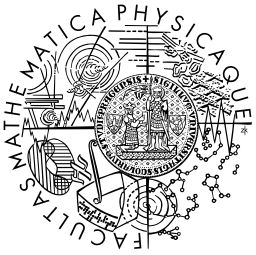


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Willy Svoboda

Metoda upraveného skóre pro lineární model s chybami v regresorech

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. Mgr. Michal Kulich, Ph.D.

Studijní program: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2024

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Velké poděkování patří všem, kteří mi pomohli při mé cestě na Matfyz a podporovali mě během veškerého dosavadního studia na něm. Z celého srdce chci vyjádřit nesmírnou vděčnost mé mamce, bez které bych si nejen studijní sny nemohl plnit.

Obrovské poděkování patří i vedoucímu diplomové práce, doc. Michalovi Kulichovi, za veškeré odborné, profesionální ale i chápavé a velice nápomocné vedení této práce, byla to pro mě čest s ním spolupracovat.

Název práce: Metoda upraveného skóre pro lineární model s chybami v regresorech

Autor: Bc. Willy Svoboda

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. Mgr. Michal Kulich, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Chyba v měření regresoru v lineární regresi může obecně způsobit nejen vychýlení odhadů parametrů, ale i například potíže pro testování hypotéz. Proto se práce věnuje metodě upraveného skóre, kterou představí na základním typu chybového modelu a nadále se ji snaží rozšířit pro zobecněný chybový model, kde špatné naměření regresoru nemusí mít nestrannou chybu, ale je obecně zašuměním lineární kombinace všech regresorů. Pro odhady hlavního i chybového modelu je v práci odvozené sdružené asymptotické rozdělení, které je pak zkoumáno v simulační části při pokrývání skutečných hodnot parametrů intervaly spolehlivosti a testování hypotéz.

Klíčová slova: skóre, metoda upraveného skóre, chyba v měření regresoru, chybně naměřené vysvětlující proměnné, lineární regrese, validační skupina, korekce vychýlení, testování vlivu regresoru

Title: Corrected score method for covariate measurement error in linear models

Author: Bc. Willy Svoboda

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Mgr. Michal Kulich, Ph.D., Department of Probability and Mathematical Statistics

Abstract: Measurement error in the regressor in linear regression can generally cause not only bias in parameter estimates but also problems in hypothesis testing, for example. Therefore, this work focuses on the corrected score method, which is presented using a basic type of error model, and then attempts to extend it to a generalized error model, where the mismeasurement of the regressor may not have an unbiased error but is generally a noise of a linear combination of all regressors. For the estimates of both the primary and error models, the joint asymptotic distribution is derived in this thesis, which is then examined in the simulation section for covering the true parameter values with confidence intervals and for hypothesis testing.

Keywords: score, corrected score method, measurement error in the regressor, mismeasured explanatory variables, linear regression, validation group, bias correction, testing the effect of the regressor

Obsah

Úvod

Situace, ve které alespoň jeden z regresorů naměříme s chybou, značně vychyluje odhady parametrů, znemožňuje správné otestování vlivu regresorů, zkresluje závislosti odezvy a dalších veličin na špatně naměřený regresor či dokonce maskuje samotnou povahu dat. A to vše probíhá i pro nejjednodušší případ naměření s chybou, kdy tato chyba má nulovou střední hodnotu a nezávisí na jiných regresorech.

Práce se věnuje jednomu z možných přístupů, jak lze přistoupit k úpravám odhadů parametrů tak, aby nedocházelo ke konzistentnímu vychýlení odhadů a zároveň aby bylo možné míti intervalové odhady, které se svým pokrytím skutečné hodnoty blíží předepsané spolehlivosti. Zároveň tedy chceme, aby bylo možné testovat hypotézy o vlivu daných regresorů na odezvu, čemuž se včetně zkoumání samotných odhadů bude věnovat simulační část práce.

V 1. kapitole nejprve připomínáme obecně známé základy lineární regrese, definujeme odhad metodou nejmenších čtverců, skóre a skórovou statistiku a uvádíme některé důležité vlastnosti, o které se dále opíráme. Větší část kapitoly se zaměřuje na motivaci, proč se vůbec zabývat případy se špatně naměřenými regresory a nakolik velké důsledky může ignorování chyby v měření mít. Poukazujeme i na to, proč dochází k vychýlení odhadu metodou nejmenších čtverců. Dále se zmiňujeme o různých typech modelů pro chybu měření a krátce se zmiňujeme o odlišnostech pro Berksonův chybový model, zatímco v práci se zaměřujeme na klasické pojetí aditivní chyby. Na závěr kapitoly se věnujeme některým metodám pro analýzu dat s chybami v regresorech.

Zatímco vše převzaté spadá do 1. kapitoly, u dalších kapitol pouze přebíráme obecnou myšlenku použití metody upraveného skóre, kterou ilustrujeme ve 2. kapitole, a dále jen zobecňujeme bez využití jiné literatury, tedy vše v pozdějších kapitolách je už vlastním přínosem autora.

Ve zmíněné 2. kapitole připomínáme souvislost skóre a odhadu medotou nejmenších čtverců. Dle metody upraveného skóre dojdeme k úpravě samotného skóre tak, aby ve střední hodnotě bylo nevychýlené, a z toho pak odvodíme patřičné korekce, které naše odhady mají vůči explicitnímu vyjádření odhadů metodou nejmenších čtverců. Díky tomu i poukážeme, proč absence korekce může mít za následek vychýlení odhadů pro regresory, které jsou naměřeny správně, tedy prokázání, že chyba v měření jistého regresoru nemá vliv jen na parametr daného regresoru.

Nicméně uvažování základního chybového modelu, který předpokládá u měření regresorů chyby s nulovou podmíněnou střední hodnotou, je značně omezující. Proto ve 3. kapitole uvažujeme zobecněný chybový model, ve kterém může chyba v měření regresoru být obecně vychýlená. Odhad daného regresoru uvažujeme obecněji jako zašuměnou lineární kombinaci všech regresorů. I pro tento chybový model ukážeme aplikaci metody upraveného skóre a odvodíme příslušné korekce v explicitním vyjádření odhadů daných parametrů za předpokladu, že parametry chybového modelu jsou známé.

Situaci, kdy parametry chybového modelu nejsou známé, řešíme ve 4. kapitole pomocí využití validační skupiny. Odvozujeme soustavu odhadovacích rovnic pro parametry hlavního i chybového modelu a pro tyto modely velice podrobně a pečlivě odvozujeme asymptotické rozdělení. Toto je hlavním přínosem této práce. Jak moc je tato asymptotika použitelná pro praktické účely, zkoumáme v simulační části, kterou provádíme v 5. kapitole. Sledujeme nakolik velký vliv mají změny velikosti rozsahu a změny podílu validační skupiny vůči celkovému rozsahu. Testujeme, zda model falešně nepřisuzuje nenulový vliv regresorům, pro které jsou příslušné parametry nastaveny na 0. Ve speciálním případě, kdy uvažujeme chybu mající nulovou podmíněnou střední hodnotu, porovnáváme naše výsledky vůči odhadům, které by byly získány neupravenou metodou nejmenších čtverců.

Dále upozorňujeme, že ke zlepšení grafické podoby některých výstupů a obrázků jsme vyyužili některé příkazy navrhnuté AI, nicméně to se netýká samotného textu.

1 Motivace ke studování situací s chybně naměřeným regresorem v lineárních modelech

V této kapitole jsou převzaty znalosti zejména z [1]. Znalosti z lineární regrese jsou zde uvedeny pro úplnost a možnost je následně bez větší potřeby osvětlování využívat při zavádění metody upraveného skóre, avšak se jedná o naprosté základy, které by typický čtenář se vzděláním v oblasti matematické statistiky měl plně znát. Vlastní prací autora v této kapitole je simulování situací s vytvořením grafů pro lepší pochopení probíraného tématu, dále v podkapitole 1.2 ukázání vychýlenosti odhadu metodou nejmenších čtverců v případě, že regresory jsou měřeny s chybou, a poskytnutí literární rešerse na dané téma, které je na konci kapitoly. Jedná se o úvodní kapitolu pro seznámení se s tématem, hlavním přínosem autora jsou až další kapitoly.

1.1 Lineární regrese

Naše práce bude využívat lineární regresi, proto si nejprve zadefinujeme lineární model, odhad metodou nejmenších čtverců a provedeme jeho odvození i ukázání některých vlastností. Nakonec zadefinujeme skóre, od kterého se bude odvíjet naše práce, a odvodíme několik důležitých asymptotických rozdělení.

1.1.1 Lineární regresní model

Uvažujme posloupnost nezávislých náhodných vektorů $(Y_i, \boldsymbol{X_i}^T)^T, i = 1, \ldots, n$. Y_i se nazývá odezva, ale též v českém jazyce se využívá označení závislá proměnná. $\boldsymbol{X_i} = (X_{i1}, \ldots, X_{ip})^T$ je p-rozměrný vektor regresorů, v českém jazyce se složky tohoto vektoru nazývají též jako vysvětlující/nezávislé proměnné nebo prediktory. Na tomto místě upozorníme, že v práci budeme využívat pro vektory a matice tučné značení jako jsme využili u $\boldsymbol{X_i}$, zatímco X_i bude značit jednorozměrný případ, tedy $p = 1$.

Obvykle se $(Y_i, \boldsymbol{X_i}^T)^T$ uvažuje jako náhodný výběr z $(p + 1)$-rozměrného rozdělení. Nicméně v některých situacích se v lineární regresi uvažuje $\boldsymbol{X_i}$ jako pevně dané hodnoty a jediná náhoda je v odezvě.

Definice 1. *Nezávislá pozorování* $(Y_i, \boldsymbol{X_i}^T)^T, i = 1, \ldots, n$ *pocházejí z* lineárního regresního modelu $\iff$ $Y_i = \boldsymbol{X_i}^T\boldsymbol{\beta} + \epsilon_i$ *pro nějaké* $\boldsymbol{\beta} \in \mathbb{R}^p$, *kde chyby* ϵ_i *splňují* $E[\epsilon_i \mid \boldsymbol{X_i}] = 0$ *a* $var(\epsilon_i \mid \boldsymbol{X_i}) = \sigma_\epsilon^2$.

Obvykle se první složka $\boldsymbol{X_i}$ pokládá rovna 1. Potom pro $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ se říká, že β_1 je absolutní člen (nebo též některými využívané označení „intercept“). Absolutní člen odpovídá podmíněné střední hodnotě Y_i pro případ, kdy mimo X_{i1} jsou všechny regresory nulové. Ne vždy má tento člen interpretační smysl, avšak pokud to lze jednoduše provést, volí se vhodná parametrizace taková, že β_1 odpovídá podmíněné střední hodnotě pro zvolenou referenční skupinu.

Zvolme $\boldsymbol{x_i} \in \mathbb{R}^p$ a pro j splňující $2 \leq j \leq p$ uvažujme $\boldsymbol{x_i^j}$, který má oproti $\boldsymbol{x_i}$ zvětšenou j-tou souřadnici o 1, zatímco ostatní složky jsou shodné. Pro tuto situaci j-tý parametr vektoru $\boldsymbol{\beta}$ pomocí vztahu

$$\beta_j = \mathsf{E}[Y_i \mid \boldsymbol{x_i^j}] - \mathsf{E}[Y_i \mid \boldsymbol{x_i}]$$

interpretujeme jako změnu podmíněné střední hodnoty Y_i, když se zvýší X_j o 1, zatímco ostatní regresory zůstanou stejné. Avšak je potřeba upozornit, že ne ve všech modelech lze tuto interpretaci použít a je třeba přistupovat k tomu ad hoc. Jednoduchým příkladem je model, kde $X_3 = X_2^2$, tedy třetí regresor je druhou mocninou druhého regresoru. V tomto případě by bylo nemožné zvětšit X_2 o 1 a nechat X_3 zafixovanou na stejné hodnotě.

1.1.2 Odhad metodou nejmenších čtverců

Základním a hlavním odhadem v lineární regresi je odhad metodou nejmenších čtverců. V definici uvedeme jak zápis po jednotlivých pozorováních, tak i zápis v maticové podobě, pro který využijeme značení

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \text{ a } \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X_1}^T \\ \vdots \\ \boldsymbol{X_n}^T \end{pmatrix}.$$

Definice 2. Odhad metodou nejmenších čtverců $\hat{\boldsymbol{\beta}}$ *vektoru* $\boldsymbol{\beta}$ *je takový bod v* $\mathbb{R}^p$*, který minimalizuje součet čtverců reziduí*

$$SS_e(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(Y_i - \boldsymbol{X_i}^T\boldsymbol{\beta}\right)^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Odvodíme si vyjádření tohoto odhadu. Maticový zápis $SS_e(\boldsymbol{\beta})$ lze jednoduše přepsat do podoby

$$\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}.$$

Pro derivaci $SS_e(\boldsymbol{\beta})$ využijeme pravidla pro derivování vektorů:

$$\frac{\mathrm{d}\boldsymbol{\beta}^T\boldsymbol{C}}{\mathrm{d}\boldsymbol{\beta}} = \boldsymbol{C} \text{ a } \frac{\mathrm{d}\boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta}}{\mathrm{d}\boldsymbol{\beta}} = \boldsymbol{2A\beta},$$

pokud matice $\boldsymbol{A}$ je symetrická. Potom derivaci $SS_e(\boldsymbol{\beta})$ podle $\boldsymbol{\beta}$ položíme rovnu nulovému vektoru a danou soustavu rovnic vyřešíme pro $\boldsymbol{\beta}$:

$$\frac{\mathrm{d}SS_e(\boldsymbol{\beta})}{\mathrm{d}\boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{Y} + 2(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} \stackrel{!}{=} \boldsymbol{0_{p,1}},$$

kde nyní i později v práci využíváme značení $\boldsymbol{0_{p,q}}$ pro matici (respektive vektor, pokud právě jedno z p, q bude rovné 1) o p řádcích a q sloupcích, které obsahují jen nuly. Pak $\hat{\boldsymbol{\beta}}$ řeší soustavu p lineárních rovnic o p neznámých, kde soustavu

$$(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$$

nazveme jako soustavu normálních rovnic. Pro další podkapitolu je vhodné upozornit, že soustavu lze přepsat do podoby po složkách:

$$\sum_{i=1}^{n} \boldsymbol{X_i}\left(Y_i - \boldsymbol{X_i}^T\boldsymbol{\beta}\right) = \boldsymbol{0_{p,1}}.$$

Pokud hodnost matice $\boldsymbol{X}$ je p, pak i hodnost matice $\boldsymbol{X}^T\boldsymbol{X}$ je rovna p a tedy její inverzní matice $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ existuje. **Odhad metodou nejmenších čtverců** můžeme v takovém případě zapsat explicitně

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

Odhad metodou nejmenších čtverců zřejmě minimalizuje $SS_e(\boldsymbol{\beta})$, protože pro všechny $\boldsymbol{\beta} \in \mathbb{R}^p$ platí

$$\frac{\mathrm{d}^2 SS_e(\boldsymbol{\beta})}{\mathrm{d}\boldsymbol{\beta}\mathrm{d}\boldsymbol{\beta}^T} = +2(\boldsymbol{X}^T\boldsymbol{X}) > 0,$$

tedy $SS_e(\boldsymbol{\beta})$ je striktně konvexní pro všechny $\boldsymbol{\beta} \in \mathbb{R}^p$ a tedy $\hat{\boldsymbol{\beta}}$ je globálním minimem $SS_e(\boldsymbol{\beta})$. Odhad metodou nejmenších čtverců je nestranným odhadem $\boldsymbol{\beta}$, neboť

$$\begin{aligned}\mathsf{E}[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}] = \mathsf{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \mid \boldsymbol{X}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathsf{E}[\boldsymbol{Y} \mid \boldsymbol{X}] = \\ = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\beta}.\end{aligned}$$

S využitím vztahu $\mathrm{var}(\boldsymbol{AX}) = \boldsymbol{A}\mathrm{var}(\boldsymbol{X})\boldsymbol{A}^T$ a značení $\boldsymbol{I_n}$ pro jednotkovou matici řádu n můžeme odvodit i podmíněný rozptyl tohoto odhadu:

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \\ = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\left(\sigma_\epsilon^2\boldsymbol{I_n}\right)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \\ = \sigma_\epsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\cancel{\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}} = \sigma_\epsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.\end{aligned}$$

Jeden z důsledků známého Gaussova-Markovova tvrzení je, že odhad metodou nejmenších čtverců $\hat{\boldsymbol{\beta}}$ je nejlepší nestranný lineární odhad vektoru parametrů $\boldsymbol{\beta}$.

1.1.3 Skóre

Přidejme si předpoklad že $(Y_i, \boldsymbol{X_i}^T)^T, i = 1, \ldots, n$, je náhodný výběr a matice

$$\boldsymbol{V_X} := \mathsf{E}\,\boldsymbol{X_i}\boldsymbol{X_i}^T$$

má konečné prvky a je pozitivně definitní. Definujme **skóre** jako

$$\boldsymbol{U_i}(\boldsymbol{\beta}) = \boldsymbol{X_i}\left(Y_i - \boldsymbol{X_i}^T\boldsymbol{\beta}\right),$$

kde $\boldsymbol{\beta} \in \mathbb{R}^p$. Skóre je též možné nazývat jako skórovou funkci. Pokud sečteme skóre přes všechna pozorování, dostaneme **skórovou statistiku**:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{U_i}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X_i} \left(Y_i - \boldsymbol{X_i}^T \boldsymbol{\beta} \right).$$

Všimněme si, že skórovou statistiku lze přepsat do maticové podoby

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X_i} \left(Y_i - \boldsymbol{X_i}^T \boldsymbol{\beta} \right) = \boldsymbol{X}^T \boldsymbol{Y} - \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta},$$

ze které, pokud ji položíme rovnu nulovému vektoru, lze dostat soustavu normálních rovnic, kterou řeší odhad metodou nejmenších čtverců. Jinými slovy odhad metodou nejmenších čtverců $\hat{\boldsymbol{\beta}}$ je jednoznačným řešením soustavy

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0_{p,1}}.$$

Nepodmíněné momenty $\boldsymbol{U_i}(\boldsymbol{\beta})$ pro skutečné $\boldsymbol{\beta}$ jsou

$$\begin{aligned} \mathsf{E}[\boldsymbol{U_i}(\boldsymbol{\beta})] &= \mathsf{E}\left[\mathsf{E}[\boldsymbol{U_i}(\boldsymbol{\beta}) \mid \boldsymbol{X_i}]\right] = \mathsf{E}\left[\boldsymbol{X_i}\, \mathsf{E}\left[\left(Y_i - \boldsymbol{X_i}^T \boldsymbol{\beta}\right) \mid \boldsymbol{X_i}\right]\right] = \\ &= \mathsf{E}\left[\boldsymbol{X_i}\, \mathsf{E}[\epsilon_i \mid \boldsymbol{X_i}]\right] = \mathsf{E}\left[\boldsymbol{X_i} 0\right] = \boldsymbol{0_{p,1}}, \end{aligned}$$

$$\begin{aligned} \operatorname{var}\left(\boldsymbol{U_i}(\boldsymbol{\beta})\right) &= \mathsf{E}\left[\operatorname{var}\left(\boldsymbol{U_i}(\boldsymbol{\beta}) \mid \boldsymbol{X_i}\right)\right] + \operatorname{var}\left(\mathsf{E}[\boldsymbol{U_i}(\boldsymbol{\beta}) \mid \boldsymbol{X_i}]\right) = \\ &= \mathsf{E}\left[\operatorname{var}\left(\boldsymbol{X_i}\epsilon_i \mid \boldsymbol{X_i}\right)\right] + \operatorname{var}\left(\boldsymbol{0_{p,1}}\right) = \mathsf{E}\left[\boldsymbol{X_i} \operatorname{var}\left(\epsilon_i \mid \boldsymbol{X_i}\right) \boldsymbol{X_i}^T\right] \\ &= \sigma_\epsilon^2\, \mathsf{E}\left[\boldsymbol{X_i}\boldsymbol{X_i}^T\right] = \sigma_\epsilon^2 \boldsymbol{V_X}. \end{aligned}$$

Dostáváme tak že $\boldsymbol{U_1}(\boldsymbol{\beta}), \ldots, \boldsymbol{U_n}(\boldsymbol{\beta})$ ve skutečném $\boldsymbol{\beta}$ pochází z náhodného výběru a mají nulové střední hodnoty. Můžeme tak využít mnohorozměrnou verzi Centrální limitní věty pro nezávislé, stejně rozdělené náhodné vektory:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U_i}(\boldsymbol{\beta})\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{U_i}(\boldsymbol{\beta}) = \frac{1}{\sqrt{n}} \boldsymbol{U}(\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}_p(\boldsymbol{0_{p,1}}, \sigma_\epsilon^2 \boldsymbol{V_X}).$$

Pomocí této asymptotiky můžeme jednoduše odvodit i asymptotické rozdělení pro samotný odhad! Připomeňme, že pro odhad metodou nejmenších čtverců je skórová statistika nulová. Proto můžeme pro skutečné $\boldsymbol{\beta}$ dále upravovat:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{U}(\boldsymbol{\beta}) - \boldsymbol{U}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \left[\boldsymbol{X_i}\left(Y_i - \boldsymbol{X_i}^T \boldsymbol{\beta}\right) - \boldsymbol{X_i}\left(Y_i - \boldsymbol{X_i}^T \hat{\boldsymbol{\beta}}\right)\right] = \sum_{i=1}^{n} \boldsymbol{X_i}\boldsymbol{X_i}^T \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right),$$

což, protože se v každém sčítanci objevuje $\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$, lze po vydělení $\sqrt{n}$ poupravit na

$$\frac{1}{\sqrt{n}}\boldsymbol{U}(\boldsymbol{\beta}) = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}\boldsymbol{X_i}^T\right)\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right),$$

a nakonec můžeme tento vztah díky předpokladu plné hodnosti matice $\boldsymbol{X_i}$ přepsat do výsledného tvaru

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}\boldsymbol{X_i}^T\right)^{-1}\frac{1}{\sqrt{n}}\boldsymbol{U}(\boldsymbol{\beta}).$$

Ze zákona velkých čísel $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}\boldsymbol{X_i}^T$ konverguje k $\boldsymbol{V_X}$, zatímco asymptotiku pro $\frac{1}{\sqrt{n}}\boldsymbol{U}(\boldsymbol{\beta})$ máme vyše uvedenou. Při aplikaci Cramérovy-Slutského věty, kdy pro asymptotický rozptyl máme výpočet $\sigma_\epsilon^2\boldsymbol{V_X^{-1}}\boldsymbol{V_X}\boldsymbol{V_X^{-1}} = \sigma_\epsilon^2\boldsymbol{V_X^{-1}}$, dostáváme asymptotické rozdělení

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}_p(\boldsymbol{0_{p,1}}, \sigma_\epsilon^2\boldsymbol{V_X^{-1}}).$$

Z tohoto asymptotického rozdělení rovnou dostáváme i jako důsledek fakt, že odhad metodou nejmenších čtverců je konzistentním odhadem $\boldsymbol{\beta}$. Tyto výsledky můžeme shrnout do tvrzení.

Věta 1. *Za předpokladů uvedené v této podkapitole pro skutečné $\boldsymbol{\beta}$ platí*

1. $\frac{1}{\sqrt{n}}\boldsymbol{U}(\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}_p(\boldsymbol{0_{p,1}}, \sigma_\epsilon^2\boldsymbol{V_X})$,
2. $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}_p(\boldsymbol{0_{p,1}}, \sigma_\epsilon^2\boldsymbol{V_X^{-1}})$,
3. $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ *(odhad metodou nejmenších čtverců je konzistentním odhadem pro $\boldsymbol{\beta}$).*

Ovšem v praxi se nám bude hodit pro testování a intervaly spolehlivosti jiná asymptotika, proto si uvedeme v zobecněné podobě ještě jednu větu, kterou lze pomocí Cramérovy–Woldovy věty a Cramérovy-Slutského věty odvodit z Věty 1. Označme si $\hat{\sigma_\epsilon^2}$ jako odhad σ_ϵ^2.

Věta 2. *Za předpokladů uvedené v této podkapitole pro skutečné $\boldsymbol{\beta}$ a libovolné $\boldsymbol{c} \in \mathbb{R}^p$ platí*

$$\frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\hat{\sigma}_\epsilon\sqrt{\boldsymbol{c}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{c}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Pokud ve Větě 2 dosadíme za $\boldsymbol{c}$ vektor obsahující nuly až na j-tou souřadnici pro $1 \leq j \leq p$, pak nám věta dává asymptotické rozdělení pro β_j, respektive lze z tohoto asymptotického rozdělení odvodit asymptotické intervaly spolehlivosti pro parametr β_j.

1.2 Lze vždy zanedbat chybu v měření?

V této i v dalších podkapitolách se budeme ve velkém opírat o [1], zejména o první dvě kapitoly, které se věnují úvodu do dané problematiky. Tento zdroj se ale snaží k dané problematice přistupovat více obecně a pojmout i případy jako je logistická regrese, zatímco my se čistě zaměříme na lineární regresi, jíchž vlastností budeme plně využívat při výpočtech a odvozování vlastností v modelech, se kterými budeme pracovat v dalších kapitolách.

Během úvodních přednášek lineární regrese se předpokládá, že zatímco regresory $\boldsymbol{X_i}$ jsou naměřeny zcela správně a přesně, pouze odezva Y_i je naměřena s chybou. Nicméně co když i některá vysvětlující proměnná bude naměřena s chybou? Uvažujme jednoduchý příklad, kdy v modelu nemáme absolutní člen, máme pouze jednu vysvětlující veličinu, kterou měříme přesně, označenou jako Z_i, zatímco druhou vysvětlující proměnnou X_i neumíme naměřit přesně, máme pouze nestranný odhad W_i (zde i dále v textu automaticky myslíme u nestrannosti a středních hodnot jejich podmíněné verze na skutečných regresorech, avšak toto nebudeme typicky explicitně zdůrazňovat. Podobně budeme vynechávat zmínění podmíněnosti i u dalšího, například u rozptylu a směrodatných odchylek. Na případné odlišnosti upozorníme, například u Berksonova modelu). Máme tedy chybový model

$$W_i = X_i + \psi_i,$$

kde $\mathsf{E}[\psi_i \mid Z_i, X_i] = 0$ a $i = 1, \ldots, n, n \in \mathbb{N}$. Nechť β_1 odpovídá přesně naměřenému regresoru Z_i a β_2 regresoru X_i, který neumíme přesně naměřit. Náš skutečný model, se kterým bychom ideálně pracovali, má podobu

$$Y_i = \beta_1 Z_i + \beta_2 X_i + \epsilon_i,$$

kde $\mathsf{E}[\epsilon_i \mid Z_i, X_i] = 0$ a $i = 1, \ldots, n, n \in \mathbb{N}$. Avšak jak bylo řečeno, my X_i přímo nepozorujeme a nemůžeme tak tento model využít ani k odhadnutí parametrů ani k žádným dalším věcem, které známe a využíváme pro lineární regresi, včetně predikcí. Prvotní myšlenka by byla k těmto záležitostem využít model

$$Y_i = \beta_1 Z_i + \beta_2 W_i + \epsilon_i,$$

neboť W_i je odhad X_i. Lze absenci X_i ignorovat díky využívání W_i a vždy zanedbávat chybu v měření regresoru? Jistě že by někdo mohl v zájmu co nejmenšího vynaloženého úsilí skutečnost chybného měření regresoru plně ignorovat a pracovat, jako by bylo vše správně naměřeno bez chyby. Ovšem naše intuice nám správně říká, že by to mohlo vést ke katastrofickým důsledkům, což se také může lehce stát. Ale my máme navíc předpoklad nulovosti podmíněné střední hodnoty chyby v měření. Je to dostatečný předpoklad pro ospravedlnění aproximace $\boldsymbol{X_i}$ pomocí $\boldsymbol{W_i}$ v daném modelu? Pojďme se podívat na samotný odhad metodou nejmenších čtverců, zda je odhadem nevychýleným.

Označme si $\boldsymbol{Z}$ jako sloupcový vektor obsahující $Z_1, \ldots, Z_n$ a obdobně $\boldsymbol{W}$ pro $W_1, \ldots, W_n$. Odhad metodou nejmenších čtverců splňuje soustavu normálních rovnic s vektorem proměnných $\boldsymbol{\beta}$:

$$\left(\begin{pmatrix}\boldsymbol{Z}^T\\ \boldsymbol{W}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{Z} & \boldsymbol{W}\end{pmatrix}\right)\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{Z}^T\\ \boldsymbol{W}^T\end{pmatrix}\boldsymbol{Y},$$

což lze upravit na

$$\begin{pmatrix}\boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\boldsymbol{W}\\ \boldsymbol{W}^T\boldsymbol{Z} & \boldsymbol{W}^T\boldsymbol{W}\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{Z}^T\boldsymbol{Y}\\ \boldsymbol{W}^T\boldsymbol{Y}\end{pmatrix}$$

a přepsat do podoby po složkách

$$\begin{pmatrix}\sum_{i=1}^n Z_i^2 & \sum_{i=1}^n Z_iW_i\\ \sum_{i=1}^n Z_iW_i & \sum_{i=1}^n W_i^2\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^n Z_iY_i\\ \sum_{i=1}^n W_iY_i\end{pmatrix}.$$

Na tento vztah chceme aplikovat střední hodnotu podmíněnou na Z a X, proto proveďmě několik pomocných výpočtů. Z chybového modelu přímočaře vyplývá

$$\mathsf{E}[Z_iW_i \mid Z_i, X_i] = Z_i\,\mathsf{E}[X_i + \psi_i \mid Z_i, X_i] = Z_iX_i,$$

navíc pro jednoduchost přidejme k chybovému modelu předpoklad pro rozptyl: $\mathsf{E}[\psi_i^2 \mid Z_i, X_i] = \sigma_\psi^2$ pro $\sigma_\psi^2 > 0$, tedy

$$\mathsf{E}[W_i^2 \mid Z_i, X_i] = \mathsf{E}[X_i^2 + 2X_i\psi_i + \psi_i^2 \mid Z_i, X_i] = X_i^2 + \sigma_\psi^2.$$

Nyní zbývá provést pomocné výpočty pro pravou stranu vztahů obsahujících Y_i:

$$\mathsf{E}[Z_iY_i \mid Z_i, X_i] = Z_i\,\mathsf{E}[\beta_1Z_i + \beta_2X_i + \epsilon_i \mid Z_i, X_i] = \beta_1Z_i^2 + \beta_2Z_iX_i,$$

zatímco pro druhou složku s předpokladem $\mathsf{E}[\epsilon_i\psi_i \mid Z_i, X_i] = \rho$ vychází

$$\mathsf{E}[W_iY_i \mid Z_i, X_i] = \mathsf{E}[(X_i + \psi_i)(\beta_1Z_i + \beta_2X_i + \epsilon_i) \mid Z_i, X_i] = \beta_1Z_iX_i + \beta_2X_i^2 + \rho.$$

Celkově po aplikaci podmíněné střední hodnoty na soustavu rovnic dostáváme

$$\begin{pmatrix}\sum_{i=1}^n Z_i^2 & \sum_{i=1}^n Z_iX_i\\ \sum_{i=1}^n Z_iX_i & \sum_{i=1}^n (X_i^2 + \sigma_\psi^2)\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^n (\beta_1Z_i^2 + \beta_2Z_iX_i)\\ \sum_{i=1}^n (\beta_1Z_iX_i + \beta_2X_i^2 + \rho)\end{pmatrix},$$

čili v lepším tvaru pro porovnávání

$$\begin{pmatrix}\sum_{i=1}^n (\beta_1Z_i^2 + \beta_2Z_iX_i)\\ \sum_{i=1}^n (\beta_1Z_iX_i + \beta_2X_i^2 + \beta_2\sigma_\psi^2)\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^n (\beta_1Z_i^2 + \beta_2Z_iX_i)\\ \sum_{i=1}^n (\beta_1Z_iX_i + \beta_2X_i^2 + \rho)\end{pmatrix}.$$

Platnost rovnic v soustavě by byla jen v okamžiku, kdy by platilo buď $\beta_2 = \frac{\rho}{\sigma_\psi^2}$ a vycházelo by to nenulové, což je v praxi nemožné ani chtěné předpokládat,

nebo $\rho = 0$ a taktéž σ_ψ^2 by muselo být nulové. To by ale nutně znamenalo, že v chybovém modelu s chybou mající nulovou podmíněnou střední hodnotu není žádná variabilita, tedy k žádnému zašumění nedochází a X_i měříme naprosto přesně. To jinými slovy znamená, že i když předpokládáme nestrannost pro chybu při měření regresoru (a mající nenulový rozptyl), tak nutně dochází k vychýlení některých složek v soustavě rovnic

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Z_i^2 & \sum_{i=1}^n Z_i W_i \\ \sum_{i=1}^n Z_i W_i & \sum_{i=1}^n W_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Z_i Y_i \\ \sum_{i=1}^n W_i Y_i \end{pmatrix},$$

což ale, protože dochází k vychýlení v matici, kterou je třeba invertovat, dochází k vychýlení nejen $\hat{\beta}_2$, ale i samotné $\hat{\beta}_1$. Vidíme tedy, že i v nejjednodušším modelu s chybně naměřenou vysvětlující proměnnou dochází i při nestrannosti chyb k vychýlení odhadu metodou nejmenších čtverců! To je velký problém, který by nebylo vhodné přehlížet, pokud je možnost provést nějakou korekci!

Chyba v měření regresoru nemusí být problémem jen pro samotné odhadování parametrů, ale samozřejmě i pro samotné testování. Testy ztrácí sílu proti alternativám, že daná veličina má efekt na odezvu. Tuto situaci vidíme na Obrázku 1.1, ve kterém možnost zachytit vliv regresoru na odezvu se velmi značně vytratila. Situace byla simulována následovně. Vygenerovali jsme si 20 hodnot X_i pocházející z rozdělení $\mathcal{N}(0{,}16)$. Nastavili jsme $\beta_0 = 25$, $\beta_1 = 0$ (žadnou přesně naměřenou veličinu nepozorujeme) a $\beta_2 = 2$. Směrodatnou odchylku ϵ_i jsme nastavili na 6. Pak levý graf na obrázku odpovídá skutečnému modelu

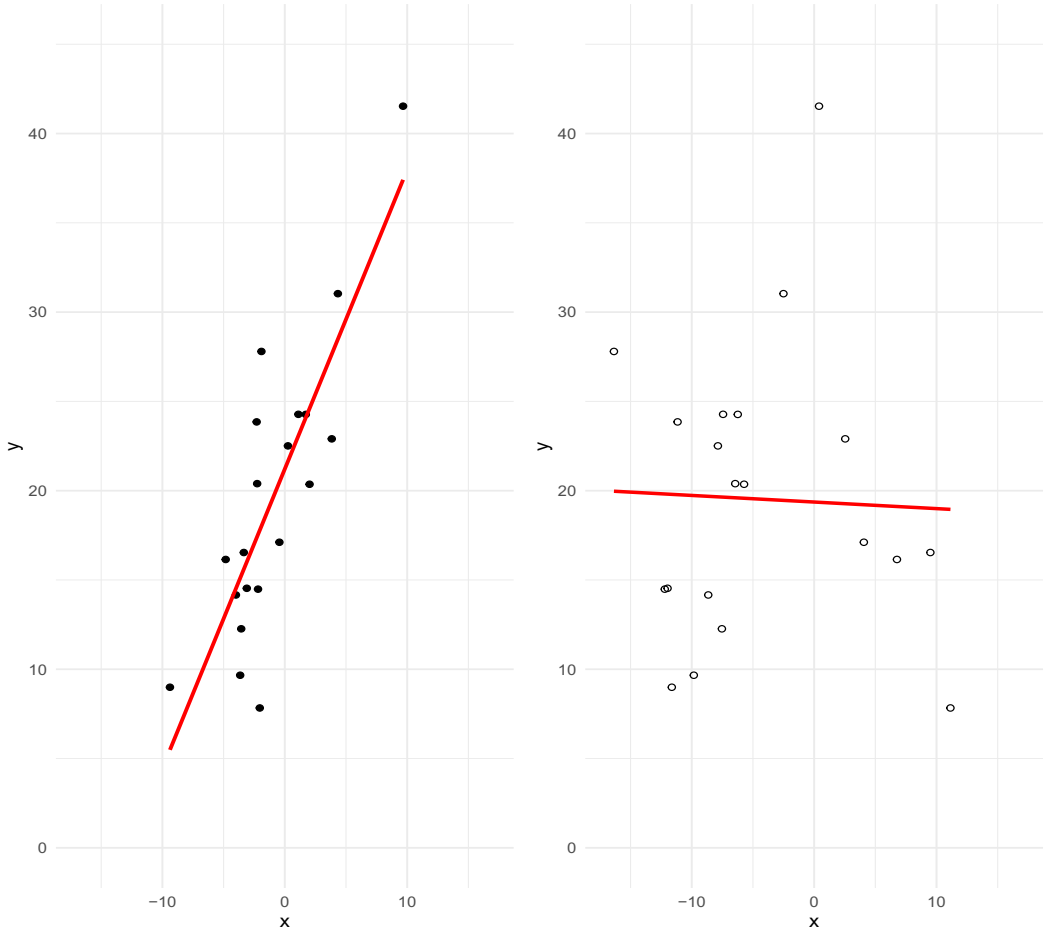
$$Y_i = \beta_0 + \beta_2 X_i + \epsilon_i,$$

kde vidíme, že β_2 je lineárním modelem odhadnuto poměrně dobře, a to zhruba dvojnásobný růst v y-ové souřadnici oproti x-ové souřadnici. Pokud ale X_i nepozorujeme přímo a máme pouze jeho odhad W_i, tak i když je nestranný, lze se lehce dostat do situace, kdy se body „rozprostřou do prostoru“ tak, že lineární trend „zploští“. Typicky tak dojde k vychýlení efektu „směrem k nule“. Pro chybu v měření jsme zvolili směrodatnou odchylku rovnu 8 a střední hodnotu jsme nechali nulovou. Graf odpovídající lineárnímu modelu

$$Y_i = \beta_0 + \beta_2 W_i + \epsilon_i$$

je na pravé straně Obrázku 1.1 a vidíme, že odhadnutý lineární trend není jen extrémně vychýlen směrem k nule, ale dokonce místo roustoucí přímky máme slabě klesající. Celkový rozptyl bodů se výrazně zvětšil oproti původní situaci. Přitom jsme v situaci s nejjednodušší strukturou chyby v měření! Došlo k ohromné ztrátě síly a celkově možnosti detekování závislosti dvou veličin. Toto ve spojení s obecným vychýlením odhadů parametrů nazývají v [1] dvojí pohroma způsobena chybou v měření. Dále seznam rozšiřují o třetí pohromu: zamaskování povahy dat.

Zamaskování povahy dat si můžeme jednoduše vysvětlit pomocí Obrázku 1.2. Opět nalevo vidíme plnými puntíky situaci, které odpovídají situaci napozorování skutečných hodnot X_i, zatímco vpravo prázdné puntíky odpovídají situaci, ve které



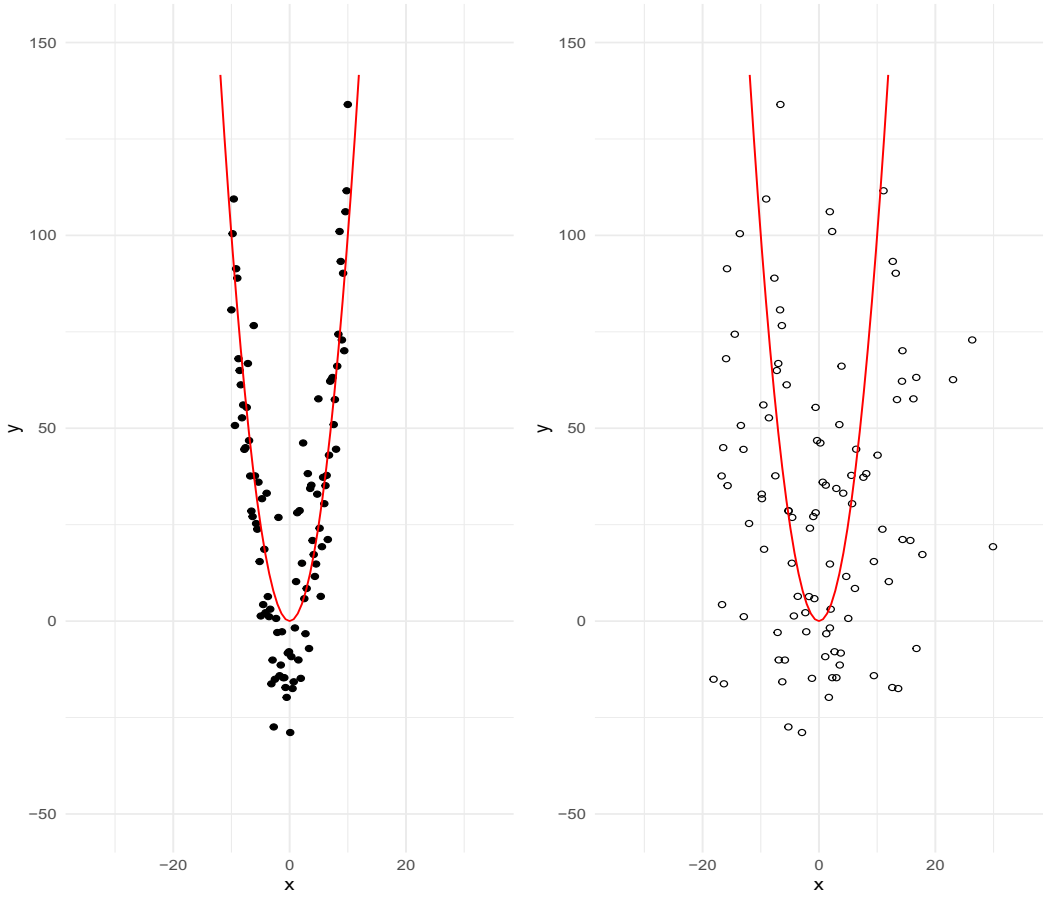
Obrázek 1.1 Chyba v měření regresoru může výrazně změnit závislost mezi veličinami a tím celkově snížit sílu testování efektů veličin na odezvu. Vlevo máme situaci, kdy pro lineární model bychom napozorovali přímo samotné X_i vyznačené plnými puntíky, zatímco vpravo se nachází prázdné body, které mají tuto veličinu odhadovat s chybou (ačkoliv s chybou mající nulovou podmíněnou střední hodnotu) v podobě W_i. Lze si všimnout větší rozptýlenosti bodů než tomu bylo původně. Pomocí jemného mřížkování lze okometricky ověřit že se y-ová souřadnice bodů mezi levým a pravým obrázkem nezměnila.

pracujeme s W_i. Na obou grafech jsme červeně vykreslili křivku $y = x^2$, neboť daných 100 bodů jsme generovali pro skutečný model podle

$$Y_i = \beta_0 + \beta_{2,1} X_i + \beta_{2,2} X_i^2 + \epsilon_i,$$

kde $\beta_0 = 0$, $\beta_{2,1} = 0$ a $\beta_{2,2} = 1$, směrodatná odchylka ϵ_i je 16. I kdybychom v levém grafu červenou parabolickou křivku nevykreslili, tak bychom stále dobře vytušili, že by bylo šikovné využít posledně zmíněný model a odhadnout parametry (i když bychom netipovali x^2, hlavní je, že stále bychom využívali parabolického trendu a to promítli do další práce s daty). Opět v pravé části máme body se stejnou y-ovou souřadnicí, avšak opět proběhlo zašumění X_i pomocí chyby s nulovou střední hodnotou a směrodatnou odchylkou 8. Z vykreslení závislosti Y_i na W_i bychom bez vynačené původní křivky vůbec z grafu nevyčetli parabolickou vlastnost původních dat a při zamaskování této skutečnosti bychom nevyužili výše zmíněný model a nejspíše by se zkoumal jen lineární efekt W_i, neboť z grafu žádnou

významnou informaci pro budování modelu nevyčteme. Chyby v regresorech tak mohou významně zamaskovat vlastnosti přesně naměřených dat a to do takové míry, že to může naprosto změnit naše modely a obecně zobtížnit vyhodnocování dat a deskriptivní analýzu využívající grafy.



Obrázek 1.2 Chyba v měření regresoru může zakrýt důležité vlastnosti dat, které by byly jinak vidět na první pohled. Opět vlevo plné puntíky odpovídají situaci, kdy bychom X_i přímo pozorovali, zatímco prázdné puntíky odpovídají situaci měření s chybou mající nulovou podmíněnou střední hodnotu, neboli kdy máme místo X_i napozorováno W_i. Zatímco u skutečných dat by šlo lehce odhadnout parabolický trend (vztah odezvy vůči vysvětlující proměnné bez chyby v odezvě odpovídá křivce $y = x^2$), u dat zatížené chybně naměřeným regresorem by parabolický trend bez žádné další znalosti vypozorovat už nešel. Lze si také všimnout že zatížením regresoru chybou došlo intuitivně k většímu rozptýlení dat.

1.3 Zobecnění na vícerozměrné případy

Rozdělme si regresory na ty, které naměříme zcela přesně (typicky sem patří věk, biologické pohlaví, stupeň nejvyššího dosaženého vzdělání, kraj/velikost obce, ve které respondent má trvalé bydliště, a další údaje, u kterých dostatečně věříme, že je lze přesně naměřit nebo respondent nemá tendenci na danou otázku odpovídat nepravdivě) a označme vektor těchto přesně „naměřených“ vysvětlujících proměnných jako $\boldsymbol{Z_i}$ pro dané pozorování, zatímco vysvětlující proměnné, které

jsou zatížené chybou v „měření“, seskupíme do vektoru $\boldsymbol{X_i}$, který avšak nepozorujeme (zde pozor, $\boldsymbol{X_i}$ stále udává skutečnou přesnou hodnotu!). Místo něho pozorujeme odhad tohoto vektoru, který budeme značit jako $\boldsymbol{W_i}$. Dále rozdělme vektor parametrů $\boldsymbol{\beta}$ na vektor parametrů pro přesně naměřené regresory $\boldsymbol{\beta_1}$ a na vektor parametrů pro nepřesně naměřené regresory $\boldsymbol{\beta_2}$ a na parametr, který bude odpovídat absolutnímu členu, označme ho β_0. Náš skutečný model tedy bude mít podobu

$$Y_i = \beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \boldsymbol{\beta_2}^T \boldsymbol{X_i} + \epsilon_i,$$

kde $\mathsf{E}[\epsilon_i \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] = 0$ a $i = 1, \ldots, n, n \in \mathbb{N}$. Avšak jak bylo řečeno, my $\boldsymbol{X_i}$ přímo nepozorujeme a nemůžeme tak tento model využít ani k odhadnutí parametrů ani k žádným dalším věcem, které známe a využíváme pro lineární regresi, včetně predikcí. Jak jsme viděli v minulé podkapitole, tak i kdyby $\boldsymbol{W_i}$ byl nestranným odhadem $\boldsymbol{X_i}$, tedy

$$\boldsymbol{W_i} = \boldsymbol{X_i} + \boldsymbol{\psi_i},$$

kde $\mathsf{E}[\boldsymbol{\psi_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] = \boldsymbol{0_{p,1}}$ a $i = 1, \ldots, n, n \in \mathbb{N}$, pracovat slepě s modelem

$$Y_i = \beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \boldsymbol{\beta_2}^T \boldsymbol{W_i} + \epsilon_i$$

by bylo sice pohodlné a jednoduché, nikoliv obecně ospravedlněné a správné. I pro nestranné chyby by model mohl vést k mnohým velkým neplechám a zcela zmařit celou práci s daty! O situaci, kdy chyby v měření nejsou nestranné (například jedná se o zašumění lineární kombinace vícera regresorů, což budeme studovat od 3. kapitoly dále) a důsledcích slepého nahrazení $\boldsymbol{W_i}$ za $\boldsymbol{X_i}$ bez žádné korekce ve výpočtech, nemluvě.

Příčiny chybného naměření regresoru mohou být různé. Může se jednat o málo přesné fyzikální měření, hrubé zaokrouhlení, ale třeba i účelové zkreslování a podhodnocování dané veličiny. V podkapitole 1.5 v [1] je zmíněna studie, ve které jednou ze zkoumaných veličin byl denní příjem kalorií. Není nijak překvapivé, že pokud záleželo jen na odpovědi respondenta, tak to bylo opravdu výrazně podhodnoceno. Když se pak histogram odpovědí porovnal s histogramem hodnot, které byly získáné pomocí speciální techniky na odhadnutí skutečného příjmu, tyto dva histogramy byly extrémně odlišné. Proto zkoumání situace měření s chybou mající nulovou střední hodnotu není rozhodně dostatečné. Hodí se zkoumati složitější chybové modely (jednomu z nich se budeme rozsáhle věnovat v pozdějších kapitolách).

1.4 Typy modelů pro chybu měření

Nejjednodušší typ chybového modelu jsme popsali v dřívější podkapitole, mající podmíněnou střední hodnotu rovnu nule. Tímto chybovým modelem, kdy samotné $\boldsymbol{X_i}$ jsou nějakým způsobem rozptýleny a dále pomocí chyby vzniklé nepřesným naměřením zatížíme data další chybou, přidáváme do rozptylu skutečných dat další rozptýlení způsobené $\boldsymbol{\psi_i}$. Data pracující s $\boldsymbol{W_i}$ tak mají celkově větší variabilitu než mají původní data s $\boldsymbol{X_i}$. Odhad se tedy skládá z přesné veličiny a chyby, čili

$$\boldsymbol{W_i} = \boldsymbol{X_i} + \boldsymbol{\psi_i}.$$

Tento chybový model lze nazvat jako klasický model aditivní chyby měření. Setkali jsme se s ním i u situací, které jsou znázorněny na Obrázcích 1.1 a 1.2. Úplně jiný typ chybového modelu se nazývá Berksonův. U něho naopak skutečný rozptyl je větší, než je rozptyl odhadu, a to právě o rozptýlení odpovídající chybě měření. Symbolicky popsáno

$$\boldsymbol{X_i} = \boldsymbol{W_i} + \boldsymbol{\psi_i},$$

avšak tady dochází k jinému podmínění chyby v měření: $\mathsf{E}[\boldsymbol{\psi_i} \mid \boldsymbol{Z_i}, \boldsymbol{W_i}] = 0$, zatímco u klasického modelu je též podmíněno $\boldsymbol{Z_i}$, ale místo $\boldsymbol{W_i}$ je podmíněno $\boldsymbol{X_i}$. Dochází tak ke změně konceptu chybového modelu a to obnáší dost změn. Proto na tomto místě upozorníme, že v dalších kapitolách pracujeme s klasickým pojetím chyby při měření (i když o něco komplikovanějším, jak bylo už zmíněno).

Klasický chybový model je intuitivnější a i přirozenější. Naopak pro Berksonův model se dokonce hůře nacházejí různé intuitivní příklady z reálné praxe, neboť většina situací spadá spíše pod klasické pojetí chybového modelu. Jeden z příkladů Berksonova chybového modelu je zaokrouhlování. Když hodnoty spojité veličiny zaokrouhlíme, body se nám shluknou. Naše odhady W_i mají menší variabilitu, neboť když k nim přidáme variabilitu odpovídající zaokrouhlení, dostáváme variabilitu skutečných dat. Neprobíhá to jen u klasického zaokrouhlování během měření, ale analogická situace nastává též, když využíváme kategorie u některého regresoru. Pro všechny jedince v rámci kategorie předpokládáme stejnou hodnotu, například průměrnou výši příjmu lidí v dané kategorii, avšak ve skutečnosti i mezi těmito lidmi je obrovský rozptyl v hodnotách a kategorický přístup tuhle variabilitu zcela ignoruje při snaze pojmout danou situaci jednodušeji a mnohdy i interpretačně snáze. Zaokrouhlování a kategorizování považujeme za nejběžnější příležitost se setkat se s Berksonovým modelem.

Berksonův model má speciální vlastnosti a různé výhody. Hlavní výhodou je nevychylování odhadů parametrů, což je v [1] zmíněno v částech 3.2.3 (strana 45) a 8.2.3 (strana 188). Avšak i pro tento model existují nějaké problémy, ale těm se v naší práci, která je zaměřena na klasický aditivní typ chyb, nebudeme dále věnovat.

Při počítání síly je potřeba pečlivě rozvážit, jestli se jedná o klasický chybový model, který snižuje sílu, nebo o Berksonův chybový model, který tím, že všechnu variabilitu přisuzuje variabilitě pro $\boldsymbol{X_i}$, obecně zvyšuje sílu pro detekci efektu regresoru, jak [1] zmiňuje v sekci 1.4.1 s názvem The Difference Between Berkson and Classical Errors: How to Gain More Power Without Trying. Zájemce o další detaily odkážeme na appendix B.1 tohoto zdroje a na podkapitolu 1.8., která je věnována ztrátě síly.

1.5 Metody pro analýzu dat s chybami v regresorech

Metod, jak se vypořádat s chybami v měření regresorů pro různé konkrétní situace, je vícero. Nezáleží jen na samotném hlavním modelu, ale situaci může

ovlivnit i chybový model, neboť chyby nemusí být nestranné, ale mohou výrazně měnit odhad a i záviset na jiných regresorech.

Metoda, kterou budeme zkoumat my, se nazývá **metoda upraveného skóre**. Vychází z toho, že vychýlení odhadu metodou nejmenších čtverců se napraví pomocí upravení skóre, které, jak jsme si ukázali v podkapitole 1.1, velice úzce souvisí se zmíněným odhadem. Zmíněné upravení skóre probíhá tak, aby podmíněná střední hodnota upraveného skóre byla nulová (respektive vektor nul). Její aplikaci na situaci se základním klasickým chybovým modelem si ukážeme hned v další kapitole.

V pozdějších kapitolách budeme uvažovat složitější chybový model, ve kterém bude nutné pracovat s dalšími parametry. Co se týká těchto parametrů chybového modelu, je několik možností. Nejjednodušší situací je, když jsou to obecně známé konstanty, například vyplývají z teoretických poznatků nebo Případně lze využít externích odhadů z jiné studie, pokud je rozumné předpokládat aplikovatelnost těchto odhadů na naší situaci (respektive design studie je shodný s našimi předpoklady). Pokud ani jedna z těchto dvou zmíněných situací nenastala, lze odhady parametrů získat pomocí validační skupiny.

Ve validační skupině totiž máme mimo odhadů $\boldsymbol{W_i}$ napozorované navíc přesné hodnoty $\boldsymbol{X_i}$. Můžeme tak odhadnout parametry v chybovém modelu a na základě toho získat rozumnější či dokonce i nestranné odhady $\hat{\boldsymbol{X}}_{\boldsymbol{i}}$ i z na první pohled nepoužitelných $\boldsymbol{W_i}$ v nevalidační skupině, tedy v té, kde $\boldsymbol{X_i}$ nepozorujeme (například z toho důvodu, že naměření přesných regresorů je velice složité a finančně nákladné, proto je možné to provést jen na malé části pozorování – očividně kdyby to šlo provést na všech, vůbec se nemusíme zabývat situací s chybou v měření regresorů a nedávalo by smysl psát tuto práci). Pomocí $\hat{\boldsymbol{X}}_{\boldsymbol{i}}$ z nevalidační skupiny a $\boldsymbol{X_i}$ z validační skupiny získáme odhady pro parametry hlavního modelu (budeme se tomu věnovat ve 4. kapitole).

Jiných metod je samozřejmě více. Jedna z nejběžnějších metod je regresní kalibrace. V kapitole 4 v [1] se lze dočíst o tom, jak odhadnout regresi nenapozorovaného $\boldsymbol{X}$ na $(\boldsymbol{Z}, \boldsymbol{W})$ označené jako $m_{\boldsymbol{X}(\boldsymbol{Z},\boldsymbol{W},\gamma)}$, neboť to závisí na parametru $\boldsymbol{\gamma}$, který je potřeba odhadnout. Pro regresi odezvy na $\boldsymbol{Z}, \boldsymbol{X}$ se nenapozorované $\boldsymbol{X}$ nahradí $m_{\boldsymbol{X}(\boldsymbol{Z},\boldsymbol{W},\hat{\gamma})}$. V tomto zdroji se o kapitolu později věnují jiné metodě s názvem SIMEX, která se zabývá extrapolací pomocí simulací.

Obecný přístup na modifikaci odhadovacích rovnic najdeme v [2], kde je věnována pozornost speciálním případům jako je kvadratická regrese, Poissonova či logistická. Regresi aditivního rizika v modelech s chybně naměřeným regresorem se věnují například v [3], proporcionálnímu riziku [4] a proporcionálnímu riziku s misklasifikovanou diskrétní vysvětlující proměnnou [5]. Alternativní přístupy pro cenzorovaná data najdeme v [6] pomocí kvantilové regrese, která je pro necenzorovanou situaci řešena v [7].

Vliv chybně naměřeného regresoru na predikci rizika se zkoumá v [8], kde lze najít i vícero grafických výstupů a tabulek pro porovnávání různých situací. Jedna z dříve uvedených „pohrom“ způsobených chybným naměřením vysvětlujících proměnných, jmenovitě ztráta síly v testech, je pro zobecněné regresní modely zkoumána společně s velikostí vzorku v [9] pomocí odvozené silofunkce založené na kvazi-věrohodnosti.

Metodu upraveného skóre zkoumají v [10], kde ale autoři zkoumají konvergenci rozvoje do řady a pak se věnují speciálním případům: a) vzácným událostem v logistické regresi, b) binární regresi pro extrémní hodnoty. My se budeme ubírat jiným směrem. Pro zobecněný chybový model, ve kterém odhad bude zašuměním lineární kombinace všech regresorů, budeme řešit odhadování všech parametrů z hlavního i chybového modelu a zkoumat sdružené asymptotické rozdělení všech těchto parametrů, ve 4. kapitole pro odvozenou asymptotiku provedeme simulační studie.

2 Metoda upraveného skóre pro klasický model

V této kapitole je převzat princip metody upraveného skóre. Odvození odhadů na model v této kapitole a další aplikace a zkoumání této metody v práci je vlastním přínosem autora.

2.1 Motivace

V této kapitole zavedeme metodu upraveného skóre pro lineární model s chybami v regresorech, kde chyba bude mít nulovou střední hodnotu vůči skutečným regresorům. Jedná se o nejjednodušší typ klasického chybového modelu, ale poslouží k ilustraci využití této metody.

Jak bylo zmíněno v předchozí kapitole, i pro velice jednoduchý model obsahující jen dva regresory, a to přesně naměřené Z_i a nenapozorované X_i, docházelo k vychýlení odhadu metodou nejmenších čtverců. Připomeňme, že tento odhad zároveň splňuje

$$\boldsymbol{U}(\hat{\boldsymbol{\beta}}) = \mathbf{0}_{\boldsymbol{p,1}},$$

kde $\boldsymbol{U}(\boldsymbol{\beta})$ je skórová statistika definovaná pomocí skóre pro jednotlivá pozorování:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{U_i}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X_i} \left(Y_i - \boldsymbol{X_i}^T \boldsymbol{\beta} \right),$$

neboť skórovou statistiku lze přepsat do maticové podoby

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X_i} \left(Y_i - \boldsymbol{X_i}^T \boldsymbol{\beta} \right) = \boldsymbol{X}^T \boldsymbol{Y} - \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta},$$

ze které, pokud ji položíme rovnu nulovému vektoru, lze dostat soustavu normálních rovnic, kterou řeší právě odhad metodou nejmenších čtverců. Proto se metoda zaměřuje na samotné skóre a to takovou úpravou, aby skóre ve střední hodnotě nebylo vychýlené. Takto upravené skóre sumarizujeme do upravené skórové statistiky. Tu následně upravíme do podoby, která bude připomínat explicitní vyjádření

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y},$$

avšak bude obsahovat korekci na dvou „místech“, které budou analogií přesně k těm místům, ve kterých jsme v podkapitole 1.2. pozorovali vychýlení. Tedy metoda upraveného skóre vyřeší vychýlení v odhadu metodou nejmenších čtverců způsobenou chybně naměřenými regresory $\boldsymbol{X_i}$.

2.2 Odvození odhadů

Mějme model

$$Y_i = \beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \boldsymbol{\beta_2}^T \boldsymbol{X_i} + \epsilon_i,$$

kde $\mathsf{E}[\epsilon_i \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] = 0$ a $i = 1, \ldots, n, n \in \mathbb{N}$. Nechť vektor vysvětlujících veličin $\boldsymbol{Z_i}$, které přesně napozorujeme, má q složek, a vektor $\boldsymbol{X_i}$ má p složek.

My ovšem $\boldsymbol{X_i}$ nepozorujeme přímo, ale máme o něm pouze informaci prostřednictvím $\boldsymbol{W_i}$, u kterého předpokládáme následující takzvaný chybový model:

$$W_i = \boldsymbol{X_i} + \boldsymbol{\psi_i},$$

kde $\mathsf{E}[\boldsymbol{\psi_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] = \boldsymbol{0_{p,1}}$, $\mathrm{var}(\boldsymbol{\psi_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}) = \boldsymbol{V(Z_i, X_i)}$ a $i = 1, \ldots, n$.

Chceme vypočítat střední hodnotu neupraveného skóre, označme ho

$$\tilde{\boldsymbol{U}}_i = \begin{pmatrix} 1 \\ \boldsymbol{Z_i} \\ \boldsymbol{W_i} \end{pmatrix} \left(Y_i - \left(\beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \boldsymbol{\beta_2}^T \boldsymbol{W_i} \right) \right).$$

Ze základních vlastností pro střední hodnotu dostáváme

$$\begin{aligned} \mathsf{E}[\tilde{\boldsymbol{U}}_i] &= \mathsf{E}\left[\mathsf{E}[\tilde{\boldsymbol{U}}_i \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] = \\ &= \begin{pmatrix} \mathsf{E}\left[\mathsf{E}[Y_i - \left(\beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \boldsymbol{\beta_2}^T \boldsymbol{W_i} \right) \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] \\ \mathsf{E}\left[\mathsf{E}[\boldsymbol{Z_i} Y_i - \boldsymbol{Z_i} \left(\beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} \right) - \boldsymbol{Z_i} \boldsymbol{\beta_2}^T \boldsymbol{W_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] \\ \mathsf{E}\left[\mathsf{E}[\boldsymbol{W_i} Y_i - \boldsymbol{W_i} \left(\beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} \right) - \boldsymbol{W_i} \boldsymbol{\beta_2}^T \boldsymbol{W_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] \end{pmatrix} = \begin{pmatrix} A_1 \\ \boldsymbol{A_2} \\ \boldsymbol{A_3} \end{pmatrix}. \end{aligned}$$

Pro lepší přehlednost proveďme jednotlivé výpočty odděleně:

$$\begin{aligned} A_1 &= \mathsf{E}\left[\beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \boldsymbol{\beta_2}^T \boldsymbol{X_i} + \mathsf{E}[\epsilon_i \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] - \beta_0 - \boldsymbol{\beta_1}^T \boldsymbol{Z_i} - \boldsymbol{\beta_2}^T \mathsf{E}[\boldsymbol{W_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] = \\ &= \mathsf{E}\left[\boldsymbol{\beta_2}^T \boldsymbol{X_i} - \boldsymbol{\beta_2}^T \mathsf{E}[\boldsymbol{W_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] = \mathsf{E}\left[\boldsymbol{\beta_2}^T \boldsymbol{X_i} - \boldsymbol{\beta_2}^T \boldsymbol{X_i} \right] = 0, \end{aligned}$$

$$\begin{aligned} \boldsymbol{A_2} &= \mathsf{E}\left[\boldsymbol{Z_i} \mathsf{E}[Y_i \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] - \boldsymbol{Z_i} \beta_0 - \boldsymbol{Z_i} \boldsymbol{\beta_1}^T \boldsymbol{Z_i} - \boldsymbol{Z_i} \boldsymbol{\beta_2}^T \mathsf{E}[\boldsymbol{W_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] = \\ &= \mathsf{E}\left[\boldsymbol{Z_i} \boldsymbol{\beta_2}^T \boldsymbol{X_i} + \boldsymbol{Z_i} \mathsf{E}[\epsilon_i \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] - \boldsymbol{Z_i} \boldsymbol{\beta_2}^T \boldsymbol{X_i} \right] = \boldsymbol{0_{q,1}}, \end{aligned}$$

$$\begin{aligned} \boldsymbol{A_3} &= \mathsf{E}\left[\mathsf{E}[\boldsymbol{W_i} (\boldsymbol{\beta_2}^T \boldsymbol{X_i} + \epsilon_i) - \boldsymbol{W_i} \boldsymbol{W_i}^T \boldsymbol{\beta_2} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] = \\ &= \mathsf{E}\left[\mathsf{E}[(\boldsymbol{X_i} + \boldsymbol{\psi_i})(\boldsymbol{\beta_2}^T \boldsymbol{X_i} + \epsilon_i) - (\boldsymbol{X_i} + \boldsymbol{\psi_i})(\boldsymbol{X_i} + \boldsymbol{\psi_i})^T \boldsymbol{\beta_2} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] = \\ &= \mathsf{E}\left[(\boldsymbol{X_i} + \mathsf{E}[\boldsymbol{\psi_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}]) \boldsymbol{\beta_2}^T \boldsymbol{X_i} + \mathsf{E}[\boldsymbol{\psi_i} \epsilon_i - (\boldsymbol{X_i} + \boldsymbol{\psi_i})(\boldsymbol{X_i} + \boldsymbol{\psi_i})^T \boldsymbol{\beta_2} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] = \\ &= \mathsf{E}\left[\boldsymbol{X_i} \boldsymbol{\beta_2}^T \boldsymbol{X_i} + \mathsf{E}[\boldsymbol{\psi_i} \epsilon_i - (\boldsymbol{X_i} \boldsymbol{X_i}^T + \boldsymbol{X_i} \boldsymbol{\psi_i}^T + \boldsymbol{\psi_i} \boldsymbol{X_i}^T + \boldsymbol{\psi_i} \boldsymbol{\psi_i}^T) \boldsymbol{\beta_2} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] = \\ &= \mathsf{E}\left[\boldsymbol{X_i} \boldsymbol{X_i}^T \boldsymbol{\beta_2} - \boldsymbol{X_i} \boldsymbol{X_i}^T \boldsymbol{\beta_2} + \mathsf{E}[\boldsymbol{\psi_i} \epsilon_i - (\boldsymbol{X_i} \boldsymbol{\psi_i}^T + \boldsymbol{\psi_i} \boldsymbol{X_i}^T + \boldsymbol{\psi_i} \boldsymbol{\psi_i}^T) \boldsymbol{\beta_2} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \right] = \\ &= \mathsf{E}\left[\mathsf{E}[\boldsymbol{\psi_i} \epsilon_i - \boldsymbol{\psi_i} \boldsymbol{\psi_i}^T \boldsymbol{\beta_2} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] - \boldsymbol{X_i} \boldsymbol{\beta_2}^T \mathsf{E}[\boldsymbol{\psi_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] - \mathsf{E}[\boldsymbol{\psi_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] \boldsymbol{X_i}^T \boldsymbol{\beta_2} \right] = \\ &= \mathsf{E}\left[\boldsymbol{\psi_i} \epsilon_i \right] - \mathsf{E}\left[\boldsymbol{V(Z_i, X_i)} \right] \boldsymbol{\beta_2} = \mathrm{cov}(\boldsymbol{\psi_i}; \epsilon_i) - \mathsf{E}\left[\boldsymbol{V(Z_i, X_i)} \right] \boldsymbol{\beta_2}. \end{aligned}$$

Vidíme, že $\boldsymbol{A_3}$ obecně není nulové a tedy celkově $\tilde{\boldsymbol{U}}_{\boldsymbol{i}}$ je vychýlené, což není žádané. Proto zavádíme upravené skóre, které bude mít nulovou střední hodnotu:

$$\boldsymbol{U_i} = \begin{pmatrix} 1 \\ \boldsymbol{Z_i} \\ \boldsymbol{W_i} \end{pmatrix} \left(Y_i - \left(\beta_0 + \boldsymbol{\beta_1^T Z_i} + \boldsymbol{\beta_2^T W_i} \right) \right) - \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \operatorname{cov}(\boldsymbol{\psi_i}; \epsilon_i) - \boldsymbol{V(Z_i, X_i)\beta_2} \end{pmatrix},$$

kde $i = 1, \ldots, n$.

Označme data pomocí $\boldsymbol{M} := (\mathbf{1_n} \mid \boldsymbol{Z} \mid \boldsymbol{W})$, kde $\mathbf{1_n}$ je sloupcový vektor obsahující n jedniček, $\boldsymbol{Z} := \left(\boldsymbol{Z_1^T}, \ \ldots, \ \boldsymbol{Z_n^T} \right)^T$ a $\boldsymbol{W} := \left(\boldsymbol{W_1^T}, \ \ldots, \ \boldsymbol{W_n^T} \right)^T$. Dále označme $\boldsymbol{Y} := \left(Y_1^T, \ \ldots, \ Y_n^T \right)^T$.

Pak by upravená skórová statistika, tedy součet upravených skór

$$\sum_{k=1}^{n} \boldsymbol{U_i},$$

měla mít pro skutečné parametry střední hodnotu rovnu nule. Tuto skórovou statistiku můžeme použít na odhadnutí našich parametrů. Pomocí vyjádření $\boldsymbol{U_i}$ můžeme maticově zapsat součet upravených skór jako

$$\boldsymbol{M^T} \left(\boldsymbol{Y} - \boldsymbol{M} \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix} \right) - \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \sum_{k=1}^{n} \left[\operatorname{cov}(\boldsymbol{\psi_i}; \epsilon_i) - \boldsymbol{V(Z_i, X_i)\beta_2} \right] \end{pmatrix}.$$

Pokud to položíme rovno vektoru nul, získáme tak soustavu rovnic pro odhadování bet, které si explicitně vyjádříme. Postupně upravujme:

$$\boldsymbol{M^T Y} - \boldsymbol{M^T M} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\boldsymbol{\beta}}_{\mathbf{2}} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \sum_{k=1}^{n} \operatorname{cov}(\boldsymbol{\psi_i}; \epsilon_i) \end{pmatrix} - \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \sum_{k=1}^{n} \boldsymbol{V(Z_i, X_i)} \hat{\boldsymbol{\beta}}_{\mathbf{2}} \end{pmatrix},$$

odhady chceme mít na levé straně, tedy

$$\boldsymbol{M^T M} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\boldsymbol{\beta}}_{\mathbf{2}} \end{pmatrix} - \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \sum_{k=1}^{n} \boldsymbol{V(Z_i, X_i)} \hat{\boldsymbol{\beta}}_{\mathbf{2}} \end{pmatrix} = \boldsymbol{M^T Y} - \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \sum_{k=1}^{n} \operatorname{cov}(\boldsymbol{\psi_i}; \epsilon_i) \end{pmatrix}.$$

Protože lze provést přepsání

$$\begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \sum_{k=1}^{n} \boldsymbol{V(Z_i, X_i)} \hat{\boldsymbol{\beta}}_{\mathbf{2}} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{0_{1,q}} & \mathbf{0_{1,p}} \\ \mathbf{0_{q,1}} & \mathbf{0_{q,q}} & \mathbf{0_{q,p}} \\ \mathbf{0_{p,1}} & \mathbf{0_{p,q}} & \sum_{k=1}^{n} \boldsymbol{V(Z_i, X_i)} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\boldsymbol{\beta}}_{\mathbf{2}} \end{pmatrix},$$

lze levou stranu soustavy rovnic zapsat ve tvaru

$$\begin{pmatrix} n & \mathbf{1}_n^T\boldsymbol{Z} & \mathbf{1}_n^T\boldsymbol{W} \\ \boldsymbol{Z}^T\mathbf{1}_n & \boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\boldsymbol{W} \\ \boldsymbol{W}^T\mathbf{1}_n & \boldsymbol{W}^T\boldsymbol{Z} & \left[\boldsymbol{W}^T\boldsymbol{W} - \sum_{k=1}^n \boldsymbol{V}(\boldsymbol{Z}_i,\boldsymbol{X}_i)\right] \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_\mathbf{1} \\ \hat{\boldsymbol{\beta}}_\mathbf{2} \end{pmatrix}.$$

Předpokládejme regularitu matice

$$\begin{pmatrix} n & \mathbf{1}_n^T\boldsymbol{Z} & \mathbf{1}_n^T\boldsymbol{W} \\ \boldsymbol{Z}^T\mathbf{1}_n & \boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\boldsymbol{W} \\ \boldsymbol{W}^T\mathbf{1}_n & \boldsymbol{W}^T\boldsymbol{Z} & \left[\boldsymbol{W}^T\boldsymbol{W} - \sum_{k=1}^n \boldsymbol{V}(\boldsymbol{Z}_i,\boldsymbol{X}_i)\right] \end{pmatrix}.$$

Pak existuje její inverze a můžeme tak vyjádřit odhad regresních parametrů:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_\mathbf{1} \\ \hat{\boldsymbol{\beta}}_\mathbf{2} \end{pmatrix} = \begin{pmatrix} n & \mathbf{1}_n^T\boldsymbol{Z} & \mathbf{1}_n^T\boldsymbol{W} \\ \boldsymbol{Z}^T\mathbf{1}_n & \boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\boldsymbol{W} \\ \boldsymbol{W}^T\mathbf{1}_n & \boldsymbol{W}^T\boldsymbol{Z} & \left[\boldsymbol{W}^T\boldsymbol{W} - \sum_{k=1}^n \boldsymbol{V}(\boldsymbol{Z}_i,\boldsymbol{X}_i)\right] \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}_n^T\boldsymbol{Y} \\ \boldsymbol{Z}^T\boldsymbol{Y} \\ \boldsymbol{W}^T\boldsymbol{Y} - \sum_{k=1}^n \operatorname{cov}(\boldsymbol{\psi}_i;\epsilon_i) \end{pmatrix}.$$

Sice se odhad liší od odhadu metodou nejmenších čtverců „jen na 2 místech“, ale je třeba si povšimnout, že jedna z těch úprav je uvnitř matice, která se invertuje. To znamená že se tato úprava projeví na všech pozicích inverzní matice a pomocí ní se ovlivní odhady všech parametrů.

2.3 Rozšíření využití metody upraveného skóre

V další kapitole si ukážeme, jak aplikovat metodu upraveného skóre v situaci, kdy chybový model je složitější, a chyba v měření regresoru obecně není nestranná. Pro jednoduchost se zaměříme jen na situaci, kdy bude právě jeden z regresorů naměřen s chybou, tedy X_i a W_i budou (jednorozměrné) náhodné veličiny. Konkrétně se zaměříme na chybový model, ve kterém W_i bude zašuměná lineární kombinace všech regresorů, tedy $\boldsymbol{Z_i}$ a X_i. Tento model je o něco obecnější a v praxi uplatnitelnější než chybový model s nestrannými chybami v měření. Ve 3. kapitole si odvodíme, že i pro tento složitější chybový model stále budou vycházet analogické korekce, ke kterým jsme došli v této kapitole. Avšak místo W_i budeme u skóre, potažmo i v opraveném odhadu bet, uvažovat jeho transformaci do $\hat{X}_i$, nestranného odhadu X_i, který bude odvozen z chybového modelu.

Ve 3. kapitole budeme parametry z chybového modelu předpokládat za známé. Může to být z několika důvodů. Jedním z nich je vyplynutí známosti hodnot těchto parametrů z teoretických poznatků, například při studiu konkrétního fyzikálního jevu. Jiným případem by bylo, kdyby se jednalo o odhady získané jinou studií, která by měla takový design, že by byly tyto výsledky „přenositelné“, avšak by bylo nutné zohlednit nejistotu těchto výsledků a přizpůsobit tomu metodu upraveného skóre. Ovšem v praxi tyto parametry nebudou známé, případně sice některá studie na jejich odhadnutí bude existovat, ale my nebudeme mít dostatečnou důvěru k jejich použití (může to být z různých důvodů či kombinace

těchto důvodů – může být pochybnost o tom, jak byla data jiné studie získána a zda jsou dostatečně reprezentativní, zda manipulace s daty byla provedena korektně, bylo využito vhodné metody na daný problém či i kdyby vše proběhlo v pořádku, tak zda jsou tyto výsledky přenositelné i na náš případ – můžeme mít odlišnost ve sběru dat či dokonce si nemusíme býti jistí, zda bude podchycen regionální vliv, tedy nebude možné považovat danou studii za reprezentativní vůči našemu regionu). Proto z praktických důvodů může být výhodnější a bezpečnější předpokládat parametry chybového modelu za neznámé. Pak ale bude potřeba se s tím nějak jinak vypořádat. Ve 4. kapitole nabídneme řešení tohoto problému pomocí validační skupiny. V této skupině budeme mít jak nepřesně naměřené W_i, tak budeme u ní mít i přesně naměřené X_i. Pomocí této validační skupiny odhadneme parametry chybového modelu, které pak budeme moci použít pro získání nestranných odhadů $\hat{X}_i$ v nevalidační skupině. Obě tyto skupiny použijeme na odhadnutí $\boldsymbol{\beta}$. Tyto všechny parametry dáme do soustavy odhadovacích rovnic. Bude tak možné odhadovat všechny parametry současně a nebude hrozit, že parametry chybového modelu budou odhadovány na skupině s jinými vlastnostmi než má skupina, na které odhadujeme parametry hlavního modelu.

Pro jednoduchost budeme předpokládat, že chyba ϵ_i v hlavním modelu je nekolerovaná s chybou ψ_i z chybového modelu. Aby bylo možné zkoumat sdruženou asymptotiku parametrů, budeme potřebovat přidat další předpoklady na chybový model a to přímo na ψ_i. Hlavním z nich bude určení parametrického modelu pro ropztyl chyb, neboť v průběhu odvozování bude nutné nejen počítání podmíněných středních hodnot, ale i derivování. Protože se budeme o rozptyl chyb zajímat a budeme chtít ho odhadovat současně společně s dalšími parametry v chybovém i hlavním modelu, budeme využívat obvyklý a přirozený předpoklad na rozptyl

$$\mathrm{var}(\psi_i \mid \boldsymbol{Z_i}, X_i) = V(\boldsymbol{Z_i}, X_i) = \sigma^2_\psi.$$

Po odvození sdružené asymptotiky naších odhadů budeme v 5. kapitole zkoumat v simulační studii, nakolik může být užitečná tato asymptotika v praxi. Budeme zkoumat, zda stále dochází k vychýlení odhadů vzniklé chybně naměřeným regresorem, či zda se naopak povedlo toto vychýlení metodou upraveného skóre úspěšně odstranit. Podíváme se i na variabilitu těchto odhadů mezi jednotlivými simulacemi a porovnáme je s teoretickou variabilitou, kterou by odhady měly mít podle odvozené asymptotiky. Pro jednotlivé parametry určíme intervaly spolehlivosti a budeme zkoumat nakolik dobře je dodržená hladina vůči dvěma praktickým aspektům. Tím prvním je samozřejmě rozsah výběru, tedy budeme zkoumat zda i pro malé n vychází všechny odhady rozumně, případně jestli aspoň vychází rozumně pro bety, které jsou pro nás primárním zájmem. V té druhé situaci budeme mít sice rozsah větší, ale budeme zkoumat jak rychle se odhady budou zhoršovat vůči zmenšující se velikosti validační skupiny vůči té nevalidační. Budeme testovat také sílu testu pro nulovou hypotézu, zda parametr u X_i v hlavním modelu je nulový či zda bude nesprávně přisuzovat vliv X_i na odezvu jako statisticky významný. Nakonec se ve speciální simulaci, ve které nastavíme parametry tak, aby chyby v měření byly nestranné, zaměříme na porovnávání odhadů $\boldsymbol{\beta}$ metodou nejmenších čtverců a námi odvozené asymptotiky využívající metodu upraveného skóre.

3 Metoda upraveného skóre pro zobecněný model

V této kapitole nepřebíráme žádné výsledky z jiného zdroje, pouze zobecňujeme využití metody upraveného skóre a vše uvedené je vlastní odvození autora.

3.1 Motivace zobecnění

Předpoklad nejzákladnějšího typu klasického chybového modelu

$$\boldsymbol{W_i} = \boldsymbol{X_i} + \boldsymbol{\psi_i},$$

kde $\mathsf{E}[\boldsymbol{\psi_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}] = \mathbf{0_{p,1}}$, $\mathrm{var}(\boldsymbol{\psi_i} \mid \boldsymbol{Z_i}, \boldsymbol{X_i}) = \boldsymbol{V(Z_i, X_i)}$ a $i = 1, \ldots, n$, může být poměrně dosti svazující a v praxi ne vždy splnitelný. Pro jednoduchost předpokládejme, že máme pouze jeden regresor X_i, u kterého neumíme „naměřit" přesnou hodnotu. Je poměrně běžné, že chyba zašumění ψ_i bude záviset na samotném X_i a nebude nestranná. Může to být například situace, kdy se ptáme respondentů na výši jejich platů. Zatímco většina lidí, která pobírá nízký příjem, nemá problém poukázat na obtížnou životní situaci spojenou s nízkým platem, uvede často bez většího přikrášlení svůj příjem, naopak bohatí lidé často svůj příjem výrazněji podhodnocují.

Situace se pak může ještě více zkomplikovat. Chyba v měření X_i může záviset i na jiném regresoru a nebo rovnou na vícero z nich či dokonce na všech! Pro jednoduchou ilustraci uveďme situaci, kdy se ptáme respondentů na jejich průměrný denní příjem, s tím, že máme navíc informace o jejich indexu tělesné hmotnosti a kategoriální veličině reprezentující míru tělesné aktivity či sportování. Zatímco lidé s nízkým indexem tělesné hmotnosti, kteří pravidelně intenzivně sportují, budou mít velkou tendenci si objem a složení své stravy více hlídat a budou umět lépe odhadovat svůj průměrný denní příjem kalorií, naopak lidé s výrazně vysokým indexem tělesné hmotnosti a prakticky žádným pohybem budou častěji mít tendenci nepřiznávat (ať doktorovi nebo dokonce i sobě) svůj reálný příjem kalorií, ale hluboce ho podstřelí. Obecně pak samotné odhady W_i v některých případech nemusí mít svou hodnotou vůbec blízko ke skutečnému X_i, nebo obecně v extrémních modelech dokonce mohou hodnoty nabývat různých řádů nebo i znamének. Je jasné, že samotné W_i je nepoužitelné, ale šlo by zrekonstruovat hodnotu X_i (získat tedy odhad, který by dával „smysl"). Toto vše vede k motivaci zobecnit využití metody upraveného skóre i na případy se složitějším chybovým modelem, v našem případě zkoumání veličiny W_i, která vznikla jako zašumění lineární kombinace některých nebo dokonce všech regresorů.

3.2 Odvození odhadů

Mějme model

$$Y_i = \beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \beta_2 X_i + \epsilon_i,$$

kde $\mathsf{E}[\epsilon_i \mid \boldsymbol{Z_i}, X_i] = 0$ a $i = 1, \ldots, n, n \in \mathbb{N}$. Nechť vektor vysvětlujících veličin $\boldsymbol{Z_i}$, které přesně napozorujeme, má q složek, X_i je náhodná veličina, kterou ale nepozorujeme přímo. Místo této náhodné veličiny máme pouze informaci o W_i. Tentokrát pro W_i místo základního klasického chybového modelu předpokládáme zobecněný model

$$W_i = \alpha_0 + \boldsymbol{\alpha_1}^T \boldsymbol{Z_i} + \alpha_2 X_i + \psi_i,$$

kde $\mathsf{E}[\psi_i \mid \boldsymbol{Z_i}, X_i] = 0$, $\mathrm{var}(\psi_i \mid \boldsymbol{Z_i}, X_i) = V(\boldsymbol{Z_i}, X_i)$ a $i = 1, \ldots, n$.

Je třeba upozornit, že oproti minulé kapitole W_i není obecně nestranným odhadem X_i. Proto neupravené skóre se bude lišit a bude potřeba ho napočítat „znovu“. V místech předchozího výskytu W_i bude odhad

$$\hat{X}_i = \frac{W_i - \alpha_0 - \boldsymbol{\alpha_1}^T \boldsymbol{Z_i}}{\alpha_2},$$

neboť v této kapitole budeme předpokládat $\alpha_0, \boldsymbol{\alpha_1}$ a α_2 známé, tedy neupravené skóre má nyní podobu

$$\tilde{\boldsymbol{U}}_{\boldsymbol{i}} = \begin{pmatrix} 1 \\ \boldsymbol{Z_i} \\ \hat{X}_i \end{pmatrix} \left(Y_i - \left(\beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \beta_2 \hat{X}_i \right) \right).$$

Vztah mezi X_i a $\hat{X}_i$ budeme opakovaně využívat, proto odvození tohoto vztahu provedeme na tomto místě velmi podrobně, později ho už budeme využívat mlčky:

$$\hat{X}_i = \frac{W_i - \alpha_0 - \boldsymbol{\alpha_1}^T \boldsymbol{Z_i}}{\alpha_2} = \frac{\cancel{\alpha_0} + \cancel{\boldsymbol{\alpha_1}^T \boldsymbol{Z_i}} + \alpha_2 X_i + \psi_i - \cancel{\alpha_0} - \cancel{\boldsymbol{\alpha_1}^T \boldsymbol{Z_i}}}{\alpha_2} = X_i + \frac{\psi_i}{\alpha_2}.$$

Z tohoto vztahu ihned vidíme, že $\hat{X}_i$ je nestranný odhad X_i, neboť podmíněná střední hodnota ψ_i je nulová. Obdobně jako v předchozí kapitole si výpočet upraveného skóre rozdělíme na 3 části:

$$\begin{aligned} \mathsf{E}[\tilde{\boldsymbol{U}}_{\boldsymbol{i}}] &= \mathsf{E}\left[\mathsf{E}[\tilde{\boldsymbol{U}}_{\boldsymbol{i}} \mid \boldsymbol{Z_i}, X_i] \right] = \\ &= \begin{pmatrix} \mathsf{E}\left[\mathsf{E}[Y_i - \left(\beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \beta_2 \hat{X}_i \right) \mid \boldsymbol{Z_i}, X_i] \right] \\ \mathsf{E}\left[\mathsf{E}[\boldsymbol{Z_i} Y_i - \boldsymbol{Z_i} \left(\beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} \right) - \boldsymbol{Z_i} \beta_2 \hat{X}_i \mid \boldsymbol{Z_i}, X_i] \right] \\ \mathsf{E}\left[\mathsf{E}[\hat{X}_i Y_i - \hat{X}_i \left(\beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} \right) - \hat{X}_i \beta_2 \hat{X}_i \mid \boldsymbol{Z_i}, X_i] \right] \end{pmatrix} = \begin{pmatrix} A_1 \\ \boldsymbol{A_2} \\ A_3 \end{pmatrix}. \end{aligned}$$

Díky nestrannosti $\hat{X}_i$ a vlastnostem podmíněné střední hodnoty se výpočet A_1 a $\boldsymbol{A_2}$ provede analogicky jako v předchozí kapitole. Proveďme výpočet pro A_2, který se liší:

$$\begin{aligned}
A_3 &= \mathsf{E}\left[\mathsf{E}[\hat{X}_i(\beta_2 X_i + \epsilon_i) - \hat{X}_i\beta_2\hat{X}_i \mid \boldsymbol{Z_i}, X_i]\right] = \\
&= \mathsf{E}\left[\mathsf{E}[\hat{X}_i\left((\beta_2 X_i+\epsilon_i) - \beta_2(X_i+\frac{\psi_i}{\alpha_2})\right) \mid \boldsymbol{Z_i}, X_i]\right] = \mathsf{E}\left[\mathsf{E}[(X_i+\frac{\psi_i}{\alpha_2})(\epsilon_i - \beta_2\frac{\psi_i}{\alpha_2}) \mid \boldsymbol{Z_i}, X_i]\right] = \\
&= \mathsf{E}\left[X_i\,\mathsf{E}[\epsilon_i - \beta_2\frac{\psi_i}{\alpha_2} \mid \boldsymbol{Z_i}, X_i] + \frac{1}{\alpha_2}\,\mathsf{E}[\psi_i(\epsilon_i - \beta_2\frac{\psi_i}{\alpha_2}) \mid \boldsymbol{Z_i}, X_i]\right] = \\
&= \mathsf{E}\left[X_i(0-0) + \frac{1}{\alpha_2}\,\mathsf{E}[\psi_i\epsilon_i \mid \boldsymbol{Z_i}, X_i] - \frac{\beta_2}{\alpha_2^2}\,\mathsf{E}[\psi_i^2 \mid \boldsymbol{Z_i}, X_i]\right] = \\
&= \frac{1}{\alpha_2}\,\mathsf{E}\left[\psi_i\epsilon_i\right] - \frac{\beta_2}{\alpha_2^2}\,\mathsf{E}\left[\mathsf{E}[\psi_i^2 \mid \boldsymbol{Z_i}, X_i]\right] = \frac{1}{\alpha_2}\mathrm{cov}(\psi_i, \epsilon_i) - \frac{\beta_2}{\alpha_2^2}\,\mathsf{E}\left[V(\boldsymbol{Z_i}, X_i)\right].
\end{aligned}$$

I pro tento chybový model vychází podobné vychýlení jako v minulé kapitole, tedy samotné skóre je vychýlené jen v poslední složce a opět se tam objevuje kovariance ψ_i a ϵ_i. Tentokrát ale oboje je vydělené α_2, respektive její druhou mocninou, což je parametr u X_i v chybovém modelu. Upravené skóre tak bude mít podobu:

$$\boldsymbol{U_i} = \begin{pmatrix} 1 \\ \boldsymbol{Z_i} \\ \hat{X}_i \end{pmatrix}\left(Y_i - \left(\beta_0 + \boldsymbol{\beta_1}^T\boldsymbol{Z_i} + \beta_2\hat{X}_i\right)\right) - \begin{pmatrix} 0 \\ \boldsymbol{0_{q,1}} \\ \frac{1}{\alpha_2}\mathrm{cov}(\psi_i;\epsilon_i) - \frac{\beta_2}{\alpha_2^2}V(\boldsymbol{Z_i}, X_i) \end{pmatrix},$$

kde $i = 1, \ldots, n$. Označme data $\boldsymbol{M} := \left(\boldsymbol{1_n} \mid \boldsymbol{Z} \mid \hat{\boldsymbol{X}}\right)$, kde $\boldsymbol{1_n}$ je sloupcový vektor obsahující n jedniček, $\boldsymbol{Z} := \left(\boldsymbol{Z_1}^T, \ \ldots, \ \boldsymbol{Z_n}^T\right)^T$ a $\hat{\boldsymbol{X}} := \left(\hat{X}_1, \ \ldots, \ \hat{X}_n\right)^T$. Dále označme $\boldsymbol{Y} := \left(Y_1, \ \ldots, \ Y_n\right)^T$. Pak by upravená skórová statistika, tedy součet upravených skór, kterou můžeme pomocí předchozího zapsat jako

$$\boldsymbol{M}^T\left(\boldsymbol{Y} - \boldsymbol{M}\begin{pmatrix}\beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2\end{pmatrix}\right) - \begin{pmatrix} 0 \\ \boldsymbol{0_{q,1}} \\ \sum_{k=1}^n\left[\frac{1}{\alpha_2}\mathrm{cov}(\psi_i;\epsilon_i) - \frac{\beta_2}{\alpha_2^2}V(\boldsymbol{Z_i}, X_i)\right] \end{pmatrix},$$

měla mít nulovou podmíněnou střední hodnotu. Opět soustavu položíme rovnu vektoru nul a chceme explicitně vyjádřit $\hat{\boldsymbol{\beta}}$, které budou danou soustavu řešit. Postupně upravujme obdobně jako v minulé kapitole:

$$\boldsymbol{M}^T\boldsymbol{M}\begin{pmatrix}\hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{1}} \\ \hat{\beta}_2\end{pmatrix} - \begin{pmatrix} 0 \\ \boldsymbol{0_{q,1}} \\ \frac{\hat{\beta}_2}{\alpha_2^2}\sum_{k=1}^n V(\boldsymbol{Z_i}, X_i) \end{pmatrix} = \boldsymbol{M}^T\boldsymbol{Y} - \begin{pmatrix} 0 \\ \boldsymbol{0_{q,1}} \\ \frac{1}{\alpha_2}\sum_{k=1}^n \mathrm{cov}(\psi_i;\epsilon_i) \end{pmatrix},$$

kde lze levou stranu přepsat do tvaru

$$\begin{pmatrix} n & \mathbf{1}_{\boldsymbol{n}}^T\boldsymbol{Z} & \mathbf{1}_{\boldsymbol{n}}^T\hat{\boldsymbol{X}} \\ \boldsymbol{Z}^T\mathbf{1}_{\boldsymbol{n}} & \boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\hat{\boldsymbol{X}} \\ \hat{\boldsymbol{X}}^T\mathbf{1}_{\boldsymbol{n}} & \hat{\boldsymbol{X}}^T\boldsymbol{Z} & \left[\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}} - \frac{1}{\alpha_2^2}\sum_{k=1}^n V(\boldsymbol{Z_i},X_i)\right] \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\boldsymbol{\beta}}_{\mathbf{2}} \end{pmatrix}.$$

Předpokládejme regularitu matice

$$\begin{pmatrix} n & \mathbf{1}_{\boldsymbol{n}}^T\boldsymbol{Z} & \mathbf{1}_{\boldsymbol{n}}^T\hat{\boldsymbol{X}} \\ \boldsymbol{Z}^T\mathbf{1}_{\boldsymbol{n}} & \boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\hat{\boldsymbol{X}} \\ \hat{\boldsymbol{X}}^T\mathbf{1}_{\boldsymbol{n}} & \hat{\boldsymbol{X}}^T\boldsymbol{Z} & \left[\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}} - \frac{1}{\alpha_2^2}\sum_{k=1}^n V(\boldsymbol{Z_i},X_i)\right] \end{pmatrix}.$$

Pak existuje její inverze a tak můžeme vyjádřit odhad bet:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} n & \mathbf{1}_{\boldsymbol{n}}^T\boldsymbol{Z} & \mathbf{1}_{\boldsymbol{n}}^T\hat{\boldsymbol{X}} \\ \boldsymbol{Z}^T\mathbf{1}_{\boldsymbol{n}} & \boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\hat{\boldsymbol{X}} \\ \hat{\boldsymbol{X}}^T\mathbf{1}_{\boldsymbol{n}} & \hat{\boldsymbol{X}}^T\boldsymbol{Z} & \left[\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}} - \frac{1}{\alpha_2^2}\sum_{k=1}^n V(\boldsymbol{Z_i},X_i)\right] \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}_{\boldsymbol{n}}^T\boldsymbol{Y} \\ \boldsymbol{Z}^T\boldsymbol{Y} \\ \hat{\boldsymbol{X}}^T\boldsymbol{Y} - \frac{1}{\alpha_2}\sum_{k=1}^n \mathrm{cov}(\psi_i;\epsilon_i) \end{pmatrix}.$$

Připomeňme, že v této kapitole jsme přepodkládali $\alpha_0, \boldsymbol{\alpha_1}$ a α_2 známé. Jak bylo řečeno na konci druhé kapitoly, ve 4. kapitole se podíváme na to, jak řešit situaci, kdy nechceme předpokládat známost parametrů v chybovém modelu. Budeme je odhadovat pomocí validační skupiny a pak si odvodíme sdruženou asymptotiku všech zkoumaných parametrů.

4 Metoda upraveného skóre s validační skupinou

V této kapitole nepřebíráme žádné výsledky z jiného zdroje (až na asymptotické rozdělení Z-odhadů, což na daném místě zmíníme při aplikování na naší situaci), pouze zobecňujeme využití metody upraveného skóre a vše uvedené je vlastní odvození autora.

4.1 Validační skupina

Připomeňme, že náš hlavní zájem je odhadnout parametry v modelu

$$Y_i = \beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \beta_2 X_i + \epsilon_i,$$

kde $\mathsf{E}[\epsilon_i \mid \boldsymbol{Z_i}, X_i] = 0$, $\boldsymbol{Z_i}$ má q složek a $i = 1, \ldots, n, n \in \mathbb{N}$, který nazýváme hlavním modelem. Ovšem X_i je regresor, který běžným měřením nenaměříme přesně, ale máme ho odhadnut s nějakou chybou ve formě W_i. Z předchozích kapitol víme, že i měření s chybou mající nulovou podmíněnou střední hodnotu vychýlí odhad metodou nejmenších čtverců – a to nejen β_2, ale všechny parametry. Proto využíváme korekci metodou upraveného skóre, abychom mohli parametry odhadovat bez vychýlení.

Dosud jsme pracovali pouze s nevalidační skupinou, ve které jsme měli napozorované $\boldsymbol{Z_i}, W_i$ a Y_i, nikoliv samotnou přesnou hodnotu X_i, o které máme jen informaci prostřednictvím zmíněného W_i. Toto nás dosud nemuselo vůbec rozrušovat, protože jsme nejdříve pracovali s nejjednodušším chybovým modelem, který by pro jednorozměrné X_i byl podoby

$$W_i = X_i + \psi_i,$$

kde $\mathsf{E}[\psi_i \mid \boldsymbol{Z_i}, X_i] = 0$, $\mathrm{var}(\psi_i \mid \boldsymbol{Z_i}, X_i) = V(\boldsymbol{Z_i}, X_i)$ a $i = 1, \ldots, n$, ve kterém nebyly žádné další neznámé parametry (až na $V(\boldsymbol{Z_i}, X_i)$), se kterými bychom se museli nějak vypořádat, zatímco ve 3. kapitole jsme sice pracovali se zobecněným chybovým modelem

$$W_i = \alpha_0 + \boldsymbol{\alpha_1}^T \boldsymbol{Z_i} + \alpha_2 X_i + \psi_i,$$

kde $\mathsf{E}[\psi_i \mid \boldsymbol{Z_i}, X_i] = 0$, $\mathrm{var}(\psi_i \mid \boldsymbol{Z_i}, X_i) = V(\boldsymbol{Z_i}, X_i)$ a $i = 1, \ldots, n$, ale $\alpha_0, \boldsymbol{\alpha_1}$ a α_2 jsme považovali za známé konstanty a s tímto předpokladem jsme odvodili, jak by měla vypadat korekce odhadu metodou nejmenších čtverců, aby nedocházelo k vychýlení. Avšak v této kapitole opouštíme od předpokladu známosti parametrů v chybovém modelu a to z několika praktických ohledů, které jsme zmínili ke konci 2. kapitoly. Všechny parametry budou odhadovány ze stejného výběru, tedy by se nemělo stát, že by parametry chybového modelu byly odhadovány na skupině s jinými vlastnostmi, než má skupina, na které budou odhadovány parametry hlavního modelu. Toto je velkou motivací proč rozšiřovat zkoumání metody upraveného skóre na situaci s validační skupinou, která má oproti nevalidační

skupině informaci navíc i o X_i, a budeme striktně vyžadovat předpoklad existence náhodné veličiny ξ_i nabývající hodnoty 1 pro validační skupinu a 0 pro nevalidační skupinu, která bude nezávislá na všem ostatním, tedy hodnotu ξ_i nebudou ovlivňovat ani hodnoty ostatních parametrů, ani odezva, ani jednotlivé regresory.

I když budeme pracovat s validační skupinou, která má informaci o přesně naměřeném X_i, stále má smysl se zabývat i nevalidační skupinou a neodhadovat $\boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta_1}^T, \beta_2\right)^T$ jen na základě validační skupiny. Je to z praktického důvodu. Validační skupina může být malá a to jak ve významu relativním, kdy v nevalidační skupině máme několikanásobně více pozorování než ve validační skupině, tak i ve významu absolutním, kdy validační i nevalidační skupina mají dohromady počet pozorování jen v řádu desítek. Samozřejmě může nastat otázka, zda nelze u všech pozorování z nevalidační skupiny naměřit X_i, které šlo naměřit ve validační skupině! Tato otázka je zcela korektní, avšak bohužel svět není tak jednoduchý a lze najít zcela rozumné důvody, proč řešit situaci, ve které máme současně validační i nevalidační skupinu.

Z reálného života si lze ihned představit situaci, kdy zkoumáme vliv několika regresorů u pacientů na odezvu, která může být faktorem pro vznik některých nemocí, například hladina cholesterolu v krvi. Je zcela zřejmé, že zatímco je mnoho regresorů, které lze „naměřit“ relativně dostatečně přesně, například věk, biologické pohlaví nebo i index tělesné hmotnosti, je zároveň několik regresorů, které se měří obtížně, například hladina některého hormonu v těle, tak i tím, že nemusí stačit naměřit ji v jeden časový okamžik, ale je potřeba měření opakovat (například „průměrný“ systolický tlak v rámci delšího časového období). Nebo se musíme spoléhat na odpověď daného pacienta, který si danou záležitost nemusí pamatovat přesně, v horším případně si výrazně přikrášlovat, jako je u průměrného denního příjmu všech kalorií. Naměření některého regresoru může být sice učiněno lepší a přesnější metodou, ale může to být extrémně finančně nákladné (a nemáme možnost na dané vyšetření poslat všechny pacienty), časově náročné a kapacitně omezené (může být zvládnutelné na dané vyšetření poslat jen setinu všech pacientů), pro pacienta poměrně otravné, náročné nebo hůře i bolestivé (dané přesnější měření je ochotna podstoupit jen část pacientů). Pro některé pacienty dokonce nemusí být složitější procedura na měření přesného X_i vůči jejich stavu vůbec přípustná a tím by tedy ani nemohli být vybráni do validační skupiny, či samozřejmě se může jednat o kombinaci vícero zmíněných faktorů. Tedy je možné získat přesně naměřené regresory jen na části pacientů.

Proto bude existovat vedle validační skupiny i nevalidační skupina, kde sice odhad W_i nemusí být numericky nijak blízko X_i, ale pomocí validační skupiny budeme moci odhadnout parametry chybového modelu tak, aby se pomocí chybového modelu dalo W_i transformovat do $\hat{X}_i$, které oproti W_i už může být dostatečně rozumným a nestranným odhadem. Všechny parametry budeme moci napsat do jedné soustavy odhadovacích rovnic, ze které půjde odhadnout jakýkoliv z parametrů, které nás zajímají, což uděláme v 2. podkapitole. Zároveň z těchto odhadovacích rovnic si následně ve 3. podkapitole odvodíme sdruženou asymptotiku našich odhadů, ze které budeme moci odvodit intervaly spolehlivosti pro jednotlivé parametry.

Pro následující podkapitoly tedy předpokládejme náhodný výběr $\xi_1, \ldots, \xi_n$, kde $\xi_i \sim \mathrm{Alt}(\pi), 0 < \pi < 1$, který bude plně nezávislý na všem ostatním v našich

modelech. Pro $\xi_i = 1$, tedy validační skupinu, si označme pro pozdější výpočty vektor s přesnými regresory u i-tého pozorování jako

$$\boldsymbol{D_i} = \begin{pmatrix} 1 \\ \boldsymbol{Z_i} \\ X_i \end{pmatrix}.$$

Pro nevalidační skupinu zřejmě $\boldsymbol{D_i}$ nemůžeme používat, protože nemáme napozorované X_i, proto v soustavě odhadovacích rovnic budeme využívat značení

$$\hat{X}_i = \frac{W_i - \alpha_0 - \boldsymbol{\alpha_1}^T \boldsymbol{Z_i}}{\alpha_2},$$

což, jak víme, je nestranným odhadem X_i. Pak pro $\xi_i = 0$, tedy situaci, kdy i-té pozorování připadne do nevalidační skupiny, zaveďme značení

$$\hat{\boldsymbol{D}}_{\boldsymbol{i}} = \begin{pmatrix} 1 \\ \boldsymbol{Z_i} \\ \hat{X}_i \end{pmatrix}.$$

Upozorněme, že i když v $\boldsymbol{D_i}$ není zahrnuté W_i, neznamená to, že W_i nebude pro i-té pozorování zaznamenáno (mimo využití v samotném odhadu $\hat{X}_i$), ale bude využíváno „separátně“ a to jen při odhadování parametrů v chybovém modelu, na kterém se bude ze zřejmých důvodů podílet jen validační skupina.

4.2 Sestavení odhadovacích rovnic

Oproti 3. kapitole přidejme pro jednoduchost předpoklad nekorelovanosti ϵ_i a ψ_i. Učiníme také předpoklad na konkrétní podobu rozptylu chyb v chybovém modelu:

$$\mathrm{var}(\psi_i \mid \boldsymbol{Z_i}, X_i) = V(\boldsymbol{Z_i}, X_i) = \sigma_\psi^2.$$

V soustavě odhadovacích rovnic budeme současně odhadovat všechny bety, alfy i σ_ψ^2. Soustavu vybudujeme po částech pro i-té pozorování. Pokud dané pozorování bude pocházet z validační skupiny, můžeme standardně využít

$$\boldsymbol{D_i} \left(Y_i - \boldsymbol{D_i}^T \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \right),$$

což je po sečtení a položení rovnou nulovému vektoru ekvivalentní soustavě normálních rovnic, které řeší odhad metodou nejmenších čtverců. Pokud naopak bude pocházet pozorování z nevalidační skupiny, ve 3. kapitole jsme si odvodili upravený tvar

$$\hat{\boldsymbol{D}}_{\boldsymbol{i}} \left(Y_i - \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \right) + \begin{pmatrix} 0 \\ \boldsymbol{0_{q,1}} \\ \frac{\beta_2}{\alpha_2^2} \sigma_\psi^2 \end{pmatrix}.$$

Když pak budeme chtít odhadovat bety, můžeme oba zápisy napsat současně díky veličině ξ_i, která jednu z variant „vynuluje“ a nechá jen ten, který odpovídá skutečně získaným informacím:

$$\xi_i \boldsymbol{D_i} \left(Y_i - \boldsymbol{D_i}^T \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \right) + (1 - \xi_i) \left[\hat{\boldsymbol{D}}_{\boldsymbol{i}} \left(Y_i - \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \right) + \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \frac{\beta_2}{\alpha_2^2} \sigma_\psi^2 \end{pmatrix} \right].$$

I když v tuto chvíli máme již odvozenou část pro odhadování bet, kterou přímo využijeme do soustavy odhadovacích rovnic, můžeme na tomto místě poukázat, že i při této „směsi modelů“ pro validační a nevalidační skupinu lze explicitní vyjádření bet provést jednoduše. Pro náš náhodný výběr položíme sumu výše zmíněného rovnou nulovému vektoru:

$$\sum_{i=1}^{n} \left(\xi_i \boldsymbol{D_i} \left(Y_i - \boldsymbol{D_i}^T \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\beta}_2 \end{pmatrix} \right) + (1 - \xi_i) \left[\hat{\boldsymbol{D}}_{\boldsymbol{i}} \left(Y_i - \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\beta}_2 \end{pmatrix} \right) + \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \frac{\hat{\beta}_2}{\hat{\alpha}_2^2} \hat{\sigma}_\psi^2 \end{pmatrix} \right] \right) = \mathbf{0_{(q+2),1}},$$

což poupravíme za chvíli s využitím značení $\mathrm{diag}(a_1, \ldots, a_k)$ pro diagonální matici řádu k, která má na diagonále postupně prvky $a_1, \ldots, a_k$. Ne vždy řád této matice budeme explicitně uvádět a to v takové situaci, kdy to bude vyplývat z kontextu jako za chvíli pro $\mathrm{diag}\left(\frac{\hat{\sigma}_\psi^2}{\hat{\alpha}_2^2}\right) = \mathrm{diag}\left(0, \ldots, 0, \frac{\hat{\sigma}_\psi^2}{\hat{\alpha}_2^2}\right)$, kde bude nenulová pouze poslední diagonální pozice (zkratku zavádíme z prostorových důvodů u výpočtů v celé kapitole. Než o něco níže uvedeme jak dostat odhady $\hat{\sigma}_\psi^2$ a $\hat{\alpha}_2^2$, dokončíme explicitní vyjádření bet pomocí těchto odhadů:

$$\sum_{i=1}^{n} \left(\xi_i \boldsymbol{D_i} \boldsymbol{D_i}^T + (1 - \xi_i) \left[\hat{\boldsymbol{D}}_{\boldsymbol{i}} \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T - \mathrm{diag}\left(\frac{\hat{\sigma}_\psi^2}{\hat{\alpha}_2^2} \right) \right] \right) \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\beta}_2 \end{pmatrix} = \sum_{i=1}^{n} \left(\xi_i \boldsymbol{D_i} + (1 - \xi_i) \hat{\boldsymbol{D}}_{\boldsymbol{i}} \right) Y_i.$$

Vidíme, že se vektor $\left(\hat{\beta}_0, \ \hat{\boldsymbol{\beta}}_{\mathbf{1}}^T, \ \hat{\beta}_2\right)^T$ opakuje v každém sčítanci a nezávisí na indexu sumy, lze ho tedy vytknout zprava. Předpokládejme regularitu matice

$$\left[\sum_{i=1}^{n} \left(\xi_i \boldsymbol{D_i} \boldsymbol{D_i}^T + (1 - \xi_i) \left[\hat{\boldsymbol{D}}_{\boldsymbol{i}} \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T - \mathrm{diag}\left(\frac{\hat{\sigma}_\psi^2}{\hat{\alpha}_2^2} \right) \right] \right) \right],$$

tedy existuje k ní matice inverzní, a explicitní odhad bet má tvar

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\beta}_2 \end{pmatrix} = \left[\sum_{i=1}^{n} \left(\xi_i \boldsymbol{D_i} \boldsymbol{D_i}^T + (1 - \xi_i) \left[\hat{\boldsymbol{D}}_{\boldsymbol{i}} \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T - \mathrm{diag}\left(\frac{\hat{\sigma}_\psi^2}{\hat{\alpha}_2^2} \right) \right] \right) \right]^{-1} \sum_{i=1}^{n} \left(\xi_i \boldsymbol{D_i} + (1 - \xi_i) \hat{\boldsymbol{D}}_{\boldsymbol{i}} \right) Y_i.$$

Tento odhad využijeme při výpočtech pro simulační část. Nyní pojďme dokončit sestavení odhadovacích rovnic. Pro získání odhadů alf využijeme přímočaře analogii k tomu, co jsme využívali pro bety u validačního výběru, čili z lineárního chybového modelu s využitím validační skupiny dostáváme skóre

$$\xi_i \boldsymbol{D_i} \left(W_i - \boldsymbol{D_i}^T \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2 \end{pmatrix} \right).$$

Pro odhad rozptylu použijeme odvození z přímočarého odhadu

$$\hat{\sigma}^2_\psi = \frac{1}{\sum_{i=1}^n \xi_i} \sum_{i=1}^n \xi_i \left(W_i - \boldsymbol{D_i}^T \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2 \end{pmatrix} \right)^2,$$

tedy pro jedno pozorování z validační skupiny odpovídající tvar odhadovací rovnice bude

$$\left(\left(W_i - \boldsymbol{D_i}^T \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2 \end{pmatrix} \right)^2 - \sigma^2_\psi \right).$$

Celkově při spojení všech částí dostáváme pro jedno pozorování

$$\boldsymbol{U_i} = \begin{pmatrix} \xi_i \boldsymbol{D_i} \left(Y_i - \boldsymbol{D_i}^T \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \right) + (1 - \xi_i) \left[\hat{\boldsymbol{D}}_{\boldsymbol{i}} \left(Y_i - \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \right) + \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \frac{\beta_2}{\alpha_2^2} \sigma^2_\psi \end{pmatrix} \right] \\ \xi_i \boldsymbol{D_i} \left(W_i - \boldsymbol{D_i}^T \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2 \end{pmatrix} \right) \\ \xi_i \left(\left(W_i - \boldsymbol{D_i}^T \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2 \end{pmatrix} \right)^2 - \sigma^2_\psi \right) \end{pmatrix}.$$

Odhadovací rovnice sestavíme tak, že sečteme všechny $\boldsymbol{U_i}$ a položíme to rovnou nulovému vektoru:

$$\sum_{i=1}^n \boldsymbol{U_i} = \mathbf{0_{(2q+5),1}}.$$

Vidíme, že soustava odhadovacích rovnic dává smysl, rozumně odhaduje všech $2q + 5$ parametrů a podmíněná (respektive i nepodmíněná) střední hodnota je pro skutečné parametry rovna nulovému vektoru (což kdyby neplatilo, tak by položení rovnosti s nulovým vektorem byl naprostý nesmysl a musela by se provést korekce nebo změna odhadování některého parametru). Samozřejmě si můžeme povšimnout, že rovnost by platila i po vydělení n. Pak

$$\frac{1}{n} \sum_{i=1}^n \boldsymbol{U_i}$$

není nic jiného než průměr pro $\boldsymbol{U_1}, \ldots, \boldsymbol{U_n}$, které jsou zřejmě stejně rozdělené a nezávislé. To je výborný základ pro studování asymptotiky a následné odvozování, čemuž se budeme věnovat v další podkapitole.

4.3 Asymptotické vlastnosti

4.3.1 Motivace

Označme si vektor skutečných hodnot všech zkoumaných parametrů (bety, alfy a σ_ψ^2) jako $\boldsymbol{\gamma}$. Ke konci předchozí kapitoly jsme si ukázali podobu odhadovacích rovnic a označme si odhad, který je řeší, jako $\hat{\boldsymbol{\gamma}}$, tedy víme že

$$\sum_{i=1}^{n} \boldsymbol{U_i}\left(\hat{\boldsymbol{\gamma}}\right) = \mathbf{0_{(2q+5),1}}$$

a pro skutečné $\boldsymbol{\gamma}$ platí

$$\mathsf{E}\, \boldsymbol{U}_i\left(\boldsymbol{\gamma}\right) = \mathbf{0_{(2q+5),1}}.$$

Dále vidíme, že $\boldsymbol{U_1}\left(\boldsymbol{\gamma}\right), \ldots, \boldsymbol{U_n}\left(\boldsymbol{\gamma}\right)$ jsou zřejmě stejně rozdělené a nezávislé. Předpokládejme, že jejich druhé momenty existují a existuje regulární rozptylová matice

$$\boldsymbol{\Sigma} := \operatorname{var}\left(\boldsymbol{U_i}\left(\boldsymbol{\gamma}\right)\right).$$

Z mnohorozměrné Centralní limitní věty dostáváme

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{U_i}\left(\boldsymbol{\gamma}\right) \xrightarrow{d} \mathcal{N}_{2q+5}\left(\mathbf{0_{2q+5,1}}, \boldsymbol{\Sigma}\right).$$

Máme tak odvozenou asymptotiku pro upravenou skórovou statistiku. Pro praktické účely se ale hodí mít odvozenou asymptotiku pro samotný odhad vektoru parametrů $\hat{\boldsymbol{\gamma}}$. Podrobný důkaz konzistence a asymptotické normality $\hat{\boldsymbol{\gamma}}$ v této práci nebudeme uvádět. Lze jej získat ověřením podmínek pro asymptotickou normalitu Z-odhadů, viz například [11] nebo [12].

Při aplikaci věty o asymptotickém rozdělení Z-odhadů dostáváme

$$\sqrt{n}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right) \xrightarrow{d} \mathcal{N}_{2q+5}\left(\mathbf{0_{2q+5,1}}, \boldsymbol{G^{-1}\Sigma(G^{-1})}^T\right),$$

kde (mimo dalších dodatečných předpokladů pro aplikaci věty) předpokládáme regularitu

$$\boldsymbol{G} := -\,\mathsf{E}\,\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\gamma}^T}\boldsymbol{U_i}\left(\boldsymbol{\gamma}\right).$$

Tato asymptotika je pro nás důležitá, proto v této práci ukážeme explicitní vyjádření asymptotické matice a na základě tohoto asymptotického rozdělení budeme chtít provést simulační studie. Abychom tak mohli učinit, nejprve si v další sekci odvodíme explicitní vyjádření inverzní matice $\boldsymbol{G^{-1}}$, následně si odvodíme vyjádření pro rozptylovou matici $\boldsymbol{\Sigma}$ a odvodíme podobu pro asymptotickou rozptylovou matici u asymptotiky pro $\boldsymbol{\gamma}$, označme ji $\boldsymbol{\Sigma}_\gamma := \boldsymbol{G^{-1}\Sigma(G^{-1})}^T$.

4.3.2 Inverzní matice

Protože v této sekci bude provedena spousta výpočtů včetně derivování podle různých parametrů, na začátek pro větší přehlednost uvedeme všechny důležité vztahy a značení, které bude potřeba mít dostatečně osvojené při následujících výpočtech a odvozování. Připomeňme nejprve náš hlavní model

$$Y_i = \beta_0 + \boldsymbol{\beta_1}^T \boldsymbol{Z_i} + \beta_2 X_i + \epsilon_i,$$

kde $\mathsf{E}[\epsilon_i \mid \boldsymbol{Z_i}, X_i] = 0$, $\boldsymbol{Z_i}$ má q složek a $i = 1, \ldots, n, n \in \mathbb{N}$. Předpokládáme zobecněný chybový model

$$W_i = \alpha_0 + \boldsymbol{\alpha_1}^T \boldsymbol{Z_i} + \alpha_2 X_i + \psi_i,$$

kde $\mathsf{E}[\psi_i \mid \boldsymbol{Z_i}, X_i] = 0$, $\mathrm{var}(\psi_i \mid \boldsymbol{Z_i}, X_i) = \sigma_\psi^2$ a $i = 1, \ldots, n$, zatímco předpokládáme nekorelovanost ψ_i a ϵ_i. Dále předpokládáme náhodný výběr $\xi_1, \ldots, \xi_n$, kde $\xi_i \sim \mathrm{Alt}(\pi), 0 < \pi < 1$, který bude plně nezávislý na všem ostatním. Připomeňme značení $\xi_i = 0$ pro nevalidační skupinu, ve které pro i-té pozorování známe $Y_i, \boldsymbol{Z_i}$ a W_i, zatímco u validační skupiny s $\xi_i = 1$ pozorujeme oproti nevalidační skupině navíc i X_i. Pro nevalidační skupinu využíváme nestranný odhad

$$\hat{X}_i = \frac{W_i - \alpha_0 - \boldsymbol{\alpha_1}^T \boldsymbol{Z_i}}{\alpha_2}$$

a dále pro nevalidační, respektive validační skupinu jsme zavedli značení

$$\hat{\boldsymbol{D}}_{\boldsymbol{i}} = \begin{pmatrix} 1 \\ \boldsymbol{Z_i} \\ \hat{X}_i \end{pmatrix}, \text{ respektive } \boldsymbol{D_i} = \begin{pmatrix} 1 \\ \boldsymbol{Z_i} \\ X_i \end{pmatrix}.$$

Pro příspěvek i-tého pozorování do odhadovacích rovnic opět vypusťme argument u značení a mějme na místě parametrů jejich skutečné hodnoty:

$$\boldsymbol{U_i} = \begin{pmatrix} \xi_i \boldsymbol{D_i} \left(Y_i - \boldsymbol{D_i}^T \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \right) + (1 - \xi_i) \left[\hat{\boldsymbol{D}}_{\boldsymbol{i}} \left(Y_i - \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \right) + \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \frac{\beta_2}{\alpha_2^2} \sigma_\psi^2 \end{pmatrix} \right] \\ \xi_i \boldsymbol{D_i} \left(W_i - \boldsymbol{D_i}^T \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2 \end{pmatrix} \right) \\ \xi_i \left(\left(W_i - \boldsymbol{D_i}^T \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2 \end{pmatrix} \right)^2 - \sigma_\psi^2 \right) \end{pmatrix}.$$

Pro praktické účely si označme prvních $q+2$ složek vektoru $\boldsymbol{U_i}$ jako $\boldsymbol{N_i}$, tedy

$$\boldsymbol{N_i} = \xi_i \boldsymbol{D_i} \left(Y_i - \boldsymbol{D_i}^T \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \right) + (1 - \xi_i) \left[\hat{\boldsymbol{D}}_{\boldsymbol{i}} \left(Y_i - \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \right) + \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \frac{\beta_2}{\alpha_2^2} \sigma_\psi^2 \end{pmatrix} \right].$$

Pro lepší zápis zaveďme značení pro vektory alf, bet a též značení pro vektor všech parametrů, tedy nechť

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix}, \boldsymbol{\alpha} = \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2 \end{pmatrix}, \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \\ \sigma_\psi^2 \end{pmatrix}.$$

Naším cílem je vyjádřit matici

$$\boldsymbol{G} = -\,\mathsf{E}\,\frac{\mathrm{d}\boldsymbol{U_i}}{\mathrm{d}\boldsymbol{\gamma}},$$

což provedeme za pomoci několika následujících pomocných výpočtů z důvodu přehlednosti a srozumitelnosti. Zejména pro úpravy ve výpočtu

$$-\,\mathsf{E}\,\frac{\mathrm{d}\boldsymbol{N_i}}{\mathrm{d}\boldsymbol{\alpha}}$$

se bude hodit mít připravené mezivýpočty pro vyjádření nejkomplikovanější části obsahující alfy v $\hat{X}_i$, zejména mít vyjádřenou derivaci

$$\frac{\mathrm{d}\hat{\boldsymbol{D}}_{\boldsymbol{i}}\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}}{\mathrm{d}\boldsymbol{\alpha}},$$

proto začneme se základními výpočty pro $\hat{X}_i$. Připomeňme, že

$$\hat{X}_i = \frac{W_i - \alpha_0 - \boldsymbol{\alpha_1^T Z_i}}{\alpha_2} = X_i + \frac{\psi_i}{\alpha_2}.$$

Protože $\mathsf{E}[\psi_i \mid \boldsymbol{Z_i}, X_i] = 0$, pak dostáváme

$$\mathsf{E}\,\hat{X}_i = \mathsf{E}\,X_i + \mathsf{E}\left[\frac{1}{\alpha_2}\,\mathsf{E}[\psi_i \mid \boldsymbol{Z_i}, X_i]\right] = \mathsf{E}\,X_i,$$

$$\begin{aligned}\mathsf{E}\,\hat{X}_i^2 &= \mathsf{E}\left(X_i + \frac{\psi_i}{\alpha_2}\right)^2 = \mathsf{E}\,X_i^2 + \frac{2}{\alpha_2}\,\mathsf{E}\,X_i\psi_i + \frac{1}{\alpha_2^2}\,\mathsf{E}\,\psi_i^2 = \\ &= \mathsf{E}\,X_i^2 + \frac{2}{\alpha_2}\,\mathsf{E}\,[X_i\,\mathsf{E}[\psi_i \mid \boldsymbol{Z_i}, X_i]] + \frac{1}{\alpha_2^2}\sigma_\psi^2 = \mathsf{E}\,X_i^2 + \frac{1}{\alpha_2^2}\sigma_\psi^2.\end{aligned}$$

Parciální derivace $\hat{X}_i$ lze vypočítat jednoduše:

$$\frac{\partial \hat{X}_i}{\partial \alpha_0} = -\frac{1}{\alpha_2},$$

$$\frac{\partial \hat{X}_i^2}{\partial \alpha_0} = 2\hat{X}_i\left(-\frac{1}{\alpha_2}\right) = -\frac{2}{\alpha_2}\hat{X}_i,$$

dále pro $l = 1, \ldots, q$:

$$\frac{\partial \hat{X}_i}{\partial \alpha_{1,l}} = -\frac{Z_{i,l}}{\alpha_2},$$

$$\frac{\partial \hat{X}_i^2}{\partial \alpha_{1,l}} = -\frac{2Z_{i,l}}{\alpha_2}\hat{X}_i,$$

a konečně

$$\frac{\partial \hat{X}_i}{\partial \alpha_2} = -1\frac{W_i - \alpha_0 - \boldsymbol{\alpha_1}^T\boldsymbol{Z_i}}{\alpha_2^2} = -\frac{1}{\alpha_2}\hat{X}_i,$$

$$\frac{\partial \hat{X}_i^2}{\partial \alpha_2} = -2\frac{(W_i - \alpha_0 - \boldsymbol{\alpha_1}^T\boldsymbol{Z_i})^2}{\alpha_2^3} = -\frac{2}{\alpha_2}\hat{X}_i^2.$$

Nyní máme všechny drobné úpravy pro $\hat{X}_i$ přichystané, proto můžeme přistoupit k parciálním derivacím $-\hat{\boldsymbol{D}}_{\boldsymbol{i}}\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}$ podle alf. Nejprve si ale ještě uveďme maticovou podobu pro $\hat{\boldsymbol{D}}_{\boldsymbol{i}}\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T$:

$$\hat{\boldsymbol{D}}_{\boldsymbol{i}}\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T = \begin{pmatrix} 1 & \boldsymbol{Z_i}^T & \hat{X}_i \\ \boldsymbol{Z_i} & \boldsymbol{Z_i}\boldsymbol{Z_i}^T & \hat{X}_i\boldsymbol{Z_i} \\ \hat{X}_i & \hat{X}_i\boldsymbol{Z_i}^T & \hat{X}_i^2 \end{pmatrix},$$

což při derivování podle alf vynuluje všechny pozice mimo poslední řádek a poslední sloupec. Spočtěme jednotlivé parciální derivace:

$$\frac{\partial - \hat{\boldsymbol{D}}_{\boldsymbol{i}}\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}}{\partial \alpha_0} = \begin{pmatrix} 0 & \boldsymbol{0_{1,q}} & \frac{1}{\alpha_2} \\ \boldsymbol{0_{q,1}} & \boldsymbol{0_{q,q}} & \frac{1}{\alpha_2}\boldsymbol{Z_i} \\ \frac{1}{\alpha_2} & \frac{1}{\alpha_2}\boldsymbol{Z_i}^T & \frac{2}{\alpha_2}\hat{X}_i \end{pmatrix}\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} = \frac{1}{\alpha_2}\begin{pmatrix} \beta_2 \\ \beta_2\boldsymbol{Z_i} \\ \beta_0 + \boldsymbol{\beta_1}^T\boldsymbol{Z_i} + 2\beta_2\hat{X}_i \end{pmatrix},$$

dále pro $l = 1, \ldots, q$:

$$\frac{\partial - \hat{\boldsymbol{D}}_{\boldsymbol{i}}\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}}{\partial \alpha_{1,l}} = \frac{1}{\alpha_2}\begin{pmatrix} 0 & \boldsymbol{0_{1,q}} & Z_{i,l} \\ \boldsymbol{0_{q,1}} & \boldsymbol{0_{q,q}} & Z_{i,l}\boldsymbol{Z_i} \\ Z_{i,l} & Z_{i,l}\boldsymbol{Z_i}^T & 2Z_{i,l}\hat{X}_i \end{pmatrix}\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} = \frac{Z_{i,l}}{\alpha_2}\begin{pmatrix} \beta_2 \\ \beta_2\boldsymbol{Z_i} \\ \beta_0 + \boldsymbol{\beta_1}^T\boldsymbol{Z_i} + 2\beta_2\hat{X}_i \end{pmatrix},$$

na závěr parciální derivace podle α_2, kde ve všech pozicích po derivování vychází $\frac{\hat{X}_i}{\alpha_2}$. které ihned vytkneme

$$\frac{\partial - \hat{\boldsymbol{D}}_{\boldsymbol{i}}\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}}{\partial \alpha_2} = \frac{\hat{X}_i}{\alpha_2}\begin{pmatrix} 0 & \boldsymbol{0_{1,q}} & 1 \\ \boldsymbol{0_{q,1}} & \boldsymbol{0_{q,q}} & \boldsymbol{Z_i} \\ 1 & \boldsymbol{Z_i}^T & 2\hat{X}_i \end{pmatrix}\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} = \frac{\hat{X}_i}{\alpha_2}\begin{pmatrix} \beta_2 \\ \beta_2\boldsymbol{Z_i} \\ \beta_0 + \boldsymbol{\beta_1}^T\boldsymbol{Z_i} + 2\beta_2\hat{X}_i \end{pmatrix}.$$

Nyní už máme o něco blíže, abychom se ve výpočtech vrátili o jedno patro „výše“ a začali jsme zkoumat střední hodnotu parciálních derivací $\boldsymbol{N_i}$ podle alf. Ještě ale předtím si předpřipravíme jeden výpočet a to konkrétně pro poslední složku posledně spočtených výpočtů. Vidíme, že se od této složky v rámci parciální derivace $\boldsymbol{N_i}$ odečte odpovídající násobek odezvy Y_i, tedy například pro parciální derivaci podle α_0 dojde k úpravě

$$\begin{aligned}
&\frac{1}{\alpha_2}\,\mathsf{E}\left(\beta_0 + \boldsymbol{\beta_1}^T\boldsymbol{Z_i} + 2\beta_2\hat{X}_i - Y_i\right) = \\
=&\frac{1}{\alpha_2}\left(\cancel{\beta_0} + \cancel{\boldsymbol{\beta_1}^T\,\mathsf{E}\,\boldsymbol{Z_i}} + 2\beta_2\,\mathsf{E}\,\hat{X}_i - \cancel{\beta_0} - \cancel{\boldsymbol{\beta_1}^T\,\mathsf{E}\,\boldsymbol{Z_i}} - \beta_2\,\mathsf{E}\,X_i\right) = \beta_2\,\mathsf{E}\,X_i.
\end{aligned}$$

Obdobně při variantě s parciální derivací podle $\alpha_{1,l}$ pro $l = 1, \ldots, q$, dojde k analogickému postupu, pouze hned na začátku přepíšeme výraz na střední hodnotu podmíněné střední hodnoty a vytkneme $Z_{i,l}$, protože je to měřitelnou funkcí podmínky. Ihned se tak dojde k výsledku $\beta_2\,\mathsf{E}\,Z_{i,l}X_i$. Komplikovanější to bude v případě s parciální derivací podle α_2, u výpočtu je lepší postupovat pomaleji:

$$\begin{aligned}
&\frac{1}{\alpha_2}\,\mathsf{E}\,\hat{X}_i\left(\beta_0 + \boldsymbol{\beta_1}^T\boldsymbol{Z_i} + 2\beta_2\hat{X}_i - Y_i\right) = \frac{1}{\alpha_2}\,\mathsf{E}\,\hat{X}_i\left(2\beta_2\hat{X}_i - \beta_2 X_i - \epsilon_i\right) = \\
=&\frac{1}{\alpha_2}\,\mathsf{E}\left[\mathsf{E}\left[\left(X_i + \frac{\psi_i}{\alpha_2}\right)\left(2\beta_2 X_i + 2\beta_2\frac{\psi_i}{\alpha_2} - \beta_2 X_i - \epsilon_i\right) \mid \boldsymbol{Z_i}, X_i\right]\right] = \\
=&\frac{1}{\alpha_2}\mathsf{E}\left[X_i\,\mathsf{E}\left[\beta_2 X_i + 2\beta_2\frac{\psi_i}{\alpha_2} - \epsilon_i \mid \boldsymbol{Z_i}, X_i\right] + \mathsf{E}\left[\frac{\psi_i}{\alpha_2}\left(\beta_2 X_i + 2\beta_2\frac{\psi_i}{\alpha_2} - \epsilon_i\right) \mid \boldsymbol{Z_i}, X_i\right]\right] = \\
=&\frac{1}{\alpha_2}\,\mathsf{E}\left[\beta_2 X_i^2 + \frac{1}{\alpha_2}\,\mathsf{E}\left[\beta_2 X_i\psi_i + 2\beta_2\frac{\psi_i^2}{\alpha_2} - \epsilon_i\psi_i \mid \boldsymbol{Z_i}, X_i\right]\right] = \\
=&\frac{1}{\alpha_2}\,\mathsf{E}\left[\beta_2 X_i^2 + \frac{2\beta_2}{\alpha_2^2}\,\mathsf{E}\left[\psi_i^2 \mid \boldsymbol{Z_i}, X_i\right] - \frac{1}{\alpha_2}\,\mathsf{E}\left[\epsilon_i\psi_i \mid \boldsymbol{Z_i}, X_i\right]\right] = \\
=&\frac{1}{\alpha_2}\beta_2\,\mathsf{E}\,X_i^2 + \frac{2\beta_2}{\alpha_2^3}\sigma_\psi^2 - \frac{1}{\alpha_2^2}\mathrm{cov}\left(\epsilon_i, \psi_i\right) = \frac{1}{\alpha_2}\beta_2\,\mathsf{E}\,X_i^2 + \frac{2\beta_2}{\alpha_2^3}\sigma_\psi^2.
\end{aligned}$$

Tímto máme vše připravené pro jednoduché spočtení $-\,\mathsf{E}\,\frac{\mathrm{d}\boldsymbol{N_i}}{\mathrm{d}\boldsymbol{\alpha}}$. Část $\boldsymbol{N_i}$ odpovídající validační skupině se zderivuje na nulu, neboť v sobě vůbec neobsahuje alfy. Také zde využijeme předpoklad nezávislosti ξ_i vůči všemu ostatnímu:

$$\begin{aligned}
-\,\mathsf{E}\,\frac{\partial\boldsymbol{N_i}}{\partial\alpha_0} &= -\,\mathsf{E}\,(1-\xi_i)\left[Y_i\begin{pmatrix}0\\ \mathbf{0_{q,1}}\\ -\frac{1}{\alpha_2}\end{pmatrix} + \frac{1}{\alpha_2}\begin{pmatrix}\beta_2\\ \beta_2\boldsymbol{Z_i}\\ \beta_0 + \boldsymbol{\beta_1}^T\boldsymbol{Z_i} + 2\beta_2\hat{X}_i\end{pmatrix}\right] \\
&= -\,\mathsf{E}\,(1-\xi_i)\,\mathsf{E}\left[\frac{1}{\alpha_2}\begin{pmatrix}\beta_2\\ \beta_2\boldsymbol{Z_i}\\ \beta_0 + \boldsymbol{\beta_1}^T\boldsymbol{Z_i} + 2\beta_2\hat{X}_i - Y_i\end{pmatrix}\right] \\
&= -\frac{1-\pi}{\alpha_2}\begin{pmatrix}\beta_2\\ \beta_2\,\mathsf{E}\,\boldsymbol{Z_i}\\ \beta_2\,\mathsf{E}\,X_i\end{pmatrix} = -\frac{1-\pi}{\alpha_2}\beta_2\begin{pmatrix}1\\ \mathsf{E}\,\boldsymbol{Z_i}\\ \mathsf{E}\,X_i\end{pmatrix},
\end{aligned}$$

dále pro $l = 1, \ldots, q$:

$$
\begin{aligned}
-\,\mathsf{E}\,\frac{\partial \boldsymbol{N_i}}{\partial \alpha_{1,l}} &= -\,\mathsf{E}\,(1-\xi_i)\,\mathsf{E}\left[\frac{Z_{i,l}}{\alpha_2}\begin{pmatrix}\beta_2\\ \beta_2\boldsymbol{Z_i}\\ \beta_0+\boldsymbol{\beta_1}^T\boldsymbol{Z_i}+2\beta_2\hat{X}_i-Y_i\end{pmatrix}\right]\\
&= -\frac{1-\pi}{\alpha_2}\beta_2\begin{pmatrix}\mathsf{E}\,Z_{i,l}\\ \mathsf{E}\,Z_{i,l}\boldsymbol{Z_i}\\ \mathsf{E}\,Z_{i,l}X_i\end{pmatrix},
\end{aligned}
$$

a u posledního výpočtu opět opatrněji, tentokrát se v $\boldsymbol{N_i}$ derivuje také vektor s korekcí, která v sobě obsahuje α_2:

$$
\begin{aligned}
-\,\mathsf{E}\,\frac{\partial \boldsymbol{N_i}}{\partial \alpha_2} &= -\,\mathsf{E}\,(1-\xi_i)\,\mathsf{E}\left[\frac{\hat{X}_i}{\alpha_2}\begin{pmatrix}\beta_2\\ \beta_2\boldsymbol{Z_i}\\ \beta_0+\boldsymbol{\beta_1}^T\boldsymbol{Z_i}+2\beta_2\hat{X}_i-Y_i\end{pmatrix}-\begin{pmatrix}0\\ \mathbf{0_{q,1}}\\ 2\frac{\beta_2}{\alpha_2^3}\sigma_\psi^2\end{pmatrix}\right]\\
&= -\,(1-\pi)\,\mathsf{E}\left[\frac{\beta_2}{\alpha_2}\begin{pmatrix}\mathsf{E}\,X_i\\ \mathsf{E}\,X_i\boldsymbol{Z_i}\\ \mathsf{E}\,X_i^2+\frac{2}{\alpha_2^2}\sigma_\psi^2\end{pmatrix}-\frac{\beta_2}{\alpha_2}\begin{pmatrix}0\\ \mathbf{0_{q,1}}\\ \frac{2}{\alpha_2^2}\sigma_\psi^2\end{pmatrix}\right] = -\frac{1-\pi}{\alpha_2}\beta_2\begin{pmatrix}\mathsf{E}\,X_i\\ \mathsf{E}\,X_i\boldsymbol{Z_i}\\ \mathsf{E}\,X_i^2\end{pmatrix}.
\end{aligned}
$$

Tímto máme hotové výpočty s parciálními derivacemi $\boldsymbol{N_i}$ podle alf, nyní postupme na výpočty s parciálními derivacemi podle bet:

$$
\begin{aligned}
-\,\mathsf{E}\,\frac{\partial \boldsymbol{N_i}}{\partial \beta_0} &= \mathsf{E}\,\xi_i\boldsymbol{D_i}\boldsymbol{D_i}^T\begin{pmatrix}1\\ \mathbf{0_{q,1}}\\ 0\end{pmatrix}+\mathsf{E}\,(1-\xi_i)\,\hat{\boldsymbol{D}}_i\hat{\boldsymbol{D}}_i^T\begin{pmatrix}1\\ \mathbf{0_{q,1}}\\ 0\end{pmatrix}=\\
&= \mathsf{E}\,\xi_i\,\mathsf{E}\,\boldsymbol{D_i}+\mathsf{E}\,(1-\xi_i)\,\mathsf{E}\,\hat{\boldsymbol{D}}_i = \pi\begin{pmatrix}1\\ \mathsf{E}\,\boldsymbol{Z_i}\\ \mathsf{E}\,X_i\end{pmatrix}+(1-\pi)\begin{pmatrix}1\\ \mathsf{E}\,\boldsymbol{Z_i}\\ \mathsf{E}\,\hat{X}_i\end{pmatrix}=\begin{pmatrix}1\\ \mathsf{E}\,\boldsymbol{Z_i}\\ \mathsf{E}\,X_i\end{pmatrix},
\end{aligned}
$$

dále pro $l = 1, \ldots, q$:

$$
\begin{aligned}
-\,\mathsf{E}\,\frac{\partial \boldsymbol{N_i}}{\partial \beta_{1,l}} &= \mathsf{E}\,\xi_i\,\mathsf{E}\,Z_{i,l}\boldsymbol{D_i}+\mathsf{E}\,(1-\xi_i)\,\mathsf{E}\,Z_{i,l}\hat{\boldsymbol{D}}_i=\\
&= \begin{pmatrix}\mathsf{E}\,Z_{i,l}\\ \mathsf{E}\,Z_{i,l}\boldsymbol{Z_i}\\ \pi\,\mathsf{E}\,Z_{i,l}X_i+(1-\pi)\,\mathsf{E}\,Z_{i,l}\hat{X}_i\end{pmatrix}=\begin{pmatrix}\mathsf{E}\,Z_{i,l}\\ \mathsf{E}\,Z_{i,l}\boldsymbol{Z_i}\\ \mathsf{E}\,Z_{i,l}X_i\end{pmatrix},
\end{aligned}
$$

a na závěr

$$
\begin{aligned}
-\mathsf{E}\,\frac{\partial \boldsymbol{N_i}}{\partial \beta_2} &= \mathsf{E}\,\xi_i\,\mathsf{E}\,X_i\boldsymbol{D_i} + \mathsf{E}\,(1-\xi_i)\left[\mathsf{E}\,\hat{X}_i\hat{\boldsymbol{D}}_{\boldsymbol{i}} - \begin{pmatrix} 0 \\ \mathbf{0_{q,1}} \\ \frac{1}{\alpha_2^2}\sigma_\psi^2 \end{pmatrix}\right] = \\
&= \begin{pmatrix} \pi\,\mathsf{E}\,X_i + (1-\pi)\,\mathsf{E}\,\hat{X}_i \\ \pi\,\mathsf{E}\,X_i\boldsymbol{Z_i} + (1-\pi)\,\mathsf{E}\,\hat{X}_i\boldsymbol{Z_i} \\ \pi\,\mathsf{E}\,X_i^2 + (1-\pi)\,\mathsf{E}\,\hat{X}_i^{\,2} - (1-\pi)\frac{1}{\alpha_2^2}\sigma_\psi^2 \end{pmatrix} \\
&= \begin{pmatrix} \mathsf{E}\,X_i \\ \mathsf{E}\,X_i\boldsymbol{Z_i} \\ \pi\,\mathsf{E}\,X_i^2 + (1-\pi)\left[\mathsf{E}\,X_i^2 + \frac{1}{\alpha_2^2}\sigma_\psi^2\right] - (1-\pi)\frac{1}{\alpha_2^2}\sigma_\psi^2 \end{pmatrix} = \begin{pmatrix} \mathsf{E}\,X_i \\ \mathsf{E}\,X_i\boldsymbol{Z_i} \\ \mathsf{E}\,X_i^2 \end{pmatrix}.
\end{aligned}
$$

Poslední chybějící výpočet $\boldsymbol{N_i}$ obsahuje parciální derivaci podle σ_ψ^2, což vyjde nenulová jen poslední složka a to $-\frac{1-\pi}{\alpha_2^2}\beta_2$. Tímto máme výpočty s $\boldsymbol{N_i}$ hotové. Dále chceme provést analogické výpočty pro další část $\boldsymbol{U_i}$ a to pro

$$
\xi_i\boldsymbol{D_i}\left(W_i - \boldsymbol{D_i}^T\begin{pmatrix}\alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2\end{pmatrix}\right).
$$

Protože tato část obsahuje z parametrů jen alfy, tak parciální derivace podle jiných parametrů budou obsahovat nuly. A i samotné výpočty obsahující parciální derivace podle alf jsou velice snadné. Proto zde ukážeme výpočet jen pro jednu variantu a to s parciální derivací podle α_2:

$$
\mathsf{E}\,\xi_i\boldsymbol{D_i}\boldsymbol{D_i}^T\begin{pmatrix}0 \\ \mathbf{0_{q,1}} \\ 1\end{pmatrix} = \mathsf{E}\,\xi_i\,\mathsf{E}\,X_i\boldsymbol{D_i} = \pi\begin{pmatrix}\mathsf{E}\,X_i \\ \mathsf{E}\,X_i\boldsymbol{Z_i} \\ \mathsf{E}\,X_i^2\end{pmatrix}.
$$

Poslední zbývající část $\boldsymbol{U_i}$ je

$$
\xi_i\left(\left(W_i - \boldsymbol{D_i}^T\begin{pmatrix}\alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2\end{pmatrix}\right)^2 - \sigma_\psi^2\right).
$$

Opět parciální derivace podle bet jsou nulové, zatímco u výpočtu s parciální derivací podle σ_ψ^2 dostáváme ihned π. Výpočet pro alfy bude opět jednoduchý a opět předvedeme jen variantu s parciální derivací α_2:

$$
\mathsf{E}\,\xi_i\,\mathsf{E}\,2X_i\left(W_i - \boldsymbol{D_i}^T\begin{pmatrix}\alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2\end{pmatrix}\right) = \pi\,\mathsf{E}\,2X_i\psi_i = 0.
$$

Tímto máme odvozené vše pro následující větu, jen si pro lepší reprezentaci odvozené matice zaveďme značení $\lambda := -\frac{1-\pi}{\alpha_2}\beta_2$.

Věta 3. *Za předpokladů uvedených v této podkapitole má matice* $\boldsymbol{G} = -\,\mathsf{E}\,\frac{\mathrm{d}\boldsymbol{U_i}}{\mathrm{d}\boldsymbol{\gamma}}$, *kde*

$$\boldsymbol{U_i} = \begin{pmatrix} \xi_i \boldsymbol{D_i}\left(Y_i - \boldsymbol{D_i}^T\boldsymbol{\beta}\right) + (1-\xi_i)\left[\hat{\boldsymbol{D}}_{\boldsymbol{i}}\left(Y_i - \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}\right) + \begin{pmatrix}\mathbf{0_{(q+1),1}} \\ \frac{\beta_2}{\alpha_2^2}\sigma_\psi^2\end{pmatrix}\right] \\ \xi_i \boldsymbol{D_i}\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right) \\ \xi_i\left(\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right)^2 - \sigma_\psi^2\right) \end{pmatrix},$$

podobu

$$\begin{pmatrix}
1 & \mathsf{E}\,Z_{i,1} & \dots & \mathsf{E}\,Z_{i,q} & \mathsf{E}\,X_i & \lambda & \lambda\,\mathsf{E}\,Z_{i,1} & \dots & \lambda\,\mathsf{E}\,Z_{i,q} & \lambda\,\mathsf{E}\,X_i & 0 \\
\mathsf{E}\,\boldsymbol{Z_i} & \mathsf{E}\,Z_{i,1}\boldsymbol{Z_i} & \dots & \mathsf{E}\,Z_{i,q}\boldsymbol{Z_i} & \mathsf{E}\,X_i\boldsymbol{Z_i} & \lambda\,\mathsf{E}\,\boldsymbol{Z_i} & \lambda\,\mathsf{E}\,Z_{i,1}\boldsymbol{Z_i} & \dots & \lambda\,\mathsf{E}\,Z_{i,q}\boldsymbol{Z_i} & \lambda\,\mathsf{E}\,X_i\boldsymbol{Z_i} & \mathbf{0_{q,1}} \\
\mathsf{E}\,X_i & \mathsf{E}\,Z_{i,1}X_i & \dots & \mathsf{E}\,Z_{i,q}X_i & \mathsf{E}\,X_i^2 & \lambda\,\mathsf{E}\,X_i & \lambda\,\mathsf{E}\,Z_{i,1}X_i & \dots & \lambda\,\mathsf{E}\,Z_{i,q}X_i & \lambda\,\mathsf{E}\,X_i^2 & \frac{\lambda}{\alpha_2} \\
0 & 0 & \dots & 0 & 0 & \pi & \pi\,\mathsf{E}\,Z_{i,1} & \dots & \pi\,\mathsf{E}\,Z_{i,q} & \pi\,\mathsf{E}\,X_i & 0 \\
\mathbf{0_{q,1}} & \mathbf{0_{q,1}} & \dots & \mathbf{0_{q,1}} & \mathbf{0_{q,1}} & \pi\,\mathsf{E}\,\boldsymbol{Z_i} & \pi\,\mathsf{E}\,Z_{i,1}\boldsymbol{Z_i} & \dots & \pi\,\mathsf{E}\,Z_{i,q}\boldsymbol{Z_i} & \pi\,\mathsf{E}\,X_i\boldsymbol{Z_i} & 0 \\
0 & 0 & \dots & 0 & 0 & \pi\,\mathsf{E}\,X_i & \pi\,\mathsf{E}\,Z_{i,1}X_i & \dots & \pi\,\mathsf{E}\,Z_{i,q}X_i & \pi\,\mathsf{E}\,X_i^2 & 0 \\
0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \pi
\end{pmatrix}.$$

Důkaz. Podrobné odvození jednotlivých pozic matice $\boldsymbol{G}$ bylo již provedeno výše. □

Povšimněme si, že matici z Věty 3 lze zapsat do zjednodušené a praktičtější podoby pro další manipulování, což se bude hodit, protože jak bylo zmíněno ke konci předchozí sekce, potřebujeme inverzní matici k $\boldsymbol{G}$. Povšimněme si, že lze $\boldsymbol{G}$ rozdělit na 9 bloků, přičemž většina nenulových prvků je součástí 3 bloků. Tyto tři bloky se od sebe liší jen násobkem. Zavedeme-li značení

$$\boldsymbol{A} := \begin{pmatrix} 1 & \mathsf{E}\,Z_{i,1} & \dots & \mathsf{E}\,Z_{i,q} & \mathsf{E}\,X_i \\ \mathsf{E}\,\boldsymbol{Z_i} & \mathsf{E}\,Z_{i,1}\boldsymbol{Z_i} & \dots & \mathsf{E}\,Z_{i,q}\boldsymbol{Z_i} & \mathsf{E}\,X_i\boldsymbol{Z_i} \\ \mathsf{E}\,X_i & \mathsf{E}\,Z_{i,1}X_i & \dots & \mathsf{E}\,Z_{i,q}X_i & \mathsf{E}\,X_i^2 \end{pmatrix} = \mathsf{E}\,\boldsymbol{D_i}\boldsymbol{D_i}^T,$$

můžeme pomocí blokového zápisu jednodušeji a přehledněji pracovat s

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{A} & \lambda\boldsymbol{A} & \begin{pmatrix}\mathbf{0_{(q+1),1}} \\ \frac{\lambda}{\alpha_2}\end{pmatrix} \\ \mathbf{0_{(q+2),(q+2)}} & \pi\boldsymbol{A} & \mathbf{0_{(q+2),1}} \\ \mathbf{0_{1,(q+2)}} & \mathbf{0_{1,(q+2)}} & \pi \end{pmatrix}.$$

Při předpokladu regulárnosti matice $\boldsymbol{A}$ je ihned vidět regularita $\boldsymbol{G}$, neboť $0 < \pi < 1$. K ivertování této matice využijeme následující lemma o invertování horní blokově diagonální matice.

Lemma 4. *Nechť čtvercová matice $\boldsymbol{J}$ řádu m a čtvercová matice $\boldsymbol{L}$ řádu n jsou regulární a $\boldsymbol{K}$ je matice typu $m \times n$. Pak inverzní maticí k matici*

$$\begin{pmatrix} \boldsymbol{J} & \boldsymbol{K} \\ \boldsymbol{0}_{n,m} & \boldsymbol{L} \end{pmatrix}$$

je matice

$$\begin{pmatrix} \boldsymbol{J}^{-1} & -\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{L}^{-1} \\ \boldsymbol{0}_{n,m} & \boldsymbol{L}^{-1} \end{pmatrix}.$$

Důkaz. Důkaz je triviální, stačí vynásobit matice:

$$\begin{pmatrix} \boldsymbol{J} & \boldsymbol{K} \\ \boldsymbol{0}_{n,m} & \boldsymbol{L} \end{pmatrix} \begin{pmatrix} \boldsymbol{J}^{-1} & -\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{L}^{-1} \\ \boldsymbol{0}_{n,m} & \boldsymbol{L}^{-1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_m & -\boldsymbol{J}\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{L}^{-1} + \boldsymbol{K}\boldsymbol{L}^{-1} \\ \boldsymbol{0}_{n,m} & \boldsymbol{I}_n \end{pmatrix}.$$

□

Nyní už můžeme uvést dlouho kýžený výsledek této sekce a to tvar $\boldsymbol{G}^{-1}$.

Věta 5. *Za předpokladů shodných s předpoklady Věty 3 doplněných o regularitu matice $\boldsymbol{A}$ má inverzní matice k matici $\boldsymbol{G}$ tvar*

$$\boldsymbol{G}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} & \frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2}\boldsymbol{A}^{-1} & \boldsymbol{A}^{-1}\begin{pmatrix} \boldsymbol{0}_{(q+1),1} \\ \frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2^2} \end{pmatrix} \\ \boldsymbol{0}_{(q+2),(q+2)} & \frac{1}{\pi}\boldsymbol{A}^{-1} & \boldsymbol{0}_{(q+2),1} \\ \boldsymbol{0}_{1,(q+2)} & \boldsymbol{0}_{1,(q+2)} & \frac{1}{\pi} \end{pmatrix}.$$

Důkaz. K dokázání tvrzení dvakrát aplikujeme Lemma 4. Poprvé ho využijeme pro podmatici

$$\boldsymbol{B} := \begin{pmatrix} \boldsymbol{A} & \lambda\boldsymbol{A} \\ \boldsymbol{0}_{(q+2),(q+2)} & \pi\boldsymbol{A} \end{pmatrix}.$$

Dostáváme tak

$$\boldsymbol{B}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} & -\frac{\lambda}{\pi}\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^{-1} \\ \boldsymbol{0}_{(q+2),(q+2)} & \frac{1}{\pi}\boldsymbol{A}^{-1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}^{-1} & -\frac{\lambda}{\pi}\boldsymbol{A}^{-1} \\ \boldsymbol{0}_{(q+2),(q+2)} & \frac{1}{\pi}\boldsymbol{A}^{-1} \end{pmatrix}.$$

Nyní můžeme aplikovat Lemma 4 s dosazením $\boldsymbol{B}$ za $\boldsymbol{J}$. Spočtěme pravý horní blok v inverzní matici:

$$-\boldsymbol{B}^{-1}\begin{pmatrix} \boldsymbol{0}_{(q+1),1} \\ \frac{\lambda}{\alpha_2} \\ \boldsymbol{0}_{(q+2),1} \end{pmatrix}\frac{1}{\pi} = -\frac{1}{\pi}\begin{pmatrix} \boldsymbol{A}^{-1} \\ \boldsymbol{0}_{(q+2),(q+2)} \end{pmatrix}\begin{pmatrix} \boldsymbol{0}_{(q+1),1} \\ \frac{\lambda}{\alpha_2} \end{pmatrix} = -\begin{pmatrix} \boldsymbol{A}^{-1}\begin{pmatrix} \boldsymbol{0}_{(q+1),1} \\ \frac{\lambda}{\pi\alpha_2} \end{pmatrix} \\ \boldsymbol{0}_{(q+2),1} \end{pmatrix}.$$

Po dosazení $\lambda = -\frac{1-\pi}{\alpha_2}\beta_2$ dostáváme tvrzení věty.

□

4.3.3 Rozptylová matice

Máme odvozenou podobu inverzní matice $\boldsymbol{G}^{-1}$, zbývá odvození samotné rozptylové matice. Avšak oproti minulé sekci zesílíme předpoklad nekorelovanosti ϵ_i a ψ_i na předpoklad nezávislosti a přidáme tři předpoklady na momenty chyb:

$$\begin{aligned}\mathsf{E}[\epsilon_i^2 \mid \boldsymbol{Z_i}, X_i] &= \sigma_\epsilon^2,\\ \mathsf{E}[\psi_i^3 \mid \boldsymbol{Z_i}, X_i] &= \gamma_{\psi,3},\\ \mathsf{E}[\psi_i^4 \mid \boldsymbol{Z_i}, X_i] &= \gamma_{\psi,4},\end{aligned}$$

kde $\sigma_\epsilon^2 > 0, \gamma_{\psi,4} > 0, \gamma_{\psi,3} \in \mathbb{R}$.

Věta 6. *Za předpokladů této kapitoly má* $\Sigma = var(\boldsymbol{U_i})$*, kde*

$$\boldsymbol{U_i} = \begin{pmatrix} \xi_i \boldsymbol{D_i}\left(Y_i - \boldsymbol{D_i}^T\boldsymbol{\beta}\right) + (1-\xi_i)\left[\hat{\boldsymbol{D}}_{\boldsymbol{i}}\left(Y_i - \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}\right) + \begin{pmatrix}\mathbf{0_{(q+1),1}}\\ \frac{\beta_2}{\alpha_2^2}\sigma_\psi^2\end{pmatrix}\right] \\ \xi_i \boldsymbol{D_i}\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right) \\ \xi_i\left(\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right)^2 - \sigma_\psi^2\right)\end{pmatrix},$$

podobu

$$\Sigma = \begin{pmatrix} \boldsymbol{V_1} & \mathbf{0_{(q+2),(q+2)}} & \mathbf{0_{(q+2),1}} \\ \mathbf{0_{(q+2),(q+2)}} & \pi\sigma_\psi^2\boldsymbol{A} & \pi\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i} \\ \mathbf{0_{1,(q+2)}} & \pi\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i}^T & \pi\left(\gamma_{\psi,4} - \sigma_\psi^4\right)\end{pmatrix},$$

kde $\boldsymbol{V_1} = \sigma_\epsilon^2\boldsymbol{A} + \frac{1-\pi}{\alpha_2^2}\left[\beta_2^2\sigma_\psi^2\boldsymbol{A} + \boldsymbol{L_\psi} + \frac{\beta_2^2}{\alpha_2^2}diag\left(\sigma_\psi^4\right)\right]$ *pro* $\boldsymbol{L_\psi}$ *definované jako*

$$\begin{pmatrix} 0 & \mathbf{0_{1,q}} & \frac{\beta_2^2}{\alpha_2}\gamma_{\psi,3} \\ \mathbf{0_{q,1}} & \mathbf{0_{q,q}} & \frac{\beta_2^2}{\alpha_2}\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{Z_i} \\ \frac{\beta_2^2}{\alpha_2}\gamma_{\psi,3} & \frac{\beta_2^2}{\alpha_2}\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{Z_i}^T & \left[2\frac{\beta_2^2}{\alpha_2}\gamma_{\psi,3}\,\mathsf{E}\,X_i + \sigma_\epsilon^2\sigma_\psi^2 + \frac{\beta_2^2}{\alpha_2^2}\gamma_{\psi,4}\right]\end{pmatrix}.$$

Důkaz. Protože $\mathsf{E}\,\boldsymbol{U_i} = \mathbf{0_{(2q+5),1}}$, pak

$$\Sigma = \mathrm{var}(\boldsymbol{U_i}) = \mathsf{E}\,\boldsymbol{U_i}\boldsymbol{U_i}^T = \begin{pmatrix}\boldsymbol{V_1} & \boldsymbol{V_2} & \boldsymbol{V_3}\\ \boldsymbol{V_2}^T & \boldsymbol{V_4} & \boldsymbol{V_5}\\ \boldsymbol{V_3}^T & \boldsymbol{V_5}^T & V_6\end{pmatrix},$$

kde jednotlivé bloky odpovídají součinu daných částí $\boldsymbol{U_i}$, tedy bloky nahoře a uprostřed mají $q+2$ řádků, ve spodní řadě mají jen jeden řádek. Obdobně bloky vlevo a uprostřed mají po $2+q$ sloupcích, zatímco bloky vpravo mají jen jeden sloupec. Nejobtížnější bude spočítání $\boldsymbol{V_1}$, kterým začneme.

Výpočet výrazně zjednodušší fakt, že $\xi_i(1-\xi_i)=0$, a to platí vždy, neb ξ_i pochází z alternativního rozdělení. Dále využijeme výpočet z 2. kapitoly, konkrétně

$$\mathsf{E}\left[\hat{\boldsymbol{D}}_{\boldsymbol{i}}\left(Y_i-\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}\right)\right]=-\begin{pmatrix}\mathbf{0_{(q+1),1}}\\ \frac{\beta_2}{\alpha_2^2}\sigma_\psi^2\end{pmatrix},$$

respektive při přepsání matice druhých mocnin na rozptylovou matici hned v první rovnosti využijeme ekvivalentní zápis

$$\mathsf{E}\left[\hat{\boldsymbol{D}}_{\boldsymbol{i}}\left(Y_i-\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}\right)+\begin{pmatrix}\mathbf{0_{(q+1),1}}\\ \frac{\beta_2}{\alpha_2^2}\sigma_\psi^2\end{pmatrix}\right]=\mathbf{0_{(q+2),1}}.$$

Pak, s využitím značení z předchozí sekce $\boldsymbol{A}:=\mathsf{E}\,\boldsymbol{D_i}\boldsymbol{D_i}^T$, lze upravit:

$$\begin{aligned}
\boldsymbol{V_1}&=\mathsf{E}\,\xi_i^2\,\mathsf{E}\left[\left(Y_i-\boldsymbol{D}_{\boldsymbol{i}}^T\boldsymbol{\beta}\right)^2\boldsymbol{D_i}\boldsymbol{D_i}^T\right]+[\mathsf{E}(1-\xi_i)^2]\mathrm{var}\left(\hat{\boldsymbol{D}}_{\boldsymbol{i}}\left(Y_i-\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}\right)+\begin{pmatrix}\mathbf{0_{(q+1),1}}\\ \frac{\beta_2}{\alpha_2^2}\sigma_\psi^2\end{pmatrix}\right)\\
&=\mathsf{E}\,\xi_i\,\mathsf{E}\left[(\epsilon_i)^2\,\boldsymbol{D_i}\boldsymbol{D_i}^T\right]+[\mathsf{E}\,(1-\xi_i)]\mathrm{var}\left(\hat{\boldsymbol{D}}_{\boldsymbol{i}}\left(Y_i-\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}\right)\right)\\
&=\pi\,\mathsf{E}\left[\mathsf{E}[\epsilon_i^2\mid\boldsymbol{Z_i},X_i]\boldsymbol{D_i}\boldsymbol{D_i}^T\right]+(1-\pi)\,\mathrm{var}\left(\hat{\boldsymbol{D}}_{\boldsymbol{i}}\left(\beta_2X_i+\epsilon_i-\beta_2\hat{X}_i\right)\right)\\
&=\pi\sigma_\epsilon^2\,\mathsf{E}\left[\boldsymbol{D_i}\boldsymbol{D_i}^T\right]+(1-\pi)\,\mathrm{var}\left(\hat{\boldsymbol{D}}_{\boldsymbol{i}}\left(\beta_2X_i+\epsilon_i-\beta_2(X_i+\frac{\psi_i}{\alpha_2})\right)\right)\\
&=\pi\sigma_\epsilon^2\boldsymbol{A}+(1-\pi)\,\mathrm{var}\left(\left(\epsilon_i-\beta_2\frac{\psi_i}{\alpha_2}\right)\hat{\boldsymbol{D}}_{\boldsymbol{i}}\right).
\end{aligned}$$

Rozptylovou matici si vypočítáme zvlášť. Nejprve se bude hodit uvést následující úpravu

$$\hat{\boldsymbol{D}}_{\boldsymbol{i}}\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T=\begin{pmatrix}1&\boldsymbol{Z_i}^T&\hat{X}_i\\ \boldsymbol{Z_i}&\boldsymbol{Z_i}\boldsymbol{Z_i}^T&\hat{X}_i\boldsymbol{Z_i}\\ \hat{X}_i&\hat{X}_i\boldsymbol{Z_i}^T&\hat{X}_i^2\end{pmatrix}=\boldsymbol{D_i}\boldsymbol{D_i}^T+\frac{\psi_i}{\alpha_2}\begin{pmatrix}0&\mathbf{0_{1,q}}&1\\ \mathbf{0_{q,1}}&\mathbf{0_{q,q}}&\boldsymbol{Z_i}\\ 1&\boldsymbol{Z_i}^T&2X_i+\frac{\psi_i}{\alpha_2}\end{pmatrix}$$

a připomeňme si značení $\mathrm{diag}\left(\sigma_\psi^4\right)=\mathrm{diag}\left(0,\ldots,0,\sigma_\psi^4\right)$, kde bude v matici nenulová pouze poslední diagonální pozice s hodnotou σ_ψ^4 (rozměr matice uvádět ve značení nebudeme, v kontextu výpočtu je zřejmý). Počítejme dále:

$$\begin{aligned}
&\mathrm{var}\left(\left(\epsilon_i-\beta_2\frac{\psi_i}{\alpha_2}\right)\hat{\boldsymbol{D}}_{\boldsymbol{i}}\right)=\\
&=\mathsf{E}\left(\epsilon_i-\beta_2\frac{\psi_i}{\alpha_2}\right)^2\hat{\boldsymbol{D}}_{\boldsymbol{i}}\hat{\boldsymbol{D}}_{\boldsymbol{i}}^T-\begin{pmatrix}\mathbf{0_{(q+1),1}}\\ \frac{\beta_2}{\alpha_2^2}\sigma_\psi^2\end{pmatrix}\begin{pmatrix}\mathbf{0_{1,(q+1)}}&\frac{\beta_2}{\alpha_2^2}\sigma_\psi^2\end{pmatrix}=\\
&=\mathsf{E}\left(\epsilon_i^2-2\epsilon_i\psi_i\frac{\beta_2}{\alpha_2}+\psi_i^2\frac{\beta_2^2}{\alpha_2^2}\right)\left[\boldsymbol{D_i}\boldsymbol{D_i}^T+\frac{\psi_i}{\alpha_2}\begin{pmatrix}0&\mathbf{0_{1,q}}&1\\ \mathbf{0_{q,1}}&\mathbf{0_{q,q}}&\boldsymbol{Z_i}\\ 1&\boldsymbol{Z_i}^T&2X_i+\frac{\psi_i}{\alpha_2}\end{pmatrix}\right]-\frac{\beta_2^2}{\alpha_2^4}\mathrm{diag}\left(\sigma_\psi^4\right).
\end{aligned}$$

Vypočítejme střední hodnotu $\psi_i^k \left(\epsilon_i^2 - 2\epsilon_i\psi_i\frac{\beta_2}{\alpha_2} + \psi_i^2\frac{\beta_2^2}{\alpha_2^2}\right)$, kde $k = 0, 1, 2$, využívajíce vlastností podmíněné střední hodnoty a nezávislosti ϵ_i a ψ_i:

$$\begin{aligned}
\mathsf{E}\left[\mathsf{E}\left[\epsilon_i^2 - 2\epsilon_i\psi_i\frac{\beta_2}{\alpha_2} + \psi_i^2\frac{\beta_2^2}{\alpha_2^2} \mid \boldsymbol{Z_i}, X_i\right]\right] &= \sigma_\epsilon^2 + \frac{\beta_2^2}{\alpha_2^2}\sigma_\psi^2,\\
\mathsf{E}\left[\mathsf{E}\left[\psi_i\left(\epsilon_i^2 - 2\epsilon_i\psi_i\frac{\beta_2}{\alpha_2} + \psi_i^2\frac{\beta_2^2}{\alpha_2^2}\right) \mid \boldsymbol{Z_i}, X_i\right]\right] &= \frac{\beta_2^2}{\alpha_2^2}\gamma_{\psi,3},\\
\mathsf{E}\left[\mathsf{E}\left[\psi_i^2\left(\epsilon_i^2 - 2\epsilon_i\psi_i\frac{\beta_2}{\alpha_2} + \psi_i^2\frac{\beta_2^2}{\alpha_2^2}\right) \mid \boldsymbol{Z_i}, X_i\right]\right] &= \sigma_\epsilon^2\sigma_\psi^2 + \frac{\beta_2^2}{\alpha_2^2}\gamma_{\psi,4}.
\end{aligned}$$

Pak u předchozího výpočtu dostáváme

$$\mathsf{E}\left(\epsilon_i^2 - 2\epsilon_i\psi_i\frac{\beta_2}{\alpha_2} + \psi_i^2\frac{\beta_2^2}{\alpha_2^2}\right)\boldsymbol{D_i}\boldsymbol{D_i}^T = \left(\sigma_\epsilon^2 + \frac{\beta_2^2}{\alpha_2^2}\sigma_\psi^2\right)\boldsymbol{A},$$

$$\begin{aligned}
&\mathsf{E}\left(\epsilon_i^2 - 2\epsilon_i\psi_i\frac{\beta_2}{\alpha_2} + \psi_i^2\frac{\beta_2^2}{\alpha_2^2}\right)\frac{\psi_i}{\alpha_2}\begin{pmatrix} 0 & \mathbf{0_{1,q}} & 1\\ \mathbf{0_{q,1}} & \mathbf{0_{q,q}} & \boldsymbol{Z_i}\\ 1 & \boldsymbol{Z_i}^T & 2X_i + \frac{\psi_i}{\alpha_2}\end{pmatrix} =\\
&= \frac{1}{\alpha_2}\begin{pmatrix} 0 & \mathbf{0_{1,q}} & \frac{\beta_2^2}{\alpha_2^2}\gamma_{\psi,3}\\ \mathbf{0_{q,1}} & \mathbf{0_{q,q}} & \frac{\beta_2^2}{\alpha_2^2}\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{Z_i}\\ \frac{\beta_2^2}{\alpha_2^2}\gamma_{\psi,3} & \frac{\beta_2^2}{\alpha_2^2}\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{Z_i}^T & \left[2\frac{\beta_2^2}{\alpha_2^2}\gamma_{\psi,3}\,\mathsf{E}\,X_i + \frac{1}{\alpha_2}\left(\sigma_\epsilon^2\sigma_\psi^2 + \frac{\beta_2^2}{\alpha_2^2}\gamma_{\psi,4}\right)\right]\end{pmatrix} =\\
&= \frac{1}{\alpha_2^2}\begin{pmatrix} 0 & \mathbf{0_{1,q}} & \frac{\beta_2^2}{\alpha_2}\gamma_{\psi,3}\\ \mathbf{0_{q,1}} & \mathbf{0_{q,q}} & \frac{\beta_2^2}{\alpha_2}\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{Z_i}\\ \frac{\beta_2^2}{\alpha_2}\gamma_{\psi,3} & \frac{\beta_2^2}{\alpha_2}\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{Z_i}^T & \left[2\frac{\beta_2^2}{\alpha_2}\gamma_{\psi,3}\,\mathsf{E}\,X_i + \sigma_\epsilon^2\sigma_\psi^2 + \frac{\beta_2^2}{\alpha_2^2}\gamma_{\psi,4}\right]\end{pmatrix} =: \frac{1}{\alpha_2^2}\boldsymbol{L_\psi}.
\end{aligned}$$

Celkově tak platí

$$\begin{aligned}
\boldsymbol{V_1} &= \pi\sigma_\epsilon^2\boldsymbol{A} + (1-\pi)\,\mathrm{var}\left(\left(\epsilon_i - \beta_2\frac{\psi_i}{\alpha_2}\right)\hat{\boldsymbol{D}}_{\boldsymbol{i}}\right) =\\
&= \pi\sigma_\epsilon^2\boldsymbol{A} + (1-\pi)\left[\left(\sigma_\epsilon^2 + \frac{\beta_2^2}{\alpha_2^2}\sigma_\psi^2\right)\boldsymbol{A} + \frac{1}{\alpha_2^2}\boldsymbol{L_\psi} - \frac{\beta_2^2}{\alpha_2^4}\mathrm{diag}\left(\sigma_\psi^4\right)\right]\\
&= \sigma_\epsilon^2\boldsymbol{A} + \frac{1-\pi}{\alpha_2^2}\left[\beta_2^2\sigma_\psi^2\boldsymbol{A} + \boldsymbol{L_\psi} - \frac{\beta_2^2}{\alpha_2^2}\mathrm{diag}\left(\sigma_\psi^4\right)\right].
\end{aligned}$$

Tímto máme vyjádřen nejsložitější blok matice Σ a to z hlediska nejen samotného výpočtu, ale i samotného zápisu. Ostatní výpočty jsou už triviální:

$$\boldsymbol{V_2} = \mathsf{E}\,\xi_i^2\,\mathsf{E}\left(Y_i - \boldsymbol{D_i}^T\boldsymbol{\beta}\right)\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right)\boldsymbol{D_i}\boldsymbol{D_i}^T = \pi\,\mathsf{E}\,\epsilon_i\psi_i\boldsymbol{D_i}\boldsymbol{D_i}^T = \boldsymbol{0_{(q+2),(q+2)}},$$

$$\boldsymbol{V_3} = \mathsf{E}\,\xi_i^2\,\mathsf{E}\,\epsilon_i\left(\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right)^2 - \sigma_\psi^2\right)\boldsymbol{D_i} = \pi\,\mathsf{E}\,\epsilon_i\left(\psi_i^2 - \sigma_\psi^2\right) = \boldsymbol{0_{(q+2),1}},$$

$$\boldsymbol{V_4} = \mathsf{E}\,\xi_i^2\,\mathsf{E}\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right)^2\boldsymbol{D_i}\boldsymbol{D_i}^T = \pi\,\mathsf{E}\,\psi_i^2\boldsymbol{D_i}\boldsymbol{D_i}^T = \pi\sigma_\psi^2\boldsymbol{A},$$

$$\begin{aligned}\boldsymbol{V_5} &= \mathsf{E}\,\xi_i^2\,\mathsf{E}\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right)\left(\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right)^2 - \sigma_\psi^2\right)\boldsymbol{D_i} = \\ &= \pi\,\mathsf{E}\,\psi_i\left(\psi_i^2 - \sigma_\psi^2\right)\boldsymbol{D_i} = \pi\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i},\end{aligned}$$

$$\begin{aligned}\boldsymbol{V_6} &= \mathsf{E}\,\xi_i^2\,\mathsf{E}\left(\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right)^2 - \sigma_\psi^2\right)^2 = \pi\,\mathsf{E}\left(\psi_i^2 - \sigma_\psi^2\right)^2 = \\ &= \pi\,\mathsf{E}\left(\psi_i^4 - 2\psi_i^2\sigma_\psi^2 + \sigma_\psi^4\right) = \pi\left(\gamma_{\psi,4} - 2\sigma_\psi^2\sigma_\psi^2 + \sigma_\psi^4\right) = \pi\left(\gamma_{\psi,4} - \sigma_\psi^4\right).\end{aligned}$$

□

Máme odvozenou pomocí Vět 5 a 6 inverzní matici $\boldsymbol{G^{-1}}$ i rozptylovou matici Σ, tedy můžeme konečně vyjádřit asymptotickou rozptylovou matici $\boldsymbol{\Sigma_\gamma} := \boldsymbol{G^{-1}}\boldsymbol{\Sigma}(\boldsymbol{G^{-1}})^T$ pro samotný odhad $\hat{\boldsymbol{\gamma}}$. Pro příjemnější násobení připomeneme pod sebou obě matice:

$$\Sigma = \begin{pmatrix} \boldsymbol{V_1} & \boldsymbol{0_{(q+2),(q+2)}} & \boldsymbol{0_{(q+2),1}} \\ \boldsymbol{0_{(q+2),(q+2)}} & \pi\sigma_\psi^2\boldsymbol{A} & \pi\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i} \\ \boldsymbol{0_{1,(q+2)}} & \pi\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i}^T & \pi\left(\gamma_{\psi,4} - \sigma_\psi^4\right) \end{pmatrix},$$

$$\boldsymbol{G}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} & \frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2}\boldsymbol{A}^{-1} & \boldsymbol{A}^{-1}\begin{pmatrix}\boldsymbol{0_{(q+1),1}} \\ \frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2^2}\end{pmatrix} \\ \boldsymbol{0_{(q+2),(q+2)}} & \frac{1}{\pi}\boldsymbol{A}^{-1} & \boldsymbol{0_{(q+2),1}} \\ \boldsymbol{0_{1,(q+2)}} & \boldsymbol{0_{1,(q+2)}} & \frac{1}{\pi} \end{pmatrix}.$$

Nejprve vypočteme $\boldsymbol{\Sigma}(\boldsymbol{G^{-1}})^T$, což má tvar

$$\begin{pmatrix} \boldsymbol{V_1}\boldsymbol{A}^{-1} & \boldsymbol{0_{(q+2),(q+2)}} & \boldsymbol{0_{(q+2),1}} \\ \boldsymbol{P_1} & \sigma_\psi^2\boldsymbol{I_{q+2}} & \gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i} \\ \boldsymbol{P_2} & \gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i}^T\boldsymbol{A}^{-1} & \left(\gamma_{\psi,4} - \sigma_\psi^4\right) \end{pmatrix},$$

kde $\boldsymbol{P_1}$ a $\boldsymbol{P_2}$ vyjádříme zvlášť:

$$\begin{aligned}
\boldsymbol{P_1} &= \left[\pi\sigma_\psi^2\boldsymbol{A}\right]\left[\frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2}\boldsymbol{A}^{-1}\right] + \left[\pi\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i}\right]\left[\left(\mathbf{0_{1,(q+1)}} \quad \tfrac{1-\pi}{\pi}\tfrac{\beta_2}{\alpha_2^2}\right)\boldsymbol{A}^{-1}\right] = \\
&= \left[(1-\pi)\frac{\beta_2}{\alpha_2}\sigma_\psi^2\boldsymbol{I_{q+2}}\right] + \left[\gamma_{\psi,3}(1-\pi)\frac{\beta_2}{\alpha_2^2}\left(\mathbf{0_{(q+2),(q+1)}} \quad \mathsf{E}\,\boldsymbol{D_i}\right)\boldsymbol{A}^{-1}\right] \\
&= (1-\pi)\frac{\beta_2}{\alpha_2}\left[\sigma_\psi^2\boldsymbol{I_{q+2}} + \frac{\gamma_{\psi,3}}{\alpha_2}\left(\mathbf{0_{(q+2),(q+1)}} \quad \mathsf{E}\,\boldsymbol{D_i}\right)\boldsymbol{A}^{-1}\right],
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{P_2} &= \left[\pi\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i^T}\right]\left[\frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2}\boldsymbol{A}^{-1}\right] + \left[\pi\left(\gamma_{\psi,4}-\sigma_\psi^4\right)\right]\left[\left(\mathbf{0_{1,(q+1)}} \quad \tfrac{1-\pi}{\pi}\tfrac{\beta_2}{\alpha_2^2}\right)\boldsymbol{A}^{-1}\right] \\
&= (1-\pi)\frac{\beta_2}{\alpha_2}\left[\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i^T} + \left(\mathbf{0_{1,(q+1)}} \quad \tfrac{\left(\gamma_{\psi,4}-\sigma_\psi^4\right)}{\alpha_2}\right)\right]\boldsymbol{A}^{-1}.
\end{aligned}$$

A na závěr výpočet $\boldsymbol{\Sigma_\gamma} = \boldsymbol{G^{-1}\Sigma(G^{-1})}^T$, kde už většinu bloků budeme muset z prostorových důvodů vyjádřit zvlášť:

$$\begin{pmatrix}
\boldsymbol{R_1} & \boldsymbol{R_2} & \boldsymbol{R_3} \\
\boldsymbol{R_4} & \frac{1}{\pi}\sigma_\psi^2\boldsymbol{A}^{-1} & \frac{1}{\pi}\gamma_{\psi,3}\boldsymbol{A}^{-1}\,\mathsf{E}\,\boldsymbol{D_i} \\
\boldsymbol{R_5} & \frac{1}{\pi}\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i^T}\boldsymbol{A}^{-1} & \frac{1}{\pi}\left(\gamma_{\psi,4}-\sigma_\psi^4\right)
\end{pmatrix},$$

kde jednotlivé bloky jsou:

$$\begin{aligned}
\boldsymbol{R_1} &= \boldsymbol{A^{-1}V_1A^{-1}} + \frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2}\boldsymbol{A}^{-1}\boldsymbol{P_1} + \boldsymbol{A}^{-1}\begin{pmatrix}\mathbf{0_{(q+1),1}} \\ \frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2^2}\end{pmatrix}\boldsymbol{P_2} = \\
&= \boldsymbol{A^{-1}V_1A^{-1}} + \frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2}\boldsymbol{A}^{-1}\left[\boldsymbol{P_1} + \begin{pmatrix}\mathbf{0_{(q+1),1}} \\ \frac{1}{\alpha_2}\end{pmatrix}\boldsymbol{P_2}\right] = \\
&= \boldsymbol{A^{-1}V_1A^{-1}} + \frac{(1-\pi)^2}{\pi}\frac{\beta_2^2}{\alpha_2^2}\boldsymbol{A}^{-1}\Big[\sigma_\psi^2\boldsymbol{I_{q+2}} + \frac{\gamma_{\psi,3}}{\alpha_2}\left(\mathbf{0_{(q+2),(q+1)}} \quad \mathsf{E}\,\boldsymbol{D_i}\right)\boldsymbol{A}^{-1} + \\
&\quad + \begin{pmatrix}\mathbf{0_{(q+1),1}} \\ \frac{1}{\alpha_2}\end{pmatrix}\left[\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i^T} + \left(\mathbf{0_{1,(q+1)}} \quad \tfrac{\left(\gamma_{\psi,4}-\sigma_\psi^4\right)}{\alpha_2}\right)\right]\boldsymbol{A}^{-1}\Big] = \\
&= \boldsymbol{A^{-1}V_1A^{-1}} + \frac{(1-\pi)^2}{\pi}\frac{\beta_2^2}{\alpha_2^3}\boldsymbol{A}^{-1}\Big[\alpha_2\sigma_\psi^2\boldsymbol{I_{q+2}} + \Big[\gamma_{\psi,3}\left(\mathbf{0_{(q+2),(q+1)}} \quad \mathsf{E}\,\boldsymbol{D_i}\right) + \\
&+ \gamma_{\psi,3}\begin{pmatrix}\mathbf{0_{(q+1),(q+2)}} \\ \mathsf{E}\,\boldsymbol{D_i^T}\end{pmatrix} + \mathrm{diag}\left(\frac{\left(\gamma_{\psi,4}-\sigma_\psi^4\right)}{\alpha_2}\right)\Big]\boldsymbol{A}^{-1}\Big],
\end{aligned}$$

$$\boldsymbol{R_2} = \frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2}\left[\sigma_\psi^2\boldsymbol{I_{q+2}} + \frac{\gamma_{\psi,3}}{\alpha_2}\boldsymbol{A}^{-1}\begin{pmatrix}\mathbf{0_{(q+1),(q+2)}} \\ \mathsf{E}\,\boldsymbol{D_i^T}\end{pmatrix}\right]\boldsymbol{A}^{-1},$$

$$\boldsymbol{R_3} = \frac{1-\pi}{\pi}\frac{\beta_2}{\alpha_2}\boldsymbol{A}^{-1}\left[\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i} + \left(\gamma_{\psi,4}-\sigma_\psi^4\right)\begin{pmatrix}\mathbf{0_{(q+1),1}} \\ \frac{1}{\alpha_2}\end{pmatrix}\right],$$

$$\boldsymbol{R_4} = \frac{1}{\pi}\boldsymbol{A}^{-1}\boldsymbol{P_1} = \frac{(1-\pi)}{\pi}\frac{\beta_2}{\alpha_2}\boldsymbol{A}^{-1}\left[\sigma_\psi^2\boldsymbol{I_{q+2}} + \frac{\gamma_{\psi,3}}{\alpha_2}\left(\boldsymbol{0_{(q+2),(q+1)}} \quad \mathsf{E}\,\boldsymbol{D_i}\right)\boldsymbol{A}^{-1}\right] = \boldsymbol{R_2}^T,$$

$$\boldsymbol{R_5} = \frac{1}{\pi}\boldsymbol{P_2} = \frac{(1-\pi)}{\pi}\frac{\beta_2}{\alpha_2}\left[\gamma_{\psi,3}\,\mathsf{E}\,\boldsymbol{D_i}^T + \left(\boldsymbol{0_{1,(q+1)}} \quad \frac{\left(\gamma_{\psi,4}-\sigma_\psi^4\right)}{\alpha_2}\right)\right]\boldsymbol{A}^{-1} = \boldsymbol{R_3}^T.$$

I když matice $\boldsymbol{\Sigma_\gamma}$ je na první pohled vizuálně velice nepříjemnou maticí, jejíž bloky jsme museli zčásti určit odděleně kvůli prostorovým důvodům, jedná se stále o velice důležitou matici, neboť je to asymptotická rozptylová matice sdruženého vektoru odhadů alf, bet a σ_ψ^2, tedy odhadu $\hat{\boldsymbol{\gamma}}$! Výsledek této kapitoly zformulujme do věty, budeme úzce s tím pracovat v další kapitole.

Věta 7. *Za předpokladů této kapitoly jsou odhady ze soustavy odhadovacích rovnic* $\sum_{i=1}^n \boldsymbol{U_i} = \boldsymbol{0_{(2q+5),1}}$, *kde*

$$\boldsymbol{U_i} = \begin{pmatrix} \xi_i\boldsymbol{D_i}\left(Y_i - \boldsymbol{D_i}^T\boldsymbol{\beta}\right) + (1-\xi_i)\left[\hat{\boldsymbol{D}}_{\boldsymbol{i}}\left(Y_i - \hat{\boldsymbol{D}}_{\boldsymbol{i}}^T\boldsymbol{\beta}\right) + \begin{pmatrix}\boldsymbol{0_{(q+1),1}} \\ \frac{\beta_2}{\alpha_2^2}\sigma_\psi^2\end{pmatrix}\right] \\ \xi_i\boldsymbol{D_i}\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right) \\ \xi_i\left(\left(W_i - \boldsymbol{D_i}^T\boldsymbol{\alpha}\right)^2 - \sigma_\psi^2\right) \end{pmatrix},$$

konzistentními odhady skutečných parametrů a asymptotická rozptylová matice vektoru odhadů $\hat{\boldsymbol{\gamma}}$ *je* $\boldsymbol{\Sigma_\gamma}$ *(explicitní tvar uveden výše). Platí toto sdružené asymptotické rozdělení:*

$$\sqrt{n}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right) \xrightarrow{d} \mathcal{N}_{2q+5}\left(\boldsymbol{0_{2q+5,1}}, \boldsymbol{\Sigma_\gamma}\right).$$

Důkaz. Důkaz neuvádíme, lze jej provést ověřením podmínek pro asymptotickou normalitu Z-odhadů, viz [11] nebo [12]. Tvar asymptotické rozptylové matice $\boldsymbol{\Sigma_\gamma}$ byl odvozen před Větou s využitím Vět 5 a 6. □

Odhadnout $\boldsymbol{\Sigma_\gamma}$ lze buď přímo podle explicitního vyjádření, nebo pomocí vztahu $\boldsymbol{\Sigma_\gamma} = \boldsymbol{G^{-1}\Sigma(G^{-1})}^T$, kde odhadnout jednotlivé matice je výpočetně o něco jednodušší. Pro matici $\boldsymbol{G^{-1}}$ je potřeba odhadnout $\boldsymbol{A^{-1}}$ a π (parametry v $\boldsymbol{\gamma}$ se odhadnou pomocí odpovídajících složek v $\hat{\boldsymbol{\gamma}}$). Odhadnutí π se udělá přímočaře pomocí celkového rozsahu a velikosti validační skupiny

$$\hat{\pi} = \frac{\sum_{i=1}^n \xi_i}{n},$$

zatímco pro odhadnutí $\boldsymbol{A^{-1}}$ lze využít

$$\widehat{\boldsymbol{A^{-1}}} \cong \hat{\boldsymbol{A}}^{\boldsymbol{-1}},$$

kde na odhadnutí $\boldsymbol{A} = \mathsf{E}\,\boldsymbol{D_i}\boldsymbol{D_i}^T$ se použije výběrová kovarianční matice. Obdobně, nyní jen s využitím validační skupiny, u matice Σ použijeme pro odhadnutí $\mathsf{E}\,\boldsymbol{D_i}$ výběrový průměr $\boldsymbol{D_i}$, průměrováním

$$\left(W_i - \boldsymbol{D_i^T}\hat{\boldsymbol{\alpha}}\right)^l$$

získáme pro $l = 3$ odhad třetího momentu ψ_i, tedy $\gamma_{\psi,3}$, respektive pro $l = 4$ odhad čtvrtého momentu $\gamma_{\psi,4}$. Dále σ_ϵ^2 získáme z validační skupiny zprůměrováním

$$\left(Y_i - \boldsymbol{D_i^T}\hat{\boldsymbol{\beta}}\right)^2.$$

Díky Větě 7 nemáme oproti klasickým článkům jen ukázánou korekci a o kolik se zlepší odhad, ale současně dostáváme odhady parametrů v chybovém modelu a přes asymptotickou rozptylovou matici i provázanost vůči parametrům hlavního modelu. Díky explicitnímu vyjádření můžeme s výsledky plně pracovat v simulačních studiích a zkoumat jak moc dobré aproximace dosahujeme i v situacích, kdy máme malé rozsahy, případně malé pravděpodobnosti π. Pomocí Cramérovy-Woldovy věty lze získat asymptotiku pro jednotlivé parametry a s využitím Cramérovy-Slutského věty lze získat i intervaly spolehlivosti pro jednotlivé parametry. V simulační části tedy lze zkoumat i dodržování hladiny pokrytí. Je dobré si povšimnout, že v asymptotické rozptylové matici jsou prvky této matice dělené π, tedy čím menší pravěpodobnost, že dané pozorování bude pocházet z validační skupiny, tím zvětšující se rozptyl daného odhadu. Jak moc se odhady zhorší si ukážeme v následující kapitole, která je věnována simulačním studiím.

5 Simulační část

V této kapitole nepřebíráme žádné výsledky z jiného zdroje, pouze využíváme poznatků ze 4. kapitoly, které podrobujeme simulační studii a vše uvedené je vlastní prací autora.

5.1 Úvod

V předchozí kapitole jsme si odvodili soustavu odhadovacích rovnic a ukázali jsme (mimo alf, pro které je to standardní odhad metodou nejmenších čtverců) explicitní vyjádření pro jednotlivé parametry. V podkapitole 4.3 jsme pak odvodili a explicitně vyjádřili asymptotické rozdělení pro $\hat{\boldsymbol{\gamma}}$ odhadující $\boldsymbol{\gamma}$, kde

$$\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \\ \sigma_\psi^2 \end{pmatrix} \text{ pro } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_1} \\ \beta_2 \end{pmatrix} \text{ a } \boldsymbol{\alpha} = \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha_1} \\ \alpha_2 \end{pmatrix}.$$

I když je asymptotické rozdělení velmi podrobně odvozené, je samozřejmě relevantní otázkou nakolik je tato asymptotika užitečná pro praktické využití. V reálném světě se můžeme setkat s několika případy, ve kterých může nastat otázka, zda lze asymptotiku ještě využít, či už je natolik špatná, aby se od využití této asymptotiky v dané situaci raději upustilo. Budeme proto zkoumat několik situací a všímat si při jakých hodnotách stále dostáváme nějakým způsobem rozumné výsledky.

Jedna z běžných situací je případ, kdy máme rozsah výběru příliš malý (například kvůli tomu, že onemocnění je vzácné a lze mít kvůli tomu jen malé množství lidí ve studii) a tak se tento problém snažíme „zachraňovat“ dostatečně velkým podílem validační skupiny, například my budeme v podkapitole 5.3 pracovat s hodnotou $\pi = 0.5$, a zkoumat, zda pro rozsah o nižších stovek či dokonce vyšších desítek pozorování bude asymptotika stále dávat použitelné odhady. Jinou situací je problém, kdy i když můžeme mít rozsah výrazně větší, tak naopak je z nějakého závažného důvodu velkým problémem poskytnout dostatečnému množství pozorování přesné měření regresoru, respektive budeme mít malý podíl validační skupiny. Abychom mohli zkoumat opravdu malé hodnoty podílu validační skupiny, kde π může být v řádu setin či vyšších jednotkách tisícin, budeme v podkapitole 5.4 uvažovat dostatečně velký rozsah, a to $n = 10\ 000$.

Dále v kapitole 5.5 budeme zkoumat testování nulové hypotézy. Podíváme se, zda model bude správně přisuzovat prakticky nulový vliv X_i, pokud daný regresor bude skutečně roven nule, či zda díky korelovanosti regresorů bude chybně přisuzovat větší vliv X_i. Také budeme zkoumat podobnou situaci pro zjednodušenou situaci, kdy chyby ψ_i budou nestranné a budeme porovnávat naše odhady s odhady získanými standardní metodou nejmenších čtverců. Vše v této kapitole bude prováděno na hladině 0.05, respektive budeme zkoumat 95% intervaly spolehlivosti.

5.2 Simulace dat a metody

Abychom zkoumali, jak moc je dobrá naše odvozená asymptotika v praxi, je potřeba provést několik simulací. V této podkapitole popíšeme jakou metodou jsme simulovali data pro naše simulace. Primární popis bude věnován především podkapitolám 5.3 a 5.4 vyjma nastavení π a velikosti rozsahu n. Na odlišnosti v nastavení parametrů pro podkapitolu testování hypotéz upozorníme na příslušných místech v dané podkapitolu, avšak generování $\boldsymbol{Z_i}$ a X_i probíhá naprosto shodně s popisem v této podkapitole. Pro jednoduchost místo dvojpozičního indexování, což bylo doposud u parametrů pro $\boldsymbol{Z_i}$, který nyní bude trojsložkový, přejdeme u alf a bet k indexování od 0 (pro absolutní člen) po 4 (pro parametr příslušný k X_i). Náš hlavní model bude tak mít podobu

$$Y_i = \beta_0 + \beta_1 Z_{i,1} + \beta_2 Z_{i,2} + \beta_3 Z_{i,3} + \beta_4 X_i + \epsilon_i,$$

kde $\mathsf{E}[\epsilon_i \mid \boldsymbol{Z_i}, X_i] = 0$, $\mathrm{var}(\epsilon_i \mid \boldsymbol{Z_i}, X_i) = \sigma_\epsilon^2 > 0$, kde $\boldsymbol{Z_i}$ má 3 složky a $i = 1, \ldots, n, n \in \mathbb{N}$ (typicky $n = 1\ 000$), zatímco náš chybový model je

$$W_i = \alpha_0 + \alpha_1 Z_{i,1} + \alpha_2 Z_{i,2} + \alpha_3 Z_{i,3} + \alpha_4 X_i + \psi_i,$$

kde $\mathsf{E}[\psi_i \mid \boldsymbol{Z_i}, X_i] = 0$, $\mathrm{var}(\psi_i \mid \boldsymbol{Z_i}, X_i) = \sigma_\psi^2 > 0$ a $i = 1, \ldots, n$.

Samotné regresory mezi sebou nemusí být nezávislé, naopak v praxi často spolu dosti souvisí a my jsme v celé práci nikde nepožadovali jejich nezávislost nebo nekorelovanost, proto při generování dat vložíme mezi regresory závislosti. Jediná nezávislost mezi čímkoliv jiným je pro ξ_i, které generujeme z alternativního rozdělení s pravděpodobností π (například u testování hypotéz používáme hodnotu 0.25), že bude patřit do validační skupiny, jinak bude v nevalidační skupině. Nezávisle na této veličině generujeme postupně vše ostatní, začneme $Z_{i,1}$, které pochází z rovnoměrného rozdělení na intervalu $(0, 1)$.

Na základě hodnoty $Z_{i,1}$ se přímo ovlivní pravděpodobnost napozorování hodnoty 1 u $Z_{i,2}$, které pochází z alternativního rozdělení. Konkrétní předpis je

$$Z_{i,2} \sim \mathrm{Alt}\left(0.3 + \frac{2 \cdot Z_{i,1}}{5}\right),$$

tedy pravděpodobnost napozorování hodnoty 1, respektive 0 pro $Z_{i,2}$ leží v intervalu $(0.3, 0.7)$. Poslední regresor, který je pro validační i nevalidační skupinu naměřen zcela přesně, je $Z_{i,3}$, které vznikne pomocí $Z_{i,1}$ a $Z_{i,2}$ prostřednictvím normálního rozdělení následujícím způsobem:

$$Z_{i,3} \sim \mathcal{N}(25 + 15Z_{i,1} - 2Z_{i,2},\ 3).$$

Obdobně modelujeme i X_i, které s $Z_{i,3}$ závisí jen nepřímo a to prostřednictvím předchozích regresorů, konkrétně

$$X_i \sim \mathcal{N}(40 - 28Z_{i,1} + 4Z_{i,2},\ 10).$$

Hodnotu X_i generujeme i pro nevalidační skupinu, protože na základě této hodnoty se bude generovat odhad W_i a odezva Y_i, a to podle výše zmíněných

modelů. Pro podkapitoly 5.3 a 5.4 budou skutečné hodnoty daných parametrů nastaveny následovně:

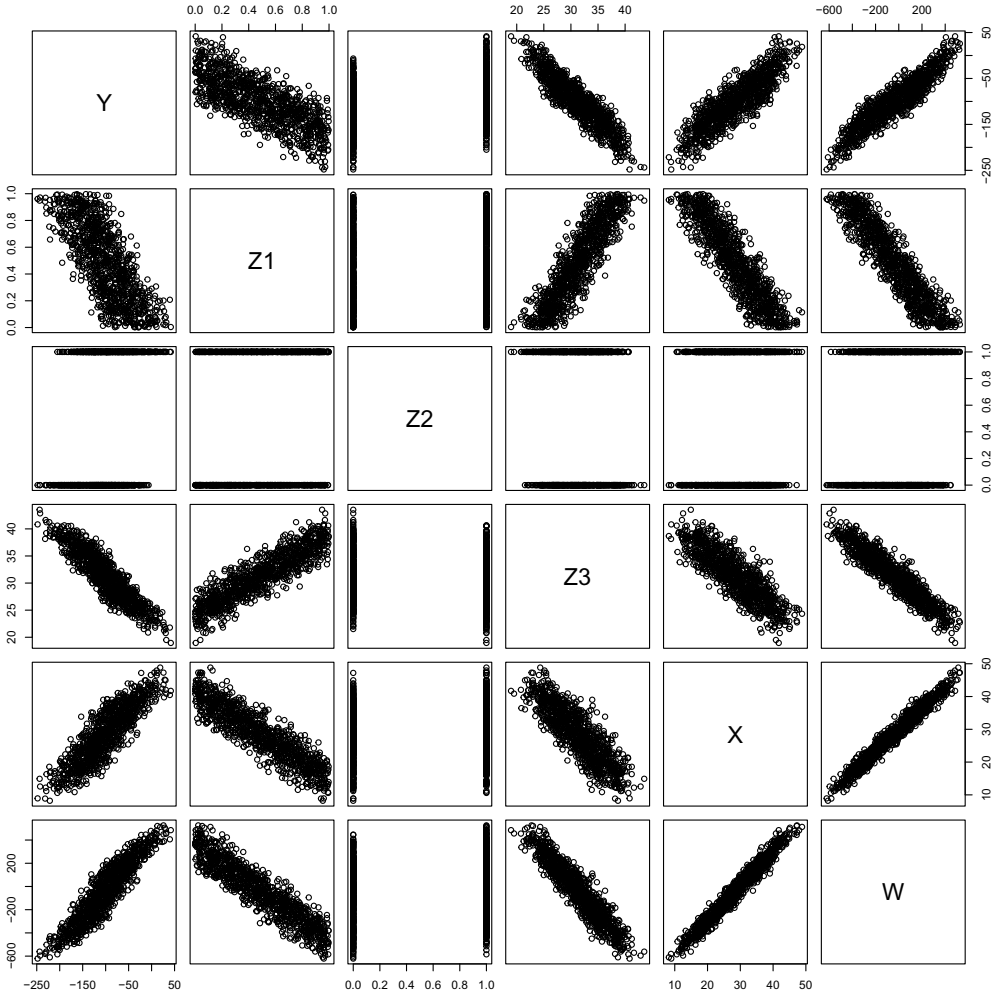
$$\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix} = \begin{pmatrix}40\\ 100\\ 20\\ -10\\ 4\end{pmatrix}, \begin{pmatrix}\alpha_0\\ \alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4\end{pmatrix} = \begin{pmatrix}12\\ -6\\ 15\\ -20\\ 20\end{pmatrix}, \sigma_\epsilon^2 = 64 \text{ a } \sigma_\psi^2 = 49.$$

I když generování dat je plně popsáno, hodí se pro lepší představu o datech přidat vizualizaci dat. Na obrázku 5.1 vidíme scatter ploty mezi jednotlivými regresory, špatně naměřenou W_i a Y_i. Až na scatter ploty s $Z2$, zjednodušeným značením pro $Z_{i,2}$, která je binární (většina hodnot se pro 0 a 1 překrývají, i když si lze všimnout, že hodnoty v 0 jsou oproti hodnotám v 1 mírně posunuty jedním směrem), si lze všimnout poměrně jasné lineární závislosti. Máme tak celkově výrazněji navzájem závislé veličiny, a to buď s rostoucím nebo klesajícím trendem. Nejvýraznější lineární trend se zdá být mezi X_i a W_i. Je rostoucí, což bychom obvykle čekali, že pokud X_i budeme očekávat větší, tak i příslušný odhad W_i by se měl zvýšit, avšak je třeba v tomto místě nepřehlédnout škálu jednotlivých veličin a všimnout si, že zatímco X_i nabývá typicky hodnot mezi 10 a 50, tak W_i se pohybuje nejen v jiném řádu, ale dokonce velká část je záporná. Napozorování hodnoty -600 není určitě dobré odhadnutí hodnoty cca. 10 i přes všechen vizuálně krásně lineárně rostoucí trend, který tak ale působí díky přeškálování osy. Takto šílené odhady W_i používáme úmyslně, abychom ukázali, že i pro na první pohled naprosto nesmyslné a zcela nepoužitelné odhady lze přeci jen pomocí námi odvozených postupů ukázat, že i tyto odhady lze použít a dostat místy možná až neočekávatelně dobré odhady parametrů, což postupně probereme v dalších podkapitolách.

Z validační skupiny odhadneme s využitím soustavy odhadovacích rovnic ze 4. kapitoly nejprve alfy a σ_ψ^2, pomocí alf a W_i pak vypočteme $\hat{X}_i$ pro nevalidační skupinu a obě skupiny následně používáme pro výpočet bet dle odvozeného explicitního vyjádření v podkapitole 4.2.

Abychom mohli získat intervaly spolehlivosti pro jednotlivé parametry, vypočteme asymptotickou matici $\boldsymbol{\Sigma}_\gamma$, a to buď pomocí vyjádření ve Větě 7, nebo využití předchozích Vět 5 a 6 k odhadům inverzní matice $\boldsymbol{G}^{-1}$ a rozptylové matice $\boldsymbol{\Sigma}$ a využitím vztahu $\boldsymbol{\Sigma}_\gamma = \boldsymbol{G^{-1}\Sigma(G^{-1})}^T$, což jsme během simulování využili i my. Pak s použitím Cramérovy-Woldovy věty a Cramérovy-Slutského věty ihned získáváme směrodatné odchylky potřebné pro intervaly spolehlivosti.

Seznámení s podobou výstupu simulací provedeme přímo na konkrétním vyhodnocení ihned v další podkapitole, zde jen dodáme, že počet simulací pro jednotlivé simulované situace je 10 000, a připomeneme, že vše v této kapitole se provádí na hladině 0.05, respektive zkoumáme 95% intervaly spolehlivosti.



Obrázek 5.1 Scatter ploty mezi regresory, špatně W_i a odezvou. Vyobrazená simulace má nastavené parametry na hodnotách, které odpovídají podkapitolám pro situaci s malým rozsahem a s malým podílem validační skupiny, ve kterých se bude akorát postupně měnit jen rozsah a pravděpodobnost toho, že dané pozorování pochází z validační skupiny. Pro vyobrazenou simulaci jsme použili rozsah o velikosti 1 000.

5.3 Situace s malým rozsahem

V praxi velmi častým případem je situace, kdy máme malý rozsah. Získat více pozorování může být extrémně složité z několika ohledů. Kromě finančních důvodů to mohou být například kapacitní a časová omezení určující strop, kolik pozorování je nejvýše možné dostat, případně v lékařských studiích můžeme zkoumat velmi vzácné onemocnění, kde počet jedinců, které by bylo možné zařadit do studie, je velice nízký. Asymptotiky zřejmě fungují velice dobře pro velké rozsahy, avšak jak moc dobře použitelná je asymptotika pro malý rozsah, to bývá různé. Proto jeden z hlavních cílů této kapitoly je zkoumání nakolik se bude asymptotika zhoršovat s klesajícím rozsahem.

Aby se dalo lépe vypořádat s problémem nízkého rozsahu, je dobré se snažit udržovat podíl validační skupiny dostatečně velký. Proto v této podkapitole budeme pracovat s hodnotou $\pi = 0.5$. Než ovšem začneme studovat situaci

s nízkými rozsahy, podíváme se jak moc dobře funguje asymptotika pro jednu z obvyklých větších velikostí rozsahu, a to pro $n = 1\ 000$. Pomocí Tabulky 5.1 si postupně v následujících odstavcích představíme podobu výstupu výsledků pro 10 000 simulací.

Hlavní parametry zájmu, bety, máme uvedené hned v prvních 5 řádcích. Připomeňme, že práce je zaměřená na metodu upraveného skóre za účelem odstranění vychýlení v betách způsobené chybou v měření v některém regresoru. Samozřejmě, abychom mohli využívat i nevalidační skupinu, potřebujeme z W_i dostat odhad $\hat{X}_i$, proto se zajímáme i o alfy a σ^2_ψ, které máme uvedené ve spodních řádcích tabulky. V každé z 10 000 simulací jsme dostali odhady těchto parametrů a průměr těchto odhadů přes všechny simulace jsme uvedli do druhého sloupce tabulky hned vedle skutečných hodnot. Při porovnávání těchto dvou sloupců vidíme, že pro většinu sloupců je rozdíl pouze v řádu setin nebo dokonce tisícin, tedy až na σ^2_ψ, které se liší v řádu desetin. Tedy mimo poslední parametr žádné výrazné vychýlení v průměru odhadů parametrů není, což by nás mělo těšit, protože motivací ke studiu metody upraveného skóre bylo právě vychýlení v betách.

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	40.000	39.959	5.023	4.979	0.945
β_1	100.000	99.951	3.352	3.325	0.947
β_2	20.000	20.001	0.692	0.689	0.950
β_3	-10.000	-9.998	0.150	0.149	0.946
β_4	4.000	4.000	0.082	0.082	0.945
α_0	12.000	12.015	6.128	6.053	0.946
α_1	-6.000	-6.022	4.059	4.042	0.946
α_2	15.000	15.009	0.843	0.838	0.948
α_3	-20.000	-20.000	0.183	0.181	0.947
α_4	20.000	19.999	0.100	0.099	0.948
σ^2_ψ	49.000	48.622	3.111	3.057	0.938

Tabulka 5.1 Výsledky simulace pro n = 1 000 a $\pi = 0.50$.

Na tomto místě podotkneme proč σ^2_ψ je vychýlen, konkrétně podhodnocen. Připomeňme, že jsme pro odhadnutí tohoto parametru zvolili pro jednoduchost přímočarý odhad bez korekce o počet parametrů, abychom v $\boldsymbol{U_i}$ měli vyjádření neobsahující velikost rozsahu a dalo se s tímto vyjádřením následně lépe manipulovat v rámci 4. kapitoly. Nicméně je to parametr, který přímo v samotném odhadu bet nefiguruje a zkoumáme ho spíše jako dodatečnou informaci o chybovém modelu. Proto nás zhoršené výsledky právě u tohoto parametru tak nebudou znervózňovat, naopak to bude očekávatelné při výrazném zmenšování validační skupiny, neboť absence výše zmíněné korekce bude klesajícím rozsahem čím dál patrnější a bude vést k většímu a výraznějšímu podhodnocení. Tím ale budou

intervaly spolehlivosti pro σ^2_ψ méně a méně pokrývat skutečnou hodnotu a dojde tak ke zhoršování relativního pokrytí, což se projeví u pozdějších tabulek.

Vraťme se k Tabulce 5.1. Vypočítali jsme směrodatnou odchylku pro to, jak moc se liší odhady parametrů přes jednotlivé simulace. Směrodatnou odchylku pro 10 000 odhadů parametru můžeme najít ve třetím sloupci, který jsme nazvali „Empirická SD“. Oproti tomu hodnoty ve čtvrtém sloupci s názvem „Průměrná SD“ vznikly následovně: Na matici $\mathbf{\Sigma}_\gamma$ se při teoretickém odvození intervalu spolehlivosti pro daný parametr aplikuje Cramérova-Woldova věta a následně Cramér-Slutského. V intervalu spolehlivosti tak figuruje směrodatná odchylka, kterou už ale máme z předchozího počítání připravenou. Stačí si pro i-tý parametr vzít i-tý diagonální prvek z našeho odhadu matice $\mathbf{\Sigma}_\gamma$ a odmocnit ho. Avšak to je opět hodnota pro jednu konkrétní simulaci, tedy i zde zprůměrujeme tyto hodnoty přes všechny simulace a dostáváme tak průměrný odhad směrodatné odchylky (dále už budeme používat zjednodušené označení dle názvu sloupce). Pokud asymptotika funguje dostatečně dobře, oba sloupečky věnované SD by se neměly moc lišit. Což tomu tak skutečně je, pro některé parametry je ten rozdíl dokonce v řádu tisícin, zatímco v řádu desetin je to jen nepatrně u některých. V posledním sloupci tabulky je relativní pokrytí, které odpovídá počtu simulací (z těch všech 10 000), ve kterých dříve zmíněný interval spolehlivosti pro daný parametr pokryl skutečnou hodnotu. V tabulce vidíme, že se tak stalo ve více než 94.5 % u všech bet a alf, což je velice výborné vzhledem k blízkosti předepsaných 95 %. Dále už začneme využívat Tabulku 5.2 a začneme porovnávat změny dané snížením rozsahu.

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	40.000	40.000	7.134	7.054	0.948
β_1	100.000	100.002	4.688	4.706	0.950
β_2	20.000	19.987	0.987	0.975	0.947
β_3	-10.000	-10.000	0.212	0.211	0.948
β_4	4.000	4.000	0.115	0.116	0.949
α_0	12.000	12.070	8.795	8.575	0.946
α_1	-6.000	-5.980	5.789	5.720	0.947
α_2	15.000	15.000	1.195	1.186	0.947
α_3	-20.000	-20.002	0.263	0.256	0.941
α_4	20.000	20.000	0.143	0.140	0.945
σ^2_ψ	49.000	48.213	4.324	4.250	0.930

Tabulka 5.2 Výsledky simulace pro n = 500 a $\pi = 0.50$.

Zkoumali jsme fungování asymptotiky i pro rozsah 750, ale tento případ se nijak výrazně nelišil od situace s $n = 500$, proto tu jen uvádíme druhou zmíněnou situaci v Tabulce 5.2. I když oproti předchozí tabulce máme hodnoty pro poloviční rozsah, výsledky jsou naprosto uspokojivé. Sloupečky s SD se od sebe stále o moc neliší a o tolik výrazně ještě nevzrostly, ještě jsou v řádu jednotek, zatímco průměry

odhadů parametrů se nejen nezhoršily, ale pro mnoho parametrů je rozdíl řádem menší než tisícina. Mimo α_2 jsou všechny pokrytí alf a bet stále alespoň 0.945. Dříve zmiňované podhodnocování σ^2_ψ se začíná více promítat, a to nejen na průměru odhadů, i když to ještě není v řádu jednotek, ale i na postupně zmenšujícím se relativním pokrytí, které zatím ještě je stále slušných 0.930.

Když opět změníme rozsah, a to o aktuální polovinu, tak výsledky simulací se trošku zhorší. V Tabulce 5.3 už vidíme, že pro $n = 250$ nastává odlišnost průměru odhadů od skutečných hodnot o více než desetinu, avšak stále pro mnoho parametrů je rozdíl v řádu tisícin. Vzhledem k tomu, že parametry jsou většinou v řádu desítek, jedná se stále o relativně malé vychýlení. Dále si můžeme povšimnout, že některé empirické a potažmo průměrné SD narostly do řádu desítek. Nicméně mezi danými sloupečky jsou rozdíly většinou stále v řádu setin a tisícin a celkově relativně malé vůči absolutním hodnotám až na řádek pro σ^2_ψ, kde ten rozdíl od této tabulky dále (hned pro další situaci začne být ten rozdíl drastický) začíná být i vizuálně poutavý, podhodnocení odhadu parametru je dokonce už v řádu jednotek a relativní pokrytí kleslo těsně pod hranici 0.900. U jiných parametrů je relativní pokrytí stále excelentní. Zatímco až na α_4 je pro alfy pokrytí kolem 0.945, tak pro bety je stále větší, mimo 0.943 pro β_4.

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	40.000	39.784	10.070	10.036	0.948
β_1	100.000	99.873	6.787	6.700	0.945
β_2	20.000	20.015	1.394	1.385	0.945
β_3	-10.000	-9.992	0.302	0.300	0.946
β_4	4.000	4.000	0.167	0.165	0.943
α_0	12.000	12.038	12.445	12.160	0.945
α_1	-6.000	-6.069	8.264	8.116	0.945
α_2	15.000	15.008	1.695	1.679	0.944
α_3	-20.000	-19.999	0.371	0.364	0.945
α_4	20.000	19.998	0.203	0.199	0.940
σ^2_ψ	49.000	47.426	6.130	5.838	0.899

Tabulka 5.3 Výsledky simulace pro n = 250 a $\pi = 0.50$.

Při snížení rozsahu na $n = 100$ se začnou alfy „odtrhávat“ od bet a viditelně došlo ke zhoršení oproti betám. V Tabulce 5.4 vidíme u alf snížení pokrytí na hodnoty kolem 0.931, zatímco u bet jsou to stále hodnoty 0.945 nebo 0.943, tedy naprosto senzačně odolávají zmenšenému rozsahu. Co naopak pochopitelně neumí odolat zmenšení rozsahu, je velikost SD, které v obou sloupcích viditelně narostlo, avšak relativní rozdíly nejsou alespoň u bet velké, u alf už se sloupce začínají lišit. Zhoršení u parametru σ^2_ψ je velice výrazné, sloupce pro SD se výrazně liší, pokrytí opadá dokonce na hodnotu 0.818.

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	40.000	40.048	16.368	16.195	0.943
β_1	100.000	99.971	10.772	10.811	0.945
β_2	20.000	20.015	2.231	2.223	0.945
β_3	-10.000	-10.000	0.485	0.484	0.945
β_4	4.000	3.998	0.267	0.266	0.943
α_0	12.000	12.187	20.727	19.402	0.931
α_1	-6.000	-6.108	13.769	12.954	0.931
α_2	15.000	14.999	2.835	2.665	0.932
α_3	-20.000	-20.001	0.618	0.581	0.929
α_4	20.000	19.996	0.338	0.318	0.932
σ^2_ψ	49.000	45.068	9.713	8.403	0.818

Tabulka 5.4 Výsledky simulace pro n = 100 a $\pi = 0.50$.

Při dalším snížení rozsahu, tentokrát na $n = 50$, je vidět v Tabulce 5.5 zhoršení u všech parametrů. Je třeba si uvědomit, že při mnoha simulacích velikost validační skupiny je menší než 25 a začne to ovlivňovat i bety prostřednictvím odhadů $\hat{X}_i$. Pokrytí alf se pohybuje kolem 0.910, zatímco u bet pokrytí se ještě drží na 0.933 a výše, zatímco rozdíly průměrů odhadů od skutečných hodnot jsou stále v řádu desetin, v některých případech dokonce setin nebo tisícin. Asymptotika je tedy pro tento rozsah ještě víceméně použitelná pro bety, zatímco pro alfy už ne. Podhodnocení σ^2_ψ je zde natolik výrazné, že už o jeho zkoumání pro tak malé rozsahy ani není třeba uvažovat, což se ale vzhledem k absenci korekce očekávalo. Připomeneme, že hlavním předmětem zájmu jsou bety, pro které se ukázalo, že i pro malé rozsahy v řádu vyšších desítek stále dostáváme ne tak špatné výsledky pomocí odvozené asymptotiky ze 4. kapitoly.

Je dobré si povšimnout, že výrazně nižší relativní pokrytí je v těch řádcích, ve kterých se empirická SD relativně výrazně liší od průměrné SD. Naopak u bet, ve kterých se sloupce pro SD relativně liší málo, je to pokrytí stále poměrně vyšší. Toho si lze všimnout i v dřívějších tabulkách.

Ještě lze poznamenat, že jsme zkoušeli simulovat situaci s velikostí rozsahu 30. Je to ale tak malý rozsah, že simulování neproběhlo pro všech 10 000 iterací, ale zastavilo se někdy po více než 2 500 simulací, kdy během výpočtu nebyla splněna potřebná regulárnost a nebyl v rozptylové matici odhadnut rozptyl odhadu pro σ^2_ψ. Je nutné si uvědomit, že pro tak malý rozsah velikost validační skupiny může být kriticky malá. Avšak můžeme pro zajímavost uvést, že relativní pokrytí pro proběhnuté simulace není tak výrazně špatné u bet, které se pohybuje kolem hodnoty 0.926, avšak u alf kolem 0.862, σ^2_ψ dokonce jen sotva přes polovinu.

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	40.000	40.133	23.890	23.758	0.934
β_1	100.000	99.758	16.006	15.870	0.936
β_2	20.000	20.010	3.283	3.236	0.935
β_3	-10.000	-9.995	0.715	0.710	0.936
β_4	4.000	3.994	0.394	0.390	0.933
α_0	12.000	11.707	31.668	27.837	0.910
α_1	-6.000	-5.912	20.941	18.581	0.911
α_2	15.000	14.993	4.325	3.794	0.906
α_3	-20.000	-19.997	0.944	0.834	0.913
α_4	20.000	20.005	0.518	0.456	0.910
σ^2_ψ	49.000	40.819	13.028	9.898	0.689

Tabulka 5.5 Výsledky simulace pro n = 50 a $\pi = 0.50$.

5.4 Situace s malým podílem validační skupiny

Další možnou situací, která může nastat, je, že i když můžeme mít rozsah nevalidační skupiny výrazně velký, tak možnosti pro i-té pozorování přesně naměřit X_i je dosti komplikované či i kapacitně nebo hlavně finančně omezené. Také dané vyšetření pro přesnější naměření může být nepříjemné až dosti bolestivé či to nemusí být vůbec vhodné provádět pro většinu lidí. Tedy jsme v situaci, kdy máme malý podíl validační skupiny.

Abychom mohli dostatečně zkoumat, zda asymptotika funguje i pro π v řádu setin nebo vyšších tisícin, zvolili jsme kvůli tomu větší rozsah, který činí $n = 10\,000$. Pak pro $\pi = 0.005$ v rámci jedné simulace je střední hodnota velikosti validační skupiny 50, zatímco pro $n = 1\,000$ by bylo velice nebezpečné odhadovat pro skoro všechny pozorování $\hat{X}_i$ z cca 5 hodnot X_i a W_i v miniaturní validační skupině.

Generování i nastavení parametrů je kompletně shodné s předchozí kapitolou, jediná odlišnost je ve velikosti rozsahu n, které je navíc pevné, zatímco π se bude postupně zmenšovat a budeme zkoumat pro které hodnoty přestane být aproximace dobrá. Připomeňme, že v asymptotické rozptylové matici ve Větě 7 jsou nenulové pozice vyděleny π, tedy zmenšujícím se podílem validační skupiny se budou zvyšovat hodnoty v asymptotické rozptylové matici a respektive i zvětšovat směrodatné odchylky, které využíváme pro intervaly spolehlivosti parametrů.

Provedli jsme a vyhodnotili více situací podle hodnot π, avšak ne všechny zde uvedeme. Všechny 4 varianty, kdy π jsme pokládaly rovno hodnotám 0.50 (využívali jsme v minulé podkapitole, akorát při malém rozsahu), 0.25 (budeme využívat v následující podkapitole), 0.15 a 0.10, dávají velice podobné výsledky. Proto z nich uvedeme jen jeden případ, a to $\pi = 0.25$, neboť zrovna tuto velikost podílu validační skupiny budeme používat při testování hypotéz, avšak stejně tak jsme mohli vybrat kteroukoliv ze zbývajících. Samozřejmě jsme provedli

ještě simulaci pro nižší hodnoty π než doposud zmíněné, ale v nich už se začíná projevovat postupné zhoršování, proto je uvedeme postupně později.

O dost lepší situaci, než jsme viděli na začátku předchozí podkapitoly, kdy sice n bylo desetkrát menší než nyní, ale π bylo rovné polovině, můžeme vidět v Tabulce 5.6 zobrazující situaci pro $n = 10\ 000$ a $\pi = 0.25$. Zatímco naše odhady působily jako antikonzervativní, kdy relativní pokrytí bylo buď splněno 0.950 nebo častěji bylo menší, nyní poprvé a rovnou u několika parametrů dostáváme relativní pokrytí větší než jsme si předepsali. Z 11 parametrů jenom u 4 máme v tabulce menší hodnoty a to jen mírně. Dokonce i pokrytí pro σ^2_ψ je 0.950, zatímco vychýlení průměru odhadů je naprosto zanedbatelné, tedy naprosto skvěle fungující asymptotika i pro tento parametr!

Rozdíly mezi prvním a druhým sloupcem a rozdíly mezi třetím a čtvrtým sloupcem pro SD jsou naprosto minimální, často v řádu tisícin nebo i menší, v tabulce nepostřehnutelné. SD pro jednotlivé parametry je celkově velice malé. Obecně tak zvýšení rozsahu oproti minulé podkapitole velice pomohlo pro naprosto krásně fungující asymptotiku, a to i když π se snížilo na 0.25 nebo případně i výše zmíněnou hodnotu 0.10, kdy daná tabulka by vypadala principově podobně, pouze by bylo o trochu vyšší SD.

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	40.000	40.001	1.619	1.618	0.949
β_1	100.000	99.995	1.068	1.080	0.954
β_2	20.000	20.002	0.226	0.224	0.946
β_3	-10.000	-10.000	0.048	0.048	0.952
β_4	4.000	4.000	0.026	0.027	0.949
α_0	12.000	11.974	2.699	2.704	0.949
α_1	-6.000	-5.995	1.810	1.804	0.950
α_2	15.000	15.001	0.376	0.375	0.952
α_3	-20.000	-20.000	0.081	0.081	0.950
α_4	20.000	20.000	0.044	0.044	0.950
σ^2_ψ	49.000	48.924	1.380	1.382	0.950

Tabulka 5.6 Výsledky simulace pro n = 10 000 a $\pi = 0.25$.

První ze zkoumaných hodnot π, kdy se začne objevovat nějaká netriviální změna, je hodnota 0.05. V Tabulce 5.7 si můžeme povšimnout prvního většího poklesu relativního pokrytí a to nepřekvapivě právě u σ^2_ψ, zatímco pokrytí ostatních parametrů je nejen stále kolem hodnoty 0.950, ale u 5 parametrů pozorujeme i pro takové malé π hodnoty větší než je předepsané pokrytí, což je velice dobrým poukázaním využitelnosti námi odvozené asymptotiky i pro situace s menším podílem validační skupiny. O kolik dále budeme moci zmenšovat tento podíl, si za chvíli ukážeme pomocí dalších tabulek.

I když je tentokrát celkový rozsah oproti předchozí kapitole velký, je třeba si uvědomit, že na odhadu σ^2_ψ se nepodílí nikoliv 10 000 pozorování, ale pouze ty, které pochází z validační skupiny, což pro $\pi = 0.05$ odpovídá zhruba kolem 500. To je srovnatelně velké s validační skupinou odpovídající Tabulce 5.1, kde došlo k podobnému vychýlení průměru odhadů a relativní pokrytí bylo též srovnatelné. Je tedy potřeba mít pro další hodnoty π na paměti, že snižujícím se podílem validační skupiny se bude výrazněji a výrazněji podhodnocovat σ^2_ψ (kvůli absenci korekce o počet parametrů, kterou jsme už zmínili v předchozí podkapitole).

Mimo zmíněné obtíže s posledním parametrem mají bety i alfy naprosto krásné výsledky. Průměry odhadů se liší často v řádu tisícin, případně setin, což obdobně platí i pro odlišnosti sloupců pro SD. Obecně hodnoty pro SD se oproti situaci s $\pi = 0.25$ o moc nezvýšily. Proto pojďme opět snížit podíl validační skupiny, opět na pětinu, tedy nyní 0.01.

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	40.000	40.012	1.933	1.957	0.951
β_1	100.000	99.991	1.300	1.306	0.950
β_2	20.000	20.004	0.271	0.270	0.946
β_3	-10.000	-10.000	0.058	0.058	0.952
β_4	4.000	4.000	0.032	0.032	0.951
α_0	12.000	11.953	6.038	6.050	0.951
α_1	-6.000	-5.963	4.053	4.037	0.947
α_2	15.000	15.003	0.845	0.838	0.946
α_3	-20.000	-20.000	0.180	0.181	0.952
α_4	20.000	20.001	0.100	0.099	0.950
σ^2_ψ	49.000	48.547	3.112	3.049	0.936

Tabulka 5.7 Výsledky simulace pro n = 10 000 a $\pi = 0.05$.

Pomocí Tabulky 5.8 vidíme, že při zmenšení podílu validační skupiny na setinu, tedy ve střední hodnotě mající kolem 100 pozorování ve validační skupině, se začíná projevovat zhoršení výsledků i pro alfy a bety. U každého z nich je relativní pokrytí menší než předepsaných 0.950, avšak až na β_2 mají bety relativní pokrytí aspoň 0.945, u alf je to kolem hodnoty 0.943, což je stále pěkné. SD jsou u bet stále maximálně v řádu nižších jednotek, avšak u bet SD rostou, a to dokonce přes 13.500 pro α_2. Podhodnocení σ^2_ψ lze už jasně poznat a relativní pokrytí je pouhých 0.880.

Poslední hodnotou π, pro kterou zobrazíme výsledky pomocí Tabulky 5.9, je hodnota 0.005. To odpovídá očekávané velikosti validační skupiny zhruba 50. Hlavní parametry zájmu, bety, mají stále poněkud dobré pokrytí, a to kolem 0.935, zatímco pro některé alfy se snížilo pokrytí na hodnoty kolem 0.930. SD výrazně narůstají u alf, kde pro dva parametry jsou v řádu desítek, zatímco u bet jsou stále

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	40.000	40.024	3.177	3.157	0.949
β_1	100.000	99.949	2.156	2.108	0.945
β_2	20.000	20.010	0.446	0.435	0.943
β_3	-10.000	-9.999	0.095	0.094	0.947
β_4	4.000	3.999	0.053	0.052	0.945
α_0	12.000	11.690	13.945	13.620	0.944
α_1	-6.000	-5.947	9.442	9.088	0.939
α_2	15.000	15.000	1.944	1.878	0.940
α_3	-20.000	-19.995	0.419	0.407	0.943
α_4	20.000	20.004	0.230	0.223	0.942
σ^2_ψ	49.000	46.919	6.844	6.403	0.880

Tabulka 5.8 Výsledky simulace pro n = 10 000 $\pi = 0.01$.

maximálně jen nižší jednotky. Obecně situace s SD vypadá poměrně dobře a stabilněji i pro tak malý podíl validační skupiny, minimálně u bet je situace o dost lepší, než tomu tak bylo ke konci předchozí podkapitoly.

Parametr σ^2_ψ má podobné výsledky jakožto bylo pro malý rozsah s podobně velkou validační skupinou, vizte Tabulku 5.4. Mimo tento podhodnocený parametr jsou průměry odhadů ostatních parametrů velice blízko skutečným hodnotám. I pro $\pi = 0.005$ stále máme u několika parametrů rozdíl v řádu tisícin, u dvou dokonce ještě menší.

Zdá se, že malý podíl validační skupiny není problémem pro dobré fungování asymptotiky, dokud je validační skupina v absolutním číslu dostatečně velká, tedy má alespoň kolem 50 pozorování.

Při zkoumání hodnoty $\pi = 0.002$ se už objevila jedna simulace obsahující jeden NaN, avšak ostatní simulace nebyly pozastaveny. Nicméně validační skupina je v absolutních číslech pro některé simulace tak výrazně malá, že docházelo k výrazně horšímu pokrytí u všech parametrů. Rozdíl mezi skutečnou hodnotou a průměrem odhadů se lišil v řádu desetin.

Nicméně pokud bychom chtěli pořádněji studovat ještě výrazněji menší podíly validační skupiny, aniž by nastávaly problémy nutně zapříčiněnými příliš malou validační skupinou v absolutním smyslu, bylo by potřeba celkový rozsah výrazně zvětšit, například na 100 000. To ale pro běžnou praxi už není obvykle možné.

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	40.000	40.125	4.367	4.233	0.942
β_1	100.000	99.899	2.976	2.824	0.935
β_2	20.000	20.026	0.615	0.580	0.932
β_3	-10.000	-10.000	0.132	0.126	0.935
β_4	4.000	3.997	0.072	0.070	0.937
α_0	12.000	11.797	20.487	19.448	0.935
α_1	-6.000	-5.914	13.983	12.968	0.931
α_2	15.000	15.042	2.869	2.672	0.928
α_3	-20.000	-20.000	0.621	0.582	0.932
α_4	20.000	20.004	0.338	0.319	0.929
σ^2_ψ	49.000	44.918	9.667	8.359	0.817

Tabulka 5.9 Výsledky simulace pro n = 10 000 a $\pi = 0.005$.

5.5 Testování nulové hypotézy

Navazujeme na předchozí podkapitoly, avšak tentokrát se chceme zaměřit na testování nulové hypotézy $\beta_4 = 0$ vůči alternativě $\beta_4 \neq 0$ pro situaci, kdy pracujeme se zobecněným chybovým modelem, a testování nulové hypotézy $\beta_1 = 0$ vůči alternativě $\beta_1 \neq 0$ pro situaci, kdy pracujeme s klasickým základním typem chybového modelu mající chybu s nulovou podmíněnou střední hodnotou. Konkrétně budeme zkoumat, zda pro nulovou hypotézu bude chyba prvního typu dodržena, tedy nebude docházet k častějšímu zamítání nulové hypotézy, než by mělo.

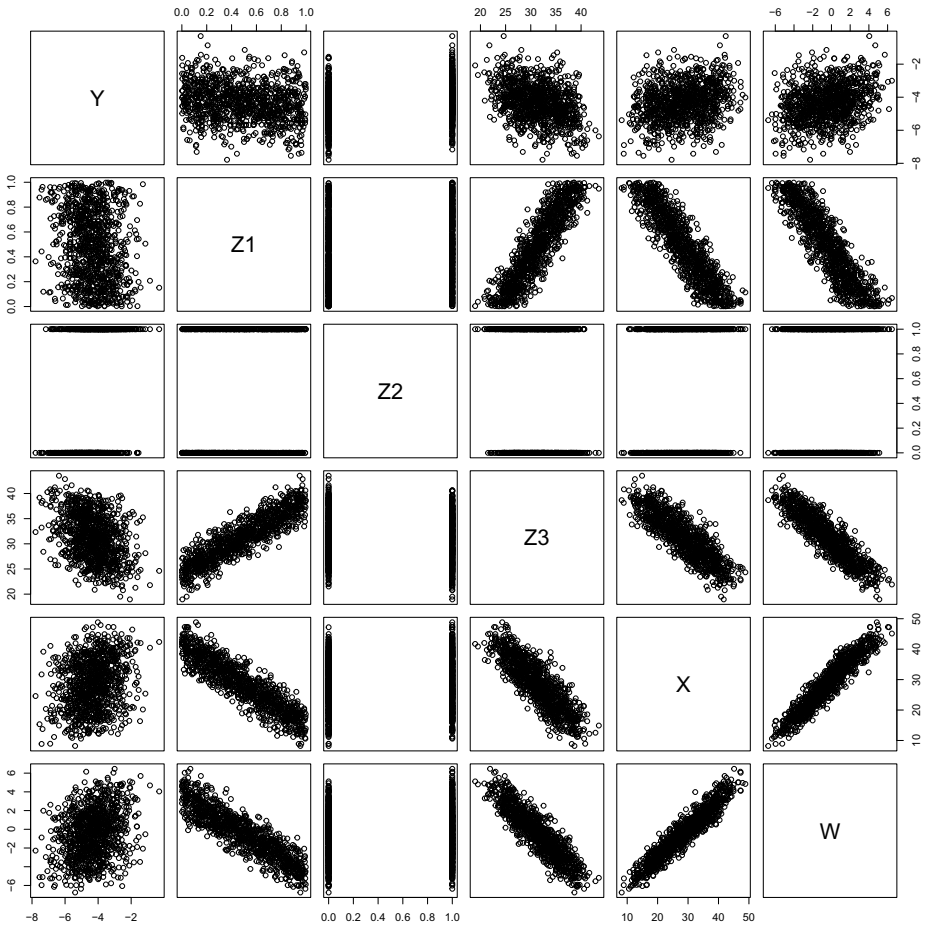
Dále u druhé situace budeme porovnávat výsledky i s neupraveným odhadem metodou nejmenších čtverců (který by byl zjevně nesmyslný pro porovnávání v prvním případě, kdy chyba není nestranná, ale silně vychýlená), kdy jsme už v dřívějších kapitolách zmínili, že parametr pro zašuměný regresor X_i bude vychýlen směrem k nule, tak se zaměříme na intervaly spolehlivosti, zda budou pokrývat skutečnou hodnotu parametru pro první regresor.

5.5.1 Testování nulové hypotézy pro zobecněný model

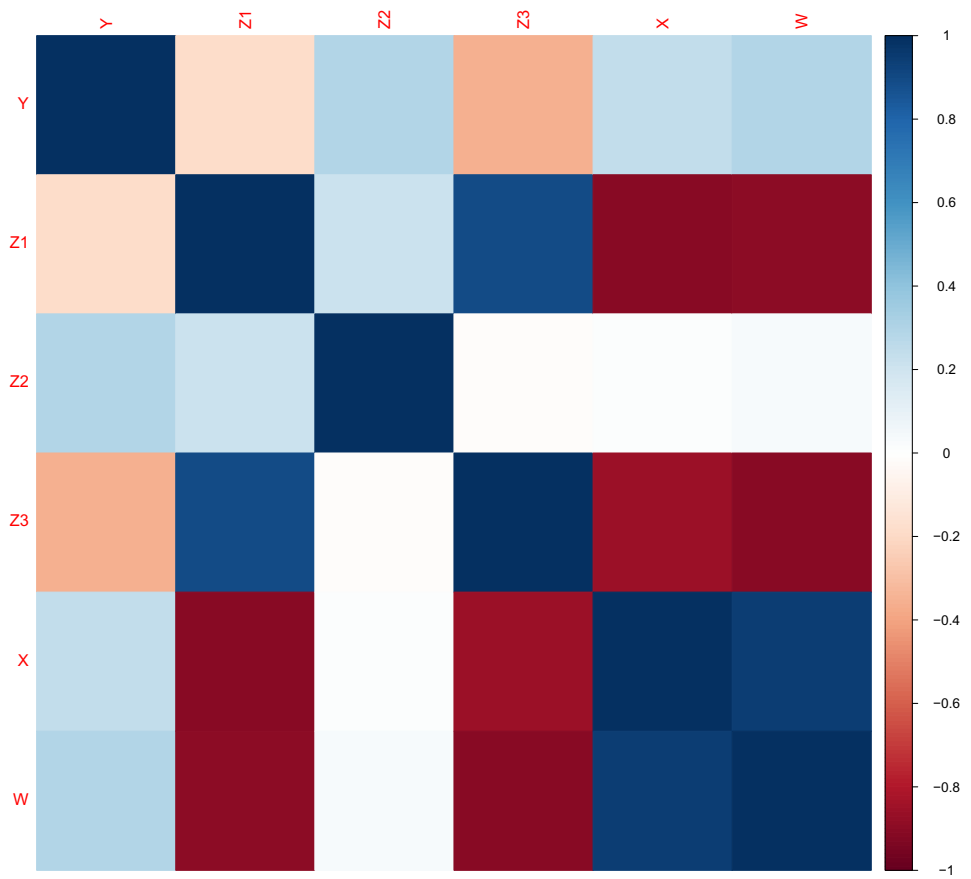
Protože pro předchozí nastavení parametrů hlavního a chybového modelu jsme dostávali poměrně úzké intervaly spolehlivosti pro dané parametry, přiblížíme všechny parametry blíže k nule tak, aby stále scatter ploty na Obrázku 5.2 a korelace odezvy vůči ostatním proměnným, které jsou znázorněny na Obrázku 5.3, poukazovaly aspoň na nějaký vztah odezvy vůči většině proměnným, i kdyby byl jen mírně rostoucí nebo mírně klesající, zejména aby byla nějaká korelace mezi X_i a odezvou. Zajímá nás, zda i přes všechny vztahy mezi proměnnými bude model schopen poukázat na nevýznamný vliv X_i na odezvu, či zda bude chybně přisuzovat větší vliv. Proto nastavme parametry například na hodnoty

$$\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix} = \begin{pmatrix}0.8\\ 2\\ 0.4\\ -0.2\\ 0\end{pmatrix}, \begin{pmatrix}\alpha_0\\ \alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4\end{pmatrix} = \begin{pmatrix}0.12\\ -0.06\\ 0.15\\ -0.20\\ 0.20\end{pmatrix}, \sigma_\epsilon^2 = 1 \text{ a } \sigma_\psi^2 = 0.5.$$

Vídíme, že až na některé dvojice s binární veličinou jsou korelace viditelně nenulové. Korelace X_i s ostatními nebinárními regresory jsou velmi značné, i se samotnou odezvou Y_i má X_i a i W_i kladnou korelaci, zatímco i u scatter plotu by se též očekával rostoucí trend s ohledem na ostatní scatter ploty. Je také dobré si povšimnout, že došlo k vychýlení odhadů X_i. Zatímco X_i se pohybovalo v řádu desítek, chybou zatížené W_i jsou v řádu jednotek, velká část je záporná. Došlo tak k výraznému podhodnocení X_i a bez jakýchkoliv úprav by v samotném odhadování bet nedávalo smysl pracovat se samotnými W_i a tvářit se, že se jedná o X_i. Přesto ale pomocí našich úprav se dá i s touto situací rozumně pracovat.



Obrázek 5.2 Scatter ploty mezi regresory, odezvou Y_i a W_i odhadující X_i podle zobecněného chybového modelu pro testování nulové hypotézy $\beta_4 = 0$.



Obrázek 5.3 Korelace mezi regresory, odezvou Y_i a W_i odhadující X_i podle zobecněného chybového modelu pro testování nulové hypotézy $\beta_4 = 0$.

Tuto záležitost můžeme vyřešit poměrně lehce a využít již známého postupu. Můžeme opět vygenerovat 10 000 simulací o rozsahu $n = 1\ 000$ a podílem validační skupiny $\pi = 0.25$ a pro každou jednotlivou simulaci zkoumat, zda interval spolehlivosti pro parametr β_4 pokrývá hodnotu 0. Z duality testování a intervalů spolehlivosti víme, že pokud interval spolehlivosti nebude pokrývat danou hodnotu, tak lze zamítnout nulovou hypotézu. Zde tedy by se zamítla nulová hypotéza o nulovosti efektu X_i na odezvu ve prospěch alternativy, že X_i má vliv na hodnotu odezvy Y_i.

Protože stále pracujeme na hladině 0.05, zamítnutí nulové hypotézy by mělo být jen v 5 % případů, tedy pokrývání nuly intervalem spolehlivosti by mělo nastávat ve zhruba 95 % simulacích. Pokud se podíváme na Tabulku 5.10, tak zjistíme, že z 10 000 simulací byla 0 pokryta intervaly spolehlivosti zhruba v 95.1 %. Samotný průměr odhadů je v tabulce neodlišitelný od nulové hodnoty, tedy žádné systematické vychýlení se neprojevilo. Zjistili jsme tak, že i přes všechny korelace a závislosti model umí nepřisuzovat falešně vliv regresoru, který je zatížen chybou při měření, a to když je chyba obecně vychýlená. Na speciální situaci, kdy je chyba nestranná a budeme porovnávat se standardním odhadem nejmenších čtverců, se podíváme v další sekci.

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	0.800	0.792	0.731	0.736	0.951
β_1	2.000	1.996	0.494	0.498	0.952
β_2	0.400	0.400	0.094	0.094	0.950
β_3	-0.200	-0.200	0.019	0.019	0.948
β_4	0.000	0.000	0.014	0.014	0.951
α_0	0.120	0.118	0.878	0.866	0.947
α_1	-0.060	-0.063	0.591	0.578	0.943
α_2	0.150	0.152	0.122	0.120	0.946
α_3	-0.200	-0.200	0.026	0.026	0.947
α_4	0.200	0.200	0.015	0.014	0.943
σ^2_ψ	0.500	0.492	0.045	0.044	0.926

Tabulka 5.10 Výsledky simulace pro testování nulové hypotézy pro zobecněný chybový model.

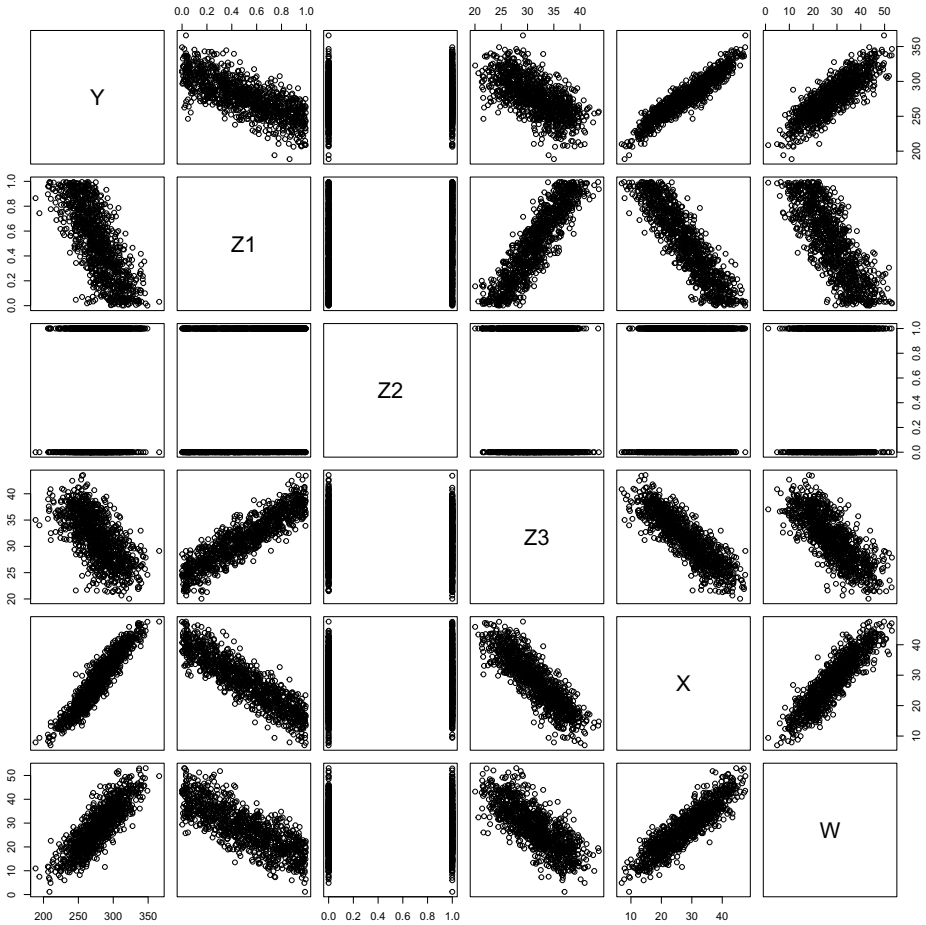
5.5.2 Testování nulové hypotézy pro klasický model

Jak bylo zmíněno výše, nyní za účelem porovnávání našich odhadů a relativního pokrytí skutečné hodnoty vůči nijak upraveného odhadu metodou nejmenších čtverců budeme uvažovat klasický typ chybového modelu, ve kterém bude chyba mít nulovou podmíněnou střední hodnotu. Protože budeme uvažovat chybu v měření jen u jednoho regresoru, je to speciálním případem pro náš zobecněný model, ve kterém všechny alfy vynulujeme kromě α_4, která bude rovna 1, abychom dostali patřičný vztah

$$W_i = X_i + \psi_i.$$

Jak bylo řečeno a ukázáno na Obrázku 1.1, regresor zatížen chybou v měření, sic mající nulovou podmíněnou střední hodnotu, se typicky „rozprostře“ do prostoru a dojde k vychýlení příslušného regresoru směrem k nule. Oproti tomu u parametrů pro ostatní regresory obecně nevíme k jakému vychýlení dojde a nakolik velké bude. Proto dává smysl místo nulové hypotézy $\beta_4 = 0$ vůči alternativě $\beta_4 \neq 0$ uvažovat testování nulového efektu jiného regresoru, například testování $\beta_1 = 0$ vůči alternativě $\beta_1 \neq 0$. Zároveň zvětšíme rozptyl chyby, aby důsledky chybného naměření byly dostatečně vidět. Nastavme proto hodnoty parametrů na

$$\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix} = \begin{pmatrix}10\\ 0\\ 3\\ 4\\ 5\end{pmatrix}, \begin{pmatrix}\alpha_0\\ \alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4\end{pmatrix} = \begin{pmatrix}0\\ 0\\ 0\\ 0\\ 1\end{pmatrix}, \sigma^2_\epsilon = 1 \text{ a } \sigma^2_\psi = 16.$$



Obrázek 5.4 Scatter ploty mezi regresory, odezvou Y_i a W_i odhadující X_i podle klasického typu chybového modelu, kde chyba má nulovou podmíněnou střední hodnotu.

Na obrázku 5.4 si můžeme povšimnout, že i přes opětovné vynulování vlivu jednoho z regresorů je opět korelace s odezvou zjevná. Přesto při pohledu na Tabulku 5.11 je zjevné, že i když relativní pokrytí pro σ_ψ^2 dosahuje kvůli menšímu podhodnocení hodnoty 0.926, odhady získanými ze soustavy odhadovacích rovnic obsahující korekci skóre a využití odvozené asymptotiky jsou natolik dobré, že i při vyšším rozptylu chyby v měření dostáváme v průměru odhady poměrně blízké skutečným hodnotám a velice dobré relativní pokrytí skutečných hodnot.

Jenom pro β_0 a β_1 dostáváme rozdíl průměru odhadů a skutečné hodnoty v řádu vyšších desetin, avšak je třeba upozornit na vysokou směrodatnou odchylku pro oba dva odhady. I přes vysoké hodnoty je empirická SD relativně dosti podobná průměrné SD, což je příjemným indikátorem toho, že asymptotika ještě pro takto velké σ_ψ^2 stále dobře funguje, což je v souladu s tím, že relativní pokrytí pro oba parametry je stále vyšší než 0.940.

Celkově výsledky simulace ukazují, že pokrývání skutečné hodnoty β_1 poměrně dobře odpovídá předepsané hladině. Z duality když 0 je pokryta intervalem spolehlivosti, tak se nezamítá nulová hypotéza ve prospěch alternativy o nenulovém vlivu daného regresoru. Model tedy nepřisuzuje falešně větší vliv regresoru v nějak výrazně větším množství případů než bychom chtěli vůči předepsané hladině.

	Skutečná hodnota	Průměr odhadů	Empirická SD	Průměrná SD	Relativní pokrytí
β_0	10.000	10.798	33.793	33.098	0.941
β_1	0.000	-0.654	23.276	22.713	0.942
β_2	3.000	3.101	3.946	3.897	0.947
β_3	4.000	4.002	0.646	0.644	0.949
β_4	5.000	4.980	0.743	0.721	0.938
α_0	0.000	-0.019	4.966	4.900	0.947
α_1	0.000	-0.012	3.335	3.271	0.944
α_2	0.000	0.007	0.690	0.678	0.946
α_3	0.000	0.001	0.149	0.147	0.946
α_4	1.000	1.000	0.082	0.080	0.944
σ^2_ψ	16.000	15.741	1.425	1.392	0.926

Tabulka 5.11 Výsledky simulace pro testování nulové hypotézy pro klasický chybový model s chybou v měření mající nulovou střední hodnotu.

Nyní se podívejme co by se stalo, pokud bychom žádnou korekci v odhadech nepoužili a chtěli bychom slepě věřit, že chyba v měření X_i není natolik významná, abychom nemohli naivně používat W_i v roli X_i. V programu `R` jsme si manuálně vypočetli odhad metodou nejmenších čtverců, který jsme uvedli v sekci 1.1.2, a intervaly spolehlivosti jsme spočetli na základě Věty 2, kde jsme za $\boldsymbol{c}$ dosadili vektor obsahující nuly až na 1 v druhé složce odpovídající parametru β_1.

Dle očekávání došlo k vychýlení většiny odhadů, což můžeme vidět v Tabulce 5.12. Zatímco parametr β_4 byl skutečně vychýlen směrem k 0 o zhruba polovinu skutečné hodnoty, odhady prvních třech parametrů jsou silně vychýlené, a to k větším hodnotám, mimo β_1, které místo hodnot blízkých nule má průměr odhadů -76.401. Všechny odhady prvních třech parametrů mají rozdíl průměru odhadů od skutečných hodnot větší než 10, v případě β_0 je to dokonce o více než 100.

Lze si ale zároveň povšimnout, že empirické směrodatné odchylky jsou poměrně malé, a to dokonce výrazně menší než jaké byly v Tabulce 5.11. To celkově svědčí, že sice odhady pro jednotlivé parametry mají vůči různým simulacím malou variabilitu, avšak konzistentně to odhaduje úplně něco jiného, než bychom si přáli.

Vzhledem k velkým vychýlením a malým empirickým SD není žádného divu, že ani v jedné z 10 000 simulací nebyla skutečná hodnota pokryta intervalem spolehlivosti. To znamená, že pro každou simulaci se chybně zamítá nulová hypotéza ve prospěch alternativní hypotézy.

Ukázalo se, že i když předpokládáme nejjednodušší typ chybového modelu, tak pro daný rozptyl nestranných chyb dává odhad metodou nejmenších čtverců chybně rozhodnutí o nenulovém vlivu jednoho z regresorů, a to dokonce pro každou simulaci! Oproti tomu náš odhad založený na metodě upraveného skóre nepřisuzuje chybně v nijak výrazně zvětšené míře nenulový vliv danému regresoru, ale naopak 94.2 % intervalů spolehlivosti pokrývá nulu.

	Skutečná hodnota	Průměr odhadů	Empirická SD
β_0	10.000	118.996	6.511
β_1	0.000	-76.401	4.296
β_2	3.000	13.912	0.949
β_3	4.000	4.003	0.216
β_4	5.000	2.273	0.088

Tabulka 5.12 Výsledky simulace pro testování nulové hypotézy pro odhad metodou nejmenších čtverců.

5.6 Shrnutí simulační části

Cílem práce bylo otestovat nakolik je použitelná asymptotika odvozená ve 4. kapitole. To jsme provedli na několika typech simulací (pro každý typ jsme provedli 10 000 iterací simulace) a zároveň jsme se postup včetně samotného generování dat snažili vysvětlit tak podrobně, aby bylo pro čtenáře možné replikovat výsledky.

Hlavním objektem zájmu bylo zkoumání, nakolik přesně jsou odhadovány bety v hlavním modelu, zda stále dochází k nějakému systematickému vychýlení (či zda naopak korekce metodou upraveného skóre byla dostatečná) a zda relativní pokrytí skutečných hodnot jednotlivých parametrů intervaly spolehlivosti se blížilo k námi předepsané spolehlivosti 0.950. Souběžně s tím jsme pozorovali i výsledky parametrů chybového modelu, zejména alfy, avšak předmětem zájmu metody upraveného skóre jsou bety, respektive náprava v jejich odhadování, proto budeme zkoumat i případy, kdy pro bety stále máme slušné výsledky, i když pro alfy už nastalo poměrné zhoršení. U parametru σ^2_ψ dochází výrazně dříve k horším výsledkům, což je zapříčiněno tím, že parametr je odhadován bez korekce, bez které zejména pro nižší n dochází k podhodnocování tohoto parametru. Avšak absence této korekce nám umožnila ve 4. kapitole pracovat s $\boldsymbol{U_i}$ i bez toho, aby se tam projevovala velikost rozsahu validační skupiny.

První situací, kterou jsme se zabývali, byl malý celkový rozsah obou skupin n. Předpokládali jsme, že hlavním problémem je obecně získat další pozorování, ale není tak výrazně problémové provést přesnější měření zhruba na polovině subjektů, tedy za účelem boje proti nízkým rozsahům jsme předpokládali $\pi = 0.500$, jinými slovy zhruba stejnou velikost validační skupiny jako je velikost nevalidační skupiny. Tato situace stále má smysl, protože oproti situaci naměření přesných hodnot u všech subjektů ušetří u poloviny celkového rozsahu n náklady, které by obnášelo přesnější měření oproti běžnému měření (případně jiné komplikace, které byly v kapitole zmíněny).

Při velikosti $n = 250$ jsou výsledky pro bety i alfy stále obstojně fungující s relativním pokrytím kolem 0.945, i když pro některé parametry začínají být směrodatné odchylky větší (nemluvě o σ^2_ψ, u kterého se podhodnocování začíná projevovat už pro $n = 1\ 000$). Relativní pokrytí pro bety je stále podobné i při snížení rozsahu na $n = 100$, avšak výsledky pro alfy se začínají zhoršovat a pro $n = 50$ je pro ně asymptotika už hůře fungující aproximací, kde relativní

pokrytí opadlo na hodnoty kolem 0.910. Je třeba ale mít na paměti, že při $n = 50$ má validační skupina v mnoha simulacích méně než 25 pozorování, což alfy počítáme jenom pomocí validační skupiny. Avšak i při tak malém rozsahu jsou průměry odhadů prakticky nevychýlené od skutečných hodnot bet, relativní pokrytí je 0,933 nebo vyšší.

Dalším typem situace, kterou jsme podrobně zkoumali, je malý podíl validační skupiny. Zde naopak předpokládáme, že není problém získat dostatečně mnoho pozorování pro nevalidační skupinu, avšak je problémové a náročné provést přesné naměření X_i. Proto jsme zkoumali různé postupně se zmenšující hodnoty pro π a to i v řádu tisícin, což aby bylo možné zkoumat, uvažovali jsme $n = 10\ 000$. Ukázalo se, že zkoumané hodnoty π od 0.100 do 0.500 dávají prakticky stejné výsledky, pouze se mění SD, neboť v asymptotické rozptylové matici jsou jednotlivé pozice děleny π.

První menší změna nastává pro $\pi = 0.050$, což ale je jen první výrazné zhoršení relativního pokrytí pro σ^2_ψ, které opadlo na 0.936, ale dále se zhoršuje, neboť klesajícím π se zmenšuje velikost validační skupiny, tedy absence korekce se více projevuje a tím se podhodnocení zvýrazňuje. Avšak odhady ostatních parametrů náramně uspokojivě fungují i pro $\pi = 0.010$, kdy se pro alfy začínají výsledky mírně zhoršovat, ale stále je relativní pokrytí kolem 0.940 nebo vyšší. Špatné aproximování pomocí odvozené asymptotiky se začíná projevovat až pro tisíciny, kdy třeba pro $\pi = 0.005$ je relativní pokrytí u bet kolem hodnot 0.935, avšak průměr odhadů nevykazuje žádné systematické výchylky od skutečných hodnot. Je dobré si uvědomit, že pro π mající hodnoty v řádu tisícin dostáváme už validační skupinu v řádu desítek pozorování, tedy opět hlavně parametry chybového modelu mají málo informací k odhadování, respektive zhoršuje se i samotný odhad $\hat{X}_i$.

Zdá se, že samotný podíl velikosti validační skupiny k celkovému rozsahu až tak není důležitý, jakožto samotná velikost skupin, zejména té validační. Výsledky v obou typech simulací jsou si mimo SD (které je pro výrazně větší rozsah nevalidační skupiny výrazně menší) podobné pro stejně velké očekávané velikosti validační skupiny. Nejspíše by tedy šlo dále zmenšovat podíl validační skupiny a zkoumat π mající hodnoty v desetitisícinách, avšak zřejmě by bylo zapotřebí mít velikost validační skupiny alespoň vyšší desítky ne-li nižší stovky, což by ale znamenalo mít n výrazně vyšší než 10 000.

Dále jsme provedli testování hypotéz, kdy se ukázalo, že i když jsou regresory více mezi sebou závislé a X_i nemá na odezvu přímý vliv (příslušný parametr v hlavním modelu je roven nule), tak i při zašumění tohoto regresoru, tedy obdržením odhadu W_i, který je na první pohled nerozumným a vychýleným odhadem, tak přesto lze nejen získat nestranný odhad $\hat{X}_i$ a mít celkově dobré výsledky simulací, ale v 95.1 % případů interval spolehlivosti pro β_4 pokryl nulu, tedy nulová hypotéza se nezamítla ve prospěch alternativy. Ukázalo se tudíž, že pro takovou situaci dokáže model nepřisuzovat falešně vliv pro regresor s chybným měřením, pokud na odezvu vliv skutečně nemá.

Abychom mohli porovnávat naše výsledky s odhady metodou nejmenších čtverců a dávalo by to smysl, uvažovali jsme základní typ klasického chybového modelu mající nulovou podmíněnou střední hodnotu chyb. Ukázalo se, že pro testování vlivu ostatních regresorů sice dostáváme vysoké hodnoty SD, ale skutečné

hodnoty jsou intervaly spolehlivosti pokryty rozumně dobře, zatímco odhady některých parametrů neupravené metody nejmenších čtverců jsou výrazně a konzistentně vychýleny, zamítající nulovou hypotézu ve prospěch nenulového vlivu daného regresoru a to nejen chybně, ale zamítající dokonce pro každou provedenou simulaci. Ukazuje se tedy, že i pro tak jednoduchou situaci se dá dobře využít výsledků 4. kapitoly a vede to i přes vyšší SD k rozumnějším odhadům, minimálně těm intervalovým.

Závěr

Na začátku práce bylo poukázáno na to, že i situace, kde je regresor naměřen s chybou mající nulovou střední hodnotu, vede k vychýleným odhadům parametrů a tím i k potížím pro testování hypotéz, změnění závislostí, které už nemusí být ani postřehnutelné a celkově zamaskování povahy dat. Tyto všechny pohromy mohou nastat i pro nejjednodušší chybový model, nemluvě o složitějších modelech.

Práce se tak zaměřuje na odstranění vychýlení odhadů parametrů a respektive umožnění testování hypotéz vlivu regresorů. Toho docílila pomocí metody upraveného skóre, která je založena na korekci skóre tak, aby nebylo vychýlené, což se následně projeví i na úpravách odhadů parametrů, které jsou porovnávány vůči neupraveným odhadům metodou nejmenších čtverců.

Práce se nezaměřila na využití metody jen pro základní chybový model, ale aplikaci rozšířila na zobecněný chybový model, kde měření regresoru odpovídá zašumění lineární kombinace všech (nebo respektive jen některých podle nastavených parametrů) regresorů. Soustava odhadovacích rovnic byla tak sestavena nejen pro samotné parametry hlavního modelu, ale i pro parametry chybového modelu, a to za pomoci validační skupiny. Pro tyto odhady bylo velice podrobně a pečlivě odvozeno sdružené asymptotické rozdělení.

Zmíněné asymptotické rozdělení bylo podrobeno zkoumání v simulační části. Bylo zjištěno, že pro dobré výsledky simulací není až tak důležitým faktorem samotný podíl velikosti validační skupiny vůči celkovému rozsahu (až na velikost směrodatných odchylek), jakožto samotná velikost rozsahů skupin, zejména validační skupiny. Zdá se, že pokud validační skupina má alespoň 100 pozorování, relativní pokrytí skutečných hodnot parametrů hlavního modelu je dostatečně dobré pro použití odvozené asymptotiky.

V simulační části bylo též provedeno testování hypotéz o nulovém vlivu daného regresoru, a to jak pro samotný regresor zatížený chybou při měření, tak i pro regresor, který byl vždy naměřen zcela přesně. Pro oba dva případy se ukázalo, že model nepřikládá falešně danému regresoru vliv, pokud ho skutečně nemá, i když samotné regresory jsou mezi sebou závislé. Pří testování hypotézy o nulovém vlivu vždy přesně naměřené veličiny byly parametry nastaveny tak, aby to odpovídalo nejjednoduššímu typu chybového modelu a dávalo smysl to porovnávat s výsledky pro odhady získané neupravenou metodou nejmenších čtverců. Zatímco neupravené odhady vykazovaly zcela evidentní konzistentní vychýlení, upravené odhady měly relativní pokrytí skutečných hodnot intervaly spolehlivosti dostatečně dobré. Celkově se tak ukázalo, že využití metody upraveného skóre je v lineární regresi vhodný způsob odstranění vychýlení v odhadování parametrů a získání možnosti testovat vliv regresorů na odezvu.

Literatura

1. Carroll, Raymond J.; Ruppert, David; Stefanski, Leonard A.; Crainiceanu, Ciprian M. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. 2006.
2. BUONACCORSI, J. P. A modified estimating equation approach to correcting for measurement error in regression. *Biometrika*. 1996, roč. 83, č. 2, s. 433–440. ISSN 0006-3444. Dostupné z DOI: 10.1093/biomet/83.2.433.
3. Kulich, Michal; Lin, D. Y. Additive Hazards Regression with Covariate Measurement Error. *Journal of the American Statistical Association*. 2000, roč. 95, č. 449, s. 238–248. Dostupné z DOI: 10.1080/01621459.2000.10473917.
4. HU, CHENGCHENG; LIN, D. Y. Cox Regression with Covariate Measurement Error. *Scandinavian Journal of Statistics*. 2002, roč. 29, č. 4, s. 637–655. Dostupné z DOI: https://doi.org/10.1111/1467-9469.00310.
5. Zucker, David M.; Spiegelman, Donna. Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Statistics in Medicine*. 2008, roč. 27, č. 11, s. 1911–1933. Dostupné z DOI: https://doi.org/10.1002/sim.3159.
6. Yuanshan Wu, Yanyuan Ma; Yin, Guosheng. Smoothed and Corrected Score Approach to Censored Quantile Regression With Measurement Errors. *Journal of the American Statistical Association*. 2015, roč. 110, č. 512, s. 1670–1683. Dostupné z DOI: 10.1080/01621459.2014.989323.
7. Wang, Huixia Judy; Stefanski, Leonard A.; Zhu, Zhongyi. Corrected-loss estimation for quantile regression with covariate measurement errors. *Biometrika*. 2012, roč. 99, č. 2, s. 405–421. ISSN 0006-3444. Dostupné z DOI: 10.1093/biomet/ass005.
8. Khudyakov, Polyna; Gorfine, Malka; Zucker, David; Spiegelman, Donna. The impact of covariate measurement error on risk prediction. *Statistics in Medicine*. 2015, roč. 34, č. 15, s. 2353–2367. Dostupné z DOI: https://doi.org/10.1002/sim.6498.
9. Tosteson, Tor D.; Buzas, Jeffrey S.; Demidenko, Eugene; Karagas, Margaret. Power and sample size calculations for generalized regression models with covariate measurement error. *Statistics in Medicine*. 2003, roč. 22, č. 7, s. 1069–1082. Dostupné z DOI: https://doi.org/10.1002/sim.1388.
10. Buzas, J.S.; Stefanski, L.A. A note on corrected-score estimation. *Statistics Probability Letters*. 1996, roč. 28, č. 1, s. 1–8. ISSN 0167-7152. Dostupné z DOI: https://doi.org/10.1016/0167-7152(95)00074-7.
11. Omelka, Marek. *Mathematical Statistics 3* [Online]. 2024. Dostupné také z: https://www.karlin.mff.cuni.cz/~omelka/Soubory/nmst424/nmst424_course-notes.pdf. Accessed: 2024-07-17.
12. Kang, Hyunseung. *Causal Inference: Estimation via Z Estimators*. 2023. Dostupné také z: https://pages.cs.wisc.edu/~hyunseung/stat992/MZEstimator.pdf. Accessed: 2024-07-17.

Seznam obrázků

Seznam tabulek