr volit při výběru transformace zachycuje následující tabulka (Tabulka 2.) [4].

Tabulka 2: Výběr transformace. [vlastní zpracování]

Tvar závislosti	**Výběr transformace**
rostoucí, konvexní	postupujeme s x nahoru (k vyšším mocninám), nebo s y dolů (k nižším mocninám)
klesající, konvexní	postupujeme s x dolů (k nižším mocninám), nebo s y dolů (k nižším mocninám)
rostoucí, konkávní	postupujeme s x dolů (k nižším mocninám), nebo s y nahoru (k vyšším mocninám)
klesající, konkávní	postupujeme s x nahoru (k vyšším mocninám), nebo s y nahoru (k vyšším mocninám)

Tvar závislosti je odhadován na základě praxe a dat.

1.10 Nelineární regresní modely

Často se nám může stát, že nejsme schopni popsat vysvětlovanou proměnnou tak, aby na parametrech závisela lineárně. V takovém případě mluvíme o nelineárním regresním modelu [2].

Dále si představíme několik známých nelineárních modelů. Tyto modely se využívají především při analyzování časových řad.

1.10.1 Exponenciální regresní funkce

Tento model je velice často využívaný pro svou jednoduchost [8].

$$f(x; \alpha, \beta) = \alpha\beta^x \tag{1.47}$$

1.10.2 Modifikovaný exponenciální trend

$$f(x;\alpha,\beta,\gamma) = \alpha + \beta\gamma^x \tag{1.48}$$

Tento model je využívaný v situacích, kdy podíly sousedních hodnot prvních diferencí analyzované řady jsou přibližně konstantní [8].

1.10.3 Logistický trend

Tento trend spadá do skupiny S-křivek, které jsou symetrické okolo inflexního bodu [9].

$$f(x;\alpha,\beta,\gamma) = \frac{1}{\alpha + \beta\gamma^x} \tag{1.49}$$

1.10.4 Gompertzova křivka

Gompertzova křivka patří mezi S-křivky, které nejsou symetrické kolem inflexního bodu [8].

$$f(x;\alpha,\beta,\gamma) = e^{\alpha+\beta\gamma^x} \tag{1.50}$$

1.10.5 Kompartmentový model

$$f(x;\alpha,\beta,\gamma) = \gamma\frac{\beta}{\alpha-\beta}(e^{\beta x} - e^{\alpha x}), \tag{1.51}$$

kde $x \geq 0$ a α, β, γ jsou neznámé kladné parametry a $\alpha \neq \beta$ [2].

1.10.6 Michaelisův - Mentenův model

$$f(x;\theta_1,\theta_2) = \frac{\theta_1 x}{\theta_2 + x}, \tag{1.52}$$

kde $x \geq 0$ [2].

Pokud použijeme

$$\frac{1}{y} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1}\frac{1}{x},$$

dostaneme Michaelisův - Mentenův model v linearizovaném tvaru [2].

Množství nelineárních regresní funkcí můžeme vhodnou transformací převést na funkci lineární v parametrech. Takové modely nazýváme linearizovatelné [4]. Mezi linearizovatelné funkce můžeme zařadit například exponenciální regresní funkci. Tento trend můžeme linearizovat pomocí logaritmů převrácených hodnot [8].

2 POPIS PROBLÉMU

Tato část se zabývá popisem problému, který budeme zkoumat dále pomocí regresní analýzy.

Teoretické poznatky budeme aplikovat na problém vlivu a optimalizace parametrů obrábění na šířku řezné spáry z titanové slitiny Ti-6Al-4V. Abychom získali představu o tomto problému, rozebereme si materiál a metody obrábění, které tento problém zkoumá. Popíšeme si titan, jeho vlastnosti a slitiny titanu. Dále se zaměříme na námi zkoumanou slitinu titanu Ti-6Al-4V. Rozebereme stručně elektroerozivní metody obrábění a zaměříme se na technologii WEDM.

2.1 Titan

Titan je využíván od roku 1948, ale objeven byl již koncem 18. století. Tento lehký kov se vyskytuje ve dvou krystalových modifikacích, tzv. hexagonální α-fáze a kubická β-fáze [10].

Obrázek 1: Titan.[11]

Titan a jeho slitiny využíváme především proto, že mají největší měrnou pevnost ze všech kovových materiálů. Další výhodnou vlastností titanových slitin je korozní odolnost. Mezi další specifické vlastnosti titanových slitin patří biokompatibilita a odolnost vůči působení zvýšených teplot [12]. Z těchto důvodů se tyto materiály využívají často v letectví a kosmickém průmyslu. Titanové slitiny nahradili slitiny hliníku u nadzvukových letounů [13]. Díky chemické odolnosti se titan využívá i ve zdravotnictví a chemickém průmyslu [10].

Velkým nedostatkem titanu je obtížná a nákladná výroba a zpracování. Titan reaguje s plyny při teplotách nad 600°C, což znamená, že se znehodnocuje. Proto tavení, odlévání

i svařování musí být prováděno v ochranné argonové atmosféře nebo ve vakuu [10].

V následující tabulce (Tabulka 3.) můžeme vidět přehled základních fyzikálních vlastností titanu.

Tabulka 3: Základní fyzikální vlastnosti titanu. [14]

Značka	Ti
Atomové číslo	22
Skupina	$IV.B$
Elektronová konfigurace	$[Ar]3d^2 4s^2$
Oxidační čísla ve sloučeninách	$+3\ +4$
Elektronegativita	$1,54$
Relativní atomová hmotnost	$47,867$
Teplota tání	$1668°C$
Teplota varu	$3287°C$
Tepelná vodivost	$11,4\ W/m.K$
Koeficient tepelné roztažnosti	$8,41\ \mu\ m/m.K$
Měrný elektrický odpor	$420\ n\ \Omega$.m
Hustota	$4500\ kg/m^3$

2.2 Slitiny titanu

Existuje více než 100 různých slitin titanu. Komerčně se však využívá pouze cca 25% z nich. Celková produkce je tvořena z cca 25% čistého titanu a cca z 50% slitiny Ti-6Al-4V. Podle fázového složení dělíme titanové slitiny na α slitiny, $\alpha + \beta$ slitiny a β slitiny [10].

V první skupině, tedy v α slitinách se objevuje především hliník. Nejčastější slitinou této skupiny je Ti-5Ald-2,5Sn [10]. Tyto slitiny se využívají například pro lopatky leteckých motorů nebo na nádrže na tekutý dusík [12].

Do druhé skupiny slitin $\alpha + \beta$ patří právě zkoumaná slitina Ti-6Al-4V [10]. Tyto slitiny se využívají na součásti leteckých motorů nebo na sportovní nářadí, do tenisových raket nebo rámů kol [12].

Slitiny β jsou nejdražší a mají nejvyšší pevnost ze všech titanových slitin. Příkladem takové slitiny je Ti-10V-2Fe-3Al [10]. Tyto slitiny můžeme najít v kosmickém průmyslu, energetice, těžebním a automobilovém průmyslu. Často jsou však využívány i pro ortopedické implantáty [12].

2.2.1 Slitina Ti-6Al-4V

Slitina Ti-6Al-4V se označuje také jako Grade 5 nebo Ti 6-4. Jde o nejpoužívanější titanovou slitinu. Využívá se hlavně proto, že má výbornou pevnost, je korozně odolná, tvářitelná, obrobitelná a svařitelná. Je možné ji také tepelně zpracovávat. Je stabilní až do 400°C [12].

Tuto slitinu vyvinuli vědci v padesátých letech pro výrobu lopatek plynových turbín. Dnes ji najdeme nejčastěji v různých částech letadel a ve vybavení sonarů a ponorek. Také závodní vozy F1 obsahují tuto slitinu a to již od osmdesátých let [15].

Podle konkrétní aplikace se pak mění přesné obsahy jednotlivých prvků ve slitině. V tabulce níže (Tabulka 4.) jsou však uvedeny mezní obsahy prvků [16].

Tabulka 4: Mezní obsahy prvků ve sloučenině Ti-6Al-4V. [16]

Prvek	**Obsah prvku [hm.%]**
Hliník	5,75 - 6,75
Vanad	3,5 - 4,5
Železo	max 0,25
Kyslík	max 0,20
Dusík	max 0,05
Vodík	max 0,015
Uhlík	max 0,08

S rostoucím obsahem dusíku a kyslíku se zvyšuje pevnost slitiny, naopak když snížíme obsah dusíku, kyslíku a hliníku, dojde ke zlepšení tažnosti a odolnosti vůči korozi pod napětím [16].

2.3 Elektroerozivní metody obrábění

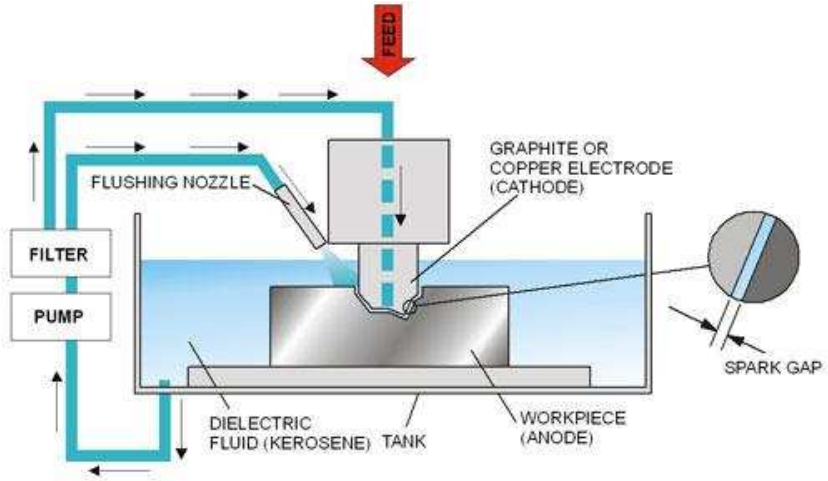
Elektroerozivní metody obrábění, ve zkratce EDM (Electrical Discharge Machining), jsou technologie obrábění, které využívají tepelnou energii vodivých materiálů, která vzniká při elektrickém výboji mezi elektrodami. Elektrody jsou na nástroji a obráběném materiálu [17].

Tyto metody můžeme podle technologických možností dělit na:

- hloubení nebo tvarové elektroerozivní obrábění (EDM Sikning),
- drátové řezání (WEDM - Wire Electrical Discharge Machining),
- broušení (EDG - Electrical Discharge Grinding) [17].

2.3.1 Princip elektroerozivního obrábění

Obrábění touto metodou probíhá mezi dvěma elektrodami. Ty jsou ponořeny do dielektrické kapaliny, to je kapalina elektricky nevodivá, a odděleny od sebe jiskrovou mezerou (0,01 - 0,5 mm). V místě nejsilnějšího elektrického napěťového pole vznikne výboj mezi elekrodami. Ty umožňují přechod jiskry mezi nástrojem a obroubkem. V kanále se tvoří plasmové pásmo, které prudce ohřívá, taví a částěčně i odpařuje materiál na povrchu obroubku. Tlak kovových par a průnik dielektrika do uzavřeného prostoru vymrští roztavený kov [18]. Tento proces je znázorněn na obrázku níže (Obrázek 2.).



Obrázek 2: Proces EDM.[19]

2.4 Elektrojiskrové řezání drátovou elektrodou

Elektrojiskrové řezání drátovou elektrodou neboli WEDM (Wire electrical discharge machining je dnes nejrozšířenější metodou elektroerozivního obrábění. Tato metoda využívá tepelný proces a umožňuje přesné obrábění tvarově složitých dílců s různou tvrdostí a s potřebou ostrých hran [20].

Drátová elektroda je z tenkého drátu, který je odvíjen z jedné cívky na druhou. Tento drát prochází přes vodící ústrojí. Tento drátek neboli nástrojová elektroda je zapojena jako katoda a obrobek se zapojuje jako anoda. Spojení je tedy přímá polarita [20].

Mezi nástrojovou elektrodou a obrobkem vznikne pracovní mezera a řez požadovaného tvaru a to vlivem elektroeroze. Drátová elektroda musí být řádně napnutá a vyrovnaná, aby byl řez co nejpřesnější. Proto se drátové elektrody před řezem kalibrují v diamantových průvlacích [20].

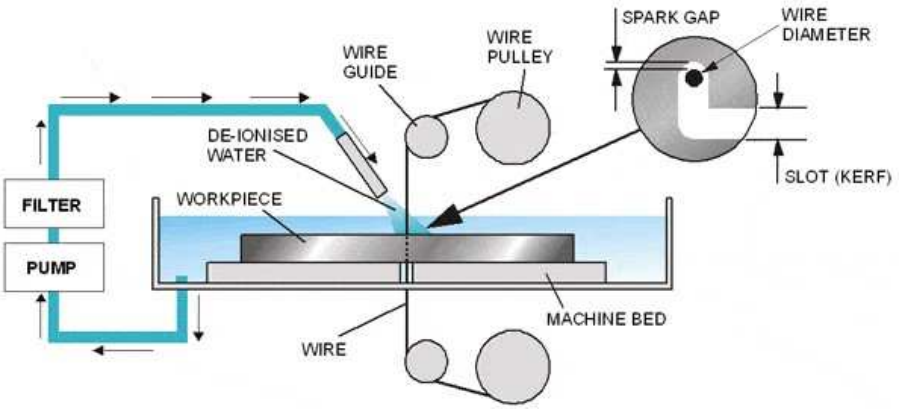
Pracovní pohyb může vykonávat jak horní, tak dolní řezací hlava nebo i obrobek. Tento pohyb je vykonáván podle předem naprogramované dráhy [20].

Obrobek i drátová elektroda jsou ponořeny v dielektrické kapalině. Ta působí jednak

jako chlazení, ale také vyplavuje drobné částice vyerodovaného materiálu z řezu [20].

Při odpařování erodovaného materiálu vzniká mezera. Šířka této mezery, přesněji řezné spáry závisí na druhu drátové elektrody, na druhu dielektrické kapaliny, obráběném materiálu a hlavně na nastavení parametrů stroje. Šířku řezné spáry je potřeba sledovat, protože ovlivní konečný rozměr obráběné součástky [21].

Celý proces WEDM je znázorněn na obrázku níže (Obrázek 3.).

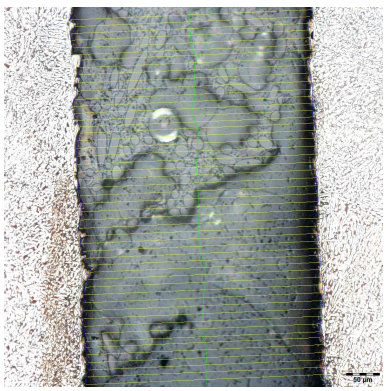


Obrázek 3: Proces WEDM.[19]

2.5 Průběh měření dat

Data [24] byla získána z hranolu slitiny Ti-6Al-4V o tloušťce 18 mm. Na základě experimentu bylo vyrobeno 33 vzorků z titanové slitiny. Z těchto vzorků byly vytvořeny metalografické výbrusy příčných řezů. Byla zde použita metoda WEDM. Na těchto preparátech se poté měřila šířka mezery řezu neboli width of kerf.

Pomocí programu pro obrazovou analýzu byla každá mezera řezu, tedy šířka mezery řezu (width of kerf) měřena na 50-ti místech.



Obrázek 4: 50 měření šířky mezery jednoho řezu.[24]

Stroj, který prováděl proces WEDM, byl CNC stroj MAKINO EU64.

Obrázek 5: CNC stroj MAKINO EU64.[22]

Za drátovou elektrodu byl zvolen PENTA CUT E. Tento drát je tvořen z 60% z mědi a z 40% ze zinku. Drát má průměr 0,25 mm.

Vzorky byly ponořeny do deionizované vody, který sloužila jako dielektrické médium a také odstranila zbytky v mezeře mezi drátovou elektrodou a obrobkem během procesu.

Plánovaný experiment byl založen na 5-ti parametrech WEDM stroje. Tyto parametry sloužily jako vstupní faktory. Mezi zkoumané parametry CNC stroje patří gap voltage, wire speed, pulse on time, pulse off time, discharge current.

Pro jednoduchší manipulaci byly zvoleny zkratky pro všechny parametry.

WK . . . . . width of kerf je šířka mezery řezu
GAV . . . . . gap voltage (V)
PON . . . . . pulse on time (μs)
POFF . . . . . pulse off time (μs)
WS . . . . . wire speed (m/min)
DC . . . . . discharge current (A)

3 ANALÝZA DAT A TVORBA MODELŮ

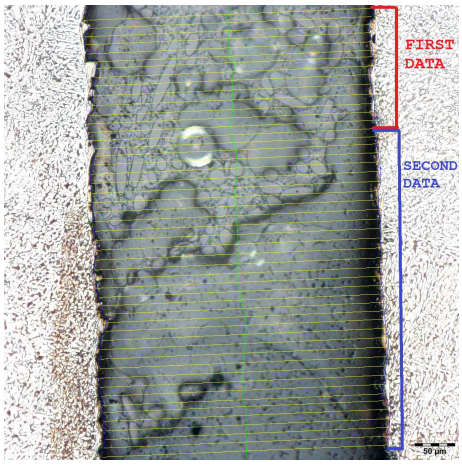
V této části budeme analyzovat data a zabývat se tvorbou regresních modelů. Data budeme analyzovat z různých pohledů. Všechny testy budeme provádět za pomocí softwaru Minitab.

Cílem je vyzkoumat, zda mezi šířkou mezery řezu a parametry CNC stroje je nějaká závislost a popřípadě jaká. Dále však chceme zjistit, zda se řez chová jinak na začátku než zbylé části řezu. Proto vytvoříme několik datových souborů a budeme je analyzovat.

První podčást této kapitoly bude řešit datový soubor MEAN DATA, kde pro každý vzorek provedeme zprůměrování hodnot padesáti měření šířky mezery každého řezu. To znamená, že pro 33 řezů získáme 33 šířek mezer řezů. Tuto část rozebereme podrobně a vysvětlíme na ni postup tvorby modelů, další části popíšeme méně detailně.

Druhá podčást se bude zabývat datovým souborem TOTAL DATA, kde budeme brát v potaz každé měření šířky mezery řezu. Pro každý vzorek máme 50 měření. Získáme tak soubor o velikosti 33x50 vstupů a výstupů.

Další podčást budeme věnovat pouze první části měření šířek mezer řezů, tedy začátku řezu. Zvolíme vždy prvních 15 hodnot. Znázorněno je to na obrázku níže (Obrázek 6.). Takový výběr provedeme pro každý jeden řez. Tak získáme datový soubor FIRST DATA o velikosti 15x33 řádků.

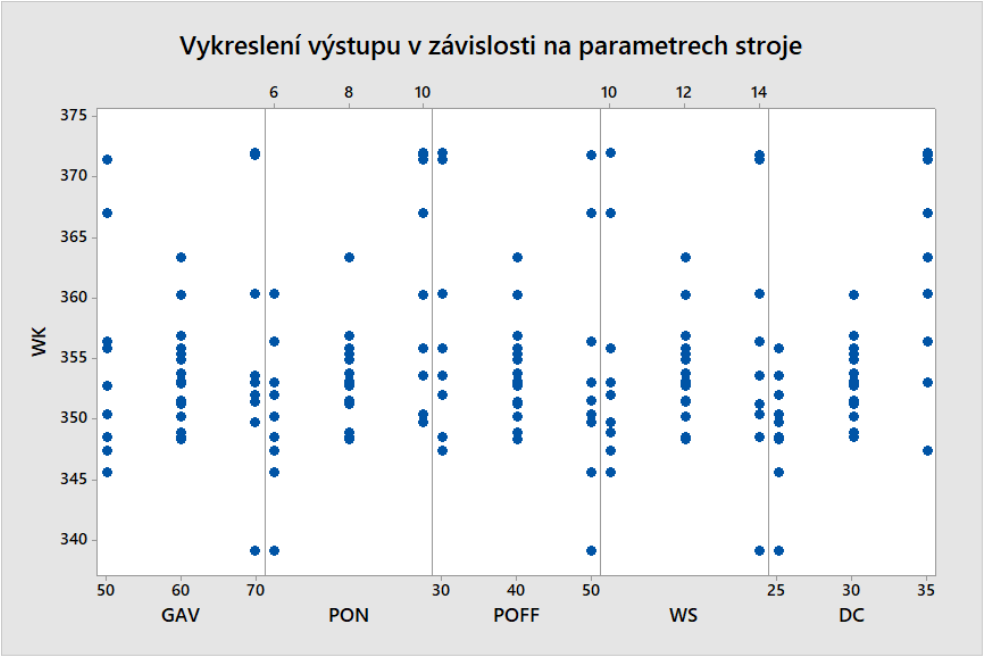


Obrázek 6: Rozdělení měření šířek mezery jednoho řezu.[vlastní zpracování]

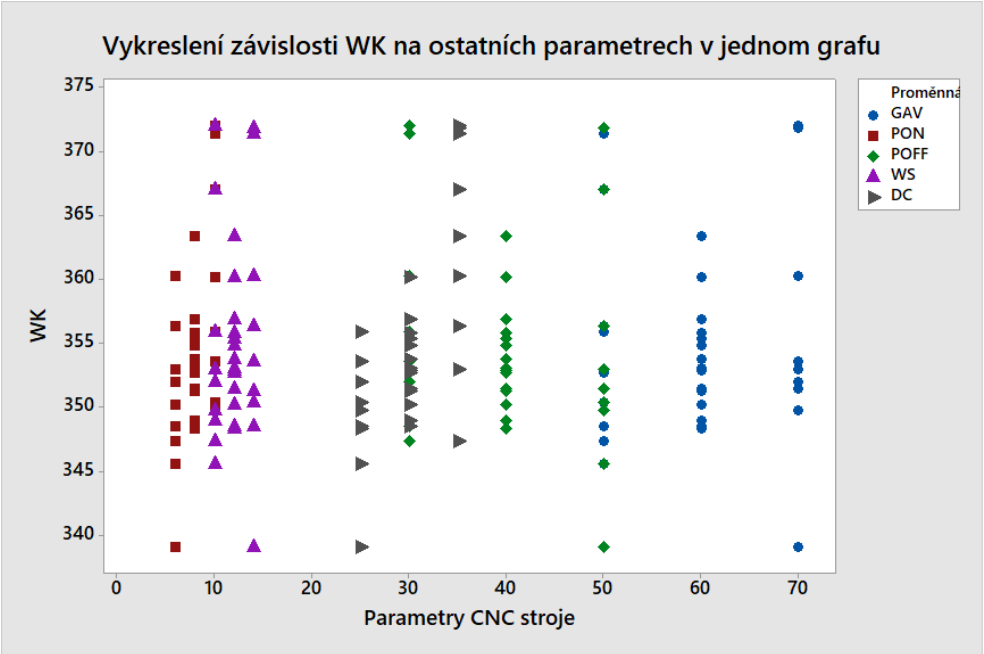
V poslední podčásti se zaměříme na druhou část měření šířek mezer řezů. Zde budeme brát zbylých 35 hodnot. Tím získáme datový soubor SECOND DATA o velikosti 35x33 řádků.

3.1 Analýza a tvorba modelu pro datový soubor MEAN DATA

Tento datový soubor obsahuje 33 položek. Celkem 33 nastavení parametrů stroje a k nim naměřené šířky mezery řezů. Na grafu níže jsou vykreslená data (Graf 1.). Tento bodový graf znázorněňuje závislost šířky mezery řezu na pěti parametrech CNC stroje.



Graf 1: Závislost WK na ostatních parametrech.[vlastní zpracování]

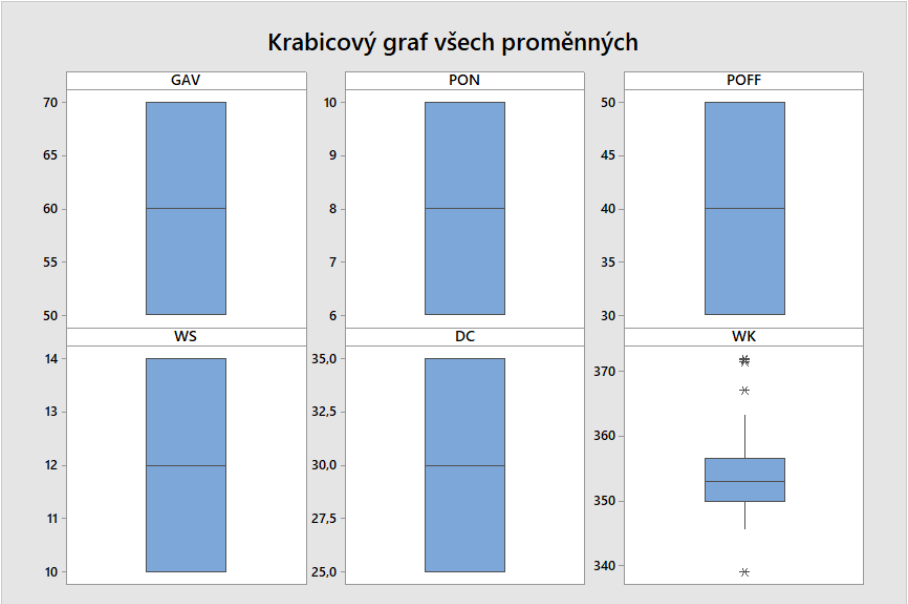


Graf 2: Závislosti WK a parametrů v jednom grafu.[vlastní zpracování]

Jelikož se jedná o vícerozměrnou regresi, není jednoduché odhadnout tvar závislosti

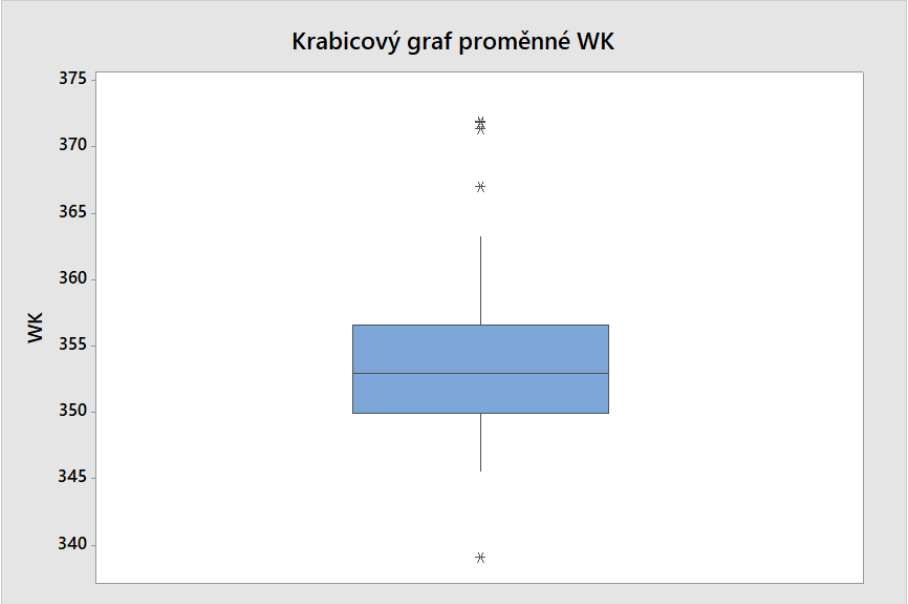
z grafu. Na bodovém grafu výše můžeme vidět všechny data v jednom grafu (Graf 2.).

Další graf je krabicový neboli boxplot graf (Graf 3.). Z grafu lze vidět, že pouze výstupní, závisle proměnná ma odlehlé hodnoty (outliery). Tento jev se zde objevuje, protože jde o plánovaný experiment a volené parametry již byly brány v jakýchsi mezích a z hodnot, které byly určeny jako vhodné a smysluplné.



Graf 3: Krabicový graf všech proměnných.[vlastní zpracování]

Pro lepší viditelnost si vykreslíme krabicový graf pouze pro proměnnou WK (Graf 4.). Proměnná WK je výstupní proměnnou, není planovaná a proto může obsahovat i odlehlé hodnoty.



Graf 4: Krabicový graf proměnné WK.[vlastní zpracování]

Provedeme test na odlehlé hodnoty proměnné WK. Nulová hypotéza je, že všechny hodnoty dat jsou v pořádku a alternativní, že menší nebo větší hodnoty dat jsou odlehlé hodnoty. Test provádíme na hladině významnosti $\alpha = 0,05$ opět pomocí Minitabu, jde o Grubbsův test pro Outliery.

Tabulka 5: Test odlehlých hodnot proměnné WK.[vlastní zpracování]

Variable	N	Mean	StDev	Min	Max	G	P
WK	33	354,61	7,55	339,06	371,97	2,30	0,562

Jak můžeme vidět (Tabulka 5.), $P - hodnota$ je vyšší něž hladina významnosti α. Proto můžeme přijmout nulovou hypotézu. To znamená, že podle tohoto testu proměnná nemá žádné odlehlé hodnoty.

S pomocí softwatu Minitab nyní vytvoříme vícerozměrný regresní model. Zvolíme podmínku, že chceme všechny parametry stroje do modelu zahrnout, požadujeme také konstantní člen. Dále umožníme, aby se v modelu objevili jakékoliv kombinace každých dvou parametrů (Obrázek 7.).

```
Terms included in every model: GAV; PON; POFF; WS; DC

Candidate terms: GAV; PON; POFF; WS; DC; GAV*GAV; PON*PON; POFF*POFF;
WS*WS; DC*DC; GAV*PON; GAV*POFF; GAV*WS; GAV*DC; PON*POFF; PON*WS; PON*DC;
POFF*WS; POFF*DC; WS*DC
```

Obrázek 7: Proměnné a jejich kombinace zahrnuté v modelu.[vlastní zpracování]

Z povahy problému jsou všechny vstupní i výstupní proměnné spojité a ne diskrétní. Tuto informaci také v softwaru nastavujeme.

Pro tvorbu modelu volíme krokovou metodu regrese. Ta bude tedy začínat se všemi parametry a s konstantou. Postupně provede v každém kroku zestupný výběr a poté vzestupný. Přidávat i odebírat z modelu proměnné budeme při $\alpha = 0,1$.

Kroková regrese proběhla v sedmi krocích. Jak je vidět z níže uvedeného (Tabulka 6.), výsledný model má nejvyšší koeficient determinace R^2. Mallowsovo C_p má nejblíže počtu regresorů a konstantě, je jich 12 a C_p je velmi blízko této hodnotě. V takovém případě podmodel úplného modelu je platný.

Přidáním každé z proměnných, které software do modelu přidal se vždy adjungovaný koeficient determinace zvýšil. Daná proměnná tedy model zlepšila.

Tabulka 6: Postupné kroky tvorby modelu.[vlastní zpracování]

Stepwise Selection of Terms

	---Step 1---		---Step 2---		---Step 3---		---Step 4---	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	291,7		352,8		425,1		521,5	
GAV	0,0430	0,651	0,0430	0,611	0,0430	0,578	0,0430	0,534
PON	2,773	0,000	-4,87	0,082	-4,87	0,058	-4,87	0,036
POFF	-0,1384	0,153	-0,1384	0,109	-0,1384	0,081	-0,1384	0,054
WS	0,314	0,509	0,314	0,458	-5,71	0,028	-5,71	0,016
DC	1,331	0,000	-0,707	0,340	-3,12	0,014	-9,64	0,001
PON*DC			0,2547	0,008	0,2547	0,004	0,2547	0,002
WS*DC					0,2007	0,020	0,2007	0,011
DC*DC							0,1088	0,013
POFF*DC								
GAV*DC								
GAV*POFF								
S	3,98952		3,54044		3,23371		2,89345	
R-sq	76,46%		82,15%		85,68%		88,99%	
R-sq(adj)	72,10%		78,03%		81,67%		85,32%	
R-sq(pred)	57,59%		64,95%		69,84%		74,33%	
Mallows' Cp	59,80		42,27		32,15		22,78	

	-----Step 5----		-----Step 6----		------Step 7-----	
	Coef	P	Coef	P	Coef	P
Constant	555,0		602,1		571,5	
GAV	0,0430	0,509	-0,743	0,067	-0,233	0,592
PON	-4,87	0,027	-4,87	0,019	-4,87	0,013
POFF	-0,975	0,027	-0,975	0,019	-0,210	0,680
WS	-5,71	0,011	-5,71	0,007	-5,71	0,004
DC	-10,76	0,000	-12,33	0,000	-12,33	0,000
PON*DC	0,2547	0,001	0,2547	0,001	0,2547	0,000
WS*DC	0,2007	0,007	0,2007	0,005	0,2007	0,003
DC*DC	0,1088	0,009	0,1088	0,006	0,1088	0,003
POFF*DC	0,0279	0,052	0,0279	0,039	0,0279	0,027
GAV*DC			0,0262	0,051	0,0262	0,037
GAV*POFF					-0,01275	0,042
S	2,71704		2,54358		2,35357	
R-sq	90,70%		92,20%		93,63%	
R-sq(adj)	87,06%		88,66%		90,29%	
R-sq(pred)	76,55%		79,87%		86,15%	
Mallows' Cp	18,92		15,76		12,87	

Při testu významnosti modelu $P - hodnota$ vychází menší než $0,001$ a aby vztah existoval, je potřeba, aby tato hodnota byla nižší než je zvolená α, takže v tomto případě vztah existuje a je významný.

Spolu s vícerozměrnou regresní analýzou provádíme v softwaru také analýzu rozptylu (Tabulka 7.).

Tabulka 7: Analýza rozptylu.[vlastní zpracování]

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	11	1708,99	155,363	28,05	0,000
GAV	1	1,64	1,642	0,30	0,592
PON	1	41,14	41,143	7,43	0,013
POFF	1	0,97	0,972	0,18	0,680
WS	1	56,53	56,528	10,20	0,004
DC	1	157,34	157,343	28,40	0,000
DC*DC	1	60,49	60,493	10,92	0,003
GAV*POFF	1	26,01	26,010	4,70	0,042
GAV*DC	1	27,46	27,458	4,96	0,037
PON*DC	1	103,84	103,836	18,75	0,000
POFF*DC	1	31,14	31,136	5,62	0,027
WS*DC	1	64,48	64,481	11,64	0,003
Error	21	116,33	5,539		
Lack-of-Fit	15	102,94	6,863	3,08	0,086
Pure Error	6	13,39	2,231		
Total	32	1825,31			

Až na dvě $P - hodnoty$ jsou všechny nižší než je zvolená α. To znamená, že mezi těmito proměnnými a závisle proměnnou šířkou mezery řezu existuje významný vztah. Jelikož však máme podmínku zahrnout všechny parametry, musíme akceptovat i dva parametry s vyšší $P - hodnotou$.

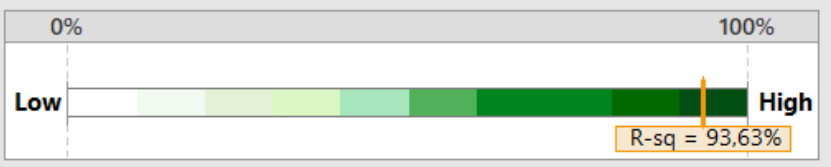
Nyní si ukážeme jakési shrnutí modelu (Tabulka 8.).

Tabulka 8: Hodnoty shrnující vhodnost modelu.[vlastní zpracování]

S	R-sq	R-sq(adj)	R-sq(pred)
2,35357	93,63%	90,29%	86,15%

S nám říká, že směrodatná odchylka dat oproti modelu je přibližně 2,35. Tuto hodnotu bychom mohli použít při porovnávání modelů. Model s nižší hodnotou bude lépe vystihovat problém, bude přesnější.

Hodnota R^2 udává kolik procent problému je vysvětlováno tímto modelem. Tedy 93,63% dat je popsáno tímto modelem. Toto číslo je velmi dobré.



Obrázek 8: Koeficient determinace modelu.[vlastní zpracování]

$R^2(adj)$ využíváme při porovnávání modelů s různým počtem proměnných.

Hodnota $R^2(pred)$ říká, jak dobře bude model predikovat šířky mezery řezů pro nová pozorování. Model vykazuje tuto hodnotu ve výši 86,15%, což je výborné.

Nyní si rozebereme tabulku (Tabulka 9.) s koeficienty regresního modelu pro všechny zahrnuté proměnné. Tyto koeficienty jsou uvedeny v prvním sloupci (Tabulka 9.).

Hodnota $SE\ Coef$ z tabulky (Tabulka 9.) udává přesnost odhadu koeficientů. Jedná se o standardní chybu, čím menší je, tím přesnější je odhad.

Další využívanou hodnotou je $P-hodnota$ (Tabulka 9.). Tu potřebujeme při testování, zda je daný koeficient modelu významně různý od 0, jde o test statistické významnosti regresního koeficientu. Opět až na dva parametry je to v pořádku a jelikož tyto parametry GAV a POFF musí být v modelu zahrnuty, tento nedostatek nebudeme řešit.

Tabulka 9: Koeficienty regresního modelu.[vlastní zpracování]

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	571,5	48,9	11,69	0,000	
GAV	-0,233	0,428	-0,54	0,592	59,50
PON	-4,87	1,79	-2,73	0,013	41,50
POFF	-0,210	0,502	-0,42	0,680	82,00
WS	-5,71	1,79	-3,19	0,004	41,50
DC	-12,33	2,31	-5,33	0,000	434,80
DC*DC	0,1088	0,0329	3,30	0,003	317,80
GAV*POFF	-0,01275	0,00588	-2,17	0,042	59,50
GAV*DC	0,0262	0,0118	2,23	0,037	82,00
PON*DC	0,2547	0,0588	4,33	0,000	59,50
POFF*DC	0,0279	0,0118	2,37	0,027	59,50
WS*DC	0,2007	0,0588	3,41	0,003	82,00

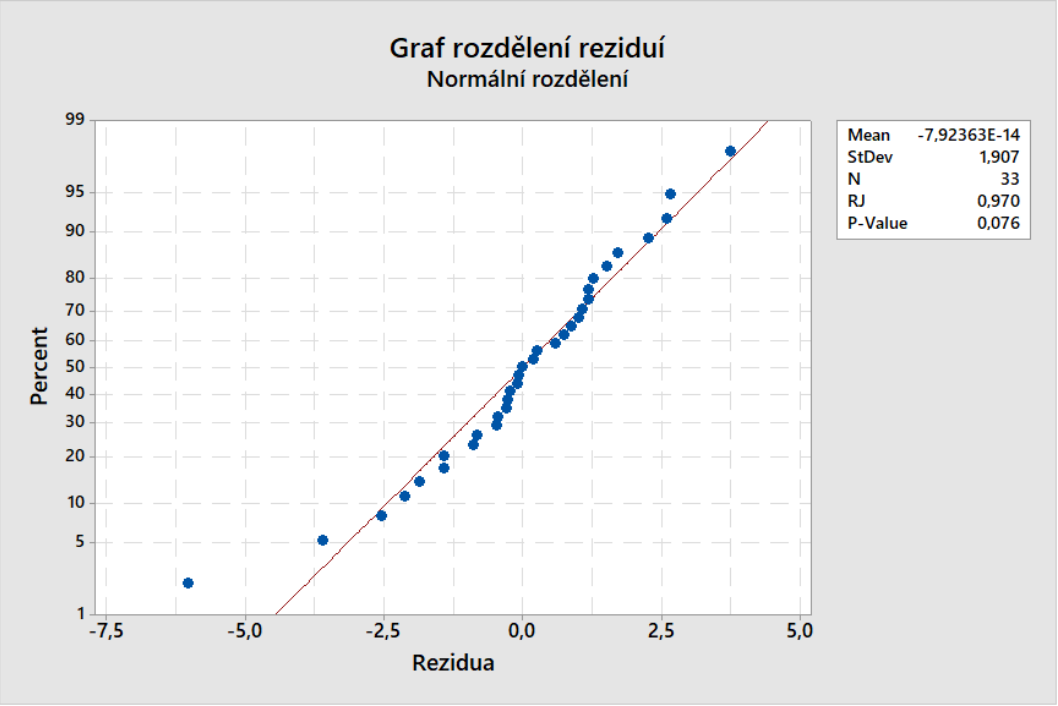
Nyní můžeme zapsat výsledný model, který jsme získali vícerozměrnou regresí (Obrázek 9.).

```
WK    = 571,5 - 0,233 GAV - 4,87 PON - 0,210 POFF - 5,71 WS
        - 12,33 DC + 0,1088 DC*DC - 0,01275 GAV*POFF
        + 0,0262 GAV*DC + 0,2547 PON*DC + 0,0279 POFF*DC
        + 0,2007 WS*DC
```

Obrázek 9: Výsledná rovnice popisující model.[vlastní zpracování]

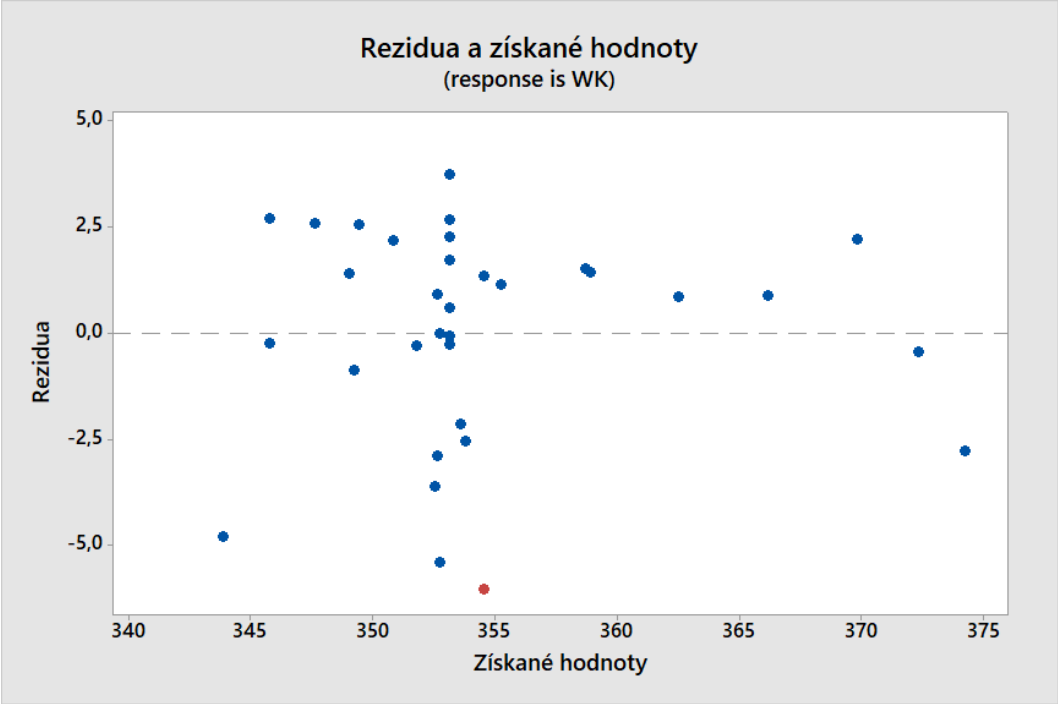
Pro model jsme vypočítali Durbin-Watsonovu statistiku. Ta vyšla 1,49071, což je hodnota vysoká a blízká dvěma, proto nás netrápí autokorelace.

Dalším předpokladem, který po modelu požadujeme je normalita reziduí (Graf 5.). Pomocí Shapiro-Wilkova testu jsem získali $P-hodnotu$ vyšší než α. Hypotézu o normálním rozdělení tedy nezamítáme. Rezidua modelu mají tedy přibližně normální rozdělení. Grafické znázornění tohoto rozdělení je na následujícím grafu (Graf 5.).



Graf 5: Test rozdělení reziduí pro náš model.[vlastní zpracování]

Následující graf zobrazuje rezidua v závislosti na získaných hodnotách šířky mezery řezu (Graf 6.).



Graf 6: Rezidua v závislosti na získaných hodnotách.[vlastní zpracování]

V grafu (Graf 6.) byl zaznačen červeně jeden bod, který má velké reziduum. Dané pozorování zjistíme i pomocí softwaru. Přestože jsme na začátku prováděli test na nalezení odlehlých hodnot, našli jsme teď nějaké (Tabulka 10.). Proto pozorování s vysokými hodnotami reziduí prověříme znovu a jinak.

Tabulka 10: Odlehlá pozorování.[vlastní zpracování]

Obs	WK	Fit	Resid	Std Resid
2	348,48	354,51	-6,03	-2,74

Provedeme odstranění tohoto pozorování a vypočítáme regresní model znovu. Pokud v modelu dojde ke zlepšení, pozorování do modelu již nevrátíme. V našem modelu došlo ke zlepšení všech parametrů, ale objevilo se další pozorování s velkým reziduem. Proces odstranění byl proveden ještě jednou, odstraněny byly tedy celkem 3 pozorování s velkými rezidui.

Výsledný model má nejlepší hodnoty a neobsahuje žádná odlehlá pozorování. Výsledný model vznikal v těchto krocích (Tabulka 11., 12.).

Tabulka 11: Kroky tvorby modelu - 1.část.[vlastní zpracování]

	Step 1		Step 2		Step 3		Step 4	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	292,5		353,6		425,9		507,6	
GAV	0,0662	0,481	0,0662	0,411	0,0662	0,35	0,058	0,369
PON	2,773	0	-4,87	0,061	-4,87	0,035	-4,87	0,022
POFF	-0,1845	0,057	-0,1845	0,029	-0,1845	0,014	-0,1764	0,011
WS	0,314	0,49	0,314	0,421	-5,71	0,015	-5,71	0,009
DC	1,331	0	-0,707	0,302	-3,12	0,007	-8,65	0,002
PON*DC			0,2547	0,005	0,2547	0,002	0,2547	0,001
WS*DC					0,2008	0,01	0,2007	0,006
DC*DC							0,0922	0,027
POFF*DC								
GAV*DC								
GAV*POFF								
WS*WS								
PON*PON								
GAV*WS								
S	3,80771		3,25798		2,85761		2,59718	
R-sq	80,35%		86,21%		89,86%		92,00%	
R-sq(adj)	76,26%		82,62%		86,63%		88,95%	
R-sq(pred)	61,37%		69,50%		75,12%		77,41%	
Mallows' Cp	186,49		127,47		91,58		71,24	

	Step 5		Step 6		Step 7		Step 8	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	541,1		588,2		557,6		537,6	
GAV	0,058	0,322	-0,728	0,034	-0,218	0,509	-0,216	0,466
PON	-4,87	0,013	-4,87	0,006	-4,87	0,002	-4,87	0,001
POFF	-1,013	0,01	-1,013	0,005	-0,248	0,522	-0,25	0,471
WS	-5,71	0,005	-5,71	0,002	-5,71	0,001	6,41	0,241
DC	-9,76	0	-11,33	0	-11,33	0	-14,83	0
PON*DC	0,2547	0	0,2547	0	0,2547	0	0,2547	0
WS*DC	0,2007	0,003	0,2007	0,001	0,2007	0	0,2007	0
DC*DC	0,0922	0,016	0,0922	0,008	0,0922	0,003	0,1504	0
POFF*DC	0,0279	0,028	0,0279	0,015	0,0279	0,006	0,0279	0,003
GAV*DC			0,0262	0,021	0,0262	0,009	0,0262	0,004
GAV*POFF					-0,0128	0,01	-0,0128	0,005
WS*WS							-0,505	0,031
PON*PON								
GAV*WS								
S	2,3507		2,09081		1,78027		1,5905	
R-sq	93,76%		95,31%		96,78%		97,57%	
R-sq(adj)	90,95%		92,84%		94,81%		95,86%	
R-sq(pred)	79,80%		83,46%		90,42%		90,15%	
Mallows' Cp	54,95		40,81		27,53		21,27	

Tabulka 12: Kroky tvorby modelu - 2.část.[vlastní zpracování]

	Step 9		Step 10	
	Coef	P	Coef	P
Constant	506,4		477,8	
GAV	-0,217	0,391	0,26	0,346
PON	-13,64	0,001	-13,64	0
POFF	-0,249	0,402	-0,249	0,32
WS	11,92	0,028	14,31	0,004
DC	-12,62	0	-12,62	0
PON*DC	0,2547	0	0,2547	0
WS*DC	0,2007	0	0,2007	0
DC*DC	0,1137	0,003	0,1137	0,001
POFF*DC	0,0279	0,001	0,0279	0
GAV*DC	0,0262	0,001	0,0262	0
GAV*POFF	-0,0128	0,002	-0,0128	0
WS*WS	-0,735	0,002	-0,735	0,001
PON*PON	0,548	0,015	0,548	0,005
GAV*WS			-0,0398	0,013
S		1,3549		1,13312
R-sq		98,34%		98,91%
R-sq(adj)		96,99%		97,90%
R-sq(pred)		91,71%		95,70%
Mallows' Cp		15,26		11,32

Tento model vznikal tedy 10 kroků. V modelu se opravdu všechny hodnoty zlepšily. I hodnoty koeficientů determinace stouply. Tento model vysvětluje náš problém dokonce z 98,91% (Tabulka 13.).

Tabulka 13: Hodnoty shrnující vhodnost modelu.[vlastní zpracování]

S	R-sq	R-sq(adj)	R-sq(pred)
1,13312	98,91%	97,90%	95,70%

Dále si opět uvedeme analýzu rozptylu pro nově vzniklý model (Tabulka 14.). Až na dvě $P - hodnoty$ jsou všechny nižší než je zvolená α. Mezi těmito proměnnými a závisle proměnnou existuje významný vztah. Naší podmínkou je však zahrnout všech 5 parametrů stroje, musíme akceptovat i parametry GAV a POFF s vyšší $P - hodnotou$.

Tabulka 14: Analýza rozptylu.[vlastní zpracování]

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	14	1751,58	125,113	97,44	0,000
GAV	1	1,21	1,214	0,95	0,346
PON	1	29,92	29,923	23,31	0,000
POFF	1	1,36	1,361	1,06	0,320
WS	1	14,80	14,803	11,53	0,004
DC	1	69,53	69,534	54,16	0,000
PON*PON	1	13,63	13,633	10,62	0,005
WS*WS	1	24,50	24,496	19,08	0,001
DC*DC	1	22,92	22,917	17,85	0,001
GAV*POFF	1	26,01	26,010	20,26	0,000
GAV*WS	1	10,11	10,112	7,88	0,013
GAV*DC	1	27,46	27,458	21,38	0,000
PON*DC	1	103,84	103,836	80,87	0,000
POFF*DC	1	31,14	31,136	24,25	0,000
WS*DC	1	64,48	64,481	50,22	0,000
Error	15	19,26	1,284		
Lack-of-Fit	10	11,59	1,159	0,76	0,670
Pure Error	5	7,67	1,533		
Total	29	1770,84			

Koeficienty tohoto modelu jsme také přepočítali. Až na dva koeficienty jsou všechny statisticky významné. U GAV a POFF jsou $P - hodnoty$ jsou vyšší než zvolená α a proto nezamítáme hypotézu, že jsou tyto koeficienty nulové. Tyto koeficienty však z modelu nemůžeme vynechat (Tabulka 15.).

Tabulka 15: Koeficienty modelu.[vlastní zpracování]

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	477,8	28,5	16,74	0,000	
GAV	0,260	0,267	0,97	0,346	94,32
PON	-13,64	2,82	-4,83	0,000	447,48
POFF	-0,249	0,242	-1,03	0,320	77,35
WS	14,31	4,21	3,40	0,004	995,45
DC	-12,62	1,72	-7,36	0,000	1031,45
PON*PON	0,548	0,168	3,26	0,005	408,51
WS*WS	-0,735	0,168	-4,37	0,001	915,99
DC*DC	0,1137	0,0269	4,22	0,001	915,99
GAV*POFF	-0,01275	0,00283	-4,50	0,000	56,24
GAV*WS	-0,0398	0,0142	-2,81	0,013	79,67
GAV*DC	0,02620	0,00567	4,62	0,000	79,67
PON*DC	0,2547	0,0283	8,99	0,000	59,50
POFF*DC	0,02790	0,00567	4,92	0,000	57,17
WS*DC	0,2007	0,0283	7,09	0,000	82,00

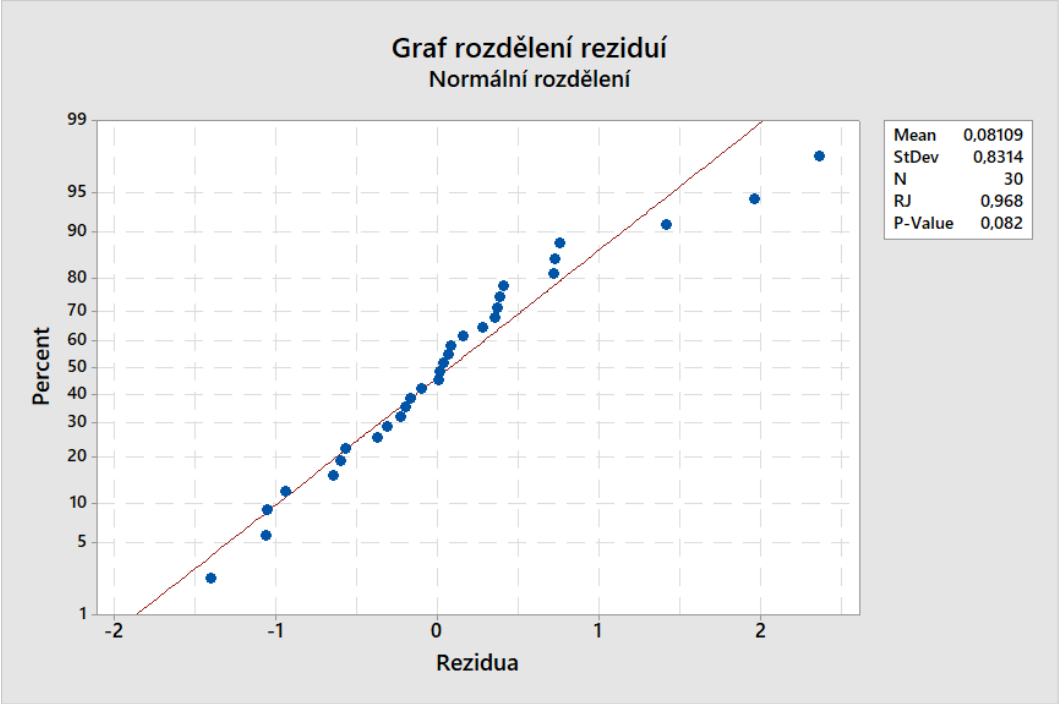
Výsledný model má tvar (Obrázek 10.):

```
WK   =  477,8 + 0,260 GAV - 13,64 PON - 0,249 POFF
        + 14,31 WS - 12,62 DC + 0,548 PON*PON
        - 0,735 WS*WS + 0,1137 DC*DC
        - 0,01275 GAV*POFF - 0,0398 GAV*WS
        + 0,02620 GAV*DC + 0,2547 PON*DC
        + 0,02790 POFF*DC + 0,2007 WS*DC
```

Obrázek 10: Výsledný vztah popisující model - MEAN.[vlastní zpracování]

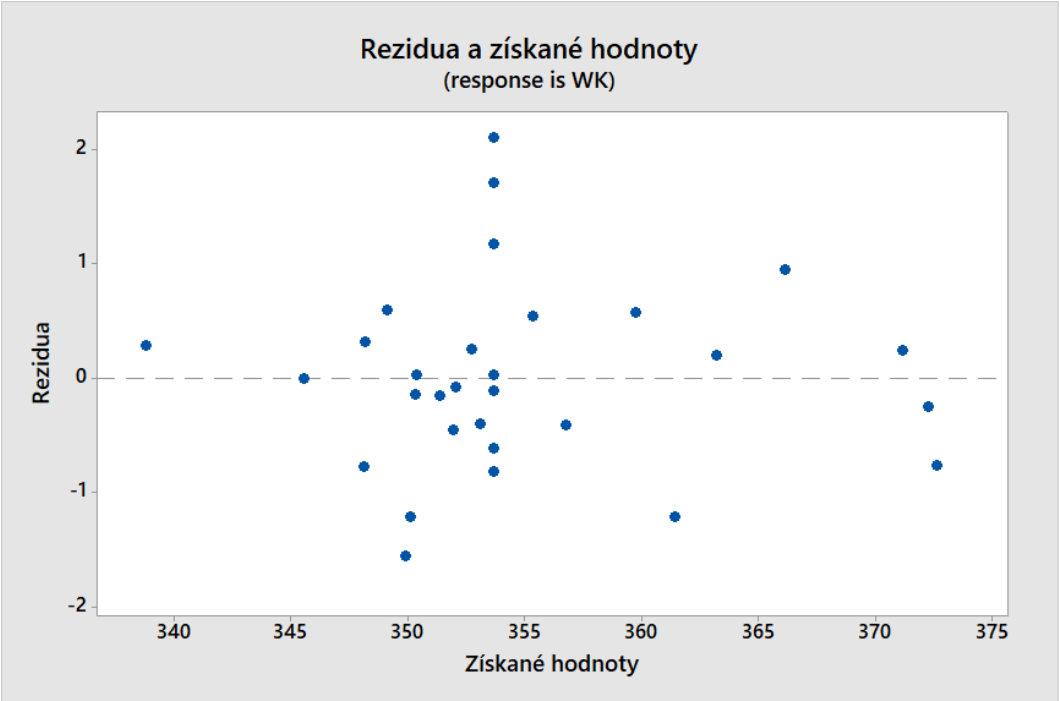
Durbin-Watsonova statistika v tomto modelu vyšla 1,87592, což je opět blízké dvěma a tím pádem není problém s autokorelací.

Normalitu opět otestujeme Shapiro-Wilkovým testem (Graf 7.). Opět je $P - hodnota$ vyšší než zvolená hladina významnosti a nezamítáme tedy hypotézu, že rezidua mají normální rozdělení.



Graf 7: Test rozdělení reziduí pro náš model.[vlastní zpracování]

Nyní vykreslíme závislost reziduí na výstupních hodnotách (Graf 8.).

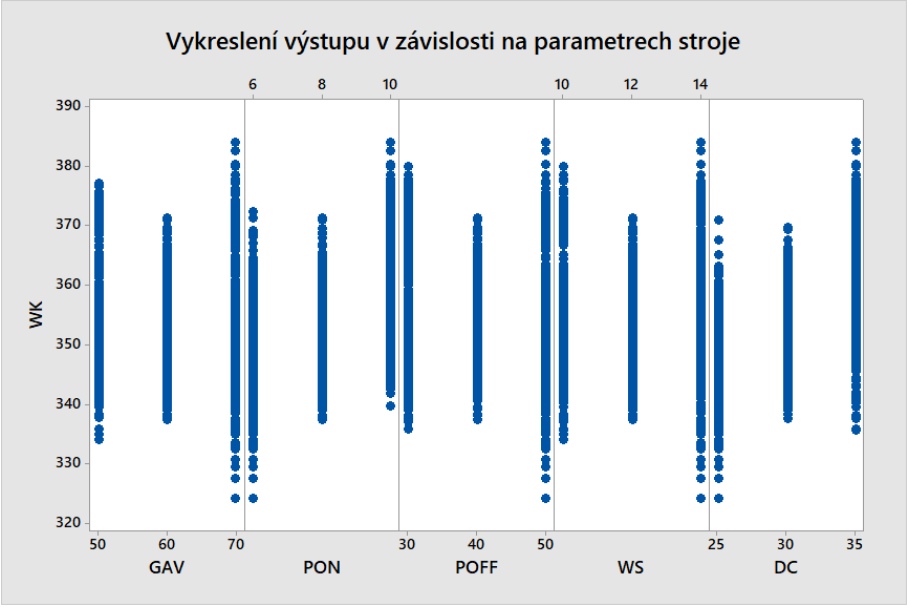


Graf 8: Rezidua v závislosti na získaných hodnotách.[vlastní zpracování]

Tento model splňuje všechny předpoklady a nejlépe vystihuje vztah mezery šířky řezu na nastavení CNC stroje.

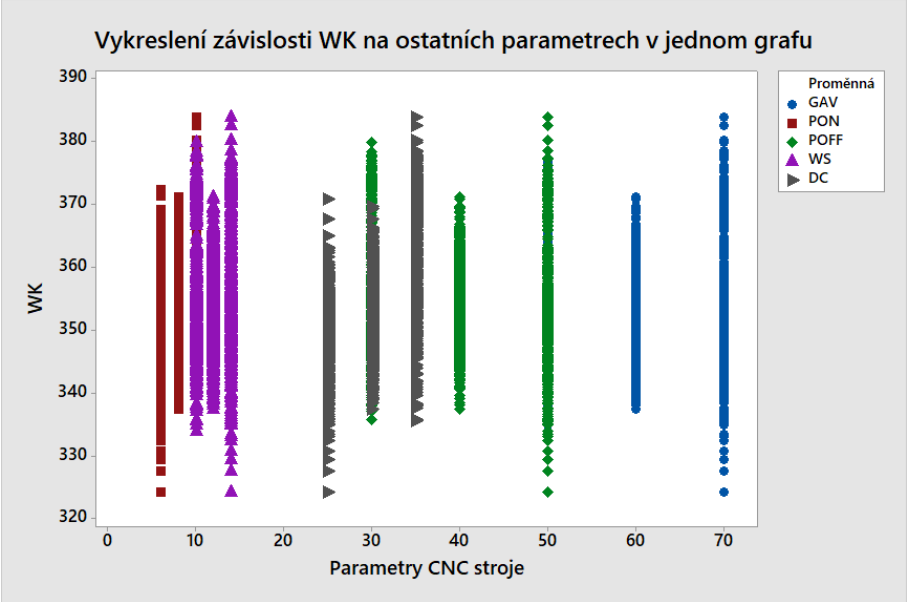
3.2 Analýza a tvorba modelu pro datový soubor TOTAL DATA

Tento datový soubor obsahuje 1650 položek, nastavení parametrů stroje a pro ně šířku mezery řezu. Na grafu níže jsou data v bodovém grafu, který znázorněňuje závislost šířky mezery řezu na našich pěti parametrech CNC stroje (Graf 9.).



Graf 9: Závislost WK na ostatních parametrech.[vlastní zpracování]

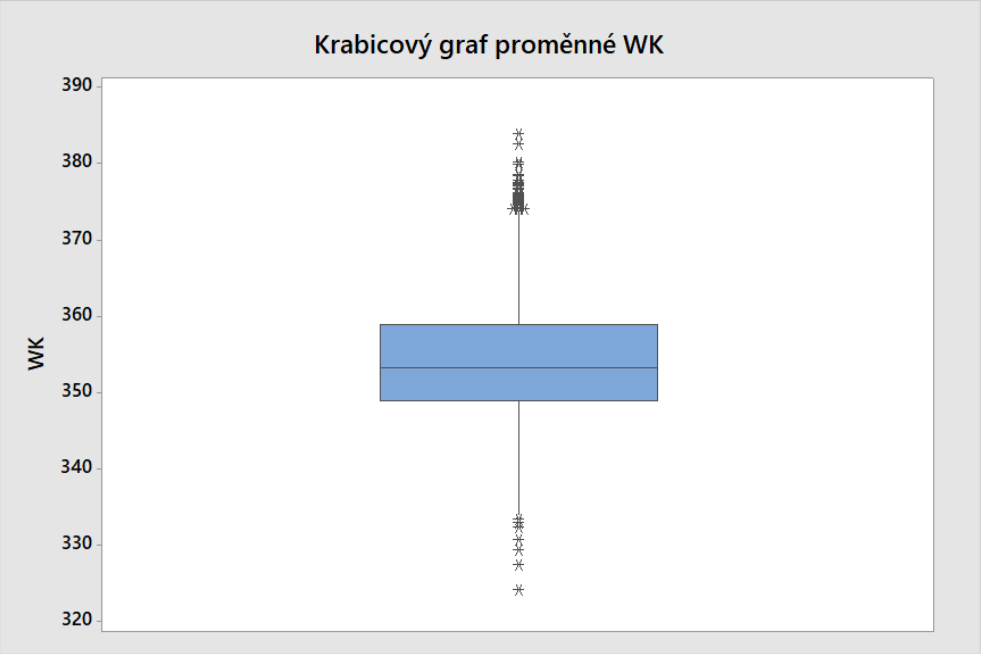
Vykreslíme ještě jeden bodový graf, kde budou veškerá data v jednom grafu (Graf 10.).



Graf 10: Závislosti WK a parametrů v jednom grafu.[vlastní zpracování]

Následujícím grafem je krabicový graf pouze pro proměnnou WK (Graf 11.). Jak

jsme již uváděli v předchozí části, nastavení CNC stroje, tedy našich 5 parametrů nebude vykazovat známky odlehlých hodnot, jelikož se jedná o plánovaný experiment, proto se podíváme pouze na proměnnou WK.



Graf 11: Krabicový graf proměnné WK.[vlastní zpracování]

Provedeme test na odlehlé hodnoty proměnné WK. Nulová hypotéza je, že všechny hodnoty dat jsou v pořádku a alternativní, že menší nebo větší hodnoty dat jsou odlehlé hodnoty.

Tabulka 16: Test odlehlých hodnot proměnné WK.[vlastní zpracování]

Variable	N	Mean	StDev	Min	Max	G	P
WK	1650	354,49	8,79	324,12	383,90	3,45	0,894

Můžeme přijmout nulovou hypotézu. To znamená, že podle tohoto testu proměnná nemá žádné odlehlé hodnoty na hladině významnosti α=0,05 (Tabulka 16.).

S pomocí Minitabu vytvoříme vícerozměrný regresní model obdobným způsobem jako v předchozí části. Využijeme opět krokovou regresi s hladinou významnosti $\alpha = 0,1$. Kroková regrese proběhla v 11 krocích a získali jsme díky ní rovnici pro výsledný model (Obrázek 11.).

```
WK   = 523,6 + 0,243 GAV - 17,30 PON + 1,481 POFF + 8,80 WS - 14,84 DC
       + 0,777 PON*PON - 0,02115 POFF*POFF - 0,505 WS*WS + 0,1505 DC*DC
       - 0,01275 GAV*POFF - 0,03970 GAV*WS + 0,02620 GAV*DC
       + 0,2547 PON*DC + 0,02791 POFF*DC + 0,2008 WS*DC
```

Obrázek 11: Výsledný vztah popisující model - TOTAL.[vlastní zpracování]

Všechny koeficienty zahrnuté v modelu jsou statisticky významné na hladině významnosti α až na koeficient u proměnné GAV, tu však v modelu vyžadujeme (Tabulka 17.).

Tabulka 17: Koeficienty modelu.[vlastní zpracování]

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	523,6	17,7	29,57	0,000	
GV	0,243	0,163	1,49	0,136	100,00
PON	-17,30	1,79	-9,65	0,000	484,28
POFF	1,481	0,373	3,97	0,000	524,78
WS	8,80	2,68	3,29	0,001	1078,24
DC	-14,84	1,09	-13,64	0,000	1114,24
PON*PON	0,777	0,107	7,25	0,000	445,92
POFF*POFF	-0,02115	0,00429	-4,93	0,000	445,92
WS*WS	-0,505	0,107	-4,71	0,000	999,39
DC*DC	0,1505	0,0171	8,78	0,000	999,39
GV*POFF	-0,01275	0,00173	-7,37	0,000	59,50
GV*WS	-0,03970	0,00864	-4,59	0,000	82,00
GV*DC	0,02620	0,00346	7,58	0,000	82,00
PON*DC	0,2547	0,0173	14,73	0,000	59,50
POFF*DC	0,02791	0,00346	8,07	0,000	59,50
WS*DC	0,2008	0,0173	11,62	0,000	82,00

Model vystihuje náš problém z 69,36% (Tabulka 18.).

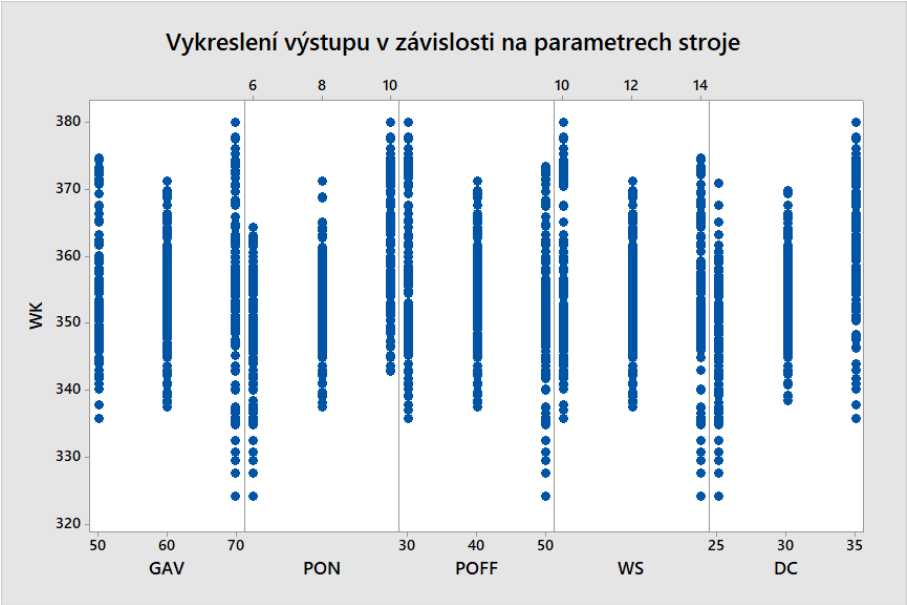
Tabulka 18: Hodnoty shrnující vhodnost modelu.[vlastní zpracování]

S	R-sq	R-sq(adj)	R-sq(pred)
4,89010	69,36%	69,08%	68,78%

Při tvorbě modelu bylo odhaleno 88 odlehlých pozorování. Postupně byla každá odlehlá hodnota odstraněna a model přepočítán. Odstranění však u žádného pozorování nezlepšilo hodnoty modelu významně.

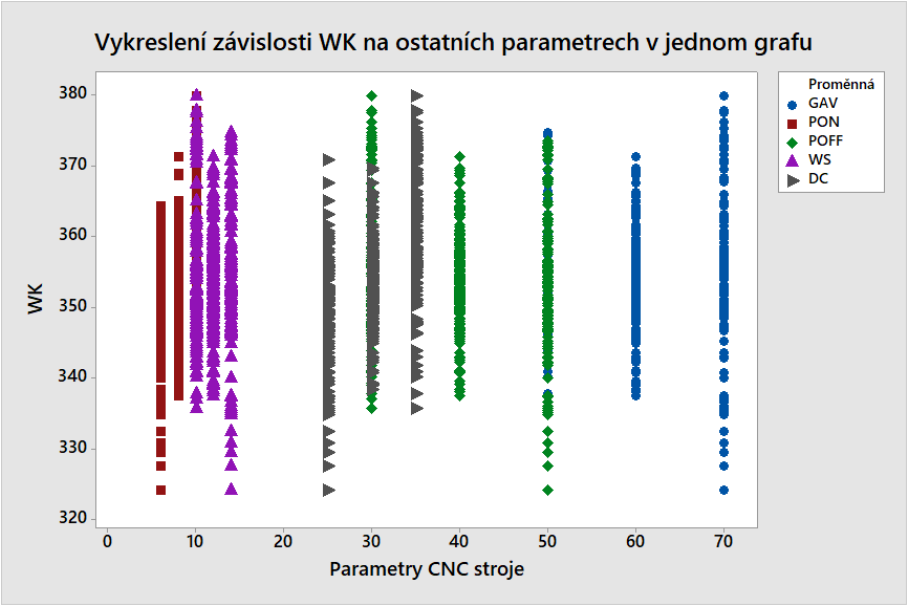
3.3 Analýza a tvorba modelu pro datový soubor FIRST PART DATA

Tento datový soubor obsahuje 495 položek. Tento datový soubor však zkoumá, jak se chová řez na začátku. Proto jsme z celkového datového souboru vzali pouze počáteční měření. Na grafu níže jsou vykreslená data (Graf 12.). Tento bodový graf znázorněňuje závislost šířky mezery řezu na našich pěti parametrech CNC stroje.



Graf 12: Závislost WK na ostatních parametrech.[vlastní zpracování]

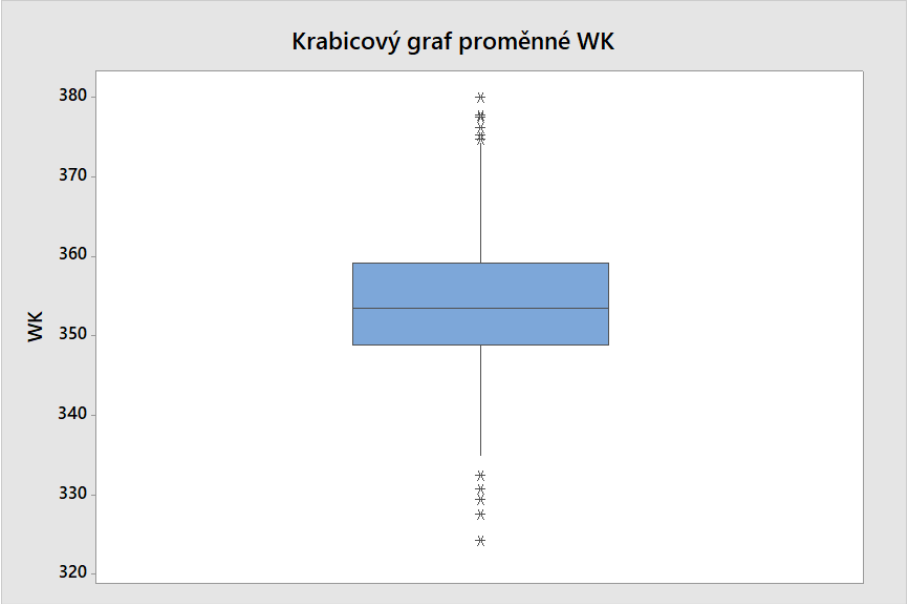
Na bodovém grafu níže můžeme vidět všechny data v jednom grafu (Graf 13.).



Graf 13: Závislosti WK a parametrů v jednom grafu.[vlastní zpracování]

Další graf (Graf 14.) je krabicový graf pro proměnnou WK a můžeme z něj vidět, že

výstupní, závisle proměnná má odlehlé hodnoty (outliery).



Graf 14: Krabicový graf proměnné WK.[vlastní zpracování]

Pro jistotu provedeme test na odlehlé hodnoty proměnné WK. Nulová hypotéza je, že všechny hodnoty dat jsou v pořádku a alternativní, že menší nebo větší hodnoty dat jsou odlehlé hodnoty.

Tabulka 19: Test odlehlých hodnot proměnné WK.[vlastní zpracování]

Variable	N	Mean	StDev	Min	Max	G	P
WK	495	354,23	9,02	324,12	379,91	3,34	0,393

P-hodnota je vyšší něž hladina významnosti α. Proto můžeme přijmout nulovou hypotézu a tedy podle tohoto testu proměnná nemá žádné odlehlé hodnoty.

S pomocí Minitabu vytvoříme vícerozměrný regresní model obdobným způsobem, jakým jsme vytvářeli předchozí modely. Postup tvorby modelu nebudeme vysvětlovat detailně jako tomu bylo při prvním datovém souboru MEAN DATA. Po 11 krocích krokové regrese jsme získali následující model (Obrázek 12.).

```
WK  =  471,0 + 0,510 GAV - 17,44 PON + 3,417 POFF + 0,75 WS
       - 10,94 DC + 0,935 PON*PON - 0,03829 POFF*POFF
       + 0,0774 DC*DC - 0,01616 GAV*POFF - 0,0942 GAV*WS
       + 0,04372 GAV*DC + 0,1854 PON*DC - 0,0325 POFF*WS
       + 0,02773 POFF*DC + 0,1981 WS*DC
```

Obrázek 12: Výsledný vztah popisující model- FIRST.[vlastní zpracování]

Koeficienty modelu tedy jsou (Tabulka 20.):

Tabulka 20: Koeficienty modelu.[vlastní zpracování]

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	471,0	31,0	15,22	0,000	
GAV	0,510	0,298	1,71	0,088	100,00
PON	-17,44	3,14	-5,55	0,000	444,19
POFF	3,417	0,683	5,00	0,000	525,19
WS	0,75	1,49	0,50	0,617	100,00
DC	-10,94	1,91	-5,74	0,000	1024,05
PON*PON	0,935	0,187	5,00	0,000	405,55
POFF*POFF	-0,03829	0,00747	-5,12	0,000	405,55
DC*DC	0,0774	0,0299	2,59	0,010	908,91
GAV*POFF	-0,01616	0,00316	-5,11	0,000	59,50
GAV*WS	-0,0942	0,0158	-5,96	0,000	82,00
GAV*DC	0,04372	0,00632	6,92	0,000	82,00
PON*DC	0,1854	0,0316	5,87	0,000	59,50
POFF*WS	-0,0325	0,0158	-2,06	0,040	59,50
POFF*DC	0,02773	0,00632	4,39	0,000	59,50
WS*DC	0,1981	0,0316	6,27	0,000	82,00

Všechny tyto koeficienty jsou statisticky významné až na koeficient u proměnné WS. Jeho $P - hodnota$ je vyšší než zvolená hladina α. Parametr WS je v modelu vyžadován a proto tento nedostatek zanedbáme.

Jak můžeme vidět níže (Tabulka 21.), model vysvětluje náš problém ze 71,45%.

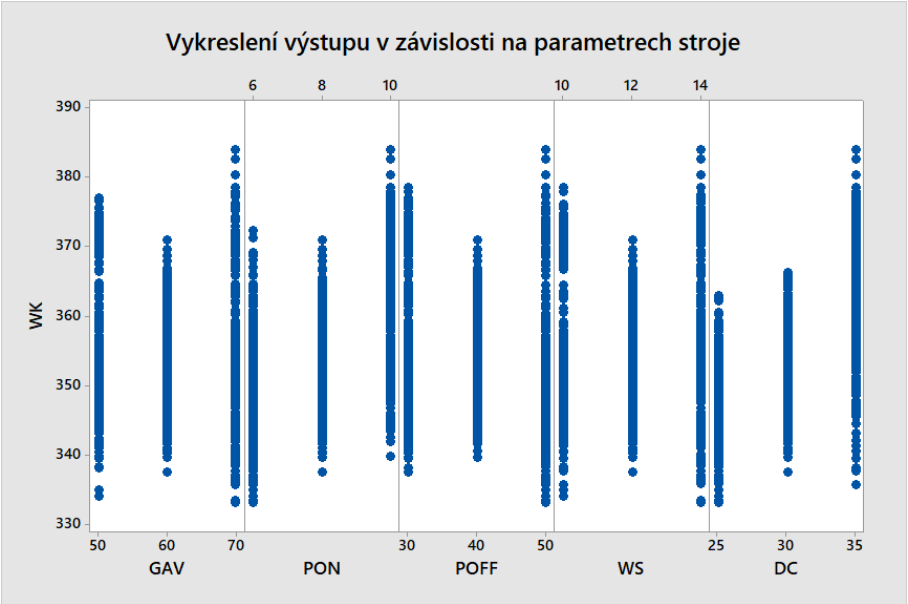
Tabulka 21: Hodnoty shrnující vhodnost modelu.[vlastní zpracování]

S	R-sq	R-sq(adj)	R-sq(pred)
4,89551	71,45%	70,56%	69,51%

Při tvořbě modelu bylo však odhaleno 27 pozorování, které vykazují známky velkých reziduí. Každá z těchto hodnot byla postupně odstraněna z dat a bylo zjišťováno, zda to zlepší vlastnosti modelu. Každou z 27 hodnot jsme otestovali a při odstranění žádné z nich nedošlo k výraznému zlepšení. Při odstranění dokonce všech 27 pozorování se model také výrazně nevylepší. Proto tato pozorování v modelu zanecháme i přes velká rezidua.

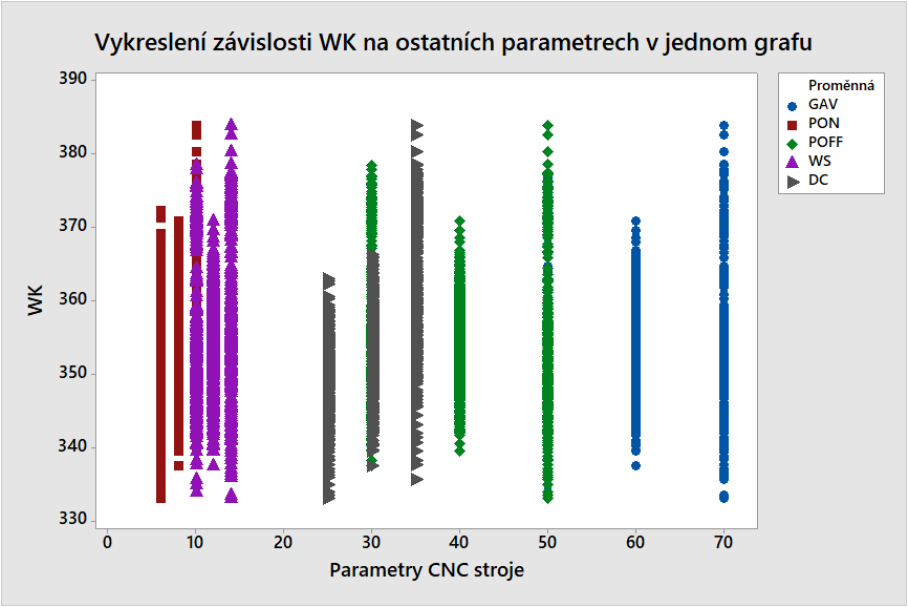
3.4 Analýza a tvorba modelu pro datový soubor SECOND PART DATA

Tento datový soubor obsahuje 1155 položek. Jde o zkoumání konce řezu. Na grafu níže (Graf 15.) jsou data v bodovém grafu, který znázorněňuje závislost šířky mezery řezu na našich pěti parametrech CNC stroje.



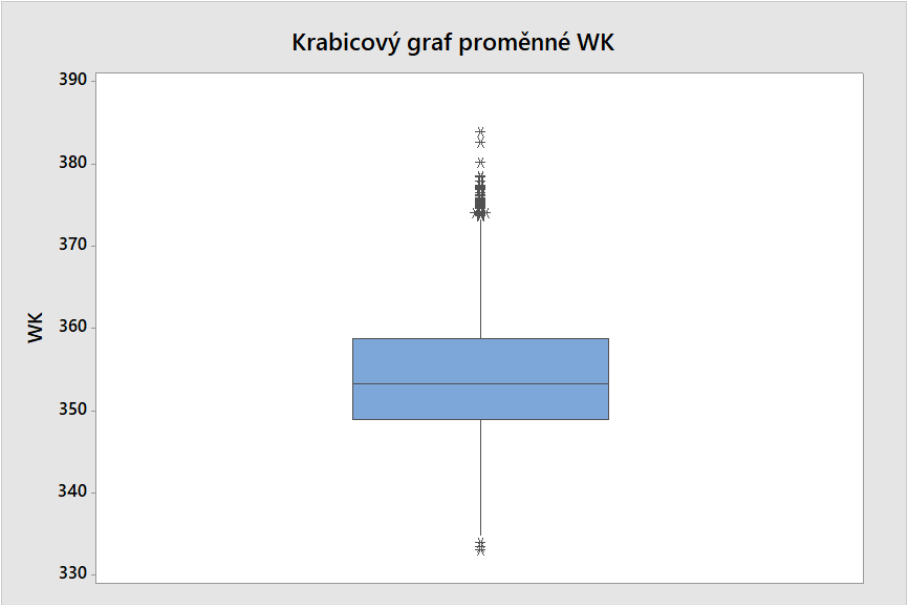
Graf 15: Závislost WK na ostatních parametrech.[vlastní zpracování]

Vykreslíme opět bodový graf (Graf 16.), kde budou všechny data v jednom grafu.



Graf 16: Závislosti WK a parametrů v jednom grafu.[vlastní zpracování]

Vykreslíme krabicový graf (Graf 17.) pouze pro proměnnou WK, opět vidíme, že mohou existovat nějaká odlehlá pozorování.



Graf 17: Krabicový graf proměnné WK.[vlastní zpracování]

Pro jistotu provedeme test na odlehlé hodnoty proměnné WK. Nulová hypotéza je, že všechny hodnoty dat jsou v pořádku a alternativní, že menší nebo větší hodnoty dat jsou odlehlé hodnoty.

Tabulka 22: Test odlehlých hodnot proměnné WK.[vlastní zpracování]

Variable	N	Mean	StDev	Min	Max	G	P
WK	1155	354,60	8,70	333,01	383,90	3,37	0,849

Jak můžeme vidět z tabulky (Tabulka 22.), $P - hodnota$ je vyšší něž hladina významnosti α, takže můžeme přijmout nulovou hypotézu. Podle tohoto testu proměnná nemá žádné odlehlé hodnoty na hladině významnosti $\alpha = 0,05$.

Opět vytvoříme vícerozměrný regresní model s využitím Minitabu. Výsledný model (Obrázek 13.) byl nalezen po 10 krocích.

```
WK   =  556,1 - 0,067 GAV - 16,86 PON + 0,894 POFF + 9,93 WS
        - 16,28 DC + 0,686 PON*PON- 0,01475 POFF*POFF - 0,643 WS*WS
        + 0,1781 DC*DC - 0,01129 GAV*POFF + 0,01870 GAV*DC
        + 0,2844 PON*DC + 0,02799 POFF*DC + 0,2020 WS*DC
```

Obrázek 13: Výsledný vztah popisující model - SECOND.[vlastní zpracování]

Tento model vystihuje náš problém ze 70,78% (Tabulka 23.).

Tabulka 23: Hodnoty shrnující vhodnost modelu.[vlastní zpracování]

S	R-sq	R-sq(adj)	R-sq(pred)
4,72976	70,78%	70,42%	70,06%

Koeficienty tohoto modelu jsou (Tabulka 24.):

Tabulka 24: Koeficienty modelu.[vlastní zpracování]

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	556,1	19,2	29,02	0,000	
GAV	-0,067	0,145	-0,46	0,643	59,50
PON	-16,86	2,07	-8,13	0,000	484,28
POFF	0,894	0,432	2,07	0,039	524,78
WS	9,93	3,04	3,27	0,001	1037,74
DC	-16,28	1,26	-12,94	0,000	1114,24
PON*PON	0,686	0,124	5,53	0,000	445,92
POFF*POFF	-0,01475	0,00496	-2,98	0,003	445,92
WS*WS	-0,643	0,124	-5,19	0,000	999,39
DC*DC	0,1781	0,0198	8,98	0,000	999,39
GAV*POFF	-0,01129	0,00200	-5,65	0,000	59,50
GAV*DC	0,01870	0,00400	4,68	0,000	82,00
PON*DC	0,2844	0,0200	14,23	0,000	59,50
POFF*DC	0,02799	0,00400	7,00	0,000	59,50
WS*DC	0,2020	0,0200	10,11	0,000	82,00

Všechny tyto koeficienty jsou statisticky významné, kromě koeficientu u proměnné GAV, ten však v modelu požadujeme. Ostatní $P - hodnoty$ jsou nižší než je zvolená hladina významnosti α.

Při tvorbě modelu bylo odhaleno 57 pozorování s vysokou hodnotou rezidua. Každé z těchto pozorování bylo postupně odebráno z modelu, ten se však odebráním významně nezlepšil.

4 ZÁVĚREČNÉ ZHODNOCENÍ MODELŮ

V této části se pokusíme modely jakýmsi způsobem zhodnotit a porovnat modely datových souborů FIRST DATA a SECOND DATA. Chceme totiž zjistit, zda se šířka mezery řezu v první části řezu odlišuje od druhé části řezu.

V tabulce (Tabulka 25.) můžeme vidět vypsané všechny regresní koeficienty modelů vytvořených pro všechny datové soubory.

Tabulka 25: Koeficienty všech modelů.[vlastní zpracování]

Term	MEAN DATA Coef	TOTAL DATA Coef	FIRST DATA Coef	SECOND DATA Coef
Constant	477,8	523,6	471	556,1
GAV	0,26	0,243	0,51	-0,067
PON	-13,64	-17,3	-17,44	-16,86
POFF	-0,249	1,481	3,417	0,894
WS	14,31	8,8	0,75	9,93
DC	-12,62	-14,84	-10,94	-16,28
PON*PON	0,548	0,777	0,935	0,686
POFF*POFF	–	-0,02115	-0,03829	-0,01475
WS*WS	-0,735	-0,505	–	-0,643
DC*DC	0,1137	0,1505	0,0774	0,1781
GAV*POFF	-0,01275	-0,01275	-0,01616	-0,01129
GAV*WS	-0,0398	-0,0397	-0,0942	–
GAV*DC	0,0262	0,0262	0,04372	0,0187
PON*DC	0,2547	0,2547	0,1854	0,2844
POFF*DC	0,0279	0,02791	-0,0325	0,02799
POFF*WS	–	–	0,02773	–
WS*DC	0,2007	0,2008	0,1981	0,202

Žádný z modelů neobsahuje stejné báze. Z tohoto důvodu nemůžeme modely porovnávat pomocí testů shody dvou modelů. Když se však i přesto podíváme na koeficienty všech modelů, tak až na proměnnou WS jsou hodnoty koeficientů u všech modelů velice podobné.

Modely můžeme porovnat pomocí střední kvadratické chyby, popřípadě koeficientu determinace, ale to nám pouze říká, který z modelů je vhodnější. Naše modely jsou však tvořeny vždy z jiného datového souboru a proto by toto porovnání nebylo vhodné pro zjišťování, zda se šířka mezeru řezu na začátku řezu chová jinak než na zbytku řezu.

Díky koeficientu determinace ale víme, že model pro MEAN DATA nejlépe vystihuje náš problém. Tento datový soubor byl však upravován, hodnoty byly zprůměrovány,

což zkresluje skutečnost. Model pro datový soubor TOTAL DATA mnohem lépe zachycuje reálnou situaci. Tento soubor je také větší než datový soubor MEAN DATA.

Tabulka 26: Shrnující hodnoty všech modelů.[vlastní zpracování]

	MEAN DATA	TOTAL DATA	FIRST DATA	SECOND DATA
R-sq	98,91%	69,36%	71,45%	70,78%
R-sq(adj)	97,90%	69,08%	70,56%	70,42%
R-sq(pred)	95,70%	68,78%	69,51%	70,06%
S	1,13312	4,8901	4,89551	4,72976
velikost datového souboru	33	1650	495	1155

Jak je vidět z tabulky (Tabulka 26.), všechny modely popisují data z více než 65%. Modely můžeme také využít pro predikci hodnot, jejich predikční procento je také vysoké a přesahuje 65% ve všech případech. Pokud to bude třeba, je možné do modelu zadat nastavení CNC stroje a predikovat tak šířku mezery řezu, která vznikne s tímto nastavením. Modely pro TOTAL, FIRST a SECOND DATA mají hodnoty koeficientů determinace a středních kvadratických chyb velice vyrovnané.

4.1 Porovnání modelů pro FIRST a SECOND DATA

Jedním z cílů je zjistit, zda se šířka mezery řezu na začátku řezu chová jinak, než na zbylé části řezu. Proto vezmeme modely vytvořené pro obě části řezu, tu počáteční a zbylou, jde o datové soubory FIRST a SECOND DATA. Do těchto modelů vložíme datový soubor MEAN DATA, tedy naměřená data plánovaného experimentu, celkem 33 kombinací hodnot parametrů CNC stroje a necháme modelem vypočítat hodnotu šířky mezery řezu pro každou z kombinací nastavení CNC stroje. Vypočtené šířky mezery řezu od sebe odečteme. Pro tyto rozdíly vytvoříme model, kde zahrneme všech 5 parametrů nastavení CNC stroje. Na tomto modelu poté otestujeme hypotézu, zda jsou regresní koeficienty všechny nulové a průměr rozdílů také. Pokud by to tak bylo, modely se chovají stejně.

Hodnoty WK pro každé z 33 kombinací nastavení vyšly pro model FIRST znatelně nižší než pro model SECOND. Model FIRST predikuje tedy nižší šířky mezery řezu pro stejná nastavení než model SECOND a to vždy.

Pro rozdíl šířek mezer řezů byl nalezen jednoduchý model obsahující pouze parametry CNC stroje a ne jejich kombinace s koeficienty, které jsou uvedeny v tabulce níže (Tabulka 27.). Při testování celého modelu jsme získali $P - hodnotu = 0,072$ a hypotézu nulovosti všech koeficientů a nulovéhé průměru nemůžeme zamítnout, protože je

$P - hodnota$ vyšší než zvolená $\alpha = 0,05$. Modely se tedy chovají stejně, pouze šířka mezery řezu je na začátku menší.

Tabulka 27: Koeficienty modelu pro rozdíl WK.[vlastní zpracování]

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0,49	6,75	-0,07	0,942	
GAV	0,0024	0,0569	0,04	0,967	1,00
PON	0,434	0,285	1,52	0,139	1,00
POFF	-0,0502	0,0569	-0,88	0,386	1,00
WS	-0,817	0,285	-2,87	0,008	1,00
DC	-0,050	0,114	-0,44	0,664	1,00

Pokud se však podíváme na testy jednotlivých koeficientů (Tabulka 27.), hypotézu o nulovosti proměnné WS musíme zamítnout.

To znamená, že podle vytvořených modelů je šířka mezery řezu na začátku obvykle menší než na zbylé části řezu. CNC stroj tedy zpočátku vytváří menší šířku mezery řezu, tvoří se tedy menší odpad a součástka má v tomto místě větší rozměr.

4.2 Optimální nastavení parametrů stroje

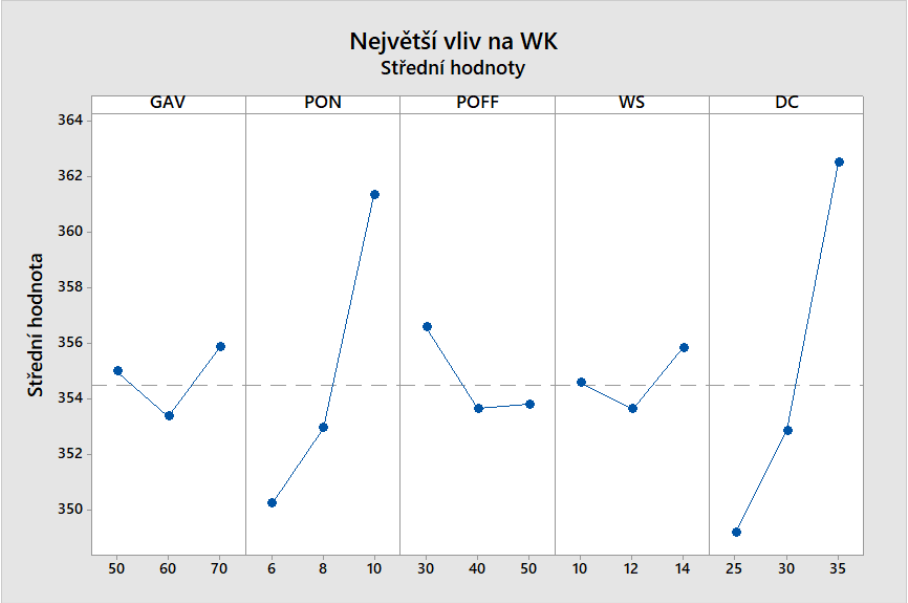
Dalším cílem práce je nalézt optimální parametry nastavení CNC stroje při procesu WEDM, které zajistí nejnižší hodnoty WK, tedy šířky mezery řezu. Takové optimální nastavení budeme hledat pomocí modelů vytvořených na datových souborech MEAN a TOTAL DATA.

Oba tyto modely predikují nejmenší hodnoty šířky mezery řezu pro nastavení v tabulce níže (Tabulka 28.). V tabulce je uvedena i konkrétní predikovaná hodnota šířky mezery řezu pro dané nastavení.

Tabulka 28: Optimální nastavení CNC stroje.[vlastní zpracování]

	GAV	PON	POFF	WS	DC	MIN WK
MEAN	70	6	50	14	25	304,65
TOTAL	70	6	50	14	25	338,93

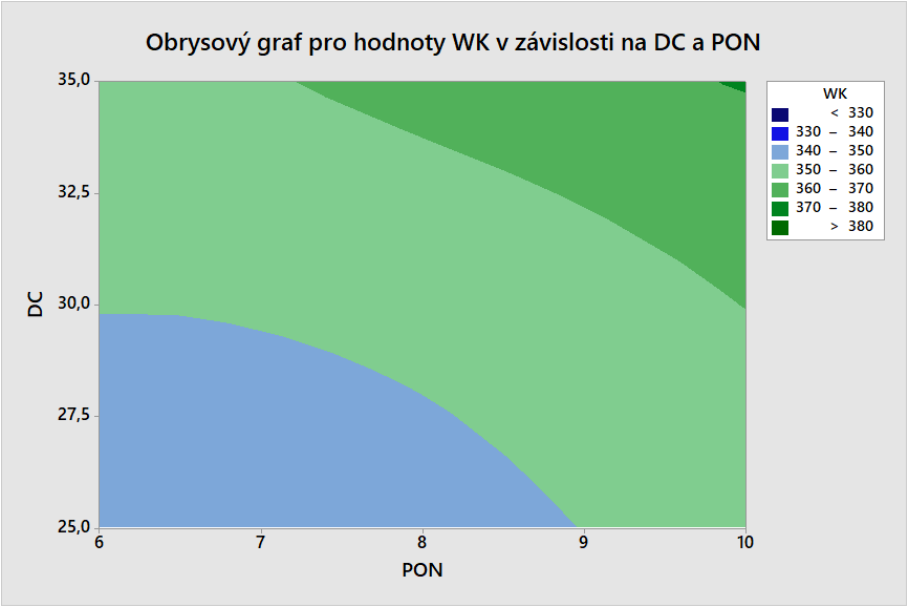
Pomocí grafu největšího vlivu na šířku mezery řezu (Graf 18.) zjistíme, které parametry nastavení CNC stroje ovlivňují šířku mezery řezu nejvíce.



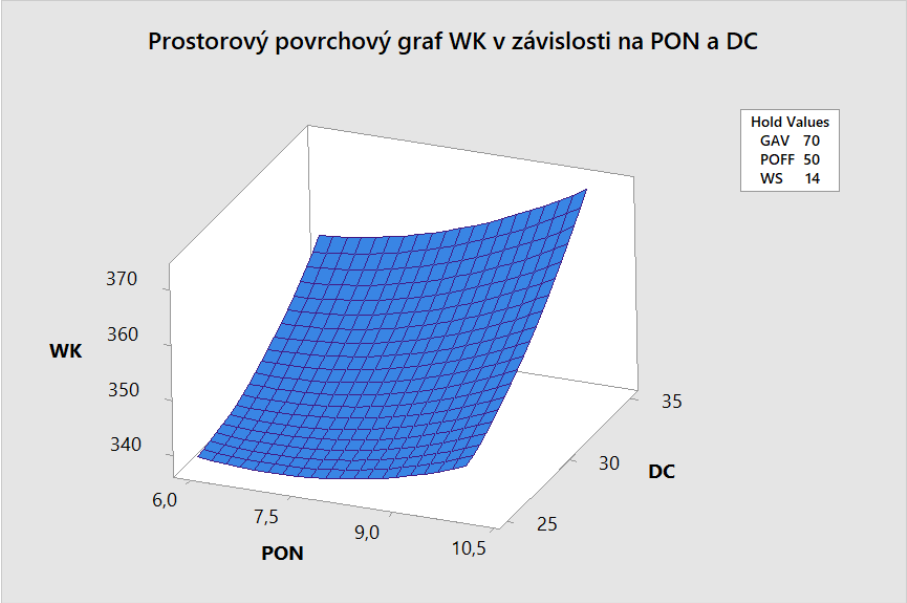
Graf 18: Vliv nastavení CNC stroje na WK.[vlastní zpracování]

Tento graf byl získán z datového souboru TOTAL DATA, ale výpočet byl proveden pro všechny datové soubory a u všech je výsledek stejný. Jak je z grafu (Graf 18.) vidět, parametry PON a DC jsou nejvlivnějšími parametry. Ty nejvíce ovlivní šířku mezery řezu. Bylo by tedy dobré zaměřit se na nejlepší nastavení těchto dvou parametrů.

Obrysový a povrchový graf (Grafy 19., 20.) znázorňují různé kombinace nastavení parametrů DC a PON a hodnotu šířky mezery řezu WK při takovém nastavení. Z těchto grafů se dá také určit nejoptimálnější nastavení.



Graf 19: Obrysový graf kombinací nastavení DC a PON.[vlastní zpracování]



Graf 20: Povrchový graf kombinací nastavení DC a PON.[vlastní zpracování]

Optimální nastavení je opět stejné jako při předchozím zjištění (Tabulka 28.). Hodnoty $DC = 25$ a $PON = 6$ nastavené na CNC stroji by měly zajistit nejměnší šířku mezery řezu obráběné součástky při procesu WEDM.

ZÁVĚR

Cílem této práce bylo popsat teoretické základy pro regresní analýzu, tvorbu vícerozměrného regresního modelu a nelineární regresní modely. Dále uvést reálný proces, na kterém se bude aplikovat regresní analýza. Následně je cílem vyhodnotit výsledky získané regresní analýzou.

V této diplomové práci je uveden souhrn pro pochopení regresní analýzy. První kapitola obsahuje teoretické základy nad rámec použitých v dalších kapitolách. Jsou zde uvedeny i metody, které nebyly dále využity. Teoretická část slouží k jakémusi souhrnnému ucelení poznatků o vícerozměrné regresní analýze.

Druhá kapitola je věnována elektroerozivním metodám obrábění, technologii WEDM a titanu a jeho slitinám. Následně je nastíněno získání dat pro tuto práci. Jedná se o měření šířky mezery řezu při různých nastaveních CNC stroje, který provádí proces WEDM na titanových slitinách Ti-6Al-4V.

Třetí kapitola popisuje tvorbu jednotlivých vícerozměrných regresních modelů. Byly vytvořeny celkem čtyři datové soubory, které popisují vždy určitý případ. MEAN DATA soubor popisuje zjednodušený problém, TOTAL DATA bere v úvahu nejvíce dat a je tak nejblíže k realitě, soubor FIRST DATA popisuje začátek řezu CNC stroje a SECOND DATA zbylou část řezu. Pro každý datový soubor se podařilo nalézt model popisující závislost šířky mezery řezu na nastavení CNC stroje s přesností větší než 65%. Tvorba vícerozměrného regresního modelu je popsána podrobně na prvním souboru a dále je uvedena pouze zkráceně. K této regresní analýze bylo využito softwaru Minitab.

Poslední, čtvrtá kapitola obsahuje závěrečné zhodnocení modelů a získaných výsledků. Nalezli jsme závislost mezi šířkou mezery řezu a nastavením CNC stroje při technologii WEDM. Díky této závislosti jsme zjistili, že šířka mezery řezu je jiná při počáteční fázi řezu než na zbytku řezu. Toto zjištění může být využito pro úpravu procesu WEDM a výroba tak může být zpřesněna. Určili jsme parametry procesu WEDM, tedy CNC stroje, které mají největší vliv na šířku mezery řezu. Zjistili jsme také optimální nastavení CNC stroje, které zajistí nejmenší šířku mezery řezu a tak nejefektivnější a nejpřesnější výrobu součástek z titanových slitin.

Stanovených cílů této diplomové práce bylo tedy dosaženo. Práce může dále posloužit i jako metodický nástroj pro návrhy vícerozměrných regresních modelů obdobného typu a k ucelení poznatků týkajících se regresní analýzy.

SEZNAM POUŽITÝCH ZDROJŮ

[1] ANDĚL, Jiří. *Základy matematické statistiky*. 2., opr. vyd. Praha: Matfyzpress, 2007. ISBN 80-737-8001-1.

[2] ZVÁRA, Karel. *Regrese*. Praha: Matfyzpress, 2008. ISBN 978-80-7378-041-8.

[3] REKTORYS, Karel. *Přehled užité matematiky.* 2., opr. vyd. Praha: Nakladatelství technické literatury, 1968, 1136 s. Česká matice technická (SNTL).

[4] ZVÁRA, Karel. *Regresní analýza.* Praha: Academia, 1989. ISBN 80-200-0125-5.

[5] SHAPIRO, S. S. a M. B. WILK. *An Analysis of Variance Test for Normality (Complete Samples)*. DOI: 10.2307/2333709. ISBN 10.2307/2333709. Dostupné také z: http://www.jstor.org/stable/2333709?origin=crossref

[6] BELSLEY, David A., Edwin KUH a Roy E. WELSCH. *Regression diagnostics: identifying influential data and sources of collinearity*. New York: Wiley, c1980. ISBN 9780471058564.

[7] BOX, G.; COX, D. *An analysis of transformations*. Journal of the Royal Statistical Society. Series B (Methodological), ročník 26, č. 2, 1964: s. 211–252, ISSN 0035-9246.

[8] HINDLS, Richard. *Statistika pro ekonomy.* 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[9] KROPÁČ, Jiří. *Statistika B: jednorozměrné a dvourozměrné datové soubory, regresní analýza, časové řady.* 2., dopl. vyd. Brno: Vysoké učení technické v Brně, Fakulta podnikatelská, 2009. ISBN 978-80-214-3295-6.

[10] VOJTĚCH, Dalibor. *Kovové materiály*. 1. vyd. Praha: VŠCHT, 2006, 185 s. ISBN 80- 708-0600-1.

[11] BARTHELMY, David. *Mineral Species containing Titanium*. In: Mineralogy database [online]. c2018 [cit. 2018-04-13]. Dostupné z: http://webmineral.com/

[12] BÁRTKOVÁ, D. *Vysokocyklová únava titanové slitiny Ti6Al4V.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2013. 97s. Vedoucí diplomové práce prof. Ing. Stanislav Věchet, CSc..

[13] Pevnost. In: Letecký ústav: Fakulta strojního inženýrství. [online]. [cit. 2018-04-10]. Dostupné z: http://lu.fme.vutbr.cz/ucebnice/opory/stress.php

[14] LUTJERING, G. a J. WILLIAMS. *Titanium*. New York: Springer, c2003, x, 379 p. Engineering materials. ISBN 35-404-2990-5.

[15] VEIGA, C., DAVIM , J.P. a LOUREIRO, A.J.R.. *Properties and aplications of titanium alloys: A brief review*. In: Reviews on advanced materials science. Coimbra: Reviews on Advanced Materials Science, 2012, 133 - 148. ISBN 1605-8127ISSN 1605-8127. Dostupné z: http:www.ipme.ruejournalsRAMSno_2321205_23212_veiga.pdf

[16] HENRY, Scott D, Kathleen S DRAGOLICH a DIMATTEO. *Fatigue data book: light structural alloys*. Materials Park, OH: ASM International, c1995, viii, 397 p. ISBN 08-717- 0507-9.

[17] MAŇKOVÁ, Ildikó. *Progresívne technológie.* 1. vyd. Košice: Vienala, 2000. ISBN 80-709-9430-4.

[18] LIEMERT, Gaston, František DRÁBEK, Josef ONDRA a Ivan VAVŘÍK. *Obrábění*. 1. vyd. Praha: SNTL. 1974, 351 s.

[19] MAHTO, Dalgobind & SINGH, Narinder. (2016). *Experimental Study of Process Parameters through Dissimilar Form of Electrodes in EDM Machining*. 10.13140/RG.2.1.2558.2320.

[20] MIČIETOVÁ, Anna. *Nekonvenčné metódy obrábania*. vyd. Žilina: EDIS, 2001. ISBN 80-7100-853-2.

[21] PARASHAR, Vishal, A. REHMAN, J.L. BHAGORIA a Y.M. PURI. Kerfs width analysis for wire cut electro discharge machining of SS 304L using design of experiments. Indian Journal of Science and Technology. 2010, 3(4). ISSN 0974- 6846.

[22] Divize Formy. MOTOR JIKOV GROUP [online]. [cit. 2018-04-13]. Dostupné z: http://www.motorjikov.com/spolecnosti/motor-jikov-fostron/divize-formy/

[23] MOURALOVÁ, K., KOVÁŘ, J. a SLIWKOVÁ, P. (2016). *Evaluation of S-parameters on the surface of titanium alloy Ti-6Al-4V and Al99.5 machined by WEDM*. MM Science Journal. 2016. 1537-1540. 10.17973MMSJ.2016_12_2016107.

[24] MOURALOVÁ, K.; KOVÁŘ, J.; PROKEŠ, T.; BEDNÁŘ, J.; MATOUŠEK, R.; KLAKURKOVÁ, L. *Statistical evaluation width of kerf after WEDM by Analysis of Variance*. Mendel Journal series, 2016, vol. 2016, no. 22, p. 301-304. ISSN: 1803-3814.

[25] MOURALOVÁ, Kateřina. *Moderní technologie drátového elektroerozivního řezání kovových slitin*. Dizertační práce. VUT, Brno, 98s., 2015.

SEZNAM POUŽITÝCH ZKRATEK

WEDM	elektroerojiskrové řezání drátovou elektrodou/ wire electrical discharge machining
EDM	elektroerozivní obrábění / electrical discharge machining
EDG	elektroerozivní broušení / electrical discharge grinding
CNC	počítačově číslicové řízení / computer numerical controlled
Ti-6Al-4V	titanová slitina Grade 5
WK	šířka mezery řezu / width of kerf
GAV	gap voltage (V)
PON	pulse on time (μs)
POFF	pulse off time (μs)
WS	wire speed (m/min)
DC	discharge current (A)

SEZNAM GRAFŮ

Graf 1: ZÁVISLOST WK NA OSTATNÍCH PARAMETRECH 50
Graf 2: ZÁVISLOST WK A PARAMETRŮ V JEDNOM GRAFU 50
Graf 3: KRABICOVÝ GRAF VŠECH PROMĚNNÝCH 51
Graf 4: KRABICOVÝ GRAF PROMĚNNÉ WK 51
Graf 5: TEST ROZDĚLENÍ REZIDUÍ PRO NÁŠ MODEL 56
Graf 6: REZIDUA V ZÁVISLOSTI NA ZÍSKANÝCH HODNOTÁCH 57
Graf 7: TEST ROZDĚLENÍ REZIDUÍ PRO NÁŠ MODEL 62
Graf 8: REZIDUA V ZÁVISLOSTI NA ZÍSKANÝCH HODNOTÁCH 62
Graf 9: ZÁVISLOST WK NA OSTATNÍCH PARAMETRECH 63
Graf 10: ZÁVISLOST WK A PARAMETRŮ V JEDNOM GRAFU 63
Graf 11: KRABICOVÝ GRAF PROMĚNNÉ WK 64
Graf 12: ZÁVISLOST WK NA OSTATNÍCH PARAMETRECH 66
Graf 13: ZÁVISLOST WK A PARAMETRŮ V JEDNOM GRAFU 66
Graf 14: KRABICOVÝ GRAF PROMĚNNÉ WK 67
Graf 15: ZÁVISLOST WK NA OSTATNÍCH PARAMETRECH 69
Graf 16: ZÁVISLOST WK A PARAMETRŮ V JEDNOM GRAFU 69
Graf 17: KRABICOVÝ GRAF PROMĚNNÉ WK 70
Graf 18: VLIV NASTAVENÍ CNC STROJE NA WK 75
Graf 19: OBRYSOVÝ GRAF KOMBINACÍ NASTAVENÍ DC A PON 75
Graf 20: POVRCHOVÝ GRAF KOMBINACÍ NASTAVENÍ DC A PON 76

SEZNAM OBRÁZKŮ

Obrázek 1: TITAN 43
Obrázek 2: PROCES EDM 46
Obrázek 3: PROCES WEDM 47
Obrázek 4: 50 MĚŘENÍ ŠÍŘKY MEZERY JEDNOHO ŘEZU 47
Obrázek 5: CNC STROJ MAKINO EU64 48
Obrázek 6: ROZDĚLENÍ MĚŘENÍ ŠÍŘEK MEZERY JEDNOHO ŘEZU 49
Obrázek 7: PROMĚNNÉ A JEJICH KOMBINACE ZAHRNUTÉ V MODELU 52
Obrázek 8: KOEFICIENT DETERMINACE MODELU 55
Obrázek 9: VÝSLEDNÁ ROVNICE POPISUJÍCÍ MODEL 56
Obrázek 10: VÝSLEDNÝ VZTAH POPISUJÍCÍ MODEL - MEAN 61
Obrázek 11: VÝSLEDNÝ VZTAH POPISUJÍCÍ MODEL - TOTAL 65
Obrázek 12: VÝSLEDNÝ VZTAH POPISUJÍCÍ MODEL - FIRST 67
Obrázek 13: VÝSLEDNÝ VZTAH POPISUJÍCÍ MODEL - SECOND 70

SEZNAM TABULEK

Tabulka 1: VÝBĚR MODELU A SUBMODELU 25
Tabulka 2: VÝBĚR TRANSFORMACE 41
Tabulka 3: ZÁKLADNÍ FYZIKÁLNÍ VLASTNOSTI TITANU 44
Tabulka 4: MEZNÍ OBSAHY PRVKŮ VE SLOUČENINĚ Ti-6Al-4V 45
Tabulka 5: TEST ODLEHLÝCH HODNOT PROMĚNNÉ WK 52
Tabulka 6: POSTUPNÉ KROKY TVORBY MODELU 53
Tabulka 7: ANALÝZA ROZPTYLU 54
Tabulka 8: HODNOTY SHRNUJÍCÍ VHODNOST MODELU 54
Tabulka 9: KOEFICIENTY REGRESNÍHO MODELU 55
Tabulka 10: ODLEHLÁ POZOROVÁNÍ 57
Tabulka 11: KROKY TVORBY MODELU - 1. ČÁST 58
Tabulka 12: KROKY TVORBY MODELU - 2. ČÁST 59
Tabulka 13: HODNOTY SHRNUJÍCÍ VHODNOST MODELU 59
Tabulka 14: ANALÝZA ROZPTYLU 60
Tabulka 15: KOEFICIENTY MODELU 61
Tabulka 16: TEST ODLEHLÝCH HODNOT PROMĚNNÉ WK 64
Tabulka 17: KOEFICIENTY MODELU 65
Tabulka 18: HODNOTY SHRNUJÍCÍ VHODNOST MODELU 65
Tabulka 19: TEST ODLEHLÝCH HODNOT PROMĚNNÉ WK 67
Tabulka 20: KOEFICIENTY MODELU 68
Tabulka 21: HODNOTY SHRNUJÍCÍ VHODNOST MODELU 68
Tabulka 22: TEST ODLEHLÝCH HODNOT PROMĚNNÉ WK 70
Tabulka 23: HODNOTY SHRNUJÍCÍ VHODNOST MODELU 71
Tabulka 24: KOEFICIENTY MODELU 71
Tabulka 25: KOEFICIENTY VŠECH MODELŮ 72
Tabulka 26: SHRNUJÍCÍ HODNOTY VŠECH MODELŮ 73
Tabulka 27: KOEFICIENTY MODELU PRO ROZDÍL WK 74
Tabulka 28: OPTIMÁLNÍ NASTAVENÍ CNC STROJE 74