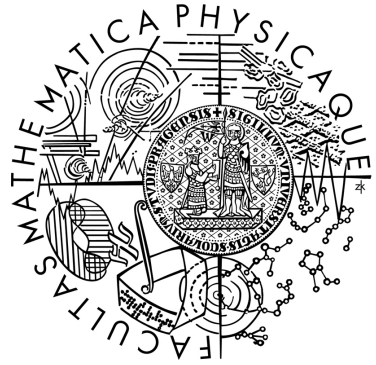
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Andrea Pacáková

Analýza změny od počáteční hodnoty ke konečné

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. Mgr. Michal Kulich, Ph.D.
Studijní program: Matematika
Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

2010

Na tomto místě chci poděkovat svému vedoucímu diplomové práce, panu doc. Mgr. Michalovi Kulichovi, Ph.D., za podporu, trpělivou pomoc, za čas, který této práci věnoval, a za veškeré připomínky, které značně dopomohly k jejímu vzniku.

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V *Praze* dne *16. dubna 2010* *Andrea Pacáková*

Obsah

Název práce: *Analýza změny od počáteční hodnoty ke konečné*
Autor: *Andrea Pacáková*
Katedra: *Katedra pravděpodobnosti a matematické statistiky*
Vedoucí diplomové práce: *doc. Mgr. Michal Kulich, Ph.D.*
e-mail vedoucího: *kulich@karlin.mff.cuni.cz*

Abstrakt: *Tato práce se zabývá srovnáním tří různých odhadů efektu léčby v klinických randomizovaných studiích, jejichž cílem je porovnat změnu v rozdělení určité veličiny mezi dvěma ošetřeními. Uvedené odhady vznikly na základě předpokladu o platnosti nějakého modelu. V práci zjišťujeme vlastnosti těchto odhadů za platnosti každého z daných modelů. Zabýváme se konzistencí odhadů a jejich asymptotickými rozděleními a následně srovnáváme odhady na základě jejich asymptotických rozptylů. Ve většině případů lze srovnání provést obecně a tam, kde to nelze, uvádíme některé konkrétní příklady. Nakonec provádíme simulační studii, která potvrzuje teoretické závěry a rozšiřuje poznatky tam, kde nešly teoretické výpočty obecně provést.*

Klíčová slova: *randomizovaná studie, počáteční a konečná hodnota, odhad efektu léčby, porušení předpokladů modelu*

Title: *Analysis of change from baseline to post-intervention value*
Author: *Andrea Pacáková*
Department: *Department of Probability and Mathematical Statistits*
Supervisor: *doc. Mgr. Michal Kulich, Ph.D.*
Supervisor's e-mail address: *kulich@karlin.mff.cuni.cz*

Abstract: *The aim of the present work is to compare three different estimators of a treatment effect in clinical randomized studies. The purpose of these studies is to compare the change of a distribution of certain variable between two attendances. Mentioned estimators were developed from the assumption of validity of some model. In this work we gather properties of the estimators when each of all given models is valid. We deal with the consistency of the estimators and with their asymptotic distributions and then we compare the estimators on the basis of their asymptotic variances. In the most of cases is possible to make the comparison in general. In the case when it is not possible, we show a few particular examples. Eventually, we accomplish the simulation study, which certifies theoretical conclusions and extends pieces of knowledge in the cases when it was not possible to make theoretical computation in general.*

Keywords: *randomized study, initial and post-intervention value, estimator of a treatment effect, breach of the assumptions of a model*

Kapitola 1

Úvod

V klinických studiích, které slouží k ověření účinnosti nových léčebných postupů, nás zajímá změna rozdělení nějaké náhodné veličiny, například systolického krevního tlaku. Tato veličina je měřena dvakrát - na začátku a na konci studie. Existuje několik metod, jak analyzovat vztah mezi počáteční hodnotou nějaké náhodné veličiny, měřenou na začátku studie, a její hodnotou konečnou, měřenou na konci. Na základě zmíněných postupů analýzy je počítán efekt léčby.

Výše uvedené klinické studie jsou randomizované, to znamená, že jednotky, které pozorujeme (pacienti), jsou náhodně rozděleny do dvou skupin. Pacientům jedné ze skupin je poskytnuta léčba, pacientům druhé skupiny není, dostanou například placebo. Druhá skupina je tedy kontrolní.

Zmíněný problém se objevuje v centru pozornosti mnoha lidí, zejména pokud jde o srovnávání daných metod analýzy, přičemž se názory na optimální přístup velmi liší. Například Stephen Senn v článku Senn (2006) prosazuje metodu ANCOVA (analýza kovariance, viz 3.3 na straně 13) proti metodě SACS (simple analysis of change score = jednoduchá analýza změny stavu, viz 3.2 na straně 11). Diskuzi ale vede na základě toho, že připouští, že střední hodnoty v obou skupinách na začátku nemusí být shodné, což při randomizaci nenastává.

V článku Tu and Gilthorpe (2007) jsou zase uvedena různá řešení problémů, které se vyskytují, chceme-li použít metodu, která je obdobou modelu 3.3 na straně 13. Autoři tohoto článku vedou o nejlepším řešení problému polemiku s Funatogawa (2007).

Cílem této práce je vystavět dané metody od základů a srozumitelně je porovnat.

Kapitola 2

Situace

Pacienti jsou na začátku studie náhodně rozděleni do experimentální skupiny s léčbou E (experimental) a kontrolní skupiny bez léčby C (control). Ve skupině E je n_E pacientů, ve skupině C je n_C pacientů. Předpokládejme, že pravděpodobnost, že je pacient přiřazen do skupiny E, je $\pi \in (0,\, 1)$. Není-li pacient určen do skupiny E, jde do skupiny C. Náhodná veličina, která randomizaci popisuje, je indikátor přiřazení pacienta i do skupiny E, tedy $\mathsf{I}_{[i\epsilon E]}$. Platí pro ni

$$\mathsf{I}_{[i\epsilon E]} \sim \mathsf{Alt}(\pi) \quad \forall i = 1, \ldots, n,$$

kde

$$n = n_C + n_E.$$

Označíme

$$q_n = \frac{n_E}{n_C} = \frac{\sum_{i=1}^{n} \mathsf{I}_{[i\epsilon E]}}{n - \sum_{i=1}^{n} \mathsf{I}_{[i\epsilon E]}}. \tag{2.1}$$

Ze zákona velkých čísel plyne $\frac{1}{n}\sum_{i=1}^{n} \mathsf{I}_{[i\epsilon E]} \xrightarrow{P} \pi, \quad n \longrightarrow \infty$. Spočítáme limitu q_n a označíme ji q:

$$q_n \xrightarrow{P} \frac{\pi}{1-\pi} = q.$$

Předpokládejme, že $q \in (0, \infty)$. Proto když $n_C \longrightarrow \infty$, tak i $n_E \longrightarrow \infty$.

Náhodnou veličinu Y, která nás zajímá, měříme nejprve v čase, který označíme 0, a podruhé v čase 1. Tato dvě měření provádíme u každého pacienta z obou skupin E a C.

Máme tedy nezávislé dvojice pozorování

$$\left(Y_{0i}^{E}, Y_{1i}^{E}\right), \quad i = 1, \ldots, n_E,$$

$$\left(Y_{0i}^{C}, Y_{1i}^{C}\right), \quad i = 1, \ldots, n_C,$$

kde horní index znamená příslušnost pacienta do skupiny, 0 a 1 znamenají čas měření a i je index pacienta v dané skupině.

Randomizace nám zaručuje, že rozdělení náhodné veličiny na počátku je stejné v obou skupinách, tedy

$$\mathcal{L}\left(Y_{0i}^{C}\right) = \mathcal{L}\left(Y_{0i}^{E}\right).$$

Označme

$$\mu_0 = \mathsf{E}\, Y_{0i}^{C} = \mathsf{E}\, Y_{0i}^{E}, \quad \sigma_0^2 = \mathsf{var}\, Y_{0i}^{C} = \mathsf{var}\, Y_{0i}^{E} \in (0, \infty). \tag{2.2}$$

Ve výpočtech budeme předpokládat, že tyto hodnoty známe. V praxi je odhadneme z dat. To, co nás zajímá, je efekt léčby, tedy vlastně nějaký „rozdíl“ v konečných hodnotách ve skupinách E a C, který budeme označovat jako δ. Existují tři různé přístupy k modelování této situace. Na jejich základě je spočítán efekt léčby.

Kapitola 3

Modely a chování odhadů efektu léčby za platnosti správných modelů

3.1 Model I

První přístup se vůbec nezabývá hodnotami na počátku. Konečnou hodnotu bere prostě jako náhodnou veličinu se střední hodnotou, která se liší v obou skupinách. Model I má tedy tvar

$$\begin{aligned} Y_{1i}^{C} &= \mu + \epsilon_i^{I,C}, \\ Y_{1i}^{E} &= \mu + \delta + \epsilon_i^{I,E} \end{aligned}$$

za podmínek

$$\begin{gathered} \mathsf{E}\,\epsilon_i^{I,C} = 0, \quad i = 1, \ldots, n_C, \\ \mathsf{E}\,\epsilon_i^{I,E} = 0, \quad i = 1, \ldots, n_E, \\ \mathsf{var}\,\epsilon_i^{I,C} = \sigma_{I,C}^2, \quad i = 1, \ldots, n_C, \qquad \mathsf{var}\,\epsilon_i^{I,E} = \sigma_{I,E}^2, \quad i = 1, \ldots, n_E \end{gathered}$$

a vzájemné nezávislosti náhodných chyb ϵ.

Důležité zde je, že náhodné chyby mohou nějak záviset na počáteční hodnotě a nejspíše i závisí, protože na počáteční hodnotě zřejmě závisí i hodnota konečná. Nic o této závislosti ale nevíme. Tento model je velmi obecný a platí vždy v tom smyslu, že se každá náhodná veličiná dá napsat jako součet její střední hodnoty a jiné veličiny se střední hodnotou nulovou.

Odhad δ získaný z modelu I je

$$\hat{\delta}^I = \bar{Y}_1^E - \bar{Y}_1^C, \tag{3.1}$$

kde $\bar{Y}_1^E = \frac{1}{n_E}\sum_{i=1}^{n_E} Y_{1i}^E$ a $\bar{Y}_1^C = \frac{1}{n_C}\sum_{i=1}^{n_C} Y_{1i}^C$.

3.1.1 Odhad efektu léčby $\hat{\delta}^I$ za platnosti modelu I

Odhad $\hat{\delta}^I$ je za platnosti modelu I nevychýlený. Spočítáním jeho rozptylu

$$\begin{aligned}\mathsf{var}\,\hat{\delta}^I &= \mathsf{var}\,\left(\bar{Y}_1^E - \bar{Y}_1^C\right)\\ &= \mathsf{var}\,\bar{Y}_1^C + \mathsf{var}\,\bar{Y}_1^E = \frac{1}{n_C^2}\sum_{i=1}^{n_C}\mathsf{var}\,Y_{1i}^C + \frac{1}{n_E^2}\sum_{i=1}^{n_E}\mathsf{var}\,Y_{1i}^E = \frac{1}{n_C}\sigma_{I,C}^2 + \frac{1}{n_E}\sigma_{I,E}^2\end{aligned}$$

zjistíme, že je i konzistentní. Při $n_C \longrightarrow \infty$ totiž konverguje výraz na pravé straně do nuly a konzistence plyne z věty 7.6 Anděl (2007). Ve výpočtu jsme použili předpoklad nezávislosti jednotlivých pozorování.

Chceme-li testovat hypotézu H_0 proti alternativě H_1

$$H_0 : \delta = 0 \quad \text{vs.} \quad H_1 : \delta \neq 0,$$

jde vlastně o dvouvýběrový test porovnávající střední hodnoty za předpokladu, že nejsou shodné rozptyly. Použijeme Welchovu testovou statistiku z knihy Anděl (2003)

$$t_I = \frac{\bar{Y}_1^E - \bar{Y}_1^C}{\sqrt{\frac{S_{I,C}^2}{n_C} + \frac{S_{I,E}^2}{n_E}}},$$

kde

$$S_{I,C}^2 = \frac{1}{n_C - 1}\sum_{i=1}^{n_C}\left(Y_{1i}^C - \bar{Y}_1^C\right)^2, \quad S_{I,E}^2 = \frac{1}{n_E - 1}\sum_{i=1}^{n_E}\left(Y_{1i}^E - \bar{Y}_1^E\right)^2. \tag{3.2}$$

Jsou-li n_C a n_E hodně velká, můžeme pro nalezení kritického oboru použít asymptotickou normalitu statistiky t_I. Z centrální limitní věty totiž víme, že $\sqrt{n_C}(\bar{Y}_1^C - \mu) \xrightarrow{d} \mathsf{N}\left(0, \sigma_{I,C}^2\right)$ (toto je schematický zápis, který znamená konvergenci v distribuci k náhodné veličině, která má normální rozdělení s parametry 0 a $\sigma_{I,C}^2$, budeme ho používat i níže) a také $\sqrt{n_E}(\bar{Y}_1^E - (\mu+\delta)) \xrightarrow{d} \mathsf{N}\left(0, \sigma_{I,E}^2\right)$. Z toho, že $q_n \longrightarrow q \in (0,\infty)$ a nezávislosti jednotlivých pozorování plyne, že

$$\sqrt{n_C}\left(\begin{pmatrix}\bar{Y}_1^C\\ \bar{Y}_1^E\end{pmatrix} - \begin{pmatrix}\mu\\ \mu+\delta\end{pmatrix}\right) \xrightarrow{d} \mathsf{N}\left(0, \begin{pmatrix}\sigma_{I,C}^2 & 0\\ 0 & \frac{\sigma_{I,E}^2}{q}\end{pmatrix}\right).$$

Odtud za platnosti H_0, kdy $\delta = 0$, platí

$$\frac{\bar{Y}_1^E - \bar{Y}_1^C}{\sqrt{\frac{\sigma_{I,C}^2}{n_C} + \frac{\sigma_{I,E}^2}{q \cdot n_C}}} \xrightarrow{d} \mathsf{N}(0,1). \tag{3.3}$$

Dále protože jsou (3.2) konzistentní odhady $\sigma_{I,C}^2$ a $\sigma_{I,E}^2$, ze Sluckého věty vyplývá, že

$$t_I \xrightarrow{d} \mathsf{N}(0,1).$$

Hypotézu H_0 tedy zamítneme na hladině α, pokud $|t_I| \geq u_{1-\frac{\alpha}{2}}$.

Intervalový odhad pro δ s pravděpodobností pokrytí $1-\alpha$ založený na t_I je

$$\left(\hat{\delta}^I - u_{1-\frac{\alpha}{2}} \sqrt{\frac{S_{I,C}^2}{n_C} + \frac{S_{I,E}^2}{n_E}},\ \hat{\delta}^I + u_{1-\frac{\alpha}{2}} \sqrt{\frac{S_{I,C}^2}{n_C} + \frac{S_{I,E}^2}{n_E}} \right).$$

Protože $\hat{\delta}^I$ je nestranný odhad δ, lze vyhledem k tomu, že $n/n_C = (1+q_n) \longrightarrow 1+q$ konvergenci v distribuci (3.3) pro skutečné (i nenulové) δ přepsat jako asymptotickou normalitu odhadu $\hat{\delta}^I$:

$$\sqrt{n}\left(\hat{\delta}^I - \delta\right) \xrightarrow{d} \mathsf{N}\left(0, V_{\delta^I}\right),$$

kde

$$V_{\delta^I} = (q+1) \cdot \left(\sigma_{I,C}^2 + \frac{1}{q} \cdot \sigma_{I,E}^2 \right). \tag{3.4}$$

3.2 Model II

Druhý model porovnává, jak se v jednotlivých skupinách liší změna mezi konečnou a počáteční hodnotou. Předpokládá, že změna nezávisí na počáteční hodnotě, a efekt léčby δ bere jako rozdíl ve středních hodnotách této změny. Model II zapíšeme

$$Y_{1i}^C - Y_{0i}^C = \gamma + \epsilon_i^{II,C},$$
$$Y_{1i}^E - Y_{0i}^E = \gamma + \delta + \epsilon_i^{II,E},$$

kde

$$\mathsf{E}\,\epsilon_i^{II,C} = 0,\ \mathsf{E}\,(\epsilon_i^{II,C} | Y_{0i}^C) = 0 \quad i = 1, \ldots, n_C,$$
$$\mathsf{E}\,\epsilon_i^{II,E} = 0,\ \mathsf{E}\,(\epsilon_i^{II,E} | Y_{0i}^E) = 0 \quad i = 1, \ldots, n_E,$$
$$\mathsf{var}\,\epsilon_i^{II,C} = \sigma_{II,C}^2, \quad \mathsf{var}\,\epsilon_i^{II,E} = \sigma_{II,E}^2$$

a $\epsilon_i^{II,C}$, $\epsilon_i^{II,E}$ jsou vzájemně nezávislé. Dále protože změna nezávisí na počáteční hodnotě, nezávisí na ní ani náhodná chyba $\epsilon_i^{II,C}$ nebo $\epsilon_i^{II,E}$.

Model II je velmi specifický. To, že v něm změna mezi konečnou a počáteční hodnotou na počáteční hodnotě nezávisí, přesně určuje tvar závislosti konečné hodnoty na počáteční (stačí převést počáteční hodnotu doprava). Pokud ve skutečnosti změna na počáteční hodnotě závisí, neboli je závislost konečné a počáteční hodnoty jiná, než ji určuje model II, promítne se to do předpokladu nezávislosti náhodné chyby na počáteční hodnotě, který bude porušen.

Odhad δ z modelu II je

$$\hat{\delta}^{II} = \left(\bar{Y}_1^E - \bar{Y}_0^E\right) - \left(\bar{Y}_1^C - \bar{Y}_0^C\right),$$

kde $\bar{Y}_1^E = \frac{1}{n_E}\sum_{i=1}^{n_E} Y_{1i}^E$, $\bar{Y}_1^C = \frac{1}{n_C}\sum_{i=1}^{n_C} Y_{1i}^C$, $\bar{Y}_0^E = \frac{1}{n_E}\sum_{i=1}^{n_E} Y_{0i}^E$ a $\bar{Y}_0^C = \frac{1}{n_C}\sum_{i=1}^{n_C} Y_{0i}^C$.

3.2.1 Odhad efektu léčby $\hat{\delta}^{II}$ za platnosti modelu II

Platí-li model II, je $\mathsf{E}\,\hat{\delta}^{II} = \delta$, takže jde opět o nestranný odhad a protože

$$\mathsf{var}\,\hat{\delta}^{II} = \frac{1}{n_C}\sigma_{II,C}^2 + \frac{1}{n_E}\sigma_{II,E}^2,$$

jde také o odhad konzistentní.

Test hypotézy o nulovém efektu léčby

$$H_0 : \delta = 0 \quad \mathsf{vs.} \quad H_1 : \delta \neq 0$$

je shodný s dvouvýběrovým testem o shodě středních hodnot v náhodných výběrech $(Y_{1i}^E - Y_{0i}^E, i = i, \ldots, n_E)$ a $(Y_{1i}^C - Y_{0i}^C, i = i, \ldots, n_C)$ při různých rozptylech. Testová statistika je opět Welchova

$$t_{II} = \frac{\left(\bar{Y}_1^E - \bar{Y}_0^E\right) - \left(\bar{Y}_1^C - \bar{Y}_0^C\right)}{\sqrt{\frac{S_{II,C}^2}{n_C} + \frac{S_{II,E}^2}{n_E}}},$$

kde

$$S_{II,C}^2 = \frac{1}{n_C - 1}\sum_{i=1}^{n_C}\left(Y_{1i}^C - Y_{0i}^C - \left(\bar{Y_1^C} - \bar{Y_0^C}\right)\right)^2,$$

$$S_{II,E}^2 = \frac{1}{n_E - 1}\sum_{i=1}^{n_E}\left(Y_{1i}^E - Y_{0i}^E - \left(\bar{Y_1^E} - \bar{Y_0^E}\right)\right)^2.$$

Pro dostatečně velká n_C a n_E lze pro nalezení kritického oboru opět použít asymptotickou normalitu. H_0 zamítáme, pokud $|t_{II}| \geq u_{1-\frac{\alpha}{2}}$.

$1-\alpha$ procentní interval spolehlivosti pro δ založený na t_{II} je

$$\left(\hat{\delta}^{II} - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2_{II,C}}{n_C} + \frac{S^2_{II,E}}{n_E}},\ \hat{\delta}^{II} + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2_{II,C}}{n_C} + \frac{S^2_{II,E}}{n_E}} \right).$$

Jako v případě modelu I lze asymptotickou normalitu odhadu $\hat{\delta}^{II}$ zapsat takto:

$$\sqrt{n}\left(\hat{\delta}^{II} - \delta_0\right) \xrightarrow{d} \mathsf{N}\left(0, V_{\delta^{II}}\right),$$

kde

$$V_{\delta^{II}} = (q+1)\left(\sigma^2_{II,C} + \frac{1}{q} \cdot \sigma^2_{II,E}\right).$$

3.3 Model III

Model III na rozdíl od modelu II, který vlastně předpokládá lineární závislost konečné hodnoty na počáteční se směrnicí 1, přechází k obecné lineární závislosti. Zapíšeme ho

$$Y^C_{1i} = \alpha + \beta \cdot Y^C_{0i} + \epsilon^{III,C}_i,$$
$$Y^E_{1i} = \alpha + \beta \cdot Y^E_{0i} + \delta + \epsilon^{III,E}_i,$$

za podmínek

$$\mathsf{E}\epsilon^{III,C}_i = 0,\ \mathsf{E}\left(\epsilon^{III,C}_i | Y^C_{0i}\right) = 0, \quad i = 1, \ldots, n_C,$$
$$\mathsf{E}\,\epsilon^{III,E}_i = 0,\ \mathsf{E}\left(\epsilon^{III,E}_i | Y^E_{0i}\right) = 0, \quad i = 1, \ldots, n_E,$$
$$\mathsf{var}\,\epsilon^{III,C}_i = \sigma^2_{III,C}, \quad \mathsf{var}\,\epsilon^{III,E}_i = \sigma^2_{III,E}$$

a $\epsilon^{III,C}_i$, $\epsilon^{III,E}_i$ jsou vzájemně nezávislé a nezávisí ani na Y^C_{0i}, Y^E_{0i}.

Nezávislost náhodné chyby na počáteční hodnotě je opět splněna jen v případě, že skutečná závislost konečné hodnoty na počáteční je lineární. Snadnou úpravou se model III může přepsat na tvar, ze kterého je vidět, že je ekvivalentní s lineární závislostí změny na počáteční hodnotě, o které se mluví v článku Tu and Gilthorpe (2007). Model III je obecnější než model II a méně obecný než model I.

Odhad δ z modelu III je klasický odhad metodou nejmenších čtverců, avšak na základě modelu s různými rozptyly. Abychom ho mohli spočítat, model III si přepíšeme následovně

$$Y_{1i} = \alpha + \beta \cdot Y_{0i} + \delta \cdot \mathsf{I}_{[i \epsilon E]} + \epsilon^{III}_i, \quad i = 1, \ldots, n, \tag{3.5}$$

kde $\mathsf{E}\,\epsilon_i^{III} = 0$, $\mathsf{E}\,(\epsilon_i^{III}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}) = 0$, $\mathsf{var}\,\left(\epsilon_i^{III}|\mathsf{I}_{[i\epsilon E]}\right) = \sigma^2_{III,C} + \mathsf{I}_{[i\epsilon E]} \cdot \left(\sigma^2_{III,E} - \sigma^2_{III,C}\right)$ a ϵ_i^{III} nezávisí na počáteční hodnotě.

Regresní matice X a vektor odezvy $\mathbb{Y}$ vypadají takto

$$X = \begin{pmatrix} 1 & Y_{01} & \mathsf{I}_{[1\epsilon E]} \\ \vdots & \vdots & \vdots \\ 1 & Y_{0n} & \mathsf{I}_{[n\epsilon E]} \end{pmatrix}, \quad \mathbb{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n} \end{pmatrix}.$$

Odhad parametrů modelu III metodou nejmenších čtverců je

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\delta}^{III} \end{pmatrix} = (X^T X)^{-1} X^T \mathbb{Y},$$

kde

$$X^T X = \begin{pmatrix} n & \sum_{i=i}^{n} Y_{0i} & n_E \\ \sum_{i=1}^{n} Y_{0i} & \sum_{i=1}^{n} Y_{0i}^2 & \sum_{i\epsilon E} Y_{0i} \\ n_E & \sum_{i\epsilon E} Y_{0i} & n_E \end{pmatrix}, \quad X^T \mathbb{Y} = \begin{pmatrix} \sum_{i=i}^{n} Y_{1i} \\ \sum_{i=i}^{n} Y_{0i} \cdot Y_{1i} \\ \sum_{i\epsilon E} Y_{1i} \end{pmatrix}.$$

Pro samotný výpočet odhadu parametru δ je třeba zinvertovat matici $X^T X$. Je to zdlouhavý výpočet, který vede na složitý vzorec. Pro jeho výpočet se dá použít například věta A10 z knihy Anděl (2007):

Věta 1. *[Anděl (2007)] Nechť* $\begin{pmatrix} A & B \\ B^T & D \end{pmatrix}$ *je symetrická pozitivně definitní bloková matice taková, že bloky* A *a* D *jsou čtvercové. Pak je také* $Q = A - BD^{-1}B^T$ *regulární a pozitivně definitní a platí*

$$\begin{pmatrix} A & B \\ B^T & D \end{pmatrix}^{-1} = \begin{pmatrix} Q^{-1} & -Q^{-1}BD^{-1} \\ -D^{-1}B^TQ^{-1} & D^{-1} + D^{-1}B^TQ^{-1}BD^{-1} \end{pmatrix}.$$

Důkaz věty 1 je v knize Anděl (2007).

Nepotřebujeme celou inverzi matice $X^T X$, pro získání odhadu $\hat{\delta}^{III}$ nám stačí poslední řádek, který následně vynásobíme s vektorem $X^T \mathbb{Y}$. Matici $X^T X$ rozdělíme na bloky takto: $\left(\begin{array}{ccc|c} n & \sum_{i=i}^{n} Y_{0i} & & n_E \\ \sum_{i=1}^{n} Y_{0i} & \sum_{i=1}^{n} Y_{0i}^2 & & \sum_{i\epsilon E} Y_{0i} \\ \hline n_E & \sum_{i\epsilon E} Y_{0i} & & n_E \end{array}\right)$. Odhad $\hat{\delta}^{III}$ vyšel jako dlouhý výraz, uvádět ho zde nebudeme.

3.3.1 Odhad efektu léčby $\hat{\delta}^{III}$ za platnosti modelu III

$\hat{\delta}^{III}$ je jakožto odhad MNČ za platnosti modelu III nestranný.

Nejprve uvedeme větu, která popisuje asymptotické rozdělení odhadů regresních parametrů za předpokladu, že jsou v klasickém regresním modelu porušeny předpoklady o shodě rozptylů. V následujících úvahách označíme $\boldsymbol{\beta}$ vektor regresních parametrů, je třeba odlišit $\boldsymbol{\beta}$ od β v našem modelu III.

Věta 2. *Mějme nezávislé, stejně rozdělené náhodné vektory* $(\mathbb{X}_i, Y_i), \quad i = 1, \ldots, n$. *Nechť platí* $\mathsf{E}\,[Y_i|\mathbb{X}_i] = \mathbb{X}_i^T\boldsymbol{\beta}_0$ *a nechť mají* $\mathbb{X}_i$ *konečné čtvrté momenty a* Y_i *konečné rozptyly. Označme*

$$U_i(\boldsymbol{\beta}) = \mathbb{X}_i \left(Y_i - \mathbb{X}_i^T\boldsymbol{\beta}\right).$$

Nechť je $\hat{\boldsymbol{\beta}}$ *odhad parametru* $\boldsymbol{\beta}$ *takový, který řeší rovnici*

$$\sum_{i=1}^{n} U_i(\boldsymbol{\beta}) = 0. \tag{3.6}$$

Potom je tento odhad konzistentní a platí pro něj

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathsf{N}\left(0, D^{-1}VD^{-1}\right), \tag{3.7}$$

kde

$$D = \mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T, \quad V = \mathsf{var}\,U_i(\boldsymbol{\beta}_0).$$

Důkaz. Konzistence odhadu plyne z toho, že je $\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum \mathbb{X}_i\mathbb{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum \mathbb{X}_i Y_i\right)$ spojitou transformací náhodných veličin, které splňují zákon velkých čísel.

Ukážeme, že platí asymptotické rozdělení $\hat{\boldsymbol{\beta}}$. Pro $U_i(\boldsymbol{\beta}_0)$ platí

$$\mathsf{E}\,U_i(\boldsymbol{\beta}_0) = \mathsf{E}\,\mathsf{E}\,[U_i(\boldsymbol{\beta}_0)|\mathbb{X}_i] = \mathsf{E}\,\left(\mathbb{X}_i\mathbb{X}_i^T\boldsymbol{\beta}_0 - \mathbb{X}_i\mathbb{X}_i^T\boldsymbol{\beta}_0\right) = 0$$

a

$$-\mathsf{E}\,\frac{\partial U_i}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = \mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T \quad \forall\boldsymbol{\beta}.$$

Z mnohoroměrné Lévyho-Lindebergovy centrální limitní věty plyne, že

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(\boldsymbol{\beta}_0) \xrightarrow{d} N\left(0, V\right), \tag{3.8}$$

a ze zákona velkých čísel pro stejně rozdělené náhodné veličiny plyne, že

$$-\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_i}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right) \xrightarrow{P} \mathsf{E}\,\mathbb{X}_i\mathbb{X}_i{}^T. \tag{3.9}$$

Pomocí konečného Taylorova rozvoje rozepíšeme

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(\hat{\boldsymbol{\beta}}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(\boldsymbol{\beta}_0) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_i}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\cdot\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right).$$

Vzhledem k tomu, že pro $\hat{\boldsymbol{\beta}}$ je výraz na levé straně nulový ($\hat{\boldsymbol{\beta}}$ řeší rovnici (3.6)), platí, že

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right) = -\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_i}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(\boldsymbol{\beta}_0).$$

Z (3.8) a (3.9) plyne tvrzení. □

Chceme-li zjistit asymptotické rozdělení odhadu $\hat{\delta}^{III}$, použijeme věty 2 a 1. Protože potřebujeme jen pravý dolní prvek asymptotické varianční matice, stačí nám spočítat matici V a poslední sloupec a poslední řádek matice D^{-1}. Rozptyl rozložíme

$$V = \mathsf{var}\,U_i(\boldsymbol{\beta}_0) = \mathsf{var}\,\mathsf{E}\left[U_i(\boldsymbol{\beta}_0)\,|\mathbb{X}_i\right] + \mathsf{E}\,\mathsf{var}\left[U_i(\boldsymbol{\beta}_0)\,|\mathbb{X}_i\right] = 0 + \mathsf{E}\,\mathsf{var}\left[U_i(\boldsymbol{\beta}_0)\,|\mathbb{X}_i\right],$$

je tedy roven střední hodnotě podmíněného rozptylu.

Nyní se vraťme konkrétně k našemu modelu III. Rozepišme $U_i(\boldsymbol{\beta}_0)$ ve tvaru (3.5), kdy vektor parametrů $\boldsymbol{\beta}_0 = (\alpha,\beta,\delta)^T$,

$$U_i(\alpha,\beta,\delta) = \begin{pmatrix} Y_{1i}-\alpha-\beta Y_{0i}-\delta \mathsf{I}_{[i\epsilon E]} \\ Y_{0i}Y_{1i}-\alpha Y_{0i}-\beta Y_{0i}^2-\delta Y_{0i}\mathsf{I}_{[i\epsilon E]} \\ \mathsf{I}_{[i\epsilon E]}Y_{1i}-\alpha \mathsf{I}_{[i\epsilon E]}-\beta Y_{01}\mathsf{I}_{[i\epsilon E]}-\delta \mathsf{I}_{[i\epsilon E]} \end{pmatrix}. \tag{3.10}$$

Nejprve spočítáme podmíněnou varianční matici, potom její střední hodnotu. Prvky vektoru $U_i(\alpha,\beta,\delta)$ budeme značit číslem v závorce psaným dolním indexem. Platí

$$\begin{aligned}
\mathsf{var}\left[U_{i(1)}(\alpha,\beta,\delta)|Y_{0i},\mathsf{I}_{[i\epsilon E]}\right] &= \sigma^2_{III,E}\mathsf{I}_{[i\epsilon E]} + \sigma^2_{III,C}(1-\mathsf{I}_{[i\epsilon E]}) \\
&= \mathsf{I}_{[i\epsilon E]}(\sigma^2_{III,E}-\sigma^2_{III,C}) + \sigma^2_{III,C}, \\
\mathsf{var}\left[U_{i(2)}(\alpha,\beta,\delta)|Y_{0i},\mathsf{I}_{[i\epsilon E]}\right] &= Y_{0i}^2\left(\mathsf{I}_{[i\epsilon E]}(\sigma^2_{III,E}-\sigma^2_{III,C}) + \sigma^2_{III,C}\right), \\
\mathsf{var}\left[U_{i(2)}(\alpha,\beta,\delta)|Y_{0i},\mathsf{I}_{[i\epsilon E]}\right] &= \mathsf{I}_{[i\epsilon E]}\left(\mathsf{I}_{[i\epsilon E]}(\sigma^2_{III,E}-\sigma^2_{III,C}) + \sigma^2_{III,C}\right) \\
&= \mathsf{I}_{[i\epsilon E]}\sigma^2_{III,E}, \\
\mathsf{cov}\left[(U_{i(1)}(\alpha,\beta,\delta),U_{i(2)}(\alpha,\beta,\delta))|Y_{0i},\mathsf{I}_{[i\epsilon E]}\right] &= Y_{0i}\left(\mathsf{I}_{[i\epsilon E]}(\sigma^2_{III,E}-\sigma^2_{III,C}) + \sigma^2_{III,C}\right), \\
\mathsf{cov}\left[(U_{i(2)}(\alpha,\beta,\delta),U_{i(3)}(\alpha,\beta,\delta))|Y_{0i},\mathsf{I}_{[i\epsilon E]}\right] &= Y_{0i}\mathsf{I}_{[i\epsilon E]}\left(\mathsf{I}_{[i\epsilon E]}(\sigma^2_{III,E}-\sigma^2_{III,C}) + \sigma^2_{III,C}\right), \\
\mathsf{cov}\left[(U_{i(1)}(\alpha,\beta,\delta),U_{i(3)}(\alpha,\beta,\delta))|Y_{0i},\mathsf{I}_{[i\epsilon E]}\right] &= \mathsf{I}_{[i\epsilon E]}\sigma^2_{III,E}.
\end{aligned}$$

Varianční matice $U_i(\beta_0)$ matice má tedy tvar

$$\begin{aligned}
&\mathsf{var}\, U_i(\beta_0) = \mathsf{E}\,\mathsf{var}\,\left[U_i\left(\beta_0\right)|\mathbb{X}_i\right] \\
&= \begin{pmatrix} \pi(\sigma^2_{III,E} - \sigma^2_{III,C}) + \sigma^2_{III,C} & \mu_0\left(\pi(\sigma^2_{III,E} - \sigma^2_{III,C}) + \sigma^2_{III,C}\right) & \pi\sigma^2_{III,E} \\ \mu_0\left(\pi(\sigma^2_{III,E} - \sigma^2_{III,C}) + \sigma^2_{III,C}\right) & \mathsf{E}\,Y^2_{0i}\left(\pi(\sigma^2_{III,E} - \sigma^2_{III,C}) + \sigma^2_{III,C}\right) & \mu_0\pi\sigma^2_{III,E} \\ \pi\sigma^2_{III,E} & \mu_0\pi\sigma^2_{III,E} & \pi\sigma^2_{III,E} \end{pmatrix}.
\end{aligned}$$

Dále spočítejme poslední řádek a poslední sloupec matice $D^{-1} = (\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T)^{-1}$. Matice $D = \mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T$ je rovna

$$\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T = \mathsf{E}\begin{pmatrix} 1 & Y_{0i} & \mathsf{I}_{[i\epsilon E]} \\ Y_{0i} & Y^2_{0i} & \mathsf{I}_{[i\epsilon E]}Y_{0i} \\ \mathsf{I}_{[i\epsilon E]} & \mathsf{I}_{[i\epsilon E]}Y_{0i} & \mathsf{I}_{[i\epsilon E]} \end{pmatrix} = \left(\begin{array}{cc|c} 1 & \mu_0 & \pi \\ \mu_0 & \mathsf{E}\,Y^2_{0i} & \pi\mu_0 \\ \hline \pi & \pi\mu_0 & \pi \end{array}\right). \tag{3.11}$$

Poslední matice je rozdělená na bloky tak, jak je použijeme pro výpočet inverze pomocí věty 1. Spočítejme matici Q z bloků A, B, D z této věty,

$$Q = A - BDB^T = \begin{pmatrix} 1 & \mu_0 \\ \mu_0 & \mathsf{E}\,Y^2_{0i} \end{pmatrix} - \begin{pmatrix} \pi \\ \pi\mu_0 \end{pmatrix}\frac{1}{\pi}(\pi, \pi\mu_0) = \begin{pmatrix} 1-\pi & \mu_0(1-\pi) \\ \mu_0(1-\pi) & \mathsf{E}\,Y^2_{0i} - \pi\mu_0^2 \end{pmatrix}$$

a označme $\eta = 1 - \pi$. Dále potřebujeme inverzi matice Q,

$$Q^{-1} = \frac{1}{|Q|}\begin{pmatrix} \mathsf{E}\,Y^2_{0i} - \pi\mu_0^2 & -\mu_0\eta \\ -\mu_0\eta & \eta \end{pmatrix} = \frac{1}{\eta(\mathsf{E}\,Y^2_{0i} - \mu^2)}\begin{pmatrix} \mathsf{E}\,Y^2_{0i} - \pi\mu_0^2 & -\mu_0\eta \\ -\mu_0\eta & \eta \end{pmatrix}.$$

Prvky inverzní matice spočítáme následovně:

$$\begin{aligned}
D^{-1} + D^{-1}B^TQ^{-1}BD^{-1} &= \frac{1}{\pi} + \frac{1}{\pi}(\pi, \mu_0\pi)\frac{1}{\eta(\mathsf{E}\,Y^2_{0i} - \mu^2)} \\
&\quad \cdot \begin{pmatrix} \mathsf{E}\,Y^2_{0i} - \pi\mu_0^2 & -\mu_0\eta \\ -\mu_0\eta & \eta \end{pmatrix}\begin{pmatrix} \pi \\ \pi\mu_0 \end{pmatrix}\frac{1}{\pi} = \frac{1}{\pi} + \frac{1}{\eta}, \\
-D^{-1}B^TQ^{-1} &= \ldots = \left(\frac{1}{\eta}, 0\right), \\
-Q^{-1}B^TD^{-1} &= \begin{pmatrix} -\frac{1}{\eta} \\ 0 \end{pmatrix}.
\end{aligned}$$

Pro matici D tedy platí

$$D^{-1} = \left(\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T\right)^{-1} = \begin{pmatrix} \cdot & \cdot & -\frac{1}{\eta} \\ \cdot & \cdot & 0 \\ -\frac{1}{\eta} & 0 & \frac{1}{\pi} + \frac{1}{\eta} \end{pmatrix}.$$

Potřebujeme pravý dolní prvek matice $\left(\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T\right)^{-1}\mathsf{var}\,U_i(\beta_0)\left(\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T\right)^{-1}$, spočítáme tedy nejprve poslední řádek matice $\left(\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T\right)^{-1}\mathsf{var}\,U_i(\beta_0)$ vynásobením posledního řádku matice $\left(\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T\right)^{-1}$ se sloupci matice $\mathsf{var}\,U_i(\beta_0)$. Navíc vhledem k 0 v posledním sloupci matice $\left(\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T\right)^{-1}$ nepotřebujeme prostřední prvek tohoto řádku. Vyjde nám

$$\left[\left(\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T\right)^{-1}\mathsf{var}\,U_i(\beta_0)\right]_{3,} = \left(\sigma^2_{III,E} - \sigma^2_{III,C}, \cdot, \sigma^2_{III,E}\right)$$

a nakonec dostaneme

$$\left[\left(\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T\right)^{-1}\mathsf{var}\,U_i(\beta_0)\left(\mathsf{E}\,\mathbb{X}_i\mathbb{X}_i^T\right)^{-1}\right]_{3,3} = \frac{1}{\pi}\sigma^2_{III,E} + \frac{1}{\eta}\sigma^2_{III,C}.$$

Pro odhad $\hat{\delta}^{III}$ a skutečné δ tedy platí

$$\sqrt{n}\left(\hat{\delta}^{III} - \delta\right) \xrightarrow{d} \mathbb{N}\left(0, V_\delta^{III}\right),$$

kde

$$V_{\delta^{III}} = \frac{1}{\pi}\sigma^2_{III,E} + \frac{1}{1-\pi}\sigma^2_{III,C} = (q+1)\left(\frac{1}{q}\sigma^2_{III,E} + \sigma^2_{III,C}\right). \tag{3.12}$$

Chceme-li testovat hypotézu

$$H_0 : \delta = 0 \quad \text{vs.} \quad H_1 : \delta \neq 0,$$

musíme odhadnout rozptyl $V_{\delta^{III}}$. Matici D odhadneme pomocí $\hat{D} = \frac{1}{n}\sum_{i=1}^n \mathbb{X}_i\mathbb{X}_i^T$, matici V odhadneme empirickou varianční maticí $\hat{V} = \frac{1}{n}\sum_{i=1}^n U_i(\hat{\alpha},\hat{\beta},\hat{\delta})U_i(\hat{\alpha},\hat{\beta},\hat{\delta})^T$ a označíme $\hat{V}_\delta^{III} = \left(\hat{D}^{-1}\hat{V}\hat{D}^{-1}\right)_{3,3}$.

Použijeme testovou statistiku

$$t_{III} = \frac{\sqrt{n}\hat{\delta}^{III}}{\hat{V}_{\delta^{III}}},$$

její rozdělení je asymptoticky normované normální, což plyne opět z konzistence odhadu rozptylu, zákona velkých čísel a ze Sluckého věty. Hypotézu tedy zamítneme na hladině α, pokud $|t_{III}| \geq u_{1-\frac{\alpha}{2}}$.

$1-\alpha$ procentní interval spolehlivosti pro δ založený na t_{III} je

$$\left(\hat{\delta}^{III} - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{V}_\delta^{III}}{n}},\ \hat{\delta}^{III} + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{V}_\delta^{III}}{n}}\right).$$

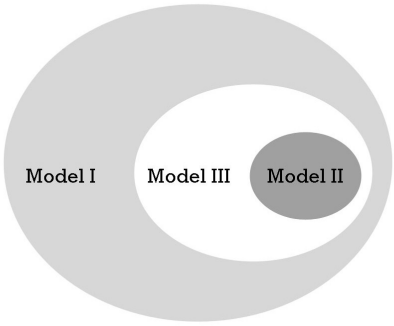
Kapitola 4

Chování odhadů za platnosti jiného modelu, než ze kterého vznikly

V praxi se může stát, že platí jeden z uvedených tří modelů, ale my se rozhodneme použít odhad efektu léčby na základě modelu jiného. V následujících odstavcích prozkoumáme takto vzniklé případy.

Situace, kdy platí jednotlivé modely, jsou do sebe zanořeny jako na obrázku 4.1.

Obrázek 4.1: Vnoření modelů



4.1 Data z modelu II

Zajímá nás, jak se v tomto případě budou chovat odhady $\hat{\delta}^{III}$ a $\hat{\delta}^{I}$.

Když platí model II, platí i model III s parametrem $\beta = 1$ a model I, který platí vždy. Odhady $\hat{\delta}^{III}$ a $\hat{\delta}^{I}$ jsou tedy nestranné a konzistentní. Vyjádříme rozptyly $V_{\delta^{III}}$ a $V_{\delta^{I}}$ pomocí parametrů modelu II.

4.1.1 Chování odhadu $\hat{\delta}^{III}$ za platnosti modelu II

Model III s parametrem $\beta = 1$ je totožný s modelem II. Proto se shodují i rozpyly náhodných chyb, platí $\sigma^2_{II,C} = \sigma^2_{III,C}$ a $\sigma^2_{II,E} = \sigma^2_{III,E}$. Asymptotický rozptyl $\sqrt{n}\left(\hat{\delta}^{III} - \delta\right)$ za platnosti modelu II označíme $W^{II}_{\hat{\delta}^{III}}$ a (3.12) přepíšeme

$$W^{II}_{\hat{\delta}^{III}} = V_{\delta^{III}} = (q+1)\left(\sigma^2_{II,C} + \frac{1}{q}\sigma^2_{II,E}\right).$$

Asymptotické rozdělení odhadu $\hat{\delta}^{III}$ za platnosti modelu II tedy popíšeme

$$\sqrt{n}\left(\hat{\delta}^{III} - \delta\right) \xrightarrow{d} \mathsf{N}\left(0, W^{II}_{\hat{\delta}^{III}}\right).$$

Odhad $\hat{\delta}^{III}$ má stejné vlastnosti (nestrannost, konzistence) a asymptotický rozptyl jako odhad $\hat{\delta}^{II}$, v tomto smyslu jsou za platnosti modelu II stejně dobré.

4.1.2 Chování odhadu $\hat{\delta}^{I}$ za platnosti modelu II

Nyní se podívejme na odhad $\hat{\delta}^{I}$. Asymptotický rozptyl $\sqrt{n}\left(\hat{\delta}^{I} - \delta\right)$ za platnosti modelu II označíme $W^{II}_{\hat{\delta}^{I}}$. Protože při platnosti modelu II platí i model I, je roven $V_{\delta^{I}}$. Vzhledem k tomu, že v tomto případě je

$$\sigma^2_{I,C} = \mathsf{var}\, Y^C_{1i} = \mathsf{var}\, Y^C_{0i} + \mathsf{var}\, \epsilon^{II,C}_i = \sigma^2_0 + \sigma^2_{II,C}$$

a

$$\sigma^2_{I,E} = \mathsf{var}\, Y^E_{1i} = \mathsf{var}\, Y^E_{0i} + \mathsf{var}\, \epsilon^{II,E}_i = \sigma^2_0 + \sigma^2_{II,E},$$

pro odhad $\hat{\delta}^{I}$ platí

$$\sqrt{n}\left(\hat{\delta}^{I} - \delta\right) \xrightarrow{d} \mathsf{N}\left(0, W^{II}_{\hat{\delta}^{I}}\right),$$

kde

$$W^{II}_{\hat{\delta}^{I}} = (q+1)\left(\left(\sigma^2_{II,C} + \sigma^2_0\right) + \frac{1}{q}\cdot\left(\sigma^2_{II,E} + \sigma^2_0\right)\right) = V_{\delta^{II}} + \frac{(q+1)^2}{q}\sigma^2_0. \tag{4.1}$$

Odhad $\hat{\delta}^{I}$ je sice také nestranný a konzistentní jako odhad $\hat{\delta}^{II}$ a odhad $\hat{\delta}^{III}$, má však větší rozptyl. Proto je v tomto případě horší než oba dva zmíněné.

4.2 Data z modelu III

Když platí model III, platí i model I, ale neplatí obecně model II.

4.2.1 Chování odhadu $\hat{\delta}^{II}$ za platnosti modelu III

Pro odhad $\hat{\delta}^{II}$ musíme za platnosti modelu III znovu ověřit nestrannost a konzistenci.

$$\begin{aligned}\mathsf{E}\,\hat{\delta}^{II} &= \mathsf{E}\,\left(\bar{Y}_1^E - \bar{Y}_0^E\right) - \left(\bar{Y}_1^C - \bar{Y}_0^C\right)\\ &= \frac{1}{n_E}\sum_{i=1}^{n_E}\mathsf{E}\,\left(Y_{1i}^E - Y_{0i}^E\right) - \frac{1}{n_C}\sum_{i=1}^{n_C}\mathsf{E}\,\left(Y_{1i}^C - Y_{0i}^C\right)\\ &= \frac{1}{n_E}\sum_{i=1}^{n_E}\left(\alpha+\delta+(\beta-1)\mathsf{E}\,Y_{0i}^E\right) - \frac{1}{n_C}\sum_{i=1}^{n_C}\left(\alpha+(\beta-1)\mathsf{E}\,Y_{0i}^C\right) = \delta,\end{aligned}$$

$$\begin{aligned}\mathsf{var}\,\hat{\delta}^{II} &= \mathsf{var}\,\left(\bar{Y}_1^E - \bar{Y}_0^E\right) - \left(\bar{Y}_1^C - \bar{Y}_0^C\right)\\ &= \frac{1}{n_E^2}\sum_{i=1}^{n_E}\mathsf{var}\,\left(Y_{1i}^E - Y_{0i}^E\right) + \frac{1}{n_C^2}\sum_{i=1}^{n_C}\mathsf{var}\,\left(Y_{1i}^C - Y_{0i}^C\right)\\ &= \frac{1}{n_E^2}\sum_{i=1}^{n_E}\mathsf{var}\left(\alpha+\delta+(\beta-1)Y_{0i}^E+\epsilon_i^{III,E}\right) + \frac{1}{n_C^2}\sum_{i=1}^{n_C}\mathsf{var}\left(\alpha+(\beta-1)Y_{0i}^C+\epsilon_i^{III,C}\right)\\ &= \frac{1}{n_E^2}\sum_{i=1}^{n_E}\left((\beta-1)^2\sigma_0^2+\sigma_{III,E}^2\right) + \frac{1}{n_C^2}\sum_{i=1}^{n_C}\left((\beta-1)^2\sigma_0^2+\sigma_{III,C}^2\right)\\ &= \frac{q_n+1}{n_E}\left(\beta-1\right)^2\sigma_0^2 + \frac{1}{n_E}\sigma_{III,E}^2 + \frac{1}{n_C}\sigma_{III,C}^2.\end{aligned}$$

Vidíme, že i v tomto případě je odhad $\hat{\delta}^{II}$ nestranný a konzistentní. Zároveň pro jeho asymptotické rozdělení platí

$$\sqrt{n}\left(\hat{\delta}^{II}-\delta\right) \xrightarrow{d} \mathsf{N}\left(0, W_{\hat{\delta}^{II}}^{III}\right),$$

kde

$$W_{\hat{\delta}^{II}}^{III} = (q+1)\left(\sigma_{III,C}^2 + \frac{1}{q}\cdot\sigma_{III,E}^2\right) + \frac{(q+1)^2}{q}\cdot\left(\beta-1\right)^2\sigma_0^2,$$

což se dá napsat jako

$$W_{\hat{\delta}^{II}}^{III} = V_{\delta^{III}} + \frac{(q+1)^2}{q}\cdot\left(\beta-1\right)^2\sigma_0^2.$$

To znamená, že v případě platnosti modelu III má sice odhad $\hat{\delta}^{II}$ stejné vlastnosti (nestrannost, konzistence) jako odhad $\hat{\delta}^{III}$, stejný asymptotický rozptyl má ale jen pokud parametr $\beta = 1$, což dělá z modelu III model II, jinak má rozptyl větší. V tomto smyslu je tedy horší.

4.2.2 Chování odhadu $\hat{\delta}^{I}$ za platnosti modelu III

Protože platí zároveň model I, je odhad $\hat{\delta}^{I}$ nestranný a konzistentní. Přepíšeme asymptotický rozptyl V_{δ^I} (3.4) pomocí parametrů modelu III.

$$\begin{aligned} V_{\delta^I} &= (q+1)\cdot\left(\mathsf{var}\, Y_{1i}^{C} + \frac{1}{q}\cdot \mathsf{var}\, Y_{1i}^{C}\right) \\ &= (q+1)\cdot\left(\beta^2 \mathsf{var}\, Y_{0i}^{C} + \mathsf{var}\, \epsilon_i^{III,C} + \frac{1}{q}\cdot\left(\beta^2 \mathsf{var}\, Y_{0i}^{E} + \mathsf{var}\, \epsilon_i^{III,E}\right)\right) \\ &= (q+1)\cdot\left(\sigma^2_{III,C} + \frac{1}{q}\cdot\sigma^2_{III,E}\right) + \frac{(q+1)^2}{q}\beta^2\sigma_0^2 = W^{III}_{\hat{\delta}^I} \end{aligned}$$

Vzhledem k (3.12) se dá asymptotický rozptyl $W^{III}_{\hat{\delta}^I}$ napsat jako

$$W^{III}_{\hat{\delta}^I} = V_{\delta^{III}} + \frac{(q+1)^2}{q}\beta^2\sigma_0^2.$$

Odhad $\hat{\delta}^{I}$ má tedy v této situaci stejný rozptyl jako $\hat{\delta}^{III}$ jen v případě, že $\beta = 0$, jinak je jeho rozptyl větší a je v tomto smyslu horší.

Srovnáme-li $W^{III}_{\hat{\delta}^{II}}$ a $W^{III}_{\hat{\delta}^I}$, vidíme, že za platnosti modelu III je rozptyl $\hat{\delta}^{II}$ menší než rozptyl odhadu $\hat{\delta}^{I}$, pokud

$$\frac{(q+1)^2}{q}\beta^2\sigma_0^2 > \frac{(q+1)^2}{q}\cdot(\beta-1)^2\sigma_0^2,$$

což po úpravě vyjde jako $\beta > \frac{1}{2}$. Je-li $\beta = \frac{1}{2}$, mají oba odhady stejný rozptyl a je-li $\beta < \frac{1}{2}$, je vhledem k velikosti rozptylu lepší $\hat{\delta}^{I}$ než $\hat{\delta}^{II}$.

4.3 Data z modelu I

Model I je obecný zápis náhodné veličiny, rozhodně tedy nemusí platit model II ani model III, tvar závislosti může být libovolný.

4.3.1 Chování odhadu $\hat{\delta}^{II}$ za platnosti modelu I

U odhadu $\hat{\delta}^{II}$ musíme opět ověřit nestrannost a konzistenci.

$$\begin{aligned}\mathsf{E}\,\hat{\delta}^{II} &= \mathsf{E}\,\left(\bar{Y}_1^E - \bar{Y}_0^E\right) - \left(\bar{Y}_1^C - \bar{Y}_0^C\right)\\ &= \frac{1}{n_E}\sum_{i=1}^{n_E}\mathsf{E}\,\left(Y_{1i}^E - Y_{0i}^E\right) - \frac{1}{n_C}\sum_{i=1}^{n_C}\mathsf{E}\,\left(Y_{1i}^C - Y_{0i}^C\right)\\ &= \frac{1}{n_E}\sum_{i=1}^{n_E}\left(\mu+\delta-\mu_0\right) - \frac{1}{n_C}\sum_{i=1}^{n_C}\left(\mu-\mu_0\right) = \delta,\end{aligned}$$

tedy i v případě modelu I je odhad $\hat{\delta}^{II}$ nestranný. Dále

$$\begin{aligned}\mathsf{var}\,\hat{\delta}^{II} &= \mathsf{var}\,\left(\bar{Y}_1^E - \bar{Y}_0^E\right) - \left(\bar{Y}_1^C - \bar{Y}_0^C\right)\\ &= \frac{1}{n_E^2}\sum_{i=1}^{n_E}\mathsf{var}\,\left(Y_{1i}^E - Y_{0i}^E\right) + \frac{1}{n_C^2}\sum_{i=1}^{n_C}\mathsf{var}\,\left(Y_{1i}^C - Y_{0i}^C\right)\\ &= \frac{1}{n_E^2}\sum_{i=1}^{n_E}\left(\mathsf{var}\,Y_{1i}^E + \mathsf{var}\,Y_{0i}^E - 2\mathsf{cov}\,\left(Y_{0i}^E, Y_{1i}^E\right)\right)\\ &\quad + \frac{1}{n_C^2}\sum_{i=1}^{n_C}\left(\mathsf{var}\,Y_{1i}^C + \mathsf{var}\,Y_{0i}^C - 2\mathsf{cov}\,\left(Y_{0i}^C, Y_{1i}^C\right)\right)\\ &= \frac{q_n+1}{n_E}\sigma_0^2 + \frac{1}{n_E}\sigma_{I,E}^2 + \frac{1}{n_C}\sigma_{I,C}^2 - 2\left(\frac{1}{n_E}\mathsf{cov}\,\left(Y_{0i}^E, Y_{1i}^E\right) + \frac{1}{n_C}\mathsf{cov}\,\left(Y_{0i}^C, Y_{1i}^C\right)\right),\end{aligned}$$

odhad $\hat{\delta}^{II}$ je také konzistentní. Dále pro jeho asymptotické rozdělení platí

$$\sqrt{n}\left(\hat{\delta}^{II} - \delta_0\right) \xrightarrow{d} \mathsf{N}\left(0, W_{\hat{\delta}^{II}}^I\right),$$

kde

$$\begin{aligned}W_{\hat{\delta}^{II}}^I &= \frac{(q+1)^2}{q}\sigma_0^2 + (q+1)\left(\frac{1}{q}\sigma_{I,E}^2 + \sigma_{I,C}^2\right)\\ &\quad - 2(q+1)\left(\frac{1}{q}\mathsf{cov}\,\left(Y_{0i}^E, Y_{1i}^E\right) + \mathsf{cov}\,\left(Y_{0i}^C, Y_{1i}^C\right)\right)\\ &= V_{\delta^I} + \frac{(q+1)^2}{q}\sigma_0^2 - 2(q+1)\left(\frac{1}{q}\mathsf{cov}\,\left(Y_{0i}^E, Y_{1i}^E\right) + \mathsf{cov}\,\left(Y_{0i}^C, Y_{1i}^C\right)\right). \quad (4.2)\end{aligned}$$

V případě, kdy platí model I, je lepší odhad $\hat{\delta}^I$ než odhad $\hat{\delta}^{II}$, pokud

$$0 < \frac{(q+1)^2}{q}\sigma_0^2 - 2(q+1)\left(\frac{1}{q}\mathsf{cov}\left(Y_{0i}^E, Y_{1i}^E\right) + \mathsf{cov}\left(Y_{0i}^C, Y_{1i}^C\right)\right),$$

tedy pokud

$$\frac{\mathsf{cov}\left(Y_{0i}^E, Y_{1i}^E\right)}{\sigma_0^2} < \frac{1}{2}, \quad \text{a} \quad \frac{\mathsf{cov}\left(Y_{0i}^C, Y_{1i}^C\right)}{\sigma_0^2} < \frac{1}{2}.$$

Jsou-li obě nerovnosti obrácené, je vhledem k velikosti rozptylu lepší odhad $\hat{\delta}^{II}$ než $\hat{\delta}^I$ a je-li místo nerovností rovnost, jsou oba odhady v tomto smyslu stejně dobré.

V případě, že

$$\mathsf{var}\, Y_{1i}^E = \mathsf{var}\, Y_{1i}^C = \sigma_0^2,$$

je ve skutečnosti kritériem pro porovnávání velikosti rozptylů korelační koeficient mezi konečnou a počáteční hodnotou. V tomto případě je-li korelace blízká 1, model I se blíží modelu II a je lepší použít pro odhadování $\hat{\delta}^{II}$.

4.3.2 Vyjádření modelu I v jiném tvaru

Abychom se mohli podívat, jak se chová odhad $\hat{\delta}^{III}$, vyjádříme si model I jinak. Konečné hodnoty rozepíšeme

$$\begin{aligned} Y_{1i} &= \mathsf{E}\left(Y_{1i}|Y_{0i}, i \in C\right) + \left(Y_{1i} - \mathsf{E}\left(Y_{1i}|Y_{0i}, i \in C\right)\right), \quad i \in C, \\ Y_{1i} &= \mathsf{E}\left(Y_{1i}|Y_{0i}, i \in E\right) + \left(Y_{1i} - \mathsf{E}\left(Y_{1i}|Y_{0i}, i \in E\right)\right), \quad i \in E \end{aligned}$$

a označíme

$$h(Y_{0i}) = \mathsf{E}\left(Y_{1i}|Y_{0i}, i \in C\right).$$

Dále také označíme

$$\epsilon_i^* = Y_{1i} - \mathsf{E}\left(Y_{1i}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right).$$

Pro ϵ_i^* platí

$$\mathsf{E}\,\epsilon_i^* = \mathsf{E}\left(Y_{1i} - \mathsf{E}\left(Y_{1i}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right)\right) = 0,$$

proto

$$\mathsf{E}\,h(Y_{0i}) = \mathsf{E}\,Y_{1i} - \mathsf{E}\,\epsilon_i^* = \mu, \quad i \in C.$$

Dále, nechť platí

$$\mathsf{E}\left(Y_{1i}|Y_{0i}, i \in E\right) = h(Y_{0i}) + \delta, \quad i \in E.$$

Dohromady dostaneme

$$Y_{1i} = h(Y_{0i}) + \delta\mathsf{I}_{[i\epsilon E]} + \epsilon_i^* \tag{4.3}$$

a střední hodnota Y_{1i} je rovna

$$\mathsf{E}\, Y_{1i} = \mathsf{E}\, h(Y_{0i}) + \delta \mathsf{E}\, \mathsf{I}_{[i\epsilon E]} + \mathsf{E}\, \epsilon_i^*. = \mu + \delta \cdot \pi.$$

Ověříme shodu zápisů modelu I. Když rozepíšeme

$$h(Y_{0i}) = \mathsf{E}\, h(Y_{0i}) + \xi_i(Y_{0i}), \quad \mathsf{E}\, \xi_i(Y_{0i}) = 0,$$

dostaneme

$$Y_{1i} = \mathsf{E}\, h(Y_{0i}) + \xi_i(Y_{0i}) + \epsilon_i^*, \quad i \in C,$$

tedy $\xi_i(Y_{0i}) + \epsilon_i^* = \epsilon_i^{I,C}, \quad i \in C$ a platí $\mathsf{E}\, \epsilon_i^{I,C} = 0$.

Pro skupinu E máme

$$Y_{1i} = h(Y_{0i}) + \delta + \epsilon_i^* = \mathsf{E}\, h(Y_{0i}) + \delta + \xi_i(Y_{0i}) + \epsilon_i^*, \quad i \in E,$$

kde $\xi_i(Y_{0i}) + \epsilon_i^* = \epsilon_i^{I,E}, \quad i \in E$ a $\mathsf{E}\, \epsilon_i^{I,E} = 0$.

Rozptyl ϵ_i^* může obecně záviset na počáteční hodnotě i na skupině, kam je pacient přiřazený, neboť

$$\epsilon_i^* = \epsilon_i^{I,C} + \mathsf{I}_{[i\epsilon E]}(\epsilon_i^{I,E} - \epsilon_i^{I,C}) - \xi_i(Y_{0i}),$$

proto označíme

$$\mathsf{var}\, (\epsilon_i^* | Y_{0i}, \mathsf{I}_{[i\epsilon E]}) = \psi\left(Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right). \tag{4.4}$$

Dále zjistíme, čemu se v novém zápisu modelu I rovnají původní parametry $\mu, \sigma^2_{I,C}$ a $\sigma^2_{I,E}$.

$$\mu = \mathsf{E}\, h(Y_{0i}),$$

dále protože

$$\sigma^2_{I,C} = \mathsf{var}\, \epsilon_i^{I,C} = \mathsf{var}\, (Y_{1i} - \mu | i \in C) = \mathsf{var}\, (h(Y_{0i}) + \epsilon_i^* - \mu | i \in C),$$
$$\sigma^2_{I,E} = \mathsf{var}\, \epsilon_i^{I,E} = \mathsf{var}\, (Y_{1i} - \mu - \delta \mathsf{I}_{[i\epsilon E]} | i \in E)$$
$$= \mathsf{var}\, (h(Y_{0i}) + \delta \mathsf{I}_{[i\epsilon E]} + \epsilon_i^* - \mu - \delta \mathsf{I}_{[i\epsilon E]} | i \in E),$$

stačí zjistit, čemu se rovná jeden z parametrů $\sigma^2_{I,C}, \sigma^2_{I,E}$. Druhý bude obdobou s podmíněním druhou skupinou. Nejprve provedeme pomocné výpočty. Protože

$$\mathsf{E}\, (\epsilon_i^* | i \in C, Y_{0i}) = 0,$$

je

$$\mathsf{E}\, (\epsilon_i^{*2} | i \in C, Y_{0i}) = \mathsf{E}\, \left[(\epsilon^* - \mathsf{E}\, [\epsilon_i^* | i \in C, Y_{0i}])^2 | i \in C, Y_{0i}\right]$$
$$= \mathsf{var}\, (\epsilon_i^* | i \in C, Y_{0i}) = \psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]} = 0).$$

Odsud máme

$$\begin{aligned}\mathsf{E}\left[(h(Y_{0i})+\epsilon_i^*)^2|i\in C\right] &= \mathsf{E}_{Y_{0i}}\mathsf{E}\left[(h(Y_{0i})+\epsilon_i^*)^2|Y_{0i}, i\in C\right]\\ &= \mathsf{E}_{Y_{0i}}\left[h(Y_{0i})^2+2h(Y_{0i})\mathsf{E}\left(\epsilon_i^*|Y_{0i}, i\in C\right)+\mathsf{E}\left(\epsilon_i^{*2}|Y_{0i}, i\in C\right)\right]\\ &= \mathsf{E}_{Y_{0i}}\left[h(Y_{0i})^2+\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]}=0)\right] = \mathsf{E}\,h(Y_{0i})^2+\mathsf{E}\left[\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})|\mathsf{I}_{[i\epsilon E]}=0\right].\end{aligned}$$

Toto použijeme při výpočtu parametru $\sigma^2_{I,C}$,

$$\begin{aligned}\sigma^2_{I,C} &= \mathsf{var}\left[h(Y_{0i})+\epsilon_i^*-\mu|i\in C\right]\\ &= \mathsf{E}\left[(h(Y_{0i})+\epsilon_i^*)^2|i\in C\right]-\left(\mathsf{E}\left(h(Y_{0i})+\epsilon_i^*|i\in C\right)\right)^2\\ &= \mathsf{E}\,h(Y_{0i})^2+\mathsf{E}\left[\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})|\mathsf{I}_{[i\epsilon E]}=0\right]-\mu^2.\end{aligned}$$

Parametr $\sigma^2_{I,E}$ je roven

$$\sigma^2_{I,E} = \mathsf{E}\,h(Y_{0i})^2+\mathsf{E}\left[\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})|\mathsf{I}_{[i\epsilon E]}=1\right]-\mu^2.$$

Rozptyly konečných hodnot jsou tedy v obou skupinách skutečně konstantní.

4.3.3 Chování odhadu $\hat{\delta}^{III}$ za platnosti modelu I

Podívejme se nyní na odhad $\hat{\delta}^{III}$. Nechť je Y_{1i} náhodná veličina z modelu I, pak jde rozepsat i ve tvaru modelu III, ale náhodné veličiny ϵ_i^{III} nebudou splňovat předpoklady uvedené v modelu III. Nejprve proto zjistíme, jakou mají ϵ_i^{III} střední hodnotu, platí-li model I. Rozepíšeme tedy Y_{1i} jako v modelu III (3.5) a zároveň v nové formulaci modelu I (4.3). Musíme ale rozlišit efekty léčby δ v obou modelech. Označme tedy prozatím $\Delta = \delta$ v modelu III a napišme

$$Y_{1i} = \alpha+\beta Y_{0i}+\Delta\mathsf{I}_{[i\epsilon E]}+\epsilon_i^{III} = h(Y_{0i})+\delta\mathsf{I}_{[i\epsilon E]}+\epsilon_i^*.$$

Odsud

$$\begin{aligned}\epsilon_i^{III} &= h(Y_{0i})+\delta\mathsf{I}_{[i\epsilon E]}+\epsilon_i^*-\alpha-\beta Y_{0i}-\Delta\mathsf{I}_{[i\epsilon E]},\\ \mathsf{E}\left(\epsilon_i^{III}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) &= h(Y_{0i})-(\alpha+\beta Y_{0i})+(\delta-\Delta)\mathsf{I}_{[i\epsilon E]},\\ \mathsf{E}\,\epsilon_i^{III} &= \mathsf{E}\,\mathsf{E}\left(\epsilon_i^{III}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = \mu-(\alpha+\beta\mu_0)+(\delta-\Delta)\pi.\end{aligned}$$

Předpoklad z modelu III (3.5) o nulovosti podmíněných středních hodnot $\mathsf{E}\left(\epsilon_i^{III}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right)$ je porušen v případě, že $\delta\neq\Delta$ nebo $h(x)$ není lineární funkce.

Nyní chceme zjistit, jaké vlastnosti má odhad $\hat{\delta}^{III}$. K tomu využijeme definice a dvě věty, převzaté z diplomové práce Drábková (2009). Říkají nám, jak se chovají odhady parametrů metodou maximální věrohodnosti, když platí jiné rozdělení, než ze kterého tvoříme věrohodnostní funkci.

Definice 1. [White (1982)] Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení se spojitou distribuční funkcí G na měřitelném euklidovském prostoru Ω a nechť $F_\theta(x)$ je rodina distribučních funkcí, jejichž hustoty $f_\theta(x)$ jsou měřitelné v x pro každé $\theta \in \Theta$, kde Θ je kompaktní množina parametrů, a spojité v θ pro všechna $x \in \Omega$. $\mathbb{X} = (X_1, \ldots, X_n)^T$, g je hustota příslušná G. Pak

$$L_n(\mathbb{X}, \theta) = \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_i)$$

nazýváme kvazi-logaritmickou věrohodnostní funkcí a

$$\hat{\theta}_n^* = \mathsf{argmax}\,(L_n(\mathbb{X}, \theta),\ \theta \in \Theta)$$

kvazi-maximálně věrohodným odhadem parametru θ.

Věta 3. *[White (1982)] Za podmínek definice 1 kvazi-maximálně věrohodný odhad $\hat{\theta}_n^*$ existuje a je měřitelný pro všechna $n \in \mathbb{N}$.*

Definice 2 (White (1982))**.** Nechť jsou P a Q dvě rozdělení pravděpodobnosti s hustotami p a q. Pak

$$K(p, q) = \mathsf{E}_P \log \frac{p(x)}{q(x)}$$

nazýváme Kullbackovou-Leiblerovou vzdáleností rozdělení Q od P.

Věta 4. *[White (1982)] Nechť platí předpoklady definice 1 a dále nechť*

(i) existuje $\mathsf{E} \log g(X_n), n \in \mathbb{N}$ *a* $|\log f_\theta(x)| \leq m(x)$ *pro všechna* $\theta \in \Theta$*, kde* $\int m(x) \mathsf{d}G(x)$ *existuje a*

(ii) $K(g, f_\theta)$ *má jediné minimum v bodě* θ^**.*

Pak platí $\hat{\theta}_n^* \xrightarrow{s.j.} \theta^*$, $n \longrightarrow \infty$ *pro skoro všechy posloupnosti* X_n*.*

Důkazy vět 3 a 4 lze nalézt v článku White (1982).

Předpokládejme, že existuje derivace $\frac{\partial}{\partial \theta} \log f_\theta(x)$. V tomto případě můžeme říci, že $\hat{\theta}_n^*$ řeší rovnici

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_\theta(X_i) = 0.$$

Přidejme dále předpoklad, že můžeme zaměnit integrál a derivaci $\frac{\partial}{\partial \theta} \mathsf{E}_G \log f_\theta(X) = \mathsf{E}_G \frac{\partial}{\partial \theta} \log f_\theta(X)$.

Chceme-li proto nalézt θ^*, které minimalizuje $K(g, f_\theta) = \mathsf{E}_G \log \frac{g(X)}{f_\theta(X)} = \mathsf{E}_G \log g(X) - \mathsf{E}_G \log f_\theta(X)$, nalezneme θ^* řešením rovnice

$$\mathsf{E}_G \frac{\partial}{\partial \theta} \log f_\theta(X) = 0. \tag{4.5}$$

Vraťme se nyní k našemu případu, kdy platí model I (zastupuje ho rozdělení G z definice 1) a neplatí model III. Předpokládejme, že rozdělění F_θ z definice 1 je nyní normální regresní model s vektorem parametrů $(\alpha, \beta, \Delta)^T$, vysvětlovaná proměnná je konečná hodnota a vysvětlující proměnné jsou počáteční hodnota a indikátor příslušnosti do skupiny E. Vektor parametrů $(\alpha, \beta, \Delta)^T$ tedy přebírá roli parametru θ a neplatné rozdělení F_θ je v podstatě náš model III s tím, že mu přibyly předpoklady o normalitě a o shodě rozptylů. V případě normálního regresního modelu je maximálně věrohodný odhad jeho parametrů roven odhadu metodou nejmenších čtverců. Dále víme, že jsou $\hat{\alpha}, \hat{\beta}$ a $\hat{\delta}^{III}$ z modelu III odhady metodou nejmenších čtverců, a jsou tedy shodné s odhady metodou maximální věrohodnosti z modelu III doplněného o předpoklady shody rozptylů a normality (který zastupuje rozdělení uvedené v definici 1 jako F_θ). Abychom zjistili, k čemu tyto odhady konvergují, vyřešíme rovnici (4.5), která je v našem případě shodná s rovnicí

$$\mathsf{E}_{modelI} U_i(\alpha, \beta, \Delta) = 0,$$

kde U_i je definováno v (3.10) s tím, že δ nahradíme Δ. E_{modelI} znamená, že budeme střední hodnotu počítat za platnosti modelu I. Prvky vektoru budeme značit číslem v závorce psaným dolním indexem. Vyřešme tedy soustavu rovnic

$$\mathsf{E}\, U_{i(1)} = \mathsf{E}\, \epsilon_i^{III} = \mu - (\alpha + \beta\mu_0) + (\delta - \Delta)\pi = 0,$$

$$\begin{aligned}\mathsf{E}\, U_{i(2)} = \mathsf{E}\, Y_{0i}\epsilon_i^{III} = \mathsf{E}\, \mathsf{E}\, (Y_{0i}\epsilon_i^{III} | Y_{0i}, \mathsf{I}_{[i\epsilon E]}) = \mathsf{E}\, (Y_{0i} \mathsf{E}\, (\epsilon_i^{III} | Y_{0i}, \mathsf{I}_{[i\epsilon E]})) \\ = \mathsf{E}\, Y_{0i} h(Y_{0i}) - \alpha\mu_0 - \beta(\mu_0^2 + \sigma_0^2) + \mu_0(\delta - \Delta)\pi = 0,\end{aligned}$$

$$\mathsf{E}\, U_{i(3)} = \mathsf{E}\, (\mathsf{I}_{[i\epsilon E]} \mathsf{E}\, (\epsilon_i^{III} | Y_{0i}, \mathsf{I}_{[i\epsilon E]})) = \pi\mu - \pi(\alpha + \beta\mu_0) + (\delta - \Delta)\pi = 0$$

vzhledem k α, β a Δ. Z první a třetí rovnice vidíme, že $\delta - \Delta = 0$ a že $\mu - (\alpha + \beta\mu_0) = 0$. Ze druhé rovnosti dosazením μ za $(\alpha + \beta\mu_0)$ dostáváme, že $\mathsf{E}\, Y_{0i} h(Y_{0i}) = \mu_0\mu + \beta\sigma_0^2$. Odtud celkově dostaneme, že

$$\begin{gathered}\hat{\delta}^{III} \xrightarrow{P} \delta, \\ \hat{\beta} \xrightarrow{P} \frac{\mathsf{E}\, Y_{0i} h(Y_{0i}) - \mu_0\mu}{\sigma_0^2} = \frac{\mathsf{cov}\ (Y_{0i}, h(Y_{0i}))}{\sigma_0^2} \quad \text{a} \\ \hat{\alpha} \xrightarrow{P} \mu - \beta\mu_0.\end{gathered} \tag{4.6}$$

Poznamenejme, že kovariance $\mathsf{cov}\,(Y_{0i}, h(Y_{0i}))$ je rovna $\mathsf{cov}\,(Y_{0i}, Y_{1i})$. Ověříme to, nejprve spočítáme

$$\mathsf{E}\,\epsilon_i^* Y_{0i} = \mathsf{E}\,\mathsf{E}\,(\epsilon_i^* Y_{0i}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}) = \mathsf{E}\,(Y_{0i}\mathsf{E}\,(\epsilon_i^*|Y_{0i}, \mathsf{I}_{[i\epsilon E]})) = 0,$$

a tedy

$$\begin{aligned}\mathsf{cov}\,(Y_{0i}, Y_{1i}) &= \mathsf{cov}\,(h(Y_{0i}) + \delta\mathsf{I}_{[i\epsilon E]} + \epsilon_i^*, Y_{0i}) \\ &= \mathsf{cov}\,(h(Y_{0i}), Y_{0i}) + (\mathsf{E}\,\epsilon_i^* Y_{0i} - \mathsf{E}\,\epsilon_i^* \mathsf{E}\,Y_{0i}) + \mathsf{cov}\,(\delta\mathsf{I}_{[i\epsilon E]}, Y_{0i}) = \mathsf{cov}\,(h(Y_{0i}), Y_{0i}).\end{aligned}$$

Je tedy jedno, zda uvedeme, že $\hat\beta$ konverguje k výrazu $\mathsf{cov}\,(Y_{0i}, Y_{1i})/\sigma_0^2$ nebo k výrazu $\mathsf{cov}\,(Y_{0i}, h(Y_{0i}))/\sigma_0^2$.

Důležité pro nás je, že odhad $\hat\delta^{III}$ odhaduje skutečné δ z modelu I. Nyní se budeme zabývat asymptotickým rozptylem odhadu $\hat\delta^{III}$.

Abychom zjistili, jaké je v případě platnosti modelu I asymptotické rozdělení odhadu $\hat\delta^{III}$, použijeme větu 2. Potřebujeme spočítat varianční matici $D^{-1}VD^{-1}$ v limitních hodnotách odhadů parametrů modelu III. Matici D máme uvedenou v (3.11), matici V opět rozložíme $V = \mathsf{var}\,U_i = \mathsf{E}\,\mathsf{var}\,\left(U_i|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) + \mathsf{var}\,\mathsf{E}\,\left(U_i|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right)$. Spočítáme nejprve podmíněný rozptyl a pak jeho střední hodnotu,

$$\begin{aligned}\mathsf{var}\,\left(U_{i(1)}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) &= \mathsf{var}\,\left(\epsilon_i^{III}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = \mathsf{var}\,\left(h(Y_{0i}) + \epsilon_i^* - \alpha - \beta Y_{0i}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) \\ &= \mathsf{var}\,\left(\epsilon_i^*|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = \psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]}), \\ \mathsf{var}\,\left(U_{i(2)}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) &= \mathsf{var}\,\left(Y_{0i}\epsilon_i^*|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = Y_{0i}^2\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]}), \\ \mathsf{var}\,\left(U_{i(3)}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) &= \mathsf{I}_{[i\epsilon E]}\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]}), \\ \mathsf{cov}\,\left((U_{i(1)}, U_{i(2)})|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) &= \mathsf{cov}\,\left((\epsilon_i^*, Y_{0i}\epsilon_i^*)|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = Y_{0i}\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]}), \\ \mathsf{cov}\,\left((U_{i(2)}, U_{i(3)})|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) &= \mathsf{cov}\,\left((Y_{0i}\epsilon_i^*, \mathsf{I}_{[i\epsilon E]}\epsilon_i^*)|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = Y_{0i}\mathsf{I}_{[i\epsilon E]}\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]}), \\ \mathsf{cov}\,\left((U_{i(1)}, U_{i(3)})|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) &= \mathsf{cov}\,\left((\epsilon_i^*, \mathsf{I}_{[i\epsilon E]}\epsilon_i^*)|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = \mathsf{I}_{[i\epsilon E]}\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]}).\end{aligned}$$

Dostáváme tedy

$$\mathsf{E}\,\mathsf{var}\,\left(U_i|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = \begin{pmatrix} \mathsf{E}\,\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]}) & \mathsf{E}\,(Y_{0i}\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})) & \mathsf{E}\,(\mathsf{I}_{[i\epsilon E]}\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})) \\ \mathsf{E}\,(Y_{0i}\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})) & \mathsf{E}\,(Y_{0i}^2\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})) & \mathsf{E}\,(\mathsf{I}_{[i\epsilon E]}Y_{0i}\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})) \\ \mathsf{E}\,(\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})\mathsf{I}_{[i\epsilon E]}) & \mathsf{E}\,(\mathsf{I}_{[i\epsilon E]}Y_{0i}\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})) & \mathsf{E}\,(\mathsf{I}_{[i\epsilon E]}\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})) \end{pmatrix}. \quad (4.7)$$

Dále spočítejme podmíněnou střední hodnotu.

$$\begin{aligned}
\mathsf{E}\left(U_{i(1)}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) &= \mathsf{E}\left(\epsilon_i^{III}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = h(Y_{0i}) - (\alpha + \beta Y_{0i}), \\
\mathsf{E}\left(U_{i(2)}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) &= \mathsf{E}\left(Y_{0i}\epsilon_i^{III}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = Y_{0i}\left(h(Y_{0i}) - (\alpha + \beta Y_{0i})\right), \\
\mathsf{E}\left(U_{i(3)}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) &= \mathsf{E}\left(\mathsf{I}_{[i\epsilon E]}\epsilon_i^{III}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = \mathsf{I}_{[i\epsilon E]}\left(h(Y_{0i}) - (\alpha + \beta Y_{0i})\right).
\end{aligned} \tag{4.8}$$

Výpočet varianční matice vektoru podmíněných středních hodnot nelze provést obecně, potřebujeme znát tvar funkce $h(Y_{0i})$. V příští kapitole se tedy zaměříme na konkrétní teoretické příklady, kdy budeme znát marginální rozdělení Y_{0i} a funkci $h(Y_{0i})$ a $\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})$.

Kapitola 5

Teoretické příklady

I za platnosti modelu I bychom chtěli spočítat rozptyl odhadu $\hat{\delta}^{III}$, respektive asymptotický rozptyl $\sqrt{n}\left(\hat{\delta}^{III}-\delta\right)$ a následně ho porovnat s rozptylem odhadu $\hat{\delta}^{I}$, respektive s asymptotickým rozptylem $\sqrt{n}\left(\hat{\delta}^{I}-\delta\right)$. Toto ale není v obecném případě v našich silách, proto se to pokusíme provést alespoň v některých konkrétních případech, kdy je porovnáme i s rozptylem odhadu $\hat{\delta}^{II}$, respektive s rozptylem $\sqrt{n}\left(\hat{\delta}^{II}-\delta\right)$. Zvolíme si rozdělení náhodné veličiny Y_{0i}, funkce $h(Y_{0i})$ a $\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})$. Je jedno, jestli dále zvolíme střední hodnotu π indikátoru $\mathsf{I}_{[i\epsilon E]}$ nebo číslo q, protože platí $q = \pi/(1-\pi)$.

5.1 Příklad - normální rozdělení

Předpokládejme, že Y_{0i} pochází z normálního rozdělení, $\mathcal{L}(Y_{0i}) = \mathsf{N}(\mu_0, \sigma_0^2)$. Dále nechť funkce $h(x) = x^2$ a nechť je $\psi(x, y) = a + by$ lineární funkce indikátoru příslušnosti do skupiny E, rozptyly jsou tedy rozdílné pouze v jednotlivých skupinách. Dále si zvolíme nějaké q, z něhož dopočítáme π. Uvedená situace odpovídá modelu I.

Asymptotické rozptyly $\sqrt{n}\left(\hat{\delta}^{I}-\delta\right)$ a $\sqrt{n}\left(\hat{\delta}^{II}-\delta\right)$ jsou v tomto případě označeny V_{δ^I} a $W^I_{\delta^{II}}$ a uvedeny v (3.4) a (4.2), asymptotický rozptyl $\sqrt{n}\left(\hat{\delta}^{III}-\delta\right)$ je označen $W^I_{\delta^{III}}$. Abychom je mohli spočítat, potřebujeme nejprve zjistit, co jsou parametry $\mu, \sigma^2_{I,C}, \sigma^2_{I,E}, \alpha$

a β. Spočítáme

$$\begin{aligned}
\mu &= \mathsf{E}\, h(Y_{0i}) = \mu_0^2 + \sigma_0^2, \\
\sigma_{I,C}^2 &= \mathsf{E}\, h(Y_{0i})^2 + \mathsf{E}\, \left[\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]}) | \mathsf{I}_{[i\epsilon E]} = 0\right] - \mu^2 = \mathsf{E}\, Y_{0i}^4 + a - \mu^2, \\
\sigma_{I,E}^2 &= \mathsf{E}\, h(Y_{0i})^2 + \mathsf{E}\, \left[\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]}) | \mathsf{I}_{[i\epsilon E]} = 1\right] - \mu^2 = \mathsf{E}\, Y_{0i}^4 + a + b - \mu^2, \\
\beta &= \frac{\mathsf{cov}\left(h(Y_{0i}), Y_{0i}\right)}{\sigma_0^2} = \frac{\mathsf{E}\, Y_{0i}^3 - \mathsf{E}\, Y_{0i}^2 \mu_0}{\sigma_0^2}, \\
\alpha &= \mu - \beta\mu_0.
\end{aligned}$$

Pro výpočet rozptylu $W_{\delta^{III}}^I$ nás budou zajímat matice D a V. Matice D je uvedena výše (3.11), v případě normálního rozdělení je rovna

$$D = \begin{pmatrix} 1 & \mu_0 & \pi \\ \mu_0 & \mu_0^2 + \sigma_0^2 & \mu_0\pi \\ \pi & \mu_0\pi & \pi \end{pmatrix}.$$

Dále první člen součtu, který tvoří matici $V = \mathsf{E}\,\mathsf{var}\,\left(U_i | Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) + \mathsf{var}\,\mathsf{E}\,\left(U_i | Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right)$ (viz (4.7)), je roven

$$\mathsf{E}\,\mathsf{var}\,\left(U_i | Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = \begin{pmatrix} a + b\pi & a\mu_0 + b\mu_0\pi & a\pi + b\pi \\ a\mu_0 + b\mu_0\pi & (\mu_0^2 + \sigma_0^2)(a + b\pi) & \mu_0(a\pi + b\pi) \\ a\pi + b\pi & \mu_0(a\pi + b\pi) & a\pi + b\pi \end{pmatrix}.$$

Nyní musíme spočítat varianční matici vektoru podmíněných středních hodnot (4.8) pro náš případ normálního rozdělení a funkci $h(x)$. Počítejme

$$\begin{aligned}
&\mathsf{var}\,\mathsf{E}\,\left(U_{i(1)} | Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = \mathsf{var}\,\left(h(Y_{0i}) - (\alpha + \beta Y_{0i})\right) \\
&\quad = \mathsf{var}\, Y_{0i}^2 + \beta^2 \mathsf{var}\, Y_{0i} - 2\beta\mathsf{cov}\left(Y_{0i}^2, Y_{0i}\right) = \mathsf{E}\, Y_{0i}^4 - (\mathsf{E}\, Y_{0i}^2)^2 + \beta^2\sigma_0^2 - \beta(\mathsf{E}\, Y_{0i}^3 - \mathsf{E}\, Y_{0i}^2),
\end{aligned}$$

podobné výsledky, vyjádřené pomocí momentů normálního rozdělení, vyjdou pro další prvky varianční matice. Všechny momenty normálního rozdělení se dají vyjádřit pomocí μ_0 a σ_0^2.

Dále vypočítáme rozptyly V_{δ^I} a $W_{\delta^{II}}^I$ uvedené ve (3.4) a (4.2), abychom je mohli porovnat s rozptylem $W_{\delta^{III}}^I$.

Žádný z asymtotických rozptylů není závislý na skutečném δ.

Výpočty jsme naprogramovali do programu R, volili jsme různé parametry μ_0, σ_0^2, a, b a q, z něhož jsme spočítali π. Výsledky uvádíme v následujících tabulkách.

V prvním řádku v prvním políčku jsou použité funkce $h(x)$ a $\psi(x, y)$. Ve druhém políčku jsou námi zadané parametry rozdělení počáteční hodnoty a indikátoru příslušnosti

Funkce	Parametry	Rozptyly
$h(x) = x^2$ $\psi(x, y) = a + by$	$\mu_0 = 0.5, \sigma_0^2 = 16$ $\pi = 0.33$ $a = 2, b = 2$	$V_\delta^I = 2391$ $W_{\delta^{II}}^I = 2319$ $W_{\delta^{III}}^I = 2319$
Sendvičová varianční matice parametrů modelu III		α a β
$\begin{matrix} 811.04 & -80.08 & -771 \\ -80.08 & 160.17 & 0 \\ -771 & 0 & 2319 \end{matrix}$		$\alpha = 15.75, \beta = 1$

Tabulka 5.1: Příklad - normální rozdělení, malá střední hodnota počáteční hodnoty

Funkce	Parametry	Rozptyly
$h(x) = x^2$ $\psi(x, y) = a + by$	$\mu_0 = 10, \sigma_0^2 = 16$ $\pi = 0.33$ $a = 2, b = 2$	$V_\delta^I = 31119$ $W_{\delta^{II}}^I = 28311$ $W_{\delta^{III}}^I = 2319$
Sendvičová varianční matice parametrů modelu III		α a β
$\begin{matrix} 16787.67 & -1601.67 & -771 \\ -1601.67 & 160.17 & 0 \\ -771 & 0 & 2319 \end{matrix}$		$\alpha = -84, \beta = 20$

Tabulka 5.2: Příklad - normální rozdělení, vyšší střední hodnota počáteční hodnoty

do skupiny E. Ve třetím políčku jsou asymptotické rozptyly $\sqrt{n}\left(\hat{\delta}^I - \delta\right)$, $\sqrt{n}\left(\hat{\delta}^{II} - \delta\right)$ a $\sqrt{n}\left(\hat{\delta}^{III} - \delta\right)$. Ve druhém řádku je uvedena sendvičová asymptotická varianční matice odhadů parametrů modelu III ve tvaru (3.7), platí-li daný model I. Její pravý dolní prvek je roven $W_{\delta^{III}}^I$. V druhém políčku druhého řádku tabulky jsou uvedeny parametry α a β ze třetího modelu spočítané na základě daného modelu I, viz (4.6).

Jak vidíme v tabulce 5.1, zvolíme-li velmi malou střední hodnotu μ_0, asymptotické rozptyly jsou si velmi podobné. Rozptyly $W_{\delta^{II}}^I$ a $W_{\delta^{III}}^I$ jsou shodné, protože odhad na základě modelu III s $\beta = 1$ je roven odhadu na základě modelu II.

Tabulka 5.2 nám ukazuje, že zvýšíme-li μ_0, rozptyly $W_{\delta^{II}}^I$ a V_{δ^I} se výrazně zvětší, zatímco $W_{\delta^{III}}^I$ zůstává stále stejný.

Kdybychom ještě zvyšovali střední hodnotu μ_0, kvadratická závislost by se posouvala stále více k vyšším hodnotám Y_{0i} a Y_{1i}, tedy „doprava“ a tím by se více „linearizovala“ s vysokou směrnicí, model III by se tedy ze všech tří modelů stal výrazně nejlepším a stejně

Funkce	Parametry	Rozptyly
$h(x) = x^2$ $\psi(x, y) = a + by$	$\mu_0 = 50, \sigma_0^2 = 16$ $\pi = 0.33$ $a = 2, b = 2$	$V_\delta^I = 722319$ $W_{\delta^{II}}^I = 707991$ $W_{\delta^{III}}^I = 2319$
Sendvičová varianční matice parametrů modelu III		α a β
$\begin{matrix} 401187.67 & -8008.33 & -771 \\ -8008.33 & 160.17 & 0 \\ -771 & 0 & 2319 \end{matrix}$		$\alpha = -2484, \beta = 100$

Tabulka 5.3: Příklad - normální rozdělení, ještě vyšší střední hodnota počáteční hodnoty

Funkce	Parametry	Rozptyly
$h(x) = x^2$ $\psi(x, y) = a + by$	$\mu_0 = 10, \sigma_0^2 = 16$ $\pi = 0.33$ $a = 2, b = 2000$	$V_\delta^I = 37113$ $W_{\delta^{II}}^I = 34305$ $W_{\delta^{III}}^I = 8313$
Sendvičová varianční matice parametrů modelu III		α a β
$\begin{matrix} 20950.17 & -2017.92 & -771 \\ -2017.92 & 201.79 & 0 \\ -771 & 0 & 8313 \end{matrix}$		$\alpha = -84, \beta = 20$

Tabulka 5.4: Příklad - normální rozdělení, větší rozptyl chyb ve skupině E

i odhad $\hat{\delta}^{III}$, což by se projevilo ještě větším rozdílem rozptylů, jak ukazuje tabulka 5.3.

Asymptotické rozptyly se také zvýší, pokud se výrazně zvětší rozptyl chyb ve skupině E (tzn. větší b). Zvláště se to projeví na V_{δ^I}.

Zajímavé je, že jsme náhodou zvolili příklad, kdy rozptyl V_{δ^I} vůbec nezávisí na střední hodnotě μ_0.

Poznamenejme, že kdychom zvolili jinou funkci $\psi(x, y)$, která by závisela i na první proměnné, rozptyl V_{δ^I} by na μ_0 závisel. Rozdíl by se objevil v matici (4.7).

V tomto příkladě je vhledem k velikosti rozptylu nejlepší odhad $\hat{\delta}^{III}$, jehož asymptotický rozptyl se nemění na základě střední hodnoty počáteční hodnoty. Odhad $\hat{\delta}^I$ je stejně dobrý, je-li střední hodnota μ_0 nulová.

5.2 Příklad - rovnoměrné rozdělení

Nyní předpokládejme, že Y_{0i} je z rovnoměrného rozdělení, $\mathcal{L}(Y_{0i}) = \mathsf{R}(a,b)$. Budeme brát taková a, b, že $P(Y_{0i} \geq 0) = 1$. Funkci h můžeme proto zvolit $h(x) = \sqrt{x}$. Dále vezměme $\psi(x,y) = x^2 + cy$, rozptyl náhodných chyb je tedy větší tam, kde je větší počáteční a konečná hodnota. Dále si zvolíme nějaké q, z něhož opět spočítáme π. Abychom nemuseli střední hodnotu a rozptyl Y_{0i} stále vyjadřovat pomocí a a b, dopočítáme nejprve μ_0 a σ_0^2,

$$\mu_0 = \frac{a+b}{2},$$
$$\sigma_0^2 = \frac{(b-a)^2}{12}.$$

I v tomto případě chceme získat asymptotické rozptyly V_{δ^I}, $W^I_{\delta^{II}}$ a $W^I_{\delta^{III}}$, z nichž jsou první dva uvedené ve vzorcích (3.4) a (4.2) a třetí dostaneme ze sendvičové varianční matice. Spočítáme proto nejprve parametry $\mu, \sigma^2_{I,C}, \sigma^2_{I,E}, \alpha$ a β.

$$\mu = \mathsf{E}\, h(Y_{0i}) = \mathsf{E}\, \sqrt{Y_{0i}} = \frac{2(b^{\frac{3}{2}} - a^{\frac{3}{2}})}{b-a},$$
$$\sigma^2_{I,C} = \mathsf{E}\, h(Y_{0i})^2 + \mathsf{E}\, \left[\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})|\mathsf{I}_{[i\epsilon E]} = 0\right] - \mu^2 =$$
$$\mathsf{E}\, Y_{0i} + \mathsf{E}\, Y_{0i}^2 - \mu^2 = \mu_0 + \frac{a^2+ab+b^2}{3} - \mu^2,$$
$$\sigma^2_{I,E} = \mu_0 + \frac{a^2+ab+b^2}{3} + c - \mu^2,$$
$$\beta = \frac{\mathsf{cov}\,(h(Y_{0i}), Y_{0i})}{\sigma_0^2} = \frac{\mathsf{E}\, Y_{0i}^{\frac{3}{2}} - \mathsf{E}\, Y_{0i}^{\frac{1}{2}} \mu_0}{\sigma_0^2},$$
$$\alpha = \mu - \beta\mu_0.$$

Matice D je v případě rovnoměrného rozdělení rovna

$$D = \begin{pmatrix} 1 & \mu_0 & \pi \\ \mu_0 & \mathsf{E}\, Y_{0i}^2 & \mu_0\pi \\ \pi & \mu_0\pi & \pi \end{pmatrix}$$

a střední hodnota podmíněné varianční matice U_i je rovna

$$\mathsf{E}\, \mathsf{var}\, \left(U_i | Y_{0i}, \mathsf{I}_{[i\epsilon E]}\right) = \begin{pmatrix} \mathsf{E}\, Y_{0i}^2 + c\pi & \mathsf{E}\, Y_{0i}^3 + c\pi \mathsf{E}\, Y_{0i} & \pi(\mathsf{E}\, Y_{0i}^2 + c) \\ \mathsf{E}\, Y_{0i}^3 + c\pi \mathsf{E}\, Y_{0i} & \mathsf{E}\, Y_{0i}^4 + c\pi \mathsf{E}\, Y_{0i}^2 & \pi(\mathsf{E}\, Y_{0i}^3 + c\mathsf{E}\, Y_{0i}) \\ c(\pi(\mathsf{E}\, Y_{0i}^2 + c) & \pi(\mathsf{E}\, Y_{0i}^3 + c\mathsf{E}\, Y_{0i}) & \pi(\mathsf{E}\, Y_{0i}^2 + c) \end{pmatrix},$$

s tím, že momenty rovnoměrného rozdělení dopočítáme pomocí parametrů a a b. Dále je potřeba spočítat varianční matici vektoru podmíněných středních hodnot $\mathsf{var}\,\mathsf{E}\,(U_i|Y_{0i}, \mathsf{I}_{[i\epsilon E]})$, jejichž obecný tvar nalezneme v (4.8)

$$\begin{aligned}\mathsf{var}\,\mathsf{E}\,(U_{i(1)}|Y_{0i}, \mathsf{I}_{[i\epsilon E]}) &= \mathsf{var}\,(h(Y_{0i}) - (\alpha + \beta Y_{0i})) \\ &= \mu_0 - (\mathsf{E}\,Y_{0i}^{\frac{1}{2}})^2 - 2\beta(\mathsf{E}\,Y_{0i}^{\frac{3}{2}} - \mathsf{E}\,Y_{0i}^{\frac{1}{2}}\mu_0) + \beta^2\sigma_0^2,\end{aligned}$$

podobně se počítají i další prvky varianční matice. Jednotlivé momenty se opět snadno vyjádří pomocí parametrů a a b.

Výpočty jsme opět naprogramovali do programu R, volili jsme různé hodnoty parametrů a, b, q a c.

V tabulce 5.5 vidíme, že $W^I_{\delta^{III}}$ je opět o něco lepší než ostatní dva rozptyly, i když rozdíl mezi ním a V_{δ^I} je nepatrný.

Funkce	Parametry	Rozptyly
$h(x) = \sqrt{x}$ $\psi(x,y) = cx^2 + dy$	$a = 10, b = 20$ $\pi = 0.41, c = 1, d = 0$ $\mu_0 = 15, \sigma_0^2 = 8.33$	$V^I_\delta = 963.92$ $W^I_{\delta^{II}} = 989.36$ $W^I_{\delta^{III}} = 963.33$
Sendvičová varianční matice parametrů modelu III		α a β
$\begin{matrix} 5976.68 & -402 & -396.67 \\ -402 & 28.8 & 0 \\ -396.67 & 0 & 963.33 \end{matrix}$		$\alpha = 1.9, \beta = 0.13$

Tabulka 5.5: Příklad - rovnoměrné rozdělení, menší počáteční hodnota

V tabulce 5.6 můžeme vidět, že zvětšíme-li a a b o stejné číslo, všechny asymptotické rozptyly se zvětší. Rozdíl mezi V_{δ^I} a $W^I_{\delta^{III}}$ se zvětší, ale rozdíl mezi V_{δ^I} a $W^I_{\delta^{II}}$ zůstane přibližně stejný. Protože se ale oba výrazně zvětšily, je více zanedbatelný. β je bližší nule, což potvrzuje, že se modely I a III, a tedy i odhady $\hat{\delta}^{III}$ a $\hat{\delta}^I$ kvalitativně (vzhledem k velikosti rozptylu) přiblížily.

V tabulce 5.7 vidíme, co se stane, zvětšíme-li šířku intervalu (a, b). Odhady $\hat{\delta}^I$ a $\hat{\delta}^{III}$ zůstanou vyhledem k rozptylu téměř stejně dobré, ale rozptyl odhadu $\hat{\delta}^{II}$ se výrazně zvětší.

Snížíme-li rozptyly chyb ve skupině E (tedy zmenšíme d), asymptotické rozptyly se sníží rovnoměrně přibližně o konstantu, jak můžeme vidět v tabulce 5.8.

Poslední dvě tabulky 5.9 a 5.10 ukážou, jak moc se zvětší asymptotické rozptyly, když je rozdělení indikátoru $\mathsf{I}_{[i\epsilon E]}$ takové, že je buď výrazně více lidí ve skupině E, nebo naopak ve skupině C. V obou případech se rozptyly zvětší rovnoměrně o stejnou konstantu.

Funkce	Parametry	Rozptyly
$h(x) = \sqrt{x}$ $\psi(x, y) = cx^2 + dy$	$a = 510, b = 520$ $\pi = 0.41, c = 0.01, d = 0$ $\mu_0 = 515, \sigma_0^2 = 8.33$	$V_\delta^I = 10950.36$ $W_{\delta^{II}}^I = 10983.25$ $W_{\delta^{III}}^I = 10950.35$
Sendvičová varianční matice parametrů modelu III		α a β
$\begin{matrix} 84411834.77 & -163908.02 & -4508.97 \\ -163908.02 & 318.29 & 0 \\ -4508.97 & 0 & 10950.35 \end{matrix}$		$\alpha = 11.35, \beta = 0.02$

Tabulka 5.6: Příklad - rovnoměrné rozdělení, větší počáteční hodnota

Funkce	Parametry	Rozptyly
$h(x) = \sqrt{x}$ $\psi(x, y) = cx^2 + dy$	$a = 10, b = 1000$ $\pi = 0.41, c = 0.01, d = 0$ $\mu_0 = 505, \sigma_0^2 = 81675$	$V_\delta^I = 14117.39$ $W_{\delta^{II}}^I = 334526.48$ $W_{\delta^{III}}^I = 13908.34$
Sendvičová varianční matice parametrů modelu III		α a β
$\begin{matrix} 8094.13 & -14.78 & -5726.96 \\ -14.78 & 0.05 & 0 \\ -5726.96 & 0 & 13908.34 \end{matrix}$		$\alpha = 8.7, \beta = 0.02$

Tabulka 5.7: Příklad - rovnoměrné rozdělení, větší šířka intervalu pro počáteční hodnotu

V tomto příkladě je vzhledem k velikosti rozptylu nejlepší odhad $\hat{\delta}^{III}$, i když v mnoha případech je rozdíl mezi jeho rozptylem a rozptylem odhadu $\hat{\delta}^I$ zanedbatelný.

Funkce	Parametry	Rozptyly
$h(x) = \sqrt{x}$ $\psi(x, y) = cx^2 + dy$	$a = 10, b = 20$ $\pi = 0.41, c = 1, d = -100$ $\mu_0 = 15, \sigma_0^2 = 8.33$	$V_\delta^I = 721.06$ $W_{\delta^{II}}^I = 746.51$ $W_{\delta^{III}}^I = 720.48$
Sendvičová varianční matice parametrů modelu III		α a β
$\begin{matrix} 4864.91 & -327.88 & -396.67 \\ -327.88 & 23.86 & 0 \\ -396.67 & 0 & 720.48 \end{matrix}$		$\alpha = 1.9, \beta = 0.13$

Tabulka 5.8: Příklad - rovnoměrné rozdělení, menší rozptyl chyb ve skupině E

Funkce	Parametry	Rozptyly
$h(x) = \sqrt{x}$ $\psi(x, y) = cx^2 + dy$	$a = 10, b = 20$ $\pi = 0.09, c = 1, d = 0$ $\mu_0 = 15, \sigma_0^2 = 8.33$	$V_\delta^I = 2825.05$ $W_{\delta^{II}}^I = 2899.62$ $W_{\delta^{III}}^I = 2823.34$
Sendvičová varianční matice parametrů modelu III		α a β
$\begin{matrix} 5836.68 & -402 & -256.67 \\ -402 & 28.8 & 0 \\ -256.67 & 0 & 2823.34 \end{matrix}$		$\alpha = 1.9, \beta = 0.13$

Tabulka 5.9: Příklad - rovnoměrné rozdělení, více lidí ve skupině C

Funkce	Parametry	Rozptyly
$h(x) = \sqrt{x}$ $\psi(x, y) = cx^2 + dy$	$a = 10, b = 20$ $\pi = 0.91, c = 1, d = 0$ $\mu_0 = 15, \sigma_0^2 = 8.33$	$V_\delta^I = 2825.05$ $W_{\delta^{II}}^I = 2899.62$ $W_{\delta^{III}}^I = 2823.34$
Sendvičová varianční matice parametrů modelu III		α a β
$\begin{matrix} 8146.68 & -402 & -2566.67 \\ -402 & 28.8 & 0 \\ -2566.67 & 0 & 2823.34 \end{matrix}$		$\alpha = 1.9, \beta = 0.13$

Tabulka 5.10: Příklad - rovnoměrné rozdělení, více lidí ve skupině E

Kapitola 6

Simulační studie

V této kapitole se zaměříme na výpočty založené na simulacích počáteční hodnoty z nějakého rozdělení a konečné hodnoty z nějakého modelu. Následně porovnáme jednotlivé odhady $\hat{\delta}^I$, $\hat{\delta}^{II}$ a $\hat{\delta}^{III}$ na základě jejich směrodatných chyb nebo empirických rozptylů. Výpočty provedeme v programu *R*.

Data získáme tak, že pro zadané n_C a q vygenerujeme $n = n_C + qn_C$ počátečních hodnot z nějakého zadaného rozdělení, dostaneme tak Y_{0i}, $i = 1, \ldots, n$. Dále určíme, které počáteční hodnoty budou patřit do skupiny E, přiřadíme tedy každé hodnotu indikátoru $\mathsf{I}_{[i\epsilon E]}$. Následně zvolíme δ a pro každou dvojici $(Y_{0i}, \mathsf{I}_{[i\epsilon E]})$ spočítáme hodnotu zvolených funkcí $h(Y_{0i}) + \delta\mathsf{I}_{[i\epsilon E]}$ a $\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})$. Nakonec spočítáme konečnou hodnotu Y_{1i} se střední hodnotou $h(Y_{0i}) + \delta\mathsf{I}_{[i\epsilon E]}$ a rozptylem $\psi(Y_{0i}, \mathsf{I}_{[i\epsilon E]})$ podle zvoleného rozdělení.

Na základě těchto dat spočítáme odhady $\hat{\delta}^I$, $\hat{\delta}^{II}$, a $\hat{\delta}^{III}$. Dále spočítáme odhady jejich asymptotických rozptylů, které odmocníme. Odmocninu odhadu asymptotického rozptylu nazveme směrodatnou chybou. V případě odhadu asymptotického rozptylu odhadu $\hat{\delta}^{III}$ se podíváme nejprve na naivní odhad, spočtený na základě klasického regresního modelu, a dále na rozptyl ze sendvičové varianční matice. Tu spočítáme dvěma způsoby, a to na základě násobku matic $\hat{D}^{-1}\hat{V}\hat{D}^{-1}$, kde $\hat{V}$ je empirická varianční matice vektoru $U_i(\hat{\alpha}, \hat{\beta}, \hat{\delta}^{III})$, a matice $\hat{D} = \frac{1}{n}\sum(1, Y_{0i}, \mathsf{I}_{[i\epsilon E]})^T(1, Y_{0i}, \mathsf{I}_{[i\epsilon E]})$. Dále se podíváme na sendvičovou varianční matici spočítanou pomocí funkce *gee* z balíku *gee*.

Nakonec na základě odhadů $\hat{\delta}^I$, $\hat{\delta}^{II}$, a $\hat{\delta}^{III}$, jejich směrodatných chyb (v případě $\hat{\delta}^{III}$ spočtených na základě každé ze tří variančních matic - naivní, sendvičové pomocí násobku matic a sendvičové z funkce *gee*) a normální aproximace spočítáme devadesátipětiprocentní konfidenční intervaly pro δ,

$$\left(\hat{\delta} - \sqrt{v}u_{0.975},\ \hat{\delta} + \sqrt{v}u_{0.975}\right),$$

kde $\hat{\delta}$ je některý z odhadů δ, v je odhad jeho asymptotického rozptylu a $u_{0.975}$ je kvantil normálního rozdělení. Zapamatujeme si jen to, zda intervaly pokryjí skutečné δ.

Toto celé budeme opakovat 10000krát. Následně spočítáme průměr a empirický rozptyl každého z odhadů parametru δ. Empirický rozptyl odmocníme, aby šel porovnat se směrodatnými chybami. Dále spočítáme průměry všech směrodatných chyb a pro každý z intervalů spolehlivosti zjistíme, jaký podíl z opakování pokryl skutečné δ.

U každého příkladu uvedeme tabulku s výsledky. V prvním řádku se v ní objeví zadané parametry rozdělení počáteční hodnoty a indikátoru příslušnosti do skupiny. Pro úplnost uvedeme q i π s tím, že jsme zadávali q. V prvním řádku tabulky budou také parametry funkcí ψ a h. Ve druhém řádku se objeví průměry odhadů $\hat{\delta}^I$, $\hat{\delta}^{II}$ a $\hat{\delta}^{III}$ a ve třetím odmocniny jejich empirických rozptylů. V dalším řádku uvedeme směrodatné chyby (ve skutečnosti jejich půměry ze všech opakování) s tím, že pro odhad $\hat{\delta}^{III}$ v tabulce uvedeme všechny tři způsoby odhadu, jak jsme je zmiňovali výše. V následujícím řádku uvedeme odhady pravděpodobnosti pokrytí skutečného δ intervaly spolehlivosti. V posledním řádku uvedeme pro úplnost průměry odhadů $\hat{\alpha}$ a $\hat{\beta}$ z modelu III.

6.1 Konečná hodnota pocházející z modelu II

Nejprve se podívejme na situaci, kdy konečná hodnota pochází z modelu II. Pro model II v zápisu (4.3) a (4.4) platí:

- $h(x) = x$,
- $\psi(x, y) = c + dy$.

Rozdělení počáteční a náhodných chyb ϵ_i^* závisí na naší volbě.

V tabulce 6.1 jsme zvolili počáteční hodnotu z gamma rozdělení tak, aby její střední hodnota a rozptyl odpovídaly μ_0 a σ_0^2 uvedeným v tabulce. Rozdělení chyb ϵ_i^* je také gamma. Vidíme, že na základě velikosti směrodatných chyb či empirických rozptylů jsou nejlepší odhady $\hat{\delta}^{II}$ a $\hat{\delta}^{III}$, největší směrodatnou chybu má $\hat{\delta}^{I}$. Naivní odhad směrodatné chyby $\hat{\delta}^{III}$ se trochu liší od sendvičových, což se promítne i do menší pravděpodobnosti pokrytí intervalem spolehlivosti.

Zvýšíme-li rozptyly konečných hodnot (zvýšíme c a d), rozdíly mezi směrodatnými chybami $\hat{\delta}^{I}$ a $\hat{\delta}^{II}$ se stanou méně podstatné, jak uvádí tabulka 6.2. Ze vzorce (4.1) vidíme, že se jejich asymptotické rozptyly liší o konstantu (závisející pouze na rozptylu počáteční hodnoty a na q) a je-li tato konstanta zanedbatelná oproti celému rozptylu, dá se říci, že

jsou rozptyly přibližně stejně velké. V tomto případě jsou tedy všechny odhady δ přibližně stejně dobré. Pro počáteční hodnotu i ϵ_i^* je opět použito gamma rozdělení.

6.2 Konečná hodnota pocházející z modelu III

Pro model III v zápisu (4.3) a (4.4) platí:

- $h(x) = a + bx$,
- $\psi(x, y) = c + dy$.

Rozdělení počáteční a konečné hodnoty opět závisí na naší volbě.

V tabulkách 6.3 a 6.4 jsme pro počáteční hodnotu zvolili exponenciální rozdělení se střední hodnotou μ_0 a pro ϵ_i^* rozdělení normální. Vzhledem k velikosti směrodatných chyb je v obou případech nejlepší odhad $\hat{\delta}^{III}$. Rozdíl nastává ve velikosti směrodatných chyb odhadů $\hat{\delta}^{II}$ a $\hat{\delta}^{I}$. Zatímco je-li parametr $b = 2$ (větší než 0.5, tabulka 6.3), má značně větší směrodatnou chybu odhad $\hat{\delta}^{I}$ a naopak, je-li $b = 0.2$ (menší než 0.5, tabulka 6.4), směrodatná chyba $\hat{\delta}^{I}$ se výrazně zmenšila a větší má odhad $\hat{\delta}^{II}$. Toto souhlasí se závěrem sekce 4.2.

6.3 Konečná hodnota pocházející z modelu I

V modelu I je funkce $h(x)$ i $\psi(x, y)$ obecná. Je tedy na nás, jak je zvolíme, stejně jako rozdělení počáteční hodnoty a náhodných chyb.

Jako první se podíváme na případ, kdy

- počáteční hodnota je z gamma rozdělení,
- ϵ_i^* pochází z normálního rozdělení,
- $h(x) = \sqrt{x}$,
- $\psi(x, y) = cx^4 + dy$.

Výsledky jsou uvedeny v tabulce 6.5. Všechny tři odhady jsou vzhledem k velikosti směrodatných chyb přibližně stejně dobré. Vzhledem k tomu, že je rozptyl náhodných chyb velmi závislý na počáteční hodnotě i na indikátoru příslušnosti do skupiny E, je rozdíl mezi asymptotickým naivním a sendvičovým rozptylem $\hat{\delta}^{III}$ a také v pravděpodobnosti pokrytí daným konfidenčním intervalem.

V tabulce 6.6 jsou výsledky situace, kdy

Parametry	$\mu_0 = 100, \sigma_0^2 = 4, n_C = 500, q = 0.7, \pi = 0.41,$ $c = 20, d = 10, \delta = 5$				
	$\hat{\delta}^I$	$\hat{\delta}^{II}$	$\hat{\delta}^{III}$		
Odhady	5.00	5.00	5.00		
Odmocniny empirických rozptylů	0.46	0.36	0.36		
			maticově	*gee*	*lm*
Směrodatné chyby	0.45	0.36	0.36	0.36	0.35
Pravděpodobnosti pokrytí δ intervalem spolehlivosti	0.950	0.945	0.946	0.945	0.936
Odhady α a β	$\hat{\alpha} = -0.01,\ \hat{\beta} = 1.00$				

Tabulka 6.1: Simulační studie - model II, menší rozptyl konečné hodnoty

Parametry	$\mu_0 = 100, \sigma_0^2 = 4, n_C = 500, q = 0.7, \pi = 0.41,$ $c = 500, d = 50, \delta = 5$				
	$\hat{\delta}^I$	$\hat{\delta}^{II}$	$\hat{\delta}^{III}$		
Odhady	4.98	4.98	4.98		
Odmocniny empirických rozptylů	1.61	1.58	1.58		
			maticově	*gee*	*lm*
Směrodatné chyby	1.63	1.60	1.60	1.60	1.59
Pravděpodobnosti pokrytí δ intervalem spolehlivosti	0.955	0.955	0.956	0.955	0.954
Odhady α a β	$\hat{\alpha} = -0.04,\ \hat{\beta} = 1.00$				

Tabulka 6.2: Simulační studie - model II, větší rozptyl konečné hodnoty

Parametry	$\mu_0 = 10, n_C = 500, q = 0.7, \pi = 0.41,$ $a = 50, b = 2, c = 10, d = 5, \delta = 10$				
	$\hat{\delta}^I$	$\hat{\delta}^{II}$	$\hat{\delta}^{III}$		
Odhady	10.02	10.01	10.00		
Odmocniny empirických rozptylů	1.41	0.73	0.24		
			maticově	*gee*	*lm*
Směrodatné chyby	1.42	0.74	0.24	0.24	0.24
Pravděpodobnosti pokrytí δ intervalem spolehlivosti	0.954	0.952	0.953	0.952	0.943
Odhady α a β	$\hat{\alpha} = 50.00,\ \hat{\beta} = 2.00$				

Tabulka 6.3: Simulační studie - model III, směrnice větší než 0.5

Parametry	$\mu_0 = 10, n_C = 500, q = 0.7, \pi = 0.41,$ $a = 50, b = 0.2, c = 10, d = 5, \delta = 5$				
	$\hat{\delta}^I$	$\hat{\delta}^{II}$	$\hat{\delta}^{III}$		
Odhady	5.00	5.00	5.00		
Odmocniny empirických rozptylů	0.30	0.61	0.24		
			maticově	*gee*	*lm*
Směrodatné chyby	0.28	0.61	0.24	0.24	0.24
Pravděpodobnosti pokrytí δ intervalem spolehlivosti	0.950	0.953	0.950	0.949	0.940
Odhady α a β	$\hat{\alpha} = 50.00,\ \hat{\beta} = 0.2$				

Tabulka 6.4: Simulační studie - model III, směrnice menší než 0.5

- počáteční hodnota je z normálního rozdělení,
- ϵ_i^* pochází z normálního rozdělení,
- $h(x) = \log(x+1)$,
- $\psi(x, y) = \sqrt{x} + dy$.

Vidíme, že směrodatné chyby $\hat{\delta}^{III}$ se opět liší, stejně tak i pravděpodobnosti pokrytí devadesátipětiprocentních intervalů spolehlivosti. Pravděpodobnost pokrytí nejvíce vzdálenou 0.95 má konfidenční interval na základě naivního odhadu asymptotického rozptylu. Z odhadů parametru δ je nejhorší $\hat{\delta}^{II}$, který má dvojnásobnou směrodatnou chybu než ostatní dva. Dá se to vysvětlit tím, že závislost konečné hodnoty na počáteční se v číslech kolem střední hodnoty počáteční hodnoty μ_0 „linearizuje“ s velmi malou směrnicí blízkou nule, tedy vzdálenou od 1. Model II proto není vůbec vhodný a stejně tak ani odhad δ na jeho základě. Naopak model I je vhodný stejně jako model III s velmi malým nezáporným parametrem β.

Situaci, kdy

- počáteční hodnota je z exponenciálního rozdělení,
- ϵ_i^* pochází z gamma rozdělení,
- $h(x) = x^2$,
- $\psi(x, y) = \sqrt{x} + dy$,

popisuje tabulka 6.7. Nejmenší směrodatnou chybu má odhad $\hat{\delta}^{III}$, odhady $\hat{\delta}^{II}$ a $\hat{\delta}^{I}$ jsou oproti němu „špatné“ s tím, že o něco lepší je odhad $\hat{\delta}^{II}$. Na rozdíl od minulého případu se to dá vysvětlit tím, že v těchto datech při proložení závislosti konečné hodnoty na počáteční přímkou by směrnice β byla hodně veliká, vhodný je tedy model III a odhad parametru δ na jeho základě. Zároveň model II popisuje situaci o něco lépe než model I. Kdyby se zmenšila střední hodnota počáteční hodnoty μ_0, rozdíl mezi směrodatnými chybami odhadů parametru δ by nebyl tak velký.

Podívejme se nyní na extrémní případ, kdy střední hodnota konečné hodnoty vůbec nezávisí na počáteční. Uvažujme situaci, kdy

- počáteční hodnota je z logaritmicko-normálního rozdělení,
- ϵ_i^* pochází z normálního rozdělení,

Parametry	$\mu_0 = 10, \sigma_0^2 = 1, n_C = 500, q = 0.7, \pi = 0.41,$ $c = 0.01, d = 50, \delta = 5$				
	$\hat{\delta}^I$	$\hat{\delta}^{II}$	$\hat{\delta}^{III}$		
Odhady	5.00	5.00	5.00		
Odmocniny empirických rozptylů	0.81	0.81	0.81		
			maticově	*gee*	*lm*
Směrodatné chyby	0.81	0.81	0.81	0.81	0.79
Pravděpodobnosti pokrytí δ intervalem spolehlivosti	0.954	0.954	0.956	0.954	0.945
Odhady α a β	$\hat{\alpha} = 1.59,\ \hat{\beta} = 0.16$				

Tabulka 6.5: Simulační studie - model I, $h(x) = \sqrt{x}$

Parametry	$\mu_0 = 100, \sigma_0^2 = 4, n_C = 500, q = 0.7, \pi = 0.41,$ $d = -5, \delta = 5$				
	$\hat{\delta}^I$	$\hat{\delta}^{II}$	$\hat{\delta}^{III}$		
Odhady	5.00	5.00	5.00		
Odmocniny empirických rozptylů	0.17	0.33	0.17		
			maticově	*gee*	*lm*
Směrodatné chyby	0.17	0.33	0.17	0.17	0.20
Pravděpodobnosti pokrytí δ intervalem spolehlivosti	0.949	0.948	0.950	0.948	0.962
Odhady α a β	$\hat{\alpha} = 3.62,\ \hat{\beta} = 0.01$				

Tabulka 6.6: Simulační studie - model I, $h(x) = \log(x + 1)$

- $h(x) = a$,
- $\psi(x, y) = cx^2 + dy$.

Výsledky jsou v tabulce 6.8, v tomto případě jsou výjimečně μ_0 a σ_0 parametry normálního rozdělení zlogaritmované počáteční hodnoty. Vzhledem k velikosti směrodatné chyby je nejhorší odhad $\hat{\delta}^{II}$. Naopak odhady $\hat{\delta}^{I}$ a $\hat{\delta}^{III}$ jsou stejně dobré. Dá se to vysvětlit tím, že závislost mezi konečnou a počáteční hodnotou není lineární s nenulovou směrnicí, směrnice 1 z modelu I je tudíž špatně a stejně tak i odhad $\hat{\delta}^{II}$. Naopak, odhadne-li model III směrnici nulou, je odhad $\hat{\delta}^{III}$ stejně dobrý jako $\hat{\delta}^{I}$.

Parametry	$\mu_0 = 5, n_C = 500, q = 0.7, \pi = 0.41,$ $d = 5, \delta = 50$				
	$\hat{\delta}^I$	$\hat{\delta}^{II}$	$\hat{\delta}^{III}$		
Odhady	50.05	50.05	50.03		
Odmocniny empirických rozptylů	7.75	7.44	3.46		
			maticově	*gee*	*lm*
Směrodatné chyby	7.81	7.50	3.39	3.39	3.44
Pravděpodobnosti pokrytí δ intervalem spolehlivosti	0.950	0.950	0.952	0.951	0.954
Odhady α a β	$\hat{\alpha} = -49.54,\ \hat{\beta} = 19.88$				

Tabulka 6.7: Simulační studie - model I, $h(x) = x^2$

Parametry	$\mu_0 = 5, \sigma_0^2 = 0.5, n_C = 500, q = 0.7, \pi = 0.41,$ $a = 200, c = 0.001, d = 5, \delta = 150$				
	$\hat{\delta}^I$	$\hat{\delta}^{II}$	$\hat{\delta}^{III}$		
Odhady	150.00	149.92	150.00		
Odmocniny empirických rozptylů	0.44	6.28	0.44		
			maticově	*gee*	*lm*
Směrodatné chyby	0.44	6.26	0.44	0.44	0.44
Pravděpodobnosti pokrytí δ intervalem spolehlivosti	0.950	0.949	0.950	0.949	0.946
Odhady α a β	$\hat{\alpha} = 200.01,\ \hat{\beta} = 0.00$				

Tabulka 6.8: Simulační studie - model I, $h(x) = a$

Kapitola 7

Závěr

Cílem této práce bylo porovnat tři způsoby odhadu efektu léčby v klinických randomizovaných studiích. Pacientům byla změřena hodnota jisté veličiny na začátku, dále byli náhodně rozděleni do dvou skupin, z nichž jedna dostala léčbu a druhá ne, a ta samá veličina jim byla změřena na konci. Za efekt léčby jsme považovali rozdíl ve střední hodnotě náhodné veličiny změřené na konci u pacieta, který léčbu dostal, a u toho, který ne. Omezili jsme se jen na případ, kdy byl efekt léčby konstantní, tedy stejný pro všechny jedince, kteří danou léčbu dostali. Dále jsme předpokládali, že zkoumaná veličina má v každé skupině konstantní rozptyl. Odhady efektu léčby byly spočteny na základě tří modelů, uvedených v sekcích 3.1 (model I), 3.2 (model II) a 3.3 (model III), označili jsme je $\hat{\delta}^I, \hat{\delta}^{II}$ a $\hat{\delta}^{III}$. Situace, kdy platí jednotlivé modely, jsou do sebe vnořeny jako na obrázku 4.1. Všechny tři odhady efektu léčby jsme porovnávali vždy za situace, kdy platí nějaký z modelů. Za kritéria pro porovnání jsme vzali nejprve základní vlastnost odhadu - konzistenci a dále asymptotický rozptyl jednotlivých odhadů. Přesným rozdělením odhadů jsme se nezabývali, to nebylo v našich silách, zkoumali jsme vždy jen rozdělení asymptotické.

Ve všech devíti případech (tři způsoby odhadu za platnosti tří různých modelů) jsme zjistili, že konzistence platí. Kritériem pro to, který odhad je za dané situace nejlepší, byl tedy asymptotický rozptyl. V kapitolách 3 a 4 jsme vždy spočítali asymptotický rozptyl veličin $\sqrt{n}\left(\hat{\delta} - \delta\right)$, kde n je počet pozorování, δ je efekt léčby a $\hat{\delta}$ je nějaký jeho odhad za platnosti některého ze tří modelů. V případě platnosti modelu I jsme ale asymptotický rozptyl odhadu $\sqrt{n}\left(\hat{\delta}^{III} - \delta\right)$, který je sendvičový, nemohli vyjádřit obecně, proto jsme se v kapitole 5 podívali jen na některé konkrétní příklady. Za platnosti modelu II vyšel nejlépe odhad $\hat{\delta}^{II}$, který byl ale shodný s odhadem $\hat{\delta}^{III}$. Za platnosti modelu III byl nejlepší odhad $\hat{\delta}^{III}$. Za platnosti modelu I v jednotlivých příkladech v kapitole 5 vyšel také nejlépe odhad $\hat{\delta}^{III}$ s tím, že odhad $\hat{\delta}^I$ byl v některých případech skoro stejně dobrý. V

kapitole 6 jsme se podívali na simulace z jednotlivých modelů a opět jsme porovnávali jednotlivé odhady na základě odhadů jejich rozptylů (tentokrát odmocněných, tedy vlastně na základě směrodatných chyb). Když jsme simulovali z modelu II nebo z modelu III, výsledky odpovídaly teoretickým závěrům. Při simulacích z modelu I vyšel vždy nejlépe odhad $\hat{\delta}^{III}$, jehož sendvičový asymptotický rozptyl byl nejmenší. Pokud se zároveň nedal vztah mezi konečnou a počáteční hodnotou proložit přímkou s nenulovou směrnicí, stejně dobrý byl i odhad $\hat{\delta}^{I}$.

Z důvodů uvedených v předchozím odstavci bych v situaci, kterou jsme se zabývali, tedy je-li konstantní efekt léčby a jsou-li v obou léčebných skupinách konstantní roztyly, doporučila vždy použít odhad efektu léčby $\hat{\delta}^{III}$. Pouze jsme-li si jisti, že data pocházejí z modelu II, můžeme použít odhad $\hat{\delta}^{II}$, který je snadněji spočitatelný i bez statitických programů. Odhad $\hat{\delta}^{I}$ bych doporučila použít jen v případě, kdy víme, že při proložení lineární regrese daty je odhad směrnice nulový. To znamená v případě, kdy máme například k dispozici obrázek dat, ze kterého jasně vidíme, že linearita není možná, nebo v případě, kdy někdo už lineární regresi použil a odhad směrnice, který mu vyšel, byl nulový. Odhad $\hat{\delta}^{I}$ je spočitatelný nejsnadněji. Máme-li ale k dispozici jakýkoli statistický program, tak při použití $\hat{\delta}^{III}$ chybu určitě neuděláme, což se při použití ostatních dvou odhadů říci nedá.

Opět ale musíme zdůraznit, že jsou tyto výsledky získané na základě dvou předpokladů: konstantní efekt léčby a konstatní rozptyly ve skupinách. Případ, kdy je efekt léčby u každého jedince jiný, bychom doporučili k dalšímu zkoumání. Stejně tak situaci, kdy ve skupinách nejsou konstatní rozptyly, anebo případ, kdy platí obojí. Dále by bylo zajímavé podívat se na situaci, kdy je měření počáteční nebo konečné hodnoty prováděno s chybou.

Vraťme se nakonec k článkům, které byly motivem k sepsání této práce. Článek Senn (2006) reaguje na názory, které upřednostňují model II (simple analysis of change score, „SACS“) proti modelu III (analysis of covariance, „ANCOVA“) v případě, kdy nejsou shodné střední hodnoty v obou skupinách E a C. Ukazuje, že není nezbytnou podmínkou pro vhodnost modelu III, aby se střední hodnoty shodovaly. Protože jsme se ale zabývali randomizovanými studiemi, které mají základní předpoklad shodu rozdělení na počátku v obou skupinách, řešili jsme vlastně jiný problém, než je ve článku Senn (2006) uvedený.

Článek Tu and Gilthorpe (2007) zase pojednává o tom, že by počáteční hodnota neměla být regresorem pro změnu mezi konečnou a počáteční hodnotou (změna je chápána jako rozdíl mezi konečnou a počáteční hodnotou). Jako důvod se zde uvádí chyba měření počáteční a konečné hodnoty, která se promítne i do změny. Chybový člen počáteční hodnoty, který se objeví v regresoru i odezvě, tak ovlivní jejich vztah. Článek dále shrnuje a na základě simulací porovnává různá řešení navržená v literatuře. V naší práci jsme se ale nezabývali případem, kdy dochází k chybě měření, proto pro nás nebyly výsledky v tomto článku využitelné.

Literatura

Anděl J, 2003. *Statistické metody*. Matfyzpress.

Anděl J, 2007. *Základy matematické statistiky*. Matfyzpress.

Drábková A, 2009. Vliv chyb měření na tvar regresní funkce v nelineárním modelu. Master's thesis, Univerzita Karlova, Matematicko-fyzikální fakulta.

Funatogawa I, 2007. LETTER TO EDITOR - Change from baseline and analysis of covariance revisited. *Statistics in medicine* **26**: 3205–3212.

Senn S, 2006. Change from baseline and analysis of covariance revisited. *Statistics in medicine* **25**: 4334–4344.

Tu YK, Gilthorpe MS, 2007. Revisiting the relation between change and initial value: A review and evaluation. *Statistics in medicine* **26**: 443–457.

White R, 1982. Maximum likelihood estimation of misspecified models. *Econometrica* **50**: 1–26.