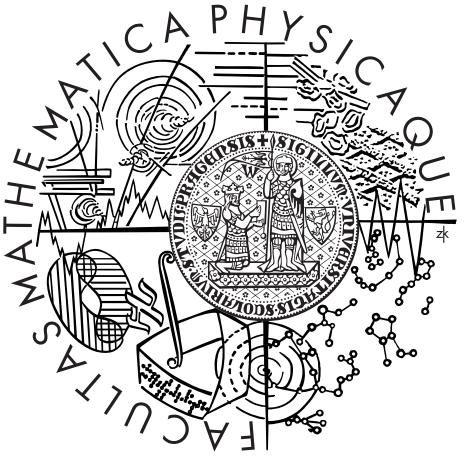
Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Ksenia Molodkina

Krokové metody v lineární regresi a jejich vlastnosti

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Milan Bašta, Ph.D.

Studijní program: Matematika

Studijní obor: Finanční matematika

Praha 2014

Na tomto místě bych velmi rada poděkovala Mgr. Milanovi Baštovi, Ph.D. za cenné připomínky a čas, věnovaný konzultacím.

Prohlašuji, že jsem tuto bakalářskou práci vypracovala 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: Krokové metody v lineární regresi a jejich vlastnosti

Autor: Ksenia Molodkina

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Milan Bašta, Ph.D., Katedra statistiky a pravděpodobnosti Fakulty informatiky a statistiky Vysoké školy ekonomické v Praze

Abstrakt: Předložená bakalářská práce se zabývá krokovými metodami v lineární regresi a jejich vlastnostmi. Krokové metody jsou jedním z nástrojů výběru vysvětlujících proměnných v rámci výstavby modelů lineární regrese. Teoretická část práce shrnuje teorii lineárních regresních modelů, popisuje odhadování parametrů těchto modelů metodou nejmenších čtverců a uvádí vlastnosti těchto odhadů. Následně jsou definovány krokové metody lineární regrese, konkrétně backward selection (metoda sestupného výběru), forward selection (metoda vzestupného výběru) a stepwise regression (hybridní metoda). Praktická část zahrnuje simulační studie, pomocí kterých budou ilustrovány vybrané vlastnosti metod krokové regrese.

Klíčová slova: backward selection, forward selection, stepwise regression

Title: Stepwise methods in linear regression and their properties

Author: Ksenia Molodkina

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Milan Bašta, Ph.D., Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague

Abstract: This thesis is concerned with methods of stepwise linear regression and their properties. Methods of the stepwise regression are techniques of building a linear regression model. The theoretical part of the thesis summarizes the theory of linear regression models, describes the estimation of the parameters using the method of least squares and represents the properties of these estimates. Thereafter it defines stepwise linear regression, specifically backward elimination, forward selection, and stepwise regression by itself. The practical part of the thesis includes simulation studies, which will illustrate selected properties of the stepwise linear regression.

Keywords: backward selection, forward selection, stepwise regression

Obsah

1 Teoretická část **7**

1.1 Pojem regresní analýza 7
1.2 Metoda nejmenších čtverců 8
1.3 Normální lineární model 12
1.4 Testy hypotéz a intervaly spolehlivosti 12
1.4.1 Dílčí t-testy 12
1.5 Metody krokové regrese 13
1.6 Aplikace metod krokové regrese 15
1.6.1 Backward selection (sestupný výběr) 15
1.6.2 Forward selection (vzestupný výběr) 16
1.6.3 Stepwise regression (kroková regrese) 17

2 Praktická část **20**

2.1 Popis dat 20
2.2 Postup zpracování dat 21
2.2.1 Implementace metody backward selection 22
2.2.2 Implementace metody forward selection 22
2.2.3 Implementace metody stepwise regression 22
2.2.4 Monte Carlo simulace 22
2.3 Výsledky simulačních studií 23
2.3.1 Vyhodnocení vychýlení 23
2.3.2 Vyhodnocení relativní četnosti detekce správné konfigurace 27
2.3.3 Vyhodnocení chyby predikce 34
2.3.4 Souhrn výsledků 36

Úvod

Lineární regrese je důležitá statistická metoda ke zkoumání závislosti proměnných. Cílem této bakalářské práce je seznamit čtenáře s nástroji výběru nezávisle proměnných v rámci výstavby modelů lineární regrese. Konkrétně se v rámci této práce soustředíme na metody krokové regrese.

Práce je rozdělena na část teoretickou a část praktickou. První část práce shrnuje teorii lineární regrese. Je zaměřena na definování lineárního regresního modelu, odhadování parametrů lineárního modelu metodou nejmenších čtverců, důležité vlastnosti těchto odhadů, s nimi související testy hypotéz o hodnotách parametrů regresní funkce. Následuje úvod do krokové regrese, která je jedním ze způsobů, jak nalézt množinu regresorů do lineárního regresního modelu. Jsou definovány tři metody krokové regrese (forward selection, backward selection a stepwise regression) a taky princip jejich fungování.

Praktická část práce je věnována vlastnostem metod krokové regrese a jejich použitelnosti při budování lineárních modelů. Součástí této časti jsou Monte Carlo simulace, které jsou prováděny ve statistickém softwaru R. Na základě výsledků, získaných ze simulačních studií, se pak hodnotí některé vybrané vlastnosti krokových metod. Výsledky jsou prezentovány numericky a graficky.

1 Teoretická část

1.1 Pojem regresní analýza

Lineární regrese se používá ve statistice ke zkoumání vztahu mezi spojitou veličinou Y a vektorem $\boldsymbol{X}$, který může obsahovat jednu nebo více spojitých nebo diskrétních veličin. Předpokládáme, že hodnota vektoru $\boldsymbol{X}$ může ovlivňovat střední hodnotu Y. Úkolem regresní analýzy je prozkoumat závislost střední hodnoty Y na komponentách $\boldsymbol{X}$ a co nejlépe vyjádřit charakter této závislosti. Pokud by komponenty $\boldsymbol{X}$ byly náhodnými veličinami, můžeme se zabývat podmíněným rozdělením náhodné veličiny Y při pevné hodnotě $\boldsymbol{x}$ náhodného vektoru $\boldsymbol{X}$. Střední hodnota $\mathsf{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ se v tom případě nazývá *regresní funkce*.

Jako příklady regresních závislostí uveďme:

1. Závislost výšky člověka na jeho váze, věku, pohlaví.
2. Závislost porodní hmotnosti dítěte na době těhotenství, pořadí narození dítěte a věku matky.
3. Závislost brzdné dráhy automobilu na jeho rychlosti, hmotnosti, hloubce vzorku pneumatiky.

Data sestávají z n nezávislých pozorování vektorů $(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n$, kde každé $\boldsymbol{X}_i$ má $p < n$ složek $(X_{i1}, \ldots, X_{ip})$.

Základními pojmy regresní analýzy jsou regresní model a regresní funkce zahrnující následující proměnné:

- Náhodnou veličinu Y_i, která se nazývá *odezva*. Alternativní název: *závisle proměnná, regresand*
- Komponenty vektoru $\boldsymbol{X}_i$, které nazýváme *regresory*. Alternativní název: *nezávisle proměnná*.
- Nenáhodné veličiny β_i. Nazýváme je *regresní parametry*.

Definice 1 (Lineární regresní model)**.** *Řekneme, že data $(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n$, splňují lineární regresní model, střední hodnoty nekorelovaných náhodných veličin $(Y_1, \ldots, Y_n)$ lze popsat takto*

$$\mathsf{E}Y_i = \sum_{k=1}^{p} \beta_k X_{ik} = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \tag{1}$$

Nezávisle proměnné $X_{ik}, k = 1, \ldots, p$ jsou nenáhodné veličiny a jejich hodnoty jsou přesně známé. Regresní koeficienty $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ jsou neznámé konstanty.

Poznámka 1. V regresní analýze většinou volíme $X_{i1} = 1$ pro všechna $i = 1, \ldots, n$. Parametr β_1 pak nazýváme absolutní člen.

Lineární regresní model můžeme přepsat pomocí rovnic

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \tag{2}$$

kde ε_i je náhodná složka lineárního modelu.

Lineární regresní model lze rovněž přepsat do maticového tvaru

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3}$$

kde $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\mathsf{T}, \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ a

$$\mathbb{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix},$$

přičemž $\boldsymbol{\beta} \in \mathbb{R}^p$ a $r(\mathbb{X}) = p < n$. [1] Požaduje se, aby matice $\mathbb{X}$ měla lineárně nezávislé sloupce.

Vektor $\mathbb{X}\boldsymbol{\beta}$ je nenáhodný. Předpokládá se, že $\mathsf{E}\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}$ a $\mathsf{Var}\boldsymbol{Y} = \sigma^2\mathbb{I}$, kde σ^2 je neznámý parametr a $\mathbb{I}$ je jednotková matice n-tého řádu.

Rovnosti výše kladou jisté nároky na chování vektoru $\boldsymbol{\varepsilon}$. Tyto nároky se dají rovněž považovat za předpoklady platnosti lineárního regresního modelu:

1. Regresní parametry β_k mohou nabývat libovolných hodnot pro $k = 1, \ldots, p$.
2. Regresní matice $\mathbb{X}$ má hodnost $r(\mathbb{X}) = p$ a $n > p$. Tato podmínka vyžaduje, aby mezi vyskytujícími se nezávisle proměnnými nebyla funkční lineární závislost, tedy v matici $\mathbb{X}$ nesmí existovat lineárně závislé sloupce. Matice $\mathbb{X}^\mathsf{T}\mathbb{X}$ je potom regulární.
3. Náhodná složka má nulovou střední hodnotu $\mathsf{E}\boldsymbol{\varepsilon} = \boldsymbol{0}$. Tato podmínka znamená, že náhodná složka nepůsobí systematickým způsobem na hodnoty zá vislé veličiny $\boldsymbol{Y}$.
4. Pro varianční matici $\boldsymbol{\varepsilon}$ platí $\mathsf{Var}\boldsymbol{\varepsilon} = \sigma^2\mathbb{I}$ kde σ^2 je neznámý, ale pevný parametr.

1.2 Metoda nejmenších čtverců

K odhadu neznámých parametrů lineárního regresního modelu se nejčastěji používá metoda nejmenších čtverců. Tento postup vychází z minimalizace součtů druhých mocnin vertikální vzdálenosti vysvětlované proměnné Y_i od regresní nadroviny. Metoda se snaží o co nejlepší proložení nadroviny množinou dat $(Y_i, \boldsymbol{X}_i)$.

Nechť je dán lineární regresní model s daty $(Y_i, \boldsymbol{X}_i)$. Odhadujeme regresní parametry $\beta_1, \ldots, \beta_p$ metodou nejmenších čtverců z podmínky minimalizace výrazu

[1] $r(\mathbb{X})$ značí hodnost matice $\mathbb{X}$.

$$\sum_{i=1}^{n}\left[Y_i - \sum_{k=1}^{p} X_{ik}\beta_k\right]^2,$$

maticově ten výraz můžeme přepsat jako $(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})$. Odhad $\boldsymbol{\beta}$ lze najít položením první derivace výrazu podle $\boldsymbol{\beta}$ rovné nule. Označíme odhad jako $\widehat{\boldsymbol{\beta}}$.

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) = \tag{4}$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}}[\boldsymbol{Y}^\mathsf{T}\boldsymbol{Y} - \boldsymbol{Y}^\mathsf{T}\mathbb{X}\boldsymbol{\beta} - (\mathbb{X}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{Y} + (\mathbb{X}\boldsymbol{\beta})^\mathsf{T}\mathbb{X}\boldsymbol{\beta}] = \tag{5}$$

$$= -\mathbb{X}^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) - [(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})^\mathsf{T}\mathbb{X}]^\mathsf{T} = \tag{6}$$

$$= -2\mathbb{X}^\mathsf{T}\boldsymbol{Y} + 2\mathbb{X}^\mathsf{T}\mathbb{X}\boldsymbol{\beta} = \tag{7}$$

$$= \mathbb{X}^\mathsf{T}\boldsymbol{Y} - \mathbb{X}^\mathsf{T}\mathbb{X}\widehat{\boldsymbol{\beta}} = 0 \tag{8}$$

$$\mathbb{X}^\mathsf{T}\boldsymbol{Y} = \mathbb{X}^\mathsf{T}\mathbb{X}\widehat{\boldsymbol{\beta}} \tag{9}$$

Tímto krokem obdržíme soustavu p lineárních rovnic o p neznámých, kterou řeší právě odhadovaný regresní parametr $\widehat{\boldsymbol{\beta}}$.

Jelikož matice $\mathbb{X}$ má plnou hodnost p, matice $\mathbb{X}^\mathsf{T}\mathbb{X}$ je regulární (čtvercová, $p \times p$), s hodností p. Z toho plyne existence právě jedno řešení soustavy (7)

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y}. \tag{10}$$

$\widehat{\boldsymbol{\beta}}$ musí být globálním minimem, protože funkce $(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})$ je konvexní v $\widehat{\boldsymbol{\beta}}$.

Věta 1. *Odhady metodou nejmenších čtverců jsou* $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y}$.

Důkaz. Ze vztahu (6) vyplývá, že $\mathbb{X}^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) = 0$.

Dostáváme

$$(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) = \tag{11}$$

$$= [(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) + (\mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\boldsymbol{\beta})]^\mathsf{T}[(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) + (\mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\boldsymbol{\beta})] = \tag{12}$$

$$= (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\mathsf{T}\mathbb{X}^\mathsf{T}\mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{13}$$

Platí

$$(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\mathsf{T}\mathbb{X}^\mathsf{T}\mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \geq (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}). \tag{14}$$

Víme, že matice $\mathbb{X}^\mathsf{T}\mathbb{X}$ je regulární, takže je pozitivně definitní. To znamená, že z nerovnosti (11) se stává rovnost, právě když $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. □

Definice 2 (Vektor vyrovnaných hodnot)**.** *Vektor* $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y}$ *se nazývá vektor vyrovnaných hodnot. Po složkách zapsáno*

$$\widehat{Y}_i = \boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}} = \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + \cdots + \widehat{\beta}_p X_{ip}. \tag{15}$$

Vektor $\widehat{\boldsymbol{Y}}$ může být považován za nejlepší aproximaci vektoru $\boldsymbol{Y}$ vzniklou pomocí lineárních kombinací sloupců matice $\mathbb{X}$.

Definice 3 (Projekční matice metody nejmenších čtverců)**.** *Výraz $\mathbb{H} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$ nadále budeme nazývat projekční maticí, lze psát $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$.*

Poznámka 2. Projekční matice $\mathbb{H}$ je také idempotentní[2], symetrická a její hodnost je rovna $r(\mathbb{H}) = r(\mathbb{X}) = p < n$.

Definice 4 (Vektor reziduí)**.** *Odchýlení pozorované hodnoty veličiny $\boldsymbol{Y}$ od vektoru vyrovnaných hodnot $\widehat{\boldsymbol{Y}}$, tj.*

$$\boldsymbol{u} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} \tag{16}$$

nazveme vektorem reziduí.

Definice 5 (Reziduální součet čtverců)**.** *Čtverec rozdílu $\boldsymbol{Y}$ a $\widehat{\boldsymbol{Y}}$, tj.*

$$S_e = \boldsymbol{u}^\mathsf{T}\boldsymbol{u} = \sum_{i=1}^{n} \boldsymbol{u}_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 \tag{17}$$

se nazve reziduálním součtem čtverců.

Definice 6 (Reziduální rozptyl)**.** *Reziduální rozptyl zavedeme jako $s^2 = \frac{S_e}{n-p}$.*

Definice 7 (Nestranný odhad)**.** *Odhad $\widehat{\boldsymbol{\beta}}$ se nazývá nestranný (nevychýlený), jestliže jeho střední hodnota je rovna hodnotě odhadovaného parametru $\mathsf{E}\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.*

Definice 8 (Vychýlení)**.** *Rozdíl $\mathsf{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ nazýváme vychýlením odhadu $\widehat{\boldsymbol{\beta}}$.*

Poznámka 3. Vychýlení odhadu $\widehat{\boldsymbol{\beta}}$ nadále budeme označovat jako $\mathsf{Bias}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$.

Věta 2. *V lineárním regresním modelu platí:*

(i) Odhad $\widehat{\boldsymbol{\beta}}$ dle metody nejmenších čtverců je nestranným odhadem regresního parametru $\boldsymbol{\beta}$.

(ii) Variance regresního parametru jest $\mathsf{Var}\widehat{\boldsymbol{\beta}} = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$.

Důkaz. Ad (i)

$$\mathsf{E}\widehat{\boldsymbol{\beta}} = \mathsf{E}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathsf{E}\boldsymbol{Y} = \tag{18}$$

$$= (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\beta}. \tag{19}$$

[2]Pro idempotentní matici platí $\mathbb{H}\mathbb{H} = \mathbb{H}$.

Ad (ii)

$$\begin{aligned} \mathsf{Var}\widehat{\boldsymbol{\beta}} &= \left(\mathbb{X}^{\mathsf{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathsf{T}}\left(\mathsf{Var}\boldsymbol{Y}\right)\mathbb{X}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1} = & (20)\\ &= (\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}(\sigma^2\mathbb{I})\mathbb{X}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1} = (\sigma^2)(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}. & (21)\end{aligned}$$

□

Poznámka 4. První část věty tvrdí, že odhad $\widehat{\boldsymbol{\beta}}$ není systematicky podhodnocen ani nadhodnocen.

Definice 9 (Nejlepší nestranný odhad)**.** *Odhad vektoru $\boldsymbol{Y}$ se nazve nejlepším nestranným lineárním odhadem vektoru $\boldsymbol{Y}$, pokud pro jakýkoli jiný nestranný lineární odhad $\widetilde{\boldsymbol{Y}}$ veličiny $\boldsymbol{Y}$ platí*

$$\mathsf{Var}\widetilde{\boldsymbol{Y}} - \mathsf{Var}\widehat{\boldsymbol{Y}} \geqq 0, \tag{22}$$

tj. rozdíl uvedených dvou variančních matic musí být matice pozitivně semidefinitní.

Poznámka 5. Jinými slovy, nejlepší nestranný lineární odhad je takový, který má minimální rozptyl ze všech nestranných lineárních odhadů.

Věta 3 (Gauss-Markov)**.** *Nechť je dán lineární model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Potom $\widehat{\boldsymbol{Y}}$ je nejlepším nestranným lineárním odhadem (NNLO) střední hodnoty $\boldsymbol{Y}$ a zároveň platí $\mathsf{Var}\widehat{\boldsymbol{Y}} = \sigma^2\mathbb{H}$.*

Důkaz. (i) Odhad $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$ je lineární.

(ii) Nestrannost odhadu $\widehat{\boldsymbol{Y}}$ plyne z následující řady rovností

$$\mathsf{E}\widehat{\boldsymbol{Y}} = \mathsf{E}\mathbb{H}\boldsymbol{Y} = \mathbb{H}\mathsf{E}\boldsymbol{Y} = \mathbb{H}\mathbb{X}\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta} = \boldsymbol{Y}. \tag{23}$$

Zaveďme nyní jiný nestranný lineární odhad střední hodnoty $\boldsymbol{Y}$ v obecnějším tvaru $\widetilde{\boldsymbol{Y}} = \mathbb{B}\boldsymbol{Y} + \boldsymbol{a}$, kde $\mathbb{B}$ je matice parametrů a $\boldsymbol{a}$ je konstantní vektor. Ukažme, že pokud má být odhad $\widetilde{\boldsymbol{Y}}$ nestranný, pak pro zmíněné parametry musí platit $\boldsymbol{a} = \boldsymbol{0}$ a $\mathbb{B}\mathbb{X} = \mathbb{X}$.

O střední hodnotě závislé veličiny $\boldsymbol{Y}$ víme, že $\mathsf{E}\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}$, neboť dle definice lineárního regresního modelu $\boldsymbol{\varepsilon} = \boldsymbol{0}$. Aby byl odhad $\widetilde{\boldsymbol{Y}}$ nestranný, musí střední hodnota $\widetilde{\boldsymbol{Y}}$ být shodná se střední hodnotou $\boldsymbol{Y}$, tedy $\mathsf{E}\widetilde{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{\beta}$. Jiný výraz pro střední hodnotu $\widetilde{\boldsymbol{Y}}$ získáme přímočarou úpravou

$$\mathsf{E}\widetilde{\boldsymbol{Y}} = \mathsf{E}\left(\mathbb{B}\boldsymbol{Y} + \boldsymbol{a}\right) = \mathbb{B}\mathsf{E}\boldsymbol{Y} + \boldsymbol{a} = \mathbb{B}\mathbb{X}\boldsymbol{\beta} + \boldsymbol{a}. \tag{24}$$

Mají-li se rovnat oba výše zmíněné výsledky pro všechna $\boldsymbol{\beta}$, dojdeme k závěru, že vektor $\boldsymbol{a}$ musí být nulový a $\mathbb{B}$ splňuje $\mathbb{B}\mathbb{X} = \mathbb{X}$.

Z identity $\mathbb{B}\mathbb{X} = \mathbb{X}$ postupným nasobením zprava maticemi $\left(\mathbb{X}^{\mathsf{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathsf{T}}$ a $\mathbb{X}$, získáme ekvivalentní rovnici $\mathbb{B}\mathbb{H} = \mathbb{H}$, kde místo matice $\mathbb{X}$ vystupuje matice $\mathbb{H}$.

Spočtěme ještě varianční matici $\widetilde{\boldsymbol{Y}}$

$$\begin{aligned} \mathsf{Var}\widetilde{\boldsymbol{Y}} &= \mathbb{B}(\mathsf{Var}\boldsymbol{Y})\mathbb{B}^{\mathsf{T}} = \mathbb{B}(\sigma^2\mathbb{I})\mathbb{B}^{\mathsf{T}} = \sigma^2\mathbb{B}\mathbb{B}^{\mathsf{T}} \\ &= \sigma^2\left[\mathbb{H} + (\mathbb{B}-\mathbb{H})\right]\left[\mathbb{H} + (\mathbb{B}-\mathbb{H})\right]^{\mathsf{T}} = \sigma^2\mathbb{H} + \sigma^2\left(\mathbb{B}-\mathbb{H}\right)\left(\mathbb{B}-\mathbb{H}\right)^{\mathsf{T}}, \end{aligned} \tag{25}$$

kde bylo využito $\mathbb{B}\mathbb{H} = \mathbb{H}$.

Výsledek srovnáme s variancí odhadu $\widehat{\boldsymbol{Y}}$

$$\mathsf{Var}\widehat{\boldsymbol{Y}} = \mathsf{Var}\,\mathbb{H}\boldsymbol{Y} = \sigma^2\mathbb{H}, \tag{26}$$

a získáme $\mathsf{Var}\widetilde{\boldsymbol{Y}} - \mathsf{Var}\widehat{\boldsymbol{Y}} = \sigma^2\left(\mathbb{B}-\mathbb{H}\right)\left(\mathbb{B}-\mathbb{H}\right)^T \geq 0$. Jinak řečeno, rozdíl variancí obou odhadů je pozitivně semidefinitní matice, což potvrzuje platnost Gauss-Markovovy věty, že $\widehat{\boldsymbol{Y}}$ je nejlepším možným lineárním odhadem vektoru $\boldsymbol{Y}$. □

1.3 Normální lineární model

Doposud jsme pracovali s modelem, který vymezoval pouze střední hodnotu $\mathsf{E}\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}$ a variační matici $\mathsf{Var}\boldsymbol{Y} = \sigma^2\mathbb{I}$ vektoru $\boldsymbol{Y}$. Uvedené předpoklady můžeme stručně zapsat jako $\boldsymbol{Y} \sim (\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbb{I})$. Dále budeme předpokládat *normální lineární model*, kde náhodný vektor $\boldsymbol{Y}$ má mnohorozměrné normální rozdělení s parametry $\mathsf{E}\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}$ a $\mathsf{Var}\boldsymbol{Y} = \sigma^2\mathbb{I}$, respektive $\boldsymbol{Y} \sim N(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbb{I})$. To znamená, že platí $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I})$. Tento předpoklad nám umožní určit rozdělení jednotlivých statistik a ukázat jejich užitečné vlastnosti.

Věta 4. *Necht' $\boldsymbol{Y} \sim N(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbb{I})$. Potom platí $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1})$.*

Důkaz. Důkaz lze najít v knize (Anděl, 2007, str. 83) □

Věta 5. *Necht' $\boldsymbol{Y} \sim N(\mathbb{X}\beta, \sigma^2\mathbb{I})$. Potom:*

(i) *Platí $\frac{S_e}{\sigma^2} \sim {\chi^2}_{(n-p)}$*

(ii) *Platí $\widehat{\boldsymbol{Y}} \sim N(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbb{H})$.*

(iii) *Platí $\boldsymbol{u} \sim N(\mathbf{0}, \sigma^2(\mathbb{I}-\mathbb{H}))$*

(iv) *Náhodné vektory $\widehat{\boldsymbol{\beta}}$ a $\boldsymbol{u}$ jsou nezávislé.*

Důkaz. Důkaz lze najít v knize (Zvára, 1989, str. 62) □

1.4 Testy hypotéz a intervaly spolehlivosti

1.4.1 Dílčí t-testy

Věta 6. *Označme v_{ij} prvky matice $\left(\mathbb{X}^{\mathsf{T}}\mathbb{X}\right)^{-1}$. Necht' $T_i = \frac{\widehat{\beta}_i - \beta_i}{\sqrt{s^2 v_{ii}}}$.*

Pak pro každé $i = 1, \ldots, p$ platí $T_i \sim t_{n-p}$.

Důkaz. Z věty 4 plyne, že $\frac{\widehat{\beta}_i-\beta_i}{\sqrt{\sigma^2 v_{ii}}} \sim N(0,1)$. Za platností bodů (i) a (iv) z věty 5 má veličina

$$\frac{\frac{\widehat{\beta}_i-\beta_i}{\sqrt{\sigma^2 v_{ii}}}\sqrt{n-p}}{\sqrt{\frac{S_e}{\sigma^2}}} = \frac{\widehat{\beta}_i-\beta_i}{\sqrt{s^2 v_{ii}}} = T_i \tag{27}$$

Studentovo rozdělení t_{n-p}. □

Větu 7 můžeme použit ke stanovení intervalu spolehlivosti pro parametr β_i a testování hypotéz o tomto parametru.

Na hladině významnosti α testujeme nulovou hypotézu ve tvaru

$$H_0 : \beta_i = \beta_i^0 \tag{28}$$

proti alternativní hypotéze

$$H_1 : \beta_i \neq \beta_i^0. \tag{29}$$

Použijeme testovou statistiku

$$T_i = \frac{\widehat{\beta}_i - \beta_i^0}{\sqrt{s^2 v_{ii}}}, \tag{30}$$

která má za platností $H_0 : \beta_i = \beta_i^0$ rozdělení t_{n-p}.

Kritický obor

H_0 budeme zamítat na hladině významnosti α ve prospěch H_1 právě tehdy, když

$$|T_i| \geq t_{n-p}(1-\alpha/2), \tag{31}$$

kde $t_{n-p}(1-\alpha/2)$ značí $(1-\alpha/2)$-tý kvantil Studentova t-rozdělení o $n-p$ stupních volnosti.

Nejčastějším případem je $\beta_i^0 = 0$. Ověřujeme, zda $\boldsymbol{Y}$ vůbec závisí na i-tém sloupci matice $\mathbb{X}$. Pokud se ukáže, že pro konkrétní i nelze zamítnout nulovou hypotézu, je třeba zvážit setrvání příslušné nezávisle proměnné v modelu.

Oboustranný interval spolehlivosti pro β_i na hladině $1-\alpha$ by vyšel

$$P\left[(\widehat{\beta}_i - t_{n-p}(1-\alpha/2)s\sqrt{v_{ii}} < \beta_i < \widehat{\beta}_i + t_{n-p}(1-\alpha/2)s\sqrt{v_{ii}})\right] = 1-\alpha. \tag{32}$$

1.5 Metody krokové regrese

Poměrně často se setkáváme s lineárními regresními modely $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ s velkým počtem regresorů v matici $\mathbb{X}$. Velký počet regresorů má za následek, že pro některé para-

metry β_i, $i = 1, \ldots, p$ nelze zamítnout $H_0 : \beta_i = 0$. To znamená, že vypuštěním některých z regresorů (i-tých sloupců matice $\mathbb{X}$) můžeme přejít k jednoduššímu modelu.

V následující části práce uvedeme postupy, které lze použít při hledání vhodného lineárního modelu, a některé z jejich charakteristik.

Jedním ze způsobů, jak nalézt optimální podmnožiny nezávisle proměnných do lineárního regresního modelu, jsou metody tzv. postupného výběru, jejichž reprezentanty jsou:

(i) backward selection (sestupný výběr)

(ii) forward selection (vzestupný výběr)

(iii) stepwise regression (kroková regrese).

Backward selection (sestupný výběr)

Nejprve se spočítá úplný model, do kterého jsou zařazeny všechny potenciální nezávisle proměnné. Postupně v každém kroku se pak z modelu vyřazuje nezávisle proměnná, která v daném modelu nejméně přispívá k vysvětlení závisle proměnné. Pro každou proměnnou testujeme hypotézu, zda v tomto modelu je regresní parametr u dané proměnné nulový. K rozhodování se používá t-statistika, vyřazuje se vždy proměnná s nejmenší absolutní hodnotou t-statistiky. Končí se tehdy, když všechny t-statistiky pro vyloučení jsou v absolutní hodnotě větší, než předem zvolené číslo t^{**}.

Forward selection (vzestupný výběr)

Jde o opak předchozího postupu. Vycházíme z prázdné množiny nezávisle proměnných a postupně v každém kroku se do ní přídá taková nezávisle proměnná, jejíž příspěvek k vysvětlení závisle proměnné je největší. U proměnné, kterou vložíme do modelu, testujeme hypotézu, zda v tomto modelu je regresní parametr u dané proměnné nulový. Proměnná, která je zařazena do modelu, v něm trvale zůstává. V daném kroku vložíme takovou proměnnou, u níž je absolutní hodnota t-statistiky největší. Skončíme, když hodnoty všech t-statistik pro zařazení jsou v absolutní hodnotě menší, než nějaké předem zvolené číslo t^*.

Stepwise regression (kroková regrese)

Kroková regrese kombinuje oba právě popsané postupy. Vzestupný výběr je v každém kroku kombinován s pokusem o zjednodušení pomocí sestupného výběru.
Je důležité, aby v průběhu výběru nedošlo k zacyklení algoritmu, kdy bude právě vložená proměnná okamžitě vyloučena, poté znovu vložena, vyloučena atd. Proto musí platit $t^* > t^{**}$.

K rozhodování, zda i-tá nezávisle proměnná bude přítomná v lineárním modelu, může být také použita příslušná p-hodnota testu hypotézy $\beta_i = 0$. Získanou p-hodnotu budeme porovnávat s předem zvolenou hladinou významnosti. Rozhodování na základě p-hodnoty je ve svém výsledku ekvivalentní výše popsanému postupu s t-statistikou, budeme mu nadále

dávat přednost při implementaci metod krokové regrese v této práci. Detailní postup rozhodování na základě p-hodnoty je uveden v následující kapitole (2.6) v kontextu praktické ukázky aplikace krokových metod.

1.6 Aplikace metod krokové regrese

V této kapitole si na konkrétních datech detailně ukážeme způsob fungování metod backward selection, forward selection a stepwise regression. Praktická aplikace všech metod krokové regrese byla realizována na databázi údajů o 27 onkologicky nemocných pacientech. Závisle proměnná ($Y_i = Remission$) je remise rakoviny, nezávisle proměnné ($X_{i1} = Faktor1, X_{i2} = Faktor2, X_{i3} = Faktor3$) jsou rizikové faktory pro vznik rakoviny. Data byla převzata z článku [9]. Soubor dat lze najít v elektronické příloze k práci.

1.6.1 Backward selection (sestupný výběr)

Algoritmus

Nejprve se odhaduje plný model a následně se posuzuje, zda z daného modelu lze vyloučit nějakou nezávisle proměnnou. Rozhoduje se na základě výsledků dílčích t-testů, vždy se najde nezávisle proměnná s maximální p-hodnotou. Kritériem pro vyloučení je to, zda maximální p-hodnota je větší než předem zadaná hodnota $PtoStay$. Funkce skončí ve chvíli, kdy v modelu nezůstane žádná nezávisle proměnná, anebo už nezbude v modelu nezávisle proměnná s p-hodnotou větší než předem zvolená hodnota $PtoStay$.

Vychozí hodnotu $PtoStay$ nastavíme na 0.05.

KROK 1

Odhadneme úplný model, zjistíme, zda je možné z daného modelu vyloučit nezávisle proměnnou.

	Estimate	Std.error	t-value	p-value
$Faktor1$	0.2450	0.2564	0.956	0.349
$Faktor2$	-0.1162	0.4451	-0.261	0.796
$Faktor3$	0.2887	0.2191	1.318	0.200

Na základě výsledků dílčích t-testů největší p-hodnotu = 0.796 můžeme pozorovat při testování hypotézy $\beta_2 = 0$. Platí $0.796 > 0.05$, tj. p-hodnota je větší než hodnota $PtoStay$. To znamená, že nemůžeme zamítnout hypotézu $\beta_2 = 0$. Proměnnou $Faktor2$ tím vyloučíme z modelu.

KROK 2

Odhadneme model bez proměnné $Faktor2$.

	Estimate	Std.error	t-value	p-value
$Faktor1$	0.1978	0.1783	1.109	0.278
$Faktor3$	0.2578	0.1810	1.425	0.167

Největší p-hodnotu ve druhém kroku získáme při testování hypotézy $\beta_1 = 0$, rovná se 0.278. Platí $0.278 > 0.05$. Nemůžeme zamítnout hypotézu $\beta_1 = 0$, nezávisle proměnnou $Faktor1$ lze vyloučit z modelu.

KROK 3

Odhadneme model bez proměnné $Faktor1$. V modelu zbývá jenom jedna nezávisle proměnná $Faktor3$.

	Estimate	Std.error	t-value	p-value
$Faktor3$	0.4250	0.1006	4.224	0.00026

P-hodnota je menší než hodnota $PtoStay$, $(0.00026 < 0.05)$, zamítáme hypotézu $\beta_3 = 0$. V posledním kroku se z modelu nic nevylučuje. Proměnná $Faktor3$ jako jediná zůstává ve výsledném modelu.

Ve výsledku metody backward selection jsme obdrželi následující lineární model:

$$Remission = 0.425 * Faktor3. \tag{33}$$

1.6.2 Forward selection (vzestupný výběr)

Algoritmus

Vycházíme z prázdné množiny nezávisle proměnných, do níž se pak v každém kroku přidá proměnná ze souboru nezařazených kandidátů. Rozhoduje se na základě výsledků dílčích t-testů, lineární model se odhaduje s každým kandidátem postupně, hledáme nezávisle proměnnou s minimální p-hodnotou. Kritériem pro zařazení je to, zda minimální p-hodnota je menší než předem zadaná hodnota $PtoEnter$. Funkce skončí ve chvíli, kdy v souboru kandidátů nezůstane žádná proměnná, anebo už nezbude v modelu nezávisle proměnná s p-hodnotou menší než hodnota $PtoEnter$.

Vychozí hodnotu $PtoEnter$ nastavíme na 0.1.

KROK 1

Odhadneme lineární model s každou proměnnou ze souboru kandidátů postupně, zjistíme, zda je možné přidat nějakou proměnnou do modelu.

	Estimate	Std.error	t-value	p-value
$Faktor1$	0.4094	0.1006	4.068	0.000392
$Faktor2$	0.5846	0.1508	3.877	0.000643
$Faktor3$	0.4250	0.1006	4.224	0.00026

Na základě výsledků dílčích t-testů nejmenší p-hodnotu = 0.00026 můžeme pozorovat při testování hypotézy $\beta_3 = 0$. Platí $0.00026 < 0.1$, tj. p-hodnota je menší než hodnota $PtoEnter$. Zamítneme hypotézu $\beta_3 = 0$, proměnnou $Faktor3$ přidáme do modelu.

KROK 2

Odhadneme nový model postupně s každou proměnnou ze souboru kandidátů, proměnná $Faktor3$ je z toho souboru vyloučena.

	Estimate	Std.error	t-value	p-value
$Faktor1$	0.1978	0.1783	1.109	0.278
$Faktor2$	0.1839	0.3150	0.584	0.565

Nejmenší p-hodnotu získáme při testování hypotézy $\beta_1 = 0$, rovná se 0.278. P-hodnota je větší než hodnota $PtoEnter$, $(0.278 > 0.1)$. Nemůžeme zamítnout hypotézu $\beta_1 = 0$, proměnná $Faktor1$ se nepřidává do modelu. Algoritmus skončí, protože v souboru kandidátů se nenašla proměnná s p-hodnotou menší než hodnota $PtoEnter$.

Ve výsledku metody forward selection jsme obdrželi následující lineární model:

$$Remission = 0.425 * Faktor3. \tag{34}$$

1.6.3 Stepwise regression (kroková regrese)

Algoritmus

Vzestupný výběr je v každém kroku kombinován s pokusem o zjednodušení pomocí sestupného výběru. Vycházíme z prázdné množiny nezávisle proměnných, do níž se pak v prvním kroku přidá proměnná ze souboru nezařazených kandidátů. Rozhoduje se na základě výsledků dílčích t-testů, odhadujeme model s každým kandidátem postupně, hledáme proměnnou s minimální p-hodnotou. Kritériem pro zařazení je to, zda minimální p-hodnota je menší než předem zadaná hodnota $PtoEnter$. V dalším kroku odhadujeme celý daný model, a následně posuzujeme, zda můžeme z něj vyloučit nějakou nezávisle proměnnou. Rozhoduje se opět na základě výsledků dílčích t-testů, hledá se nezávisle proměnná s maximální p-hodnotou. Kritériem pro vyloučení je to, zda maximální p-hodnota je větší než předem zadaná hodnota $PtoStay$. Dále se takto po krocích opakuje zařazování a vyřazování, dokud v souboru kandidátů nezůstane žádná nezávisle proměnná, anebo už nezbude v souboru kandidatů nezávisle proměnná s minimální p-hodnotou menší než hodnota $PtoEnter$.

Abychom předešli nekonečnému cyklu vyřazování a zařazování proměnných, vždy musí být splněná nerovnost $PtoStay > PtoEnter$.

Vychozí hodnoty: $PtoStay = 0.1$ a $PtoEnter = 0.05$

KROK 1

Odhadneme lineární model s každou proměnnou ze souboru kandidátů postupně, zjístíme, zda je možné přidat nějakou proměnnou do modelu.

	Estimate	Std.error	t-value	p-value
$Faktor1$	0.4094	0.1006	4.068	0.000392
$Faktor2$	0.5846	0.1508	3.877	0.000643
$Faktor3$	0.4250	0.1006	4.224	0.00026

Můžeme vidět, že na základě výsledků dílčích t-testů nejmenší p-hodnotu $= 0.00026$ můžeme pozorovat při testování hypotézy $\beta_3 = 0$. Platí $0.00026 < 0.05$, tj. p-hodnota je menší než hodnota $PtoEnter$. Zamítneme hypotézu $\beta_3 = 0$, proměnnou $Faktor3$ přidáme do modelu.

KROK 2

Odhadneme celý lineární model, zjistíme, zda je možné z daného modelu vyloučit nezávisle proměnnou $Faktor3$.

	Estimate	Std.error	t-value	p-value
$Faktor3$	0.4250	0.1006	4.224	0.00026

P-hodnota je menší než hodnota $PtoStay$, $(0.00026 < 0.1)$, zamítáme hypotézu $\beta_3 = 0$. Nezávisle proměnná $Faktor3$ zůstává v modelu.

KROK 3

Odhadneme model postupně s každou proměnnou ze souboru kandidátů, proměnná $Faktor3$ je z toho souboru vyloučena.

	Estimate	Std.error	t-value	p-value
$Faktor1$	0.1978	0.1783	1.109	0.278
$Faktor2$	0.1839	0.3150	0.584	0.565

Nejmenší p-hodnotu ve druhém kroku získáme při testování hypotézy $\beta_1 = 0$, rovná se 0.278. P-hodnota je větší než hodnota $PtoEnter$, $(0.278 > 0.05)$. Nemůžeme zamítnout

hypotézu $\beta_1 = 0$, proměnná $Faktor1$ se nepřidává do modelu. Algoritmus skončí, protože v souboru kandidátů se nenašla proměnná s p-hodnotou menší než hodnota $PtoEnter$.

Ve výsledku metody stepwise regression jsme obdrželi následující lineární model:

$$Remission = 0.425 * Faktor3. \tag{35}$$

2 Praktická část

V rámci této části práce ukážeme aplikaci výše popsaných metod krokové regrese na simulovaných datech. Simulace se provádí v statistickém softwaru R. Výpočet simulací je založený na metodě Monte Carlo. Zdrojové kódy simulací lze najít v elekronické příloze k práci.

Během simulací se opakovaně na základě lineárního regresního modelu $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ generují konkrétní data pomocí předem zvoleného vektoru $\boldsymbol{\beta}$. Na tato data $(Y_i, \boldsymbol{X}_i)$ budeme postupně aplikovat metody krokové regrese (backward selection, forward selection nebo stepwise regression), metodou nejmenších čtverců odhadneme výsledný model a tím získáme vektor odhadnutých regresních parametrů $\widehat{\boldsymbol{\beta}}$ pro každý soubor dat.

Naším cílem je pak zjístit, jestli vektor $\widehat{\boldsymbol{\beta}}$ je pro tento regresní model „kvalitním“ odhadem regresních parametrů. Budeme vycházet z toho, že vektor $\widehat{\boldsymbol{\beta}}$ by se měl co nejvíce přibližovat předem známému vektoru $\boldsymbol{\beta}$.

Dále budeme taky zkoumat, jak se mění „kvalita“ odhadnutého regresního modelu v závislosti na nastaveních Monte Carlo simulace - parametrech simulovaných modelů. Souhrn výsledků bude následně prezentován pomocí tabulek a grafů.

2.1 Popis dat

Nastavení a parametry pro simulace regresních modelů byly převzaty z článku [7], a již byly v tomto článku použity k testování jiných regresních metod (např. pro úplný model).

Simulace provádíme na základě regresního modelu

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \tag{36}$$

kde

$\boldsymbol{\beta}$ je vektor regresních parametrů délky p, je předem volené nastavení simulace.

$\mathbb{X}$ je matice nezávisle proměnných. Hodnoty nezávisle proměnných jsou generovány náhodným výběrem o rozsahu n z vícerozměrného (konkrétně p-rozměrného) normálního rozdělení, jež má nulový vektor středních hodnot a kovarianční matici, jejíž prvek na pozici $ij, i, j = 1, \ldots, p$, označený jako ρ_{ij}, je dán jako $\rho^{|i-j|}$, kde ρ je předem volený parametr simulace.

$$\mathbb{COV} = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{pmatrix} \tag{37}$$

Náhodná složka tohoto modelu se počítá jako součin $\sigma\boldsymbol{\epsilon}$. $\boldsymbol{\epsilon}$ je výběr z normovaného

normálního rozdělení o rozsahu n. σ je směrodatná odchylka náhodné složky a jedná se o předem volený parametr simulace.

$\boldsymbol{Y}$ je vektor závisle proměnných délky n.

Při simulaci budeme využívat tři nastavení parametrů $\boldsymbol{\beta}$, jež jsou uvedena v tabulce 1.

Vektory regresních parametrů $\boldsymbol{\beta}$ jsou voleny tak, aby uživatel měl možnost odzkoušet, zda vlastnosti dané metody jsou ovlivněny nastavením regresních parametrů, konkrétně počtem nulových hodnot ve vektoru $\boldsymbol{\beta}$. Pro první soubor dat platí, že výsledný vektor závisí pouze na první, druhé a páté vstupní proměnné. V druhém případě platí, že všechny vstupní proměnné se podílí na výsledku rovným dílem. V posledním případě jenom jedna proměnná přispívá k budování regresního modelu.

	$\boldsymbol{\beta}$
Soubor 1	(3, 1.5, 0, 0, 2, 0, 0, 0)
Soubor 2	(0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)
Soubor 3	(5, 0, 0, 0, 0, 0, 0, 0)

Tabulka 1: *Variace vektoru regresních parametrů*

Pro každé ze tří možných nastavení parametru $\boldsymbol{\beta}$ studujeme vlastnosti krokových metod pro různé hodnoty parametrů σ, n, ρ, *PtoStay* (u metody backward selection), *PtoEnter* (u metody forward selection) a kombinaci hodnot parametrů *PtoStay* a *PtoEnter* (u metody stepwise regression). Výchozí hodnoty těchto parametrů jsou uvedeny v tabulce 2. Výchozí hodnoty těchto parametrů budou v rámci Monte Carlo simulace posléze měněny tak, abychom mohli zkoumat, jak vlastnosti metod krokové regrese na těchto hodnotách závisejí.

σ	n	ρ	*PtoStay*	*PtoEnter*	*PtoStay*, *PtoEnter*
1	50	0.5	0.05	0.1	0.1, 0.05

Tabulka 2: *Výchozí nastavení parametrů*

Každá Monte Carlo simulace obsahuje matici nezávisle proměnných $\mathbb{X}$ o rozměru $n \times p$, kde n je počet pozorování a p je počet proměnných. Počet proměnných p v našich simulacích je vždy rovno 8. Délka vektoru závisle proměnných $\boldsymbol{Y}$ je potom n.

2.2 Postup zpracování dat

V rámci této práce byly implementovány metody krokové regrese: backward selection, forward selection a stepwise regression. V statistickém softwaru R jsme naprogramovali funkce
`backward(x, y, PtoStay)`,
`forward(x, y, PtoEnter)` a
`stepwise(x, y, PtoEnter, PtoStay)`.
Zdrojové kódy těchto funkcí lze najít v elektronické příloze k práci.

2.2.1 Implementace metody backward selection

Funkce `backward(x, y, PtoStay)` je implementací metody backward selection. Na vstupu dostává $\mathbb{X}$, vektor závisle proměnných $\boldsymbol{Y}$ a hodnotu *PtoStay* (zvolí si uživatel). Vrací konfiguraci vektoru regresních parametrů $\boldsymbol{\beta}$ - vektor logických hodnot $\boldsymbol{jeTam}$ s prvky TRUE a FALSE. Počet prvků ve vektoru $\boldsymbol{jeTam}$ se rovná počtu nezávisle proměnných v matici $\mathbb{X}$. TRUE na pozici vektoru $\boldsymbol{jeTam}$ znamená, že daná vysvětlující proměnná zůstala v modelu, FALSE - že byla vyloučena.

2.2.2 Implementace metody forward selection

Funkce `forward(x, y, PtoEnter)` je implementací metody forward selection. Na vstupu dostává $\mathbb{X}$, vektor závisle proměnných $\boldsymbol{Y}$ a hodnotu PtoEnter (zvolí si uživatel). Vrací konfiguraci vektoru regresních parametrů $\boldsymbol{\beta}$ - vektor logických hodnot $jeTam$ s prvky TRUE a FALSE. Počet prvků ve vektoru $\boldsymbol{jeTam}$ se rovná počtu nezávisle proměnných v matici $\mathbb{X}$. TRUE na pozici vektoru $\boldsymbol{jeTam}$ znamená, že daná vysvětlující proměnná byla zařazena do modelu, FALSE - že nebyla.

2.2.3 Implementace metody stepwise regression

Funkce `stepwise(x, y, PtoEnter, PtoStay)` je implementací metody stepwise, která je kombinací metod backward a forward. Na vstupu dostává $\mathbb{X}$, vektor závisle proměnných $\boldsymbol{Y}$ a hodnoty *PtoEnter* a *PtoStay* (zvolí si uživatel) a vrací konfiguraci vektoru regresních parametrů $\boldsymbol{\beta}$ - vektor logických hodnot $\boldsymbol{jeTam}$ s prvky TRUE a FALSE. Počet prvků se rovná počtu nezávisle proměnných v matici $\mathbb{X}$. TRUE na pozici vektoru $\boldsymbol{jeTam}$ znamená, že proměnná je přítomná v lineárním modelu, FALSE - že není.

2.2.4 Monte Carlo simulace

Dále v statistickém softwaru R jsme naprogramovali funkci

`MCsimulace(beta, sim, sigma, n, rho, PtoStay, PtoEnter)`.

Zdrojový kód této funkce lze najít v elektronické příloze k práci. Kvůli reprodukovatelnosti výsledků při simulacích byla použíta funkce `set.seed(555)`.

Funkce `MCsimulace(beta, sim, sigma, n, rho, PtoStay, PtoEnter)` na vstupu dostává příslušné parametry (vektor regresních parametrů $\boldsymbol{\beta}$, počet simulací sim, směrodatnou odchylku náhodné složky regresního modelu σ, počet pozorování n, sílu lineární závislosti mezi nezávisle proměnnými ρ, parametry implementovaných metod krokové regrese *PtoStay* a *PtoEnter*). Za použití výše popsaných funkcí `backward(x, y, PtoStay)`, `forward(x, y, PtoEnter)`, `stepwise(x, y, PtoEnter, PtoStay)` generuje regresní modely, počítá a vrací charakteristiky vytvořených regresních modelů: vychýlení odhadů regresních parametrů, relativní četnost detekce správné konfigurace vektoru regresních parametrů

$\boldsymbol{\beta}$ a chybu predikce. Chybu predikce nepřímo počítáme prostřednictvím výrazu

$$\mathsf{ME} = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathsf{T}} \mathbb{COV}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{38}$$

Funkce `MCsimulace(beta, sim, sigma, n, rho, PtoStay, PtoEnter)` taky vyhodnocuje přesnost provedených Monte Carlo simulací a vrací odhad této hodnoty. Přesnost Monte Carlo simulací posoudíme pomocí směrodatné chyby odhadu, kterou spočítáme následujícím způsobem: $\mathsf{SE} = \frac{S_{sim}}{\sqrt{sim}}$, kde sim je počet simulací a S_{sim} je výběrová směrodatná odchylka pro soubor simulovaných dat. Čím je hodnota SE bližší nule, tím je výsledek simulací přesnější. Ze vzorce lze vidět, že velký počet simulovaných dat zvýší přesnost Monte Carlo simulací.

Jak jsme již zmiňovali, v rámci Monte Carlo simulací budeme potřebovat měnit parametry lineárních modelů. Z tohoto důvodu jsme naprogramovali další funkce, uvnitř kterých se bude navíc měnit jeden z parametrů. Tyto funkce budou taky vytvářet regresní modely pomocí metod krokové regrese a počítat jejich charakteristiky (vychýlení odhadů regresních parametrů, relativní četnost detekce správné konfigurace vektoru regresních parametrů, chybu predikce, odhad přesnosti) pro každou hodnotu ze zadaného rozsahu daného parametru.

Funkce `MCsigma(beta, sim, sigmaV, n, rho, PtoStay, PtoEnter)` na vstupu dostává proměnné $\boldsymbol{\beta}$, sim, n, ρ, $PtoStay$, $PtoEnter$ a taky vektor $sigmaV$. Do vektoru $sigmaV$ uložíme hodnoty, kterých bude nabývat parametr σ.

Funkce `MCn(beta, sim, sigma, nV, rho, PtoStay, PtoEnter)` na vstupu dostává proměnné $\boldsymbol{\beta}$, sim, σ, ρ, $PtoStay$, $PtoEnter$ a vektor nV. Do vektoru nV uložíme hodnoty, kterých bude nabývat parametr n.

Funkce `MCrho(beta, sim, sigma, n, rhoV, PtoStay, PtoEnter)` na vstupu dostává proměnné $\boldsymbol{\beta}$, sim, σ, n, $PtoStay$, $PtoEnter$ a vektor $rhoV$. Do vektoru $rhoV$ uložíme hodnoty, kterých bude nabývat parametr ρ.

Funkce `MCptoenter(beta, sim, sigma, n, rho, PtoStayV)` na vstupu dostává proměnné $\boldsymbol{\beta}$, sim, σ, n, ρ a vektor hodnot $PtoStayV$, kterých bude nabývat parametr $PtoStay$.

Funkce `MCptoenter(beta, sim, sigma, n, rho, PtoEnterV)` na vstupu dostává proměnné $\boldsymbol{\beta}$, sim, σ, n, ρ a vektor hodnot $PtoEnterV$, kterých bude nabývat parametr $PtoEnter$.

Zdrojové kódy těchto funkcí lze najít v elektronické příloze k práci. Kvůli reprodukovatelnosti výsledků při simulacích byla použita funkce `set.seed(555)`.

2.3 Výsledky simulačních studií

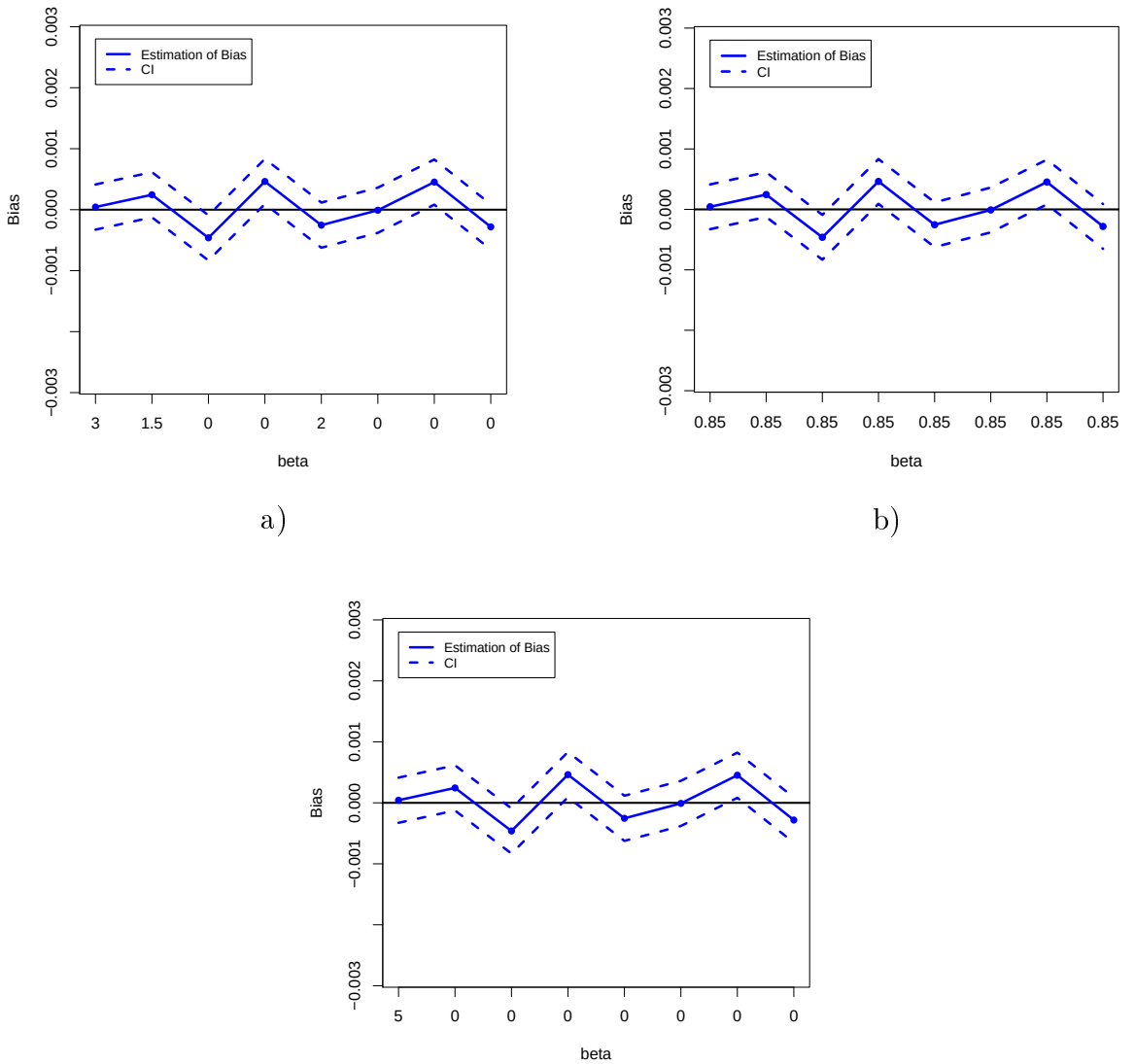
2.3.1 Vyhodnocení vychýlení

Vychýlení odhadů regresních parametrů je jedna z charakteristik regresních modelů, která se vyhodnocuje pomocí Monte Carlo simulací. Vektor vychýlení počítáme jako rozdíl střední hodnoty všech odhadů $\mathsf{E}\widehat{\boldsymbol{\beta}}$ a vstupního vektoru $\boldsymbol{\beta}$. Vzhledem k tomu, že počet simulací je konečný, nelze hodnoty vychýlení odhadů získat úplně přesně. Abychom zachytili tuto

nepřesnost, pro získané hodnoty vychýlení budeme konstruovat intervaly spolehlivosti.

Nejprve budeme zkoumat, jak se chová vychýlení v plném modelu. Regresní parametry v plném modelu jsme odhadovali metodou nejmenších čtverců. Víme, že tato metoda poskytuje nevychýlené odhady.

Obrázek 1 ukazuje, jak vypadá vychýlení jednotlivých parametrů ve vektoru $\widehat{\boldsymbol{\beta}}$. Znázorníme vychýlení odhadů pro všechny tři vektory $\boldsymbol{\beta}$ (viz. tabulka 1) při výchozím nastavení všech ostatních parametrů (viz. tabulka 2). Tečkovanou čárou jsou označeny 95% intervaly spolehlivosti. Můžeme vidět, že intervaly pokrývají hodnotu 0 s koeficientem spolehlivosti 95%. Z toho usoudíme, že odhady regresních parametrů v plném modelu jsou skutečně nevychýlené.



a) b)

c)

Obrázek 1: *Znázornění vychýlení odhadů v plném modelu*

Dále se podíváme na vychýlení v lineárních modelech, vzniklých pomocí metod krokové

regrese. Ted' budeme zkoumat chování vychýlení celého vektoru $\widehat{\boldsymbol{\beta}}$, nikoliv jeho jednotlivých složek. Vezmeme absolutní hodnotu vektoru vychýlení a spočítáme výběrový průměr složek daného vektoru. Tím získáme hodnotu vychýlení pro vektor $\widehat{\boldsymbol{\beta}}$. Jelikož v rámci simulace nemůžeme získat přesné odhady hodnot vychýlení, stejně jako předtím budeme konstruovat 95% intervaly spolehlivosti.

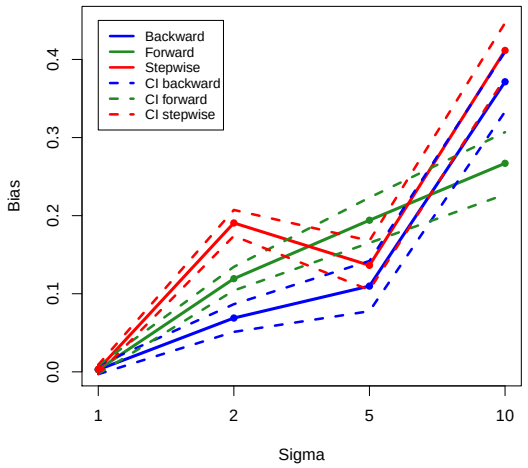
Pomocí Monte Carlo simulací dále zkusíme prozkoumat závislost hodnoty vychýlení na různých parametrech lineárního modelu. Simulace provádíme za podmínky, že při zkoumání vlivu jednoho parametru na vychýlení ostatní parametry budou nabývat výchozích hodnot (viz. tabulka 2).

Vyhodnocení vychýlení budeme provádět pro nenulové regresní parametry. Z tohoto důvodu jako vstupní vektor regresních parametrů pro tuto simulaci si zvolíme vektor, který neobsahuje nulové hodnoty, a to $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$, který je představený v tabulce 2 jako druhé nastavení vektoru regreních parametrů. Ostatní parametry necháváme nastavené výchozím způsobem podle tabulky 2. Výsledky simulací budeme vykreslovat do grafů v závislosti na tom parametru, který se momentálně mění.

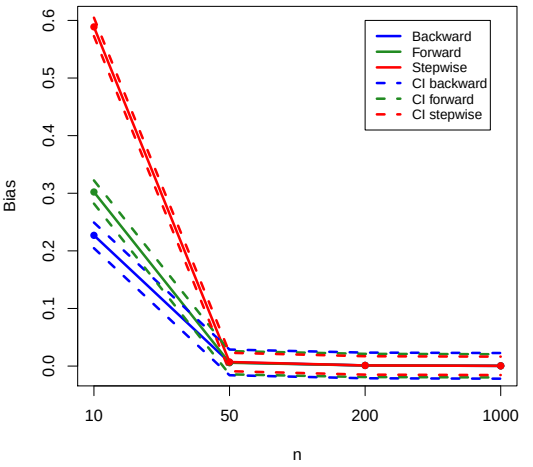
Obrázek 2 a) ukazuje, jak se mění hodnota vychýlení v závislosti na parametru σ. Směrodatná odchylka náhodné složky v rámci této simulace bude nabývat hodnot 1, 2, 5 a 10. U všech metod krokové regrese při hodně malých hodnotách σ vychýlení je téměř nulové. S rostoucí hodnotou parametru σ lze očekávat růst hodnoty vychýlení. To se vysvětluje tím, že vyšší hodnoty parametru σ zvětší vliv náhodné složky simulovaného regresního modelu a tím sníží sílu dílčích t-testů. Nízká síla testů vede k tomu, že $H_0 : \beta_i = 0$ bude chybně nezamítnuta. V důsledku nezamítnutí nulové hypotézy jsou v odhadnutém vektoru regresních parametrů nulové hodnoty, tj. odhady jsou vychýlené.

Obrázek 2 b) ukazuje, jak se mění hodnota vychýlení v závislosti na parametru n. Počet pozorování se bude měnit takto: 10, 50, 200 a 1000. Na obrázku 2 b) můžeme vidět, že s rostoucí hodnotou parametru n hodnota vychýlení odhadů regresních parametrů klesá. Čím je výběrový soubor větší, tím je vyšší síla dílčích t-testů, protože máme více informací o souboru dat. Testy s vyšší silou budou s větší pravděpodobností správně zamítat hypotézu $H_0 : \beta_i = 0$, která ve skutečnosti neplatí, tj. vychýlení odhadů regresních parametrů bude klesat. Pro velmi vysoké hodnoty počtu pozorování vychýlení je téměř nulové.

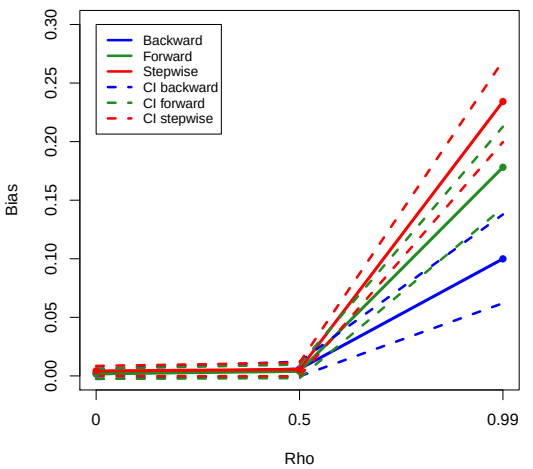
Obrázek 2 c) ukazuje, jak se mění hodnota vychýlení v závislosti na parametru ρ. Simulujeme data s nulovou ($\rho = 0$), slabou ($\rho = 0.5$) a silnou ($\rho = 0.99$) lineární závislostí. Čím menší je závislost mezi $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$, $i, j = 1, \ldots, 8$, tím větší je pravděpodobnost, že ve více případech hypotéza $H_0 : \beta_i = 0$ bude zamítnuta. Na obrázku 2 c) lze vidět, že s rostoucí hodnotou síly lineární závislosti mezi $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$ roste hodnota vychýlení. Při silné lineární závislosti mezi $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$ v odhadnutém vektoru $\widehat{\boldsymbol{\beta}}$ jsou přítomné nulové hodnoty, tj. odhady jsou vychýlené.



a) σ



b) n

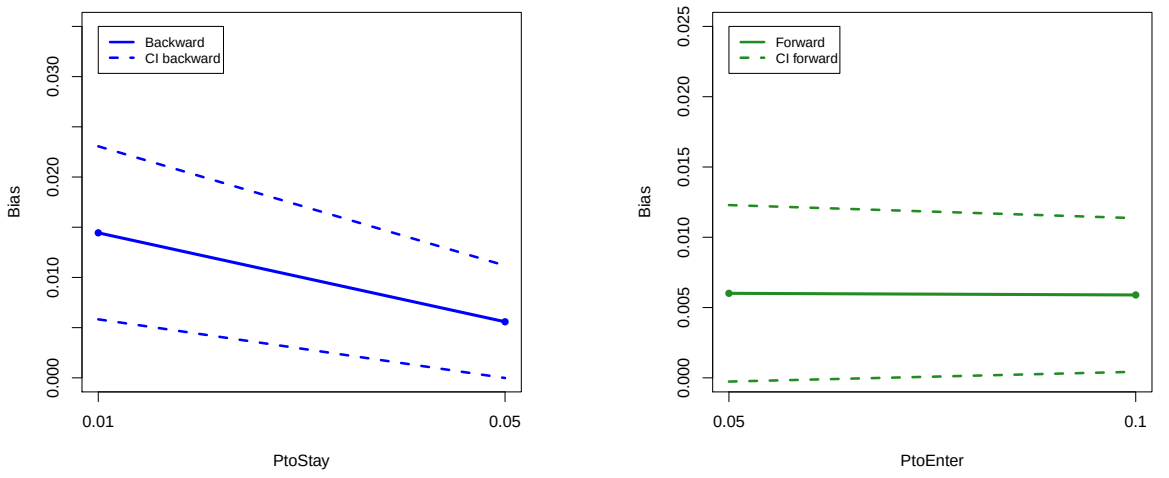


c) ρ

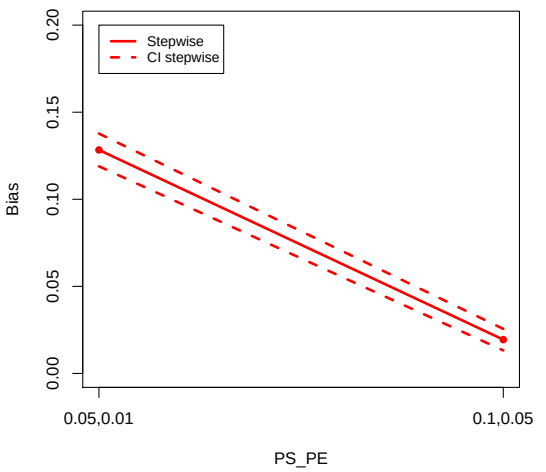
Obrázek 2: *Znázornění vychýlení odhadů v závislosti na parametrech σ, n, ρ*

Obrázek 3 znázorňuje závislosti hodnoty vychýlení na parametrech *PtoStay* a *PtoEnter*. Pro každou metodu vykreslíme vývoj vychýlení zvlášť, v závislosti na parametru příslušné funkce, parametr *PtoStay* nabývá hodnot 0.01 a 0.05, parametr *PtoEnter* nabývá hodnot 0.05 a 0.1. Všechny metody reagují na změnu parametrů podobně, hodnota vychýlení klesá s rostoucími hodnotami parametrů *PtoStay* a *PtoEnter*. To souvisí s tím, že vyšší hodnoty *PtoStay* a *PtoEnter* zvětší sílu dílčích t-testů, to vede k tomu, že hypotéza $H_0 : \beta_i = 0$ se bude častěji zamítat. Vychýlení odhadů regresních parametrů při velkých hodnotách parametrů *PtoStay* a *PtoEnter* bude velmi nízké.

Můžeme dospět k závěru, že metody krokové regrese budou vždy poskytovat vychýlené odhady regresních parametrů. Vychýlení odhadů těsně souvisí se silou a výsledky dílčích t-testů, a vzniká v důsledku vyřazování a zařazování proměnných do lineárního modelu.



a) b)



c)

Obrázek 3: *Znázornění vychýlení odhadů v závislosti na parametrech PtoStay a PtoEnter*

2.3.2 Vyhodnocení relativní četnosti detekce správné konfigurace

Relativní četnost detekce správné konfigurace vektoru regresních parametrů je další charakteristika, kterou se snažíme vyhodnotit pomocí Monte Carlo simulací. Konfigurace vektoru regresních parametrů je vektor logických hodnot délky p s prvky TRUE a FALSE. TRUE na i-té pozici vektoru konfigurace znamená, že i-tá proměnná je přítomná v odhadnutém lineárním modelu, FALSE - že není. Průběžně při odhadování simulovaných modelů budeme počítat ty vektory odhadnutých parametrů $\widehat{\boldsymbol{\beta}}$, které mají stejnou konfiguraci, jakou má vstupní vektor $\boldsymbol{\beta}$. Pak vydělíme počet vektorů se správnou konfigurací celkovým počtem provedených simulací a tak zjistíme relativní četnost správné konfigurace vektoru regresních

parametrů. [3]

Pomocí Monte Carlo simulací dále budeme vyšetřovat závislost relativní četnosti detekce správné konfigurace na různých parametrech modelu. Simulace se zase budou provádět za podmínky, že při zkoumání vlivu jednoho parametru ostatní parametry budou nabývat výchozích hodnot (viz. tabulka 2).

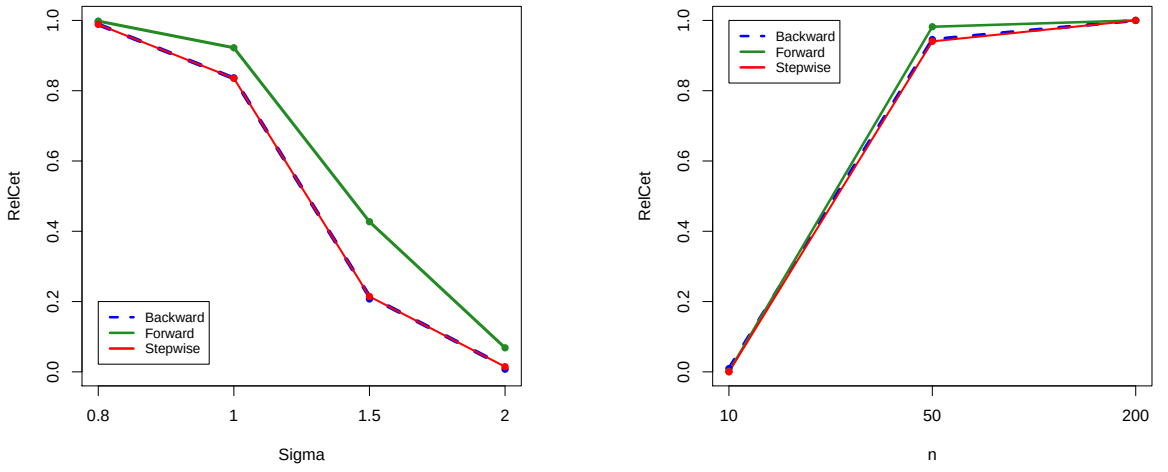
Nejprve zkusíme prozkoumat vektor konfigurace, který obsahuje na všech pozicích logickou hodnotu TRUE. Proto si jako vstupní vektor regresních parametrů zvolíme $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$, který je představený v tabulce 1 jako druhé nastavení vektoru regreních parametrů. Ostatní parametry necháváme nastavené defaultně podle tabulky 2. Výsledky simulací budeme vykreslovat do grafů v závislosti na tom parametru, který se momentálně bude měnit.

Obrázek 4 a) ukazuje, jak se mění relativní četnosti detekce správné konfigurace $\boldsymbol{\beta}$ v závislosti na parametru σ. Směrodatná odchylka náhodné složky v rámci této simulace bude nabývat hodnot 0.8, 1, 1.5 a 2. S rostoucí hodnotou parametru σ lze očekávat pokles těchto relativních četností u všech metod krokové regrese. Vyšší hodnoty parametru σ zvětší vliv náhodné složky simulovaného regresního modelu a tím sníží sílu dílčích t-testů. Nízká síla testů vede k tomu, že $H_0 : \beta_i = 0$ je chybně nezamítnuta. V důsledku nezamítnutí nulové hypotézy v odhadnutých vektorech regresních parametrů jsou přítomné nulové hodnoty. Do vektorů konfigurace regresních parametrů se dostanou logické hodnoty FALSE, což sníží relativní četnosti detekce správné konfigurace.

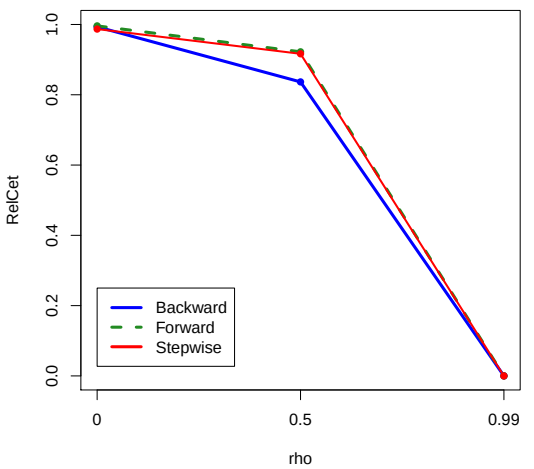
Obrázek 4 b) ukazuje, jak se mění hodnoty relativní četnosti detekce správné konfigurace $\boldsymbol{\beta}$ v závislosti na parametru n. Počet pozorování se bude měnit takto: 10, 50 a 200. Na obrázku 4 b) můžeme vidět, že s rostoucí hodnotou parametru n hodnoty relativních četností u všech metod krokové regrese rostou. Čím je výběrový soubor větší, tím je vyšší síla dílčích t-testů, protože máme více informací o souboru dat. Testy s vyšší silou s větší pravděpodobností správně zamítnou hypotézu $H_0 : \beta_i = 0$, která ve skutečnosti neplatí. Potom konfigurace u více odhadnutých vektorů regresních parametrů obsahuje hodnoty TRUE na všech pozicích, což zvýší relativní četnosti detekce správné konfigurace pro všechny metody krokové regrese.

Obrázek 4 c) ukazuje, jak se mění hodnoty relativních četností detekce správné konfigurace $\boldsymbol{\beta}$ v závislosti na parametru ρ. Opět simulujeme data s nulovou ($\rho = 0$), slabou ($\rho = 0.5$) a silnou ($\rho = 0.99$) lineární závislostí. Na obrázku 4 c) lze vidět, že s rostoucí hodnotou síly lineární závislosti mezi $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$ klesají relativní četnosti. Čím menší je závislost mezi $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$, tím větší je pravděpodobnost, že ve více případech hypotéza $H_0 : \beta_i = 0$ bude zamítnuta, a tím větších hodnot budou nabývat relativní četnosti detekce správné konfigurace.

[3] Uvedenou relativní četnost je možno považovat za odhad pravděpodobnosti detekce správné konfigurace. Opět by tudiž bylo možné konstruovat intervaly spolehlivosti, ale pro jednoduchost vyhodnocení výsledků tak není činěno, vzhledem k tomu, že příslušné odhady pravděpodobnosti jsou relativně přesné a vzhledem k tomu, že nám jde především o získání kvalitativních (orientačních) představ o závislosti pravděpodobnosti detekce správné konfigurace.



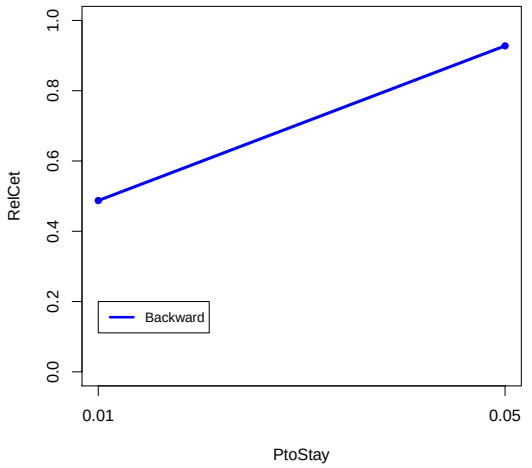
a) σ b) n



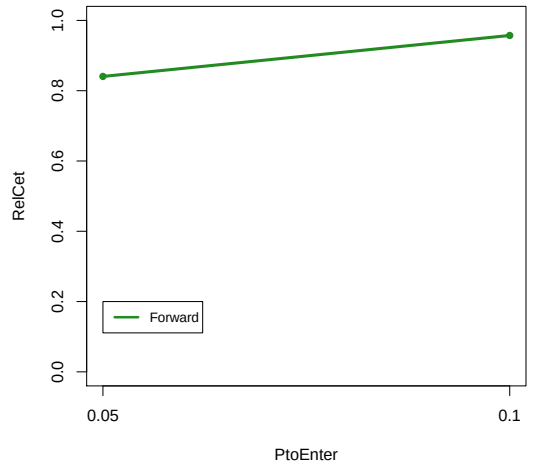
c) ρ

Obrázek 4: *Znázornění vývoje relativní četnosti v závislosti na parametrech σ, n a ρ*

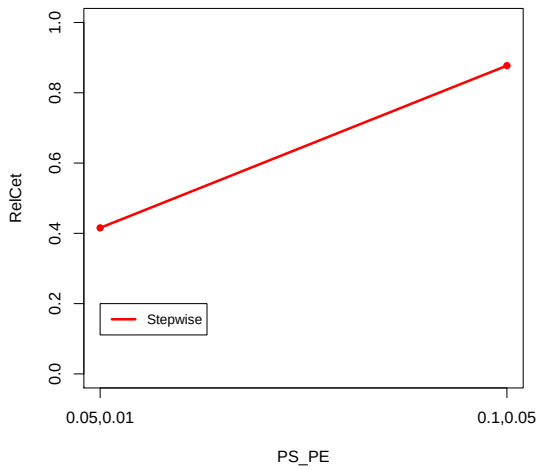
Na obrázku 5 jsou znázorněny závislosti relativní četnosti detekce správné konfigurace $\boldsymbol{\beta}$ na parametrech *PtoStay* a *PtoEnter*. Pro každou metodu vykreslíme vývoj relativních četností zvlášť', v závislosti na parametru příslušné funkce. Parametr *PtoStay* nabývá hodnot 0.01 a 0.05, parametr *PtoEnter* nabývá hodnot 0.05 a 0.1. S rostoucími hodnotami parametrů *PtoStay* a *PtoEnter* se zvětšuje šance, že víc vektorů odhadnutých regresních parametrů bude mít stejnou konfiguraci jako vstupní vektor $\boldsymbol{\beta}$. Hodnoty relativních četností klesají s rostoucími hodnotami parametrů *PtoStay* a *PtoEnter*. To souvisí s tím, že vyšší hodnoty *PtoStay* a *PtoEnter* zvětší sílu dílčích t-testů, což vede k zamítnutí hypotézy $H_0 : \beta_i = 0$. To znamená, že při dost vysokých hodnotách parametrů *PtoStay* a *PtoEnter* obdržíme ve výsledku správnou konfiguraci vektoru $\widehat{\boldsymbol{\beta}}$.



a)



b)



c)

Obrázek 5: *Znázornění vývoje relativní četnosti v závislosti na parametrech PtoStay a PtoEnter*

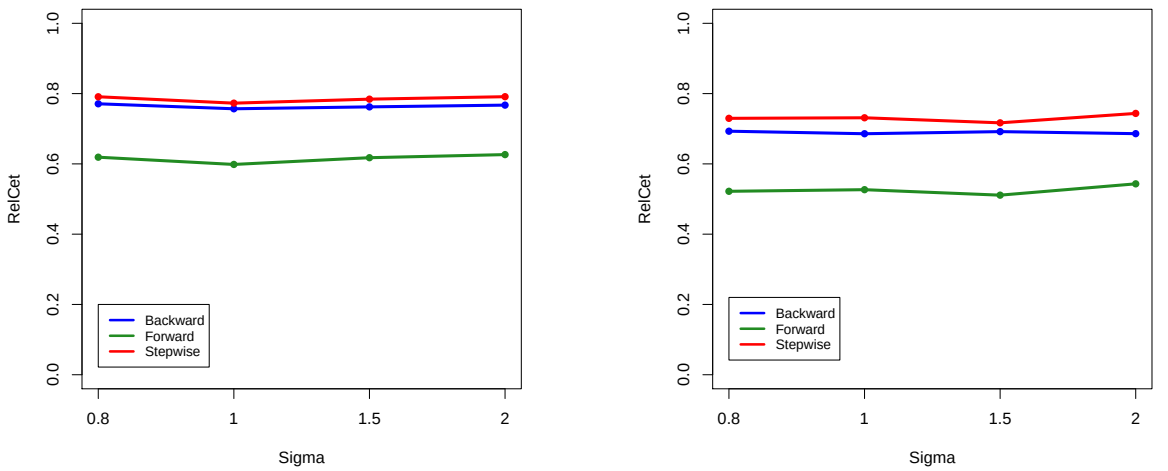
Doposud jsme analyzovali závislosti vlastností metod krokové regrese pomocí vektoru, jehož konfigurace neobsahovala žádnou logickou hodnotu FALSE. Dále zkusíme zadat jako vstupní vektor regresních parametrů $\boldsymbol{\beta}$ následující vektory:

- vektor $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$, jehož vektor konfigurace obsahuje na první, druhé a páté pozicích logickou hodnotu TRUE a na všech ostatních pozicích logickou hodnotu FALSE. Tento vektor je představený v tabulce 1 jako první nastavení vektoru regresních parametrů.

- vektor $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)$, jehož vektor konfigurace obsahuje na první pozici logickou hodnotu TRUE a na všech ostatních pozicích logickou hodnotu FALSE. Tento vektor je představený v tabulce 1 jako třetí nastavení vektoru regresních parametrů.

Ostatní parametry necháváme nastavené výchozím způsobem podle tabulky 2. V rámci simulací budeme vždy měnit jeden parametr.

Na obrázku 6 je znázorněno, jak se vyvíjí relativní četnosti správné detekce obou vektorů $\boldsymbol{\beta}$ v závislosti na parametru σ. Parametr σ bude nabývat hodnot 0.8, 1, 1.5 a 2. Průběh funkcí relativních četností pro všechny metody krokové regrese by se dal popsat jako víceméně konstantní. To znamená, že pro ostatní parametry, nabývající svých výchozích hodnot dle tabulky 2, relativní četnosti u zadaných vektorů nejsou citlivé na změnu parametru σ. To se vysvětluje tím, že vysoké hodnoty parametru σ ovlivňují výsledek dílčích t-testů a vedou k nezamítnutí hypotézy $H_0 : \beta_i = 0$. Potom konfigurace odhadnutých vektorů $\widehat{\boldsymbol{\beta}}$ bude obsahovat logické hodnoty FALSE. Vzhledem k tomu, že v konfiguracích zadaných vektorů $\boldsymbol{\beta}$ jsou taky přítomné hodnoty FALSE, rostoucí hodnota parametru σ nepovede k poklesu relativních četností jako v předchozím případě.

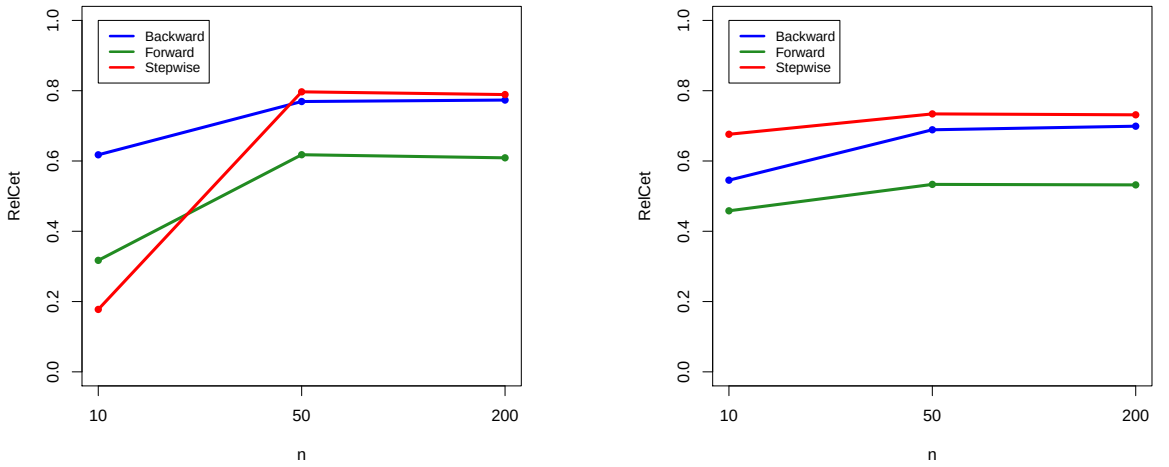


a) $\boldsymbol{\beta}$ = (3,1.5,0,0,2,0,0,0) b) $\boldsymbol{\beta}$ = (5,0,0,0,0,0,0,0)

Obrázek 6: *Znázornění vývoje relativní četnosti v závislosti na parametru σ*

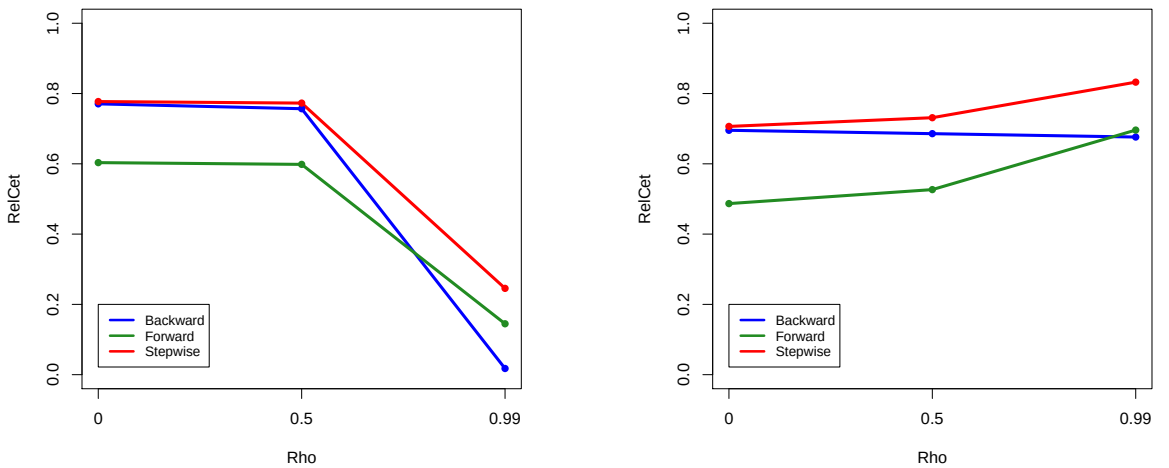
Na obrázku 7 je znázorněno, jak se vyvíjí relativní četnosti detekce správné konfigurace obou vektorů $\boldsymbol{\beta}$ v závislosti na parametru n. Studovaný rozsah parametru n je 10, 50 a 200. Můžeme vidět, že pro parametry, nabývající svých výchozích hodnot dle tabulky 2, relativní četnosti u všech krokových metod jsou méně citlivé v závislosti na parametru n. Funkce relativních četností jsou pro oba dva zadané vektory $\boldsymbol{\beta}$ lehce rostoucí. Souvisí to opět s výsledky dílčích t-testů. Čím je výběrový soubor větší, tím častěji se bude zamítat nulová hypotéza $H_0 : \beta_i = 0$, tj. ve výsledném vektoru konfigurace odhadnutých vektorů $\widehat{\boldsymbol{\beta}}$ budou zahrnuty logické hodnoty TRUE. Jelikož konfigurace zadaných vektorů $\boldsymbol{\beta}$ obsahují větší počet logických hodnot FALSE, můžeme usoudit, že větší počet pozorování v tomto případě nepovede k výraznému růstu relativních četností detekce správné konfigurace.

Na obrázku 8 můžeme vidět, jak se mění hodnoty relativních četností detekce správné konfigurace $\boldsymbol{\beta}$ v závislosti na parametru ρ. Studovaný rozsah parametru je 0, 0.5 a 0.99.



a) $\boldsymbol{\beta} = (3,1.5,0,0,2,0,0,0)$ b) $\boldsymbol{\beta} = (5,0,0,0,0,0,0,0)$

Obrázek 7: *Znázornění vývoje relativní četnosti v závislosti na parametru n*



a) $\boldsymbol{\beta} = (3,1.5,0,0,2,0,0,0)$ b) $\boldsymbol{\beta} = (5,0,0,0,0,0,0,0)$

Obrázek 8: *Znázornění vývoje relativní četnosti v závislosti na parametru ρ*

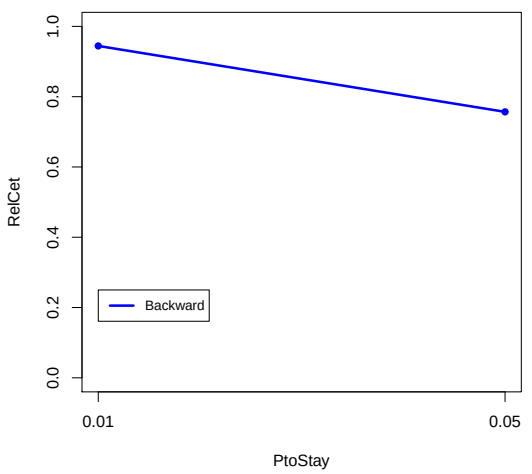
Na obrázku 8 a) lze vidět, že pro všechny parametry, nabývající svých výchozích hodnot dle tabulky 2, s rostoucí hodnotou parametru ρ funkce relativních četností lehce klesají. Čím menší je lineární závislost mezi $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$, tím větší je pravděpodobnost, že ve více případech hypotéza $H_0 : \beta_i = 0$ bude zamítnuta, tj. ve výsledném vektoru konfigurace odhadnutého vektoru $\widehat{\boldsymbol{\beta}}$ budou zahrnuty logické hodnoty TRUE. Jelikož konfigurace vektoru $\boldsymbol{\beta}$ obsahuje větší počet logických hodnot FALSE, můžeme usoudit, že větší hodnota parametru ρ v tomto případě vede k nevýraznému poklesu relativních četností detekce správné konfigurace.

Na obrázku 8 b) můžeme dokonce vidět, že s rostoucí hodnotou parametru ρ relativní četnosti lehce rostou. Platí to opět pro všechny ostatní parametry, nabývající svých výcho-

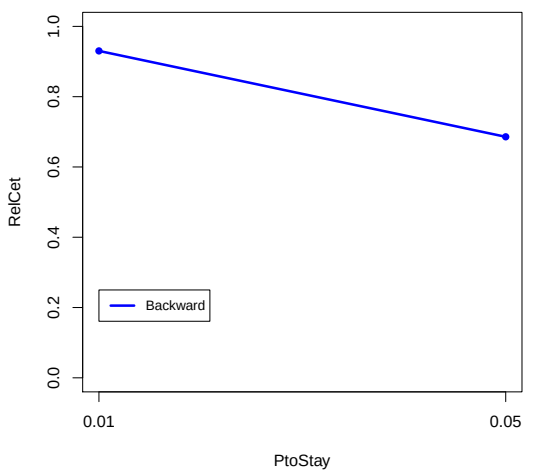
zích hodnot dle tabulky 2. Při silné lineární závislosti mezi $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$ ve vektoru konfigurace odhadnutého vektoru $\widehat{\boldsymbol{\beta}}$ se vyskytuje větší počet logických hodnot FALSE. Konfigurace zadaného vektoru $\boldsymbol{\beta}$ obsahuje hodnoty FALSE téměř na všech pozicích, proto relativní četnosti detekce správné konfigurace s rostoucí hodnotou parametru ρ budou v případě tohoto vektoru růst.

Na obrázku 9 je znázorněn vývoj relativních četností správné detekce obou vektorů v závislosti na parametrech *PtoStay* a *PtoEnter*. Pro každou metodu vykreslíme vývoj relativních četností zvlášť, v závislosti na parametru příslušné funkce. Parametr *PtoStay* nabývá hodnot 0.01 a 0.05, parametr *PtoEnter* nabývá hodnot 0.05 a 0.1. Všechny metody reagují na změnu parametrů stejně, přímky jsou klesající. To souvisí s tím, že vyšší hodnoty *PtoStay* a *PtoEnter* vedou k zamítnutí hypotézy $H_0 : \beta_i = 0$. To znamená, že při dost vysokých hodnotách parametrů *PtoStay* a *PtoEnter* obdržíme ve výsledku konfigurace vektorů odhadnutých regresních parametrů, ve kterých jsou zahrnuty logické hodnoty TRUE. Jelikož konfigurace zadaných vektorů $\boldsymbol{\beta}$ obsahují větší počet logických hodnot FALSE, můžeme usoudit, že zvýšení parametrů *PtoStay* a *PtoEnter* v tomto případě nepovede k růstu relativních četností detekce správné konfigurace, ale naopak k jejich poklesu.

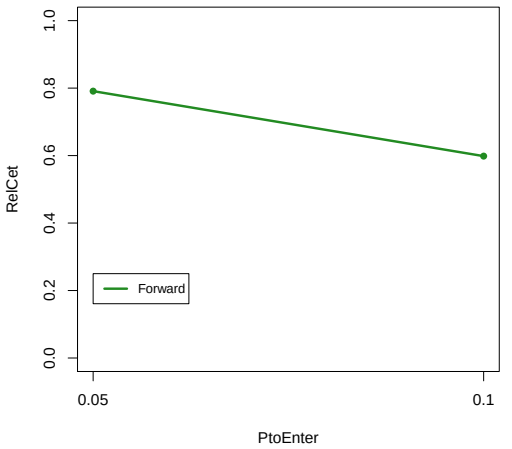
V rámci simulačních studií v této podkapitole jsme počítali relativní četnosti detekce správné konfigurace vektoru regresních parametrů. Pomocí této chatakteristiky jsme se snažili posoudit „kvalitu“ odhadnutých lineárních modelů, které jsme získali s použitím metod krokové regrese. Po provedené analýze můžeme dospět k závěru, že metody krokové regrese ne vždy dokážou poskytnout odhady regresních parametrů se správnou konfigurací. To souvisí se silou a výsledky testů hypotéz, které používáme při budování lineárního modelu.



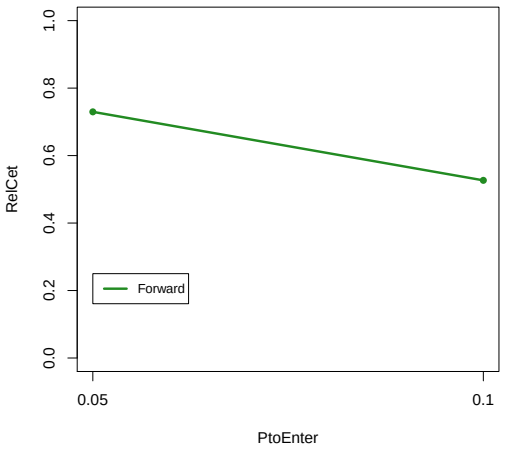
a) $\boldsymbol{\beta}$ = (3,1.5,0,0,2,0,0,0)



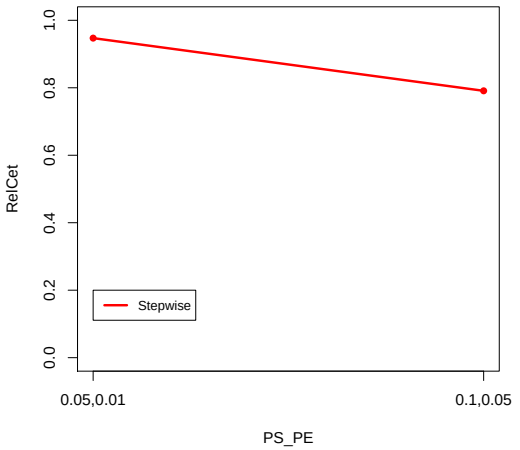
b) $\boldsymbol{\beta}$ = (5,0,0,0,0,0,0,0)



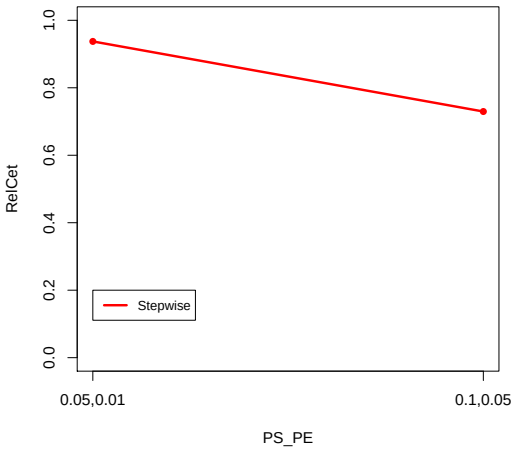
c) $\boldsymbol{\beta} = (3,1.5,0,0,2,0,0,0)$



d) $\boldsymbol{\beta} = (5,0,0,0,0,0,0,0)$



e) $\boldsymbol{\beta} = (3,1.5,0,0,2,0,0,0)$



f) $\boldsymbol{\beta} = (5,0,0,0,0,0,0,0)$

Obrázek 9: *Znázornění vývoje relativní četnosti v závislosti na parametrech PtoStay a PtoEnter*

2.3.3 Vyhodnocení chyby predikce

Chyba predikce je další vlastnost simulovaných regresních modelů, kterou se v této práci budeme zabývat. Chybu predikce odhadnutého modelu budeme v rámci této práce počítat prostřednictvím rovnice(38). Hodnotu ME budeme počítat jak pro modely, které jsme získali s použitím všech tří metod krokové regrese, tak pro plný model.

Pomocí Monte Carlo simulací budeme zkoumat, jak se mění chyba predikce v závislosti na parametrech σ a n. Pomocí různých kombinací těchto dvou parametrů budeme simulovat různé regresní modely. Zavedeme charakteristiku σ^2/n. Cílem je prozkoumat, jak ovlivní predikci modelů variace tohoto poměru. Připomeneme si, že všechny ostatní parametry budou nabývat výchozích hodnot dle tabulky 2.

Nejprve zkusíme vyhodnotit chybu predikce pro případ, kdy vektor regresních parametrů neobsahuje nulové hodnoty. Proto si jako vstupní vektor regresních parametrů zvolíme

$\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$, který je představený v tabulce 1 jako druhé nastavení vektoru regresních parametrů.

σ^2/n	Plný Model	Backward	Forward	Stepwise
0.08	0.75	1.82	1.47	1.9
0.5	4.70	7.03	5.98	7.25
1	13.95	12.88	10.74	12.14
2.5	47.55	25.11	20.99	18.00
4.9	93.19	41.09	34.89	25.28

Tabulka 3: *Chyba predikce pro* $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$

Z tabulky 3 lze vidět, že pro menší poměry σ^2/n chyba predikce je mírně nižší u plného modelu než u metod krokové regrese, kdežto pro větší poměry σ^2/n metody krokové regrese fungují lépe než plný model. To znamená, že v případě velkého σ a malého n lze tudíž ocekávat, že odhadnuté lineární modely, jež nám vrátí metody krokové regrese, budou z hlediska predikce lepší než plný model, a to i přesto, ze všechny skutečné parametry $\boldsymbol{\beta}$ jsou nenulové.

Dále zkusíme zadat jako vstupní vektor regresních parametrů $\boldsymbol{\beta}$ vektory, které budou obsahovat nulové hodnoty:

- vektor $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$. Tento vektor je představený v tabulce 1 jako první nastavení vektoru regresních parametrů.
- vektor $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)$. Tento vektor je představený v tabulce 1 jako třetí nastavení vektoru regresních parametrů.

σ^2/n	Plný Model	Backward	Forward	Stepwise
$1 * 10^{-6}$	$1.9 * 10^{-7}$	$1.4 * 10^{-7}$	$1.3 * 10^{-7}$	$9.6 * 10^{-8}$
0.001	0.008	0.004	0.005	0.004
0.08	0.75	0.36	0.42	0.35
0.5	4.70	3.92	3.66	3.86
1	13.95	9.2	8.32	7.98
2.5	47.55	25.65	17.86	18.50
4.9	93.19	43.9	31.18	27.85

Tabulka 4: *Chyba predikce pro* $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$

σ^2/n	Plný Model	Backward	Forward	Stepwise
$1 * 10^{-6}$	$1.9 * 10^{-7}$	$1.3 * 10^{-7}$	$1.2 * 10^{-7}$	$9.2 * 10^{-8}$
0.001	0.008	0.003	0.004	0.004
0.08	0.75	0.18	0.22	0.22
0.5	4.70	1.60	2.01	1.38
1	13.95	4.03	4.31	2.87
2.5	47.55	18.81	13.87	11.23
4.9	93.19	39.52	27.18	24.00

Tabulka 5: *Chyba predikce pro* $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)$

Z tabulek 4 a 5 lze vidět, že pro zadané vektory význam metod krokové regrese z hlediska predikce může být ještě větší. I pro hodně malé hodnoty poměru σ^2/n metody krokové regrese mají lepší výsledky než plný model. Můžeme si taky povšimnout, že pro ten vektor $\boldsymbol{\beta}$, který obsahuje větší počet nulových hodnot, metody krokové regrese mají z hlediska predikce výrazně lepší výsledky. Můžeme dospět k závěru, že v případě, kdybychom měli lineární model s malým počtem pozorovaných dat, která by navíc byla zašumělá, můžeme z hlediska predikce dát metodám krokové regrese přednost před plným modelem.

2.3.4 Souhrn výsledků

V této podkapitole jsou shrnuty vlastnosti metod krokové regrese, které jsme zkoumali v praktické části bakalářské práce pomocí simulačních studií.

Jako první vlastnost jsme studovali vychýlení odhadů regresních parametrů, získaných použitím metod krokové regrese. Provedené Monte Carlo simulace ukázaly, že metody krokové regrese poskytují vychýlené odhady regresních parametrů. Taky jsme zjistili, jak závisí vychýlení odhadů vektorů $\boldsymbol{\beta}$ na různých parametrech regresních modelů. Vždy nás zajímala závislost vychýlení na studovaném rozsahu jednoho z parametrů, zatímco ostatní parametry nabývali výchozích hodnot dle tabulky 2.

Dospěli jsme k závěru, že hodnota vychýlení pro všechny metody bude klesat, pokud budeme snižovat hodnotu parametru σ, zvyšovat hodnotu parametru n, snižovat hodnotu parametru ρ nebo zvyšovat hodnoty parametrů *PtoStay* a *PtoEnter*. Tyto závislosti vzniknou v důsledku toho, že výše uvedené parametry můžou ovlivnit sílu a výsledky testů hypotéz, které provádíme při budování regresního modelu pomocí metod krokové regrese.

Další studovanou vlastností byla relativní četnost detekce správné konfigurace vektoru regresních parametrů. Tuto vlastnost jsme zkoumali pomocí tří různých nastavení vektoru regresních parametrů z tabulky 1. Ve výsledku simulačních studií jsme zjistili, jak se chová relativní četnost v závislosti na rozsahu jednoho z parametrů regresního modelu, zatímco ostatní parametry nabývají výchozích hodnot dle tabulky 2. Pro druhé nastavení vektoru regresních parametrů (viz. tabulka 1), které neobsahuje nulové hodnoty, jsme obdrželi následující závislosti: relativní četnost pro všechny metody bude růst, pokud budeme snižovat hodnotu parametru σ, svyšovat hodnotu parametru n, snižovat hodnotu parametru ρ nebo

zvyšovat hodnoty parametrů *PtoStay* a *PtoEnter*. Pro první a třetí nastavení regresních parametrů (viz. tabulka 1), která zahrnují větší počty nulových hodnot, jsme pomocí Monte Carlo simulací získali jiné výsledky. Chování relativních četností v závislosti na jednotlivých parametrech je u těchto vektorů složitější. Funkce relativních četností budou růst, pokud budeme zvyšovat parametr n a snižovat parametry *PtoStay* a *PtoEnter*. Pro parametr σ funkce relativních četností nejsou monotónní. Pokud budeme zvyšovat parametr ρ, relativní četnosti pro první nastavení $\boldsymbol{\beta}$ budou růst a pro třetí nastavení $\boldsymbol{\beta}$ budou klesat.

Tyto závěry opět plynou z toho, že výsledky testů hypotéz závisí na parametrech lineárních modelů a tím padem ovlivňují výsledný vektor konfigurace.

Poslední studovanou vlastností byla chyba predikce. Chování chyby predikce jsme zkoumali pro všechny metody krokové regrese a pro plný model. Pro detailnější vyhodnocení chyby predikce byla použita tři různá nastavení vektoru regresních parametrů z tabulky 1. Nejprve jsme se podívali na druhé nastavení $\boldsymbol{\beta}$, které neobsahuje nulové hodnoty. Při dost velkých hodnotách poměru σ^2/n chyby predikce u plného modelu jsou výrazně vyšší než u modelů, získaných krokovou regresí. Při nižších hodnotách poměru σ^2/n výsledky pro všechny metody krokové regrese jsou mírně horší než u plného modelu.

Pro první a třetí nastavení $\boldsymbol{\beta}$ chyby predikce u všech metod krokové regrese jsou výrazně nižší než u plného modelu. Platí to i pro velmi nizké hodnoty poměru σ^2/n. Z toho plyne, že z hlediska predikce metodám krokové regrese můžeme dát přednost před plným modelem.

Závěr

Cílem této bakalářské práce bylo charakterizovat model lineární regrese a popsat jeho vlastnosti. Ukázali jsme, jak lze odhadovat regresní parametry jak metodou nejmenších čtverců, tak pomocí metod krokové regrese. V teoretické části práce jsme zadefinovali tři metody krokové regrese (backward selection, forward selection a stepwise regression) a ilustrovali aplikaci oněch metod na konkrétních datech.

Praktická část práce je věnována Monte Carlo simulacím, sestrojeným v statistickém softwaru R. V rámci simulačních studií jsme prozkoumali vybrané vlastnosti regresních modelů, získaných za použití metod krokové regrese. Pro každou z tří metod krokové regrese jsme vypočítali vychýlení odhadů regresních parametrů, relativní četnosti detekce správné konfigurace vektoru regresních parametrů a chybu predikce. Taky jsme se snažili popsat vývoj těchto vlastností v závislosti na jednotlivých parametrech regresních modelů. Jako referenční metodu pro hodnocení a porovnání výsledků jsme zvolili plný regresní model.

Dospěli jsme k závěru, že metody krokové regrese na rozdíl od plného modelu poskytují vychýlené odhady. Vychýlení odhadů souvisí se silou testů hypotéz, které provádíme při budování regresního modelu pomocí metod krokové regrese. Výsledky testů hypotéz jsou citlivé na změnu parametrů regresních modelů. Chyby, které vznikají při testování hypotéz, ovlivňují taky výslednou konfiguraci odhadnutých parametrů. Z tohoto důvodu z hlediska testů hypotéz je lepší preferovat plný model.

Co se týče predikce, výsledky pro metody krokové regrese jsou celkově lepší než u plného modelu. Kroková regrese by mohla být preferována, pokud při výstavbě regresního modelu je nutné vybírat z mnoha nezávisle proměnných s malým počtem pozorování a větším vlivem náhodné složky. Došli jsme k závěru, že z hlediska predikce je možné metody krokové regrese doporučit před predikcí s plným modelem.

Správnost prezentovaných výsledků je možné ověřit pomocí zdrojových kódů, které jsou obsahem elektronické přílohy.

Reference

[1] Anděl, J. (1998). *Statistické metody.* Druhé přepracované vydání. Matfyzpress, Praha. ISBN 80-85863-27-8.

[2] Anděl, J. (2007). *Základy matematické statistiky.* Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

[3] Dupač, V. a Hušková, M. (2005). *Pravděpodobnost a matematická statistika.* Nakladatelství Karolinum, Praha. ISBN 978-80-246-0009-3.

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

[5] Zvára, K. (1989). *Regresní analýza.* Academia, Praha. ISBN 80-200-0125-5.

[6] Harrell, F.E. (2001). *Regression Modeling Strategies (With Applications to Linear Models, Logistic Regression, and Survival Analysis).* Springer. ISBN 978-1-4757-3462-1.

[7] Tibshirani, R. (1996). Regression Shrinkage and selection via the Lasso. *Journal of the Royal Statistical society*, 58(1), 267-288

[8] Derksen S., Keselman H.J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2), 265-282.

[9] Lee, E. T. (1974). A Computer Program for Linear Logistic Regression Analysis. *Computer Programs in Biomedicine*, 4(2), 80-92.

Seznam obrázků

1 *Znázornění vychýlení odhadů v plném modelu* 24
2 *Znázornění vychýlení odhadů v závislosti na parametrech* σ, n, ρ 26
3 *Znázornění vychýlení odhadů v závislosti na parametrech PtoStay a PtoEnter* . . 27
4 *Znázornění vývoje relativní četnosti v závislosti na parametrech* σ, n *a* ρ 29
5 *Znázornění vývoje relativní četnosti v závislosti na parametrech PtoStay a PtoEnter* 30
6 *Znázornění vývoje relativní četnosti v závislosti na parametru* σ 31
7 *Znázornění vývoje relativní četnosti v závislosti na parametru* n 32
8 *Znázornění vývoje relativní četnosti v závislosti na parametru* ρ 32
9 *Znázornění vývoje relativní četnosti v závislosti na parametrech PtoStay a PtoEnter* 34

Seznam tabulek

1 *Variace vektoru regresních parametrů* 21
2 *Výchozí nastavení parametrů* 21
3 *Chyba predikce pro* $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$ 35
4 *Chyba predikce pro* $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$ 35
5 *Chyba predikce pro* $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)$ 36