Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Viktor Szabados

Některé sekvenční postupy pro jednoduchou regresi

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	prof. RNDr. Marie Hušková, DrSc.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2014

Chcel by som sa v prvom rade poďakovať svojej vedúcej bakalárky prof. RNDr. Marie Huškovej, DrSc. za jej profesionálny prístup pri vedení mojej bakalárskej práci. Som rád, že som mohol byť práve pod jej vedením a ďakujem jej za ochotu a čas nájdený na konzultačné hodiny. Oceňujem hlavne rýchle odpisovanie na e-maily a taktiež si cením pripomienky k bakalárskej práci, vďaka ktorým sa zdvihla celková kvalita textu.

Ako ďalším by som sa rád poďakoval rodine Ješkových a Lukášovi Zavřelovi za výpomoc pri preklade abstraktu do anglického a českého jazyka.

Nakoniec ďakujem svojej rodine a kamarátom, ktorí mi pomohli s gramatikou v niektorých slovných spojeniach.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V .... dne .... Podpis autora

Název práce: Některé sekvenční postupy pro jednoduchou regresi

Autor: Viktor Szabados

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: prof. RNDr. Marie Hušková, DrSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Se sekvenčními metodami se ve statistice setkáváme v situacích, kde se snažíme udělat jen potřebný počet pozorování k vyvození důvěryhodného závěru. V této bakalářské práci seznámíme čtenáře se základními sekvenčními postupy, které se využívají v lineárním regresním modelu. V první kapitole zavedeme lineární regresní model, se kterým budeme pracovat. V druhé kapitole již budou zmiňovány sekvenční metody, které si navzájem mezi sebou porovnáme a určíme výhody a nevýhody jednotlivých sekvenčních postupů. V třetí a čtvrté kapitole budeme konstruovat intervalové odhady pro regresní parametry. Navíc budeme ještě ve čtvrté kapitole konstruovat i testy pro regresní parametry.

Klíčová slova: sekvenční postupy, jednoduchý regresní model, zastavovací pravidlo, testy, odhady

Title: Some sequential procedures in simple regression models

Author: Viktor Szabados

Department: Department of Probability and Mathematical Statistics

Supervisor: prof. RNDr. Marie Hušková, DrSc., Department of Probability and Mathematical Statistics

Abstract: Sequential methods are used in statistics, where only a limited number of observations has to be made to obtain a reliable result. In this thesis we present basic sequential procedures that are used in the linear regression model. In the first chapter we introduce the linear regression model. In the second chapter we present different sequential methods. We compare these methods with each other and determine the advantages and disadvantages of individual sequential procedures. In the third and fourth chapter we construct interval estimates for regression parameters. Additionally, in the fourth chapter we construct tests for regression parameters.

Keywords: sequential procedures, simple linear regession, stopping rules, tests, estimators

Názov práce: Niektoré sekvenčné postupy pre jednoduchú regresiu

Autor: Viktor Szabados

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: prof. RNDr. Marie Hušková, DrSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: So sekvenčnými metódami sa v štatistike stretávame v situáciách, kde sa snažíme spraviť len potrebný počet pozorovaní na vyvodenie dôveryhodného záveru. V tejto bakalárskej práci oboznámime čitateľa so základnými sekvenčnými postupmi, ktoré sú využívané v lineárnom regresnom modeli. V prvej kapitole zavedieme lineárny regresný model, s ktorým budeme pracovať. V druhej kapitole už budú spomínané sekvenčné metódy, ktoré si navzájom medzi sebou porovnáme a určíme výhody a nevýhody jednotlivých sekvenčných postupov. V tretej a štvrtej kapitole budeme konštruovať intervalové odhady pre regresné parametre. V štvrtej kapitole budeme navyše konštruovať aj testy pre regresné parametre.

Kľúčové slová: sekvenčné procedúry, jednoduchý regresný model, zastavovacie pravidlo, testy, odhady

Obsah

Úvod

Ako už názov bakalárskej práce hovorí, budeme sa zaoberať sekvenčnými postupmi v lineárnej regresii. Preto by som rád na začiatok pripomenul, čo to vlastne je sekvenčná analýza. Je to metóda, pomocou ktorej vieme vyriešiť štatistické úlohy, u ktorých nemáme dopredu stanovený pevný počet pozorovaní. Týmto typom úlohy sa zaoberajú sekvenčné metódy. Tie na základe zvoleného počtu pozorovaní zistia, či už máme dostatočne veľa pozorovaní alebo ešte nejaké musíme pridať, aby sme z nich mohli vyvodiť dôveryhodný záver.

Sekvenčné metódy sa využívajú napríklad v ekonómii, v inžinierstve alebo aj v biomedicíne. Používajú sa zväčša k tomu, aby sme spravili iba potrebný počet experimentov, a tým pádom ušetríme na nákladoch vložených do prípravy experimentov. Bohužiaľ na druhej strane nepoznáme dopredu rozsah výberu. Preto testovanie po každom experimente, či ešte máme urobiť nejaký ďalší experiment alebo nie, môže byť časovo náročným procesom. Najčastejšie ich teda využívame tam, kde sa robia experimenty jeden za druhým.

Táto bakalárska práca nadväzuje na prácu (Rusá, 2013) o dvojstupňových štatistických metódach. Budem sa preto snažiť zachovávať značenie a formu práce mojej predchodkyňe, aby sa prípadní budúci pokračovatelia v danej problematike vedeli rýchlejšie zorientovať v našich prácach.

Síce názov bakalárskej práce hovorí o skúmaní sekvenčných metód v jednoduchej lineárnej regresii, ale podarilo sa mi s pomocou mojej vedúcej rozšíriť skúmanie na všeobecný lineárny regresný model pri dodatočných predpokladov.

Práca je rozdelená do štyroch kapitol, ktoré sa ďalej rozpadajú do niekoľkých sekcií prípadne podsekcií.

V prvej kapitole si zadefinujeme a pripomenie základné vzťahy a tvrdenia, ktoré platia v lineárnom regresnom modeli. Následne si vyskúšame zostrojiť konfidenčnú množinu s danou spoľahlivosťou a objemom pre neznáme regresné parametre.

V druhej kapitole popíšeme základné sekvenčné metódy, ktoré sa zvyknú využívať v lineárnej regresii. Zväčša pôjde o dvojstupňové metódy, ktoré tvorili základ v následnom rozvíjaní nových sekvenčných metód. Na záver tejto kapitoly aplikujeme teoretické poznatky na príkladoch. Predvedieme zopár simulácii v matematickom softvéri R a dosiahnuté výsledky zo simulácii si rozanalyzujeme.

V tretej kapitole si skonštruujeme intervalové odhady s danou spoľahlivosťou a dĺžkou pre jednotlivé regresné parametre.

V štvrtej a zároveň poslednej kapitole budeme opäť konštruovať konfidenčné množiny, intervalové odhady. . . , ale pre obmenený lineárny model. Zostrojíme aj testové štatistiky na testovanie regresných parametrov.

Tento text je určený pre študentov, ktorý ukončili základný kurz Matematickej štatistiky I a II. Očakáva sa, že čitateľ je zbehlý v základných štatistických úvahách a je oboznámený s pojmami ako konfidenčná množina, F rozdelenie, idempotentná matica ....

Prajem príjemné čítanie

Viktor Szabados

Kapitola 1

Konfidenčné množiny v lineárnej regresii

Matematická štatistika, ako ju dnes poznáme, má v súčasnom živote nezastupiteľné miesto. Stala sa z nej vedná disciplína, ktorá nachádza široké uplatnenie v podnikaní, prírodných vedách a mnoho ďalších.

Medzi základné typy úloh matematickej štatistiky patrí dnes konštruovanie štatistických odhadov a testovanie hypotéz. Pod štatistické odhady patria bodové odhady, intervalové odhady, konfidenčné množiny, oblasti spoľahlivosti i konfidenčné oblasti s predpísaným koeficientom spoľahlivosti.

V tejto kapitole sa budeme zaoberať konštruovaním konfidenčnej množiny pre neznáme regresné parametre v lineárnom modeli s vopred daným koeficientom spoľahlivosti. Skôr než sa do toho pustíme, pripomeňme si základné pojmy a tvrdenia, ktoré platia v lineárnom modeli.

1.1 Lineárny regresný model

Majme náhodný vektor $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^\top$ a maticu daných čísel $\mathbf{X}_n = (x_{ij})$ typu $n \times p$ a hodnosťou p. Nech platí $\mathbf{Y}_n = \mathbf{X}_n\boldsymbol{\beta} + \mathbf{e}_n$, kde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ je vektor neznámych parametrov a $\mathbf{e}_n = (e_1, \ldots, e_n)^\top$ je náhodný vektor, ktorý splňuje:

$$\mathsf{E}\mathbf{e}_n = \mathbf{0}_n \qquad \operatorname{var}\mathbf{e}_n = \sigma^2 \boldsymbol{I}_n,$$

potom hovoríme o **lineárnom regresnom modeli**. Parameter σ^2 nepoznáme a o vektore $\mathbf{e}_n$ zväčša hovoríme ako o vektore chýb. Podmienky, ktoré naň kladieme zavádzame z dôvodu, aby pri skúmaní vektora $\mathbf{Y}_n$ nedochádzalo k systematickým chybám a jednotlivé zložky boli merané s rovnakou presnosťou. Vektor $\boldsymbol{\beta}$ sa väčšinou odhaduje pomocou metódy najmenších štvorcov. Tento odhad si označme $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\top$.

Odteraz budeme predpokladať o vektore chýb, že má $N_n(0, \sigma^2\boldsymbol{I}_n)$ rozdelenie, kde $\boldsymbol{I}_n$ značí n-rozmernú jednotkovú maticu. Z toho potom plynie, že náhodný vektor $\mathbf{Y}_n$ má normálne rozdelenie $N_n(\mathbf{X}_n\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$. Skrátene to budeme zapisovať $\mathbf{Y}_n \sim N_n(\mathbf{X}_n\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$. Predpoklad normálneho rozdelenia u vektora chýb patrí k štandardným predpokladom, ktorý naň kladieme. Potom má odhadovaný vektor $\widehat{\boldsymbol{\beta}}$, metódou najmenších štvorcov, dobré štatistické vlastnosti.

Označme si reziduálny súčet štvorcov ako $R_n = (\mathbf{Y}_n - \mathbf{X}_n\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y}_n - \mathbf{X}_n\widehat{\boldsymbol{\beta}})$ a reziduálny rozptyl $S_n^2 = R_n/(n-p)$. V nasledujúcej vete by som rád pripomenul zopár základných vlastností, ktoré platia pre vektor $\widehat{\boldsymbol{\beta}}$.

Veta 1.1. *Nech v lineárnom regresnom modeli má vektor chýb $N_n(0, \sigma^2 \boldsymbol{I}_n)$ rozdelenie a nech $\mathbf{X}_n^\top \mathbf{X}_n$ je regulárna matica, potom platí:*

a) $\widehat{\boldsymbol{\beta}} = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}\mathbf{X}_n^\top \mathbf{Y}_n$

b) $\widehat{\boldsymbol{\beta}} \sim N_n[\boldsymbol{\beta}, \sigma^2(\mathbf{X}_n^\top \mathbf{X}_n)^{-1}]$

c) R_n/σ^2 *má* χ^2_{n-p} *rozdelenie*

d) $\widehat{\boldsymbol{\beta}}$ *a* R_n *sú nezávislé*

e) $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top[\sigma^{-2}(\mathbf{X}_n^\top \mathbf{X}_n)](\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ *má* χ^2_p *rozdelenie,*

kde symbolom χ^2_p *značíme* χ^2 *rozdelenie s* p *stupňami voľnosti.*

Dôkaz. Tvrdenia $a) - d)$ sú dokázané v knihe (Anděl, 2007, str. 112-113), vety 9.1 a 9.4. Dôkaz časti $e)$ je uvedený v spomínanej knihe na str. 58, veta 4.13. □

Z vety 1.1 $c), d)$ vyplýva, že S_n^2 je nestranným odhadom parametru σ^2 a je nezávislý na $\widehat{\boldsymbol{\beta}}$.

1.2 Konfidenčná množina pre regresné parametre s danou veľkosťou

V tejto sekcii sa pokúsime nájsť konfidenčnú množinu pre regresné parametre. Majme zadané čísla $0 < d < \infty$ a $0 < \alpha < 1$. Od konfidenčnej množiny očakávame, aby odhadnuté regresné parametre v nej ležali s pravdepodobnosťou aspoň $1 - \alpha$, a aby sme voľbou d vedeli zabezpečiť ľubovoľne malý objem tejto množiny.

Najprv skonštruujeme konfidenčnú množinu pre **známy rozptyl $\boldsymbol{\sigma}^2$** v lineárnom modeli

$$\begin{aligned} &\mathbf{Y}_n = \mathbf{X}_n\boldsymbol{\beta} + \mathbf{e}_n, \qquad \mathbf{e}_n \sim N_n(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n) \text{ a} \\ &\text{hodnosť matice } \mathbf{X}_n \text{ je plná t.j. } h(\mathbf{X}_n) = p, n \geq p+1, \end{aligned} \tag{1.1}$$

kde dolný index $_n$ pri vektoroch a maticiach vysvetľuje s akým rozsahom výberu pracujeme. Zaveďme značenie $\widehat{\boldsymbol{\beta}}_n = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}\mathbf{X}_n^\top \mathbf{Y}_n$. Elipsoidná konfidenčná množina pre neznámy vektor parametrov $\boldsymbol{\beta}$ pri danom d je:

$$\mathcal{R}_n \equiv \mathcal{R}_n(d) = \{\boldsymbol{\beta}_{p\times 1} \in \mathbb{R}^p : (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top[n^{-1}(\mathbf{X}_n^\top \mathbf{X}_n)](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \leq d^2\}. \tag{1.2}$$

Výraz $n^{-1}(\mathbf{X}_n^\top \mathbf{X}_n)$ možno vyzerá trochu umelo, ale vo väčšine prípadov sa predpokladá, že $n^{-1}(\mathbf{X}_n^\top \mathbf{X}_n)$ konverguje pre $n \to \infty$ k pozitívne definitnej matici $\mathbf{A}_{p\times p}$. Poznamenajme, že voľbou $0 < d < \infty$ dokážeme zaistiť ľubovoľne malý objem $\mathcal{R}_n$. Preto hovoríme, že elipsoidná konfidenčná množina má pevnú veľkosť.

Konfidenčný koeficient pre množinu $\mathcal{R}_n$ je:

$$\begin{aligned}\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{\boldsymbol{\beta}\in\mathcal{R}_n\Big\} &= \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})^\top(\mathbf{X}_n^\top\mathbf{X}_n)(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\le nd^2\Big\} = \\ &= F\left(\frac{nd^2}{\sigma^2}\right) \qquad F(x)=\mathsf{P}(\chi_p^2\le x), x>0.\end{aligned} \tag{1.3}$$

Nech a je kritická hodnota rozdelenia χ_p^2 na hladine α, teda splňuje $F(a)=1-\alpha$. Chceme, aby koeficient spoľahlivosti bol aspoň $1-\alpha$. Preto požadujeme, aby $F(nd^2/\sigma^2)\ge F(a)$. Táto nerovnosť platí, keď $nd^2/\sigma^2\ge a$, a teda nech

$$\text{n je najmenšie prirodzené číslo} \ge \frac{a\sigma^2}{d^2} =: C. \tag{1.4}$$

Obdoba tejto podmienky sa vyskytuje aj pri testovaní hypotéz. Práve tento problém pri určení hodnoty n bol jedným z dôvodov, prečo začali vznikať sekvenčné metódy.

Ak by sme poznali σ^2 nemali by sme problém. Zo vzťahu (1.4) by sme dorátali hodnotu C (optimálny rozsah), a tým pádom by sme vedeli určiť rozsah výberu n na zostrojenie konfidenčnej množiny s koeficientom spoľahlivosti $1-\alpha$.

Bohužiaľ v lineárnom regresnom modeli nepoznáme rozptyl σ^2 a ani neexistuje konfidenčná oblasť s koeficientom spoľahlivosti $1-\alpha$, ktorú by sme vedeli zostrojiť z pevného rozsahu výberu pri danom $0<d<\infty$ a $0<\alpha<1$. Dôkaz neexistencie tejto konfidenčnej oblasti s koeficientom spoľahlivosti $1-\alpha$ sa nachádza v (Ghosh a Sen, 1991, str. 23). Ako si neskôr ukážeme, práve sekvenčné metódy poskytujú riešenie.

Myšlienka je jednoduchá. Pokúsime sa vo vzťahu $n\ge a\sigma^2/d^2$ nahradiť σ^2 nestranným odhadom S_m^2, ktorý dopočítame z prvých m pozorovaní. Tento odhad nám dopomôže k určeniu veľkosti rozsahu výberu $N\equiv N(d)$. O veličine N sa zvykne hovoriť aj ako o zastavovacej veličine (angl. stopping random variable) a jej hodnota závisí na voľbe d, dopočítaným S_m^2 a tiež na zvolenej spoľahlivosti $1-\alpha$.

Analogicky ako v prípade (1.2) navrhnime konfidenčnú množinu pre $\boldsymbol{\beta}$ so spoľahlivosťou $1-\alpha$, ale už pri výbere o veľkosti N.

$$\mathcal{R}_N\equiv\mathcal{R}_N(d)=\{\boldsymbol{\beta}\in\mathbb{R}^p : N^{-1}(\widehat{\boldsymbol{\beta}}_N-\boldsymbol{\beta})^\top(\mathbf{X}_N^\top\mathbf{X}_N)(\widehat{\boldsymbol{\beta}}_N-\boldsymbol{\beta})\le d^2\} \tag{1.5}$$

Túto elipsoidnú množinu ovplyvňuje náhodná veličina N, čo nám môže skomplikovať našu situáciu. Preto budeme predpokladať, že matica $\mathbf{X}_N$ je dopredu daná. Taktiež si musíme dať pozor pri rátaní konfidenčného koeficientu (1.3). Pomôže nám k tomu veľmi dôležité pozorovanie, ktoré si aj neskôr dokážeme.

$$\text{Pre každé pevné } n\ge m \text{ sú } \widehat{\boldsymbol{\beta}}_n \text{ a } (S_m^2, S_{m+1}^2,\dots,S_n^2) \text{ nezávislé,} \tag{1.6}$$

kde $m\ge p+1$ a veličiny $\widehat{\boldsymbol{\beta}}_n$ a $(S_m^2,\dots,S_n^2)$ sú dopočítané z modela (1.1). Uvedomme si, že hodnota náhodnej veličiny $I[N=n]$ závisí iba na $(S_m^2,S_{m+1}^2,\dots,S_n^2)$, kde symbolom $I[A]$ označujeme indikátor množiny A. Za platnosti tvrdenia (1.6) potom platí, že aj náhodné veličiny $I[N=n]$ a $\widehat{\boldsymbol{\beta}}_n$ sú nezávislé. Pokúsme sa teraz dopočítať konfidenčný koeficient pre množinu $\mathcal{R}_N$ z (1.5) pri známom σ^2.

$$\begin{aligned}&\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{\boldsymbol{\beta}\in\mathcal{R}_N(d)\Big\} = \\ &= \sum_{n=m}^{\infty}\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{N^{-1}(\widehat{\boldsymbol{\beta}}_N-\boldsymbol{\beta})^\top(\mathbf{X}_N^\top\mathbf{X}_N)(\widehat{\boldsymbol{\beta}}_N-\boldsymbol{\beta})\le d^2|N=n\Big\}\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{N=n\Big\}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=m}^{\infty} \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{n^{-1}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top(\mathbf{X}_n^\top\mathbf{X}_n)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \le d^2 | N = n\Big\}\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{N = n\Big\} \\
&\overset{\spadesuit}{=} \sum_{n=m}^{\infty} \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{n^{-1}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top(\mathbf{X}_n^\top\mathbf{X}_n)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \le d^2\Big\}\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{N = n\Big\} \\
&= \sum_{n=m}^{\infty} \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top\frac{(\mathbf{X}_n^\top\mathbf{X}_n)}{\sigma^2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \le \frac{nd^2}{\sigma^2}\Big\}\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{N = n\Big\} \qquad (1.7)\\
&= \sum_{n=m}^{\infty} F\left(\frac{nd^2}{\sigma^2}\right)\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{N = n\Big\} = \mathsf{E}_{\boldsymbol{\beta},\sigma}\left[F\left(\frac{Nd^2}{\sigma^2}\right)\right],
\end{aligned}
$$

kde $F(x) = \mathsf{P}(\chi_p^2 \le x), x > 0$.

V rovnosti $\spadesuit$ sme využili nezávislosť náhodných veličín $\widehat{\boldsymbol{\beta}}_n$ a $I[N = n]$ pre každé $n = m, m+1, \ldots$. Poďme sa na konci tejto sekcie pozrieť na sľubovaný dôkaz kľúčového pozorovania.

Nasledujúci dôkaz je prevzatý z knihy (Mukhopadhyay a Silva, 2009, str. 331).

Veta 1.2. *Pre každé pevné $n \ge m$ sú $\widehat{\boldsymbol{\beta}}_n$ a $(S_m^2, S_{m+1}^2, \ldots, S_n^2)$ nezávislé, kde $m \ge p+1$ a veličiny $\widehat{\boldsymbol{\beta}}_n$ a $(S_m^2, \ldots, S_n^2)$ sú dopočítané z modela* (1.1)

Dôkaz. Vyjadríme si reziduálny súčet štvorcov pre nejaké $k, m \le k \le n$.

$$
\begin{aligned}
R_k \equiv (\mathbf{Y}_k - \mathbf{X}_k\widehat{\boldsymbol{\beta}}_k)^\top(\mathbf{Y}_k - \mathbf{X}_k\widehat{\boldsymbol{\beta}}_k) &= \mathbf{Y}_k^\top\mathbf{Y}_k - \mathbf{Y}_k^\top\mathbf{X}_k\widehat{\boldsymbol{\beta}}_k \\
&= \mathbf{Y}_k^\top(\mathbf{I}_{k\times k} - \mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{X}_k^\top)\mathbf{Y}_k.
\end{aligned}
$$

Poznamenajme, že $\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{X}_k^\top$ je projekčná matica. Z toho potom plynie, že $\mathbf{I}_{k\times k} - \mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{X}_k^\top$ je symetrická idempotentná matica s hodnosťou $k-p$. Z teórie matíc vieme, že táto matica má vlastné číslo 0 rádu p a vlastné číslo 1 rádu $k-p$. Nech $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{k-p}$ sú ortonormálne vlastné vektory prislušné k vlastnému číslu 1. Potom môžme upraviť vzťah pre R_k nasledovne

$$
R_k = \mathbf{Y}_k^\top\left(\sum_{i=1}^{k-p}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right)\mathbf{Y}_k = \sum_{i=1}^{k-p}(\boldsymbol{\xi}_i^\top\mathbf{Y}_k)^2.
$$

Označme $\boldsymbol{\rho}_i^\top = (\boldsymbol{\xi}_i^\top : \mathbf{0}^\top)_{1\times n}$ vektor, ktorý vznikne rozšírením vektora $\boldsymbol{\xi}_i^\top$ o nulový vektor. Teda platí $R_k = \sum_{i=1}^{k-p}(\boldsymbol{\rho}_i^\top\mathbf{Y}_n)^2$. Z vety 1.1 b) platí $\widehat{\boldsymbol{\beta}}_n = B\mathbf{Y}_n$, kde $B = (\mathbf{X}_n^\top\mathbf{X}_n)^{-1}\mathbf{X}_n^\top$. Pokúsime sa ukázať, že $B\mathbf{Y}_n$ je nezávislé na $\boldsymbol{\rho}_i^\top\mathbf{Y}_n$ pre $i = 1, \ldots, k-p$. Stačí nám ukázať, že $B\boldsymbol{\rho}_i = \mathbf{0}$, $i = 1, \ldots, k-p$.

Vektor $\boldsymbol{\xi}_i$ patrí do stĺpcového priestora $\mathbf{I}_{k\times k} - \mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{X}_k^\top$. Tým pádom $\boldsymbol{\xi}_i$ patrí do ortogonálneho doplnku stĺpcového priestora $\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{X}_k^\top$. Kedže $\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}$ generuje inverz $\mathbf{X}_k$, potom stĺpcový priestor $\mathbf{X}_k$ je súhlasný so stĺpcovým priestorom $\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{X}_k^\top$.

Z toho dostávame, že $\mathbf{X}_k^\top\boldsymbol{\xi}_i = \mathbf{0}$ pre všetky $i = 1, \ldots, k-p$, teda aj $\mathbf{X}_n^\top\boldsymbol{\rho}_i = \mathbf{0}$ pre všetky $i = 1, \ldots, k-p$. Z poslednej rovnosti už ľahko nahliadneme, že aj $B\boldsymbol{\rho}_i = \mathbf{0}$, $i = 1, \ldots, k-p$. Ukázali sme, že $\widehat{\boldsymbol{\beta}}_n$ a $(R_m, R_{m+1}, \ldots, R_n)$ sú nezávislé, čo zrejme už implikuje vetu. $\square$

Kapitola 2

Sekvenčné metódy

V celej tejto kapitole budeme pracovať s lineárnym modelom definovaným v (1.1). Ukážeme si zopár sekvenčných metód, pomocou ktorých dokážeme zostrojiť konfidenčnú množinu pre parameter $\boldsymbol{\beta}$ s koeficientom spoľahlivosti aspoň $1-\alpha$. Jednotlivé sekvenčné metódy sa budú od seba líšiť voľbou zastavovacej veličiny N a ich účinnosťou.

Nasledujúce definície sú prevzaté z bakalárskej práce (Rusá, 2013, str. 9-10).

Definícia 2.1. *Sekvenčná metóda je (exaktne) konzistentná, ak pre každé pevné $d > 0$ platí*

$$\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{\boldsymbol{\beta} \in \mathcal{R}_N(d)\Big\} \geq 1-\alpha,$$

kde $\mathcal{R}_N(d)$ je konfidenčná oblasť pre $\boldsymbol{\beta}$ odvodená touto metodou.

Definícia 2.2. *Sekvenčná metóda je asymptoticky účinná (1. rádu), ak*

$$\lim_{d\to 0_+} \mathsf{E}_{\boldsymbol{\beta},\sigma}\left[\frac{N}{C}\right] = 1,$$

kde C je definované v (1.4) a N je rozsah výberu určený touto metódou.

Definícia 2.3. *Sekvenčná metóda je asymptoticky účinná (2. rádu), ak*

$$\lim_{d\to 0_+} \mathsf{E}_{\boldsymbol{\beta},\sigma}(N-C) \textit{ je konečné číslo},$$

kde C je definované v (1.4) a N je rozsah výberu určený touto metódou.

Definícia 2.4. *Sekvenčná metóda je asymptoticky konzistentná, ak pre každé $\boldsymbol{\beta},\sigma$ platí*

$$\lim_{d\to 0_+} \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{\boldsymbol{\beta} \in \mathcal{R}_N(d)\Big\} = 1-\alpha,$$

kde $\mathcal{R}_N(d)$ je konfidenčná oblasť a pre $\boldsymbol{\beta}$ odvodená touto metódou.

Poznámka. Z definícií o asymptotických účinnostiach sekvenčných metód je zrejmé, že asymptotická účinnosť 2. rádu je silnejšia ako asymptotická účinnosť 1. rádu.

2.1 Dvojstupňová metóda

Ako už bolo spomenuté, problémom pri určení celkového rozsahu výberu je neznámy parameter σ^2 v lineárnom modeli. V roku 1956 pán Healy aplikoval Steinovu dvojstupňovú metódu (z roku 1945) na lineárny regresný model za predpokladu viacrozmerného normálneho rozdelenia vektora chýb.

Myšlienkou Steinovej dvojstupňovej metódy bolo zvoliť si najprv rozsah prvého výberu (1. stupeň), pomocou ktorého odhadneme parameter σ^2. Vďaka tomuto odhadu už budeme schopný určiť rozsah výberu N (2. stupeň), a teda dokážeme zostrojiť konfidenčnú množinu pre $\boldsymbol{\beta}$ s koeficientom spoľahlivosti $1-\alpha$.

Popíšme si túto metódu podrobnejšie.

STUPEŇ 1: Majme náhodný vektor $\mathbf{Y}_m = (Y_1, \ldots, Y_m)^\top$ a maticu daných čísel $\mathbf{X}_m, (m \geq p+1)$ pomocou ktorých spočítame

$$\widehat{\boldsymbol{\beta}}_m = (\mathbf{X}_m^\top \mathbf{X}_m)^{-1} \mathbf{X}_m^\top \mathbf{Y}_m \tag{2.1}$$

$$S_m^2 = (\mathbf{Y}_m - \mathbf{X}_m \widehat{\boldsymbol{\beta}}_m)^\top (\mathbf{Y}_m - \mathbf{X}_m \widehat{\boldsymbol{\beta}}_m) \tag{2.2}$$

Nech $F_{p,m-p}(\alpha)$ označuje kritickú hodnotu F rozdelenia o p a $m-p$ stupňoch voľnosti na hladine α. Definujme

$$b \equiv b_{p,m,\alpha} = pF_{p,m-p}(\alpha). \tag{2.3}$$

Na prvý pohľad nie je jasné, prečo práve takto volíme parameter b. Je to z dôvodu, aby odhadnutý parameter $\boldsymbol{\beta}$ ležal v konfidenčnej množine $\mathcal{R}_N$ s pravdepodobnosťou aspoň $1-\alpha$, čo dokážem vo vete 2.2. Definujme zastavovaciu veličinu nasledovne:

$$N \equiv N(d) = \max\left\{m, \left\lfloor \frac{bS_m^2}{d^2} \right\rfloor + 1\right\}, \tag{2.4}$$

kde symbolom $\lfloor a \rfloor$ označujeme najväčšie celé číslo $\leq a$.

STUPEŇ 2: Ak pre náhodnú veličinu N platí

- $N = m$, potom ukončíme výber po prvom stupni a položíme

$$\widehat{\boldsymbol{\beta}}_N = \widehat{\boldsymbol{\beta}}_m \tag{2.5}$$

- $N > m$, vyberieme ďalších $N-m$ pozorovaní $Y_{m+1}, \ldots, Y_N$ a matica $\mathbf{X}_m$ sa zväčší o $N-m$ daných riadkov. Potom definujme

$$\widehat{\boldsymbol{\beta}}_N = (\mathbf{X}_N^\top \mathbf{X}_N)^{-1} \mathbf{X}_N^\top \mathbf{Y}_N \tag{2.6}$$

Poznamenajme, že pri zavedení elipsoidnej množiny $\mathcal{R}_N$ z (1.5) sme požadovali, aby matica $\mathbf{X}_N$ bola dopredu určená. Možností ako určiť $\boldsymbol{x}_{m+1}, \ldots, \boldsymbol{x}_N$ je celý rad. My budeme uvažovať, že

$$\text{Matica } \mathbf{X}_N \text{ vznikne z matice } \mathbf{X}_m \text{ postupným opakovaním riadkov z } \mathbf{X}_m. \tag{2.7}$$

Tým máme na mysli, že $k, m+k, 2m+k, \ldots$-tý riadok v matici $\mathbf{X}_N$ bude rovnaký ako k-tý riadok v matici X_m pre $k = 1 \ldots m$. S takýmto rozšírením matice $\mathbf{X}_m$ budeme pracovať aj v nasledujúcich sekciách o sekvenčných metódach.

Všimnime si, že zastavovacia veličina N z (2.4) je konečná s pravdepodobnosťou 1. Preto za konfidenčnú množinu pre $\boldsymbol{\beta}$ intuitívne volíme množinu $\mathcal{R}_N$, ktorú sme definovali v (1.5). Vo vete 2.2 overíme, že voľba tejto konfidenčnej množiny je správna v zmysle, že koeficient spoľahlivosti je aspoň $1-\alpha$.

Predtým si dokážme nejaké pomocné tvrdenia.

Veta 2.1. *Pre náhodnú veličinu N určenú vzťahom* (2.4) *a pre všetky pevné $\boldsymbol{\beta}, \sigma, p, m, \alpha$, platia nasledujúce vlastnosti:*

a) $\frac{b\sigma^2}{d^2} \leq \mathsf{E}_{\boldsymbol{\beta},\sigma}\big(N\big) \leq m + \frac{b\sigma^2}{d^2}$

b) náhodné veličiny $\widehat{\boldsymbol{\beta}}_N$ a S_m^2 sú na seba nezávislé

c) náhodná veličina $\widehat{\boldsymbol{\beta}}_N$ je nestranným odhadom $\boldsymbol{\beta}$ a jej rozptyl je $\sigma^2(\mathbf{X}_N^\top \mathbf{X}_N)^{-1}$

d) $Y \equiv (\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})^\top[\sigma^{-2}(\mathbf{X}_N^\top \mathbf{X}_N)](\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})$ má podmienené χ_p^2 rozdelenie pri pevne zvolenom S_m^2 a je naň nezávislé

e) $Z \equiv \frac{p^{-1}Y}{S_m^2\sigma^{-2}}$ má podmienné $F_{p,m-p}$ rozdelenie pri pevne zvolenom S_m^2

kde $b \equiv b_{p,m,\alpha}$ bolo určené v (2.3), *S_m^2 je dané vzťahom* (2.2) *a $\widehat{\boldsymbol{\beta}}_N$ je z* (2.6).

Dôkaz. a) Z podmienky pre zastavovaciu veličinu N platí:

$$\begin{aligned} \frac{bS_m^2}{d^2} \leq \quad N \quad &\leq m + \frac{bS_m^2}{d^2} \quad \text{s.j, teda platí aj} \\ \frac{b\,\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(S_m^2\big)}{d^2} \leq E(N) &\leq m + \frac{b\,\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(S_m^2\big)}{d^2}. \end{aligned} \tag{2.8}$$

Z vlastnosti $\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(S_m^2\big) = \sigma^2$ dostávame platnosť tvrdenia a).

b) Pracujme s modelom:

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + e_i \qquad i = 1 \ldots m, \quad e_i \sim N(0, \sigma^2) \tag{2.9}$$

Maticový (vektorový) zápis: $\mathbf{Y}_m = \mathbf{X}_m\boldsymbol{\beta} + \mathbf{e}_m \quad h(\mathbf{X}_m) = p$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mp} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_m \end{pmatrix}. \tag{2.10}$$

Z vety 1.1 a) vieme, že platí

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_m &= (\mathbf{X}_m^\top \mathbf{X}_m)^{-1}\mathbf{X}_m^\top \mathbf{Y}_m = (\mathbf{X}_m^\top \mathbf{X}_m)^{-1}\mathbf{X}_m^\top(\mathbf{X}_m\boldsymbol{\beta} + \mathbf{e}_m) \\ &= \boldsymbol{\beta} + (\mathbf{X}_m^\top \mathbf{X}_m)^{-1}\mathbf{X}_m^\top \mathbf{e}_m. \end{aligned} \tag{2.11}$$

Označme si $\mathbf{A}_m = (\mathbf{X}_m^\top \mathbf{X}_m)^{-1}$. O tejto matici vieme, že je štvorcová $(p \times p)$, symetrická, regulárna a teda má lineárne nezávislé riadky aj stĺpce. Nech a_{ij}

označuje (i,j)-tý prvok matice $\mathbf{A}_m$. Pre lepšiu predstavivosť rozpíšem maticu $\mathbf{A}_m$ a vektor $\mathbf{X}_m^\top \mathbf{e}_m$.

$$\mathbf{A}_m = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pp} \end{pmatrix} \qquad \mathbf{X}_m^\top \mathbf{e}_m = \begin{pmatrix} \sum_{i=1}^m x_{i1}e_i \\ \vdots \\ \sum_{i=1}^m x_{ip}e_i \end{pmatrix} \overset{ozn.}{=} \begin{pmatrix} E_1 \\ \vdots \\ E_p \end{pmatrix} \overset{ozn.}{=} \mathbf{E}_m. \quad (2.12)$$

Teraz si rozpíšme vzťah $\widehat{\boldsymbol{\beta}}_m = \mathbf{A}_m\mathbf{E}_m$ z (2.11) po zložkách:

$$\begin{aligned} \widehat{\beta}_{1m} &= \beta_1 + a_{11}E_1 + a_{12}E_2 + \cdots + a_{1p}E_p \\ &\vdots \\ \widehat{\beta}_{pm} &= \beta_p + a_{1p1}E_1 + a_{p2}E_2 + \cdots + a_{pp}E_p \end{aligned} \quad (2.13)$$

Tieto zložky sú z vety 1.1 d) nezávislé na S_m^2. Pre dané S_m^2 dopočítame N a rozšírme náš model (2.10) o $N-m$ nových pozorovaní: $\mathbf{Y}_N = \mathbf{X}_N\boldsymbol{\beta} + \mathbf{e}_N$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_m \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mp} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_m \\ \vdots \\ e_N \end{pmatrix}. \quad (2.14)$$

Zaveďme analogicky $\widehat{\boldsymbol{\beta}}_N, \mathbf{A}_N$ a $\mathbf{E}_N$. Pri danom S_m^2 je matica $\mathbf{X}_N$ pevne daná. Rozpíšme si $\widehat{\boldsymbol{\beta}}_N = \mathbf{A}_N\mathbf{E}_N$ podobne ako v (2.13) po zložkách. Dostaneme

$$\begin{aligned} \widehat{\beta}_{1N} &= \beta_1 + A_{11}E_1' + A_{12}E_2' + \cdots + A_{1p}E_p' \\ &\vdots \\ \widehat{\beta}_{pN} &= \beta_p + A_{p1}E_1' + A_{p2}E_2' + \cdots + A_{pp}E_p' \end{aligned} \quad (2.15)$$

Cieľom je ukázať, že tieto zložky sú nezávislé na S_m^2. Ideou dôkazu tohto tvrdenia je vyjadriť si vektor $\mathbf{E}_N$ pomocou $\mathbf{E}_m$ t.j.:

$$\mathbf{E}_N = \begin{pmatrix} E_1' \\ \vdots \\ E_p' \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N x_{i1}e_i \\ \vdots \\ \sum_{i=1}^N x_{ip}e_i \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m x_{i1}e_i + \sum_{i=m+1}^N x_{i1}e_i \\ \vdots \\ \sum_{i=1}^m x_{ip}e_i + \sum_{i=m+1}^N x_{ip}e_i \end{pmatrix} = \mathbf{E}_m + \begin{pmatrix} E_{m+1,1}^N \\ \vdots \\ E_{m+1,p}^N \end{pmatrix}$$

Vzťah (2.15) môžme teraz rozpísať ako

$$\begin{aligned} \widehat{\beta}_{1N} &= \beta_1 + A_{11}E_1 + A_{12}E_2 + \cdots + A_{1p}E_p + \sum_{i=1}^p A_{1i}E_{m+1,i}^N \\ &\vdots \\ \widehat{\beta}_{pN} &= \beta_p + A_{p1}E_1 + A_{p2}E_2 + \cdots + A_{pp}E_p + \sum_{i=1}^p A_{pi}E_{m+1,i}^N \end{aligned} \quad (2.16)$$

Zrejme $\sum_{i=1}^{p} A_{ji} E_{m+1,i}^{N}$ pre každé $j = 1 \dots p$ je nezávislá na S_m^2. Potrebujeme ukázať už len nezávislosť súčtov $\sum_{i=1}^{p} A_{ji} E_i$ pre každé $j = 1 \dots p$ na veličine S_m^2. Všimnime si, že zo vzťahu (2.13) pre $\widehat{\boldsymbol{\beta}}_m$ sme zistili nezávislosť súčtov $\sum_{i=1}^{p} a_{ji} E_i$ pre každé $j = 1 \dots p$ na veličine S_m^2. Tým pádom aj lineárna kombinácia týchto súčtov je nezávislá na S_m^2.

Keďže matica $\mathbf{A}_m$ je regulárna, tak jej riadkové vektory generujú celý p-rozmerný priestor, a preto pomocou týchto súčtov vieme nagenerovať aj jednotlivé súčty $\sum_{i=1}^{p} A_{ji} E_i$ pre každé $j = 1 \dots p$. Tým pádom sme ukázali, že $\widehat{\boldsymbol{\beta}}_N$ je nezávislé na S_m^2.

c) Priamym dôsledkom tvrdenia b) je

$$\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(\widehat{\boldsymbol{\beta}}_N | S_m^2\big) = \mathsf{E}_{\boldsymbol{\beta},\sigma}\big(\widehat{\boldsymbol{\beta}}_N\big) = \boldsymbol{\beta} \quad \mathsf{Var}_{\boldsymbol{\beta},\sigma}(\widehat{\boldsymbol{\beta}}_N | S_m^2) = \mathsf{Var}_{\boldsymbol{\beta},\sigma}(\widehat{\boldsymbol{\beta}}_N) = \sigma^2 (\mathbf{X}_N^\top \mathbf{X}_N)^{-1}.$$

d) Platnosť tohto tvrdenia vyplýva z vlastností $b), c)$ a vety 1.1 e).
e) Z tvrdenia d) vieme, že Y má podmienené χ_p^2 rozdelenie pri pevne zvolenom S_m^2 a je naň nezávislé. Z vety 1.1 c) vieme, že $(n-p)S_m^2/\sigma^2 \sim \chi_{n-p}^2$. Z rozdelenia podielu dvoch nezávislých náhodných veličín viď (Anděl, 2007, str. 61, veta 4.17.), dostávame platnosť tvrdenia e). $\square$

Veta 2.2. *Pre náhodnú veličinu $N \equiv N(d)$ určenú vzťahom* (2.4) *a pre všetky pevné $\boldsymbol{\beta}, \sigma, p, m, \alpha$, platia nasledujúce vlastnosti:*

a) $\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{\boldsymbol{\beta} \in \mathcal{R}_N\Big\} \geq 1 - \alpha$

b) $\lim_{d \to 0_+} \mathsf{E}_{\boldsymbol{\beta},\sigma}\left[\frac{N}{C}\right] = \frac{b}{a}$

c) $\lim_{d \to 0_+} \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{\boldsymbol{\beta} \in \mathcal{R}_N\Big\} = 1 - \alpha$

kde $b \equiv b_{p,m,\alpha}$ je dané v (2.3), *$\mathcal{R}_N$ je z* (1.5) *a $C = \frac{a\sigma^2}{d^2}$ bolo určené v* (1.4).

Dôkaz. a) Nech náhodná veličina Y má χ_p^2 rozdelenie a nech je nezávislá na S_m^2. Potom konfidenčný koeficient $\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{\boldsymbol{\beta} \in \mathcal{R}_N(d)\Big\}$ z (1.7) je:

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\beta},\sigma}\left[F\left(\frac{Nd^2}{\sigma^2}\right)\right] &\geq \mathsf{E}_{\boldsymbol{\beta},\sigma}\left[F\left(\frac{bS_m^2}{\sigma^2}\right)\right] \qquad N \geq \frac{bS_m^2}{d^2} \text{ s. j. z definície} \\
&= \mathsf{E}_{\boldsymbol{\beta},\sigma}\Big[\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{Y \leq bS_m^2\sigma^{-2} | N\Big\}\Big] \\
&= \mathsf{P}_{\boldsymbol{\beta},\sigma}\left\{\frac{p^{-1}Y}{S_m^2\sigma^{-2}} \leq p^{-1}b\right\} \qquad Y \sim \chi_p^2, Y \perp\!\!\!\perp S_m^2 \\
&= \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{F_{p,m-p} \leq p^{-1}b\Big\} \\
&= 1 - \alpha, \quad \text{keď } p^{-1}b = F_{p,m-p}(\alpha).
\end{aligned} \tag{2.17}$$

Preto definujeme $b \equiv b_{p,m,\alpha} = pF_{p,m-p}(\alpha)$ ako som uviedol v (2.3).
b) Predelením nerovnosti z vety 2.1 a) s C dostávame

$$\frac{b}{a} \leq \mathsf{E}_{\boldsymbol{\beta},\sigma}\big(N/C\big) \leq \frac{md^2}{\sigma^2} + \frac{b}{a}. \tag{2.18}$$

Uvedomme si, že m je pevné, a preto limitným prechodom pre d dostávame dokazované tvrdenie.
c) Zo vzťahu (2.8) vyplýva rovnosť $\lim_{d\to 0_+} d^2 N = bS_m^2$ s. j., čo je konečné kladné číslo. Zvyšok dôkazu potom vyplýva z časti a). $\square$

Zistili sme, že dvojstupňová metóda nie je asymptoticky účinná, a preto celkový výber N nie je optimálny. To bolo hlavnou príčinou k objavovaniu ďalších sekvenčných metód.

2.2 Modifikovaná dvojstupňová metóda

Modifikovaná dvojstupňová metóda vznikla vylepšením dvojstupňovej metódy so zámerom, aby sme dosiahli asymptotickú účinnosť 1. rádu (2.2). Myšlienka ako to spraviť spočíva v šikovnej voľbe rozsahu prvého výberu. Zvoľme teda ľubovoľné kladné číslo γ a definujme rozsah prvého výberu nasledovne

$$m \equiv m(d) = \max\Big\{p+1, \Big\lfloor \Big(\frac{a}{d^2}\Big)^{1/(1+\gamma)} \Big\rfloor + 1\Big\} \text{ a definujme} \tag{2.19}$$

$$N \equiv N(d) = \max\Big\{m, \Big\lfloor \frac{bS_m^2}{d^2} \Big\rfloor + 1\Big\}, \tag{2.20}$$

kde $b \equiv b_{p,m,\alpha}$ z (2.3). Analogicky ako v dvojstupňovej procedúre sa pozrieme na vzťah medzi náhodnou veličinou N a m.

Ak platí $N = m$, tak nepotrebujeme spraviť ďalšie pozorovania. Pre $N > m$, potrebujeme rozšíriť rozsah výberu o $N - m$ nových pozorovaní. Kombináciou s predošlými pozorovaniami odhadneme $\boldsymbol{\beta}$ opäť metódou najmenších štvorcov rovnako ako v (2.6).

Vidíme, že N je opäť konečné s pravdepodobnosťou 1, a preto za konfidenčnú množinu pre $\boldsymbol{\beta}$ s koeficientom spoľahlivosti $1-\alpha$ intuitívne volíme $\mathcal{R}_N$ z (1.5). Správnosť voľby konfidenčnej množiny plynie z nasledujúcej vety.

Veta 2.3. *Pre náhodnú veličinu N z (2.20) a pre všetky pevné $\beta, \sigma, p, m, \alpha$ platia nasledujúce vlastnosti:*

a) $\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{\boldsymbol{\beta} \in \mathcal{R}_N\Big\} \geq 1-\alpha;$

b) $\lim_{d\to 0_+} \mathsf{E}_{\boldsymbol{\beta},\sigma}\big[\frac{N}{C}\big] = 1;$

c) $\lim_{d\to 0_+} \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{\boldsymbol{\beta} \in \mathcal{R}_N\Big\} = 1-\alpha;$

kde $b \equiv b_{p,m,\alpha}$ z (2.3), $\mathcal{R}_N$ je z (1.5) a $C = \frac{a\sigma^2}{d^2}$ je definované v (1.4).

Dôkaz. Dôkaz časti a) a c) prebieha analogicky ako vo vete (2.2).
b) Nech náhodná veličina Y má χ_p^2 rozdelenie a nech je nezávislá na S_m^2. Zo vzťahu (2.17) sme zistili:

$F_{p,m-p} = p^{-1}Y/(S_m^2\sigma^{-2})$ a teda pre $d \to 0_+$
$F_{p,m-p}$ konverguje v distribúcii k $p^{-1}Y \Rightarrow \lim\limits_{d\to 0_+} ba^{-1} = 1.$

Zvyšok dôkazu vyplýva limitným prechodom vo vzťahu (2.18). $\square$

Pozorný čitateľ si určite všimol, že rozsah prvého výberu závisí na voľbe γ. Ak očakávame, že náhodná veličina N bude veľká, tak je logické zvoliť aj väčší rozsah prvého výberu. Na druhej strane si musíme dať pozor, aby sme nezvolili väčší rozsah ako by bol optimálny (C), ktorý je neznámy. Pre rozsah prvého výberu m platia nasledujúce vzťahy:

$$\begin{aligned} \lim_{d\to 0_+} m &= \infty \\ \lim_{d\to 0_+} (m/C) &= \sigma^{-2} \lim_{d\to 0_+} (d^2/a)^{\gamma/(1+\gamma)} = 0 \end{aligned} \tag{2.21}$$

Práve z druhého vzťahu nám vyplýva, že voľba m bude menšia ako optimálna hodnota C. Preto je vzorec pre m dobre definovaný. Voľba vhodného parametru γ sa vo veľkej väčšine prípadov dopočítava na počítači.

Príklad (Rola γ v modifikovanej dvojstupňovej procedúry)

Pre rôzne hodnoty γ budeme určovať rozsah prvého výberu pomocou vzorca (2.19). Vstupné parametre pre simuláciu úlohy volíme: d=0.10, $\alpha = 0.05$, p=1.

Postup: Zvolíme postupne 10 rôznych hodnôt pre $\gamma = 0.1, 0.2, \ldots, 1$ a výsledky pre $m = \max\{2, \lfloor (a/d^2)^{1/(1+\gamma)} \rfloor + 1\}$ zapíšeme do tabuľky. Pripomínam, že parameter a v tomto vzorci symbolizuje kritickú hodnotu rozdelenia χ^2_p na hladine α. V našom prípade je $a = \chi^2_{1,0.05} \doteq 3.841$.

Vstupné dáta:	d	0.10		
	alpha	0.05		
	gamma	0.1		
	alpha	d	gamma	m
	0.05	0.10	0.10	363
	0.05	0.10	0.20	342
	0.05	0.10	0.30	323
	0.05	0.10	0.40	306
	0.05	0.10	0.50	290
	0.05	0.10	0.60	275
	0.05	0.10	0.70	261
	0.05	0.10	0.80	248
	0.05	0.10	0.90	235
	0.05	0.10	1.00	224

Tabuľka 2.1: Rola γ v modifikovanej dvojstupňovej procedúre

Záver: Z tabuľky si môžme povšimnúť, že čím väčšie γ zvolíme, tým menšie m dostaneme. Preto sa vo väčšine prípadov nastavuje γ okolo hodnoty 0.10. $\diamondsuit$

2.3 Priama sekvenčná metóda

Medzi ďalšie metódy, ktoré zabezpečujú asymptotickú účinnosť 1. rádu, patrí priama sekvenčná metóda (angl. Purely sequential procedure). Motiváciou k vzniku tejto metódy bolo dosiahnutie asymptotickej účinnosti 2. rádu, ktorá u modifikovanej dvojstupňovej metódy nebola dosiahnutá.

Voľba celkového rozsahu výberu N sa teda nebude výrazne líšiť od dorátaného optimálneho rozsahu pri známom σ^2. Profesor Mukhopadhyay a pán Abid odvodili v roku 1986 túto metódu pre lineárny model. Popíšme si ju.

Rovnako ako v dvojstupňovej sekvenčnej metóde začneme s $m \geq p+1$ pozorovaní. Myšlienkou priamej metódy je postupne po jednom pridávať nové pozorovania a zakaždým odhadnúť neznámy parameter σ^2. Po každom pridaní určíme, či nám už aktuálny rozsah výberu stačí alebo nie. Zastavovaciu veličinu určíme rovnako ako v knihe (Mukhopadhyay a Silva, 2009, str. 319):

$$N \equiv N(d) \quad \begin{array}{l} \text{najmenšie prirodzené číslo } n \geq m, \\ \text{pre ktoré platí } n \geq aS_n^2/d^2. \end{array} \tag{2.22}$$

Pre väčšiu zreteľnosť rozpíšme túto procedúru podrobnejšie.

Majme náhodný vektor $\mathbf{Y}_m = (Y_1, \ldots, Y_m)^\top$ a maticu daných čísiel $\mathbf{X}_m$ typu $m \times p$ pre $m \geq p+1$. Pomocnou známych vzorcov dopočítame S_m^2 a overíme, či $m \geq aS_m^2/d^2$. Ak áno, tak naša procedúra skončí a $N = m$. Ak nie, tak pridáme jedno ďalšie pozorovanie Y_{m+1} a matica $\mathbf{X}_m$ sa nám rozšíri o jeden nový riadok podľa vzoru (2.7). Vylepšíme odhad σ^2 dopočítaním S_{m+1}^2 a skontrolujeme, či $m+1 \geq aS_{m+1}^2/d^2$. Ak áno, tak naša procedúra skončí a $N = m+1$. Ak nie, tak pridáme jedno ďalšie pozorovanie Y_{m+2}.... Tento postup opakujeme až kým nebude platiť $n \geq aS_n^2/d^2$.

Poznamenajme, že N je opäť konečné s pravdepodobnosťou 1, a preto za konfidenčnú množinu pre $\boldsymbol{\beta}$ volíme množinu $\mathcal{R}_N$ z (1.5). Nevýhodou tejto metódy je zrejme časová náročnosť. Napriek tomu má táto metóda aj dobré vlastnosti, ktoré si sformulujeme v nasledujúcej vete.

Veta 2.4. *Pre náhodnú veličinu N z (2.22) a pre všetky pevné $\beta, \sigma, p, m, \alpha$ platia nasledujúce vlastnosti:*

a) $\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(N\big)$ *je konečné číslo s.j.*

b) $\lim_{d\to 0_+} \mathsf{E}_{\boldsymbol{\beta},\sigma}\left[\frac{N}{C}\right] = 1$

c) $\lim_{d\to 0_+} \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big\{\boldsymbol{\beta} \in \mathcal{R}_N\Big\} = 1-\alpha$

kde $\mathcal{R}_N$ je z (1.5) a $C = \frac{a\sigma^2}{d^2}$ bolo určené v (1.4).

Dôkaz. Obdoba dôkazu sa nájde v knihe (Mukhopadhyay a Silva, 2009, str. 303). □

Priama sekvenčná metóda dokonca zaisťuje aj asymptotickú účinnosť druhého rádu. Dôležité si je však uvedomiť, že postupom akým sa volí náhodná veličina N, vyplýva nekonzistentnosť metódy. Udalosť $N = n$ totiž nie je rovnaká ako udalosť $n \geq aS_n^2/d^2$. Musí zároveň platiť, že pre všetky prirodzené čísla k, $m \leq k < n$, platí $k < aS_k^2/d^2$. Preto vzniká problém pri určení distribučnej funkcie náhodnej veličiny N, a teda aj pri dopočítaní konfidenčnéhoho koeficientu.

Našťastie sa ukázalo, že pri pridaní niekoľko ďalších pozorovaní k celkovému výberu dokážeme okrem zachovania asymptotickej účinnosti zaistiť aj konzistentnosť metódy. Zhrňme si na záver do tabuľky 2.2 základné vlastnosti sekvenčných metód, o ktorých sme sa doteraz dozvedeli.

Vlastnosti	Dvojstupňová metóda	Modifikovaná dvojstupňová metóda	Priama sekvenčná metóda*
Konzistencia	Áno	Áno	Áno
Asymptotická konzistencia	Áno	Áno	Áno
As. účinnosť 1. rádu	Nie	Áno	Áno
As. účinnosť 2. rádu	Nie	Nie	Áno

Poznámka: * Vlastnosti priamej sekvenčnej metódy už pri pridaní dodatočných pozorovaní

Tabuľka 2.2: Súhrn vlastností sekvenčných metód

Každá metóda má svoje výhody a nevýhody. Z tabuľky by sa mohlo zdať, že priama sekvenčná metóda je najlepšia, ale práve spomínaná časová náročnosť viedlo štatistikov k objavovaniu ďalších sekvenčných metód. Ak by sa chcel čitateľ dozvedieť viac o ďalších sekvenčných metódach, tak silne doporučujem knižku (Mukhopadhyay a Silva, 2009, 6. kapitola).

2.4 Teória v praxi

V predchádzajúcej sekcii sme si ukázali tri sekvenčné procedúry, ktoré nám dopomáhajú zostrojiť konfidenčnú množinu pre $\boldsymbol{\beta}$ s daným koeficientom spoľahlivosti $1-\alpha$. Poďme si ukázať ich použitie na konkrétnych príkladoch.

Pri konštrukcii konfidenčnej množiny potrebujeme poznať iba hodnotu náhodnej veličiny N. Na základe nej už ľahko určíme konfidenčnú množinu $\mathcal{R}_N$ so spoľahlivosťou $1-\alpha$ definovanú v (1.5). Na spočítanie náhodnej veličiny N budeme využívať matematický softvér R, ktorý bol naprogramovaný z dôvodu uľahčenia práce so štatistickými úlohami a dopomáha k simulácii dát a overovaní hypotéz.

Pre jednoduchosť a lepšiu predstavivosť sa budeme zaoberať iba príkladmi riadené jednoduchým lineárnym modelom. Budeme teda predpokladať n nezávislých náhodných veličín Y_i, pre ktoré platí:

$$Y_i = \beta x_i + e_i, \qquad i = 1, \ldots, n \qquad e_i \sim N(0, \sigma^2) \tag{2.23}$$

kde β je neznámy parameter a x_i sú známe konštanty. Ide vlastne o lineárny regresný model, v ktorom $\mathbf{X}_n = (x_1, \ldots, x_n)^\top$ je matica typu $n \times 1$ a vektor $\boldsymbol{\beta}$ obsahuje iba jednu zložku β. Grafická predstava tohto modelu je priamka prechádzajúca počiatkom súradnicovej sústavy. Z vety 1.1 a) dostávame

$$\beta_n = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \quad a \quad S_n^2 = \frac{R}{n-1} = \frac{\sum_{i=1}^n (Y_i - \beta_n x_i)^2}{n-1}$$

Poznámka: Keďže odhadujeme iba jeden parameter, tak už nebudem ďalej písať, že konštruujeme konfidenčnú množinu o danej veľkosti a daným konfidenčným koeficientom $1-\alpha$. Namiesto toho budem písať, že konštruujeme interval spoľahlivosti s danou dĺžkou (d) a koeficientom spoľahlivosti $(1-\alpha)$. V nasledujúcom príklade použijeme dáta prevzaté zo skrípt (Anděl, 2007, str. 190).

Príklad : Sledovali sme priehyb plastickej hmoty Y_i v závislosti na tlaku x_i. V tabuľke 2.3 uvádzame namerané hodnoty:

Tlak x_i	2	4	6	8	10	12
Priehyb Y_i	14	35	48	61	80	93

Tabuľka 2.3: Priehyb plastickej hmoty v závislosti na tlaku

Zostrojte interval spoľahlivosti pre parameter β o dĺžke d a koeficientom spoľahlivosti $1-\alpha$ pomocou dvojstupňovej metódy. Môžete predpokladať, že sa náhodná veličina Y_i riadi lineárnym modelom z (2.23).

Riešenie : Pomocou vzorca $N \equiv N(d) = \max\{m, \lfloor bS_m^2/d^2 \rfloor + 1\}$ z (2.4) dokážeme dopočítať, koľko máme ešte priehybou plastickej hmoty napozerať. Následne už ľahko zostrojíme interval spoľahlivosti pre β o dĺžke d a koeficientom spoľahlivosti $1-\alpha$. Preto sa budeme zaoberať iba určením náhodnej veličiny N.

Nastavme vstupné parametre na $d = 0{,}5$ a $\alpha = 0{,}05$. V prvej fáze máme 6 pozorovaní, preto $m = 6$, $b = F_{1,5,0.5} \doteq 6{,}608$ a

$$\beta_6 = \frac{\sum_{i=1}^{6} x_i Y_i}{\sum_{i=1}^{6} x_i^2} \doteq 7{,}857 \quad a \quad S_6^2 = \frac{\sum_{i=1}^{6}(Y_i - \beta_6 x_i)^2}{5} \doteq 4{,}714.$$

Teraz už ľahko dopočítame $N = \max\left\{6, \left\lfloor \frac{6{,}608 \cdot 4{,}714}{0{,}5^2} \right\rfloor + 1\right\} = 125.$

Záver : Potrebujem namerať ešte 119 ďalších pozorovaní, aby sme vedeli zostrojiť interval spoľahlivosti pre β s koeficientom spoľahlivosti 0,95 a dĺžke intervalu 0,5.

$\diamondsuit$

Príklad (Porovnávanie sekvenčných metód)

V tomto príklade si overíme teoretické poznatky o sekvenčných metódach. Budeme simulovať dáta v R-ku a pomocou vzorcov (2.4), (2.20) a (2.22) určíme náhodnú veličinu N pri rôznych vstupných parametroch. Výsledky zaznamenáme do prehľadnej tabuľky a spravíme malú diskusiu o nich.

Postup: Nasimulujeme si dáta pre jednoduchý lineárny model z (2.23), v ktorom budeme navyše požadovať, aby vektor chýb mal normované normálne rozdelenie. Spravíme dokopy 1000 simulácii a v každej simulácii určíme náhodnú veličinu N. Priemernú hodnotu $\overline{N}$ porovnáme s optimálnou hodnotou, ktorú získame zo vzorca $C = a\sigma^2/d^2$, kde $a = \chi^2_{1,0.05} \doteq 3.841$.

Postup budeme opakovať s rovnakými simuláciami pre každú metódu. Vstupné parametre volíme nasledovne $m = 10, \gamma = 0.1, \alpha = 0.05$ pre $d_1 = 0.50, d_2 = 0.10, d_3 = 0.05, d_4 = 0.01$. Dosiahnuté výsledky som zapísal do tabuľky 2.4.

Táto tabuľka nám potvrdila hneď viacero poznatkov o sekvenčných metódach. Nikoho asi neprekvapilo, že zmenšovaním hodnoty d (určuje objem konfidenčnej množiny), nám narastá hodnota náhodnej veličiny.

Metóda	m	d	$\overline{N}$	C	$\overline{N}$/C	$\overline{N}$-C
Dvojstupňová	10	0.50	21	16	1.313	5
Dvojstupňová	10	0.10	507	385	1.317	122
Dvojstupňová	10	0.05	2025	1537	1.318	488
Dvojstupňová	10	0.01	50597	38415	1.317	12182
Modifikovaná	12	0.50	20	16	1.250	4
Modifikovaná	224	0.10	388	385	1.008	3
Modifikovaná	789	0.05	1542	1537	1.003	5
Modifikovaná	14713	0.01	38431	38415	1.0004	16
Modifikovaná	51881	0.005	153693	153640	1.0003	53
Priama sekvenčná	10	0.50	15	16	0.938	1
Priama sekvenčná	10	0.10	383	385	0.995	2
Priama sekvenčná	10	0.05	1535	1537	0.999	2
Priama sekvenčná	10	0.01	38400	38415	0.9996	**15**

Tabuľka 2.4: Porovnávanie sekvenčných metód pre rôzne voľby d

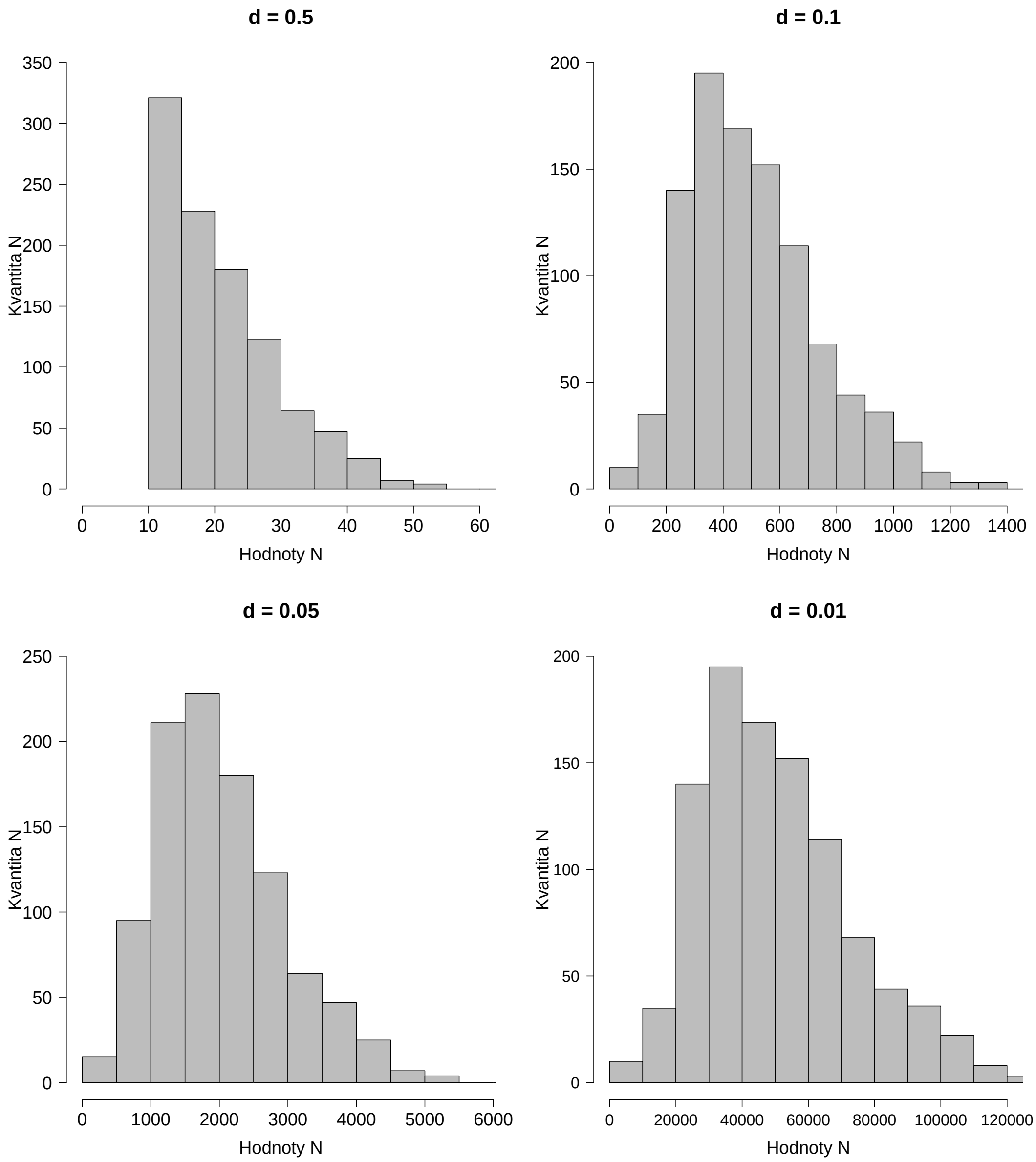
Zaujímavejšie si je všimnúť ako voľba m (rozsah prvého výberu) v modifikovanej metóde vylepší dvojstupňovú metódu. Pri dvojstupňovej metóde potrebujeme približne o 32% viac pozorovaní ako by sme v skutočnosti potrebovali t.j. je asymptoticky neúčinná. U modifikovanej a priamej sekvenčnej metódy potrebujeme približne rovnako veľa pozorovaní, čo nám potvrdzuje poznatok o asymptotickej účinnosti prvého rádu.

Všimnime si, že pre d približujúce sa k 0 sa v modifikovanej metóde rozdiel medzi náhodnou veličinou $\overline{N}$ a C pomaly zväčšuje. To znamená, že modifikovaná metóda nebude asymptoticky účinná druhého rádu. Keďže tento rozdiel výrazne nenarastá (nie ako v dvojstupňovej metóde) asymptotická účinnosť prvého rádu bude zachovaná.

U priamej metóde sa rozdiel $\overline{N} - C$ až tak nezväčšuje, čo potvrdzuje asymptotickú účinnosť 2. rádu. Ako som už spomínal priama metóda je časovo náročnou metódou. Pre zaujímavosť na samotný výpočet pre $d = 0.01$ som čakal deň a to som spravil len 45 simulácii a nie tisíc. Preto som zvýraznil v tabuľke číslo **15 tučne**. Na rozdiel od modifikovanej a dvojstupňovej metódy prebehol tento výpočet do minúty.

Na záver ponúkam zopár grafov 2.1 , v ktorých bude vidieť rozloženie náhodnej veličiny N, čo do počtu výskytov v 1000 simuláciách. Keďže pre každú metódu vyzerajú tieto grafy približne rovnako, tak uvádzam iba rozloženie náhodnej veličiny N v dvojstupňovej metóde pri rôznej voľbe d.

Z prvého grafu si môžme všimnúť, že výskyt náhodnej veličiny N bol najmä medzi hodnotami 10 a 15. Vyplýva to z nastavenia parametru $m = 10$ blízko optimálnej hodnoty $C = 16$. Viac aplikovaných príkladov sa môže zapálený čitateľ dozvedieť v knižke (Mukhopadhyay a kol., 2004).



Obr. 2.1: Rozloženie náhodnej veličiny N pre rôzne voľby d

Kapitola 3

Intervaly spoľahlivosti pre regresné parametre v lineárnom modeli

Ako už na začiatku 1. kapitoly bolo spomenuté, medzi základné úlohy matematickej štatistiky patrí konštruovanie intervalových odhadov s danou spoľahlivosťou.

Cieľom tejto kapitoly je ukázať spôsob ako skonštruovať intervalový odhad, za pomoci dvojstupňovej metódy, v ktorom budú regresné parametre ležať s danou spoľahlivosťou. Naďalej sa budeme riadiť lineárnym modelom definovaným v (1.1) a postupovať budeme obdobne ako v sekcii o dvojstupňovej metóde.

STUPEŇ 1: Majme náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_m)^\top$ a maticu daných čísiel $\mathbf{X}_{m\times p}$, pomocou ktorých spočítame

$$\widehat{\boldsymbol{\beta}}_m = (\mathbf{X}_m^\top \mathbf{X}_m)^{-1}\mathbf{X}_m^\top \mathbf{Y}_m$$
$$S_m^2 = (\mathbf{Y}_m - \mathbf{X}_m\widehat{\boldsymbol{\beta}}_m)^\top(\mathbf{Y}_m - \mathbf{X}_m\widehat{\boldsymbol{\beta}}_m)$$

Nech $z > 0$. Definujme zastavovaciu veličinu nasledovne:

$$N = \max\left\{m, \left\lfloor z^{-1}S_m^2 \right\rfloor + 1\right\}. \tag{3.1}$$

Význam voľby z vysvetlím neskôr.

STUPEŇ 2: Ak pre N z (3.1) platí

- $N > m$, spravíme ďalších $N - m$ nezávislých pozorovaní Y_{m+1},...,Y_N a matica $\mathbf{X}_m$ sa zväčší o $N - m$ nových riadkov daných čísiel v zmysle (2.7). Potom definujme:

$$\widehat{\boldsymbol{\beta}}_N = (\mathbf{X}_N^\top \mathbf{X}_N)^{-1}\mathbf{X}_N^\top \mathbf{Y}_N.$$

- $N = m$, potom ukončíme výber po prvom stupni a položíme

$$\widehat{\boldsymbol{\beta}}_N = \boldsymbol{\beta}_m$$

Nasledujúca veta nám dá nástroj ako dokážeme skonštruovať intervalový odhad pre regresné parametre v lineárnom modeli.

Veta 3.1. *Nech v_{ij}^N označuje prvok v matici $(\mathbf{X}_N^\top \mathbf{X}_N)^{-1}$na (i,j)-tej pozícii. Potom pre každé $i = 1 \dots p$ má náhodná veličina*

$$T_i = \frac{\widehat{\beta}_i^N - \beta_i}{\sqrt{S_m^2 v_{ii}^N}}$$

Studentovo t rozdelenie s $m - p$ stupňami voľnosti.

Dôkaz. Z vety 2.1 c) vieme, že pri danom S_m^2 je N pevné a $\widehat{\beta}_i^N$ má podmienené $N(\beta_i, \sigma^2 v_{ii}^N)$ rozdelenie a je nezávislé na S_m^2. Tým pádom

$$Y_i = \frac{\widehat{\beta}_i^N - \beta_i}{\sigma \sqrt{v_{ii}^N}}$$

je tiež nezávislé na S_m^2 a má podmienené $N(0,1)$ rozdelenie vzhľadom k S_m^2. Z vety 1.1 c) vieme, že $(m-p)S_m^2/\sigma^2 \sim \chi_{m-p}^2$. Preto podielom dvoch nezávislých veličín má

$$T_i = \frac{\widehat{\beta}_i^N - \beta_i}{\sqrt{S_m^2 v_{ii}^N}} = \frac{\frac{\widehat{\beta}_i^N - \beta_i}{\sigma\sqrt{v_{ii}^N}}}{\sqrt{\frac{(m-p)S_m^2}{\sigma^2(m-p)}}},$$

Studentovo t rozdelenie s $m - p$ stupňami voľnosti. □

Nech $t_{m-p}(\alpha/2)$ označuje kritickú hodnotu t rozdelenia s $m - p$ stupňami voľnosti na hladine $\alpha/2$. Potom pre každé $i = 1 \dots p$ je

$$\widehat{\beta}_i^N \mp t_{m-p}(\alpha/2) S_m \sqrt{v_{ii}^N}$$

intervalovým odhadom pre neznámy parameter β_i s koeficientom spoľahlivosti $1 - \alpha$. Ak by sme požadovali, aby tento interval spoľahlivosti pre parameter β_i bol dlhý maximálne $2d$, muselo by platiť

$$\begin{aligned} t_{m-p}(\alpha/2) S_m \sqrt{v_{ii}^N} &\leq d, \text{čo je ekvivalentné nerovnosti} \\ v_{ii}^N &\leq \frac{d^2}{t_{m-p}^2(\alpha/2) S_m^2} \end{aligned} \tag{3.2}$$

Poznamenajme, že posledná nerovnosť od určitého N nadobúda platnosti. Je to z dôvodu, že ľavá strana nám s rastúcim N klesá do nuly a pravá strana je zakaždým kladná. Preto ak pri definovaní náhodnej veličiny N v (3.1) zvolíme za z dostatočne malé číslo, dokážeme zabezpečiť, aby dĺžka intervalu spoľahlivosti bola maximálne $2d$.

Doteraz sme pracovali iba s klasickým lineárnym modelom, kde sme postupne dorábali nové pozorovania Y_i. Rád by som preto v nasledujúcej kapitole ukázal obmenený lineárny model, v ktorom budeme tiež konštruovať konfidenčné oblasti, ale aj testovať regresné parametre. Zmena bude spočívať v generovaní rovno n-rozmerných náhodných vektorov $\mathbf{Y}_n$, ktoré budú riadené danou maticou $\mathbf{X}_n$. Viac sa dozviete v nasledujúcej kapitole.

Kapitola 4

Konštrukcie konfidenčných množín a testovanie hypotéz v opakovanom lineárnom modeli

V tejto kapitole sa budeme zaoberať konštruovaním vhodnej štatistiky, pomocou ktorej by sme mohli v „opakovanom“ lineárnom modeli (vysvetlím za chvíľu) testovať regresné parametre.

Spomínané konštrukcie sú popísané v knihe (Ghosh a Sen, 1991, str. 29-31) pre všeobecnejší Aitkenov model. Základné informácie o Aitkenovom modeli si zvedavý čitateľ môže doplniť z knihy (Zvára, 2008, str. 23-25).

Predpokladajme, že máme dostatočne veľa nezávislých náhodných n-rozmerných vektorov $\mathbf{Y}_{1n}, \mathbf{Y}_{2n}, \ldots$ riadené nasledovným modelom:

$$\begin{aligned} &\mathbf{Y}_{jn} = \mathbf{X}_n \boldsymbol{\beta} + \mathbf{e}_{jn}, \text{ pre } j = 1, 2 \ldots \quad \mathbf{e}_{jn} \sim N_n(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n) \text{ a} \\ &\text{hodnosť matice } \mathbf{X}_n \text{ je } h(\mathbf{X}_n) = p, n \geq p + 1, \end{aligned} \tag{4.1}$$

V tomto modeli už nebudeme generovať nové náhodné veličiny Y_i a rozširovať maticu $\mathbf{X}_n$ o nové riadky. Namiesto toho rovno nagenerujeme n-rozmerný vektor, ktorý sa bude opakovane riadiť maticou $\mathbf{X}_n$. Preto tento lineárny model nazývame opakovaným (angl. Replicable linear model).

Skôr než zostrojíme testovú štatistiku potrebujeme najprv skonštruovať konfidenčnú množinu pre $\boldsymbol{\beta}$ s koeficientom spoľahlivosti $1 - \alpha$ a danou veľkosťou d. Postupujme podobne ako v sekvenčných metódach.

Zvoľme rozsah prvého výberu m, pre ktorý platí nerovnosť $\nu = mn - p > 1$. Majme teda m nezávislých náhodných n-rozmerných pozorovaní $\mathbf{Y}_{1n}, \ldots, \mathbf{Y}_{mn}$ (ďalej už nebudem písať dolný index $_n$) a danú maticu $\mathbf{X}_n$. Dopočítajme

$$\nu S_m^2 = \sum_{k=1}^{m} (\mathbf{Y}_k - \mathbf{X}_n \widehat{\boldsymbol{\beta}}_m)^\top (\mathbf{Y}_k - \mathbf{X}_n \widehat{\boldsymbol{\beta}}_m), \text{ kde} \tag{4.2}$$

$$\widehat{\boldsymbol{\beta}}_m = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \overline{\mathbf{Y}}_m, \quad \overline{\mathbf{Y}}_m = \frac{1}{m} \sum_{k=1}^{m} \mathbf{Y}_k. \tag{4.3}$$

Nech $z > 0$ a definujme $N = \max\{m + 1, \lfloor z^{-1} S_m^2 \rfloor + 1\}$. Vo väčšine prípadov volíme $z^{-1} = pF_{p,\nu}(\alpha)/d^2$, aby mala skonštruovaná konfidenčná množina danú veľkosť d s koeficientom spoľahlivosti aspoň $1 - \alpha$.

Nájdime $1 + N - m$ čísel $a_0, a_m, a_{m+1}, \ldots, a_N$, ktoré spĺňajú

$$a_0 + \sum_{k=m+1}^{N} a_k = 1; \qquad a_0^2/m + \sum_{k=m+1}^{N} a_k^2 = zS_m^{-2}. \tag{4.4}$$

Prvú podmienku zavádzame z dôvodu, aby $\widehat{\boldsymbol{\beta}}_N$ bol nestranným odhadom $\boldsymbol{\beta}$ (viď veta 4.2). Druhá podmienka slúži na ohraničenie rozptylu náhodnej veličiny $\widehat{\boldsymbol{\beta}}_N$, kde vhodnou voľbou z budeme schopný nastaviť silu testu. Nech

$$\mathbf{J}_N^{p\times 1} = a_0\overline{\mathbf{Y}}_m + \sum_{k=m+1}^{N} a_k\mathbf{Y}_k \tag{4.5}$$

Pri danom S_m^2 má $\mathbf{J}_N$ rozdelenie $N_n(\mathbf{X}_n\boldsymbol{\beta}, zS_m^{-2}\sigma^2\boldsymbol{I}_n)$. Dôkaz tohto tvrdenia je uvedený vo vete 4.1. Odhadnime vektor $\boldsymbol{\beta}$ metódou najmenších štvorcov pri minimalizovaní výrazu $(\mathbf{J}_N - \mathbf{X}_n\boldsymbol{\beta})^\top(\mathbf{J}_N - \mathbf{X}_n\boldsymbol{\beta})$. Z toho dostávame odhad

$$\widehat{\boldsymbol{\beta}}_N = (\mathbf{X}_n^\top\mathbf{X}_n)^{-1}\mathbf{X}_n^\top\mathbf{J}_N. \tag{4.6}$$

Pri danom S_m^2 má $\widehat{\boldsymbol{\beta}}_N$ podmienené normálne rozdelenie $N_p[\boldsymbol{\beta}, zS_m^{-2}\sigma^2(\mathbf{X}_n^\top\mathbf{X}_n)^{-1})]$ a je nezávislé na S_m^2 (viď veta 4.2). Vďaka tomuto poznatku už budeme schopný zostrojiť konfidenčnú množinu a test pre $\boldsymbol{\beta}$. Definujeme Q nasledovne

$$Q \equiv Q(\boldsymbol{\beta}) = z^{-1}p^{-1}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})^\top(\mathbf{X}_n^\top\mathbf{X}_n)(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}). \tag{4.7}$$

Z rozdelenia $\widehat{\boldsymbol{\beta}}_N$ vyplýva, že pS_m^2Q/σ^2 má podmienené χ_p^2 rozdelenie pri danom S_m^2 a je naň nezávislé. Z vety 1.1 c) dostávame, že $\nu S_m^2/\sigma^2$ má χ^2 rozdelenie s ν stupňami voľnosti. To implikuje, že Q má Fisherovo rozdelenie so stupňami voľnosti p, ν. Ak $F_{p,\nu}(\alpha)$ označuje kritickú hodnotu Fisherovho rozdelenia so stupňami voľnosti p, ν na hladine α, potom

$$\mathcal{R}_N = \{\boldsymbol{\beta}_{p\times 1} \in \mathbb{R}^p : (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_N)^\top(\mathbf{X}_n^\top\mathbf{X}_n)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_N) \le zpF_{p,\nu}(\alpha)\} \tag{4.8}$$

je konfidenčná množina pre $\boldsymbol{\beta}$ s konfidenčným koeficientom $1 - \alpha$. Tvar tejto elipsoidnej množiny je daný maticou $\boldsymbol{X}_n$. Vhodnou voľbou z dokážeme zmeniť veľkosť tejto množiny, tak ako si zaželáme. Požadovali sme, aby táto množina mala danú veľkosť d. Preto musí platiť nerovnosť:

$$zpF_{p,\nu}(\alpha) \le d^2 \quad \Rightarrow \quad z \le \frac{d^2}{pF_{p,\nu}(\alpha)}.$$

Z toho vidíme, prečo sme pri voľbe N položili $z^{-1} = pF_{p,\nu}(\alpha)/d^2$. Ak by sme teraz chceli skonštruovať test na testovanie regresných parametrov budeme musieť hodnotou z ešte „hýbať“. Potrebujeme nastaviť obmedzenie na pravdepodobnosti chýb prvého aj druhého druhu a to nám ovplyvní voľbu z. Pôjde o analógiu testovania hypotéz v práci (Rusá, 2013, str.17-18).

4.1 Testovanie hypotéz

Majme daný p-rozmerný vektor $\boldsymbol{\beta}_0$. Testujme hypotézu $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ proti alternatíve $H_1 : \boldsymbol{\beta} \ne \boldsymbol{\beta}_0$ na hladine α. Naša testová štatistika vyzerá:

$$Q_N = z^{-1}p^{-1}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0)^\top(\mathbf{X}_n^\top\mathbf{X}_n)^{-1}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0). \tag{4.9}$$

Za platnosti hypotézy H_0 má Q_N Fisherovo rozdelenie $F_{p,\nu}$. Tým pádom našu hypotézu zamietame na hladine α, keď platí $Q_N > F_{p,\nu}(\alpha)$. Z rovnakých dôvodov ako predtým má $pS_m^2 Q_N/\sigma^2$, pri pevnom S_m^2, podmienené rozdelenie $\chi^2(p|\nu_0)$ (necentrálne χ^2 rozdelenie s p stupňami voľnosti a parametrom necentrality ν_0), kde $\nu_0 = z^{-1}S_m^2\triangle/\sigma^2$ a

$$\triangle = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top(\mathbf{X}_n^\top\mathbf{X}_n)^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0). \tag{4.10}$$

Z toho potom plynie, že Q_N má podmienené rozdelenie $F(p,\nu|z^{-1}p^{-1}\triangle/\sigma^2)$ (necentrálne $F_{p,\nu}$ rozdelenie s parametrom necentrality $z^{-1}p^{-1}\triangle/\sigma^2$).

Pán Shoutir Kishore Chatterjje v knihe (Ghosh a Sen, 1991, str. 31) tvrdí, že dokážeme zvoliť také z, aby pravdepodobnosť chyby druhého druhu bola dostatočne obmedzená. Bohužiaľ sa mi nepodarilo toto tvrdenie overiť.

V nasledujúcich podsekciách budeme riešiť špeciálny prípad testovania regresných parametrov.

4.1.1 Intervalový odhad pre regresný parameter

Naznačili sme ako vieme testovať vektor neznámych parametrov. Poďme sa pozrieť ako by to vyzeralo ak by sme chceli testovať len jednu zložku tohto vektora. Majme daný odhad β_0. Testujme pre nejaké $i \in \{1,\ldots,p\}$ nulovú hypotézu $H_0 : \beta_i = \beta_0$ proti alternatíve $H_1 : \beta_i > \beta_0$ na hladine α.

Skonštruujme najprv interval spoľahlivosti pre β_i s dĺžkou $2d$ a koeficientom spoľahlivosti $1-\alpha$. Nech v_{ij} označuje prvok v matici $(\mathbf{X}_n^\top\mathbf{X}_n)^{-1}$na $(i{,}j)$-tej pozícii Definujme

$$N = \max\{m, \lfloor z^{-1}S_m^2\rfloor + 1\} \qquad z = \frac{d^2}{v_{ii}t_{m-p}^2(\alpha/2)}, \tag{4.11}$$

Uvedomme si, že hodnoty v_{ii} sa nemenia, lebo matica $\mathbf{X}_n$ je pevne daná. Definujme $\widehat{\boldsymbol{\beta}}_N$ rovnako ako v (4.6). Pripomeňme, že tento vektor má pri danom S_m^2 podmienené normálne rozdelenie $N_p[\boldsymbol{\beta}, zS_m^{-2}\sigma^2(\mathbf{X}_n^\top\mathbf{X}_n)^{-1})]$ a je na S_m^2 nezávislé. Potom analógiou vety 3.1 pre zložky vektora $\widehat{\boldsymbol{\beta}}_N$ platí:

$$T_i^N = \frac{\frac{\widehat{\beta}_i^N - \beta_i}{\sigma\sqrt{zv_{ii}}}S_m^2}{\sqrt{\frac{(m-p)S_m^2}{\sigma^2(m-p)}}} = \frac{\widehat{\beta}_i^N - \beta_i}{\sqrt{zv_{ii}}} \quad\sim\quad t_{m-p}.$$

Nech $t_{m-p}(\alpha/2)$ označuje kritickú hodnotu t rozdelenia s $m-p$ stupňami voľnosti na hladine $\alpha/2$. Potom pre každé $i = 1\ldots p$ je

$$\widehat{\beta}_i^N \mp t_{m-p}(\alpha/2)\sqrt{zv_{ii}}$$

intervalovým odhadom pre neznámy parameter β_i s koeficientom spoľahlivosti $1-\alpha$. Požadujeme, aby tento interval spoľahlivosti pre parameter β_i bol dlhý maximálne $2d$. Musí teda platiť

$$t_{m-p}(\alpha/2)\sqrt{zv_{ii}} \le d, \text{čo je ekvivalentné nerovnosti}$$
$$z \le \frac{d^2}{v_{ii}t_{m-p}^2(\alpha/2)}. \tag{4.12}$$

Z posledného vzťahu je vidieť ako sme dospeli k voľbe z pri definovaní náhodnej veličiny N v (4.11).

4.1.2 Test pre regresný parameter

Chceme skonštruovať test pre parameter β_i. Preto musíme zvoliť z tak, aby pravdepodobnosti chýb prvého aj druhého druhu boli obmedzené. Pripomeňme, že testujeme hypotézu $H_0 : \beta_i = \beta_0$ proti alternatíve $H_1 : \beta_i > \beta_0$ na hladine α. Naša testová štatistika v tomto prípade vyzerá:

$$T_i^N = \frac{\widehat{\beta}_i^N - \beta_0}{\sqrt{zv_{ii}}} \tag{4.13}$$

Za platnosti hypotézy H_0 platí $T_i^N \sim t_{m-p}$. Tým pádom našu hypotézu zamietame na hladine α, keď platí $T_i^N > t_{m-p}(\alpha)$. Ak skutočný parameter β_i poznáme, tak môžme T_i^N rozpísať nasledovne

$$T_i^N = \frac{\widehat{\beta}_i^N - \beta_i}{\sqrt{zv_{ii}}} + \frac{\beta_i - \beta_0}{\sqrt{zv_{ii}}} \tag{4.14}$$

Vieme, že $\frac{\widehat{\beta}_i^N - \beta_i}{\sqrt{zv_{ii}}}$ má t rozdelenie s $m-p$ stupňami voľnosti. Sila testu nulovej hypotézy proti jednostrannej alternatíve H_1 pre $\beta_i - \beta_0 \geq \delta, \delta > 0$ sa rovná:

$$\begin{aligned}
\mathsf{\beta}(\beta_i - \beta_0) &= \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big(T_i^N > t_{m-p}(\alpha)\Big) \\
&= \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big(\frac{\widehat{\beta}_i^N - \beta_i}{\sqrt{zv_{ii}}} + \frac{\beta_i - \beta_0}{\sqrt{zv_{ii}}} > t_{m-p}(\alpha)\Big) \\
&= \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big(\frac{\widehat{\beta}_i^N - \beta_i}{\sqrt{zv_{ii}}} > t_{m-p}(\alpha) - \frac{\beta_i - \beta_0}{\sqrt{zv_{ii}}}\Big) \\
&\geq \mathsf{P}_{\boldsymbol{\beta},\sigma}\Big(\frac{\widehat{\beta}_i^N - \beta_i}{\sqrt{zv_{ii}}} > t_{m-p}(\alpha) - \frac{\delta}{\sqrt{zv_{ii}}}\Big)
\end{aligned} \tag{4.15}$$

Vidíme, že sila testu nezávisí na σ. Chceme nájsť také z, aby platilo

$$\mathsf{P}_{\boldsymbol{\beta},\sigma}\Big(\frac{\widehat{\beta}_i^N - \beta_i}{\sqrt{zv_{ii}}} > t_{m-p}(\alpha) - \frac{\delta}{\sqrt{zv_{ii}}}\Big) = 1 - \mathsf{\beta}. \tag{4.16}$$

Vieme, že $\frac{\widehat{\beta}_i^N - \beta_i}{\sqrt{zv_{ii}}}$ má t rozdelenie s $m-p$ stupňami voľnosti. Musí teda platiť

$$\begin{aligned}
t_{m-p}(\alpha) - \frac{\delta}{\sqrt{zv_{ii}}} &= t_{m-p}(1 - \mathsf{\beta}) \text{ ekvivalentnými úpravami dostávame} \\
z &= \frac{\delta^2}{v_{ii}[t_{m-p}(\alpha) - t_{m-p}(1 - \mathsf{\beta})]^2}.
\end{aligned} \tag{4.17}$$

Z tejto voľby pre z potom platí, že $\mathsf{\beta}(\beta_i - \beta_0) \geq 1 - \mathsf{\beta}$. Tým pádom pre $\beta_i - \beta_0 \geq \delta$ je pravdepodobnosť chyby druhého druhu nanajvýš $\mathsf{\beta}$.

Na záver mojej bakalárky uvádzam preskočené dôkazy o rozdelení náhodných veličín $\mathbf{J}_N$ z (4.5) a $\widehat{\boldsymbol{\beta}}_N$ z (4.6).

Veta 4.1. *Pri danom S_m^2 má $\mathbf{J}_N$, definované v (4.5), podmienené normálne rozdelenie $N_n(\mathbf{X}_n\boldsymbol{\beta}, zS_m^{-2}\sigma^2\boldsymbol{I}_n)$ a je nezávislé na S_m^2.*

Dôkaz.

$$
\begin{aligned}
\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(\mathbf{J}_N|S_m^2\big) &= a_0\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(\overline{\mathbf{Y}}_m|S_m^2\big) + \sum_{k=m+1}^{N} a_k\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(\mathbf{Y}_k|S_m^2\big)\\
&= a_0\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(\overline{\mathbf{Y}}_m\big) + \sum_{k=m+1}^{N} a_k\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(\mathbf{Y}_k\big)\\
&= a_0\mathbf{X}_n\boldsymbol{\beta} + \sum_{k=m+1}^{N} a_k\mathbf{X}_n\boldsymbol{\beta}\\
&= \mathbf{X}_n\boldsymbol{\beta}\Big[a_0 + \sum_{k=m+1}^{N} a_k\Big] \overset{(4.4)}{=} \mathbf{X}_n\boldsymbol{\beta};\\
\mathsf{Var}_{\boldsymbol{\beta},\sigma}(\mathbf{J}_N|S_m^2) &= a_0^2\mathsf{Var}_{\boldsymbol{\beta},\sigma}\overline{\mathbf{Y}}_m + \sum_{k=m+1}^{N} a_k^2\mathsf{Var}_{\boldsymbol{\beta},\sigma}\mathbf{Y}_k\\
&= a_0^2\sigma^2\boldsymbol{I}_n/m + \sum_{k=m+1}^{N} a_k^2\sigma^2\boldsymbol{I}_n\\
&= \sigma^2\boldsymbol{I}_n(a_0^2/m + \sum_{k=m+1}^{N} a_k^2)\\
&\overset{(4.4)}{=} zS_m^{-2}\sigma^2\boldsymbol{I}_n
\end{aligned}
$$

□

Veta 4.2. *Pri danom S_m^2 má $\widehat{\boldsymbol{\beta}}_N$, definované v (4.6), podmienené normálne rozdelenie $N_p[\boldsymbol{\beta}, zS_m^{-2}\sigma^2(\mathbf{X}_n^\top\mathbf{X}_n)^{-1})]$ a je nezávislé na S_m^2.*

Dôkaz.

$$
\begin{aligned}
\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(\widehat{\boldsymbol{\beta}}_N|S_m^2\big) &= (\mathbf{X}_n^\top\mathbf{X}_n)^{-1}\mathbf{X}_n^\top\mathsf{E}_{\boldsymbol{\beta},\sigma}\big(\mathbf{J}_N|S_m^2\big)\\
&\overset{(4.1)}{=} (\mathbf{X}_n^\top\mathbf{X}_n)^{-1}\mathbf{X}_n^\top\mathbf{X}_n\boldsymbol{\beta} = \boldsymbol{\beta}
\end{aligned}
$$

$$
\begin{aligned}
\mathsf{Var}_{\boldsymbol{\beta},\sigma}(\widehat{\boldsymbol{\beta}}_N|S_m^2) &= (\mathbf{X}_n^\top\mathbf{X}_n)^{-1}\mathbf{X}_n^\top\mathsf{Var}_{\boldsymbol{\beta},\sigma}(\mathbf{J}_N|S_m^2)[(\mathbf{X}_n^\top\mathbf{X}_n)^{-1}\mathbf{X}_n^\top]^\top\\
&\overset{(4.1)}{=} (\mathbf{X}_n^\top\mathbf{X}_n)^{-1}\mathbf{X}_n^\top zS_m^{-2}\sigma^2\boldsymbol{I}_n\mathbf{X}_n[(\mathbf{X}_n^\top\mathbf{X}_n)^{-1}]^\top\\
&= (\mathbf{X}_n^\top\mathbf{X}_n)^{-1}\mathbf{X}_n^\top\mathbf{X}_n zS_m^{-2}\sigma^2(\mathbf{X}_n^\top\mathbf{X}_n)^{-1}\\
&= zS_m^{-2}\sigma^2(\mathbf{X}_n^\top\mathbf{X}_n)^{-1}.
\end{aligned}
$$

□

Záver

V tejto bakalárskej práci sme sa oboznámili so sekvenčnými postupmi, ktoré sa využívajú v lineárnom regresnom modeli. Ako bolo už naznačené v sekcii 2.1 tieto sekvenčné metódy sa vyvinuli so Steinovej dvojstupňovej metódy, ktorý v roku 1945 navrhol sekvenčný postup pre test strednej hodnoty v normálnom rozdelení. Na základe tohto postupu sa začali rozširovať štúdie o sekvenčných postupoch v rôznych oblastiach a modeloch v štatistike. Jedným z nich bol aj lineárny regresný model.

Základným sekvenčným postupom v lineárnom modeli bola dvojstupňová metóda 2.1, ktorú pán Healy v roku 1956 ako prvý publikoval. Kvôli asymptotickej neúčinnosti tejto metódy viedlo štatistikov k objavovaniu ďalších sekvenčných metód. V tejto práci sme si uviedli ešte dve metódy a to: modifikovanú dvojstupňovú metódu 2.2 a priamu sekvenčnú metódu 2.3. Tieto metódy už mali o niečo lepšie vlastnosti, ale stále neboli ideálne. Preto vznikol rad nových sekvenčných metód ako napríklad zrýchlená sekvenčná metóda (angl. Accelerated sequential procedure) alebo trojstupňová metóda (angl. Three-stage procedure). Základné informácie o zrýchlenej a trojstupňovej metóde sú pekne spísané v knihe (Mukhopadhyay a Silva, 2009, 6. kapitola).

V bakalárskej práci sme pracovali s lineárnym modelom, v ktorom sme tradične požadovali od vektora chýb, aby mal normálne rozdelenie. Vďaka tomuto predpokladu sme mohli využívať vynikajúcu vlastnosť a to nezávislosť odhadnutých regresných parametrov na reziduálnom rozptyle.

Ak by sme namiesto normálneho rozdelenia predpokladali o vektore chýb, že je z rozdelenia s neznámou distribučnou funkciou $\mathcal{F}$ s takou, že $0 < \sigma^2 = \sigma^2(\mathcal{F}) < \infty$, tak už by to bolo o dosť komplikovanejšie. Dnes už existujú dostatočne účinné sekvenčné metódy aj pre takéto všeobecnejšie rozdelenie v lineárnom modeli.

Literatúra

Anděl, J. (2007). *Statistické metody.* Čtvrté upravené vydání. Matfyzpress, Praha. ISBN 80-7378-003-8.

Ghosh, B. K. a Sen, P. K. (1991). *Handbook of Sequential Analysis.* Marcel Dekker, Inc., New York, Basel, Hong Kong. ISBN 0-8247-8408-1.

Mukhopadhyay, N. a Silva, B. M. d. (2009). *Sequential Methods and Their Applications.* CRC Press, Boca Raton. ISBN 978-1-58488-102-5.

Mukhopadhyay, N., Datta, S. a Chattopadhyay, S. (2004). *Applied Sequential Methodologies.* Marcel Dekker, Inc., New York, Basel. ISBN 0-8247-5395-X.

Rusá, Š. (2013). *Dvoustupňové statistické postupy.* Praha, Univerzita Karlova v Praze.

Zvára, K. (2008). *Regrese.* Matfyzpress, Praha. ISBN 978-80-7378-041-8.

Zoznam obrázkov

Zoznam tabuliek

Zoznam použitých skratiek

$\boldsymbol{I}_n, \boldsymbol{I}_{n\times n}$	n-rozmerná jednotková štvorcová matica
$\mathbf{A}^\top$	transponovaná matica $\mathbf{A}$
χ^2_n	χ^2 rozdelenie s n stupňami voľnosti
t_n	t rozdelenie s n stupňami voľnosti
$F_{m,n}$	rozdelenie F o m a n stupňoch voľnosti
tzv.	takzvaný
t. j.	to jest (to znamená)
$E(X)$	stredná hodnota náhodnej veličiny X
$\mathsf{P}_{\boldsymbol{\beta},\sigma}(\cdot)$	$\mathsf{P}_{\boldsymbol{\beta},\sigma}(\,\cdot \mid \boldsymbol{\beta},\sigma)$
$\mathsf{E}_{\boldsymbol{\beta},\sigma}(\cdot)$	$\mathsf{E}_{\boldsymbol{\beta},\sigma}(\,\cdot \mid \boldsymbol{\beta},\sigma)$
$\mathsf{Var}_{\boldsymbol{\beta},\sigma}(\cdot)$	$\mathsf{Var}_{\boldsymbol{\beta},\sigma}(\,\cdot \mid \boldsymbol{\beta},\sigma)$
$I[A]$	indikátor množiny A
$\lfloor a \rfloor$	najväčšie celé číslo $\leq a$
$N(\mu,\sigma)$	normálne rozdelenie so strednou hodnotou μ a rozptylom σ^2