VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNíHO INŽENÝRSTVí
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

REGRESNÉ METÓDY ODHADU VYBRANÝCH CHARAKTERISTÍK TAVENÝCH SYROV V ZÁVISLOSTI NA POMERE TAVIACICH SOLÍ

REGRESSION METHODS OF ESTIMATION OF CHOSEN PROPERTIES OF PROCESSED CHEESE WITH REGARD TO THE RELATIVE AMOUNT OF DIFFERENT TERNARY MIXTURES OF SODIUM PHOSPHATES

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. BRANISLAV PETROVIČ

VEDOUCÍ PRÁCE
SUPERVISOR

doc. RNDr. JAROSLAV MICHÁLEK, CSc.

BRNO 2013

Vysoké učení technické v Brně, Fakulta strojního inženýrství

Ústav matematiky
Akademický rok: 2012/2013

ZADÁNÍ DIPLOMOVÉ PRÁCE

student(ka): Bc. Branislav Petrovič

který/která studuje v **magisterském navazujícím studijním programu**

obor: **Matematické inženýrství (3901T021)**

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:

Regresní metody odhadu vybraných charakteristik tavených sýrů v závislosti na poměru tavicích solí

v anglickém jazyce:

Regression methods of estimation of chosen properties of processed cheese with regard to the relative amount of different ternary mixtures of sodium phosphates.

Stručná charakteristika problematiky úkolu:

Při výrobě tavených sýrů závisí jejich vlastnosti zejména na použité směsi tavicích solí a na době jejich zralosti. Pro posouzení této statistické vazby byl na Technologické fakultě Univerzity Tomaše Bati ve Zlíně proveden statistický experiment, jehož výsledky umožňují sledované vazby statisticky popsat.

Cíle diplomové práce:

Cílem práce je vybrat, popsat a algoritmizovat vhodné metody regresní analýzy, které jsou pro statistickou analýzu provedeného experimentu vhodné. Zejména se zaměřitna polynomickou regresi a případně metody nelineární. Dále popsat vlastnosti uvedených regresních metod, provést regresní diagnostiku, zaměřit se na výběr regresních parametrů (např. kroková regrese, Cp statistika) a provést příslušné statistické testy. Výsledky analýz demonstrovat graficky.

Seznam odborné literatury:

[1] Brook R,J. and Arnold G.C. Applied regression analysis and experimental design. Marcel Dekker ,INC. 1985

[2] Seber G.A.F. and Wild C.J.: Nonlinear regression. John Wiley and Sons. 1988.

[3] Weiserová E., Doudová L., Galiová L., Žák L., Michálek J., Janiša R., Buňka F.: The effect of combinations of sodium phosphates in binary mixtures on selected texture parameters of processed cheese spreads. International Dairy Journal. Volume 21, Issue 12, December 2011, Pages 979-986 2010

Vedoucí diplomové práce: doc. RNDr. Jaroslav Michálek, CSc.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2012/2013.

V Brně, dne

L.S.

prof. RNDr. Josef Šlapal, CSc. Ředitel ústavu	prof. RNDr. Miroslav Doupovec, CSc., dr. h. c. Děkan fakulty

Abstrakt
Diplomová práca sa zaoberá regresnou analýzou experimentálne nameraných dát tavených syrov. Využitá je polynomiálna regresia a výber regresorov je založený na Krokovej regresii a Mallowsovej štatistike. Následný odhad strednej hodnoty je využitý k nájdeniu najlepšej zmesi taviacich solí, z hľadiska sledovaného parametra taveného syra. Analýza experimentu aj výsledky sú dobre graficky zdokumentované.

Summary
This thesis deals with regression analysis of experimentally measured data of processed cheese. There is a polynomial regression used. The choice of regressors is based on Stepwise Regression and Mallows's Statistics. The estimation of the mean value is used to find the best mixture of the emulsifying salts with regards to the observed characteristic of the processed cheese. Analysis of the experiment and its results are well documented graphically.

Klíčová slova
lineárna regresia, Kroková regresia, Mallowsova Cp štatistika

Keywords
linear regression, Stepwise Regression, Mallows's Cp Statistics

PETROVIČ, B. *Regresné metódy odhadu vybraných charakteristík tavených syrov v závislosti na pomere taviacich solí*. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2013. 68 s. Vedoucí diplomové práce doc. RNDr. Jaroslav Michálek, CSc.

Prehlasujem, že som diplomovú prácu „Regresné metódy odhadu vybraných charakteristík tavených syrov v závislosti na pomere taviacich solí“ vypracoval samostatne s použitím odbornej literatúry, uvedenej v zozname, ktorý je súčasťou tejto práce.

Bc. Branislav Petrovič

Ďakujem vedúcemu doc. RNDr. Jaroslavovi Michálkovi, CSc. za cenné rady, venovaný čas a odborný prístup pri tvorbe tejto diplomovej práce.

Bc. Branislav Petrovič

Obsah

1. Úvod

Cieľom tejto práce je pomocou regresnej analýzy vyhodnotiť namerané data v nižšie uvedenom experimente. Práca je rozdelená do šiestich kapitol.

V prvej úvodnej kapitole je odstavec 1.1 venovaný problematike tavených syrov a uskutočnenému experimentu. Text bol vytvorený s použitím článkov [11], [7] a [5]. Druhý odstavec 1.2 je venovaný zavedeniu matematického aparátu. V tomto odstavci bol výklad usporiadaný s použitím literatúr [2], [1] a [9].

Druhá kapitola pojednáva o jednoduchšom prípade, kedy matica modelu nemá plnú stĺpcovú hodnosť. Výklad pre model úplnej hodnosti je usporiadaný s použitím zdrojov [3] a [13].

Tretia kapitola hovorí o lineárnom regresnom modeli, kedy nie je splnená úplná stĺpcová hodnosť matice modelu. V tejto kapitole sa občas porovnáva model s úplnou stĺpcovou hodnosťou s modelom neúplnou stĺpcovou hodnosťou. Výklad je vytvorený s použitím zdrojov [9] a [2].

Štvrtá kapitola pojednáva o dvoch metódach, ktoré sa využívajú pri výbere regresorov v lineárnom modeli. Ďalej hovorí o diagnostike lineárneho regresného modelu. Výklad o výbere modelu vychádza z literatúr [13], [4] a [6]. Výklad o diagnostike je usporiadaný s použitím [10] a [13].

V piatej kapitole sa nachádza aplikačná časť práce. Je tu popísané vyhodnocovanie experimentu a doložené grafickou dokumentáciou. Sú porovnávané obe metódy na výber regresorov. Kapitolu uzaviera odhad zmesi taviacich solí, ktorá najlepšie spĺňa zadané kritérium.

1.1. Charakteristika problematiky

Tavené syry

Tavený syr (v angl. processed cheese) je mliečny výrobok, ktorého hlavnou zložkou je tvrdý alebo mäkký syr (alebo ich kombinácia) a príp. tvaroh, sušené mlieko alebo maslo. Výroba taveného syra spočíva v zmiešaní syrov (aj rôznej doby zrenia alebo skladovania) s tavnými soľami (v angl. emulsifying salts) a prípadne inými zložkami a za neustáleho miešania a ohrievania privedené na vhodnú konzistenciu. Pôvodnou myšlienkou výroby syra s kombináciou taviacej soli (pôvodne citrátu sodného), bolo predĺženie trvanlivosti kvôli možnosti lepšieho predaja tohto výrobku. Neskoršie začali byť používané tiež taviace soli tvorené fosfátmi alebo ich kombináciami.

Tavený syr bol objavený v roku 1911, vo Švajčiarsku, Walterom Gerberom a Fritzom Stettlerom z Gerber and Co., ktorý rozpúšťali Švajčiarský syr[1] použitím citrátu sodného ako taviacej soli na výrobu jemného a homogénneho výrobku ([7], str. 194).

Hlavná úloha taviacich solí spočíva v procese v ktorom vápnik, ktorý viaže kazeínové bielkoviny (alebo iné hydrolyzované frakcie kazeínu[2]) vytvárajúc trojdimenzionálnu štruktúru syra, je odlúčený (z kazeínového matrixu) a nahradený sodíkovými iónmi. Prostredníctvom výmeny vápnikových iónov za sodíkové ióny, nerozpustný vápnikový para-kazeín sa zmení na viac rozpustiteľný sodíkový para-kazeín, ktorý môže pôsobiť ako

[1] Švajčiarsky syr (v angl. Swiss cheese) je známejší pod názvom Ementál.
[2] Kaseín je hlavný proteín v cicavčom mlieku.

emulgátor ([11], str. 979).

Z predchádzajúceho citovaného odseku teda vyplýva, že taviace soli môžu mať vplyv na vlastnosti výsledného produktu - taveného syra. Toto potvrdzujú aj štúdie, ktoré sú uvedené vo vyššie uvedených článkoch. Výsledné vlastnosti taveného syru sú ovplyvnené ako výrobným procesom (napr. vlastnosťami miešania, dobou ohrievania, teplotou ohrievania a následného ochladenia), tak i skladovaním (napr. teplotou a dobou skladovania). Preto je možné modelovať pomocou regresnej analýzy vybrané parametre tavených syrov (obzvlášť tvrdosti) získaných experimentálne v závislosti na pridaných taviacich soliach.

Experiment

Experiment bol uskutočnený na Fakulte technologickej Univerzity Tomáše Bati v Zlíne. Zmes taviacej soli bola zložená z troch zložiek fosforečnanových taviacich solí:

- Na_2HPO_4 - hydrogenfosforečnan sodný (v angl. disodium hydrogen phosphate)
- $Na_4P_2O_7$ - difosforečnan sodný (v angl. tetrasodium diphosphate)
- $(NaPO_3)_m$ - polyfosforečnan sodný (v angl. sodium salt of polyphosphate)

Každá zložka zmesi bola dávkovaná po 10 % zo zmesi. Vytvorením všetkých takýchto kombinácii vynikne 66 rôznych zmesí $\left(\binom{12}{10} = 66\right)$. Pre každé meranie boli získané dve hodnoty, preto celkový počet meraní je 132 pre jednu sledovanú vlastnosť, dobu zrenia a dobu skladovania.

Boli sledované tri vlastnosti tavených syrov: tvrdosť, kohezivita a relatívna lepivosť. Pre každú vlastnosť bolo nameraných 132 hodnôt po uplynutí dňoviatich rôznych dôb. Najpr sa tavené syry nechali zrieť 2, 4 alebo 8 týždňov a po tejto dobe boli skladované 2, 9 alebo 30 dní. Potom sa uskutočnili jednotlivé merania troch vlastností (každá vlastnosť vždy na osobitnom prvku).

Na výrobu taveného syra pre tento experiment boli použité suroviny: Eidamská tehla, maslo, voda a hore uvedená zmes taviacich solí (3 % z celkovej hmotnosti). Pre polyfosforečnan sodný v zmesi bolo m v rozsahu $15 - 20$. Eidamský syr s maslom boli nakrájané na menšie kusky a zohrievané a rozmieľané počas doby 1 min približne na otáčkach 4000 ot/min. Zmes taviacich solí a voda boli pridané pri otáčkach 2000 ot/min. Ohrev trval 10-12 min a počas jednej minuty bola teplota udržiavaná na $90°C$. Následne bola tavenina zabalená do malých foriem a ochladzovaná na teplotu $6°C$ počas 2 hodin. Ďalej sa výrobok nechal zrieť v miestnosti na to určenej pri teplote 9-10$°C$ (po dobu 2, 4 alebo 8 týždňov). Následné skladovanie prebiehalo pri teplote $6°C$ po dobu 2, 9 alebo 30 dní.

Úlohou tejto práce je štatistický rozbor na tvrdosti taveného syru. Preto bude uvedená definícia tvrdosti a technické prevedenie. Valcová tyč s priemerom 20mm bola pri rýchlosti 2 mm/s vtlačovaná do taveného syra do hĺbky 10 mm. Mierou tvrdosti je tu maximálna sila, ktorá bola potrebná pri prieniku do syra.

Podrobnejšie informácie o uvedenom experimetne sú uvedené v článku [5].

1.2. Základné pojmy a označenia

V tejto práci budeme pri výklade matematickej teórie predpokladať základné znalosti z kurzov pravdepodovnosti a štatistiky. V tomto odstavci budú uvedené niektoré matema-

tické pojmy, ktoré budú použité v následnom teoretickom výklade. Ďalej odstavec bude doplnený matematickými vetami z teórie matic a vektorov. Okrem miernych zmien, ktoré budú detailne uvedené, sa autor drží označenia zavedeného Andělom v literatúre [3].

Priestor všetkých elementárnych javov bude označený symbolom Ω. Nech je na priestore Ω nejaká σ-algebra $\mathcal{A}$ podmnožin Ω. Ak P je nejaká pravdepodobnostná miera na σ-algebre $\mathcal{A}$, potom trojica $(\Omega, \mathcal{A}, \mathsf{P})$ bude označovať pravdepodobnostný priestor. Z teórie pravdepodobnosti bude uvedená známa Bonferroniho nerovnosť, ktorá znie nasledovne: Pre náhodné javy $A_1, \ldots, A_n$ platí vzťah

$$\mathsf{P}(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} \mathsf{P}(A_i). \tag{1.1}$$

Zobrazenie $X : \Omega \to \mathbb{R}$ bude nazývané *náhodnou veličinou* v $(\Omega, \mathcal{A}, \mathsf{P})$ ak platí

$$\{\omega \in \Omega \mid X(\omega) \leq a\} \in \mathcal{A} \quad \text{pre} \quad \forall a \in \mathbb{R}.$$

Náhodné veličiny a vektory budú označené veľkými písmenami z konca latinky. Hustota rozdelenia náhodnej veličiny bude označená f a distribučná funkcia bude označená F. Náhodný vektor oproti náhodnej veličine bude označený tučne. Strednú hodnotu náhodnej veličiny X bude označená $\mathsf{E}X$ a rozptyl $\mathsf{D}X$. Stredná hodnota náhodného vektora $\boldsymbol{X}$ bude označená $\mathsf{E}\boldsymbol{X}$ a variančná matica $\mathsf{Var}\boldsymbol{X}$. Variančná matica je symetrická a pozitívne semidefinitná. Kovariancia dvoch náhodných vektorov alebo veličin $\boldsymbol{X}$ a $\boldsymbol{Y}$ bude označená $\mathsf{Cov}(\boldsymbol{X}, \boldsymbol{Y})$. Následne budú uvedené vzťahy, ktoré platia pre variančnú a kovariančnú maticu

$$\mathsf{Var}\boldsymbol{X} = \mathsf{E}\boldsymbol{X}\boldsymbol{X}' - (\mathsf{E}\boldsymbol{X})(\mathsf{E}\boldsymbol{X})', \quad \mathsf{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathsf{E}\boldsymbol{X}\boldsymbol{Y}' - (\mathsf{E}\boldsymbol{X})(\mathsf{E}\boldsymbol{Y})'.$$

Ďalej budú uvedené základné pravdepodobnostné rozdelenia, ktoré budú použité. Normálne rozdelenie so strednou hodnotou μ a rozptylom σ^2 bude označené $\mathsf{N}(\mu, \sigma^2)$. Formálne sem patrí aj singulárne normálne rozdelenie $\mathsf{N}(\mu, 0)$. n-rozmerné normálne rozdelenie so strednou hodnotou $\boldsymbol{\mu}$ a variančnou maticou $\sigma^2\boldsymbol{V}$ bude označené $\mathsf{N}_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{V})$. Ďalej bude uvedené označovanie pre odvodené rozdelenia z normálneho rozdelenia ako Studentovo rozdelenie s n stupňami voľnosti $\mathsf{t}(n)$, Chí-kvadrát rozdelenie s n stupňami voľnosti $\chi^2(n)$ a Fisherovo-Snedecorovo rozdelenie s n a k stupňami voľnosti $\mathsf{F}(n, k)$. Následne budú zavedené označenia jednotlivých kvantilových funkcií. Keďže všetky tieto rozdelenia majú rastúcu (pre $\chi^2(n)$ a $\mathsf{F}(n, k)$ uvažované na intervale $(0, \infty)$) a spojitú distribučnu funkciu F, bude postačovať definícia kvantilovej funkcie ako inverznej funkcie F^{-1} k distribučnej funkcii. Napríklad hodnota kvantilovej funkcie $F^{-1}(\alpha)$ (pre $0 < \alpha < 1$) a prípadne príslušné stupne voľnosti budú označené nasledovne: u_α kvantil normovaného normálneho rozdelenia, $\mathsf{t}_\alpha(n)$ kvantil studentovho rozdelenia, $\chi^2_\alpha(n)$ kvantil chí-kvadrát rozdelenia, $\mathsf{F}_\alpha(n, k)$ kvantil Fisherovho-Snedecorovho rozdelenia.

Skutočnosť, že náhodná veličina X má normálne rozdelenie $\mathsf{N}(\mu, \sigma^2)$ bude jednoducho označená zápisom $X \sim \mathsf{N}(\mu, \sigma^2)$. Nech $\boldsymbol{X} = (X_1, \ldots, X_n)'$ je náhodný vektor, $\boldsymbol{V}$ pozitívne semidefinitná matica a $\boldsymbol{\mu}$ je daný vektor. Ak pre ľubovoľný vektor $\boldsymbol{c} \in \mathbb{R}^n$ platí

$$\boldsymbol{c}'\boldsymbol{X} \sim \mathsf{N}(\boldsymbol{c}'\boldsymbol{\mu}, \sigma^2\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c}), \tag{1.2}$$

potom sa povie, že $\boldsymbol{X}$ má *n-rozmerné normálne rozdelenie* $\mathsf{N}_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{V})$ (zápisom $\boldsymbol{X} \sim \mathsf{N}_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{V})$). V prípade, že matica $\boldsymbol{V}$ je singulárna, sa pridáva prívlastok singulárne

rozdelenie. Pre vektor $\boldsymbol{a} \in \mathbb{R}^n$, maticu $\boldsymbol{B}$ typu $n \times m$ a náhodný vektor $\boldsymbol{X} \sim \mathsf{N}_m(\boldsymbol{\mu}, \sigma^2 \boldsymbol{V})$ platí

$$\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X} \sim \mathsf{N}_n(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \sigma^2 \boldsymbol{B}\boldsymbol{V}\boldsymbol{B}'), \tag{1.3}$$

ako ukázal napríklad Anděl v literatúre ([2], str. 75).

Nech je uvažované $\boldsymbol{X} \sim \mathsf{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{V})$, kde $\mathrm{h}(\boldsymbol{X}) = r \geq 1$. Ďalej nech $\boldsymbol{B}_{n \times r}$ je matica spĺňajúca $\boldsymbol{B}\boldsymbol{B}' = \boldsymbol{V}$. Potom pre náhodný vektor $\boldsymbol{Z} \sim \mathsf{N}_r(\boldsymbol{0}, \boldsymbol{I})$ platí, že rozdelenie vektora $\boldsymbol{\mu} + \sigma \boldsymbol{B}\boldsymbol{Z}$ je rovnaké ako náhodného vektora $\boldsymbol{X}$ (ako ukázal napríklad Anděl v literatúre ([2], str. 76)).

Následne budú zavedené výberové charakteristiky pre náhodný výber $X_1, \ldots, X_n$ rozsahu n. Výberový priemer označený $\bar{X}$ a výberový rozptyl označený S_X^2 budú definované vzťahmi

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} (\sum_{i=1}^{n} X_i^2 - n\bar{X}^2).$$

Nech je uvažovaný náhodný výber $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ rozsahu n z dvojrozmerného rozdelenia. Budú uvažované charakteristiky výberový priemer $\bar{X}$ a výberový rozptyl S_X^2 náhodného výberu $X_1, \ldots, X_n$. Obdobne sú uvažované charakteristiky $\bar{Y}$ a S_Y^2 pre náhodný výber $Y_1, \ldots, Y_n$. Potom je charakteristika výberová kovariancia označená S_{XY} a definovaná vzťahom

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1} (\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}).$$

A výberový korelačný koeficient označený r_{XY} je definovaný vzťahom $r_{XY} = S_{XY} / \sqrt{S_X^2 S_Y^2}$, ak $S_X^2 S_Y^2 \neq 0$.

Nech je uvažovaná postupnosť náhodných veličín $X_1, X_2, \ldots$ a náhodná veličina X. Nech sú tieto veličiny uvažované na rovnakom pravdepodobnostnom priestore $(\Omega, \mathcal{A}, \mathsf{P})$. Ďalej nech X_n má distribučnú funkciu $F_n(x)$ a X má distribučnú funkciu $F(x)$. Povie sa, že X_n konverguje k X *v distribucii*, ak $F_n(x)$ konverguje k $F(x)$ v každom bode x, v ktorom je $F(x)$ spojitá. Rovnako je možné povedať, X_n má asymptoticky rozdelenie náhodnej veličiny X.

Niekedy súčet nezávislých a rovnako rozdelených náhodných veličín môže mať rozdelenie asymptoticky normálne. O tom pojednáva nasledujúca centrálna limitná veta.

Veta 1.1. (**Lévyho-Lindebergova**) *Nech $X_1, X_2, \ldots$ sú nezávislé náhodné veličiny s rovnakým rozdelením, ktoré má strednú hodnotu μ a konečný rozptyl σ^2. Potom*

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right) \text{ konverguje v distribúcii k rozdeleniu } \mathsf{N}(0, \sigma^2). \tag{1.4}$$

Dôkaz. Viď Anděl ([2], str. 184).

Delta metóda

Nech $\hat{\theta}_n$ je štatistika, ktorá je závislá na rozsahu výberu n a má asymptoticky normálne rozdelenie $\mathsf{N}(\theta, \sigma^2/n)$. Ako bolo uvedené vyššie, platí

$$\sqrt{n} \left(\hat{\theta}_n - \theta \right) \text{ konverguje v distribúcii k } \mathsf{N}(0, \sigma^2).$$

Nech je uvažovaná funkcia g, ktorá je v bode θ najmenej dvakrát diferencovateľná. Otázka je, aké rozdelenie bude mať $g(\hat{\theta}_n)$. Ako ukazuje Agresti ([1], str. 578), platí

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \text{ konverguje v distribúcii k } \mathsf{N}(0, \sigma^2[g'(\theta)]^2).$$

Tento výsledok je nazývaný *delta metódou* pre určenie asymptotického rozdelenia. Väčšinou σ^2 aj $g'(\theta)$ je závislé na θ, asymptotický rozptyl je neznámy. Využije sa teda výberový odhad $\hat{\theta}$ parametra θ. A za predpokladu spojitosti g' i σ v bode θ, je $\sigma(\hat{\theta})g'(\hat{\theta})$ konzistentným odhadom $\sigma(\theta)g'(\theta)$. A toto má významný dôsledok pri testovaní, že platí

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right)/\sigma(\hat{\theta})\left|g'(\hat{\theta})\right| \text{ konverguje v distribúcii k } \mathsf{N}(0,1).$$

Pre veľký výber je možné vytvoriť nasledujúci interval spoľahlivosti pre $g(\theta)$ na hladine významnosti α

$$\left(g(\hat{\theta}) + u_{\alpha/2}\,\sigma(\hat{\theta})\left|g'(\hat{\theta})\right|/\sqrt{n},\;\; g(\hat{\theta}) + u_{1-\alpha/2}\,\sigma(\hat{\theta})\left|g'(\hat{\theta})\right|/\sqrt{n}\right).$$

Ďalej Agresti zovšeobecnuje delta metódu pre funkciu náhodného vektora. Nech vektor štatistík $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,m})'$ má asymptoticky m-rozmerné normálne rozdelenie $\mathsf{N}_m(\boldsymbol{\theta}, \boldsymbol{\Sigma}/n)$. Nech ďalej funkcia $g(t_1, \ldots, t_m)$ má nenulový diferenciál $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)'$ v bode $\boldsymbol{\theta}$, kde

$$\phi_i = \left.\frac{\partial g}{\partial t_i}\right|_{\boldsymbol{t}=\boldsymbol{\theta}} \quad \text{pre} \quad i = 1, 2, \ldots, m.$$

Potom

$$\sqrt{n}\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right) \text{ konverguje v distribúcii k } \mathsf{N}(0, \boldsymbol{\phi}'\boldsymbol{\Sigma}\boldsymbol{\phi}). \tag{1.5}$$

Teda pre veľký výber n je rozdelenie $g(\hat{\boldsymbol{\theta}}_n)$ blízke normálnemu so strednou hodnotou $g(\boldsymbol{\theta})$ a rozptylom $\boldsymbol{\phi}'\boldsymbol{\Sigma}\boldsymbol{\phi}/n$.

1.2.1. Prehľad výsledkov z teórie matic

V práci bude predpokladaná znalosť základných pojmov z lineárnej algebry, prevažne teórie matic a vektorov. Ďalej budú uvedené potrebné výsledky z tejto teórie. Matematická teória v celej práci bude uvažovaná na poli reálnach čísiel $\mathbb{R}$. Matice budú vždy označované veľkými písmenami latinky tučne. Potom reálne vektory budú označené malými písmenami latinky tučne. Matica bude označovaná s príslušným typom matice takto $\boldsymbol{A}_{m\times n} = (a_{ij})_{\ i=1,..,m,\ j=1,...,n}$, kde m vyjadruje počet riadkov a n počet stĺpcov matice $\boldsymbol{A}_{m\times n}$. Vektory budú vždy uvažované v stĺpcovom tvare a budú označené $\boldsymbol{a} = (a_i)_{\ i=1,..,m}$ bez typu, ktorý bude vopred uvedený. Avšak ak bude zo zadania zrejmé o aký typ matice/vektora sa jedná alebo tento typ nebude dôležitý, matica bude pre jednoduchosť označené bez typu $\boldsymbol{A}$ a pri vektore $\boldsymbol{a}$ už nebude uvadzaný typ. Pre j-tý stĺpec matice $\boldsymbol{A}$ bude označený $\boldsymbol{a}_{\bullet j}$. V prípade štvorcovej matice $\boldsymbol{A}$ bude v indexe uvedený rád tejto matice takto $\boldsymbol{A}_n$. Jednotková matica bude označená $\boldsymbol{I}$ a nulová matica bude označená $\boldsymbol{O}$. Vektor nul bude označený $\boldsymbol{0}$. Hodnosť matice $\boldsymbol{A}$ bude označená $\mathrm{h}(\boldsymbol{A})$. Determinant štvorcovej matice $\boldsymbol{A}$ bude označený $\det(\boldsymbol{A})$. Transponovaná matica k matici $\boldsymbol{A}$ bude označená $\boldsymbol{A}'$. Inverzná matica k matici $\boldsymbol{A}$ bude označená $\boldsymbol{A}^{-1}$. Stopa matice $\boldsymbol{A}$ bude označená $\mathrm{Tr}(\boldsymbol{A})$. Pre nejakú štvorcovú maticu $\boldsymbol{A}$ a ľubovolný vektor $\boldsymbol{c}$ príslušného typu platí užitočný vzťah

$$\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} = \mathrm{Tr}(\boldsymbol{A}\boldsymbol{c}\boldsymbol{c}'), \tag{1.6}$$

ktorý je možné jednoduchým rozpísaním pravej a ľavej strany potvrdiť. Ďalej bude zavedené označenie $\mathcal{M}(\boldsymbol{A}_{m\times n}) = \mathcal{M}(\boldsymbol{a}_{\bullet 1}, \ldots, \boldsymbol{a}_{\bullet n})$, ktoré bude vyjadrovať vektorový priestor vytvorený lineárnou kombináciou stĺpcov matice $\boldsymbol{A}$ (resp. generovaný $\boldsymbol{a}_{\bullet 1}, \ldots, \boldsymbol{a}_{\bullet n}$). Vektorový priestor bude označovaný $\mathcal{V}$. Bude uvažovaná euklidovská norma označená $\|\cdot\|$. Matica $\boldsymbol{A}$ spĺňajúca $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$, bude nazývaná *ortogonálnou maticou*. Nasledujúca veta bude užitočná pri dokazovaní mnohých matematických viet v tejto práci.

Veta 1.2. *Nech je uvažovaná matica $\boldsymbol{A}$ s n stĺpcami. Potom platí pre vektorové priestory $\mathcal{M}(\boldsymbol{A}')$ a $\mathcal{M}(\boldsymbol{A}'\boldsymbol{A})$ v n-rozmernom vektorovom priestore $\mathcal{V}$ nasledujúce*

$$\mathcal{M}(\boldsymbol{A}') = \mathcal{M}(\boldsymbol{A}'\boldsymbol{A}). \tag{1.7}$$

Dôkaz. Zrejme priestory $\mathcal{M}(\boldsymbol{A}')$, $\mathcal{M}(\boldsymbol{A}'\boldsymbol{A})$ sú podpriestormi $\mathcal{V}$. Ak sú oba podpriestory identické v $\mathcal{V}$, majú identické aj svoje komplementy. A platí to aj opačne. Nech $\boldsymbol{v} \in \mathcal{V}$ a zároveň platia obe naledujúce tvrdenia:

1. pre $\forall \boldsymbol{v} \notin \mathcal{M}(\boldsymbol{A}')$ platí $\boldsymbol{v}'\boldsymbol{A}' = \boldsymbol{0}$ a úpravou vyplýva platnost $\boldsymbol{v}'\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{0}$, a teda záver $\boldsymbol{v} \notin \mathcal{M}(\boldsymbol{A}'\boldsymbol{A})$
2. pre $\forall \boldsymbol{v} \notin \mathcal{M}(\boldsymbol{A}'\boldsymbol{A})$ platí $\boldsymbol{v}'\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{0}$ a úpravou vyplýva platnosť taktiež $\boldsymbol{v}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{v} = 0$, z čoho jasne vyplýva $\boldsymbol{v}'\boldsymbol{A}' = \boldsymbol{0}$ a teda záver $\boldsymbol{v} \notin \mathcal{M}(\boldsymbol{A}')$

Potom sú ich komplementy totožné a taktiež oba podpriestory $\mathcal{M}(\boldsymbol{A}')$, $\mathcal{M}(\boldsymbol{A}'\boldsymbol{A})$. □

Nepriamym dôsledkom predchádzajúcej vety je platnosť vzťahu $\mathrm{h}(\boldsymbol{A}') = \mathrm{h}(\boldsymbol{A}'\boldsymbol{A})$.

Veta 1.3. (**o skeletnom rozklade**) *Nech $h(\boldsymbol{A}_{m\times n}) = r \geq 1$. Potom existujú matice $\boldsymbol{B}_{m\times r}$ a $\boldsymbol{C}_{r\times n}$ také, že platí*

$$\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C} \text{ a } h(\boldsymbol{B}) = h(\boldsymbol{C}) = r. \tag{1.8}$$

Dôkaz. Dôkaz existencie matic $\boldsymbol{B}$ a $\boldsymbol{C}$ bude založený na ich konštrukcii. Z $\mathrm{h}(\boldsymbol{A}) = r$ vyplýva, že matica $\boldsymbol{A}$ má r nezávislých stĺpcov. Nech matica $\boldsymbol{B}$ je tvorená z r nezávislých stĺpcov matice $\boldsymbol{A}$, v takom poradí ako boli v matici $\boldsymbol{A}$. Zrejme $\mathrm{h}(\boldsymbol{B}) = r$. Nech symbol $\mathcal{I}$ značí množinu čísiel, ktoré vyjadrujú poradie stĺpcov v matici $\boldsymbol{A}$ vybratých do matice $\boldsymbol{B}$. Potom matica $\boldsymbol{C}$ má stĺpce $\boldsymbol{c}_{\bullet i}$ na pozicii $i \in \mathcal{I}$ tvorené jednou jedničkou a $r-1$ nulami, tak aby platil súčin $\boldsymbol{B}\boldsymbol{c}_{\bullet i} = \boldsymbol{a}_{\bullet i}$. Stĺpce $\boldsymbol{c}_{\bullet j}$ na pozíciach $j \in \{1, \ldots, n\} - \mathcal{I}$ sú dané koeficientami lineárnej kombinácie stĺpcov matice $\boldsymbol{B}$, tak aby platil súčin $\boldsymbol{B}\boldsymbol{c}_{\bullet j} = \boldsymbol{a}_{\bullet j}$. A platí $\mathrm{h}(\boldsymbol{C}) = r$, pretože matica $\boldsymbol{C}$ obsahuje r nezávislých stĺpcov na všetkých r pozíciach z množiny $\mathcal{I}$. A keďže bol zavedený súčin $\boldsymbol{B}\boldsymbol{c}_{\bullet i} = \boldsymbol{a}_{\bullet i}$ pre $\forall i \in \{1, \ldots, n\}$, platí taktiež súčin $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$. □

Predchádzajúci dôkaz odpovedal taktiež na jednoznačnosť. Matice $\boldsymbol{B}_{m\times r}$ a $\boldsymbol{C}_{r\times n}$ nie sú jednoznačne maticou $\boldsymbol{A}_{m\times n}$ určené, ale v prípade výberu matice $\boldsymbol{B}$ resp. $\boldsymbol{C}$ je jednoznačne daná matica $\boldsymbol{C}$ resp. $\boldsymbol{B}$.

Nech je teraz uvažovaná matica $\boldsymbol{A}$ štvorcová. Rovnica $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$ pre neznáme číslo λ, je nazývaná *charakteristickou rovnicou* matice $\boldsymbol{A}$. Korene charakteristickej rovnice matice $\boldsymbol{A}$ sú nazývané *vlastnými číslami* matice $\boldsymbol{A}$. Nenulový vektor $\boldsymbol{v}$, spĺňujúci rovnicu $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{v} = \boldsymbol{0}$ pre nejaké vlastné číslo λ matice $\boldsymbol{A}$, je nazývaný *vlastným vektorom* príslušného vlastného čísla λ matice $\boldsymbol{A}$. Bez újmy na všeobecnosti bude ďalšom texte vždy uvažovaný vlastný vektor jednotkový (teda $\|\boldsymbol{v}\| = 1$).

Ak je matica $\boldsymbol{A}$ reálna, štvorcová a symetrická a pre každý reálny nenulový vektor $\boldsymbol{c}$ príslušného typu platí nerovnosť

$$\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} \geq 0, \tag{1.9}$$

potom matica $\boldsymbol{A}$ je nazývaná *pozitívne semidefinitnou* maticou. Ak vyššie definovaná matica $\boldsymbol{A}$ spĺňa pre každý nenulový vektor c ostrú nerovnosť vo vzťahu 1.9, potom je matica $\boldsymbol{A}$ nazývaná *pozitívne definitnou* maticou.

Veta 1.4. *Nech $\boldsymbol{A}_m$ je reálna symetrická matica. Potom existuje ortogonálna matica $\boldsymbol{B}$ s vlastnosťou $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}'$ resp. $\boldsymbol{D} = \boldsymbol{B}'\boldsymbol{A}\boldsymbol{B}$, kde $\boldsymbol{D}$ je diagonálna matica.*

Dôkaz. Viď Rao ([9], str. 62). □

Následne bude ukázaný takýto rozklad $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}'$ pre maticu $\boldsymbol{A}_m$, uvedený vo veta 1.4. Nech $\lambda_1, \ldots, \lambda_m$ sú vlastné čísla matice $\boldsymbol{A}_m$, uvedené toľko krát ako je ich násobnosť (ako koreňa charakteristickej rovnice). Vlastné čísla $\lambda_1, \ldots, \lambda_m$ sa po poradí priradia na diagonálu matice $\boldsymbol{D}$. Potom do každého stĺpca $\boldsymbol{b}_{\bullet i},\ i = 1, \ldots, m$ matice $\boldsymbol{B}$ sa priradí vlastný vektor $\boldsymbol{v}_i$ príslušný vlastnému číslu λ_i. V prípade k-nasobnosti ($k \leq m$) vlastného čísla λ_i, bude v matici $\boldsymbol{B}$ uvedených všetkých n rôznych vlastných vektorov príslušnych k vlastnému číslu λ_i. Vyššie uvedenému rozkladu symetrickej matice

$$\boldsymbol{A} = \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}' = \sum_{j=1}^{m} d_{jj}\boldsymbol{b}_{\bullet j}\boldsymbol{b}'_{\bullet j} = \sum_{j=1}^{m} \lambda_j \boldsymbol{v}_j \boldsymbol{v}'_j, \tag{1.10}$$

sa hovorí *spektrálny rozklad*. Je zrejmé, že každá matica $\boldsymbol{v}_j\boldsymbol{v}'_j,\ j = 1, \ldots, m$ je symetrická matica a preto aj ich lineárna kombinácia (určená vlastnými číslami) bude opäť tvoriť symetrickú maticu $\boldsymbol{A}$. Keďže v spektrálnom rozklade 1.10 je vždy uvažovaná matica $\boldsymbol{B}$ regulárna, hodnosť matice $\boldsymbol{A}$ je rovná počtu nenulových prvkov na diagonále matice $\boldsymbol{D}$.

Bez dôkazu bude uvedené nasledujúce tvrdenie. Reálna symetrická matica má realné vlastné čísla a vlastné vektory. Pozitívne semidefinitná matica $\boldsymbol{A}_m$ má vždy nezáporné vlastné čísla, pretože v definícii zo vzťahu 1.9 a vety 1.4. vyplýva

$$\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} = \boldsymbol{c}'\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}'\boldsymbol{c} = \boldsymbol{e}'\boldsymbol{D}\boldsymbol{e} = \sum_{j=1}^{m} \lambda_j e_j^2 \geq 0, \tag{1.11}$$

kde $\boldsymbol{e} = \boldsymbol{B}'\boldsymbol{c}$. V posledné vyjadrenie vo vzťahu 1.11 platí pre ľubovoľný vektor $\boldsymbol{c}$ resp. $\boldsymbol{e}$, len ak všetky vlastné čísla λ_j sú nezáporné. Pozitívne definitná matica $\boldsymbol{A}_m$ má všetky vlastné čísla kladné. A to sa ukáže rovnako.

Ďalej bude tento odstavec venovaný špeciálnym maticiam, ktoré majú významné uplatnenie v teórii lineárnych regresných modelov. Dobrým príkladom je idempotentná alebo pseudoinverzná matica.

Idempotentná matica

Štvrcová matica $\boldsymbol{A}$ bude nazývaná *idempotentnou maticou*, ak platí $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$.

Veta 1.5. *Ak $\boldsymbol{A}$ je matica idempotentná, potom platí $h(\boldsymbol{A}) = Tr(\boldsymbol{A})$.*

Dôkaz. Dôkaz je technického charakteru. Viď Anděl ([2], str. 66). □

Pseudoinverzná matica

Nech je daná matica $\boldsymbol{A}_{m\times n}$. Potom každá matica $\boldsymbol{B}_{n\times m}$ spĺňajúca rovnosť

$$\boldsymbol{ABA} = \boldsymbol{A}, \tag{1.12}$$

bude nazývaná *pseudoinverznou maticou* k matici $\boldsymbol{A}$. Pseudoinverzná matica k matici $\boldsymbol{A}$ bude v ďalšom texte vyznačená horným indexom nasledovne $\boldsymbol{A}^-$. Len v prípade regulárnosti matice $\boldsymbol{A}$ je pseudoinverzná matica určená jednoznačne. Pre regulárnu maticu $\boldsymbol{A}$, platí rovnica $\boldsymbol{ABA} = \boldsymbol{A}$ ekvivalentne s rovnicami $\boldsymbol{AB} = \boldsymbol{I} = \boldsymbol{BA}$. A tieto rovnice platia súčasne len pre maticu $\boldsymbol{B} = \boldsymbol{A}^{-1}$ inverznú k matici $\boldsymbol{A}$.

Nech je daná sústava lineárnych rovníc

$$\boldsymbol{Ab} = \boldsymbol{y}, \text{ kde } \boldsymbol{y} \in \mathcal{M}(\boldsymbol{A}). \tag{1.13}$$

Ďalej nech je uvažovaný vzťah $\boldsymbol{AA}^-\boldsymbol{Ab} = \boldsymbol{Ab}$, z čoho vyplýva $\boldsymbol{AA}^-\boldsymbol{y} = \boldsymbol{y}$ pri platnosti 1.13. Je zrejmé, že sústavy $\boldsymbol{AA}^-\boldsymbol{y} = \boldsymbol{y}$ a 1.13 súčasne platia, ak $\boldsymbol{b} = \boldsymbol{A}^-\boldsymbol{y}$. Záver je nasledovný, pre ľubovoľnú sústavu lineárnych rovníc 1.13 je riešením $\boldsymbol{b} = \boldsymbol{A}^-\boldsymbol{y}$ (matica sústavy $\boldsymbol{A}$ môže byť singulárna i neštvorcová). Pseudoinverzná matica je teda zovšeobecnením pojmu inverznej matice pri riešení sústavy lineárnych rovníc.

Operátor ortogonálnej projekcie

Nech je uvažovaný vektorový priestor $\mathcal{V}$ konečnej dimenzie na $\mathbb{R}$. Nech je ďalej uvažovaný podpriestor $\mathcal{V}_1$ vo $\mathcal{V}$ a jeho ortogonálny komplement $\mathcal{V}_1^\perp$ vo $\mathcal{V}$ (platí $\mathcal{V}_1 \cap \mathcal{V}_1^\perp = \{\boldsymbol{0}\}$). Potom každý vektor $\boldsymbol{x} \in \mathcal{V}$ je možné jednoznačne vyjadriť dvomi vektormi $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ nasledovne $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, kde $\boldsymbol{x}_1 \in \mathcal{V}_1$ a $\boldsymbol{x}_2 \in \mathcal{V}_1^\perp$. Zobrazenie $\boldsymbol{P} : \mathcal{V} \to \mathcal{V}_1$ definované vzťahom $\boldsymbol{Px} = \boldsymbol{x}_1$ sa nazýva *ortogonálnym operátorom projekcie* na priestor $\mathcal{V}_1$. Ďalej zrejme platí $\boldsymbol{P}(a\boldsymbol{x} + b\boldsymbol{y}) = a\boldsymbol{Px} + b\boldsymbol{Py}$ pre $a,\ b \in \mathbb{R}$ a teda $\boldsymbol{P}$ je lineárnou transformáciou, ktorú je možné vyjadriť maticou. Ďalej bude označená takáto matica symbolom $\boldsymbol{P}$ a bude nazývaná *projekčnou maticou.*

Veta 1.6. *Projekčná matica $\boldsymbol{P}$ je idempotentná matica.*

Dôkaz. Zrejme platí vyššie uvedený vzťah $\boldsymbol{Px} = \boldsymbol{x}_1$ a vzťah $\boldsymbol{Px}_1 = \boldsymbol{x}_1$. Dosadením vyjadrenia $\boldsymbol{x}_1$ z prvého vzťahu do druhého je možné dostať vzťah $\boldsymbol{PPx} = \boldsymbol{Px}$, z čoho vyplýva $\boldsymbol{PP} = \boldsymbol{P}$. □

Veta 1.7. *Ak $\boldsymbol{P}$ je projekčná matica na podpriestor $\mathcal{V}_1$, potom $\boldsymbol{I} - \boldsymbol{P}$ je projekčná matica na podpriestor $\mathcal{V}_1^\perp$.*

Dôkaz. Nech je uvažovaný ľubovoľný vektor $\boldsymbol{x} \in \mathcal{V}$ a príslušný vektor $\boldsymbol{x}_1 = \boldsymbol{Px}$, kde zrejme $\boldsymbol{x}_1 \in \mathcal{V}_1$. V prípade nenulovosti vektora $\boldsymbol{x} - \boldsymbol{x}_1 = \boldsymbol{x}_2$, platí $\boldsymbol{x}_2 \notin \mathcal{V}_1$ ale $\boldsymbol{x}_2 \in \mathcal{V}_1^\perp$. Potom $\boldsymbol{x}_2 = \boldsymbol{x} - \boldsymbol{x}_1 = \boldsymbol{Ix} - \boldsymbol{Px} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}$ a teda $\boldsymbol{I} - \boldsymbol{P}$ je projekčnou maticou z priestoru $\mathcal{V}$ na podpriestor $\mathcal{V}_1^\perp$. □

Nech je uvažovaný n-rozmerný vektorový priestor $\mathcal{V}$ so skalárnym súčinom zadefinovaným pozitívne definitnou maticou $\boldsymbol{W}_n$ nasledovne $\langle\cdot,\cdot\rangle_{\boldsymbol{W}} : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$, kde dvom vektorom $\boldsymbol{x},\ \boldsymbol{y} \in \mathcal{V}$ priradí toto zobrazenie $\langle\boldsymbol{x},\ \boldsymbol{y}\rangle_{\boldsymbol{W}}$ hodnotu $\boldsymbol{x'Wy}$. V prípade, uvedenia skalárneho súčinu bez maticového indexu takto $\langle\cdot,\cdot\rangle$, bude uvažovaná matica $\boldsymbol{W}$ jednotková $\boldsymbol{I}$.

Veta 1.8. *Nech je daný vektorový priestor $\mathcal{V}$, kde je definovaný skalárny súčin pozitívne definitnou maticou $\boldsymbol{\Sigma}$. Potom matica $\boldsymbol{P}$ je operátorom ortogonálnej projekcie, práve vtedy, keď matica $\boldsymbol{P}$ je idempotentná a zároveň $\boldsymbol{\Sigma P}$ je symetrická matica.*

Dôkaz. Z vety 1.7. vyplýva, že každá projekčná matica musí mať vlastnosť idempotentnej matice. Nech sú uvažované dva ľubovoľné vektory $\boldsymbol{x},\ \boldsymbol{y} \in \mathcal{V}$, potom $\boldsymbol{Px} \in \mathcal{V}_1,\ (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y} \in \mathcal{V}_1^{\perp}$ a aby boli na seba ortogonálne, musí pre ne vždy platiť

$$0 = \langle \boldsymbol{Px},\ (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}\rangle_{\boldsymbol{\Sigma}} = \ \boldsymbol{x}'\boldsymbol{P}'\boldsymbol{\Sigma}(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}.$$

Predchádzajúci vzťah ekvivalentne platí, pre ľubovoľné vektory $\boldsymbol{x}$ a $\boldsymbol{y}$, len ak platí $\boldsymbol{P}'\boldsymbol{\Sigma}(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{O}$ a teda $\boldsymbol{P}'\boldsymbol{\Sigma P} = \boldsymbol{P}'\boldsymbol{\Sigma}$. Z tejto druhej rovnice a jej transponovanej verzie $\boldsymbol{P}'\boldsymbol{\Sigma P} = \boldsymbol{\Sigma P}$ vyplýva platnosť $\boldsymbol{P}'\boldsymbol{\Sigma} = (\boldsymbol{\Sigma P})' = \boldsymbol{\Sigma P}$ a teda symetrickosť matice $\boldsymbol{\Sigma P}$. □

V prípade skalárneho súčinu definovaného jednotkovou maticou $\boldsymbol{I}$ je matica $\boldsymbol{P}$ operátorom ortogonálnej projekcie práve vtedy, keď $\boldsymbol{P}$ je idempotentná a symetrická matica.

2. Lineárny regresný model úplnej hodnosti

Nech $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ je náhodný vektor a $\boldsymbol{X}_{n\times k}$ je matica daných reálnych čísiel, kde zároveň platí $\mathrm{h}(\boldsymbol{X}) = k$ a $k < n$. Ďalej nech pre náhodný vektor $\boldsymbol{Y}$ platí

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.1}$$

kde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ je vektor neznámych parametrov a $\boldsymbol{e} = (e_1, \ldots, e_n)'$ je náhodný vektor spĺňajúci podmienky

$$\mathsf{E}\boldsymbol{e} = \boldsymbol{0}, \quad \mathsf{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$$

s neznámym parametrom $\sigma^2 > 0$. V takomto prípade sa tiež povie, že $\boldsymbol{Y}$ sa riadi lineárnym modelom a bude nazývaný *regresným*. Pretože vektor $\boldsymbol{Y}$ závisí lineárne na prvkoch vektora $\boldsymbol{\beta}$, bude model 2.1 nazývaný *lineárnym regresným modelom*. Matica $\boldsymbol{X}$ bude nazývaná *regresnou maticou* a jej stĺpce $\boldsymbol{x}_{\bullet 1}, \ldots, \boldsymbol{x}_{\bullet k}$ budú nazývané *regresormi*. Vektor $\boldsymbol{\beta}$ bude nazývaný *vektorom regresných koeficientov*. Keďže regresná matica má rovnaký počet stĺpcov ako je jej hodnosť, bude model 2.1 nazývaný *model s úplnou stĺpcovou hodnosťou* alebo skrátene *model úplnej hodnosti*.

Podmienka $\mathrm{h}(\boldsymbol{X}) = k$ matice $\boldsymbol{X}_{n\times k}$ v modeli 2.1 určuje, že sa jedná o model úplnej hodnosti a teda stĺpce regresnej matice sú lineárne nezávislé. Zároveň ak matica $\boldsymbol{X}$ je úplnej hodnosti, potom matica $\boldsymbol{X'X}$ je regulárna. Predchádzajúce tvrdenie je tiež nepriamym dôsledkom vety 1.2. Priestor $\mathcal{M}(\boldsymbol{X})$ bude nazývaný *regresným priestorom*. Ďalej nech je zavedený pojem *reziduálneho priestoru*, ktorý bude vyjadrovať priestor kolmý na regresný priestor a bude označený $\mathcal{M}(\boldsymbol{X})^\perp$ (resp. $\mathcal{M}(\boldsymbol{X})^\perp$ je komplementárnym podpriestorom $\mathcal{M}(\boldsymbol{X})$ v $\mathbb{R}^n$).

Ďalej nech je uvažovaný všeobecnejší model úplnej hodnosti pre náhodný vektor $\boldsymbol{Z}$ tvaru

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.2}$$

s neznámym vektorom $\boldsymbol{\beta}$ a neznámym náhodným vektorom $\boldsymbol{e}$

$$\mathsf{E}\boldsymbol{e} = \boldsymbol{0}, \quad \mathsf{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{G},$$

kde $\boldsymbol{G}$ je známa pozitívne definitná matica. V prípade že matica $\boldsymbol{G}$ nie je diagonálna, budú medzi pozorovaniami korelácie. Hodnoty $\boldsymbol{Z}$ je možné lineárne transformovať nasledovne $\boldsymbol{Y} = \boldsymbol{G}_1^{-1}\boldsymbol{Z}$ a získať pôvodný model 2.1 pre $\boldsymbol{Y}$ s $\mathsf{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$, kde matica $\boldsymbol{G}_1$ je zavedená vzťahom $\boldsymbol{G} = \boldsymbol{G}_1\boldsymbol{G}_1'$). Pretože matica $\boldsymbol{G}$ odpovedá variančnej matici, má všetky vlastné čísla kladné a platí, že existuje rozklad podľa vety 1.4, potom existuje aj vyššie zadefinovaná matica $\boldsymbol{G}_1$. Takto je možné previesť všeobecnejší model $\boldsymbol{Z}$, s ľubovoľnou známou pozitívne definitnou maticou $\boldsymbol{G}$, na pôvodný prípad modelu 2.1.

Regresné koeficienty $\beta_1, \ldots, \beta_k$ sa odhadujú *metódou najmenších štvorcov*, tj. minimalizáciou funkcie $S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. Tieto odhady budú spoločne označené vektorom $\boldsymbol{b} = (b_1, \ldots, b_k)'$. Ilustratívny môže byť obrázok 3**a**, v 3. kapitole, kde je prípad uvažovaný na $\mathbb{R}^2$.

Veta 2.1. *Ak má matica $\boldsymbol{X}$ úplnú stĺpcovú hodnosť, potom odhady metódou najmenších štvorcov sú $\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$.*

Dôkaz. Nech je uvažovaný taký vektor $\boldsymbol{b}$ pre ktorý buď platí $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b} \notin \mathcal{M}(\boldsymbol{X})$ alebo $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}$. Potom samozrejme platí vzťah

$$\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{0}, \tag{2.3}$$

ktorý bude využitý v nasledujúcom odvodzovaní.

$$\begin{aligned}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) &= [(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + (\boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta})]' [(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + (\boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta})] \\ &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + (\boldsymbol{b} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{\beta}) \\ &\geq (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})\end{aligned}$$

Pretože matica $\boldsymbol{X}'\boldsymbol{X}$ je regulárna, tak je i pozitívne definitná. A práve pre túto vlastnosť, v prípade $(\boldsymbol{b} - \boldsymbol{\beta}) = \boldsymbol{0}$, nadobúda súčet štvorcov $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ svoje minimum $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})$. Sústavu lineárnych rovníc 2.3 je možné prepísať na $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$ a kvôli úplnej hodnosti $\boldsymbol{X}$, priamo dostať vysledok pre odhady $\boldsymbol{b}$. □

Je vhodné si uvedomiť, že súčin 2.3 vyjadruje podmienku: vektor $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}$ je ortogonálny na každý stĺpec matice $\boldsymbol{X}$ a teda zároveň na $\mathcal{M}(\boldsymbol{X})$. Tu môže byť vhodná geometrická predstava viď obrázok 1.

Je možné získať aj iný pohľad na metódu najmenších štvorcov (minimalizáciu funkcie $S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$) s ekvivalentným záverom, ako v dôkaze vety 2.1. Keďže funkcia $S(\boldsymbol{\beta})$ je konvexná na $\mathbb{R}^k$, pre neobmedzené odhady $\boldsymbol{b}$ musí platiť

$$\frac{\partial}{\partial b_i}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = 0, \quad i = 1, 2, \ldots, k\,,$$

a zjednodušene platí

$$\frac{\partial}{\partial b_i}(\boldsymbol{Y} - \boldsymbol{x}_{\bullet i}b_i)'(\boldsymbol{Y} - \boldsymbol{x}_{\bullet i}b_i) = 0, \quad i = 1, 2, \ldots, k\,, \tag{2.4}$$

kde $\boldsymbol{x}_{\bullet i}$ značí i-tý stĺpec matice $\boldsymbol{X}$. Každú z parciálnych derivácií zo vzťahu 2.4 je možné rozpísať nasledovne

$$\frac{\partial}{\partial b_i}(\boldsymbol{Y} - \boldsymbol{x}_{\bullet i}b_i)'(\boldsymbol{Y} - \boldsymbol{x}_{\bullet i}b_i) = -\boldsymbol{x}'_{\bullet i}(\boldsymbol{Y} - \boldsymbol{x}_{\bullet i}b_i) - (\boldsymbol{Y} - \boldsymbol{x}_{\bullet i}b_i)'\boldsymbol{x}_{\bullet i} = -2\boldsymbol{x}'_{\bullet i}(\boldsymbol{Y} - \boldsymbol{x}_{\bullet i}b_i).$$

Potom zo vzťahu 2.4 vyplýva platnosť sústavy lineárnych rovníc $\boldsymbol{x}'_{\bullet i}(\boldsymbol{Y} - \boldsymbol{x}_{\bullet i}b_i) = 0$, pre $i = 1, 2, \ldots, k$, ktorú je možné zapísať maticovo

$$\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{0} \;\text{ resp. }\; \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}. \tag{2.5}$$

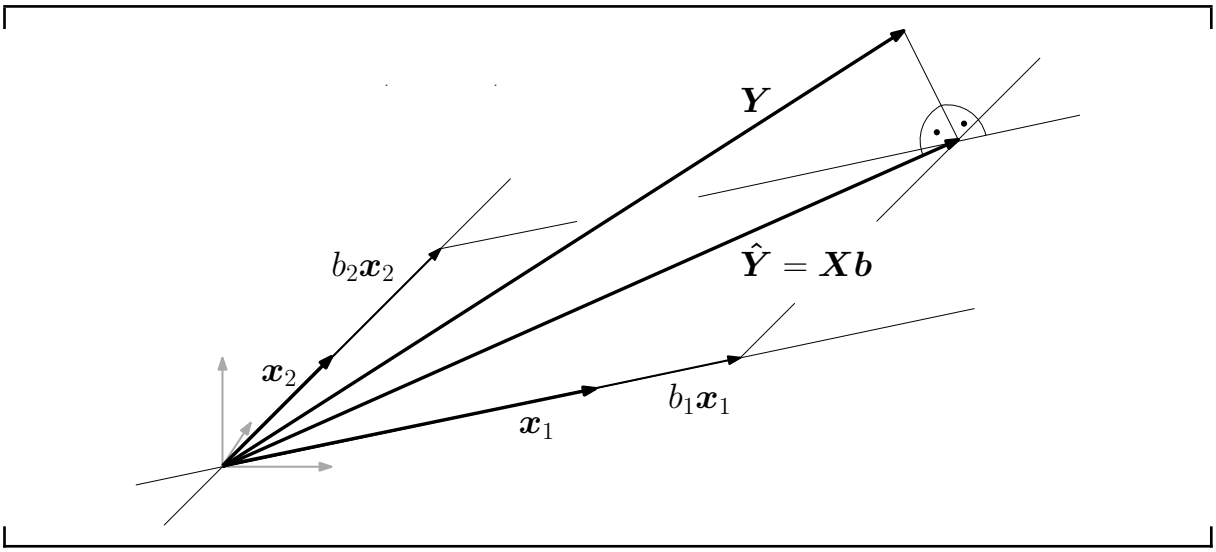
Sústava rovníc $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$, vyplývajúca taktiež z dôkazu vety 2.1., určuje teda riešenia pre odhady $\boldsymbol{b}$ regresných koeficientov $\boldsymbol{\beta}$ získaných metódou najmenších štvorcov. Je určite dobré si položiť otázku existencie riešenia, či vôbec pre takúto sústavu rovníc musí existovať riešenie. Z platnosti vzťahu $\boldsymbol{X}'\boldsymbol{Y} \in \mathcal{M}(\boldsymbol{X}')$ a vety 1.7. vyplýva platnosť $\boldsymbol{X}'\boldsymbol{Y} \in \mathcal{M}(\boldsymbol{X}'\boldsymbol{X})$. Keďže pravá strana sústavy $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$ je podľa vety 1.2. taktiež z priestoru $\mathcal{M}(\boldsymbol{X}'\boldsymbol{X})$, má táto sústava vždy riešenie. V prípade úplnej hodnosti matice $\boldsymbol{X}$ je riešenie tejto lineárnej sústavy vždy jednoznačné. Sústava lineárnych rovníc

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y} \tag{2.6}$$

bude nazývaná *sústava normálnych rovníc* alebo skrátene *normálne rovnice* (*normálna rovnica*). Vektor

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{2.7}$$

je najlepšou aproximáciou vektora $\boldsymbol{Y}$, akú je možné vytvoriť v priestore $\mathcal{M}(\boldsymbol{X})$. Z geometrickej predstavy riešenia normálnych rovníc vyplýva, že sa jedná o kolmý priemet vektora $\boldsymbol{Y}$ na $\hat{\boldsymbol{Y}} \in \mathcal{M}(\boldsymbol{X})$ a ten je určený jednoznačne. Táto situácia je zobrazená na obrázku 1.



Obrázok 1: Zobrazený je kolmý priemet vektora $\boldsymbol{Y}$ na vektor $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b} = b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2$ do regresného priestoru $\mathcal{M}(\boldsymbol{X}) = \mathcal{M}(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Vektor $\boldsymbol{b}$ je odhadom vektora $\boldsymbol{\beta}$ metódou najmenších štvorcov.

Vektor $\boldsymbol{Y}$ je lineárne transformovaný do priestoru $\mathcal{M}(\boldsymbol{X})$ na vektor $\hat{\boldsymbol{Y}}$ pomocou vzťahu 2.7. V ďalšom texte bude $\boldsymbol{H}$ označovať maticu tejto lineárnej transformácie

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

Je zrejmé, že platí vzťah $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$, totiž $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}$. Preto vektor z priestoru $\mathcal{M}(\boldsymbol{X})$ sa už touto ortogonálnou projekciou $\boldsymbol{H}$ nezmení, a teda každý stĺpec matice $\boldsymbol{X}$ sa transformuje na seba $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ z čoho vyplýva, že matica $\boldsymbol{H}$ je projekčnou maticou na priestor $\mathcal{M}(\boldsymbol{X})$. Zrejme ďalej platí tiež $(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{X} = \boldsymbol{O}$, teda že stĺpce projekčnej matice $(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$ (z vety 1.7.) sú ortogonálne na regresný priestor $\mathcal{M}(\boldsymbol{X})$. Ďalej nech $\boldsymbol{M}$ vyjadruje $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{I} - \boldsymbol{H}$.

Zhrnutím platia vzťahy

$$\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}, \quad \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{H} \quad \text{a} \quad \boldsymbol{M}\boldsymbol{X} = \boldsymbol{O}. \tag{2.8}$$

Dá sa ľahko ukázať, že matice $\boldsymbol{H}$ a $\boldsymbol{M}$ sú okrem symetrickosti aj idempotentné, keďže platí

$$\boldsymbol{H}\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{H},$$

$$\boldsymbol{M}\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{I} - \boldsymbol{H} - \boldsymbol{H} + \boldsymbol{H}\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{H} = \boldsymbol{M}.$$

Alebo inak, každý stĺpec z projekčnej matice $\boldsymbol{H}$ resp. $\boldsymbol{M}$ sa transformuje na seba do priestoru $\mathcal{M}(\boldsymbol{H})$ resp. $\mathcal{M}(\boldsymbol{M})$, teda platí $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ a $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$. A tým sa ukázalo, že

matice $\boldsymbol{H}$ a $\boldsymbol{M}$ sú symetrické a idempotentné. Podľa vety 1.8. sú obe matice operátormi ortogonálnej projektcie (na regresný a reziduálny priestor) a tým boli potvrdené doterajšie intuitívne predstavy.

Pre idempotentné matice, podľa vety 1.5., platia vzťahy

$$\mathrm{h}(\boldsymbol{H}) = \mathrm{Tr}(\boldsymbol{H}) = k, \quad \mathrm{h}(\boldsymbol{M}) = \mathrm{Tr}(\boldsymbol{M}) = n - k.$$

Matica $\boldsymbol{H}$ je symetrická a preto môže byť uvažovaný spektrálny rozklad 1.10. Keďže $\mathrm{h}(\boldsymbol{H}) = k$ platí, že matica $\boldsymbol{H}$ má práve k nenulových vlastných čísiel (i s násobnosťami) s príslušnými vlastnými vektormi. Nech $\boldsymbol{v}$ je nejaký vlastný vektor matice $\boldsymbol{H}$ príslušný k nenulovému vlastnému číslu λ. Potom platí $\boldsymbol{H}\boldsymbol{v} = \lambda\boldsymbol{v}$. Keďže $\boldsymbol{v} \in \mathcal{M}(\boldsymbol{H})$ a $\boldsymbol{H}$ je projekčná matica na $\mathcal{M}(\boldsymbol{H})$ musí platiť $\lambda = 1$. Preto nech pre $\boldsymbol{H}$ sú označené vlastné čísla nasledovne $\lambda_i = 1,\ i = 1, \ldots, k$ a $\lambda_i = 0,\ i = k+1, \ldots, n$. Potom podľa spektrálneho rozkladu 1.10 platí pre maticu $\boldsymbol{H}$

$$\boldsymbol{H} = \sum_{j=1}^{m} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j' = \sum_{j=1}^{k} \boldsymbol{v}_j \boldsymbol{v}_j' = \boldsymbol{Q}\boldsymbol{Q}', \tag{2.9}$$

kde matica $\boldsymbol{Q}_{n\times k}$ má stĺpce vyplnené vlastnými vektormi takto $\boldsymbol{Q} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$. Vektory $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ sú ortonormálne bázové vektory priestoru $\mathcal{M}(\boldsymbol{H})$ (a teda i regresného priestoru $\mathcal{M}(\boldsymbol{X})$) a preto platí $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_k$. Obdobne pre maticu $\boldsymbol{M}$ existuje $n-k$ vlastných vektorov matice $\boldsymbol{M}$ generujúcich priestor $\mathcal{M}(\boldsymbol{M})$ (a teda i reziduálneho priestoru $\mathcal{M}(\boldsymbol{X})^{\perp}$). Keďže sú priestory $\mathcal{M}(\boldsymbol{H})$ a $\mathcal{M}(\boldsymbol{M})$ na seba ortogonálne, ortonormálnou bázou sú práve vektory $\boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_n$ zo spektrálneho rozkladu 2.9 matice $\boldsymbol{H}$. Nech je teda, rovnako ako matica $\boldsymbol{Q}$, vytvorená matica $\boldsymbol{N}_{n\times n-k} = (\boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_n)$. Potom platí okrem vzťahu $\boldsymbol{M} = \boldsymbol{N}\boldsymbol{N}'$ aj vzťah $\boldsymbol{N}'\boldsymbol{N} = \boldsymbol{I}_{n-k}$. Pomocou platnosti $\boldsymbol{H} + \boldsymbol{M} = \boldsymbol{I}$ (zo vzťahu 2.8) je možné ukázať

$$\boldsymbol{I} = \boldsymbol{H} + \boldsymbol{M} = \sum_{j=1}^{k} \boldsymbol{v}_j \boldsymbol{v}_j' + \sum_{j=k+1}^{n} \boldsymbol{v}_j \boldsymbol{v}_j' = (\boldsymbol{Q}, \boldsymbol{N})(\boldsymbol{Q}, \boldsymbol{N})'. \tag{2.10}$$

A teda, že matica $(\boldsymbol{Q}, \boldsymbol{N})$ (v blokovom tvare) je skutočne ortogonálna, a spĺňa podmienky vety 1.4. Totiž len rovnosť $(\boldsymbol{Q}, \boldsymbol{N})'(\boldsymbol{Q}, \boldsymbol{N}) = \boldsymbol{I}$ je hneď zrejmá.

Náhodný vektor $\boldsymbol{b}_L$ s k-zložkami bude nazývaný *lineárnym* odhadom parametra $\boldsymbol{\beta}$ (v lineárnom regresnom modeli 2.1), ak existuje matica $\boldsymbol{B}_{k\times n}$, že

$$\boldsymbol{b}_L = \boldsymbol{B}\boldsymbol{Y}. \tag{2.11}$$

Lineárny odhad $\boldsymbol{b}_L$ bude nazývaný *nestranným* odhadom, ak platí vzťah $\mathsf{E}(\boldsymbol{b}_L) = \boldsymbol{\beta}$ pre každé $\boldsymbol{\beta} \in \mathbb{R}^k$ Nestranný lineárny odhad $\boldsymbol{b}_L$ bude nazývaný *najlepším* nestranným odhadom, ak platí pre ľubovoľný iný nestranný lineárny odhad $\boldsymbol{b}_L^*$, že matica

$$(\mathsf{Var}(\boldsymbol{b}_L^*) - \mathsf{Var}(\boldsymbol{b}_L)) \tag{2.12}$$

je pozitívne semidefinitná. Vysvetlenie tejto definície môže byť nasledujúce. Odhad $\boldsymbol{b}_L$ je najlepší, pretože sa rozptyl $\mathsf{D}(\boldsymbol{c}'\boldsymbol{b}_L)$ ľubovoľnej lineárnej kombinácie odhadov $\boldsymbol{c}'\boldsymbol{b}_L$ zmení oproti rozptylu $\mathsf{D}(\boldsymbol{c}'\boldsymbol{b}_L^*)$ o hodnotu $\boldsymbol{c}'(\mathsf{Var}(\boldsymbol{b}_L^*) - \mathsf{Var}(\boldsymbol{b}_L))\boldsymbol{c}$, ktorá je nezáporná (pretože matica 2.12 je pozitívne semidefinitnou). Teda odhad $\boldsymbol{b}_L$ je najlepší, pretože pre ľubovoľný vektor $\boldsymbol{c}$ minimalizuje rozptyl na $\mathsf{D}(\boldsymbol{c}'\boldsymbol{b}_L)$ oproti ostatným nestranným lineárnym odhadom.

Veta 2.2. *Pre odhad* $\boldsymbol{b}$ *metódou najmenších štvorcov platí* $\mathsf{E}(\boldsymbol{b}) = \boldsymbol{\beta}$, $\mathsf{Var}(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Dôkaz.

$$\begin{aligned} \mathsf{E}(\boldsymbol{b}) &= \mathsf{E}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathsf{E}\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}, \\ \mathsf{Var}(\boldsymbol{b}) &= \mathsf{Var}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\mathsf{Var}\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \end{aligned} \tag{2.13}$$

□

Teda bolo ukázané, že odhad metódou najmenších štvorcov $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ uvedený vo vete 2.1. je nestranným lineárnym odhadom vektora $\boldsymbol{\beta}$.

Veta 2.3. (**Gaussova-Markovova**) *V modeli* $\boldsymbol{Y}$*, pre ktorý platí* $\mathsf{E}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$ *a* $\mathsf{Var}\boldsymbol{Y} = \sigma^2\boldsymbol{I}$*, je* $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ *najlepším nestranným lineárnym odhadom vektora* $\boldsymbol{X}\boldsymbol{\beta}$*. Zároveň platí* $\mathsf{Var}\hat{\boldsymbol{Y}} = \sigma^2\boldsymbol{H}$.

Dôkaz. Odhad $\hat{\boldsymbol{Y}}$ je lineárny, pretože $\hat{\boldsymbol{Y}}$ je lineárnou funkciou $\boldsymbol{Y}$ ($\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$). Keďže platí

$$\mathsf{E}\hat{\boldsymbol{Y}} = \mathsf{E}\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta},$$

je odhad $\hat{\boldsymbol{Y}}$ nestranný. Ďalej nech je uvažovaný ľubovoľný iný lineárny odhad, teda napríklad $\tilde{\boldsymbol{Y}} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{Y}$. Aby bol lineárny odhad $\tilde{\boldsymbol{Y}}$ nestranný musí platiť

$$\mathsf{E}\tilde{\boldsymbol{Y}} = \mathsf{E}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{Y}) = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} \quad \text{pre každé} \quad \boldsymbol{\beta} \in \mathbb{R}^k.$$

Z toho vyplýva $\boldsymbol{a} = \boldsymbol{0}$ a $\boldsymbol{B}\boldsymbol{X} = \boldsymbol{X}$. Matica $\boldsymbol{B}$ teda transformuje stĺpce matice $\boldsymbol{X}$ na samé seba. Preto matica $\boldsymbol{B}$ taktiež lineárne transformuje lineárnu kombináciu stĺpcov $\boldsymbol{X}$ na seba. Preto $\boldsymbol{B}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ z čoho vyplýva rovnosť $\boldsymbol{B}\boldsymbol{H} = \boldsymbol{H}$. Posledná rovnosť sa dá prepísať na vzťah $(\boldsymbol{B} - \boldsymbol{H})\boldsymbol{H} = \boldsymbol{O}$, využitím idemotentnosti matice $\boldsymbol{H}$. Tento vzťah bude využitý nižšie. Pre variančnú maticu odhadu $\hat{\boldsymbol{Y}}$ platí

$$\mathsf{Var}(\hat{\boldsymbol{Y}}) = \mathsf{Var}(\boldsymbol{H}\boldsymbol{Y}) = \boldsymbol{H}\mathsf{Var}(\boldsymbol{Y})\boldsymbol{H}' = \sigma^2\boldsymbol{H}.$$

Pre variančnú maticu lineárneho nestranného odhadu $\tilde{\boldsymbol{Y}}$ platí

$$\begin{aligned} \mathsf{Var}(\tilde{\boldsymbol{Y}}) &= \boldsymbol{B}\sigma^2\boldsymbol{I}\boldsymbol{B}' = \sigma^2[(\boldsymbol{B} - \boldsymbol{H}) + \boldsymbol{H}][(\boldsymbol{B} - \boldsymbol{H}) + \boldsymbol{H}]' = \\ &= \sigma^2(\boldsymbol{B} - \boldsymbol{H})(\boldsymbol{B} - \boldsymbol{H})' + \sigma^2\boldsymbol{H}\boldsymbol{H}' = \sigma^2(\boldsymbol{B} - \boldsymbol{H})(\boldsymbol{B} - \boldsymbol{H})' + \mathsf{Var}(\hat{\boldsymbol{Y}}), \end{aligned}$$

pretože $(\boldsymbol{B} - \boldsymbol{H})\boldsymbol{H} = \boldsymbol{O}$. Aby vektor $\hat{\boldsymbol{Y}}$ bol najlepší nestranný odhad, musí platiť, že matica $(\mathsf{Var}(\tilde{\boldsymbol{Y}}) - \mathsf{Var}(\hat{\boldsymbol{Y}}))$ je pozitívne semidefinitná matica podľa vzťahu 2.12. Podľa tvaru matice $(\boldsymbol{B} - \boldsymbol{H})(\boldsymbol{B} - \boldsymbol{H})'$, sa jedná o pozitívne semidefinitnú maticu, pretože je symetrická a má nezáporné vlastné čísla. Platí totiž ekvivalentne $(\boldsymbol{B} - \boldsymbol{H})\boldsymbol{I}(\boldsymbol{B} - \boldsymbol{H})'$, kde matica $\boldsymbol{I}$ určuje nezápornosť vlastných čísiel. □

Dôležitým dôsledkom predchádzajúcej Gauss-Markovovej vety je pre odhad $\boldsymbol{b}$ vektora $\boldsymbol{\beta}$, že tento odhad je dokonca najlepším nestranným lineárnym odhadom vektora $\boldsymbol{\beta}$. Totiž odhad $\boldsymbol{b}$ je lineárnou funkciou $\hat{\boldsymbol{Y}}$ a preto je tiež najlepším nestranným lineárnym odhadom svojej strednej hodnoty, teda $\boldsymbol{\beta}$.

Minimálna hodnota funkcie $S(\boldsymbol{\beta})$ bude označená

$$RSS = S(\boldsymbol{b}) = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}),$$

a nazývaná *reziduálnym súčtom štvorcov* (z angl. Residual sum of squares). Iné vyjdarenie pre štatistiku RSS popisuje nasledujúca veta.

Veta 2.4. *Pre reziduálny súčet štvorcov platí* $RSS = \boldsymbol{Y'MY}$ *a* $RSS = \boldsymbol{Y'Y - b'X'Y}$.

Dôkaz.

$$\begin{aligned} RSS &= (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = (\boldsymbol{Y} - \boldsymbol{HY})'(\boldsymbol{Y} - \boldsymbol{HY}) = \\ &= (\boldsymbol{MY})'(\boldsymbol{MY}) = \boldsymbol{Y'M'MY} = \boldsymbol{Y'MY} = \\ &= \boldsymbol{Y'(I - H)Y} = \boldsymbol{Y'Y - Y'HY} = \boldsymbol{Y'Y} - (\boldsymbol{H'Y})'\boldsymbol{Y} = \\ &= \boldsymbol{Y'Y - \hat{Y}'Y} = \boldsymbol{Y'Y - b'X'Y}. \end{aligned}$$

□

Veta 2.5. *Náhodná veličina* $s^2 = RSS/(n-k)$ *je nestranným odhadom parametra* σ^2.

Dôkaz. Pretože platí $RSS = \boldsymbol{Y'MY}$, je môžné vyvodiť

$$\mathsf{E}(RSS) = \mathsf{E}(\boldsymbol{Y'MY}) = \mathsf{E}\left[\mathsf{Tr}(\boldsymbol{MYY'})\right] = \mathsf{Tr}\left[\boldsymbol{M}\mathsf{E}\left(\boldsymbol{YY'}\right)\right], \tag{2.14}$$

použitím vzťahu 1.6. Pre variančnú maticu platí vzťah

$$\mathsf{Var}(\boldsymbol{Y}) = \mathsf{E}(\boldsymbol{YY'}) - (\mathsf{E}\boldsymbol{Y})(\mathsf{E}\boldsymbol{Y})' = \mathsf{E}(\boldsymbol{YY'}) - \boldsymbol{X\beta}(\boldsymbol{X\beta})'. \tag{2.15}$$

Zo vzťahu 2.15 je možné vyjadriť $\mathsf{E}(\boldsymbol{YY'}) = \mathsf{Var}(\boldsymbol{Y}) + \boldsymbol{X\beta}(\boldsymbol{X\beta})'$. Použitím predchádzajúceho vzťahu $\mathsf{E}(\boldsymbol{YY'})$ v odvodzovaní $\mathsf{E}(RSS)$ vo vzťahu 2.14 sa pokračuje nasledovne

$$\begin{aligned} \mathsf{Tr}\left[\boldsymbol{M}\mathsf{E}\left(\boldsymbol{YY'}\right)\right] &= \mathsf{Tr}\left[\boldsymbol{M}(\mathsf{Var}(\boldsymbol{Y}) + \boldsymbol{X\beta}(\boldsymbol{X\beta})')\right] = \mathsf{Tr}\left[\boldsymbol{M}(\boldsymbol{I}\sigma^2 + \boldsymbol{X\beta}(\boldsymbol{X\beta})')\right] \\ &= \sigma^2\mathsf{Tr}\left(\boldsymbol{M}\right) + \mathsf{Tr}\left[\boldsymbol{MX\beta}\left(\boldsymbol{X\beta}\right)'\right] = (n-k)\sigma^2 + \mathsf{Tr}\left[\boldsymbol{O}\right] \\ &= (n-k)\sigma^2, \end{aligned}$$

kde bol využitý vyťah $\boldsymbol{MX} = \boldsymbol{O}$. □

Bude zavedené nové označenie pre vektor odhadov $\boldsymbol{u} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{MY} = \boldsymbol{Me}$ náhodných zložiek $\boldsymbol{e}$, ktorý bude nazývaný *vektorom reziduí*. Potom platí tiež iné vyjadrenia pre reziduálny súčet štvorcov, napríklad $RSS = \boldsymbol{u'u} = \|\boldsymbol{u}\|^2$.

Veta 2.6. *Pre vektor reziduí platia vzťahy* $\mathsf{E}(\boldsymbol{u}) = \boldsymbol{0}$ *a* $\mathsf{Var}(\boldsymbol{u}) = \sigma^2\boldsymbol{M}$.

Dôkaz. Na platnosť tejto vety sa dá ľahko nahliadnuť podľa nasledujúcich odvodení

$$\begin{aligned} \mathsf{E}(\boldsymbol{u}) &= \mathsf{E}(\boldsymbol{MY}) = \boldsymbol{M}\mathsf{E}\boldsymbol{Y} = \boldsymbol{MX\beta} = \boldsymbol{0}, \\ \mathsf{Var}(\boldsymbol{u}) &= \mathsf{Var}(\boldsymbol{MY}) = \boldsymbol{M}\mathsf{Var}(\boldsymbol{Y})\boldsymbol{M'} = \sigma^2\boldsymbol{M}, \end{aligned}$$

kde bol použitý vyťah $\boldsymbol{MX} = \boldsymbol{O}$. □

V regresnej matici je v prvom stĺpci často uvažovaný vektor jednotiek. Potom sa formuluje matica $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{X}_1)$, kde matica $\boldsymbol{X_1}$ je typu $n \times r$ $(r = k - 1.)$ Potom sú preformulované koeficienty $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_r)'$ a odhady $\boldsymbol{b} = (b_0, b_1, \ldots, b_r)'$.

V prípade, že bude uvažovaný vektor $\boldsymbol{1} \in \mathcal{M}(\boldsymbol{X})$, zo vzťahu 2.3 z ktorého boli odvodené normálne rovnice (teda metódu najmenších štvorcov) vyplýva

$$\boldsymbol{1}'(\boldsymbol{Y} - \boldsymbol{Xb}) = \boldsymbol{1}'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = 0 \quad \text{a z toho vyplýva} \quad \frac{1}{n}\sum Y_i = \bar{Y} = \frac{1}{n}\sum \hat{Y}_i\,.$$

Potom bude výberový korelačný koeficient $r_{Y,\hat{Y}}$ medzi vektormi $\boldsymbol{Y}$ a $\hat{\boldsymbol{Y}}$ zavedený nasledovne

$$r_{Y,\hat{Y}} = \frac{S_{Y,\hat{Y}}}{\sqrt{S_Y^2 S_{\hat{Y}}^2}} = \frac{\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})/(n-1)}{\sqrt{\sum(Y_i - \bar{Y})^2/(n-1)\sum(\hat{Y}_i - \bar{Y})^2/(n-1)}},$$

kde $S_{Y,\hat{Y}}$ je výberová kovariancia a S_Y^2, $S_{\hat{Y}}^2$ sú výberové rozptyly, podľa zavedenia v úvode.

Korelačný koeficient je dobrým ukazovateľom lineárnej závislosti. Preto pri "dobrej"aproximácii vektora $\boldsymbol{Y}$ vektorom $\hat{\boldsymbol{Y}}$ by bol očakávaný koeficient korelácie blízky 1. K vyhodnoteniu kvality preloženia lineárnym modelom sa najčastejšie používa druhá mocnina tohto korelačného koeficientu. Túto štatistiku je možné vyjadriť pomocou vektorov nasledovne

$$r_{Y,\hat{Y}}^2 = \frac{S_{Y,\hat{Y}}}{\sqrt{S_Y^2 S_{\hat{Y}}^2}} = \frac{\left(\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\right)^2}{\sum(Y_i - \bar{Y})^2 \sum(\hat{Y}_i - \bar{Y})^2} = \frac{\left((\boldsymbol{Y} - \bar{\boldsymbol{Y}})'(\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}})\right)^2}{\left\|\boldsymbol{Y} - \bar{\boldsymbol{Y}}\right\|^2 \left\|\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}\right\|^2}, \tag{2.16}$$

kde bol označený vektor $\bar{\boldsymbol{Y}} = \bar{Y}\mathbf{1}$. Vektor $\bar{\boldsymbol{Y}}$ odpovedá odhadu parametra modelu s regresnou maticou tvorenou len stĺpcom jednotiek. Vtedy je samozrejme najlepším odhadom stredná hodnota. Pre ďalšie odvodzovanie bude označený nový vektor $\boldsymbol{d} = \hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}$, ktorý bude vyjadrovať jednotlivé vzdialenosti týchto odhadov. Viac o tejto problematike (vzdialenosti rôznych odhadov) bude pojednávať kapitola 4. V odvodzovaní $r_{Y,\hat{Y}}^2$ zo vzťahu 2.16 je možné pokračovať takto

$$r_{Y,\hat{Y}}^2 = \frac{((\boldsymbol{d} + \boldsymbol{u})'\boldsymbol{d})^2}{\|\boldsymbol{d} + \boldsymbol{u}\|^2 \|\boldsymbol{d}\|^2} = \frac{(\boldsymbol{d}'\boldsymbol{d} + \boldsymbol{u}'\boldsymbol{d})^2}{\|\boldsymbol{d} + \boldsymbol{u}\|^2 \|\boldsymbol{d}\|^2} = \frac{\left(\|\boldsymbol{d}\|^2 + 0\right)^2}{\|\boldsymbol{d} + \boldsymbol{u}\|^2 \|\boldsymbol{d}\|^2} = \frac{\|\boldsymbol{d}\|^2}{\|\boldsymbol{d} + \boldsymbol{u}\|^2}, \tag{2.17}$$

kde $\boldsymbol{u}'\boldsymbol{d} = \boldsymbol{Y}'\boldsymbol{M}'\boldsymbol{d} = \boldsymbol{u}'\boldsymbol{M}\boldsymbol{d} = 0$ z platnosti $\hat{\boldsymbol{Y}}$, $\bar{\boldsymbol{Y}} \in \mathcal{M}(\boldsymbol{X})$ vyplýva $\boldsymbol{d} \in \mathcal{M}(\boldsymbol{X})$ a teda $\boldsymbol{M}\boldsymbol{d} = \boldsymbol{0}$. Vektory $\boldsymbol{u}$ a $\boldsymbol{d}$ sú na seba kolmé, pretože sa oba nachádzajú v navzájom kolmých podpriestoroch a preto platí vzťah $\langle \boldsymbol{u}, \boldsymbol{d} \rangle = \boldsymbol{u}'\boldsymbol{d} = 0$.

Nech je uvažovaný výraz $\sum_{i=1}^n (Y_i - \bar{Y})^2$, ktorý určuje súčet štvorcov odchyliek v modeli s regresnou maticou $\boldsymbol{X} = \mathbf{1}\,\bar{Y}$. Táto hodnota bude označená TSS (z angl. Total sum of squares) a bude vyjadrená vzťahmi

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

Je teda zrejmé, že v prípade regresnej matice tvorenej len stĺcom jednotiek a odhadu $\hat{\boldsymbol{Y}}$ získaného metódou najmenších štvorcov platí $RSS = TSS$. Toto nové označenie štatistiky TSS je možné využiť v ďalšom odvodzovaní štatistiky $r_{Y,\hat{Y}}^2$ zo vzťahu 2.17

$$r_{Y,\hat{Y}}^2 = \frac{\|\boldsymbol{d}\|^2}{\|\boldsymbol{d} + \boldsymbol{u}\|^2} = \frac{\|\boldsymbol{d} + \boldsymbol{u} - \boldsymbol{u}\|^2}{\|\boldsymbol{d} + \boldsymbol{u}\|^2} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}. \tag{2.18}$$

Ako uvádza Anděl v literatúre ([3], str. 82), k popisu presnosti regresného modelu sa používa *koeficient determinácie* R^2 a *adjustovaný (korigovaný) koeficient determinácie* R_{adj}^2, ktoré sú dané vzorcami

$$R^2 = 1 - \frac{RSS}{TSS}, \quad R_{adj}^2 = 1 - \frac{(n-1)RSS}{(n-r-1)TSS}. \tag{2.19}$$

Vo vzťahu 2.18 je teda ukázané, že koeficient determinácie R^2 je priamo odvodený z druhej mocniny korelačného koeficientu $r^2_{Y,\hat{Y}}$.

Menej matematický pohľad na vzorec 2.19 koeficientu determinácie: Čím je hodnota R^2 bližšia k 1 (resp. hodnota RSS je výrazne menšia ako hodnota TSS), tým lepšie model zachytáva pozorované data. V hodnote adjustovaného koeficientu determinácie sú zohľadňované jednotlivé stupne voľnosti štatistik RSS a TSS.

2.1. Normálny lineárny regresný model

Pre predchádzajúce tvrdenia nebola vyžadovaná podmienka normálneho rozdelenia vektora $\boldsymbol{Y}$. Avšak pre nasledujúce tvrdenia, už bude treba uvažovať normálne rozdelenie, keďže je testovanie hypotéz založené na znalosti rozdelenia dat. Nech je teda predpokladané, že náhodná veličina $\boldsymbol{e}$ má n-rozmerné normálne rozdelenie $\boldsymbol{e} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, z čoho vyplýva $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. Tento model bude ďalej nazývaný *normálny* lineárny regresný model.

Veta 2.7. *Pre normálny lineárny regresný model úplnej hodnosti platí*

$$\boldsymbol{b} \sim \mathsf{N}_k(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X'X})^{-1}).$$

Dôkaz. Vzťah $\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$ je známy. Matica $(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$ je maticu čísel typu $k \times n$, ktorá určuje lineárnu transformáciu vektora z normálneho rozdelenia $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. Parametre pre tento vektor boli určené vo vete 2.2. a preto podľa vzťahu 1.3 platí $\boldsymbol{b} \sim \mathsf{N}_k(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X'X})^{-1})$. □

Veta 2.8. *Pre odhady platí*

$$\hat{\boldsymbol{Y}} \sim \mathsf{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H}), \quad \boldsymbol{u} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{M}).$$

Dôkaz. Pre vektor $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$ platí, že matica čísiel $\boldsymbol{H}$ určuje jeho lineárnu transformáciu z vektora $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$. Preto aj vektor $\hat{\boldsymbol{Y}}$ má normálne rozdelenie podľa vzťahu 1.3. Z výpočtu parametrov

$$\begin{aligned} \mathsf{E}(\hat{\boldsymbol{Y}}) &= \mathsf{E}(\boldsymbol{HY}) = \boldsymbol{H}\mathsf{E}\boldsymbol{Y} = \boldsymbol{HX\beta} = \boldsymbol{X\beta}, \\ \mathsf{Var}(\hat{\boldsymbol{Y}}) &= \mathsf{Var}(\boldsymbol{HY}) = \boldsymbol{H}\mathsf{Var}(\boldsymbol{Y})\boldsymbol{H'} = \sigma^2\boldsymbol{H}, \end{aligned}$$

sa určí normálne rozdelenie $\hat{\boldsymbol{Y}} \sim \mathsf{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H})$. Podobne sa určí rozdelenie aj pre reziduá $\boldsymbol{u}$. Pre reziduá $\boldsymbol{u}$ platí vzťah $\boldsymbol{u} = \boldsymbol{Me}$, kde $\boldsymbol{M}$ je matica čísiel a vektor chýb $\boldsymbol{e}$ má rozdelenie $\boldsymbol{e} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Parametre odhadov $\boldsymbol{u}$ boli určené vo vete 2.6. a z toho vyplýva, že odhady $\boldsymbol{u}$ majú rozdelenie $\mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{M})$. □

Veta 2.9. *Platí* $RSS/\sigma^2 \sim \chi^2(n-k)$.

Dôkaz. Matica $\boldsymbol{N}_{n\times n-k}$ bola spomenutá vyššie, pri spektrálnom rozklade matice $\boldsymbol{M}$. Jej stĺpce boli ortonormálne bázové vektory reziduálneho priestoru $\mathcal{M}(\boldsymbol{X})^\perp$ a spĺňala vzťah $\boldsymbol{NN'} = \boldsymbol{M}$. Nech je uvažovaný náhodný vektor $\boldsymbol{Z} \sim \mathsf{N}_{n-k}(\boldsymbol{0}, \boldsymbol{I}_{n-k})$, potom platí vzťah

$\sigma \boldsymbol{N}\boldsymbol{Z} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{M})$. Je zrejmé (viď ostavec 1.2), že vektor reziduí $\boldsymbol{u}$ a náhodný vektor $\sigma\boldsymbol{N}\boldsymbol{Z}$ majú rovnaké rozdelenie $\mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{M})$. Pretože platí vzťah

$$\sigma^2\boldsymbol{Z}'\boldsymbol{N}'\boldsymbol{N}\boldsymbol{Z} = \sigma^2\boldsymbol{Z}'\boldsymbol{Z} \sim \sigma^2\chi^2(n-k), \text{ vyplýva z toho } \frac{RSS}{\sigma^2} = \frac{\boldsymbol{u}'\boldsymbol{u}}{\sigma^2} \sim \chi^2(n-k).$$

□

Veta 2.10. *Vektor $\boldsymbol{b}$ a veličina s^2 sú nezávislé.*

Dôkaz. Vzťahy $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ a $RSS = \boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y}$ sú známe. Anděl vo vete 4.19. literatúry ([3], str. 69), formuluje kedy kvadratická forma $\boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y}$ a náhodný vektor $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ sú nezávislé. V tomto prípade, predpoklady uvedenej Andělovej vety sú: matica $\boldsymbol{M}$ je pozitívne semidefinitná matica, $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ a zároveň platí

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I})\boldsymbol{M} = \boldsymbol{O}. \tag{2.20}$$

Bolo už ukázané vyššie, pri výpisu spektrálneho rozkladu matice $\boldsymbol{M}$, že matica $\boldsymbol{M}$ je symetrická a vlastné čísla matice $\boldsymbol{M}$ sú rovné buď 0 alebo 1. Preto matica $\boldsymbol{M}$ je pozitívne semidefinitná. Keďže $\boldsymbol{M}$ je symetrická matica a platí $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{O}$, platí tiež vzťah 2.20. Vektor $\boldsymbol{b}$ a hodnota RSS sú teda nezávislé. Z toho vyplýva tiež nezávislosť medzi vektorom $\boldsymbol{b}$ a veličinou s^2. □

2.1.1. Testy hypotéz a intervaly spoľahlivosti

Veta 2.11. *Nech v_{ij} značia prvky matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Potom platí pre náhodnú veličinu*

$$\frac{b_i - \beta_i}{\sqrt{s^2 v_{ii}}} \sim \mathsf{t}(n-k),$$

pre každé $i = 1, \ldots, k$.

Dôkaz. Z vety 2.7 vyplýva $(b_i - \beta_i)/\sqrt{\sigma^2 v_{ii}} \sim \mathsf{N}(0,1)$. Keďže $RSS/\sigma^2 \sim \chi^2(n-k)$ a vektor $\boldsymbol{b}$ s odhadom s^2 sú nezávislé, je možné urobiť záver pre náhodnú veličinu

$$\frac{(b_i - \beta_i)/\sqrt{\sigma^2 v_{ii}}}{\sqrt{\frac{RSS/\sigma^2}{n-k}}} = \frac{(b_i - \beta_i)}{\sqrt{s^2 v_{ii}}} = T_i \sim \mathsf{t}(n-k).$$

A teda, vyššie uvedená štatistika má Studentovo rozdelenie s $n-k$ stupňami voľnosti. □

Pomocou vety 2.11 sa testuje hypotéza $\mathsf{H}_0 : \beta_i = \beta_i^0$ proti alternatívnej hypotéze $\mathsf{H}_1 : \beta_i \neq \beta_i^0$, pre práve jedno $i = 1, \ldots, k$. Hypotéza H_0 je zamietnutá na hladine α, práve keď, platí

$$\frac{|(b_i - \beta_i^0)|}{\sqrt{s^2 v_{ii}}} \geq \mathsf{t}_{1-\frac{\alpha}{2}}(n-k).$$

Potom obojstranný interval spoľahlivosti pre odhad parametra β_i so spoľahlivosťou $1-\alpha$ je interval

$$\left(b_i - s\sqrt{v_{ii}}\,\mathsf{t}_{1-\frac{\alpha}{2}}(n-k),\ b_i + s\sqrt{v_{ii}}\,\mathsf{t}_{1-\frac{\alpha}{2}}(n-k)\right). \tag{2.21}$$

Ekvivalentne sa zamieta nulová hypotéza $\mathsf{H}_0 : \beta_i = \beta_i^0$ na hladine významnosti α, ak hodnota β_i^0 neleží v intervale spoľahlivosti 2.21 so spoľahlivosťou $1-\alpha$.

Najčastejšie sa testuje nulová hypotéza $\mathsf{H}_0 : \beta_i = 0$ a teda sa testuje, či je koeficient β_i významne odlišny od 0. Ak sa nulovú hypotézu H_0 zamietne, znamená to, že $\boldsymbol{Y}$ významne (na hladine α) závisí na i-tom stĺpci regresnej matice $\boldsymbol{X}$. Avšak niekedy je potrebné testovať viacero regresnych koeficientov súčasne. Preto bude uvedená nasledujúca veta, pre ktorú je potrebné rozdeliť vektor regresnych koeficientov $\boldsymbol{\beta}$ na vektor $\boldsymbol{\beta}_1$ s p zložkami a vektor $\boldsymbol{\beta}_2$ s q zložkami s odpovedajúcou maticou $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ takto

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix}, \quad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} \boldsymbol{V}_{p\times p} & \boldsymbol{U} \\ \boldsymbol{U}' & \boldsymbol{W}_{q\times q} \end{pmatrix}.$$

Veta 2.12. *Platí*

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

Dôkaz. Pre marginálne rozdelenie platí $\boldsymbol{b}_2 \sim \mathsf{N}_q(\boldsymbol{\beta}_2, \sigma^2\boldsymbol{W})$. Keďže matica $\boldsymbol{X}$'$\boldsymbol{X}$ je pozitívne definitná, potom aj matica $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ je pozitívne definitná (všetky vlastné čísla sú prevrátené hodnoty vlastných čísiel matice $\boldsymbol{X}'\boldsymbol{X}$ a teda opäť kladné). Potom submatica $\boldsymbol{W}$ je tiež pozitívne definitná a preto existuje rozklad $\boldsymbol{B}\boldsymbol{B}' = \boldsymbol{W}$, s regulárnou maticou $\boldsymbol{B}$. Nech je uvažovaný náhodný vektor $\boldsymbol{\beta_2} + \boldsymbol{B}\boldsymbol{Z} \sim \mathsf{N}_q(\boldsymbol{\beta}_2, \sigma^2\boldsymbol{W})$, kde $\boldsymbol{Z} \sim \mathsf{N}_q(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Potom je možné kvadratickú formu

$$\sigma^{-2}(\boldsymbol{b}_2 - \boldsymbol{\beta}_2)'\boldsymbol{W}^{-1}(\boldsymbol{b}_2 - \boldsymbol{\beta}_2),$$

nahradiť za nasledujúci vzťah bez toho aby sa zmenilo rozdelenie tejto kvadratickej formy, takto

$$\sigma^{-2}(\boldsymbol{B}\boldsymbol{Z})'(\boldsymbol{B}\boldsymbol{B}')^{-1}(\boldsymbol{B}\boldsymbol{Z}) = \sigma^{-2}\boldsymbol{Z}'\boldsymbol{B}'(\boldsymbol{B}\boldsymbol{B}')^{-1}\boldsymbol{B}\boldsymbol{Z}.$$

Ďalej sa ukáže, že platí $\boldsymbol{B}'(\boldsymbol{B}\boldsymbol{B}')^{-1}\boldsymbol{B} = \boldsymbol{I}$. Pri vynásobení zľava maticou $\boldsymbol{B}$, nastáva rovnosť $\boldsymbol{B} = \boldsymbol{B}$. Týmto tvrdením je rovnosť $\boldsymbol{B}'(\boldsymbol{B}\boldsymbol{B}')^{-1}\boldsymbol{B} = \boldsymbol{I}$ dokázaná, len ak matica $\boldsymbol{B}$ je regulárna. Následne platí

$$\sigma^{-2}\boldsymbol{Z}'\boldsymbol{B}'(\boldsymbol{B}\boldsymbol{B}')^{-1}\boldsymbol{B}\boldsymbol{Z} = \sigma^{-2}\boldsymbol{Z}'\boldsymbol{Z} \sim \chi^2(q).$$

Už bola ukázaná platnosť $RSS/\sigma^2 \sim \chi^2(n-k)$ a nezávislosť vektora $\boldsymbol{b}$ s hodnotou RSS, potom platí

$$\frac{\sigma^{-2}(\boldsymbol{b}_2 - \boldsymbol{\beta}_2)'\boldsymbol{W}^{-1}(\boldsymbol{b}_2 - \boldsymbol{\beta}_2)}{RSS/\sigma^2}\,\frac{n-k}{q} = \frac{1}{qs^2}(\boldsymbol{b}_2 - \boldsymbol{\beta}_2)'\boldsymbol{W}^{-1}(\boldsymbol{b}_2 - \boldsymbol{\beta}_2) \sim \mathsf{F}(q, n-k).$$

□

Dôsledok tejto vety má využitie pri testovaní nulovej hypotézu $\mathsf{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$ proti alternativnej hypotéze $\mathsf{H}_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_2^0$, kde $\boldsymbol{\beta}_2^0$ je q-prvkový vektor. Nulová hypotéza H_0 bude zamietnutá na hladine α, práve keď, vektor $\boldsymbol{\beta}_2^0$ nepatrí do konfidenčnej množiny pre $\boldsymbol{\beta}_2$ so spoľahlivosťou $1-\alpha$

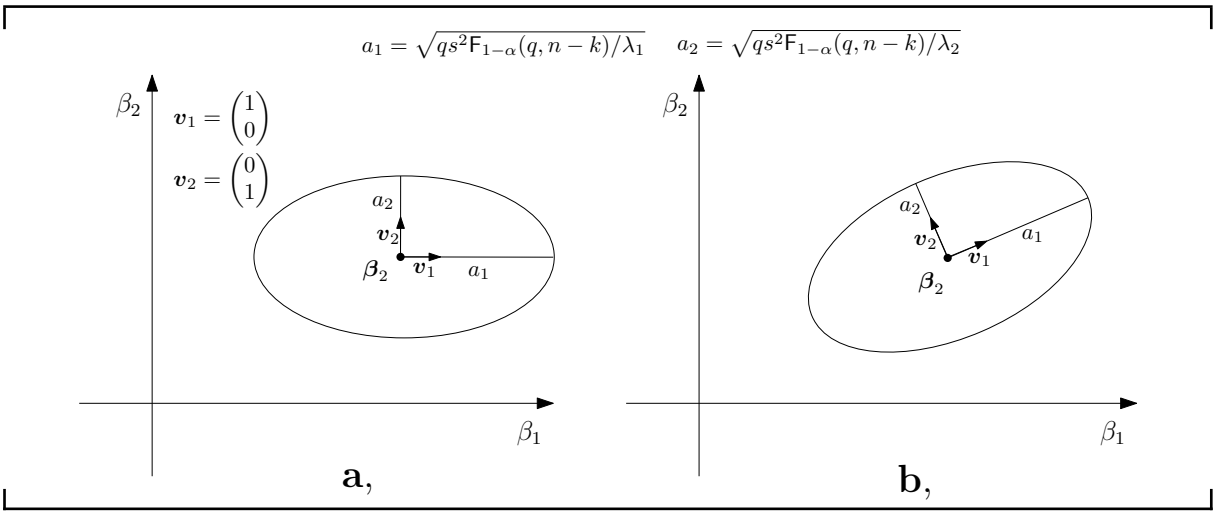
$$\left\{\boldsymbol{b} \in \mathbb{R}^q \,|\, (\boldsymbol{b} - \boldsymbol{\beta}_2)'\boldsymbol{W}^{-1}(\boldsymbol{b} - \boldsymbol{\beta}_2) < qs^2\mathsf{F}_{1-\alpha}(q, n-k)\right\}. \tag{2.22}$$

Pretože je matica $\boldsymbol{W}^{-1}$ pozitívne definitná, konfidenčná množina 2.22 je tvaru elipsoidu[1]so stredom v bode $\boldsymbol{\beta}_2$. Pomocou spektrálneho rozkladu 1.10 matice $\boldsymbol{W}^{-1} = \sum_{i=1}^{q} \lambda_i \boldsymbol{v}_i\boldsymbol{v}_i'$, je možné prepísať nerovnosť v konfidenčnej množine 2.22 nasledovne

$$(\boldsymbol{b} - \boldsymbol{\beta}_2)'\left(\sum_{i=1}^{q} \lambda_i \boldsymbol{v}_i\boldsymbol{v}_i'\right)(\boldsymbol{b} - \boldsymbol{\beta}_2) = \sum_{i=1}^{q} \lambda_i \left((\boldsymbol{b} - \boldsymbol{\beta}_2)'\boldsymbol{v}_i\right)^2 < qs^2\mathsf{F}_{1-\alpha}(q, n-k). \tag{2.23}$$

[1]Myslený je všeobecnejší pojem elipsoidu, zahrňujúci dvojrozmerný (elipsa) a viacrozmerný prípad. V euklidovskom priestore dimenzie n, spĺňujúci $\sum_{j=1}^{n} x_j^2/r_j^2 < 1$, kde x_j sú súradnice bodu a $r_j \in \mathbb{R}$.

Súčin $(\boldsymbol{b}-\boldsymbol{\beta}_2)'\boldsymbol{v}_i$ vyjadruje veľkosť premietaného vektora $(\boldsymbol{b}-\boldsymbol{\beta}_2)$ na priestor generovaný vlastným vektorom $\boldsymbol{v}_i$. V tomto prípade, vlastné čísla $\lambda_1, \ldots, \lambda_q$ figurujú ako váhy sumy uvedenej vo vzťahu 2.23. Nech najväčším vlastným číslom matice $\boldsymbol{W}^{-1}$ je λ_j, potom je elipsoid práve v smere vektora $\boldsymbol{v}_j$ najviac obmedzený. Osami elipsoidu sú vlastné vektory $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$ matice $\boldsymbol{W}^{-1}$. Pre lepšiu predstavu je uvedený obrázok 2, ktorý je uvažovaný pre $q = 2$.



Obrázok 2: Konfidenčná množina v **a**, je prípad kedy matica $\boldsymbol{W}^{-1}$ typu 2×2 je diagonálna a teda vlastné vektory sú rovnobežné s euklidoskou bázou. Konfidenčná množina v **b**, je všeobecnejším prípadom s vlastnými vektormi $\boldsymbol{v}_1$ a $\boldsymbol{v}_2$. V oboch prípadoch je vlastné číslo λ_2 väčšie, a_1 vyjadruje dĺžku hlavnej poloosy a a_2 dĺžku vedľajšej poloosy elípsy.

Väčšinou je nulová hypotéza nastavená $\mathsf{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ proti alternatívnej hypotéze $\mathsf{H}_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{0}$. Teda je testovaná situácia, že posledných q regresných koeficientov v $\boldsymbol{\beta}$ je súčasne nevýznamných pri zvolenej hladine α. V prípade, že je testovaný celý vektor $\mathsf{H}_0 : \boldsymbol{\beta} = \boldsymbol{0}$ proti alternatíve $\mathsf{H}_1 : \boldsymbol{\beta} \neq \boldsymbol{0}$, je pokladané $\boldsymbol{W} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ a je možné odvodiť vzťah

$$\frac{\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}}{ks^2} = \frac{\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}}}{ks^2} = \frac{\boldsymbol{Y}'\boldsymbol{Y} - RSS}{ks^2} \sim \mathsf{F}_{k,n-k}.$$

Je známy vzťah $\boldsymbol{b} \sim \mathsf{N}_k(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ z vety 2.7. Ďalej nech je uvažovaný ľubovoľný vektor $\boldsymbol{c} \in \mathbb{R}^k$. Potom pre odhad lineárnej kombinácie regresných koeficientov $\boldsymbol{c}'\boldsymbol{b}$ platí

$$\begin{aligned} \mathsf{E}(\boldsymbol{c}'\boldsymbol{b}) &= \boldsymbol{c}'\mathsf{E}(\boldsymbol{b}) = \boldsymbol{c}'\boldsymbol{\beta}, \\ \mathsf{Var}(\boldsymbol{c}'\boldsymbol{b}) &= \boldsymbol{c}'\mathsf{Var}(\boldsymbol{b})\boldsymbol{c} = \sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}. \end{aligned}$$

Z definície k-rozmerného normálneho rozdelenia pre náhodný vektor vo vzťahu 1.2, platí $\boldsymbol{c}'\boldsymbol{b} \sim \mathsf{N}(\boldsymbol{c}'\boldsymbol{\beta}, \sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c})$.

Veta 2.13. *Pre nenulový vektor* $\boldsymbol{c} = (c_1, \ldots, c_n)'$ *platí*

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

Dôkaz. Z vety 2.10. vyplýva, že taktiež hodnoty $\boldsymbol{c}'\boldsymbol{b}$ a s^2 sú nezávislé. Potom platí $(\boldsymbol{c}'\boldsymbol{b} - \boldsymbol{c}'\boldsymbol{\beta})/\sqrt{\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}} \sim \mathsf{N}(0,1)$ a z vety 2.9. platí $RSS/\sigma^2 \sim \chi^2(n-k)$. Preto platí

$$\frac{(\boldsymbol{c}'\boldsymbol{b} - \boldsymbol{c}'\boldsymbol{\beta})/\sqrt{\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}}{\sqrt{RSS\sigma^{-2}/(n-k)}} = \frac{\boldsymbol{c}'\boldsymbol{b} - \boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim \mathsf{t}(n-k).$$

Tvrdenie vyplýva z definovania Studentovho rozdelenia. □

Predošlá veta 2.13. je užitočná pri testovaní hypotéz o lineárnych kombináciach regresných koeficientov. Príkladom môže byť odhad individuálnej hodnoty (funkcie strednej hodnoty) v bode $\boldsymbol{c}$. Následne bude uvedený aspoň obojstranný interval spoľahlivosti pre túto hodnotu $\boldsymbol{c}'\boldsymbol{\beta}$ so spoľahlivosťou $1-\alpha$

$$\left(\boldsymbol{c}'\boldsymbol{b} - \mathsf{t}_{1-\frac{\alpha}{2}}(n-k)\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}, \boldsymbol{c}'\boldsymbol{b} - \mathsf{t}_{1-\frac{\alpha}{2}}(n-k)\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}\right).$$

Tvrdenie vety 2.13. už nebude všeobecne platné pre ľubovoľný nenulový vektor $\boldsymbol{c}$, v prípade modelu s neúplnou hodnosťou. Tým sa bude zaoberať nasledujúca kapitola.

3. Lineárny regresný model neúplnej hodnosti

Najprv bude zavedený lineárny regresný model neúplnej hodnosti, rovnako ako model plnej hodnosti 2.1.

Nech $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ je náhodný vektor a $\boldsymbol{X}_{n\times k}$ je matica daných reálnych čísiel, kde zároveň platí $\mathrm{h}(\boldsymbol{X}) = r$ a $r < k < n$. Ďalej nech sa $\boldsymbol{Y}$ riadi lineárnym modelom

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{3.1}$$

keď $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ je vektor neznámych parametrov a $\boldsymbol{e} = (e_1, \ldots, e_n)'$ je náhodný vektor spĺňajúci podmienky

$$\mathsf{E}\boldsymbol{e} = \boldsymbol{0}, \quad \mathsf{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$$

s neznámym parametrom $\sigma^2 > 0$. Bude uvažované rovnaké pomenovanie pre $\boldsymbol{Y}$, $\boldsymbol{X}$ a $\boldsymbol{\beta}$ ako v modeli 2.1. V tomto prípade, keď regresná matica $\boldsymbol{X}$ nemá úplnú stĺpcovú hodnosť, bude model 3.1 nazývaný *modelom s neúplnou stĺpcovou hodnosťou* alebo skrátene *modelom neúplnej hodnosti*.

Tak ako v predchádzajúcej kapitole, je dôležitou úlohou uvažovať také odhady $\boldsymbol{b}$ regresných koeficientov $\boldsymbol{\beta}$ aby minimalizovali funkciu $S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. Je dobré pripomenúť, že na minimalizácii tejto kvadratickej funkcii $S(\boldsymbol{\beta})$ je založená celá metóda najmenších štvorcov. Otázkou je, či tak, ako v modeli úplnej hodnosti stačí splnenie normálnych rovníc 2.6

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y},$$

pre odhady $\boldsymbol{b}$ (minimalizujúcich $S(\boldsymbol{\beta})$). Nasledujúca veto o tomto pojednáva.

Veta 3.1. *Ak matica $\boldsymbol{X}$ nemá úplnú stĺpcovú hodnosť, potom všetky odhady metódou najmenších štvorcov sú tvaru*

$$\boldsymbol{b} = \boldsymbol{b}_p + \boldsymbol{b}_h,$$

kde $\boldsymbol{b}_p = (\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y}$, $\boldsymbol{b}_h \in \mathcal{M}(\boldsymbol{X}'\boldsymbol{X})^\perp$ a $(\boldsymbol{X}'\boldsymbol{X})^-$ je ľubovoľná pseudoinverzia k $\boldsymbol{X}'\boldsymbol{X}$.

Dôkaz. Tento dôkaz bude obdobný dôkazu vety 2.1. Rovnako je uvažovaný nejaký vektor $\boldsymbol{b}$, pre ktorý bude platiť buď $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b} \notin \mathcal{M}(\boldsymbol{X})$ alebo $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}$. Taktiež platia vzťahy $\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{0}$ a následné odvodenie

$$\begin{aligned}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + (\boldsymbol{b} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{\beta}) \\ &\geq (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}).\end{aligned}$$

Keďže matica $\boldsymbol{X}'\boldsymbol{X}$ je singulárna, tak je pozitívne semidefinitná. Riešenie z rovníc $\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{0}$ teda minimalizuje funkciu $S(\boldsymbol{\beta})$ na minimum $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})$. Predchádzajúcu sústavu lineárnych rovníc je možné upraviť na tvar $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$. Avšak kvôli neúplnej hodnosti $\boldsymbol{X}$, nie je riešenie $\boldsymbol{b}$ jediné a dá sa rozdeliť na partikulárne $\boldsymbol{b}_p = (\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y}$ a homogénne riešenie $\boldsymbol{b}_h \in \mathcal{M}(\boldsymbol{X}'\boldsymbol{X})^\perp$. Totiž pre ľubovoľné riešenie partikulárne riešenie $\boldsymbol{b}_p = (\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y}$ a homogénne riešenie $\boldsymbol{b}_h \in \mathcal{M}(\boldsymbol{X}'\boldsymbol{X})^\perp$ platia vzťahy

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_p = \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y} \quad \text{a} \quad \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_h = \boldsymbol{0}.$$

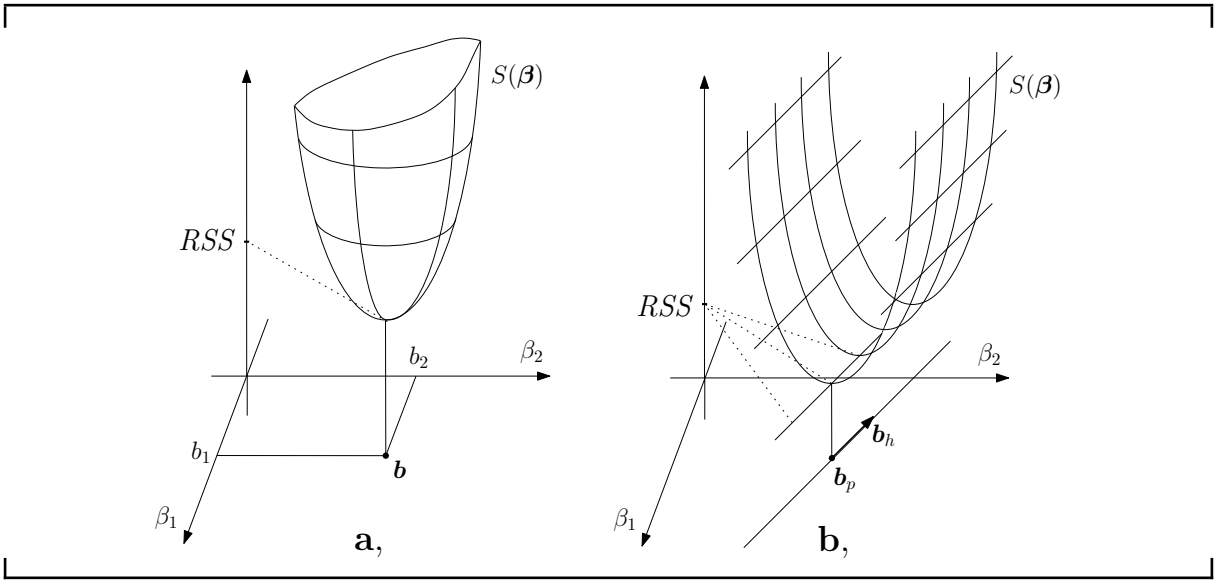
□

Je potreba pár slovami doplniť záverečné myšlienky dôkazu. Ak je matica $\boldsymbol{X}'\boldsymbol{X}$ singulárna, výpočet inverznej matice k nej nie je možný. Pre riešenie sústavy lineárnych rovníc $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$, so singulárnou maticou sústavy, bola v teórii matic zavedená pseudoinverzná matica (ako bolo uvedené v úvode v časti o pseudoiverznej matici). V tomto prípade, partikulárnym riešením normálnych rovníc sú

$$\boldsymbol{b}_p = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}.$$

V úvode bolo poznamenané, že pseudoinverzná matica $(\boldsymbol{X}'\boldsymbol{X})^{-}$ nie je jednoznačne daná pre singulárnu maticu $\boldsymbol{X}'\boldsymbol{X}$. V tomto prípade, je partikulárnych riešení $\boldsymbol{b}_p$ nekonečne veľa, ktoré sa líšia o vektor homogénneho riešenia $\boldsymbol{b}_h$. Preto ak matica $\boldsymbol{X}$ nemá úplnú stĺpcovú hodnosť, májú normálne rovnice vždy nekonečne veľa riešení tvaru $\boldsymbol{b} = \boldsymbol{b}_p + \boldsymbol{b}_h$ ako je znázornené v obrázku 3.

Minimálna hodnota funkcie $S(\boldsymbol{b})$, pre nejaké riešenie $\boldsymbol{b}$ normálnych rovníc, bude nazývaná *reziduálny súčet štvorcov* a označená *RSS* (tak ako tomu bolo v predchádzajúcej kapitole).



Obrázok 3: V obrázku **a**, je zobrazená funkcia $S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ pre rôzne odhady $\boldsymbol{\beta}$, v prípade lineárneho modelu úplnej hodnosti. V obrázku **b**, zobrazená táto funkcia pre model neúplnej hodnosti. Tu $\boldsymbol{b}_p$ vyjadruje nejaké partikulárne riešenie normálnych rovníc $\boldsymbol{b}_p = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}$.

Geometrická predstava: bez úplnej hodnosti regresnej matice, sú niektoré stĺpce matice $\boldsymbol{X}$ lineárne závislé a práve preto nie je možné jednoznačne vyjadriť odhad vektora $\boldsymbol{\beta}$. V takýchto prípadoch uľahčuje prácu s odhadmi, v modeloch neúplnej hodnosti, pojem odhadnuteľnosť parametrickej funkcie.

3.1. Odhadnuteľné parametrické funkcie

V prípade úplnej stĺpcovej hodnosti regresnej matice je odhad metódou najmenších štvorcov parametra $\boldsymbol{t}'\boldsymbol{\beta}$ rovný $\boldsymbol{t}'\boldsymbol{b}$ pre ľubovoľný vektor $\boldsymbol{t}$ daného typu. Vektor $\boldsymbol{b}$ je totiž jediným riešením sústavy normálnych rovníc. No otázkou je, ako je to v prípade nekonečného počtu

riešení vektora koeficientov $\boldsymbol{b}$ sústavy normálnych rovníc.

Nech je daný nejaký nenáhodný vektor $\boldsymbol{t} \in \mathbb{R}^k$. Potom lineárna funkcia $\boldsymbol{t}'\boldsymbol{\beta}$ bude nazývaná *odhadnuteľnou*, ak existuje jej nestranný lineárny odhad.

Veta 3.2. *Lineárna funkcia $\boldsymbol{t}'\boldsymbol{\beta}$ je odhadnuteľná práve vtedy, keď platí*

$$\boldsymbol{t} \in \mathcal{M}(\boldsymbol{X}') = \mathcal{M}(\boldsymbol{X}'\boldsymbol{X}). \tag{3.2}$$

Dôkaz. Aby pre nejaký vektor $\boldsymbol{s} \in \mathbb{R}^n$ bola lineárna funkcia $\boldsymbol{s}'\boldsymbol{Y}$ nestranným odhadom $\boldsymbol{t}'\boldsymbol{\beta}$, musí pre každé $\boldsymbol{\beta} \in R^k$ platiť

$$\mathsf{E}(\boldsymbol{s}'\boldsymbol{Y}) = \boldsymbol{s}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{t}'\boldsymbol{\beta}. \tag{3.3}$$

Odkadiaľ $\boldsymbol{t} = \boldsymbol{X}'\boldsymbol{s}$. Preto ak je lineárna funkcia $\boldsymbol{t}'\boldsymbol{\beta}$ odhadnuteľnou, potom $\boldsymbol{t} \in \mathcal{M}(\boldsymbol{X}')$ (resp. $\boldsymbol{t} \in \mathcal{M}(\boldsymbol{X}'\boldsymbol{X})$). A obrátene, ak $\boldsymbol{t} \in \mathcal{M}(\boldsymbol{X}')$ potom musí existovať vektor $\boldsymbol{s} \in \mathbb{R}^n$ taký, že platí $\boldsymbol{t} = \boldsymbol{X}'\boldsymbol{s}$. Potom platí vzťah 3.3, a preto je lineárna funkcia $\boldsymbol{t}'\boldsymbol{\beta}$ odhadnuteľná. □

Veta 3.3. *Ak je lineárna funkcia $\boldsymbol{t}'\boldsymbol{\beta}$ odhadnuteľnou, potom $\boldsymbol{t}'\boldsymbol{b}$ je nestranným lineárnym odhadom funkcie $\boldsymbol{t}'\boldsymbol{\beta}$, kde $\boldsymbol{b}$ je ľubovoľné riešenie normálnych rovníc. Zároveň je hodnota $\boldsymbol{t}'\boldsymbol{b}$ rovnaká pre všetky riešenia $\boldsymbol{b}$ normálnych rovníc.*

Dôkaz. Najprv bude ukázaný dôkaz druhého tvrdenia. Nech $\boldsymbol{b} = \boldsymbol{b}_p + \boldsymbol{b}_h$ a $\boldsymbol{b}^* = \boldsymbol{b}_p^* + \boldsymbol{b}_h^*$ sú rôzne riešenia normálnych rovníc vyjadrené podľa vety 3.1. Aby platilo $\boldsymbol{t}'\boldsymbol{b} = \boldsymbol{t}'\boldsymbol{b}^*$, musí ekvivalentne platiť $\boldsymbol{t}'(\boldsymbol{b} - \boldsymbol{b}^*) = 0$. Platí vzťah

$$\boldsymbol{t}'(\boldsymbol{b} - \boldsymbol{b}^*) = \boldsymbol{t}'(\boldsymbol{b}_p - \boldsymbol{b}_p^*) + \boldsymbol{t}'(\boldsymbol{b}_h - \boldsymbol{b}_h^*) = \boldsymbol{t}'(\boldsymbol{b}_p - \boldsymbol{b}_p^*), \tag{3.4}$$

pretože $\boldsymbol{b}_h,\ \boldsymbol{b}_h^* \in \mathcal{M}(\boldsymbol{X}'\boldsymbol{X})^\perp$ (z čoho vyplýva $(\boldsymbol{b}_h - \boldsymbol{b}_h^*) \in \mathcal{M}(\boldsymbol{X}'\boldsymbol{X})^\perp$) a zároveň $\boldsymbol{t} \in \mathcal{M}(\boldsymbol{X}'\boldsymbol{X})$. Nech $\boldsymbol{b}_p = (\boldsymbol{X}'\boldsymbol{X})_1^-\boldsymbol{X}'\boldsymbol{Y}$ a $\boldsymbol{b}_p^* = (\boldsymbol{X}'\boldsymbol{X})_2^-\boldsymbol{X}'\boldsymbol{Y}$ podľa vety 3.1., kde sú uvedené dve rôzne pseudoinverzné matice (označené indexami 1 a 2) k matici $\boldsymbol{X}'\boldsymbol{X}$. Anděl vo vete 15.c, v literatúre ([2], str. 69) uviedol, že matica

$$\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}' \tag{3.5}$$

je nezávislá na zvolenej pseudoinverzii $(\boldsymbol{X}'\boldsymbol{X})^-$. Použitím predchádzajúceho tvrdenia a vzťahu $\boldsymbol{t} = \boldsymbol{X}'\boldsymbol{s}$ (z odhadnuteľnosti $\boldsymbol{t}'\boldsymbol{\beta}$) je možné ďalej pokračovať v úprave vzťahu 3.4 takto

$$\boldsymbol{t}'(\boldsymbol{b}_p - \boldsymbol{b}_p^*) = \boldsymbol{s}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})_1^-\boldsymbol{X}'\boldsymbol{Y} - (\boldsymbol{X}'\boldsymbol{X})_2^-\boldsymbol{X}'\boldsymbol{Y}] = \boldsymbol{s}'[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})_1^-\boldsymbol{X}' - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})_2^-\boldsymbol{X}']\boldsymbol{Y} = 0.$$

Z toho vyplýva rovnosť $\boldsymbol{t}'(\boldsymbol{b} - \boldsymbol{b}^*) = 0$ resp. $\boldsymbol{t}'\boldsymbol{b} - \boldsymbol{t}'\boldsymbol{b}^* = 0$, pre ľubovoľné riešenia normálnych rovníc. Nasleduje dôkaz prvého tvrdenia. Použitím vzťahu $\boldsymbol{t} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{g}$ (z odhadnuteľnosti $\boldsymbol{t}'\boldsymbol{\beta}$) je možné vyjadriť nasledujúci vzťah

$$\mathsf{E}(\boldsymbol{t}'\boldsymbol{b}) = \mathsf{E}(\boldsymbol{g}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}) = \mathsf{E}(\boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}) = \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{t}'\boldsymbol{\beta}, \tag{3.6}$$

kde v druhej rovnosti bola využitá platnosť normálnych rovníc $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$ (keďže platí pre každé $\boldsymbol{b}$ spĺňajúce normálne rovnice). Zo vzťahu 3.6 vyplýva, že náhodná veličina $\boldsymbol{t}'\boldsymbol{b}$ je nestranným lineárnym odhadom funkcie $\boldsymbol{t}'\boldsymbol{\beta}$ (lineárnosť podľa vzťahu 2.11). □

Keďže pri úplnej hodnosti $\boldsymbol{X}$ je jednoznačne určené riešenie $\boldsymbol{b}$ normálnej rovnice, potom pre ľubovoľnú lineárnu transformáciu $\boldsymbol{t}'$ je funkcia $\boldsymbol{t}'\boldsymbol{\beta}$ odhadnuteľná.

Veta 3.4. *Ak je parametrická funkcia $\boldsymbol{t}'\boldsymbol{\beta}$ odhadnuteľná, potom $\boldsymbol{t}'\boldsymbol{b}$ je najlepší nestranný lineárny odhad parametrickej funkcie $\boldsymbol{t}'\boldsymbol{\beta}$.*

Dôkaz. Nestrannosť a lineárnosť odhadu $\boldsymbol{t}'\boldsymbol{b}$ bola ukazaná vo vete 3.3. Z odhadnuteľnosti funkcie $\boldsymbol{t}'\boldsymbol{\beta}$, vyplýva existencia $\boldsymbol{g} \in \mathbb{R}^k$ tak, že platí $\boldsymbol{t} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{g}$. Potom platí

$$\boldsymbol{t}'\boldsymbol{b} = \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}, \tag{3.7}$$

kde v druhej rovnosti boli použité normálne rovnice (a ich ľubovoľné riešenie $\boldsymbol{b}$). Zo vzťahu 3.7 vyplýva, že pre každý lineárny nestranný odhad $\boldsymbol{v}'\boldsymbol{Y}$ platí $\boldsymbol{v}'\boldsymbol{X}\boldsymbol{\beta} = \mathsf{E}(\boldsymbol{v}'\boldsymbol{Y}) = \boldsymbol{t}'\boldsymbol{\beta}$, z čoho vyplýva $(\boldsymbol{v}'\boldsymbol{X} - \boldsymbol{t}')\boldsymbol{\beta} = 0$ pre $\forall\boldsymbol{\beta}$ (teda $\boldsymbol{v}'\boldsymbol{X} = \boldsymbol{t}'$). Využitím vzťahu 3.7 sa vyjadrí rozptyl funkcie $\boldsymbol{v}'\boldsymbol{Y}$ nasledovne

$$\begin{aligned}
\mathsf{D}(\boldsymbol{v}'\boldsymbol{Y}) &= \mathsf{D}[\boldsymbol{v}'\boldsymbol{Y} - \boldsymbol{t}'\boldsymbol{b} + \boldsymbol{t}'\boldsymbol{b}] = \mathsf{D}[(\boldsymbol{v}'\boldsymbol{Y} - \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}) + \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}] = \\
&= \mathsf{D}(\boldsymbol{v}'\boldsymbol{Y} - \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}) + \mathsf{D}(\boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}) + 2\mathsf{Cov}(\boldsymbol{v}'\boldsymbol{Y} - \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}, \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}) = \\
&= \mathsf{D}(\boldsymbol{v}'\boldsymbol{Y} - \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}) + \mathsf{D}(\boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}) = \mathsf{D}(\boldsymbol{v}'\boldsymbol{Y} - \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}) + \mathsf{D}(\boldsymbol{t}'\boldsymbol{b}),
\end{aligned}$$

keďže platí rovnosť

$$\mathsf{Cov}(\boldsymbol{v}'\boldsymbol{Y} - \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}, \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}) = (\boldsymbol{v}' - \boldsymbol{g}'\boldsymbol{X}')(\boldsymbol{I}\sigma^2)\boldsymbol{X}\boldsymbol{g} = \sigma^2(\boldsymbol{v}'\boldsymbol{X} - \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{X})\boldsymbol{g} = \sigma^2(\boldsymbol{t}' - \boldsymbol{t}')\boldsymbol{g} = 0.$$

Aby odhad $\boldsymbol{t}'\boldsymbol{b}$ bol najlepším nestranným lineárnym odhadom, musí platiť, že hodnota $\mathsf{D}(\boldsymbol{v}'\boldsymbol{Y}) - \mathsf{D}(\boldsymbol{t}'\boldsymbol{b}) = \mathsf{D}(\boldsymbol{v}'\boldsymbol{Y} - \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y})$ je nezáporná. Je zrejmé, že platí

$$\mathsf{D}(\boldsymbol{v}'\boldsymbol{Y} - \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}) = (\boldsymbol{v}' - \boldsymbol{\xi}'\boldsymbol{X}')\mathsf{Var}(\boldsymbol{Y})(\boldsymbol{v} - \boldsymbol{X}\boldsymbol{g}) = \sigma^2(\boldsymbol{v} - \boldsymbol{\xi}\boldsymbol{X})'(\boldsymbol{v} - \boldsymbol{X}\boldsymbol{g}),$$

a pretože je táto kvadratická forma tvaru $\sigma^2\boldsymbol{w}'\boldsymbol{w}$ (kde $\boldsymbol{w} = (\boldsymbol{v} - \boldsymbol{X}\boldsymbol{g})$) je pre $\boldsymbol{w} \neq \boldsymbol{0}$ vždy kladná. Preto je odhad $\boldsymbol{t}'\boldsymbol{b}$ najlepším z iných nestranných lineárnych odhadov. □

Je dôležité si uvedomiť, že v prípade modelu neúplnej hodnosti, je jednoznačný odhad metódou najmenších štvorcov iba pre strednú hodnotu v bode $\boldsymbol{t} \in \mathcal{M}(\boldsymbol{X})$ (teda $\boldsymbol{t}'\boldsymbol{b}$ je odhadom $\boldsymbol{t}'\boldsymbol{\beta}$). Nech je uvažovaný lineárny model úplnej hodnosti s regresnou maticou $\boldsymbol{X}_1$ a druhý lineárny model neúplnej hodnosti s regresnou maticou $\boldsymbol{X}_2 = (\boldsymbol{X}_1, \boldsymbol{x}_3)$, kde pre vektor $\boldsymbol{x}_3$ platí $\boldsymbol{x}_3 \in \mathcal{M}(\boldsymbol{X}_1)$. Oba modely generujú rovnaké regresné priestory $\mathcal{M}(\boldsymbol{X}_1) = \mathcal{M}(\boldsymbol{X}_2)$ a preto majú aj ortogonálny priemet na tento priestor totožný (ilustruje obrázok 4). Ako bolo už uvedené, v modeli neúplnej hodnosti nie je možné jednoznačne vyjadriť tento priemet (odhadmi regresných koeficientov) ako je tomu v modeli úplnej hodnosti.

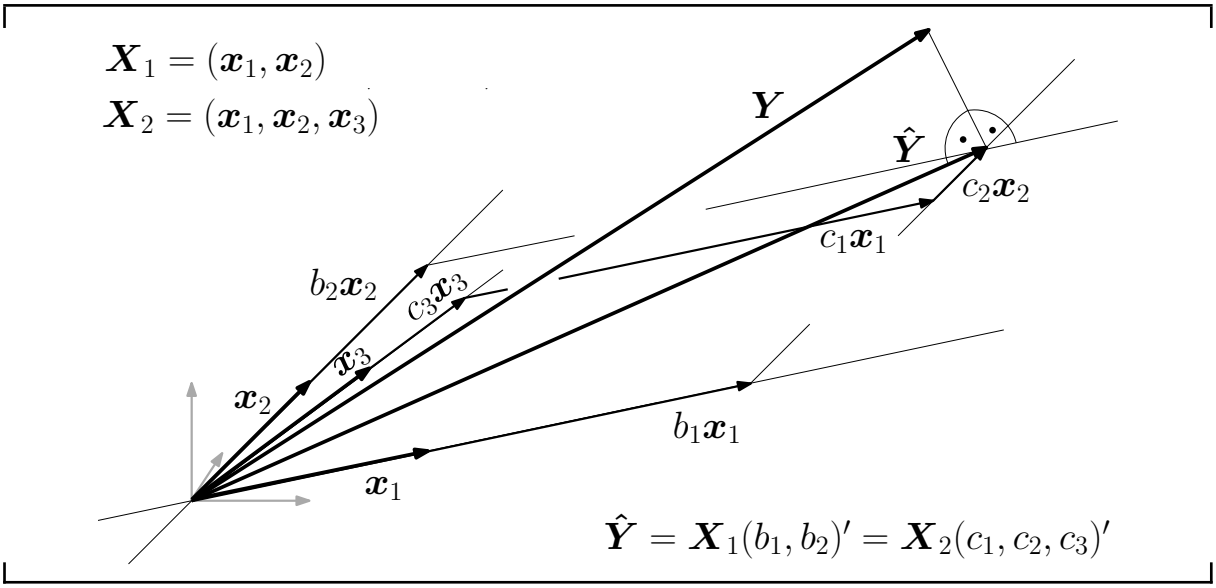
Veta 3.5. *Nech $\boldsymbol{t}'\boldsymbol{b}$ a $\boldsymbol{t}_1'\boldsymbol{b}$ sú odhady metódou najmenších štvorcov pre odhadnuteľné funkcie $\boldsymbol{t}'\boldsymbol{\beta}$ a $\boldsymbol{t}_1'\boldsymbol{\beta}$. Potom platí*

$$\mathsf{D}(\boldsymbol{t}'\boldsymbol{b}) = \sigma^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}, \quad \mathsf{Cov}(\boldsymbol{t}'\boldsymbol{b}, \boldsymbol{t}_1'\boldsymbol{b}) = \sigma^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}_1, \tag{3.8}$$

kde matica $(\boldsymbol{X}'\boldsymbol{X})^-$ je ľubovoľná pseudoinverzná matica k $\boldsymbol{X}'\boldsymbol{X}$.

Dôkaz. Z odhadnuteľnosti parametrickej funkcie $\boldsymbol{t}'\boldsymbol{\beta}$ vyplýva $\boldsymbol{t} \in \mathcal{M}(\boldsymbol{X}'\boldsymbol{X})$. Preto existuje $\boldsymbol{g}$ také že platí $\boldsymbol{t} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{g}$. Nech je uvažovaná pseudoinverzná matice $(\boldsymbol{X}'\boldsymbol{X})^-$ k matici $\boldsymbol{X}'\boldsymbol{X}$. Potom z definície pseudoinverznej matice platí rovnosť $\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{X}$. Transponovaním tohto vzťahu sa vyjadrí rovnosť

$$\boldsymbol{X}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^-]'\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{X}. \tag{3.9}$$



Obrázok 4: Obrázok ilustruje prípad, kedy regresné matice modelu úplnej a neúplnej hodnosti vytvárajú totožné priestory $\mathcal{M}(\boldsymbol{X}_1) = \mathcal{M}(\boldsymbol{X}_2)$.

Použitím poslednej rovnosti 3.9 a vzťahu $\boldsymbol{t} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{g}$ platí

$$\begin{aligned} \mathsf{D}(\boldsymbol{t}'\boldsymbol{b}) &= \mathsf{D}(\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y}) = \boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\mathsf{Var}(\boldsymbol{Y})\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^-]'\boldsymbol{t} = \\ &= \sigma^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^-]'\boldsymbol{t} = \\ &= \sigma^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^-]'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{g} = \\ &= \sigma^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}\boldsymbol{g} = \sigma^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}. \end{aligned} \tag{3.10}$$

Podľa vzťahu 3.7 odhadu $\boldsymbol{t}'\boldsymbol{b}$, odhadnuteľnej funkcie $\boldsymbol{t}'\boldsymbol{\beta}$, metódou najmeších štvorcov platí vzťah $\boldsymbol{t}'\boldsymbol{b} = \boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}$. Obdobne je možné získať vyjadrenie odhadu pre odhadnuteľnú funkciu $\boldsymbol{t}_1'\boldsymbol{\beta}$, kde $\boldsymbol{t}_1 = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{g}_1$. Potom pre kovarianciu odhadov $\boldsymbol{t}'\boldsymbol{b}$ a $\boldsymbol{t}_1'\boldsymbol{b}$ platí

$$\begin{aligned} \mathsf{Cov}(\boldsymbol{t}'\boldsymbol{b}, \boldsymbol{t}_1'\boldsymbol{b}) &= \mathsf{Cov}(\boldsymbol{g}'\boldsymbol{X}'\boldsymbol{Y}, \boldsymbol{g}_1'\boldsymbol{X}'\boldsymbol{Y}) = \boldsymbol{g}'\boldsymbol{X}'\mathsf{Var}(\boldsymbol{Y})\boldsymbol{X}\boldsymbol{g}_1 = \sigma^2\boldsymbol{g}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{g}_1 = \\ &= \sigma^2\boldsymbol{g}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}\boldsymbol{g}_1 = \sigma^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}_1. \end{aligned} \tag{3.11}$$

□

V predchádzajúcom dôkaze je možné si povšimnúť, že rozptyl odhad $\boldsymbol{t}'\boldsymbol{b}$ by bolo možné vyjadriť bez použitia psudoinverznej matice vo vzťahu 3.10, takto

$$\mathsf{D}(\boldsymbol{t}'\boldsymbol{b}) = \sigma^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t} = \sigma^2\boldsymbol{g}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}\boldsymbol{g} = \boldsymbol{g}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{g} = \boldsymbol{g}'\boldsymbol{t}.$$

Obdobne v odvodzovaní 3.11, bola kovariancia vyjadrená bez použitia pseudoinverznej matice nasledovne

$$\mathsf{Cov}(\boldsymbol{t}'\boldsymbol{b}, \boldsymbol{t}_1'\boldsymbol{b}) = \sigma^2\boldsymbol{g}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{g}_1 = \sigma^2\boldsymbol{g}'\boldsymbol{t}_1 = \sigma^2\boldsymbol{t}'\boldsymbol{g}_1.$$

Nech je uvažovaných m vektorov $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_m$, ktoré sú z priestoru $\mathcal{M}(\boldsymbol{X}')$. Potom sú linárne parametrické funkcie $\boldsymbol{t}_1'\boldsymbol{\beta}, \boldsymbol{t}_2'\boldsymbol{\beta}, \ldots, \boldsymbol{t}_m'\boldsymbol{\beta}$ odhadnuteľné a je možné ich zapisať maticovo $\boldsymbol{T}'\boldsymbol{\beta}$, kde $\boldsymbol{T} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_m)$. Nech je takto zvolená matica $\boldsymbol{T}$, potom pre vektor $\boldsymbol{T}'\boldsymbol{\beta}$ je vhodný názov - vektor odhadnuteľných parametrických funkcií. Ako príklad

je možné uviesť samotnú maticu $\boldsymbol{X}'$. Každý stĺpec matice $\boldsymbol{X}'$ je z priestoru $\mathcal{M}(\boldsymbol{X}')$ a preto vektor parametrckých funkcií $\boldsymbol{X\beta}$ je vždy odhadnuteľný. Preto najlepším odhadom vektoru $\boldsymbol{Y}$ metódou najmenších štvorcov je

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}. \tag{3.12}$$

Opäť je možné označiť maticou $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'$ a uvažovať, že je nejakou transformáciou (zohľadňujúcou metódu najmenších štvorcov pri vektore $\boldsymbol{Y}$). Anděl vo vete 15. v literatúre ([2], str. 69) uviedol, že platí vzťah

$$\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-}\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{A}, \tag{3.13}$$

kde teda matica $(\boldsymbol{A}'\boldsymbol{A})^{-}\boldsymbol{A}'$ je pseudoinverzná k matici $\boldsymbol{A}$. Ďalej profesor uvádza v tejto vete, že matica

$$\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-}\boldsymbol{A}' \tag{3.14}$$

je symetrická a táto vlastnosť nezávisí na zvolenej pseudoinverzii $(\boldsymbol{A}'\boldsymbol{A})^{-}$. Preto je matica $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'$ symetrická. A platí pre túto maticu aj idempotentnosť

$$\boldsymbol{H}\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}' = \boldsymbol{H},$$

použitím vzťahu 3.13 uvedeného vyššie. A keďže matica $\boldsymbol{H}$ je symerická aj idempotentná, podľa vety 1.8 je matica $\boldsymbol{H}$ operátorom ortogonálnej projekcie na regresný priestor $\mathcal{M}(\boldsymbol{X})$. Z idempotentnosti matice $\boldsymbol{H}$ vyplýva, že $\boldsymbol{I} - \boldsymbol{H}$ je tiež idempotentná. A keďže matica $\boldsymbol{H}$ je symetrická, je zrejmé že, aj matica $\boldsymbol{I} - \boldsymbol{H}$ je symetrická. Preto matica $\boldsymbol{I} - \boldsymbol{H}$ je projekčnou maticou, podľa vety 1.8, ale na reziduálny priestor $\mathcal{M}(\boldsymbol{X})^{\perp}$ (podľa vety 1.7). Táto projekčná matica tak, ako v predchádzajúcej kapitole, bude označovaná $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{H}$.

Z vyššie uvedeného, náhodnú veličinu RSS, je možné vyjadriť nasledovnými vzťahmi

$$RSS = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{u}'\boldsymbol{u}.$$

Veta 3.6. $\mathsf{E}(RSS) = (n-r)\sigma^2$, *takže nestranným odhadom parametra* σ^2 *je*

$$s^2 = \frac{RSS}{(n-r)},$$

kde r *označuje hodnosť matice* $\boldsymbol{X}$.

Dôkaz. Odvodenie je podobné ako v dôkaze vety 2.5. Ak platí $\mathrm{h}(\boldsymbol{X}) = r$, potom platí $\mathrm{h}(\boldsymbol{H}) = r$. Podľa vety 1.5. pre idempotentnú maticu $\boldsymbol{M}$ platí

$$\mathrm{h}(\boldsymbol{M}) = \mathrm{Tr}(\boldsymbol{M}) = \mathrm{Tr}(\boldsymbol{I} - \boldsymbol{H}) = \mathrm{Tr}(\boldsymbol{I}) - \mathrm{Tr}(\boldsymbol{H}) = n - \mathrm{h}(\boldsymbol{H}) = n - r.$$

Zo vzťahu 2.15 vyplýva $\mathsf{E}(\boldsymbol{Y}\boldsymbol{Y}') = \mathsf{Var}(\boldsymbol{Y}) + \boldsymbol{X\beta}(\boldsymbol{X\beta})'$, aj pre model neúplnej hodnosti. Táto rovnosť bude použitá v nasledujúcom vyjadrovaní strednej hodnoty RSS

$$\begin{aligned}
\mathsf{E}(RSS) &= \mathsf{E}(\boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y}) = \mathsf{E}\left[\mathsf{Tr}(\boldsymbol{M}\boldsymbol{Y}\boldsymbol{Y}')\right] = \mathsf{Tr}\left[\boldsymbol{M}\mathsf{E}\left(\boldsymbol{Y}\boldsymbol{Y}'\right)\right] = \\
&= \mathsf{Tr}\left[\boldsymbol{M}(\mathsf{Var}(\boldsymbol{Y}) + \boldsymbol{X\beta}(\boldsymbol{X\beta})')\right] = \mathsf{Tr}\left[\boldsymbol{M}(\boldsymbol{I}\sigma^2 + \boldsymbol{X\beta}(\boldsymbol{X\beta})')\right] = \\
&= \sigma^2\mathsf{Tr}\left(\boldsymbol{M}\right) + \mathsf{Tr}\left[\boldsymbol{M}\boldsymbol{X\beta}(\boldsymbol{X\beta})'\right] = (n-r)\sigma^2 + (\boldsymbol{X\beta})'\boldsymbol{M}\boldsymbol{X\beta} = \\
&= (n-r)\sigma^2,
\end{aligned}$$

kde boli použité vzťahy 1.6 a 2.15. Posledná rovnosť vyplýva z platnosti $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$. □

3.2. Normálny lineárny regresný model

Nech je uvažovaný $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ a matica $\boldsymbol{T}$ typu $k \times p$, ktorej stĺpce sú z priestoru $\mathcal{M}(\boldsymbol{X}')$ a spĺňa $\mathrm{h}(\boldsymbol{T}) = p$. Teda nebudú zbytočne uvažované nadbytočné zložky vektora parametrických funkcií $\boldsymbol{T}'\boldsymbol{\beta}$, pretože parametrické funkcie $\boldsymbol{t}_i'\boldsymbol{\beta}$ sú lineárne a ich kombináciami je možné získať ľubovoľný odhad funkcie $\boldsymbol{t}'\boldsymbol{\beta}$, kde $\boldsymbol{t} \in \mathcal{M}(\boldsymbol{T})$. Keďže každý stĺpce matice $\boldsymbol{T}$ je z priestoru $\mathcal{M}(\boldsymbol{X}')$, vektor parametrických funkcií $\boldsymbol{T}'\boldsymbol{\beta}$ je odhadnuteľný. Z odhadnuteľnosti vyplýva, že existuje matica $\boldsymbol{J}_{k\times p}$ taká, že platí

$$\boldsymbol{T} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{J}. \tag{3.15}$$

Aby platilo $\mathrm{h}(\boldsymbol{T}) = p$, vo vzťahu 3.15 musí platiť $\mathrm{h}(\boldsymbol{J}) = p$ a $\mathrm{h}(\boldsymbol{X}\boldsymbol{J}) = p$. Bodovým odhadom $\boldsymbol{T}'\boldsymbol{\beta}$ metódou najmenších štvorcov je vektor

$$\boldsymbol{T}'\boldsymbol{b} = \boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y}, \tag{3.16}$$

pretože homogénne riešenie normálnych rovníc je ortogonálne na priestor $\mathcal{M}(\boldsymbol{X}')$ a teda aj na $\mathcal{M}(\boldsymbol{T})$. Použitím vzťahu 3.15 sa ukáže nestrannosť odhadu $\boldsymbol{T}'\boldsymbol{b}$ nasledovne

$$\begin{aligned} \mathsf{E}(\boldsymbol{T}'\boldsymbol{b}) &= \boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \\ &= \boldsymbol{J}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{J}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{T}'\boldsymbol{\beta}. \end{aligned} \tag{3.17}$$

Variančná matica $\mathsf{Var}(\boldsymbol{T}'\boldsymbol{b})$ sa vyjadrí nasledovne

$$\begin{aligned} \mathsf{Var}(\boldsymbol{T}'\boldsymbol{b}) &= \mathsf{Var}(\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y}) = \sigma^2\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^-]'\boldsymbol{T} = \\ &= \sigma^2\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^-]'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{J} = \\ &= \sigma^2\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}\boldsymbol{J} = \sigma^2\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{T}, \end{aligned} \tag{3.18}$$

kde v predposlednej rovnosti bol použití vzťah 3.9. Vzťah 3.16 určuje pre odhad $\boldsymbol{T}'\boldsymbol{b}$ lineárnu transforáciu z normálne rozdeleného vektora $\boldsymbol{Y}$. Potom p-rozmerný vektor odhadov $\boldsymbol{T}$'$\boldsymbol{b}$ má tiež normálne rozdelenie

$$\boldsymbol{T}'\boldsymbol{b} \sim \mathsf{N}_p(\boldsymbol{T}'\boldsymbol{\beta}, \sigma^2\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{T}). \tag{3.19}$$

Z odvodzovania predchádzajúcich vzťahov vyplýva pre samotný odhad $\boldsymbol{b}$ rozdelenie

$$\boldsymbol{b} \sim \mathsf{N}_k((\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^-]'). \tag{3.20}$$

Variančnú maticu $\mathsf{Var}(\boldsymbol{T}'\boldsymbol{b})$ je možné vyjadriť aj bez pseudoinverznej matice $(\boldsymbol{X}'\boldsymbol{X})^-$ nasledovne

$$\mathsf{Var}(\boldsymbol{T}'\boldsymbol{b}) = \sigma^2\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{T} = \sigma^2\boldsymbol{J}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}\boldsymbol{J} = \sigma^2\boldsymbol{J}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{J}. \tag{3.21}$$

Veta 3.7. *Odhad $\boldsymbol{T}'\boldsymbol{b}$ a hodnota RSS sú navzájom nezávislé.*

Dôkaz. Platí $RSS = \boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y}$ a $\boldsymbol{T}'\boldsymbol{b} = \boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y}$. Myšlienka dôkazu je rovnaká ako pri dôkaze vety 2.10. a v tomto prípade stačí ukázať

$$\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'(\sigma^2\boldsymbol{I})\boldsymbol{M} = \boldsymbol{O}.$$

Táto rovnosť sa ukáže nasledovne

$$\begin{aligned} \boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'(\sigma^2\boldsymbol{I})\boldsymbol{M} &= \sigma^2\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}') = \\ &= \sigma^2\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}' - \sigma^2\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}' = \\ &= \sigma^2\boldsymbol{J}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}' - \sigma^2\boldsymbol{J}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}' = \\ &= \sigma^2\boldsymbol{J}'\boldsymbol{X}'(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}' - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}') = \boldsymbol{O}, \end{aligned}$$

kde v poslednej rovnosti bol využitý vzťah 3.5. □

Veta 3.8. *Platí* $RSS/\sigma^2 \sim \chi^2(n-r)$, *kde* r *označuje hodnosť matice* $\boldsymbol{X}$.

Dôkaz. Postupuje sa rovnako ako v dôkaze vety 2.9. Priestor $\mathcal{M}(\boldsymbol{X})^\perp$ má dimenziu $n-r$ a preto je možné uvažovať maticu $\boldsymbol{N}_{n\times n-r}$ spĺňajúcu $\boldsymbol{N}\boldsymbol{N}' = \boldsymbol{M}$ a $\boldsymbol{N}'\boldsymbol{N} = \boldsymbol{I}_{n-r}$. Potom je zrejmé, že vektor $\boldsymbol{u} = \boldsymbol{M}\boldsymbol{Y}$ má rovnaké rozdelenie ako $\sigma\boldsymbol{N}\boldsymbol{Z} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{M})$, kde vektor $\boldsymbol{Z} \sim \mathsf{N}_{n-r}(\boldsymbol{0}, \boldsymbol{I})$. Z toho vyplýva

$$\frac{RSS}{\sigma^2} = \frac{\boldsymbol{u}'\boldsymbol{u}}{\sigma^2} \text{ má rovnaké rozdelenie, ako } \boldsymbol{Z}'\boldsymbol{N}'\boldsymbol{N}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{Z} \sim \chi^2(n-r).$$

□

3.2.1. Testy hypotéz a intervaly spoľahlivosti

Najprv bude ukázané testovanie na jednej parametrickej funkcii. Nech je teda daná odhadnuteľná parametrická funkcia $\boldsymbol{t}'\boldsymbol{\beta}$. Jej bodový odhad metódou najmenších štvorcov bude označený $\boldsymbol{t}'\boldsymbol{b}$.

Veta 3.9. *Nech* $h(\boldsymbol{X}) = r$ *a* $\boldsymbol{t}'\boldsymbol{\beta}$ *je odhadnuteľná parametrická funkcia. Potom platí*

$$\frac{(\boldsymbol{t}'\boldsymbol{b} - \boldsymbol{t}'\boldsymbol{\beta})}{\sqrt{s^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}}} \sim \mathsf{t}(n-r).$$

Dôkaz. Zo vzťahu 3.19 vyplýva $(\boldsymbol{t}'\boldsymbol{b} - \boldsymbol{t}'\boldsymbol{\beta})/\sqrt{\sigma^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}} \sim \mathsf{N}(0,1)$. Z vety 3.8. je známa platnosť $RSS/\sigma^2 \sim \chi^2(n-r)$. A z vety 3.7. nepriamo vyplýva, že hodnota $\boldsymbol{t}'\boldsymbol{b}$ a odhad s^2 sú navzájom nezávislé veličiny. Potom platí

$$\frac{(\boldsymbol{t}'\boldsymbol{b} - \boldsymbol{t}'\boldsymbol{\beta})/\sqrt{\sigma^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}}}{\sqrt{RSS\sigma^{-2}/(n-r)}} = \frac{(\boldsymbol{t}'\boldsymbol{b} - \boldsymbol{t}'\boldsymbol{\beta})}{\sqrt{s^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}}} \sim \mathsf{t}(n-r)$$

kde bol použitý vzťah pre odhad $s^2 = RSS/(n-r)$ neznámeho parametra σ^2. □

Vetu 3.9 je možné výhodne využiť pri testovaní hypotéz. Nech pre dané číslo $\theta \in \mathbb{R}$, je uvažovaná nulová hypotéza $\mathsf{H}_0 : \boldsymbol{t}'\boldsymbol{\beta} = \theta$ proti alternatívnej hypotéze $\mathsf{H}_1 : \boldsymbol{t}'\boldsymbol{\beta} \neq \theta$. Potom H_0 bude zamietnutá na hladine významnosti α, práve keď platí

$$\frac{(\boldsymbol{t}'\boldsymbol{b} - \theta)}{\sqrt{s^2\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}}} \geq \mathsf{t}_{1-\frac{\alpha}{2}}(n-r).$$

Obojstranný interval spoľahlivosti pre odhad parametra $\boldsymbol{t}'\boldsymbol{\beta}$ so spoľahlivosťou $1-\alpha$ je

$$\left(\boldsymbol{t}'\boldsymbol{b} - s\sqrt{\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}}\,\mathsf{t}_{1-\frac{\alpha}{2}}(n-r),\ \boldsymbol{t}'\boldsymbol{b} + s\sqrt{\boldsymbol{t}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{t}}\,\mathsf{t}_{1-\frac{\alpha}{2}}(n-r)\right). \tag{3.22}$$

V prípade, že je potrebné testovať viacero odhadnuteľných parametrických funkcií súčasne, je vhodnejšie používať vektorový tvar $\boldsymbol{T}'\boldsymbol{\beta}$, kde matica $\boldsymbol{T}_{k\times p}$ bola presne definovaná v odstavci 3.2. Bodovým odhadom získaným metódou najmenších štvorcov vektora parametrických funkcií $\boldsymbol{T}'\boldsymbol{\beta}$ je vektor $\boldsymbol{T}'\boldsymbol{b}$, ktorý má podľa vzťahu 3.19 normálne rozdelenie $\mathsf{N}_p(\boldsymbol{T}'\boldsymbol{\beta}, \sigma^2\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{T})$.

Veta 3.10. *Platí*

$$\frac{1}{ps^2}(\boldsymbol{T}'\boldsymbol{b}-\boldsymbol{T}'\boldsymbol{\beta})'(\boldsymbol{T}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{T})^{-1}(\boldsymbol{T}'\boldsymbol{b}-\boldsymbol{T}'\boldsymbol{\beta}) \sim \mathsf{F}(p, n-r).$$

Dôkaz. Podľa vzťahu 3.21 platí $\mathsf{Var}(\boldsymbol{T}'\boldsymbol{b}) = \sigma^2\boldsymbol{J}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{J}$, kde na maticu $\boldsymbol{J}_{k\times p}$ bola položená podmienka $\mathrm{h}(\boldsymbol{X}\boldsymbol{J}) = p$ (kde $p \leq r$). Variančná matica $\mathsf{Var}(\boldsymbol{T}'\boldsymbol{b})$ je potom regulárna (s hodnosťou p) pozitívne definitná. Použitím vety 1.4 a kladnosti vlastných čísiel matice $\boldsymbol{J}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{J}$, je možné urobiť rozklad

$$\boldsymbol{J}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{J} = \boldsymbol{B}\boldsymbol{B}',$$

kde matica $\boldsymbol{B}$ je regulárna a jej i-tý stĺpce je $\sqrt{\lambda_i}$-násobkom vlastného vektora $\boldsymbol{v}_i$ pôvodnej matice. Potom náhodný vektor $\boldsymbol{T}'\boldsymbol{b}$ má rovnaké rozdelenie ako vektor

$$\boldsymbol{T}'\boldsymbol{\beta} + \sigma\boldsymbol{B}\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{T}'\boldsymbol{\beta}, \sigma^2\boldsymbol{J}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{J}), \tag{3.23}$$

kde vektor $\boldsymbol{Z}$ má normálne rozdelenie $\mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$. Pri následnom dosadení náhodného vektora 3.23 do vzťahu

$$\begin{aligned}\tfrac{1}{\sigma^2}(\boldsymbol{T}'\boldsymbol{b}-\boldsymbol{T}'\boldsymbol{\beta})'(\boldsymbol{B}\boldsymbol{B}')^{-1}(\boldsymbol{T}'\boldsymbol{b}-\boldsymbol{T}'\boldsymbol{\beta}) &= \tfrac{1}{\sigma^2}(\sigma\boldsymbol{B}\boldsymbol{Z})'(\boldsymbol{B}\boldsymbol{B}')^{-1}(\boldsymbol{B}\boldsymbol{Z}\sigma) = \\ &= \boldsymbol{Z}'\boldsymbol{B}'(\boldsymbol{B}\boldsymbol{B}')^{-1}\boldsymbol{B}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{I}\boldsymbol{Z} \sim \chi^2(p),\end{aligned}$$

kde pri odvodzovaní bola využitá rovnosť $\boldsymbol{B}'(\boldsymbol{B}\boldsymbol{B}')^{-1}\boldsymbol{B} = \boldsymbol{I}$. Táto rovnosť sa ukáže platnou, podobne ako v dôkaze vety 2.12. vynásobením zľava (regulárnou) maticou $\boldsymbol{B}$ (resp. sprava maticou $\boldsymbol{B}'$). Pomocou vysledkov viet 3.6. a 3.8. je možné odvodiť nasledujúci vzťah

$$\frac{\frac{1}{p\sigma^2}(\boldsymbol{T}'\boldsymbol{b}-\boldsymbol{T}'\boldsymbol{\beta})'(\boldsymbol{B}\boldsymbol{B}')^{-1}(\boldsymbol{T}'\boldsymbol{b}-\boldsymbol{T}'\boldsymbol{\beta})}{RSS\sigma^{-2}/(n-r)} = \frac{(\boldsymbol{T}'\boldsymbol{b}-\boldsymbol{T}'\boldsymbol{\beta})'(\boldsymbol{B}\boldsymbol{B}')^{-1}(\boldsymbol{T}'\boldsymbol{b}-\boldsymbol{T}'\boldsymbol{\beta})}{ps^2} \sim \mathsf{F}(p, n-r),$$

ktorý korešponduje s tvrdením tejto vety. Ľavá strana predchádzajúceho vzťahu odpovedá zadefinovanému Fisherovmu rozdeleniu $\frac{\chi^2(p)/p}{\chi^2(n-r)/(n-r)}$. □

Podľa vety 3.10. je možné testovať nulovú hypotézu $\mathsf{H}_0 : \boldsymbol{T}'\boldsymbol{\beta} = \boldsymbol{\theta}$ proti alternatívnej hypotéze $\mathsf{H}_1 : \boldsymbol{T}'\boldsymbol{\beta} \neq \boldsymbol{\theta}$, kde $\boldsymbol{\theta}$ je p-prvkový vektor. Nulová hypotéza H_0 bude zamietnutá na hladine α, práve vtedy, keď hodnota výrazu uvedeného vo vete 3.10. prekročí kvantil $\mathsf{F}_{1-\alpha}(q, n-k)$.

4. Výber modelu a diagnostika

Pri výbere "rozumného"modelu by mali byť regresory vyberané do modelu tak, aby čo najlepšie popisovali strednú hodnotu náhodného vektora $\boldsymbol{Y}$ (bez újmy na všeobecnosti, nech je uvažovaný lineárny regresný model neúplnej hodnosti 3.1). Napríklad v prípade, že regresná matica $\boldsymbol{X}$ ($\mathrm{h}(\boldsymbol{X}) = k$) má príliš veľa stĺpcov a niektoré sa len nevýznamne podieľajú na vyjadrení strednej hodnoty $\mathsf{E}(\boldsymbol{Y})$. Preto je dôležitý správny výber týchto regresorov. Teda je potrebné uvažovať menší model (takzvaný *podmodel*) s regresnou maticou $\boldsymbol{X}^*$ s hodnosťou k^*, pre ktorú platí

$$\mathcal{M}(\boldsymbol{X}^*) \subseteq \mathcal{M}(\boldsymbol{X}) \quad \text{a zároveň} \quad 0 < k^* < k, \tag{4.1}$$

kde $k = \mathrm{h}(\boldsymbol{X})$ a $k^* = \mathrm{h}(\boldsymbol{X}^*)$. Často je takýto podmodel tvorený len istými stĺpcami z $\boldsymbol{X}$. V prípade opačnej situácie sa hovorí o *nadmodeli*. Nech je teda uvažovaný model 3.1 a matica $\boldsymbol{X}^*$ spĺňajúca 4.1. Potom sa povie, že *platí podmodel* modelu 3.1, ak pre nejaký vektor regresných koeficientov $\boldsymbol{\beta}^*$ platí

$$\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{X}^*\boldsymbol{\beta}^*.$$

Nech sú označené všetky veličiny vzťahujúce sa na podmodel symbolom * v hornom indexe. Zvára v literatúre ([13], str. 30) uviedol nasledujúce matematické tvrdenie: Ak platí v modeli podmodel a $\boldsymbol{Y}$ má normálne rozdelenie, potom štatistiky $\hat{\boldsymbol{Y}}^*$ a $\hat{\boldsymbol{u}}^*$ sú nezávislé a platí

$$\frac{(RSS^* - RSS)/(k - k^*)}{RSS/(n-k)} \sim \mathsf{F}(k - k^*, n - k). \tag{4.2}$$

Dôsledok neúplného výberu regresorov

Nasledujúce pojednanie bude o situácii, kedy budú vytvorené odhady pomocou podmodelu pričom skutočný model bude širší. Nech je teda uvažovaný lineárny regresný model neúplnej hodnosti (podľa 3.1)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{4.3}$$

kde $\boldsymbol{X}_{n\times k}$ je matica daných reálnych čísiel hodnosti p a $\boldsymbol{\beta}$ vektor neznámych parametrov. A zároveň platí pre náhodný vektor $\mathsf{E}(\boldsymbol{e}) = \boldsymbol{0}$ a $\mathsf{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Avšak skutočný model sa bude riadiť rozšíreným lineárnym modelom

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{e} \tag{4.4}$$

tak, že $\boldsymbol{Z}_{n\times m}$ je matica daných reálnych čísiel a $\boldsymbol{\gamma}$ vektor m-neznámych parametrov. A zároveň pre náhodný vektor tohto rozšíreného modelu platí $\mathsf{E}(\boldsymbol{e}) = \boldsymbol{0}$ a $\mathsf{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. V ďalšom texte bude regresná matica rozšíreného modelu označená $\boldsymbol{G}_{n\times(k+m)} = (\boldsymbol{X}, \boldsymbol{Z})$ a rozšírený vektor regresných koeficientov $\boldsymbol{\alpha} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$. Potom je možné formálne prepísať model na

$$\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{\alpha} + \boldsymbol{e}. \tag{4.5}$$

Aby skutočný rozšírený model rozšíril priestor strednej hodnoty $\mathsf{E}(\boldsymbol{Y})$, na regresnú maticu $\boldsymbol{G}$ je kladená podmienka $\mathrm{h}(\boldsymbol{G}) > p$. Samozrejme tiež platí $\mathrm{h}(\boldsymbol{G}) < n$, pretože model 4.5 je lineárnym regresným modelom.

Ďalej budú označené indexom G všetky potrebné štatistiky vzťahujúce sa na rozšírený

model 4.5. $\boldsymbol{H}_G$ bude označovať projekčnú maticu na rozšírený regresný priestor $\mathcal{M}(\boldsymbol{G})$ a matica $\boldsymbol{M}_G = \boldsymbol{I} - \boldsymbol{H}_G$ bude označovať projekčnú maticu na (zúžený) reziduálny priestor $\mathcal{M}(\boldsymbol{G})^\perp$. Potom zrejme platí pre najlepší lineárny nestranný odhad vektora $\boldsymbol{Y}$ vzťah $\hat{\boldsymbol{Y}}_G = \boldsymbol{H}_G\boldsymbol{Y}$ a pre odhad chyby $\boldsymbol{u}_G = \boldsymbol{M}_G\boldsymbol{Y} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_G$. Pre reziduálny súčet štvorcov rozšíreného modelu $RSS_G = \|\boldsymbol{u}_G\|^2$. Podľa vzťahu 3.12 platí

$$\hat{\boldsymbol{Y}}_G = \boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G})^-\boldsymbol{G}'\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b}_G + \boldsymbol{Z}\boldsymbol{c}_G, \tag{4.6}$$

kde $\boldsymbol{b}_G$ a $\boldsymbol{c}_G$ sú odhady metódou najmenších štvorcov pre vektory $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$. Avšak ak je potrebné vidieť rozdiel v odhadoch regresných koeficientov medzi uvažovaným 4.3 a skutočným rozšíreným modelom, je potrebné aby odhady vektorov regresných koeficientov $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ boli nezávislé. Platí $\mathcal{M}(\boldsymbol{X}, \boldsymbol{Z}) = \mathcal{M}(\boldsymbol{X}, \boldsymbol{M}\boldsymbol{Z})$, pretože $\boldsymbol{M}$ je projekčná matica do reziduálneho priestoru $\mathcal{M}(\boldsymbol{X})^\perp$ a teda stĺpce matice $\boldsymbol{Z}$ sa premietnu do reziduálneho priestoru ako stĺpce matice $\boldsymbol{M}\boldsymbol{Z}$. Tým vznikne nezávislosť odhadov regresných koeficientov. Miesto matice $\boldsymbol{G}$ vo vzťahu 4.6 sa použije matica $(\boldsymbol{X}, \boldsymbol{M}\boldsymbol{Z})$, ktorá (ako už bolo ukázané) generuje rovnaký regresný priestor.

$$\begin{aligned}
\hat{\boldsymbol{Y}}_G &= (\boldsymbol{X}, \boldsymbol{M}\boldsymbol{Z})((\boldsymbol{X}, \boldsymbol{M}\boldsymbol{Z})'(\boldsymbol{X}, \boldsymbol{M}\boldsymbol{Z}))^-(\boldsymbol{X}, \boldsymbol{M}\boldsymbol{Z})'\boldsymbol{Y} \\
&= (\boldsymbol{X}, \boldsymbol{M}\boldsymbol{Z}) \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} \end{pmatrix}^- \begin{pmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}'\boldsymbol{M} \end{pmatrix} \boldsymbol{Y} \\
&= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{M}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Y} \\
&= \hat{\boldsymbol{Y}} + \boldsymbol{M}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{u} \\
&= \boldsymbol{X}\boldsymbol{b} + (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{u} \\
&= \boldsymbol{X}\boldsymbol{b} + (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}')\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{u} \\
&= \boldsymbol{X}(\boldsymbol{b} - (\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{u}) + \boldsymbol{I}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{u} \\
&= \boldsymbol{X}\underbrace{(\boldsymbol{b} - (\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{u})}_{\boldsymbol{b}_G} + \boldsymbol{Z}\underbrace{(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{u}}_{\boldsymbol{c}_G}
\end{aligned} \tag{4.7}$$

Použitím vlastnosti matice 3.14, na hore uvedený súčin blokových matíc, sa príde k záveru, že i tento súčin je nezávislý na zvolenej pseudoinverzii. Ďalej je zrejmé, že k matici

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} \end{pmatrix}$$

je pseudoinverzná matica

$$\begin{pmatrix} (\boldsymbol{X}'\boldsymbol{X})^- & \boldsymbol{O} \\ \boldsymbol{O} & (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^- \end{pmatrix}, \tag{4.8}$$

čo sa dá ukázať súčinom blokových matíc z definície pseudoinverznej matice. A práve použitím pseudoinverznej matice 4.8 v odvodzovaní 4.7 sa sprehľadní postup (nezávislý na zvolenej pseudoinverzii). V poslednom riadku odvodzovania 4.7 je možné si všimnúť priamy vzťah medzi odhadmi $\boldsymbol{b}$ a $\boldsymbol{b}_G$. V predchádzajúcom odvodzovaní 4.7 taktiež bola ukázaná platnosť vzťahu $\hat{\boldsymbol{Y}}_G = \hat{\boldsymbol{Y}} + \boldsymbol{M}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{u}$. Vektorom $\boldsymbol{d}$ bude ďalej označený rozdiel odhadov $\hat{\boldsymbol{Y}}_G$ a $\hat{\boldsymbol{Y}}$ takto

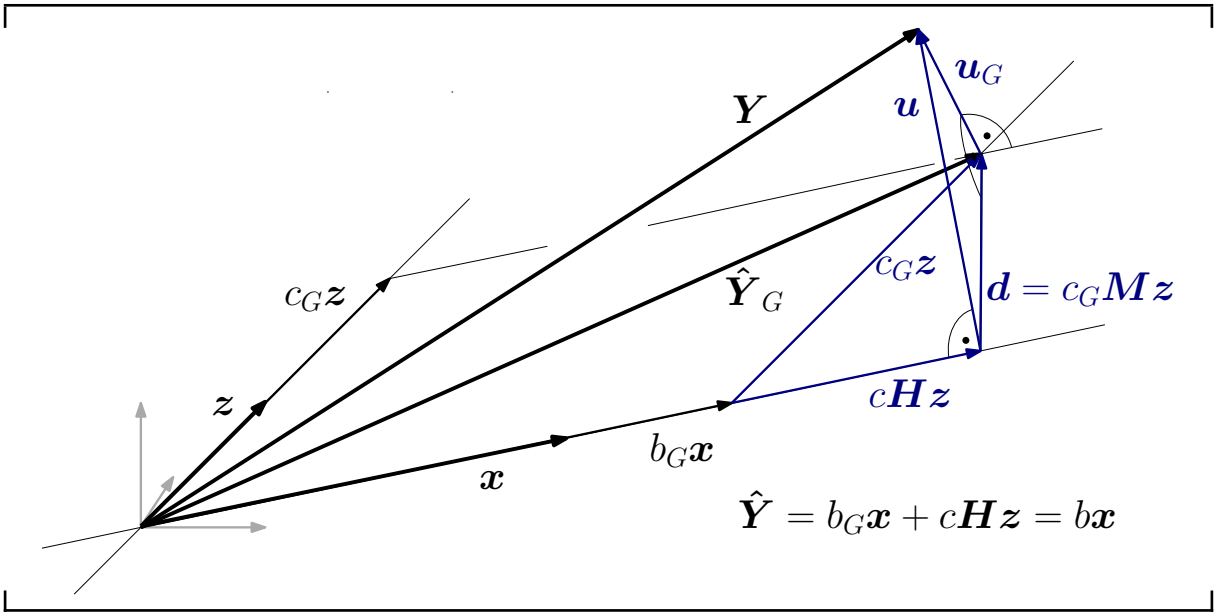
$$\boldsymbol{d} = \hat{\boldsymbol{Y}}_G - \hat{\boldsymbol{Y}} = \boldsymbol{M}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{u}. \tag{4.9}$$

Nech je odčítaný vektor $\boldsymbol{Y}$ zo vzťahu $\hat{\boldsymbol{Y}}_G = \hat{\boldsymbol{Y}} + \boldsymbol{d}$, potom platí vzťah

$$(\hat{\boldsymbol{Y}}_G - \boldsymbol{Y}) = (\hat{\boldsymbol{Y}} - \boldsymbol{Y}) + \boldsymbol{d}$$

a následne rovnosť

$$-\boldsymbol{u}_G = -\boldsymbol{u} + \boldsymbol{d} \quad \text{resp.} \quad \boldsymbol{u} = \boldsymbol{u}_G + \boldsymbol{d}. \tag{4.10}$$



Obrázok 5: Na obrázku je zachytený prípad, kedy regresná matica $\boldsymbol{X}$ je tvorená len vektorom $\boldsymbol{x}$ a matica $\boldsymbol{Z}$ je tvorená len vektorom $\boldsymbol{z}$. Označenie korešponduje so zavedením v tejto kapitole. Matica $\boldsymbol{H}$ je projekčná matica na $\mathcal{M}(\boldsymbol{x})$ a $\boldsymbol{M}$ je projekčná matica na $\mathcal{M}(\boldsymbol{x})^{\perp}$.

Z platností vzťahov $\boldsymbol{u}_G \in \mathcal{M}(\boldsymbol{G})^{\perp}$ a $\boldsymbol{d} \in \mathcal{M}(\boldsymbol{G}) \cap \mathcal{M}(\boldsymbol{X})^{\perp}$ vyplýva, že vektory $\boldsymbol{u}_G$ a $\boldsymbol{d}$ sú na seba kolmé (zjednodušená situácia je zobrazená na obrázku 5). Preto z kolmosti pre vzťah $\boldsymbol{u} = \boldsymbol{u}_G + \boldsymbol{d}$ z 4.10 platí po znormovaní a umocnení

$$\|\boldsymbol{u}\|^2 = \|\boldsymbol{u}_G + \boldsymbol{d}\|^2 = \|\boldsymbol{u}_G\|^2 + \|\boldsymbol{d}\|^2 .$$

Z predchádzajúceho priamo vyplýva

$$RSS = RSS_G + \|\boldsymbol{d}\|^2 . \tag{4.11}$$

Podľa vety 3.6. a platnosti modelu 4.5 platí

$$\mathsf{E}(RSS_G) = (n - \mathrm{h}(\boldsymbol{G}))\sigma^2. \tag{4.12}$$

Avšak v prípade $\mathsf{E}(RSS)$ platí

$$\mathsf{E}(RSS) = \mathsf{E}\,\|\boldsymbol{M}\boldsymbol{Y}\|^2 = \mathsf{E}\,\|\boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{e})\|^2 = \mathsf{E}\,\|\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{M}\boldsymbol{e}\|^2 . \tag{4.13}$$

Oba členy v norme, poslednom vyjadrení vzťahu 4.13, sa nachádzajú v navzájom ortogonálnych podpriestoroch a preto je možné ďalej pokračovať vo vyjadrovaní takto

$$\mathsf{E}(RSS) = \mathsf{E}\,\|\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}\|^2 + \mathsf{E}\,\|\boldsymbol{M}\boldsymbol{e}\|^2 = \|\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}\|^2 + (n - \mathrm{h}(\boldsymbol{X}))\sigma^2. \tag{4.14}$$

Pre odhad strednej hodnoty uvažovaného modelu $\mathsf{E}\hat{\boldsymbol{Y}}$ platí

$$\mathsf{E}\hat{\boldsymbol{Y}} = \boldsymbol{H}\mathsf{E}\hat{\boldsymbol{Y}}_G = \boldsymbol{H}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{Z}\boldsymbol{\gamma}.$$

V prípade, že rozšírený model skutočne rozširuje pôvodný model (teda neplatí $\boldsymbol{H}\boldsymbol{Z} = \boldsymbol{Z}$) je odhad strednej hodnoty $\hat{\boldsymbol{Y}}$ odhadom vychýleným. Preto platí pre vychýlenie tohto odhadu nasledujúci vzťah

$$\mathsf{bias}(\hat{\boldsymbol{Y}}) = \mathsf{E}\hat{\boldsymbol{Y}} - \mathsf{E}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{Z}\boldsymbol{\gamma} - (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}) = (\boldsymbol{H} - \boldsymbol{I})\boldsymbol{Z}\boldsymbol{\gamma} = -\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}. \quad (4.15)$$

Ako uvádza Zvára v literatúre ([13], str. 83), vychýlené odhady nie sú porovnávané pomocou ich rozptylu či variančnej matice, ale pomocou *strednej štvorcovej chyby*. Táto štatistika označovaná MSE (z angl. Mean Square Error) je všeobecne definovaná pre odhad $\hat{\theta}$ parametra θ takto

$$\mathsf{MSE}(\hat{\theta}) = \mathsf{E}(\hat{\theta} - \theta)^2 = \mathsf{D}(\hat{\theta}) + \mathsf{bias}^2(\hat{\theta}). \quad (4.16)$$

a pre vektor odhadov $\hat{\boldsymbol{\theta}}$ parametrov $\boldsymbol{\theta}$ nasledovne

$$\mathsf{MSE}(\hat{\boldsymbol{\theta}}) = \mathsf{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' = \mathsf{Var}(\hat{\boldsymbol{\theta}}) + \mathsf{bias}(\hat{\boldsymbol{\theta}})\mathsf{bias}(\hat{\boldsymbol{\theta}})'. \quad (4.17)$$

Preto v prípade dvoch nestranných odhadov ($\mathsf{bias}(\hat{\theta}) = 0$ resp. $\mathsf{bias}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}$) je ich porovnanie pomocou rozptylov (resp. variančných matic) a stredných švorcovych chýb totožné. Potom strednú štvorcovú chybu odhadu $\hat{\boldsymbol{Y}}$ pomocou vzťahu 4.15 je možno rozpísať nasledovne

$$\mathsf{MSE}(\hat{\boldsymbol{Y}}) = \mathsf{Var}(\hat{\boldsymbol{Y}}) + \mathsf{bias}(\hat{\boldsymbol{Y}})(\mathsf{bias}(\hat{\boldsymbol{Y}}))' = \sigma^2\boldsymbol{H} + \boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{M}.$$

Pretože odhad $\hat{\boldsymbol{Y}}_G$ je nevychýlený (teda $\mathsf{bias}(\hat{\boldsymbol{Y}}_G) = \boldsymbol{0}$) platí

$$\begin{aligned}
\mathsf{MSE}(\hat{\boldsymbol{Y}}_G) &= \mathsf{Var}(\hat{\boldsymbol{Y}}_G) = (\boldsymbol{X}, \boldsymbol{M}\boldsymbol{Z})\mathsf{Var}((\boldsymbol{b}_G', \boldsymbol{c}_G')')(\boldsymbol{X}, \boldsymbol{M}\boldsymbol{Z})'\boldsymbol{Y} \\
&= \sigma^2(\boldsymbol{X}, \boldsymbol{M}\boldsymbol{Z}) \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} \end{pmatrix}^- \begin{pmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}'\boldsymbol{M} \end{pmatrix} \\
&= \sigma^2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}' + \sigma^2\boldsymbol{M}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{M} \\
&= \sigma^2\boldsymbol{H} + \sigma^2\boldsymbol{M}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{M}
\end{aligned}, \quad (4.18)$$

kde bola použitá pseudoinverzia tvaru 4.8. Ako bolo konštatované vyššie, súčin blokových matíc v 4.18 je nezávislý na voľbe pseudoinverzie. Použitím 4.18 ďalej platí

$$\mathsf{MSE}(\hat{\boldsymbol{Y}}_G) - \mathsf{MSE}(\hat{\boldsymbol{Y}}) = \sigma^2(\boldsymbol{M}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{M} - \boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{M}/\sigma^2). \quad (4.19)$$

Zvára v literatúre ([13], str. 228) dokázal nasledujúcu vetu: Pre maticu $\boldsymbol{A}_{m\times n}$ a vektor $\boldsymbol{c} \in \mathbb{R}^n$ platí nerovnosť $\|\boldsymbol{A}\boldsymbol{c}\|^2 \leq 1$ práve vtedy, keď je matica

$$\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^-\boldsymbol{A}' - \boldsymbol{A}\boldsymbol{c}\boldsymbol{c}'\boldsymbol{A}' \quad (4.20)$$

pozitívne semidefinitná.

Aplikovaním tejto vety na vzťah 4.19 (položením $\boldsymbol{A} = \boldsymbol{M}\boldsymbol{Z}$ a $\boldsymbol{c} = \boldsymbol{\gamma}/\sigma$) je možné dôjsť k záveru, že matica 4.19 je pozitívne semidefinitná práve za podmienky $\|\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}/\sigma\|^2 \leq 1$. Túto nerovnosť je možné jednoduchou úpravou previesť na $\|\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}\|^2 \leq \sigma^2$ a použitím vzťahu 4.15 ďalej previesť na vzťah $\left\|\mathsf{bias}(\hat{\boldsymbol{Y}})\right\|^2 \leq \sigma^2$ (resp. na vzťah $\|\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}\|^2 \leq \sigma^2$.). Potom z toho vyplýva

$$\mathsf{MSE}(\hat{\boldsymbol{Y}}_G) - \mathsf{MSE}(\hat{\boldsymbol{Y}}) \text{ je pozitívne semidefinitná, práve keď, platí } \|\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}\| \leq \sigma. \quad (4.21)$$

To znamená, že pri dostatočne malom vychýlení odhadu strednej hodnoty je menší model prijateľnejší, pretože minimalizuje kvadráty reziduí v budúcom pozorovaní.

4.1. Výber regresorov

Nech je uvažovaná regresná matica $\boldsymbol{X}$ a vektor závislých veličin $\boldsymbol{Y}$. Úlohou je vybrať tie regresory, ktoré dobre popisujú náhodné veličiny v $\boldsymbol{Y}$. Prvou dôležitou analýzou dát je zistenie vzájomnej závislosti regresorov. Tu sa využíva korelačný koeficient (resp. korelačná matica), ktorý indikuje mieru zavislosti medzi dvoma regresormi. Napríklad môžu byť dva regresory závislé na faktore, ktorý nie je známy. V prípade, že jeden regresor $\boldsymbol{x}_{\bullet i}$ (faktor) je silne lineárne závislý na druhom $\boldsymbol{x}_{\bullet j}$, potom si odhady $\boldsymbol{b}_i$, $\boldsymbol{b}_j$ rozdelia pôsobenie jedného faktora. V tomto prípade, kedy sú stĺpce regresnej matice $\boldsymbol{X}$ lineárne nezávislé, ale majú (v nejakom zmysle) silnú lineárnu závislosť medzi sebou, sa hovorí o *multikolinearite*. Tu je možné využiť metódu hlavných komponentov alebo napríklad faktorovú analýzu.

V prípade modelu úplnej hodnosti je teda situácia výberu regresorov nasledujúca. Variančná matica odhadov pre vektor $\boldsymbol{\beta}$ podľa vzťahu 2.13 bola určený ako $\mathsf{Var}(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Nech je najprv uvažovaný najjednoduchší prípad, keď stĺpce regresnej matice sú na seba ortogonálne. Potom sú matice $\boldsymbol{X}'\boldsymbol{X}$ i $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ diagonálne a teda jednotlivé odhady b_i na sebe nezávislé. To znamená, že pri uvažovaní podmodelu zo stĺpcov regresnej matice sa ich príslušné bodové odhady b_i už nebudú meniť. V odstavci 2.1 bolo uvedené, ako t-testom sú testované jednotlivé odhady pre β_i. Avšak ani pri nezávislosti jednotlivých odhadov sa nemožno spoľahnúť na t-test pri vytváraní vhodného podmodelu. Pretože pri testovaní, jednotlivých odhadov na hladine významnosti α, pravdepodobnosť správneho výberu podmodelu z pôvodného modelu s počtom regresorov klesá. V prípade závislosti jednotlivých odhadov b_i, sa táto situácia ešte komplikuje. Pri závislosti odhadov treba pri každom vyradení stĺpca z regresnej matice $\boldsymbol{X}$ znova vypočítať odhady regresných koeficientov.

F-testom sa testuje nevýznamnosť viacerych koeficientov súčasne. Preto sa zdá byť vhodnou voľbou. Problém môže nastať v prípade, že model obsahuje veľký počet regresorov. V tomto prípade p regresorov je potrebné pomocou F-testu skontrolovať 2^p modelov. V prípade veľkého počtu modelov sa používa takzvaný hladný algoritmus (v angl. Greedy Search). Na tejto myšlienke je založená kroková regresia, ktorá bude následne popísaná.

4.1.1. Kroková regresia

Najprv budú uvedené idei jednotlivých algoritmov pre *vzostupný výber* (v angl. Forward Selection) a *zostupny výber* (v angl. Backward Elimination), na ktorých je založená kroková regresia. Vzostupný výber je založený na postunom pridávaní regresorov podľa významnosti. Začína sa teda z prázdnym modelom a dá sa zhrnúť nasledovným postupom, ktorý je stručne zhrnutý na obrázku 6.

1. Nech je množine vybraných regresorov (označenej $\mathcal{N}_{vyb}$) priradená prázdna množina $\emptyset$. A ďalej nech množine všetkých p regresorov (označenej $\mathcal{N}_{nevyb}$) je priradená množina $\{1, \ldots, p\}$, kde je každému regresoru priradené práve jedno číslo i z $1, \ldots, p$ podľa označenia koeficientu β_i. t-testom so stupňom voľnosti $n - 1$ sa vyberie najvýznamnejšie nenulový odhad b_i. Potom do množiny $\mathcal{N}_{vyb}$ sa priradí hodnota i a z množiny $\mathcal{N}_{nevyb}$ sa vyradí hodnota i.

2. Vypočíta sa hodnota RSS pre model obsahujúci regresory s indexami z $\mathcal{N}_{vyb}$, ktorá bude ďalej označená RSS_{mod}. Nech k vyjadruje počet prvkov $\mathcal{N}_{vyb}$. Obmenou vzťahu 4.2 bude testovaný tentokrát nadmodel za platnosti

$$\frac{RSS_{mod} - RSS^*}{RSS_{mod}/(n-k)} \sim \mathsf{F}(1, n-k),$$

kde RSS^* je reziduálny súčet štvorcov pre nadmodel už uvažovaného modelu (vždy rozšírený o jeden regresor s indexom z $\mathcal{N}_{nevyb}$). Je zrejmé, že pri menšej hodnote RSS^* bude skôr zamietnutá hypotéza. Preto je hľadaná minimálna hodnota RSS^* zo všetkých uvažovaných nadmodelov, ktorá bude ďalej označená RSS_{roz}.

3. Ak

$$\frac{RSS_{mod} - RSS_{roz}}{RSS_{mod}/(n-k)} > \mathsf{F}_{1-\alpha}(1, n-k),$$

- hypotéza $\beta_j = 0$ je zamietnutá, index j je priradený do množiny $\mathcal{N}_{vyb}$ a vyradený z množiny $\mathcal{N}_{nevyb}$,
- do hodnoty RSS_{mod} sa priradí hodnota RSS_{roz},
- prejde sa na krok 2,

inak hypotéza $\beta_j = 0$ nie je zamietnutá a algoritmus sa týmto ukončuje.

Zostupný výber je založený na rovnakej myšlienke, ale vychádza sa z úplného modelu, obsahujúceho všetky regresory. Potom sú postupne vyraďované z modelu regresory najmenej významné pomocou podobného F-testu.

Kroková regresia (v angl. Stepwise Regression) funguje na princípe vzostupného výberu s tým, že v každom kroku sa zároveň použije i zostupný výber. Teda v každom kroku sa kontroluje množina $\mathcal{N}_{vyb}$, či sa v nej nenachádza index už nevýznamného regresného koeficientu. Predošlý postup vzostupného výberu bude rozšírený o nasledujúce dva kroky.

4. RSS_{mod} bude priradená hodnota z kroku 3 rozšíreného modelu RSS_{roz}. Nech opäť k vyjadruje počet prvkov $\mathcal{N}_{vyb}$. Pomocou vzťahu 4.2 bude testovaný podmodel nasledovne

$$\frac{RSS^* - RSS_{mod}}{RSS_{mod}/(n-k)} \sim \mathsf{F}(1, n-k),$$

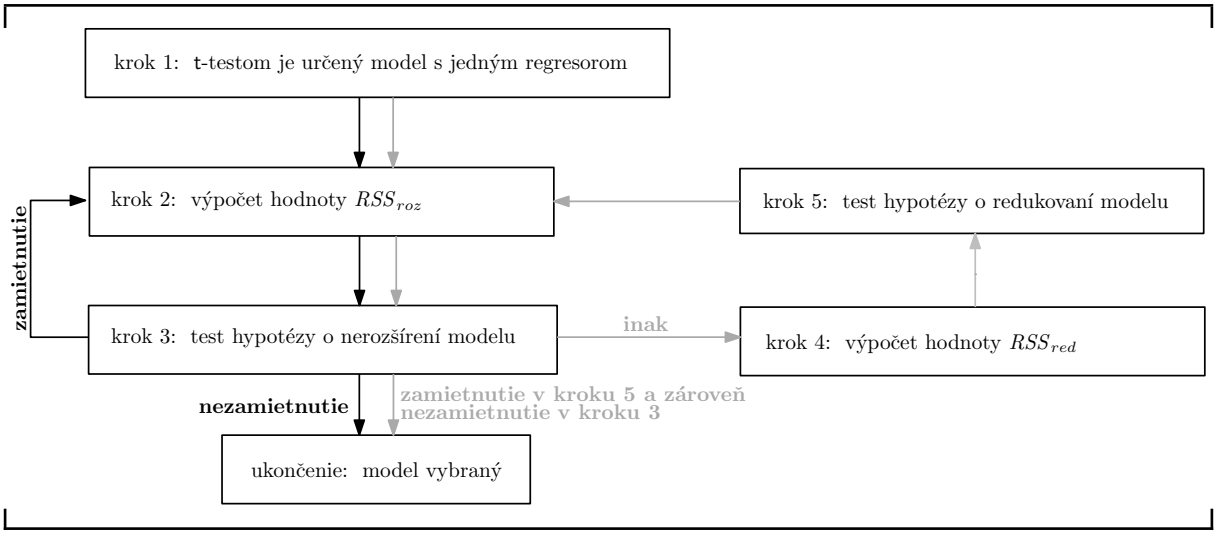
kde RSS^* je reziduálny súčet štvorcov pre podmodel už uvažovaného modelu (vždy redukovaný o jeden regresor s indexom z $\mathcal{N}_{vyb}$). Opäť je hľadaná minimálna hodnota RSS^* zo všetkých uvažovaných podmodelov, ktorá bude ďalej označená RSS_{red}.

5. Ak

$$\frac{RSS_{red} - RSS_{mod}}{RSS_{mod}/(n-k)} < \mathsf{F}_{1-\alpha}(1, n-k),$$

- nie je zamietnutá hypotéza o nulovosti vyradeného regresoru, preto je číslo určujúce tento regresor vyradený z $\mathcal{N}_{vyb}$ a priradený do $\mathcal{N}_{nevyb}$,
- do hodnoty RSS_{mod} sa priradí hodnota RSS_{red},
- prejde sa na krok 2,

inak sa prejde na krok 2.



Obrázok 6: Na obrázku je porovnaný proces vzostupneho výberu pomocou čiernych šipiek a proces krokovej regresie pomocou šedých šipiek.

Proces výberu regresorov je ukončený, ak prejde cyklus, v ktorom v krokoch 3 a 5 nebude možné rozšíriť alebo redukovať množinu vybraných regresorov $\mathcal{N}_{vyb}$.

Keďže tieto tri metódy fungujú na princípe hladného algoritmu, nemusia nájsť v danom zmysle najlepšie riešenie ani rovnaké (oproti ostatným). Napríklad vyber modelu bude závislý na zvolenej hladine významnosti α. Na túto tému Zvára uviedol v literatúre ([13], str. 144) nasledujúcu vetu: Obzvlášť pri krokovej regresii sa doporučuje vyhľadať niekoľko takmer optimálnych modelov a pokúsiť sa nájsť medzi nimi ten, ktorý má najlepšiu interpretáciu. Ďalej Jörnsten uviedla ([6], 5.prednáška) k tejto téme nasledujúcu vetu: Hladné algoritmy môžu viesť na modeli, ktoré sú ťažko interpretovateľné - regresory, ktoré sú korelované, súperia o zúčastnenie sa na tvorbe modelu a ľudské vedomosti (skúsenosti) môžu vybrať rozumné modely, čo tu štatistika nemôže.

V tomto odstavci, boli metódy výberu regresorov založené na F-teste. Rovnako tieto metódy môžu fungovať na základe iných testovacích kriterií. Známe sú napríklad Akaikeho informačné kritérium (známe pod skratkou AIC) alebo Bayesovské informačné kritérium (známe pod skratkou BIC).

4.1.2. Mallowsova C_p štatistika

Štatistika C_p je založená na odhade celkovej strednej štvorcovej chyby $\sum_{i=1}^{n} \mathsf{MSE}(\hat{Y}_i)$, ktorej je možné sa dopustiť pri uvažovaní podmodelu miesto skutočného modelu. Nech je uvažovaný zavedený model 4.3, z ktorého sa vychádzalo, avšak v skutočnosti platí rozšírený model 4.5. Model 4.3 bude teraz uvažovaný úplnej hodnosti (hodnosti k). Potom podľa 4.16 platí

$$\sum_{i=1}^{n} \mathsf{MSE}(\hat{Y}_i) = \sigma^2 \sum_{i=1}^{n} h_{ii} + (\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma})'\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma} = \sigma^2 \mathsf{Tr}(\boldsymbol{H}) + \|\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}\|^2 . \tag{4.22}$$

Zo vzťahu 4.14 platí

$$\|\boldsymbol{M}\boldsymbol{Z}\boldsymbol{\gamma}\|^2 = \mathsf{E}(RSS) - (n - \mathrm{h}(\boldsymbol{X}))\sigma^2.$$

Dosadením tohto vzťahu do 4.22 sa získa vzťah

$$\sum_{i=1}^{n} \mathsf{MSE}(\hat{Y}_i) = \sigma^2 k + \mathsf{E}(RSS) - (n-k)\sigma^2 = (2k-n)\sigma^2 + \mathsf{E}(RSS).$$

Podelením σ^2 predchádzajúceho vzťahu vznikne

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathsf{MSE}(\hat{Y}_i) = 2k - n + \frac{\mathsf{E}(RSS)}{\sigma^2}. \tag{4.23}$$

Mallowsova C_p štatistika vyjadruje hodnotu 4.23 po dosadení za σ^2 odhadom tohto parametra s_G^2 (zo skutočného rozšíreného modelu) a za $\mathsf{E}(RSS)$ je uvažovaná hodnota RSS z podmodelu nasledovne

$$\mathsf{C}_p = 2k - n + \frac{RSS}{s_G^2}. \tag{4.24}$$

Potom pre strednú hodnotu tejto štatistiky a použitím vzťahu 4.14 platí

$$\mathsf{E}(\mathsf{C}_p) = 2k - n + \frac{\mathsf{E}(RSS)}{\sigma^2} = 2k - n + \frac{\|\boldsymbol{MZ\gamma}\|^2 + (n-k)\sigma^2}{\sigma^2} = k + \frac{\|\boldsymbol{MZ\gamma}\|^2}{\sigma^2}. \tag{4.25}$$

Zo vzťahu 4.25 vyplýva, že pri malom vychýlení je stredná hodnota štatistiky C_p približne rovná k (počtu regresorov v podmodeli).

Ďalej bude uvažované, že skutočný model 4.5 je neznámy (väčšinou reálna situácia). No je uvažovaná množina p regresorov, ktorá v nejakom zmysle dobre zodpovedá modelu. Bez újmy na všeobecnosti a pre zachovanie modelu 4.3, nech medzi p regresormi sú všetky regresory z $\boldsymbol{X}$ ($p \geq k$). Mallows v tomto prípade vo vzťahu 4.24 uvažuje, že všetkých p uvažovaných regresorov vytvára nevychýlený odhad $\hat{\boldsymbol{Y}}$ ($\mathsf{bias}(\hat{\boldsymbol{Y}}) = \boldsymbol{0}$). Preto je z modelu s p regresormi počítaný odhad s_p^2 rozptylu σ^2.

Pre takýto prípad je C_p štatistika pre uvažovaný model 4.3

$$\mathsf{C}_p = 2k - n + \frac{RSS}{s_p^2}. \tag{4.26}$$

Takéto hodnoty štatistiky C_p sa vypočítajú pre všetky rôzne podmnožiny z množiny p regresorov. Pre n pozorovaní sa hodnota Mallowsovej štatistiky 4.26 mení na základe počtu k regresorov v podmodeli a hodnoty RSS podmodelu.

Navrhuje sa zobraziť hodnoty C_p štatistiky podľa veľkosti podmodelov a zároveň v grafe zobraziť úsečku z bodu $[0,0]$ do bodu $[p,p]$. Podľa vzťahu 4.25 by sa hodnoty C_p pre málo vychýlené odhady mali pohybovať v blízkosti uvedenej úsečky (pre podmodely s m regresomi je to hodnota $\mathsf{E}(\mathsf{C}_p) = m$). V prípade väčšieho počtu regresorov p sa pre každý počet regresorov v podmodeli zobrazujú iba tie najbližie k úsečke. Názorné sú grafy v 5. kapitole.

Mallows v článku [8] uvádza, čo je možné z takéhoto grafu vydedukovať. Ak regresné koeficienty sú nezávislé a tie ktoré sú nenulové, dosahujú vyšších hodnôt (vzhľadom k smerodajnej odchýlke ich odhadu), potom sú hodnoty C_p (podmodelov z nich tvorených) blízke okolo uvedenej úsečky (vyššie - až spoločne vytvárajú dobrý odhad strednej hodnoty). Ďalší prípadom je, ak sú regresory x_1, x_2 a x_3 silne korelované vzájomne a podobne korelované s $\boldsymbol{Y}$. Potom každé dva regresory je možné vylúčiť ale nie všetky tri. Preto tieto všetky uvažované modely budú okolo úsečky (nižšie - pre menšie modely), pretože každý jeden regresor môže dobre odhadnúť strednú hodnotu. Obe situácie sú umelé, ale môžu niekedy pomôcť.

4.2. Diagnostika

Diagnostické metódy pomahajú zisťovať hodnoty z nameraných dat, ktoré sú v nejakom zmysle významne vzdialené od väčšiny nameraných hodnôt. Tieto hodnoty je možné rozdeliť do dvoch základných skupín. Prvé sú nazývané *odľahlými pozorovaniami* (v angl. outliers), ktoré sú odľahlé z pohľadu závislej premennej. Druhé hodnoty sú nazývané *vplyvnými pozorovaniami* (v angl. leverage points), ktoré sú odľahlé z pohľadu nezávislej premennej. Niekedy môžu byť obe skupiny zaujímavé z fyzikálneho hľadiska. Avšak ak sú odľahlé pozorovania značne vzdialené od strednej hodnoty, vypočítanej bez týchto hodnôt, môžu významne ovplyvniť výsledné riešenie metódou najmenších štvorcov. Napríklad je možné pri takejto analýze využiť metódu Jackknife (vždy by sa vypočítala stredná hodnota bez jedného pozorovania). Práve preto, že majú významný vplyv na riešenie metódou najmenších štvorcov je dôležitá ich diagnostika. V prípade jednorozmerného modelu je situácia s diagnostikou jednoduchšia, často sa dajú obe skupiny určiť z bodového grafu. No v prípade mnohorozmerného modelu je diagnostika týchto hodnôt značne komplikovanejšia a tu sú diagnostické metódy veľmi užitočné. V tejto práci budú uvedené nasledujúce dve diagnostiky:

- diagnostika pomocou projekčnej matice $\boldsymbol{H}$
- studentizované reziduá

4.2.1. Diagnostika pomocou projekčnej matice H

Z kapitoly 2. je známe, že projekčná matica $\boldsymbol{H}$ je idempotentná a symetrická. Preto platí tiež rovnosť $\boldsymbol{H}\boldsymbol{H}' = \boldsymbol{H}$, z ktorej priamo vyplýva, že každý riadok $\boldsymbol{h}_{i\bullet}$ matice $\boldsymbol{H}$ má hodnotu $\|\boldsymbol{h}_{i\bullet}\|^2$ uvedenú v prvku h_{ii}. Zo vzťahu 2.7 je známa platnosť vzťahu $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, ktorý je možné ekvivalentne formulovať nasledovne

$$\hat{Y}_i = \boldsymbol{h}_{i\bullet}\boldsymbol{Y}, \quad i = 1, \ldots, n.$$

Z predchádzajúceho vzťahu je teda zrejmé, že jednotlivé odhady $\hat{Y}_i$ vektora $\hat{\boldsymbol{Y}}$ sú závislé na váhach podľa riadku $\boldsymbol{h}_{i\bullet}$ (na vektore pozorovaní $\boldsymbol{Y}$). Rovnako stĺpec $\boldsymbol{h}_{\bullet j}$ matice $\boldsymbol{H}$ informuje o vplyve pozorovania Y_j na všetky odhady vektora $\hat{\boldsymbol{Y}}$. A mierou tohto vplyvu pozorovania Y_j na odhady $\hat{\boldsymbol{Y}}$ môže byť diagonálny prvok h_{ii}.

Pre hodnotu h_{ii} platia nerovnosti $0 \leq h_{ii} \leq 1$. Prvá nerovnosť vyplýva z hore uvedenej normy $\|\boldsymbol{h}_{i\bullet}\|^2$. Druhá nerovnosť vyplýva z nerovnosti

$$h_{ii} = \|\boldsymbol{h}_{i\bullet}\|^2 \geq h_{ii}^2,$$

ktorá môže platiť len pre h_{ii} nezáporné a $h_{ii} \leq 1$. Nerovnosti $0 \leq h_{ii} \leq 1$ budú využité v nasledujúcej úvahe.

Pre odhad $\boldsymbol{u}$ náhodnej chyby $\boldsymbol{e}$ platí $\boldsymbol{u} = \boldsymbol{M}\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{e}$. Túto rovnosť je možné ekvivalentne formulovať nasledovne

$$u_i = (1 - h_{ii})e_i - \sum_{j \neq i} h_{ij}e_j, \quad i, j = 1, \ldots, n.$$

Z predchádzajúceho vzťahu vyplýva, že pre hodnoty h_{ii} blízke 1, náhodná chyba e_i v pozorovaní $\hat{Y}_i$ nie je vysvetlená v odhade u_i tejto chyby.

Z výsledkov Gaussovej-Markovovej vety a vety 2.6. plynú vzťahy pre variančné matice $\mathsf{Var}\hat{\boldsymbol{Y}} = \sigma^2\boldsymbol{H}$ a $\mathsf{Var}(\boldsymbol{u}) = \sigma^2\boldsymbol{M} = \sigma^2(\boldsymbol{I}-\boldsymbol{H})$. Z úvah, ktoré boli urobené v tomto odstavci, je možné usúdiť nasledujúce tvrdenie. V prípade hodnoty h_{ii} blízkej 1, je rozptyl odhadu $\hat{Y}_i$ väčší ale zároveň rozptyl odhadu u_i menší. A rovnako to platí naopak.

4.2.2. Studentizované reziduá

Najprv bude zavedené nové označenie. $\boldsymbol{X}_{(i)}$ je označenie pre maticu typu $(n-1)\times k$, ktorá vznikla odobraním i-tého riadku matice $\boldsymbol{X}_{n\times k}$. $\boldsymbol{Y}_{(i)}$ značí vektor bez i-tého pozorovania. Nech je uvažovaný normálny lineárny regresný model

$$\boldsymbol{Y}_{(i)} = \boldsymbol{X}_{(i)}\boldsymbol{\beta}_{(i)} + \boldsymbol{e}_{(i)}, \tag{4.27}$$

kde $\boldsymbol{\beta}_{(i)}$ sú regresné koeficienty a pre vektor chýb platí $\boldsymbol{e}_{(i)} \sim \mathsf{N}_{(n-1)}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Nech je ďalej uvažovaný normálny lineárny regresný model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{j}_i\gamma + \boldsymbol{e}, \tag{4.28}$$

kde $(\boldsymbol{\beta}', \gamma)$ sú regresné koeficienty, $\boldsymbol{j}_i$ je i-tý stĺpec jednotkovej matice $\boldsymbol{I}_n$ a pre vektor chýb platí $\boldsymbol{e} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Ďalej bude uvažovaný prípad $\mathrm{h}(\boldsymbol{X}_{(i)}) = \mathrm{h}(\boldsymbol{X}) = k$, ktorý znamená, že pre vyradený riadok je hodnota parametrickej funkcie $\boldsymbol{x}_{i\bullet}\boldsymbol{\beta}_{(i)}$ odhadnuteľná. Potom je možné tvrdiť, že odhady stredných hodnôt $\boldsymbol{X}_{(i)}\boldsymbol{b}_{(i)}$ a $\boldsymbol{X}\boldsymbol{b} + \boldsymbol{j}_i c$ modelov 4.27 a 4.28 metódou najmenších štvorcov sú totožné. Totiž pozorovanie i nemá v druhom modeli 4.28 žiadny vplyv, pretože odhad c koeficientu γ nadobudne takú hodnotu, aby minimalizoval reziduum v bode $\boldsymbol{x}_{i\bullet}$. Odhad c teda vyjadruje reziduum v prvom modeli 4.27 v bode $\boldsymbol{x}_{i\bullet}$.

Pre prvý model sa vypočíta odhad $\boldsymbol{b}_{(i)}$ regresných koeficientov $\boldsymbol{\beta}_{(i)}$ nasledovne

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

Potom stredná hodnota takéhoto modelu v bode $\boldsymbol{x}_{i\bullet}$ je $\hat{Y}_{(i)} = \boldsymbol{x}_{i\bullet}\boldsymbol{b}_{(i)}$. A nakoniec bude zavedená hodnota $u_{(i)} = Y_i - \hat{Y}_{(i)}$, ktorá je totožná s odhadom c v druhom modeli.

Za podmienky $\mathrm{h}(\boldsymbol{X}_{(i)}) = \mathrm{h}(\boldsymbol{X})$ je druhý model 4.28 nadmodelom modelu 4.27, ak platí $\mathrm{h}(\boldsymbol{X}_{(i)}) < \mathrm{h}((\boldsymbol{X}, \boldsymbol{j}_i))$. Toto môže byť ďalej predpokladané, inak by sa jednalo o nezaujímavý prípad, kedy nový stĺpec regresnej matice neposkytuje novú informáciu (platilo by $c = 0$ pre nadmodel). Podľa vzťahu 4.7 a označeného odhadu $\boldsymbol{c}_G$ rozširujúceho modelu totiž platí

$$\boldsymbol{c}_G = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-}\boldsymbol{Z}'\boldsymbol{u}. \tag{4.29}$$

Pre odhad c zo vzťahu 4.29 teda platí

$$c = (\boldsymbol{j}'_i\boldsymbol{M}\boldsymbol{j}_i)^{-}\boldsymbol{j}'_i\boldsymbol{u} = \frac{u_i}{m_{ii}}. \tag{4.30}$$

V predchádzajúcom vzťahu sa formálne jednalo o pseudoinverziu z regulárnej štvorcovej matice (jednoprvkovej), ktorá je totožná s inverznou maticou. Podľa vety, ktorú uviedol Zvára v literatúre ([13], str. 99), je podmienka $\mathrm{h}(\boldsymbol{X}_{(i)}) = \mathrm{h}(\boldsymbol{X})$ ekvivalentná s tvrdením $m_{ii} > 0$. Preto môže byť uvažovaný vzťah 4.30. Použitím vzťahu $\mathsf{Var}(\boldsymbol{u}) = \sigma^2\boldsymbol{M}$ z vety 2.6. je možné vyjadriť vzťah

$$\mathsf{Var}(c) = \mathsf{Var}\left(\frac{u_i}{m_{ii}}\right) = \frac{\mathsf{Var}(u_i)}{m_{ii}^2} = \frac{\sigma^2 m_{ii}}{m_{ii}^2} = \frac{\sigma^2}{m_{ii}}. \tag{4.31}$$

Keďže u_i má normálne rozdelenie a platia vzťahy 4.30 a 4.31, potom náhodná veličina $u_i/(\sigma\sqrt{m_{ii}})$ má rozdelenie $\mathsf{N}(0,1)$. Nech $s_{(i)}$ značí odhad smerodatnej odchýlky v prvom modeli, potom náhodná veličina $s^2_{(i)}/\sigma^2$ má $\chi^2(n-1-k)$ rozdelenie. Podľa vety 2.10. sú odhady $\boldsymbol{b}_{(i)}$ nezávislé na odhade $s^2_{(i)}$, preto aj odhady $\hat{Y}_{(i)} = \boldsymbol{x}_{i\bullet}\boldsymbol{b}_{(i)}$ a $s^2_{(i)}$ sú na sebe nezávislé. A keďže obe tieto štatistiky boli vypočítané bez znalosti Y_i, je možné usúdiť, že sú všetky tri Y_i, $\hat{Y}_{(i)}$ a $s^2_{(i)}$ vzájomne nezávislé. Zo vzťahu 4.30 vyplýva vzťah

$$u_i = cm_{ii} = (Y_i - \hat{Y}_{(i)})m_{ii}.$$

Z predchádzajúceho vzťahu a tvrdení vyplýva nezávislosť medzi veličinami u_i a $s^2_{(i)}$. Potom je možné zaviesť novú štatistiku

$$u_i^* = \frac{u_i/(\sigma\sqrt{m_{ii}})}{\sqrt{s^2_{(i)}/\sigma^2}} = \frac{u_i}{\sqrt{m_{ii}}s_{(i)}} \sim \mathsf{t}(n-k-1), \tag{4.32}$$

ktorá odpovedá definovanému Studentovmu rozdeleniu podľa prvého vyjadrenia. Táto štatistika sa nazýva *studentizované reziduum*. V prípade, kedy je vopred známy index i, je možné klasicky testovať t-testom odľahlosť pozorovania. Hypotéza o neodľahlosti pozorovania i je zamietnutá na hladine α ak platí $|u_i^*| > \mathsf{t}_{1-\alpha/2}(n-k-1)$. No v prípade, kedy nie je vopred známy index i a jeho výber je závislý na pozorovaní $\boldsymbol{Y}$, je potrebné testovanie mierne pozmeniť.

Nech je pre $\delta \in (0,1)$ uvažovaných n náhodných javov

$$A_i(\delta) = \{|u_i^*| > \mathsf{t}_{1-\delta/2}(n-k-1)\}, \quad i = 1,2,\ldots,n.$$

Podľa Bonferroniho nerovnosti 1.1 platí pri zvolenej hladine α

$$\mathsf{P}(\cup_{i=1}^n A_i(\alpha/n)) \leq \sum_{i=1}^n \mathsf{P}(A_i(\alpha/n)) = \alpha.$$

Ak bude vybraný index j na základe štatistík z pozorovaní $\boldsymbol{Y}$, bude hypotéza o neodľahlosti pozorovania od strednej hodnoty $\hat{\boldsymbol{Y}}$ zamietnutá ak $\left|u_j^*\right| > \mathsf{t}_{1-\alpha/(2n)}(n-k-1)$.

5. Analýza experimentu

Experimet bol vyhodnotený s použitím programovacieho jazyka a prostedia R. Ide o voľne dostupný jazyk a software, ktorý je určený primárne na štatistické výpočty a grafické zobrazovanie. Pre prácu v tomto prostredí je možné priamo využiť dostupnú konzolu. Avšak ako nadstavbové vyvojové prostredie bolo použité RStudio, ktoré je taktiež voľne šíriteľné. Toto vývojové prostredie je možno doporučiť pre jednoduchosť ovládania oproti konzole (vzhľadom aj funkcionalitou pripomína vývojové prostedie MATLABu).

Po dohode s odborníkom v danej oblasti experimentu, bol zadaný problém, ktorý bolo potrebné vyriešiť. Jednalo sa o zistenie najlepšej kombinácie fosforečnanových taviacich solí v zmesi tak, aby bola zaistená maximálna tvrdosť. Prípadne nájsť takú množinu, ktorá bude takému zadaniu s nejakou pravdepodobnosťou odpovedať.

Nezávislé premenné x_1, x_2 a x_3 budú vyjadrovať pomer určitej zložky na zmesi a teda bude platiť $x_1 + x_2 + x_3 = 1$.

V prípade vyhodnocovania experimentu troch zložiek (napr. zmesi) sa používajú ternárne diagramy. Výhodou je napríklad jednoduchá interpretovateľnosť grafických výsledkov. Avšak pre zadanú úlohu je lepšie uvažovať dve zložky, pretože sa týmto nemení stredná hodnota, ktorá má byť vyhodnocovaná. Takýto záver je možné urobiť na základe vzťahu 4.9

$$\boldsymbol{d} = \hat{\boldsymbol{Y}}_G - \hat{\boldsymbol{Y}} = \boldsymbol{M}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-}\boldsymbol{Z}'\boldsymbol{u}, \tag{5.1}$$

ktorý určoval vektor vzdialenosti dvoch odhadov, ak jeden je širší (v uvažovanom prípade bol skutočný). Nech teda regresná matica $\boldsymbol{X}$ (projekčnej matici $\boldsymbol{M}$ vo vzťahu 5.1) je z modelu

$$\mathsf{E}Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2. \tag{5.2}$$

No je uvažovaný i širší model s hotnotami x_3. Takýto model by rozšíril strednú hodnotu nasledovne

$$\mathsf{E}Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_2 x_3 + \beta_6 x_1 x_3 + \beta_7 x_1^2 + \beta_8 x_2^2 + \beta_9 x_3^2. \tag{5.3}$$

Ale člen x_3 je možné vyjadriť v tvare $1 - x_1 - x_2$, no nevznikne takto žiadny nový člen a bude platiť

$$\begin{aligned} \mathsf{E}Y \;=\; & (\beta_0 + \beta_3 + \beta_9) + (\beta_1 - \beta_3 + \beta_6 - 2\beta_9)x_1 + (\beta_2 - \beta_3 + \beta_5 - 2\beta_9)x_2 + \\ & + (\beta_7 - \beta_6 + \beta_9)x_1^2 + (\beta_4 - \beta_5 - \beta_6 + 2\beta_9)x_1 x_2 + (\beta_8 - \beta_5 + \beta_9)x_2^2. \end{aligned} \tag{5.4}$$

Ako je vidieť, takto uvažovaný model (s podmientkou $x_1 + x_2 + x_3 = 1$) totiž nerozširuje priestor strednej hodnoty. Preto by bola vo vzťahu 5.1 matica $\boldsymbol{Z}$ nulová. Ale aj keby bol uvažovaný nadmodel, kde $\boldsymbol{Z}$ by bola tvorená vektorom $\boldsymbol{x}_3$, platilo by $\boldsymbol{x}_3 \in \mathcal{M}(\boldsymbol{1}, \boldsymbol{x}_1, \boldsymbol{x}_2)$ a teda $\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{0}$ vo vzťahu 5.1.

Nech sú odhady pre strednú hodnotu 5.2 označené v hornom indexe *. Keďže 5.2 a 5.4 vyjadrujú totožné stredné hodnoty, platí $\beta_0^* = \beta_0 + \beta_3 + \beta_9$ a podobne ďalších päť vzťahov. Odhady 5.2 sú jednoznačné. Teda sa jedná o sústavu šiestich rovníc pre desať neznámych. Čo nasvedčuje tomu, že neexistuje jednoznačné vyjadrenie pre vyjadrenie strednej hodnoty 5.3.

Podľa 3D bodového grafu, kde boli zobrazené hodnoty $\boldsymbol{x}_1$ a $\boldsymbol{x}_2$ na $\boldsymbol{Y}$, je zrejmé, že odhad strednej hodnoty rovinou nebude postačujúci. Pretože je takýto graf ťažko interpretovateľný do roviny, nebude tu uvedený, ale rozloženie dát je možné pozorovať na ďalších

stránkach s už odhadnutou strednou hodnotou. Nech je teda uvažovaná stredná hodnota tvaru 5.2 i vyššieho stupňa. Za regresor x_1 bude uvažovaný pomer hydrogenfosforečnanu sodného zo zmesi a za regresor x_2 pomer difosforečnanu sodného. Pozorovania $\boldsymbol{Y}$ sú pre tvrdosť, 8 týždňov zrenia a 30 dní skladovania.

5.1. Polynóm druhého stupňa

Pomocou dvoch funkcií summary(lm(y x1 + x2)) je možné veľmi rýchlo získať základné informácie o uvedenom modeli (v tejto funkcii) v jazyku R. Ak je uvažovaný model so strednou hodnotou 5.2, potom sú k dispozícii štatistiky vo výstupe 1. Väčšina bola zavedená v 2. kapitole. Sú tu informácie o reziduách (napr. medián a kvartilové rozpätie), jednotlivé odhady regresných koeficientov metódou najmenších štvorcov a smerodajné odchýlky týchto odhadov. Následne hodnoty pre t-test a príslušné p-hodnoty. Hviezdičky určujú významnosť regresoru v modeli na hladine α podľa uvedenej legendy. Nižšie sú uvedené informácie o odhade smerodajnej odchýlky chyby modelu a koeficient determinácie R^2 (i adjustovaný R^2_{adj}) ako bol uvedený v 2.19 v prípade $\mathbf{1} \in \mathcal{M}(X)$. Posledný je F-test, ktorý testuje hypotézu, že všetky regresné koeficienty sú nulové.

Výstup z R 1: Štatistiky pre model so strednou hodnotou s regresormi do 2. stupňa

```
Call:
lm(formula = y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x2^2))

Residuals:
    Min      1Q  Median      3Q     Max
-3.1820 -0.7955  0.0686  0.7832  3.2342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.098      0.482  12.651   <2e-16 ***
x1            -1.110      2.036  -0.545   0.5866
x2            -2.132      2.036  -1.047   0.2971
I(x1^2)       -4.830      1.933  -2.499   0.0137 *
I(x1 * x2)    37.549      3.156  11.898   <2e-16 ***
I(x2^2)       -1.991      1.933  -1.030   0.3048
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.223 on 126 degrees of freedom
Multiple R-squared:  0.8213,Adjusted R-squared:  0.8142
F-statistic: 115.8 on 5 and 126 DF,  p-value: < 2.2e-16
```

F-test jasne zamieta súčasnú nulovosť všetkých regresorov. Ako významné na hladine α vyšli tri regresory. Korelačný koeficient medzi $\boldsymbol{x}_{\bullet 2}$ a $\boldsymbol{x}_{\bullet 4}$ vyšiel 0.9481194. Teda regresory x_1, x_1^2 sú silne lineárne závislé a rovnako regresory x_2, x_2^2.

Pomocou Krokovej regresie bol vybraný podmodel zložený zo 4 regresorov. Zdrojový kód použitej metódy je v prílohe. Metóda použila len výber regresorov v kroku 3 v takomto poradí: absolutný člen, x_1x_2, x_1^2 a x_2^2. Následne bude uvedené štatistiky pre tento vybraný model vo výstupe 2.

Výstup z R 2: Štatistiky pre vybraný model Krokovou regresiou

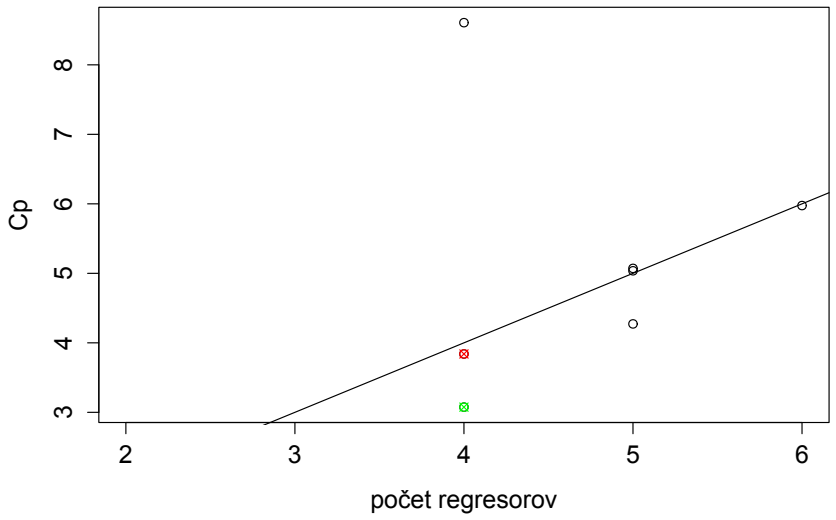
```
Call:
lm(formula = y ~ I(x1^2) + I(x1 * x2) + I(x2^2))

Residuals:
    Min      1Q  Median      3Q     Max
-3.1424 -0.7417  0.0556  0.8026  3.2432

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6902     0.1991  28.581  < 2e-16 ***
I(x1^2)      -5.5444     0.5165 -10.735  < 2e-16 ***
I(x1 * x2)   34.9663     1.4857  23.536  < 2e-16 ***
I(x2^2)      -3.8927     0.5165  -7.537 7.72e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.219 on 128 degrees of freedom
Multiple R-squared:  0.8198,Adjusted R-squared:  0.8155
F-statistic: 194.1 on 3 and 128 DF,  p-value: < 2.2e-16
```

Podľa štatistík, je vybraný model vhodný. Hodnota R^2 klesla nepatrne o 0.0015, ale model bol redukovaný o dva regresory. Všetky regresory sú významné.

V prípade Mallowsovej C_p štatistiky je uvažované, že odhad strednej hodnoty z pôvodného modelu je nestranným odhadom. Bude teda použitý vzťah 4.26

$$\mathsf{C}_p = 2k - n + \frac{RSS}{s_p^2},$$

kde $s_p = 1.223$ z výstupu 1. Graf spomínaný v časti o Mallowsovej štatistike je zobrazený v grafe 1, kde je taktiež zobrazený bod pre vybraný model na zeleno. Sú tu vyobrazené hodnoty C_p menšie ako 10. Len informatívne, najväčšia bola hodnota 578.5447. V grafe 1 je vidieť, že k strednej hodnote C_p je bližšia hodnota s počtom regresorov 4 označená na červeno. Ide o model s regresormi: absolutný člen, x_2, x_1x_2 a x_1^2. Teda oproti Krokovej regresii ($\mathsf{C}_p = 3.074860$) bol vybraný za lepší z pohľadu C_p ($\mathsf{C}_p = 3.839803$) regresor x_2 miesto x_2^2. To nie je nič prekvapivé, keďže oba regresory sú silno lineárne závislé (korelačný koeficient vyšiel 0.9481194). Základné štatistiky pre takto vybraný model sú uvedené vo výstupe 3. Keďže základné štatistiky oboch modelov sú veľmi podobné, bude ďalej uvažovaný model preferovaný na základe Mallowsovej štatistiky. V grafe 2 je uvedený 3D bodový graf s odhadnutou strednou hodnotou. V grafe 3 sú názornejšie zobrazené na-



Graf 1: Zobrazené C_p štatistiky pre pôvodný model. Zelenou farbou je vyznačený bod vybraný Krokovou regresiou.

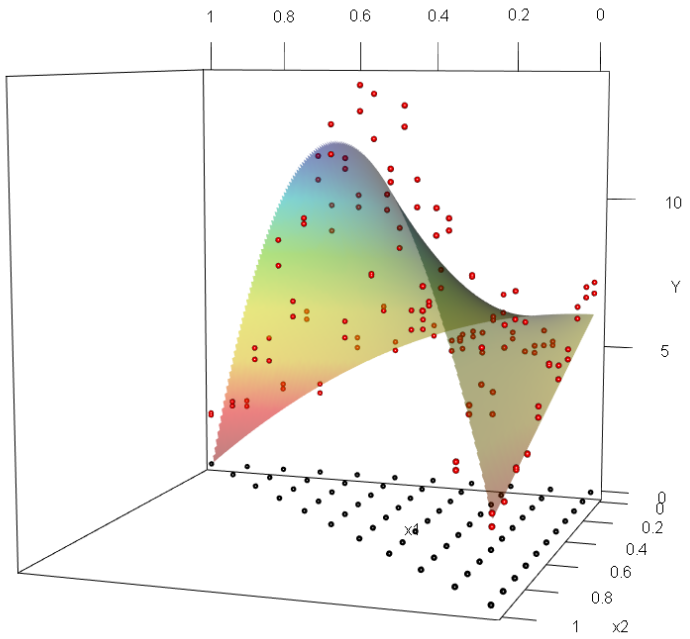
merané dáta $\boldsymbol{Y}$ s odhadovanými $\hat{\boldsymbol{Y}}$. Postupnosť zmesí (v grafe 3) je určená pre dvojicu x_1, x_2 nasledovne

$$(1,0),(0.9,0.1),(0.8,0.2),\ldots,(0.1,0.9),(0,1),(0.9,0),(0.8,0.1),\ldots,(0.1,0),(0,0.1),(0,0).$$

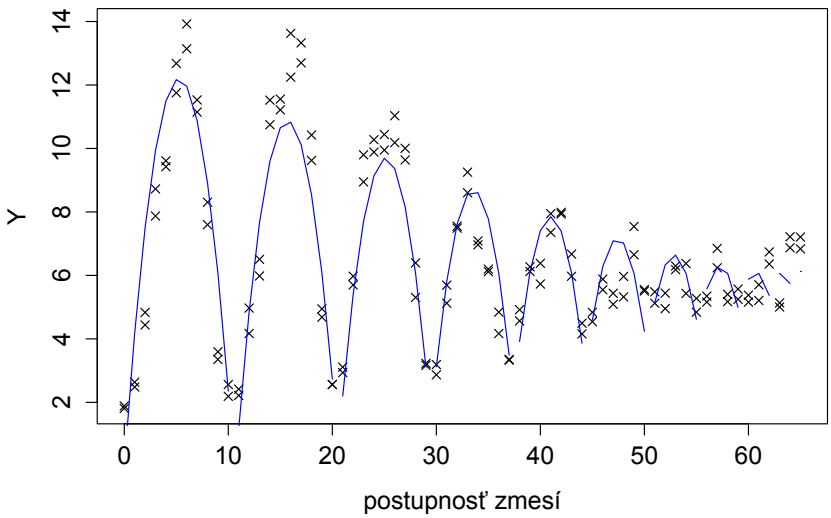
Výstup z R 3: Štatistiky pre vybraný model Krokovou regresiou

```
Call:
lm(formula = y ~ I(x2) + I(x1^2) + I(x1 * x2))

Residuals:
    Min      1Q  Median      3Q     Max
-3.0640 -0.7895  0.0642  0.7468  3.2037

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.1247     0.2376  25.776  < 2e-16 ***
I(x2)        -3.7697     0.5051  -7.463 1.14e-11 ***
I(x1^2)      -6.1307     0.5586 -10.975  < 2e-16 ***
I(x1 * x2)   37.8453     1.6264  23.269  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.222 on 128 degrees of freedom
Multiple R-squared:  0.8187,Adjusted R-squared:  0.8144
F-statistic: 192.6 on 3 and 128 DF,  p-value: < 2.2e-16
```

Z uvedených grafov 2, 3 vyplýva, že odhadnutá stredná hodnota nepokrýva dostatočne maximálne hodnoty $\boldsymbol{Y}$. Preto je vhodné, pre úlohu najdenia najväčšej tvrdosti, prejsť na model vyššieho stupňa.



Graf 2: 3D bodový graf pre model vybraný s pomocou C_p štatistiky. V grafe je zobrazená stredná hodnota určená metódou najmenších štvorcov.



Graf 3: Zobrazené sú hodnoty $\boldsymbol{Y}$ a po častiach spojitá lineárna funkcia, ktorá nadobúda rovnaké hodnoty s $\hat{\boldsymbol{Y}}$ pre zmesy. Postupnosť zmesí je určená v texte vyššie.

5.2. Polynóm tretieho stupňa

Nech je uvažovaný model so strednou hodnotou

$$\mathsf{E}Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2 + \beta_6 x_1^3 + \beta_7 x_1^2 x_2 + \beta_8 x_1 x_2^2 + \beta_9 x_2^3 \quad (5.5)$$

Základné štatistiky pre model 5.5 sú uvedené vo výstupe 4. Je zrejmé, že odhad smerodajnej odchýlky klesol, pretože sú uvažované nezávislé regresory naviac (4 oproti modelu 5.2).

Výstup z R 4: Štatistiky pre celý model

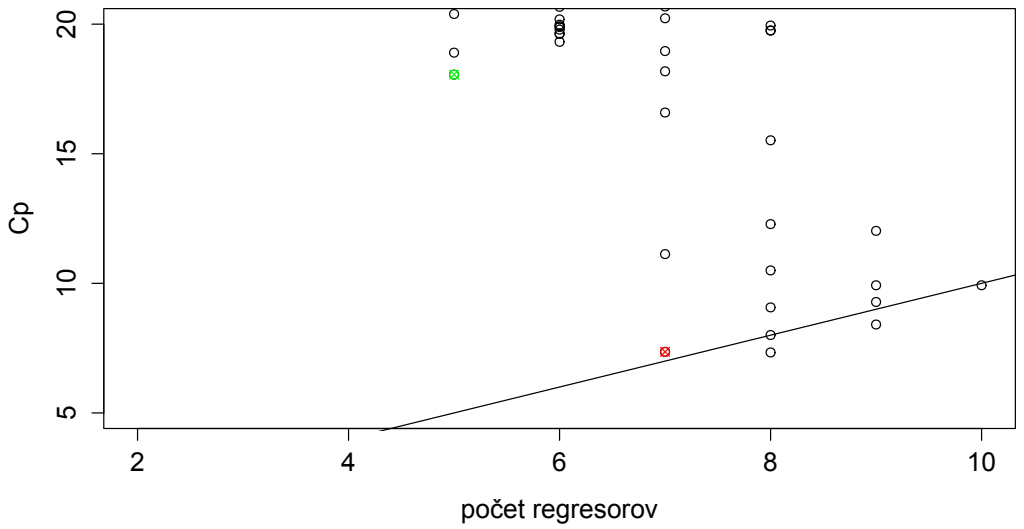
```
Call:
lm(formula = Y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x2^2) + I(x1^3) +
    I(x1^2 * x2) + I(x1 * x2^2) + I(x2^3))

Residuals:
    Min      1Q  Median      3Q     Max
-2.4686 -0.6791  0.0833  0.6273  2.5341

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.4283     0.5767  12.881  < 2e-16 ***
x1           -19.4881     4.6946  -4.151 6.16e-05 ***
x2            -9.5078     4.6946  -2.025 0.045024 *
I(x1^2)       42.4275    10.9122   3.888 0.000165 ***
I(x1 * x2)    78.2282    17.4705   4.478 1.71e-05 ***
I(x2^2)       12.7081    10.9122   1.165 0.246462
I(x1^3)      -30.2817     7.1199  -4.253 4.16e-05 ***
I(x1^2 * x2) -61.4758    16.5451  -3.716 0.000307 ***
I(x1 * x2^2) -11.5376    16.5451  -0.697 0.486914
I(x2^3)      -10.0715     7.1199  -1.415 0.159742
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.084 on 122 degrees of freedom
Multiple R-squared:  0.8641,Adjusted R-squared:  0.8541
F-statistic: 86.22 on 9 and 122 DF,  p-value: < 2.2e-16
```

Výber regresorov bude uskutočnený najprv na základe Krokovej regresie. Metóda pre $\alpha = 0.05$ postupne vybrala nasledovné regresory: absolutný člen, x_1x_2, x_3^1, $x_1x_2^2$ a x_3^2. Rovnaké regresory vybrala aj pri $\alpha = 0.1$. Výpis štatistík pre takýto model sa nachádza vo výstupe 5.

V grafe 4 sú zobrazené všetky C_p štatistiky, ktoré majú menšiu hodnotu ako 20. Je vidieť, že model vybraný Krokovou regresiou, má najlepšiu C_p štatistiku ($\mathsf{C}_p = 18.0526$) z ostatných podmodelov s piatimi regresormi. Ďalej je červenou farbou označený bod, ktorý má síce viac regresorov ale dobre dpovedá strednej hodnote Mallowsovej štatistiky ($\mathsf{C}_p = 7.356668$). Tento model zahŕňa regresory: absolutný člen, x_1, x_2, x_2^1, x_1x_2, x_3^1 a $x_1^2x_2$. Budú teda uvedené štatistiky pre tento model vo výpise 6.

Výstup z R 5: Štatistiky pre celý model

```
Call:
lm(formula = Y ~ I(x1 * x2) + I(x1^3) + I(x1 * x2^2) + I(x2^3))

Residuals:
     Min       1Q   Median       3Q      Max
-2.39084 -0.60531  0.01664  0.70441  2.67335

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.5548     0.1741  31.912  < 2e-16 ***
I(x1 * x2)    18.0726     3.3691   5.364 3.72e-07 ***
I(x1^3)       -5.0751     0.5329  -9.524  < 2e-16 ***
I(x1 * x2^2)  28.6603     6.5276   4.391 2.36e-05 ***
I(x2^3)       -5.1508     0.5915  -8.709 1.39e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.138 on 127 degrees of freedom
Multiple R-squared:  0.8439,Adjusted R-squared:  0.839
F-statistic: 171.7 on 4 and 127 DF,  p-value: < 2.2e-16
```



Graf 4: Zobrazené C_p štatistiky pre pôvodný model. Zelenou farbou je vyznačený bod vybraný Krokovou regresiou. Červenou farbou je označená štatistika modelu, ktorý je z pohľadu C_p a teda samotného výberu modelu zaujímavý.

Štatistiky druhého uvažovaného modelu vo výpise 6 sa zlepšili (napr. odhad smerodajnej odchýlky alebo hodnota R^2). Keďže je v modeli silnejšia lineárna závislosť niektorých

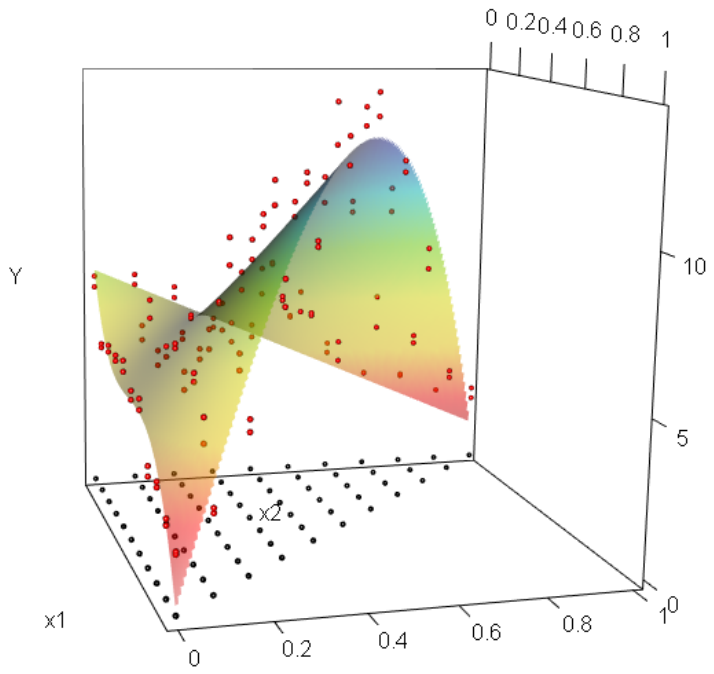
regresorov, nie je prekvapujúce, že v podstate iné regresory dokážu rovnako dobre odhadnúť strednú hodnotu. No kvôli odhadu strednej hodnoty by bolo rozumnejšie sa držať navrhovaného modelu podľa C_p. Síce má o dva regresory viac ale menšie vychýlenie podľa Mallowsovej štatistiky. Bude teda zobrazený bodový graf s odhadnutou strednou hodnotou v grafe 5. Bodový graf má body označené červenou farbou. Čiernou farbou sú označené body vyjadrujúce všetky sledované zmesy.

Výstup z R 6: Štatistiky pre vybraný model na základe C_p

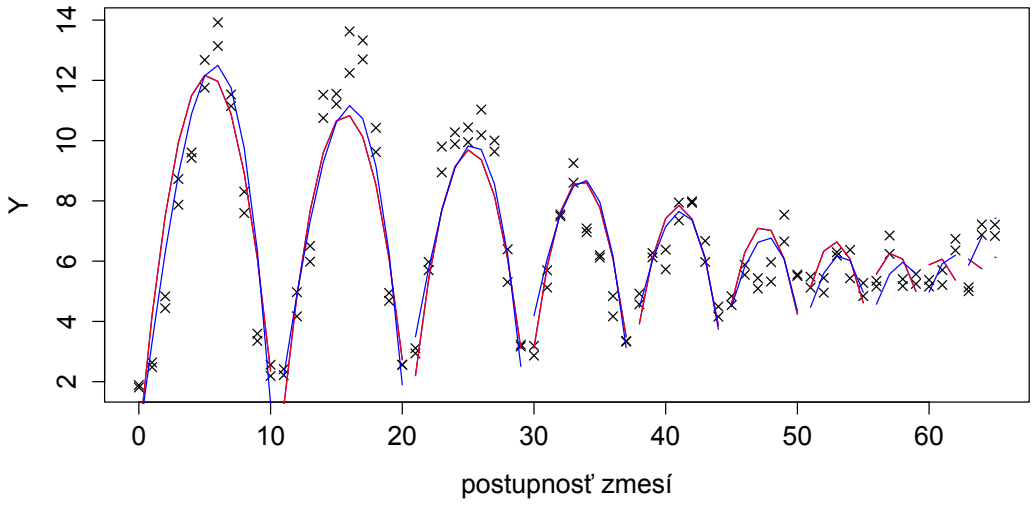
```
Call:
lm(formula = Y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x1^3) + I(x1^2 * x2))

Residuals:
    Min      1Q  Median      3Q     Max
-2.9675 -0.7213  0.1014  0.6523  2.5963

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.4256     0.3931  18.891  < 2e-16 ***
x1           -19.5647     3.9397  -4.966 2.19e-06 ***
x2            -6.1460     0.6609  -9.300 5.87e-16 ***
I(x1^2)       42.5981    10.0210   4.251 4.13e-05 ***
I(x1 * x2)    72.4134     5.8947  12.284  < 2e-16 ***
I(x1^3)      -30.3730     6.8037  -4.464 1.77e-05 ***
I(x1^2 * x2) -59.0026     9.8051  -6.018 1.82e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.086 on 125 degrees of freedom
Multiple R-squared:  0.8603,Adjusted R-squared:  0.8536
F-statistic: 128.3 on 6 and 125 DF,  p-value: < 2.2e-16
```

Ani v predchádzajúcom ani v tomto vybranom modeli studentizované reziduá na hladine $\alpha = 0.05$ nezaznamenali odľahlé pozorovania $\boldsymbol{Y}$.

V grafoch 5 a 6 je vidieť, že sa podarilo lepšie zachytiť strednú hodnotu pre zmesy s najväčšou tvrdosťou. Vzhľadom k analýze sa bude pokračovať k vyššiemu stupňu polynómu a budú pozorne sledované zmeny. Ak budú zmeny minimálne, bude preferovaný model vyberaný v tomto odstavci.



Graf 5: Zobrazený je 3D bodový graf s odhadnutou strednou hodnotou polynómom tretieho stupňa.



Graf 6: Zobrazené sú hodnoty $\boldsymbol{Y}$ a červenou farbou je označená funkcia totožná s funkciou z grafu 3 (odhad polynómom 2. stupňa). Modrou farbou je označená po častiach spojitá lineárna funkcia, ktorá nadobúda hodnoty $\hat{\boldsymbol{Y}}$ (odhad vybraného polynómu 3. stupňa) v príslušných zmesiach. Postupnosť zmesí bola už určená.

5.3. Polynóm štvrtého stupňa

Nech je uvažovaný model so strednou hodnotou

$$\begin{aligned} \mathsf{E}Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2 + \beta_6 x_1^3 + \beta_7 x_1^2 x_2 + \beta_8 x_1 x_2^2 + \\ &\quad + \beta_9 x_2^3 + \beta_1 0 x_1^4 + \beta_1 1 x_1^3 x_2 + \beta_1 2 x_1^2 x_2^2 + \beta_1 3 x_1 x_2^3 + \beta_1 4 x_2^4. \end{aligned} \tag{5.6}$$

Základné štatistiky pre tento model sú vo výstupe 7. Podľa teórie s novými lineárne nezávyslými regresormi nevzrastie hodnota *RSS* (resp. väčšinou klesne). Presnejšie za predpokladu, že nové regresory majú priemety do reziduálneho priestoru pôvodného modelu a tieto priemety nie sú všetky ortogonálne na $\boldsymbol{Y}$, potom vždy poklesne hodnota *RSS*. Preto odhad s smerodajnej odchýlky klesne a hodnota R^2 vzrastie. Toto je pozorovateľné na uvedených štatistikách (vo výstupoch) pôvodných a rozšírených modelov.

Výstup z R 7: Štatistiky pre uplný model

```
Call:
lm(formula = Y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x2^2) + I(x1^3) +
    I(x1^2 * x2) + I(x1 * x2^2) + I(x2^3) + I(x1^4) + I(x1^3 * x2) +
    I(x1^2 * x2^2) + I(x1 * x2^3) + I(x2^4))

Residuals:
     Min       1Q   Median       3Q      Max
-2.08738 -0.51084 -0.02468  0.48056  2.06610

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.4436     0.5398  11.938  < 2e-16 ***
x1               -6.1838     7.4585  -0.829 0.408739
x2                9.0341     7.4585   1.211 0.228240
I(x1^2)          12.8747    31.5518   0.408 0.683983
I(x1 * x2)     -151.1174    49.0644  -3.080 0.002580 **
I(x2^2)         -47.7718    31.5518  -1.514 0.132703
I(x1^3)         -19.0753    47.9855  -0.398 0.691707
I(x1^2 * x2)    451.6944   108.8529   4.150 6.35e-05 ***
I(x1 * x2^2)    457.8222   108.8529   4.206 5.12e-05 ***
I(x2^3)          59.6301    47.9855   1.243 0.216474
I(x1^4)           7.8460    23.7331   0.331 0.741543
I(x1^3 * x2)   -330.1447    71.7622  -4.601 1.08e-05 ***
I(x1^2 * x2^2) -445.1972   102.1194  -4.360 2.82e-05 ***
I(x1 * x2^3)   -286.3343    71.7622  -3.990 0.000116 ***
I(x2^4)         -25.7157    23.7331  -1.084 0.280798
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8616 on 117 degrees of freedom
Multiple R-squared:  0.9176,Adjusted R-squared:  0.9078
F-statistic: 93.11 on 14 and 117 DF,  p-value: < 2.2e-16
```

Ďalej bude použitá Kroková regresia. Prvýkrát metóda využila krok 5, kedy vylučovala už vybraný regresor. Metóda postupovala nasledovne:

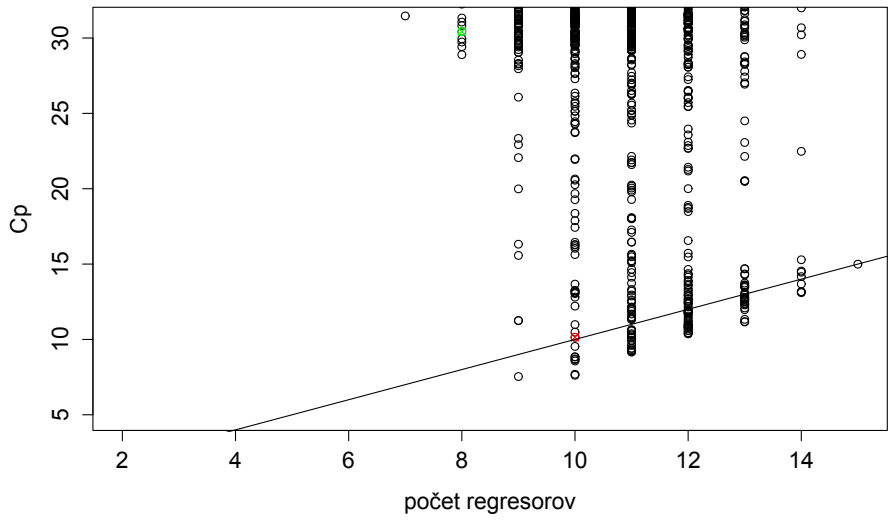
absolutný člen, x_1x_2, x_1^3, x_2^4, $x_1^3x_2$, $x_1^2x_2$, odobraný x_1x_2, $x_1x_2^2$, x_1, odobraný x_1^3, x_2, $x_1^2x_2^2$.

Štatistiky pre tento model, ktorý má 8 regresorov, sú uvedené vo výstupe 8. Prvýkrát Kroková regresia vybrala regresory, ktorých odhady koeficientov majú p-hodnoty väčšie ako 0.001. Doteraz veľmi striktný výber regresorov, založený na Krokovej regresii a F-testu ($\alpha = 0.05$), sa zoslabil dôsledkom počtu nezávislých regresorov. Napríklad veľmi významný regresor v menšom modeli, ktorý je korelovaný s novým regresorom, vo väčšom modeli už taký významný nemusí byť (vyššia p-hodnota).

Výstup z R 8: Základné štatistiky pre model vybraný Krokovou regresiou

```
Call:
lm(formula = Y ~ x1 + x2 + I(x1^2 * x2) + I(x1 * x2^2) + I(x1^3 *
    x2) + I(x1^2 * x2^2) + I(x2^4))

Residuals:
     Min       1Q   Median       3Q      Max
-1.93603 -0.52415 -0.03862  0.59403  2.47676

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.3781     0.2770  23.025  < 2e-16 ***
x1               -4.2619     0.5252  -8.114 4.08e-13 ***
x2               -2.6627     0.8382  -3.177 0.001880 **
I(x1^2 * x2)    188.6860    18.2909  10.316  < 2e-16 ***
I(x1 * x2^2)     24.3791     6.5946   3.697 0.000326 ***
I(x1^3 * x2)   -205.4114    19.7481 -10.402  < 2e-16 ***
I(x1^2 * x2^2)  -56.7671    25.6073  -2.217 0.028458 *
I(x2^4)          -2.3669     0.9168  -2.582 0.010996 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9363 on 124 degrees of freedom
Multiple R-squared:  0.8969,Adjusted R-squared:  0.8911
F-statistic: 154.1 on 7 and 124 DF,  p-value: < 2.2e-16
```

V grafe 7 sú označené C_p štatistiky do hodnoty 31. Zelenou farbou je tu označená hodnota $C_p = 30.42169$ vybraného modelu. Ďalej bola označená červenou farbou hodnota $C_p = 10.143824$, pre model s desiatimi regresormi, ktorá je s týmto počtom regresorov najbližšie k strednej hodnote C_p. Tento model má všetky regresory z modelu vybraného Krokovou regresiou a naviac regresory x_1x_2, $x_1x_2^3$. Základné štatistiky pre tento model sú uvedené vo výstupe 9. Tu je vidieť, že regresor x_2^4 má p-hodnotu približne 0.0805. Teda t-test na hladine menšej ako 0.08 by už nezamietol hypotézu o nulovosti príslušného koeficientu. Keďže je primárne záujem o odhad strednej hodnoty, bude ďalej uvažovaný model vybraný na základe C_p so všetkými desiatimi regresormi.

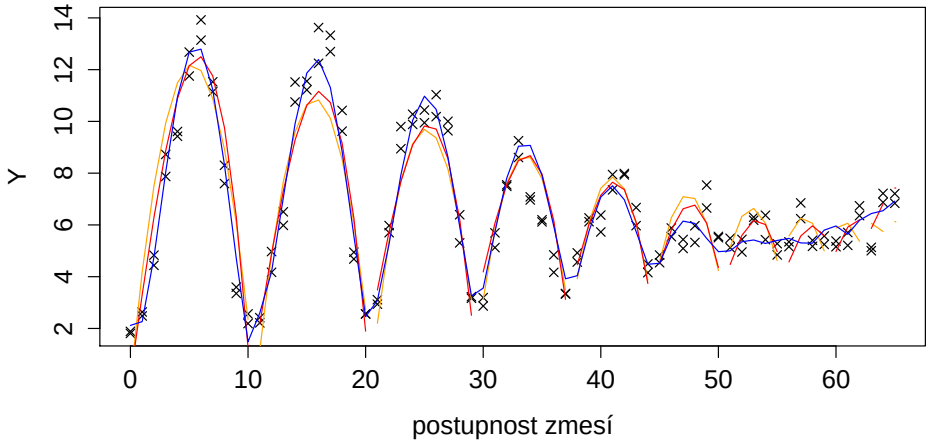


Graf 7: Zelenou farbou je označená hodnota C_p pre model vybraný Krokovou regresiou a červenou farbou vybraný nový model na základe C_p.

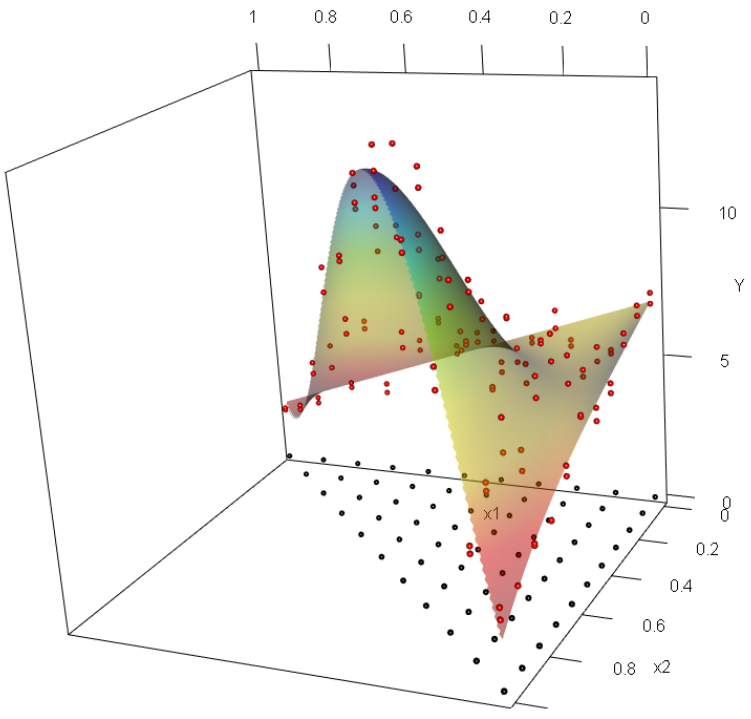
Výstup z R 9: Základné štatistiky pre model vybraný na základe C_p

```
Call:
lm(formula = Y ~ x1 + x2 + I(x1 * x2) + I(x1^2 * x2) + I(x1 * x2^2)
    + I(x1^3 * x2) + I(x1^2 * x2^2) + I(x1 * x2^3) + I(x2^4))

Residuals:
     Min       1Q   Median       3Q      Max
-2.10842 -0.54463 -0.01986  0.53136  2.06035

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.9191     0.2777  24.914  < 2e-16 ***
x1               -4.8015     0.4961  -9.678  < 2e-16 ***
x2               -3.6869     0.8712  -4.232 4.51e-05 ***
I(x1 * x2)     -109.6728    23.6793  -4.632 9.15e-06 ***
I(x1^2 * x2)    428.2087    61.3065   6.985 1.62e-10 ***
I(x1 * x2^2)    327.2563    62.0074   5.278 5.77e-07 ***
I(x1^3 * x2)   -335.9864    41.2304  -8.149 3.67e-13 ***
I(x1^2 * x2^2) -376.9328    70.8638  -5.319 4.80e-07 ***
I(x1 * x2^3)   -197.5333    42.7102  -4.625 9.40e-06 ***
I(x2^4)          -1.7454     0.9903  -1.762   0.0805 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8621 on 122 degrees of freedom
Multiple R-squared:  0.914,Adjusted R-squared:  0.9077
F-statistic: 144.1 on 9 and 122 DF,  p-value: < 2.2e-16
```

V grafoch 8 a 9 je vidieť, že sa podarilo zlepšiť odhad strednej hodnoty i pre zmesy s najvyššou tvrdosťou.

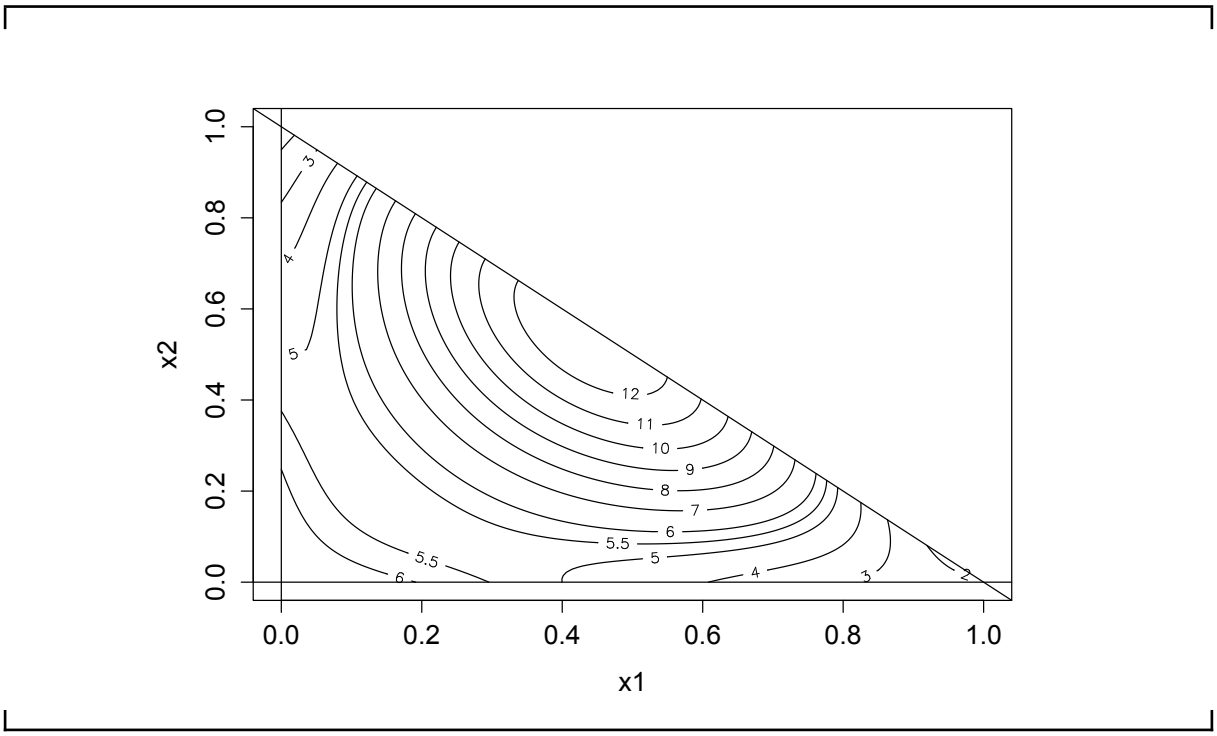


Graf 8: Funkcie označené červenou a oranžovou farbou sú funkcie z grafu 6 (oranžová odhad polynómom 2. stupňa, červená odhad polynómom 3. stupňa). Modrou farbou je označená po častiach spojitá lineárna funkcia, ktorá nadobúda hodnoty $\hat{\boldsymbol{Y}}$ (odhad vybraného polynómu 4. stupňa) v príslušných zmesiach. Postupnosť zmesí bola už určená.



Graf 9: 3D bodový graf s odhadnutou strednou hodnotou polynómom štvrtého stupňa vybraného na základe C_p.

V grafe 10 je zobrazený kontúrový graf pre odhadnutú strednú hodnotu s desiatimi regresormi. Izočiary s hodnotou 12 až po 5.5 nápadne pripomínajú tvar paraboloidu. A keďže sa práve v tejto oblasti (budú uvažované aj hraničné body) nachádza najvyššia tvrdosť, ktorú je potrebné vyšetriť, je možné modelovať strednú hodnotu v tejto oblasti pomocou paraboloidu. O tomto bude pojednávať nasledujúci odstavec.



Graf 10: Kontúrový graf odhadnutej strednej hodnoty.

5.3.1. Modelovanie tvrdosti paraboloidom

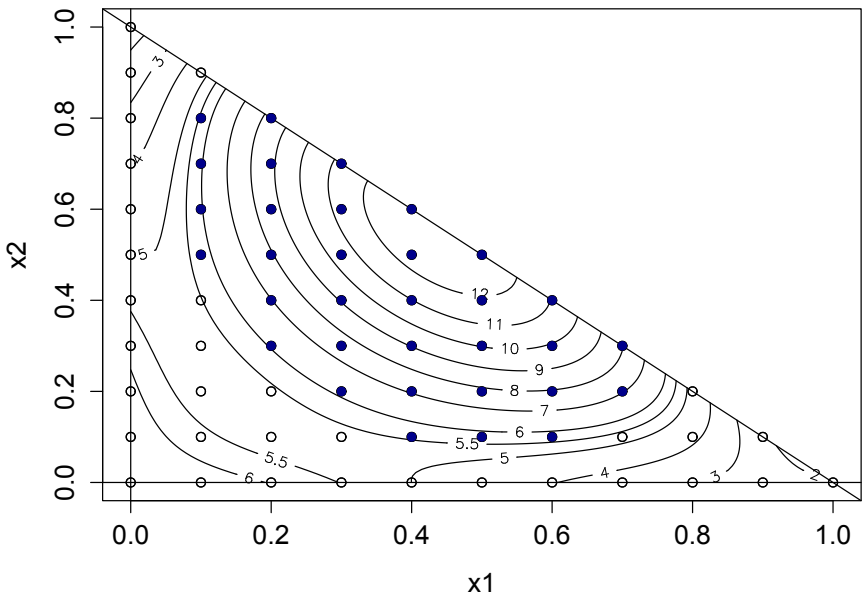
Nový model budú tvoriť hodnoty zmesí, ktoré sú označené v kontúrovom grafe 11 tmavomodrou farbou. Všetky sa nachádzajú v danej oblasti vymedzenou izočiarou (strednej hodnoty) s hodnotou 5.5. V tejto oblasti sa nachádza 33 rôznych zmesí a teda 66 pozorovaní z $\boldsymbol{Y}$. Pre tieto pozorovania bude uvažovaný model so strednou hodnotou

$$\mathsf{E}Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2.$$

Základné štatistiky pre takýto model sa nachádzajú vo výstupe 10.

Kroková regresia pre $\alpha = 0.05$, vybrala 4 regresory (absolútny člen, x_1, x_1^2 a $x_1 x_2$). C_p štatistika pre tento model vyšla 13.684728 a v grafe 12 je označená červenou farbou. Bude opäť lepšie zamerať sa kvôli strednej hodnote viac na širší model ale menej odchýlený od celkového modelu (so šiestimi regresormi). Totiž v iných prípadoch, pri vyhodnocovaní lineárneho modelu, je dôležitá interpretácia jednotlivých regresorov a ich význam v modeli. Následne sa vyhodnocuje aký majú jednotlivé regresory vplyv na celkový model a z toho sa vyhodnocujú závery. No v prípade, ktorý rieši táto práca, ide primárne o odhad strednej hodnoty a následnom najdení maximálnej tvrdosti (resp. príslušnej zmesy). Preto je dávaná prednosť metóde výberu pomocou C_p štatistiky, avšak s rozumným počtom regresorov. Táto hodnota je označená modrou farbou a nadobúda hodnoty $\mathsf{C}_p = 4.057632$.

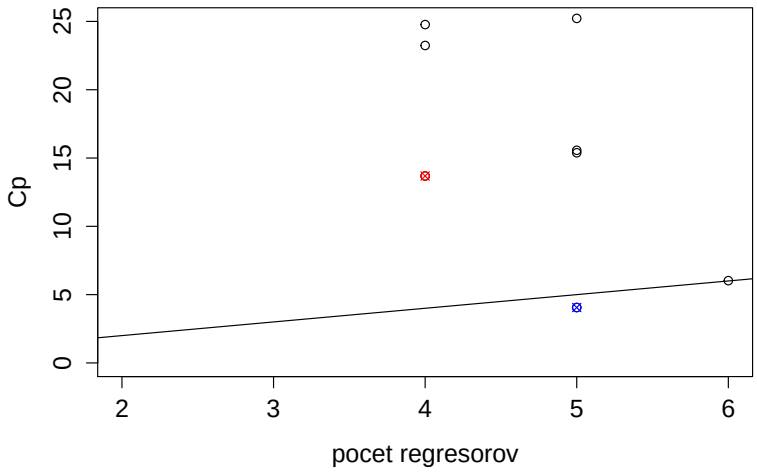
Je vidieť, že uvažovaný model má päť regresorov (nie je uvažovaný regresor x_1x_2). Pre tento model sú uvedené štatistiky vo výstupe 11.



Graf 11: Kontúrový graf s vybranými hodnotami.

Výstup z R 10: Základné štatistiky pre celý model

```
Call:
lm(formula = Y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x2^2))

Residuals:
     Min       1Q   Median       3Q      Max
-2.36715 -0.67358 -0.03526  0.65396  2.17594

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.797      3.529  -2.776  0.00732 **
x1            48.102     10.450   4.603 2.21e-05 ***
x2            31.250      9.200   3.397  0.00121 **
I(x1^2)      -45.508      8.049  -5.654 4.61e-07 ***
I(x1 * x2)    -2.243     12.049  -0.186  0.85296
I(x2^2)      -21.482      6.374  -3.370  0.00132 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.114 on 60 degrees of freedom
Multiple R-squared:  0.8346,Adjusted R-squared:  0.8208
F-statistic: 60.54 on 5 and 60 DF,  p-value: < 2.2e-16
```

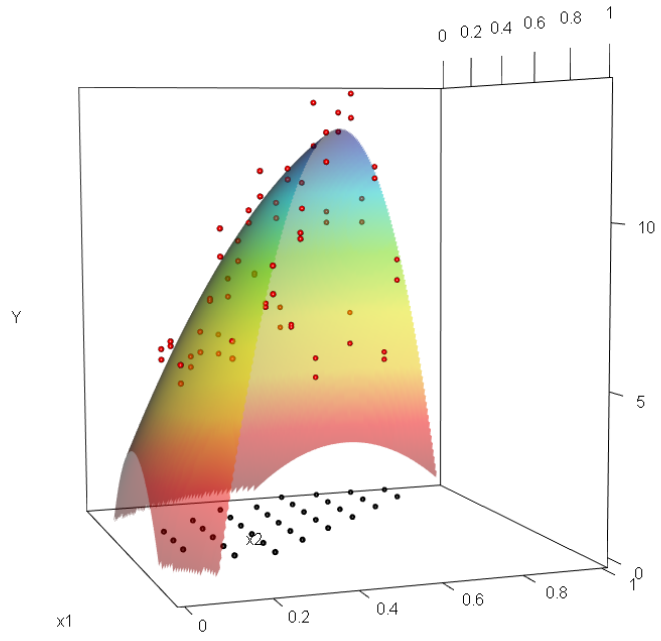


Graf 12: Modrou farbou je označený model, ktorý bol vybraný na základe C_p.

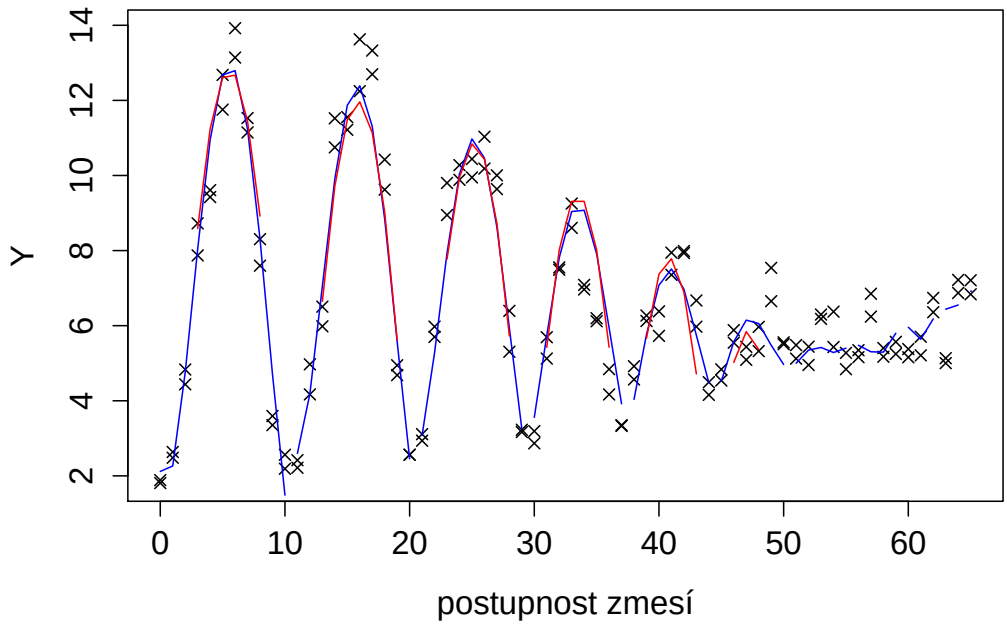
Výstup z R 11: Základné štatistiky pre vybraný model na základe C_p.

```
Call:
lm(formula = Y ~ x1 + x2 + I(x1^2) + I(x2^2))

Residuals:
     Min       1Q   Median       3Q      Max
-2.34660 -0.67076 -0.05109  0.65131  2.18474

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.168      1.015  -9.032 7.54e-13 ***
x1            46.292      3.792  12.209  < 2e-16 ***
x2            29.636      3.048   9.724 5.15e-14 ***
I(x1^2)      -44.281      4.583  -9.662 6.54e-14 ***
I(x2^2)      -20.482      3.403  -6.018 1.09e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.105 on 61 degrees of freedom
Multiple R-squared:  0.8345,Adjusted R-squared:  0.8236
F-statistic: 76.88 on 4 and 61 DF,  p-value: < 2.2e-16
```

Stredná hodnota pre tento model je zobrazená v grafe 13. V grafe 14 je vidieť, že odhad pomocou paraboloidu skutočne vyšiel dobre. Rozdiely v odhadoch sú v podstate minimálne, zrejme boli v pôvodnom modeli (4. stupňa) korigované vyšším stupňom polynómu. Keďže boli dôsledne prehodnocované zmesy (z odhadu strednej hodnoty polynómom 4. stupňa), ktoré primárne vytvárajú parabolický tvar, výrazne sa zamedzilo pákovému efektu. Je samozrejmé, že lokálny odhad strednej hodnoty (napríklad i na konvexnej množine) nemusí odpovedať odhadu urobenému na základe všetkych pozorovaní.



Graf 13: 3D bodový graf s odhadnutou strednou hodnotou.



Graf 14: Porovnanie funkcie modrej farby totožnej s funkciou modrej farby v grafe 8 (odhad polynómom 4. stupňa). Graf funkcie označený červenou farbou (počastiach spojitá a lineárna), nadobúda pre sledované zmesy stredných hodnôt odhadovaných paraboloidom.

5.4. Určenie zmesi s najväčšou tvrdosťou

Z predchádzajúceho odstavca 5.3.1 vyplýva, že na uvažovanej oblasti v grafe 11, je možné predpokladať model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + e,$$

kde $e \sim \mathsf{N}(0, \sigma^2)$. Na hranici, kde sa nachádza maximum (viď graf 13), platí $x_2 = 1 - x_1$ a preto platí

$$Y = \beta_0 + \beta_1 x_1 + \beta_2(1 - x_1) + \beta_3 x_1^2 + \beta_4(1 - x_1)^2 + e.$$

Úpravou sa prejde na tvar

$$Y = (\beta_0 + \beta_2 + \beta_4) + (\beta_1 - \beta_2 - 2\beta_4)x_1 + (\beta_3 + \beta_4)x_1^2 + e.$$

Pre hľadané maximum platí $\frac{\partial Y}{\partial x_1} = 0$, a preto

$$\frac{\partial Y}{\partial x_1} = (\beta_1 - \beta_2 - 2\beta_4) + 2(\beta_3 + \beta_4)x_1 = 0.$$

Z čoho vyplýva,

$$x_1 = \frac{-\beta_1 + \beta_2 + 2\beta_4}{2(\beta_3 + \beta_4)} = f_1(\beta_1, \beta_2, \beta_3, \beta_4).$$

Keďže z predchádzajúceho odstavca (z výstupu 11) vyplýva pre odhad metódou najmenších štvorcov $\boldsymbol{b} = (-9.168, 46.292, 29.636, -44.281, -20.482)'$, potom bodovým odhadom na hranici je

$$\hat{x}_1 = \frac{-b_1 + b_2 + 2b_4}{2(b_3 + b_4)} = \frac{-57.62}{-129.526} = 0.4448528.$$

A pre druhú zložku platí odhad

$$\hat{x}_2 = 1 - \hat{x}_1 = 1 - 0.4448528 = 0.5551472.$$

Preto bodovým odhadom pre najväčšiu tvrdosť je binárna zmes zložená

$$(\hat{x}_1, \hat{x}_2) = (0.4448528, 0.5551472),$$

kde x_1 vyjadruje pomer hydrogenfosforečnanu sodného zo zmesi a x_2 pomer difosforečnanu sodného.

Ďalej bude využitá delta metóda ktorá bola uvedená v úvode. V odstavci o normálnom lineárnom modeli úplnej hodnosti bolo vo vete 2.7 uvedené rozdelenie pre odhad regresných koeficientov

$$\boldsymbol{b} \sim \mathsf{N}_k(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Odhad $\boldsymbol{b}_n$ (z n pozorovaní) má asymptoticky k-rozmerné normálne rozdelenie $\mathsf{N}_k(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$(v tomto odstavci $k = 5$). Aby bolo možné dôjsť k podobnému záveru ako v 1.5 je potrebné vypočítať nasledujúce parciálne derivácie funkcie f_1

$$\phi_0 = \frac{\partial f_1}{\partial \beta_0} = 0, \qquad \phi_1 = \frac{\partial f_1}{\partial \beta_1} = \frac{-1}{2(\beta_3+\beta_4)},$$

$$\phi_2 = \frac{\partial f_1}{\partial \beta_2} = \frac{1}{2(\beta_3+\beta_4)}, \qquad \phi_3 = \frac{\partial f_1}{\partial \beta_3} = \frac{\beta_1-\beta_2-2\beta_4}{2(\beta_3+\beta_4)^2},$$

$$\phi_4 = \frac{\partial f_1}{\partial \beta_4} = \frac{4(\beta_3+\beta_4)-2(-\beta_1+\beta_2+2\beta_4)}{4(\beta_3+\beta_4)^2} = \frac{\beta_1-\beta_2+2\beta_3}{2(\beta_3+\beta_4)^2},$$

kde označenie $\boldsymbol{\phi} = (\phi_0, \phi_1, \phi_2, \phi_3, \phi_4)'$ korešponduje zo zavedením pri uvedení delta metódy.

Potom podľa 1.5 platí

$$\sqrt{n}\,(f_1(\boldsymbol{b}_n) - f_1(\boldsymbol{\beta})) \text{ konverguje v distribúcii k } \mathsf{N}(0, \sigma^2 \boldsymbol{\phi}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\phi}). \tag{5.7}$$

Vektor $\boldsymbol{\phi}$ je neznámy a je nahradený $\hat{\boldsymbol{\phi}} = \boldsymbol{\phi}(\boldsymbol{b}_n)$. Rozptyl σ^2 je taktiež neznámy a nahradený podľa delta metódy $s^2 = \sigma^2(\boldsymbol{b}_n)$.

Pre odhad rozptylu 5.7 platí $\mathsf{D}(\hat{x}_1)$

$$\mathsf{D}(\hat{x}_1) = \hat{\boldsymbol{\phi}}'\mathsf{Var}(\boldsymbol{b}_n)\hat{\boldsymbol{\phi}} = s^2\hat{\boldsymbol{\phi}}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\boldsymbol{\phi}},$$

kde reálny vektor

$$\hat{\boldsymbol{\phi}} = \left(0, \frac{-1}{2(b_3+b_4)}, \frac{1}{2(b_3+b_4)}, \frac{b_1-b_2-2b_4}{2(b_3+b_4)^2}, \frac{b_1-b_2+2b_3}{2(b_3+b_4)^2}\right)'.$$

Získané boli hodnoty

$$\hat{\boldsymbol{\phi}} = (0, 0.007720451, -0.007720451, 0.006868902, -0.008572)'$$

a

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 0.8432522 & -2.2224712 & -1.4219867 & 2.176086 & 1.025125 \\ -2.2224712 & 11.7659634 & -0.5005523 & -13.692503 & 1.776762 \\ -1.4219867 & -0.5005523 & 7.6026069 & 1.028647 & -8.099799 \\ 2.1760862 & -13.6925034 & 1.0286466 & 17.191063 & -1.985065 \\ 1.0251250 & 1.7767623 & -8.0997994 & -1.985065 & 9.480443 \end{pmatrix}.$$

Pre interval spoľahlivosti s bodovým odhadom $\hat{x}_1 = 0.4448528$ platí

$$\left(\hat{x}_1 - u_{1-\alpha/2}s\sqrt{\hat{\boldsymbol{\phi}}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\boldsymbol{\phi}}},\ \hat{x}_1 + u_{1-\alpha/2}s\sqrt{\hat{\boldsymbol{\phi}}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\boldsymbol{\phi}}}\right),$$

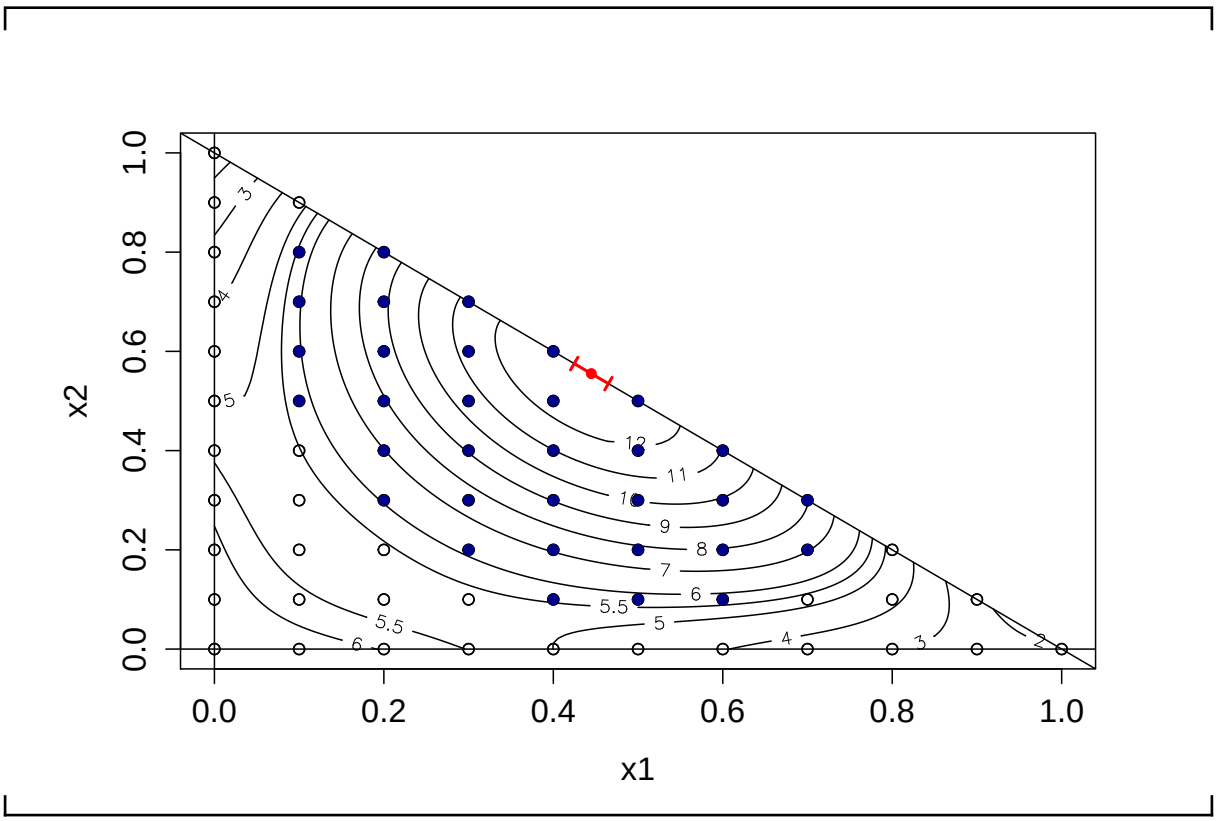
kde hodnota smerodajnej odchýlky $s = 1.105$ je určená z výstupu 11. Po dosadení hodnôt a pre $\alpha = 0.05$ výjde

$$\left(\hat{x}_1 - 1.96 \cdot 1.105\sqrt{8.701371 \cdot 10^{-5}},\ \hat{x}_1 + 1.96 \cdot 1.105\sqrt{8.701371 \cdot 10^{-5}}\right),$$

a nakoniec

$$(\hat{x}_1 - 0.02020283,\ \hat{x}_1 + 0.02020283).$$

Jedná sa teda o interval spoľahlivosti pre $\hat{x}_1$ v binárnej zmesi. V grafe 15 je zaznačený tento interval aj s bodovým odhadom.



Graf 15: Zobrazenie bodového aj intervalového odhadu do kotúrového grafu odhadnutej strednej hodnoty polynómom 4. stupňa (viď graf 11).

6. Záver

Ciele tejto diplomovej práce sa podarilo naplniť, ako po teoretickej tak i aplikačnej stránke.

V teoretickej časti práce bol v druhej kapitole precízne zavedený lineárny regresný model s úplnou stĺpcovou hodnosťou a následne v tretej kapitole model s neúplnou stĺpcovou hodnosťou. Ďalšia časť popisuje metódy pre výber regresorov a diagnostiku lineárneho regresného modelu. Celkovo sa podarilo urobiť ucelený výklad teórie s logickým usporiadaním a doplnený vhodnými obrázkami. Z nich niektoré, ako napr. obrázok 3**b**, alebo 5, neboli podľa vedomia autora publikované a môžu napomôcť pri pochopení danej látky.

V aplikačnej časti, ktorá je uvedená v piatej kapitole, sa podarilo pomocou lineárnej regresie úspešne vyhodnotiť experiment. To dokladujú aj výsledky, ktoré sú demonštrované graficky. Bola použitá polynomická regresia a výber regresorov bol založený na Krokovej regresii a Mallowsovej C_p štatistike. Následne bola odhadnutá stredná hodnota. Aby mohla byť odhadnutá zmes taviacich solí s najväčšou tvrdosťou, bola na vybranej oblasti stredná hodnota modelovaná paraboloidom. Využitím tohto odhadu bol nájdený bodový odhad a ďalej pomocou Delta metódy intervalový odhad.

Experiment bol sledovaný na ternárnych zmesiach tvorených hydrogenfosforečnanom sodným, difosforečnanom sodným a polyfosforečnanom sodným. Regresnou analýzou sa ukázalo, že najväčšia tvrdosť je dosiahnutá pri použití binárnej zmesi (8 týždňov zrenia, 30 dní skladovania). Bodovým odhadom je zmes tvorená: 44,485 % hydrogenfosforečnanu sodného a 55,515 % difosforečnanu sodného. Pre hydrogenfosforečnan sodný bol zistený 95% interval spoľahlivosti pre naj-väčšiu tvrdosť v binárnej zmesi a to interval (42.465, 46.505) v percentách.

Výsledky tejto práce budú využité pri výrobe tavených syrov.

Literatura

[1] AGRESTI, A. *Categorical Data Analysis*. 2nd edition. Hoboken: Wiley Press, 2002. ISBN 0-471-36093-7

[2] ANDĚL, J. *Matematická statistika*. 1. vyd. Praha: SNTL/ALFA, 1978.

[3] ANDĚL, J. *Základy matematické statistiky*. 3. vyd. Praha: MATFYZPRESS, 2011. ISBN 978-80-7378-162-0

[4] BROOK, R. J., ARNOLD, G. C. *Applied Regression Analysis and Experimental Designs*. 1st edition. New York: Marcel Dekker, 1985. ISBN 0-8247-7252-0

[5] BUŇKA, F., et al. The effect of ternary emulsifying salt composition and cheese maturity on the textural properties of processed cheese [online]. *International Dairy Journal*, March 2013, vol. 29, iss. 1, pages 1-7 [cit. 2013-12-04]. Dostupné na: <http://dx.doi.org/10.1016/j.idairyj.2012.09.006>

[6] JÖRNSTEN, R. *Linear Statistical Models* (lectures) [online]. Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, SE-412 96 Göteborg, Sweden. 2011. [cit. 2011-10-11]. Dostupné na:<http://www.math.chalmers.se/Stat/Grundutb/GU/MSG500/A11/>

[7] KAPOOR, R., METZGER, L. E. Process Cheese: Scientific and Technological Aspects — A Review [online]. *Comprehensive Reviews in Food Science and Food Safety*, March 2008, vol. 7, iss. 2, pages 194–214 [cit. 2012-12-10]. Dostupné na: <http://onlinelibrary.wiley.com/doi/10.1111/j.1541-4337.2008.00040.x/pdf>

[8] MALLOWS, C. L. Some Comments on Cp [online]. *Technometrics*, November 1973, vol. 15, No. 4, pages 661-675 [cit. 2013-27-04]. Dostupné na: <http://www.jstor.org/discover/10.2307/1267380?uid=3737856&uid=2&uid=4&sid=21102278590171>

[9] RAO, R. C. *Lineární metody statistické indukce a jejich aplikace*. 1. vyd. Praha: Academia nakladatelství ČSAV, 1978.

[10] STAUDTE, R. G., SHEATHER, S. J. *Robust estimation and testing*. 1st edition. Wiley Press,INC. 1990. ISBN 0-471-85547-2

[11] WEISEROVÁ, E., et al. The effect of combinations of sodium phosphates in binary mixtures on selected texture parameters of processed cheese spreads [online]. *International Dairy Journal*, December 2011, vol. 21, iss. 12, pages 979-986 [cit. 2012-12-07]. Dostupné na: <http://dx.doi.org/10.1016/j.idairyj.2011.06.006>

[12] ZVÁRA, K. *Regresní analýza*. 1. vyd. Praha: Academia nakladatelství ČSAV, 1989. ISBN 80-200-0125-5

[13] ZVÁRA, K. *Regrese*. 1. vyd. Praha: MATFYZPRESS, 2008. ISBN 978-80-7378-041-8

Príloha

Ako príloha bude uvedený iba zdrojový kód pre Krokovú regresiu, pretože táto metóda bola v odstavci 4.1.1 podrobne popísaná. K analýze experimentu teda nebola použitá v jazyku R na to určená funkcia step() (používajúca primárne kritérium AIC), ale tu uvedená verzia založená na F-teste.

V zdrojovom kóde 1 je uvedený krok 1, ktorý na základe t-testu vyberá prvý regresor. Tu sú súbežne počítané hodnoty pre všetky modely s jediným regresorom. Ako kritérium pre najviac významný regresor je uvažovaná maximálna hodnota testovacieho kritéria (v absolútnej hodnote).

Zdrojový kód 1: Kroková regresi (krok 1)

```
#1.krok
Y<-data[,13]                  #pozorovania Y
X<-cbind(matrix(1,length(x1),1),x1,x2,x1^2,x1*x2,x2^2)  #regresna matica
XY<-t(X)%*%Y
XX<-1/(diag(t(X)%*%X))
b<-XX*XY                      #vektor jednotliych odhadov beta
rezid<-X%*%diag(as.vector(b))-Y%*%t(as.vector(b))  #matica rezidui modelov
RSS<-diag(t(rezid)%*%rezid)  #RSS hodnoty pre jednotlive modely
odh_rozp<-RSS/131             #odhad rozptylu pre jednotlive modely
test<-b/sqrt(XX*odh_rozp)     #jednotlive statistiky pre t-test
Nvyb <- which.max(abs(test)) #vybraty regresor
```

V zdrojovom kóde 2 je uvedený krok 2 až 5 ako boli popísané v tejto práci. Jediným nepodstatným rozdielom je nepoužitie množiny $\mathcal{N}_{nevyb}$, ktoré bolo nahradené v zmysle komplementu množiny $\mathcal{N}_{vyb}$. Označenie je zachované podľa popísanej metódy. Všetky výpočty štatistík sú názorne zapísané vzťahmi tak, ako boli popísané v teoretickej časti.

Zdrojový kód 2: Kroková regresi (krok 2-5)

```
#2.krok
krok5<-TRUE
repeat{
  X1<-X[,Nvyb]               #regresna matica urcena Nvyb
  H1<-X1%*%solve(t(X1)%*%X1)%*%t(X1)
  M1<-diag(1,length(Y))-H1
  RSSmod<-t(Y)%*%M1%*%Y      #RSS vybranych regresorov
  RSSroz<-RSSmod
  for(j in 1:dim(X)[2]) {    #vypocet RSS pre vsetky nadmodely(o 1 regresor)
    if (!(j%in%Nvyb)) {
      X2<-X[,c(Nvyb,j)]
      H2<-X2%*%solve(t(X2)%*%X2)%*%t(X2)
      M2<-diag(1,length(Y))-H2
      RSS2<-t(Y)%*%M2%*%Y
      if (RSSroz>RSS2){
        RSSroz<-RSS2
        index<-j }
  }
  }
#3.krok
  F<-(RSSmod-RSSroz)/(RSSmod/(length(Y)-length(Nvyb)))
  #nasleduje F-test o nerozsirení modelu
  if(F>qf(0.95,1,length(Y)-length(Nvyb))){
    Nvyb<-c(Nvyb,index)
    RSSmod<-RSSroz
  }
  if((F<=qf(0.95,1,length(Y)-length(Nvyb)))&&(!krok5)){break}
#4.krok
  RSSred<-Inf
  for(j in 1:dim(X)[2]) {  #vypocet RSS pre vsetky podmodely(o 1 regresor)
    if (j%in%Nvyb){
      X3<-X[,setdiff(Nvyb,j)]
      H3<-X3%*%solve(t(X3)%*%X3)%*%t(X3)
      M3<-diag(1,length(Y))-H3
      RSS3<-t(Y)%*%M3%*%Y
      if (RSSred>RSS3){
        RSSred<-RSS3
        index<-j }
    }
  }
#5.krok
  F<-(RSSred-RSSmod)/(RSSmod/(length(Y)-length(Nvyb)))
  if(F<qf(0.95,1,length(Y)-length(Nvyb))){
    Nvyb<-setdiff(Nvyb,index)
    krok5<-TRUE }
  if(F>=qf(0.95,1,length(Y)-length(Nvyb))){krok5<-FALSE}
}
```