ROBUST'98, 115 – 128

MODEL ADITIVNÍHO RISIKA S CHYBAMI V PREDIKTORECH

Michal KULICH[1]

MFF UK, KPMS

Abstract: This paper studies the estimation of a regression parameter in the additive hazards model for censored data. The model includes one covariate measured with error. The true value of the covariate is known only in a randomly selected validation set. A surrogate covariate is assessed for all observations. A pseudoscore function that defines a consistent and asymptotically normal estimator of the regression parameter is derived.

Резюме. Ета статя занимается оценкой параметра регрессионной модели аддитивного риска для цензурированных данных. Модељ содержает одну независимую переменную измеренную с погрешностью. Действитељное значение этой переменной знакомо тољко в случайно избранной валидационной группе. Вторая запасная переменная измеряемая у всех наблюдений. В этом положении выведена псеудоскорная функция которая определяет состоятељную и асимптотически нормаљную оценку регрессионного параметра.

1. Úvod

Počet pacientů v lékařských a epidemiologických studiích potřebný k dosažení dostatečné síly vůči předem dané klinicky významné alternativě se často pohybuje v řádu tisíců a desetitisíců. Finanční náklady potřebné k provedení takových studií jsou tak velké, že mnohé otázky důležité pro efektivní prevenci a léčbu některých chorob je takřka nemožné spolehlivě zodpovědět. Podstatná část celkových nákladů je nezřídka vynaložena na jedno nebo více velmi drahých měření, které je třeba provést na každém pacientovi. Jako příklad lze jmenovat složitý laboratorní test nebo důkladné a spolehlivé vyšetření skladby potravy.

Problém drahých měření má snadné a zřejmé řešení. Stačí provést tato měření pouze na malé části výběru a náklady okamžitě klesnou. Není-li však statistická analysa takového experimentu dosti důmyslná, podstatně klesne i přesnost odhadů a síla testů. Jiné řešení spočívá ve využití náhradních proměnných, které jsou korelovány se skutečnými drahými proměnnými a jsou snáze dostupné a levněji změřitelné. Například drahý laboratorní test se dá nahradit testem jednodušším a levnějším, ale méně spolehlivým. I tak lze uspořit značné částky, ovšem vzniknou problémy s konsistencí odhadů a lze očekávat i pokles síly testů.

[1]Autor by rád poděkoval svému školiteli Dr. Danyu Linovi za trpělivost při vzájemné spolupráci na tomto projektu a Dr. Normu Breslowovi a National Wilms Tumor Study Group za laskavé poskytnutí datového souboru. Během práce na tomto článku byl autor podporován grantem GA ČR 201/98/P136.

Náš přístup kombinuje obě výše zmíněná řešení a snaží se vyhnout jejich nevýhodám. Předpokládáme, že skutečná proměnná (prediktor) je změřena pouze na validační skupině vybrané z celkové populace všech subjektů studie bernoulliovským výběrem (prostým náhodným výběrem o náhodném rozsahu). Zároveň ale předpokládáme existenci náhradního prediktoru, který můžeme chápat jako skutečný prediktor měřený s chybou, jenž je změřen na všech subjektech bez rozdílu. Vztah mezi skutečným a náhradním prediktorem je určen tzv. chybovým modelem, jenž vyžaduje, aby podmíněná střední hodnota skutečného prediktoru byla lineární funkcí náhradního prediktoru a podmíněný rozptyl nejvýše kvadratickou funkcí náhradního prediktoru.

Protože zanedbání chyb v prediktorech obvykle vede k vychýleným odhadům, je třeba v případě použití nepřesných prediktorů aplikovat speciální metody odhadu parametrů. V kontextu regrese pro censorovaná data byly takové metody zatím navrženy pouze pro Coxův model (Cox, 1972). Bohužel, všechny mají jistá omezení. Buď fungují pouze při silném censorování, kdy je jen málo událostí pozorováno (Prentice, 1982), nebo vyžadují, aby všechny prediktory byly diskrétní (Zhou a Pepe, 1995), nevedou obecně ke konsistentním odhadům (Wang a kol., 1996), anebo činí silné předpoklady o rozdělení prediktorů a chyb měření (Nakamura, 1992).

My náš problém nebudeme řešit v Coxově modelu, ale v modelu aditivního risika (Breslow a Day, 1987; Cox a Oakes, 1984; Lin a Ying, 1994). Model aditivního risika (AH model) je semiparametrický regresní model pro censorovaná data. Na rozdíl od Coxova modelu, jehož parametry lze interpretovat jako logaritmy relativního risika, parametry AH modelu udávají rozdíly risik, tj. rozdíly v počtu očekávaných událostí za jednotku času připadající na jednotkový rozdíl v prediktorech. Epidemiologové se často právě o tyto veličiny zajímají, protože mohou být důležitější pro porovnání různých exposicí nebo intervencí. AH model může být též užitečný v situacích, kdy jsou porušeny předpoklady Coxova modelu, ale také jen jako nástroj poskytující odlišný pohled na strukturu analysovaných dat.

Pro odvození konsistentního odhadu budeme používat metodu upravené skórové funkce (CS metoda). Tato metoda byla v minulosti použita při řešení analogických problémů v normální, Poissonově a gama regresi (Carroll, Ruppert a Stefanski, 1995, Kap. 6), a též v Coxově modelu (Nakamura, 1992). Jenže Coxův model je poněkud nevhodný k aplikaci CS metody, protože skórová funkce částečné věrohodnosti závisí na prediktorech příliš komplikovaným způsobem. Proto jsou zapotřebí silné předpoklady o rozdělení chyb prediktorů a výsledný odhad je i tak konsistentní pouze přibližně. Naproti tomu pseudoskórová funkce AH modelu má mnohem jednodušší tvar, jenž dovoluje přímočarou aplikaci CS metody za minimálních předpokladů.

V tomto článku definujeme upravenou pseudoskórovou funkci pro model aditivního risika s jedním prediktorem za předpokladu existence validační skupiny. Odhad regresního parametru definovaný touto pseudoskórovou funkcí je konsistentní, funguje pro širokou škálu možných chybových rozdělení, diskrétních i spojitých, a snadno se spočítá. Odvodíme zde asymptotické rozdělení navrženého odhadu a pomocí simulační studie prozkoumáme jeho chování při konečných výběrech o rozsahu obvyklém v praxi. Na závěr ukážeme příklad aplikace odhadu na reálná data.

2. Model aditivního risika

Nechť T je doba do události a C doba do censorování. Označme $X = \min(T, C)$ a $\Delta = \mathbb{1}(T \leq C)$, kde $\mathbb{1}(A)$ je indikátor události A. Nechť $\boldsymbol{Z}(\cdot)$ je p-vektor prediktorů, které mohou záviset na čase. Model aditivního risika (AH model) váže risikovou funkci doby do události T (intensitu) k vektoru prediktorů $\boldsymbol{Z}(\cdot)$ vztahem

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{Z}(t),$$

kde základní intensita $\lambda_0(t)$ je pevná nespecifikovaná funkce a $\boldsymbol{\beta}_0$ je neznámý p-vektor parametrů nezávislých na čase, které je třeba odhadnout. Označme kumulativní základní intensitu $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$.

Pozorovaná data nechť jsou dána iid trojicemi $(X_i, \Delta_i, \boldsymbol{Z}_i(t), 0 \leq t \leq X_i)$, $i = 1, \ldots, n$. Označme $N_i(t) = \mathbb{1}(X_i \leq t, \Delta_i = 1)$ čítací proces, který udává počet událostí pozorovaných u i-tého subjektu do času t. Nechť $Y_i(t) = \mathbb{1}(X_i \geq t)$ indikuje, zdali byl i-tý subjekt v čase t pozorován. Budeme předpokládat, že pozorování je ukončeno nejpozději v okamžiku $\tau < \infty$, pro nějž platí $\mathrm{P}\left[Y_i(\tau) = 1\right] > 0$. Označme si ještě $\overline{\boldsymbol{Z}}(t)$ průměrný prediktor subjektů, kteří byli v čase t pozorováni, tj. $\overline{\boldsymbol{Z}}(t) = \sum_{i=1}^n Y_i(t) \boldsymbol{Z}_i(t) \big/ \sum_{i=1}^n Y_i(t)$.

Lin a Ying (1994) definují pseudoskórovou funkci pro $\boldsymbol{\beta}$ následovně:

$$U_A(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]\, [dN_i(t) - \boldsymbol{Z}_i(t)^{\mathrm{T}} \boldsymbol{\beta} Y_i(t)\, dt].$$

Protože $\sum_{i=1}^n [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)] Y_i(t) = \mathbf{0}$, hodnotu U_A vyčíslenou ve skutečném parametru lze psát jako

$$U_A(\boldsymbol{\beta}_0) = \sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]\, dM_i(t),$$

kde $M_i(t) = N_i(t) - \int_0^t Y_i(s)\, d\Lambda_0(s) - \int_0^t Y_i(s) \boldsymbol{Z}_i(s)^{\mathrm{T}} \boldsymbol{\beta}_0\, ds$ je martingal čítacího procesu $N_i(\cdot)$. Tudíž $U_A(\boldsymbol{\beta}_0)$ je martingalový integrál.

Pseudoskórová funkce U_A definuje odhad parametru jako řešení rovnice $U_A(\boldsymbol{\beta}) = \mathbf{0}$. Označíme-li si tento odhad $\widehat{\boldsymbol{\beta}}_A$ (nazvěme jej AH odhad), dostaneme snadnou úpravou, že

$$\widehat{\boldsymbol{\beta}}_A = \frac{\sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]\, dN_i(t)}{\sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]^{\otimes 2} Y_i(t)\, dt}, \tag{1}$$

Lin a Ying ukázali, že náhodný vektor $n^{-1/2}U_A(\boldsymbol{\beta}_0)$ za podmínek regularity konverguje v distribuci k p-rozměrnému normálnímu rozdělení s nulovou střední hodnotou a varianční maticí $\Sigma_A(\boldsymbol{\beta}_0) = \mathrm{E} \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]^{\otimes 2}\, dN_i(t)$. Matice parciálních derivací $-n^{-1}U_A(\boldsymbol{\beta})$ podle $\boldsymbol{\beta}$ konverguje v pravděpodobnosti k $D_A = \mathrm{E} \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]^{\otimes 2} Y_i(t)\, dt$. Odtud plyne, že $\sqrt{n}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)$ konverguje v distribuci k p-rozměrnému normálnímu rozdělení s nulovou střední hodnotou a varianční maticí $D_A^{-1}\Sigma_A(\boldsymbol{\beta}_0)D_A^{-1}$.

3. Upravená pseudoskórová funkce

3.1. **Definice.** Nechť Z_i je skutečný a W_i náhradní prediktor. Nadále předpokládáme, že Z_i a W_i jsou jednorozměrné a nezávisí na čase, ale v závěrečné diskusi naznačíme, jak postupovat ve vícerozměrných případech. Validační skupinu budeme definovat pomocí nezávislých indikátorů výběru ξ_i s rozdělením $\mathrm{P}[\xi_i = 1] = 1 - \mathrm{P}[\xi_i = 0] = \alpha$, kde α je známá konstanta. Zavedeme značení $\overline{\xi}_i$ pro $1 - \xi_i$ a $\overline{\alpha}$ pro $1 - \alpha$. Předpokládejme dále, že ξ_i je nezávislé na Z_i, W_i, Δ_i a X_i. Je-li $\xi_i = 1$, pozorujeme $(X_i, \Delta_i, Z_i, W_i, \xi_i)$, v opačném případě pozorujeme pouze $(X_i, \Delta_i, W_i, \xi_i)$.

Budeme vyžadovat, aby náhradní prediktor W_i splňoval tyto podmínky:

(1) $\mathrm{E}[W_i \mid Z_i] = \gamma_0 + \gamma_1 Z_i$, kde $\gamma_1 \neq 0$;
(2) $\mathrm{var}[W_i \mid Z_i] = V(Z_i) = v_0 + v_1 Z_i + v_2 Z_i^2$;
(3) W_i a W_j jsou podmíněně nezávislé, je-li dáno Z_i a Z_j;
(4) W_i je podmíněně nezávislé na $N_i(t)$ a $Y_i(t)$, je-li dáno Z_i.

První dvě podmínky vyjadřují první dva podmíněné momenty W_i, je-li dáno Z_i a tím udávají model pro rozdělení chyb. Žádné jiné předpoklady o rozdělení chyb nebo prediktorů nebudeme zavádět. Rozptylová funkce $V(\cdot)$ v Podmínce 2 může být libovolná nenulová konstantní, lineární, či kvadratická funkce. Proč je toto omezení nutné, vysvětlíme později. Podmínka 3 zajišťuje nezávislost chyb. Podmínka 4 znamená, že všechen vliv W na dobu do události a do censorování musí být zprostředkován skrze Z. Parametery γ_0, γ_1, v_0, v_1 a v_2, které jsme zavedli v podmínkách 1 a 2 budeme nazývat *chybové parametry.*

Předpokládejme nejdříve, že chybové parametry jsou známy. Definujme upravený náhradní prediktor $\widehat{Z}_i = \gamma_1^{-1}(W_i - \gamma_0)$. Očividně $\mathrm{E}\left[\widehat{Z}_i \,\middle|\, Z_i\right] = Z_i$ a $\mathrm{var}\left[\widehat{Z}_i \,\middle|\, Z_i\right] = \gamma_1^{-2}V(Z_i)$. Vezmeme-li pseudoskóre U_A a dosadíme-li do něj hodnoty $\widehat{Z}_i$ za neznámá Z_i v nevalidační skupině, dostaneme naivní odhad β_0. Označíme-li $R_i = \xi_i Z_i + \overline{\xi}_i \widehat{Z}_i$, naivní odhad $\widehat{\beta}_N$ je definován jako řešení rovnice $U_N(\beta) = 0$, kde

$$U_N(\beta) = \sum_{i=1}^{n} \int_0^\tau [R_i - \bar{R}(t)]\,[dN_i(t) - R_i\beta Y_i(t)\,dt] \tag{2}$$

a $\bar{R}(t) = \sum Y_i(t)R_i \,/\sum Y_i(t)\,$.

Ačkoli $\widehat{Z}_i$ je nestranný odhad Z_i ve smyslu předchozího odstavce, naivní odhad není konsistentní, protože $\mathrm{E}\,U_N(\beta_0) \neq 0$. Abychom našli konsistentní odhad, spočteme $\mathrm{E}\,U_N(\beta_0)$ podmíněně na veškerou informaci o událostech, censorování a skutečných prediktorech. Konkrétně, označme si $\sigma(Z,N,Y)$ σ-algebru generovanou $(Z_i, X_i, \Delta_i; i = 1,\ldots,n)$. Lze snadno ukázat, že

$$\mathrm{E}\,[U_N(\beta) \mid \sigma(Z,N,Y)\,] = U_A(\beta) - \gamma_1^{-2}\beta \sum_{i=1}^{n} \overline{\xi}_i V(Z_i)X_i + o_P(n^{1/2}).$$

Člen $-\gamma_1^{-2}\beta\sum_{i=1}^{n}\overline{\xi}_i V(Z_i)X_i$ má nenulovou střední hodnotu. To on tedy způsobuje výchylku naivního odhadu. Zdá se, že lepší by bylo založit odhad na pseudoskóru, které dostaneme odečtením tohoto členu od $U_N(\beta)$. Tím ale narážíme na problém: $V(Z_i)$ závisí na neznámém Z_i. V tomto okamžiku se nám hodí Podmínka 2. Pokud totiž $V(x) = v_0 + v_1x + v_2x^2$, můžeme snadno ukázat, že $\mathrm{E}\left[V(\widehat{Z}_i) \,\middle|\, Z_i\right] = (1 + \gamma_1^{-2}v_2)V(Z_i)$. Tato rovnost nám dává návod, jak nestranně odhadnout $V(Z_i)$.

Upravenou pseudoskórovou funkci formálně definujeme jako kombinaci příspěvků validačních a nevalidačních pozorování. Nejprve však zaveďme váhu w, $0 \le w \le 1$, která sníží vliv nevalidačních pozorování na pseudoskórovou funkci. Průměrný pozorovaný prediktor potom odhadneme jako

$$\bar{R}(t,w) = \frac{\sum(\xi_i Z_i + w\overline{\xi}_i\widehat{Z}_i)Y_i(t)}{\sum(\xi_i + w\overline{\xi}_i)Y_i(t)}. \tag{3}$$

Za mírných podmínek regularity existuje deterministická funkce $e(t)$ taková, že $\sup|\bar{Z}(t)-e(t)| \to_p 0$. Lze ukázat, že potom též platí $\sup|\bar{R}(t,w)-e(t)| \to_p 0$ pro libovolné $0 \le w \le 1$. Argument w budeme obvykle vypouštět jak z $\bar{R}(t,w)$ tak z ostatních veličin závisejících na w.

Příspěvek validačního pozorování do pseudoskórové funkce nechť je dán jako

$$\psi_i^{(V)}(\beta) = \int_0^\tau [Z_i - \bar{R}(t)]\,[dN_i(t) - Z_i\beta Y_i(t)\,dt].$$

Nevalidační pozorování budou přispívat

$$\psi_i^{(NV)}(\beta) = \int_0^\tau [\widehat{Z}_i - \bar{R}(t)]\,[dN_i(t) - \widehat{Z}_i\beta Y_i(t)\,dt] + \frac{V(\widehat{Z}_i)}{\gamma_1^2 + v_2}\beta X_i,$$

kde poslední sčítanec vpravo je motivován předchozí heuristickou úvahou o střední hodnotě $U_N(\beta_0)$. Tyto dvě části zkombinujeme do upravené pseudoskórové funkce takto:

$$U_C(\beta, w) = \sum_{i=1}^n \left[\xi_i\psi_i^{(V)}(\beta) + w\overline{\xi}_i\psi_i^{(NV)}(\beta)\right].$$

Rovnice $U_C(\beta, w) = 0$ definuje odhad metodou upravené pseudoskórové funkce (CS odhad), který si označíme $\widehat{\beta}_C(w)$. Podobně jako AH odhad, i CS odhad má explicitní tvar, takže se při jeho výpočtu obejdeme bez iterativních procedur. Je-li $w = 0$, nevalidační data nejsou pro výpočet odhadu vůbec používána a CS odhad se redukuje na AH odhad spočtený z validační skupiny.

3.2. **Asymptotické rozdělení CS odhadu.** Asymptotické vlastnosti odhadu $\widehat{\beta}_C(w)$ odvodíme aproximováním $U_C(\beta_0, w)$ součtem nezávislých veličin. Samotné $U_C(\beta_0, w)$ toto nesplňuje, protože jeho jednotlivé sčítance závisejí na $\bar{R}(t)$. Za mírných podmínek na konečnost $\Lambda_0(\tau)$ a čtvrtých momentů Z a W lze ukázat, že

$$\frac{1}{\sqrt{n}}U_C(\beta_0, w) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\xi_i\widetilde{\psi}_i^{(V)}(\beta_0) + w\overline{\xi}_i\widetilde{\psi}_i^{(NV)}(\beta_0)\right] + o_P(1),$$

kde

$$\widetilde{\psi}_i^{(V)}(\beta_0) = \int_0^\tau [Z_i - e(t)]\,dM_i(t)$$

a

$$\widetilde{\psi}_i^{(NV)}(\beta_0) = \int_0^\tau [\widehat{Z}_i - e(t)]\,[dN_i(t) - Y_i(t)\,d\Lambda_0(t) - \widehat{Z}_i\beta_0 Y_i(t)\,dt] + \beta_0\frac{V(\widehat{Z}_i)}{\gamma_1^2 + v_2}X_i. \tag{4}$$

Odtud ihned plyne, že $n^{-1/2}U_C(\beta_0, w)$ konverguje v distribuci k normálnímu rozdělení s nulovou střední hodnotou a rozptylem

$$\Sigma_C(\beta_0, w) = \alpha\Sigma_A(\beta_0) + \overline{\alpha}w^2\mathrm{E}\left[\widetilde{\psi}_i^{(NV)}(\beta_0)\right]^2$$

a že $\sqrt{n}(\widehat{\beta}_C - \beta_0)$ je asymptoticky normální s nulovou střední hodnotou a rozptylem $\Sigma_C(\beta_0)/D_C^2$, kde $D_C = (\alpha + \overline{\alpha}w)D_A$ je očekávaná derivace

$n^{-1}U_C(\beta)$ s opačným znaménkem. Tuto veličinu můžeme odhadnout pomocí

$$\widehat{D}_C = \frac{1}{n}\sum_{i=1}^{n}\biggl(\xi_i \int_0^{\tau} [Z_i - \bar{R}(t)]^2 Y_i(t)\,dt + \overline{\xi}_i w \int_0^{\tau} \biggl\{[\widehat{Z}_i - \bar{R}(t)]^2 - \frac{V(\widehat{Z}_i)}{\gamma_1^2 + v_2}\biggr\} Y_i(t)\,dt\biggr).$$

Pro nalezení konsistentního odhadu $\Sigma_C(\beta_0, w)$ potřebujeme odhadnout kumulativní základní intensitu. Můžeme použít například odhad $d\widehat{\Lambda}_0(t) = dA(t) - \bar{R}(t)\widehat{\beta}_C\,dt$, kde $dA(t) = \sum(\xi_i + \overline{\xi}_i w)\,dN_i(t)/\sum(\xi_i + \overline{\xi}_i w)Y_i(t)$. Lze ukázat, že $\sup|\Lambda_0(t) - \widehat{\Lambda}_0(t)| \to_p 0$. Nyní tedy odhadneme $\Sigma_C(\beta_0, w)$ jako

$$\begin{aligned}\widehat{\Sigma}_C(\beta, w) \;=\; & \frac{1}{n}\sum_{i=1}^{n}\xi_i \int_0^{\tau} [Z_i - \bar{R}(t)]^2\,dN_i(t) + \frac{w^2}{n}\sum_{i=1}^{n}\overline{\xi}_i \biggl\{\int_0^{\tau} [\widehat{Z}_i - \bar{R}(t)] \\ & \times \Bigl[dN_i(t) - Y_i(t)\,dA(t) - [\widehat{Z}_i - \bar{R}(t)]\beta Y_i(t)\,dt\Bigr] + \beta\frac{V(\widehat{Z}_i)X_i}{\gamma_1^2 + v_2}\biggr\}^2,\end{aligned}$$

vyčíslené v $\beta = \widehat{\beta}_C$.

CS odhad je konsistentní pro libovolnou váhu $0 \le w \le 1$. Limitní rozptyl výrazu $\sqrt{n}(\widehat{\beta}_C(w) - \beta_0)$ je nicméně funkcí w. Není těžké ukázat, že tato funkce nabývá svého minima v bodě $w_{\text{opt}} \equiv \Sigma_A(\beta_0)/E[\widetilde{\psi}_i^{(NV)}]^2$. Je zřejmé, že $w_{\text{opt}} \le 1$. Váha w_{opt} definuje odhad, který je eficientní mezi všemi odhady danými vahami $0 \le w \le 1$. To znamená, že CS odhad s optimální váhou nemá nikdy větší rozptyl než odhad založený na validační skupině ($w = 0$). Optimální váhu w_{opt} můžeme odhadnout jako

$$\begin{aligned}\widehat{w}_{\text{opt}} \;=\; & \frac{1}{\alpha}\sum \xi_i \int_0^{\tau} [Z_i - \bar{R}(t)]^2\,dN_i(t) \Big/ \frac{1}{\overline{\alpha}}\sum \overline{\xi}_i \biggl\{\int_0^{\tau} [\widehat{Z}_i - \bar{R}(t)] \\ & \times \Bigl[dN_i(t) - Y_i(t)\,d\hat{\Lambda}_0(t) - \widehat{Z}_i\widehat{\beta}_C Y_i(t)\,dt\Bigr] + \widehat{\beta}_C\frac{V(\widehat{Z}_i)X_i}{\gamma_1^2 + v_2}\biggr\}^2.\end{aligned} \tag{5}$$

Protože $\widehat{w}_{\text{opt}}$ závisí na $\widehat{\beta}_C$, můžeme odhadnout oba parametry jednoduchou iterativní procedurou nebo spočteme pracovní odhad β_0 s vahou $w = 0$, dosadíme ho do výrazu pro $\widehat{w}_{\text{opt}}$, a výsledek použijeme pro výpočet konečného $\widehat{\beta}_C$. Druhý postup je možná rozumnější.

3.3. CS odhad při neznámých chybových parametrech. Nyní konečně odstraníme předpoklad, že chybové parametry jsou známy. Nechť tedy $\boldsymbol{\theta}$ značí sloupcový vektor chybových parametrů (je jich nejvýše pět) a nechť $\boldsymbol{\theta}_0$ je skutečná hodnota $\boldsymbol{\theta}$. Jsou-li chybové parametry neznámé, mohou být jednoduše odhadnuty z validační množiny jakýmkoli konsistentním odhadem a dosazeny do pseudoskórové funkce. Konsistence CS odhadu tím nebude porušena, ale asymptotický rozptyl se změní.

Ve všech uvažovaných situacích je možné chybové parametry konsistentně odhadnout kvazivěrohodnostní metodou. V nejobecnějším uvažovaném případě tak bude odhad $\widehat{\boldsymbol{\theta}}$ chybových parametrů řešením rovnice $\sum \xi_i \phi_i(\boldsymbol{\theta}) = \mathbf{0}$, kde ϕ_i je sloupcový vektor složený z $\phi_i^{(\gamma)}$ a $\phi_i^{(v)}$ definovaných jako

$$\phi_i^{(\gamma)}(\boldsymbol{\theta}) = \frac{W_i - \gamma_0 - \gamma_1 Z_i}{V(Z_i)} \, (1,\, Z_i)^{\mathrm{T}},$$

$$\phi_i^{(v)}(\boldsymbol{\theta}) = \frac{(W_i - \gamma_0 - \gamma_1 Z_i)^2 - V(Z_i)}{V^2(Z_i)} \, (1,\, Z_i,\, Z_i^2)^{\mathrm{T}}.$$

Kvazivěrohodnostní odhad je samozřejmě možné nahradit jakýmkoli jiným konsistentním odhadem, například odhadem metodou nejmenších čtverců, jedná-li se o lineární regresi s konstantním rozptylem.

Neznámé $\boldsymbol{\theta}_0$ tedy nahradíme konsistentním odhadem $\widehat{\boldsymbol{\theta}}$ a řešíme rovnici $U_C(\beta, \widehat{\boldsymbol{\theta}}) = 0$. Výsledný odhad, jenž značíme $\widehat{\beta}_C$ bez ohledu na to, je-li $\boldsymbol{\theta}_0$ známo či nikoli, zůstává konsistentní. Asymptotický rozptyl $n^{-1/2} U_C(\beta, \widehat{\boldsymbol{\theta}})$ dostaneme rozvinutím $U_C(\beta, \widehat{\boldsymbol{\theta}}) - U_C(\beta, \boldsymbol{\theta}_0)$ v Taylorovu řadu. Ukáže se, že

$$U_C(\beta_0, \widehat{\boldsymbol{\theta}}) \approx \sum \left[\xi_i \widetilde{\psi}_i^{(V)}(\beta_0) + w \overline{\xi}_i \widetilde{\psi}_i^{(NV)}(\beta_0, \boldsymbol{\theta}_0) - \frac{\overline{\alpha}}{\alpha} w \xi_i \phi_i(\theta_0)^{\mathrm{T}} \Gamma(\beta_0, \boldsymbol{\theta}_0) \right],$$

s chybou aproximace řádu $o_P(n^{1/2})$. Korekční vektor $\Gamma(\beta_0, \boldsymbol{\theta}_0)$ je definován vztahem $\Gamma^{\mathrm{T}} = D_2(\beta_0, \boldsymbol{\theta}_0) D_1^{-1}(\boldsymbol{\theta}_0)$, kde

$$D_1(\boldsymbol{\theta}) = -\mathrm{E} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \phi_i(\boldsymbol{\theta}) \quad \text{a} \quad D_2(\beta, \boldsymbol{\theta}) = -\mathrm{E} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \widetilde{\psi}_i^{(NV)}(\beta, \boldsymbol{\theta}).$$

Z toho plyne, že limitní rozptyl $\sqrt{n}(\widehat{\beta}_C - \beta_0)$ je Σ_{CN}/D_C^2, kde

$$\Sigma_{CN} = \mathrm{E} \left(\xi_i \widetilde{\psi}_i^{(V)} + w \overline{\xi}_i \widetilde{\psi}_i^{(NV)} - \frac{\overline{\alpha}}{\alpha} w \xi_i \Gamma^{\mathrm{T}} \phi_i \right)^2.$$

Vzorec pro Σ_{CN} naznačuje, jak spočítat odhad rozptylu CS odhadu v případě neznámých chybových parametrů. Konkrétní tvar korekčního vektoru Γ se liší od případu k případu, ale není těžké jej odvodit pro každou konkrétní aplikaci zvlášť.

4. Speciální případy

4.1. **Spojitý prediktor měřený s chybou o konstantním rozptylu.** Toto je nejjednodušší praktická aplikace metody upravené skórové funkce. Nechť $W_i = Z_i + \varepsilon_i$, kde $\varepsilon_1, \ldots, \varepsilon_n$ jsou náhodné veličiny s nulovou střední hodnotou, vzájemně nezávislé a nezávislé na Z_i, N_i a Y_i. Budiž $\operatorname{var} \varepsilon_i = \sigma^2$, kde σ^2 je neznámé. Označme $\varrho_i = \xi_i + \overline{\xi}_i w$ a $R_i = \xi_i Z_i + \overline{\xi}_i W_i$. CS odhad pak má tvar

$$\widehat{\beta}_C = \frac{\sum \int_0^\tau \varrho_i [R_i - \bar{R}(t)]\, dN_i(t)}{\sum \int_0^\tau \varrho_i [R_i - \bar{R}(t)]^2 Y_i(t)\, dt + w\widehat{\sigma}^2 \sum \overline{\xi}_i X_i},$$

kde $\widehat{\sigma}^2 = \sum \xi_i (W_i - Z_i)^2 / \sum \xi_i$ je řešením $\sum \xi_i \phi_i = 0$ s $\phi_i = (W_i - Z_i)^2 - \sigma^2$. Korekční faktor Γ je prostě $-\beta_0 \mathrm{E}\, X_i$.

Chování CS odhadu v této situaci jsme zkoumali pomocí simulační studie. Za populační rozdělení prediktoru Z jsme vzali $\mathrm{N}(0,1)$ useknuté v ± 1.96. Chyba měření ε měla rozdělení $\mathrm{N}(0,\sigma^2)$. Základní intensita byla konstantní a censorování bylo rovnoměrné na jistém intervalu. Jeden datový soubor se sestával z 1000 pozorování, z nichž asi 55% bylo censorováno. Pro každou situaci jsme generovali 1000 simulací a z každé jsme spočítali čtyři odhady parametru AH modelu: odhad z úplných dat daný výrazem (1); odhad založený na validační skupině ($w = 0$); naivní odhad definovaný v (2); a CS odhad $\widehat{\beta}_C(\widehat{w}_{\mathrm{opt}})$, kde $\widehat{w}_{\mathrm{opt}}$ bylo spočteno podle (5) za použití odhadu s $w = 0$ a chybový rozptyl σ^2 byl odhadnut z validačních dat. Výsledky simulací jsou shrnuty v Tabulce 1.

Tabulka ukazuje chování čtyř uvažovaných odhadů při různých velikostech validační skupiny a různých rozptylech chyby měření. Rozptyl chyb měření se pohyboval od čtvrtiny do čtyřnásobku populačního rozptylu skutečného prediktoru. Validační skupina obsahovala pětinu až polovinu všech pozorování. Podle očekávání v tabulce nacházíme podstatné vychýlení u naivního odhadu. CS odhad toto vychýlení úspěšně odstranil. Výběrová směrodatná odchylka CS odhadu byla vždy menší než u odhadu z validační skupiny, a to i v případě, že rozptyl chyb měření byl velký. Odhadnutá směrodatná odchylka CS odhadu byla v průměru dostatečně blízko výběrové směrodatné odchylce, což svědčí o tom, že odhad rozptylu rozumně funguje. Pravděpodobnost pokrytí skutečného parametru 95% intervalem spolehlivosti založeným na CS odhadu a jeho odhadnutém rozptylu se pohybovala od 0.945 do 0.961, což je uspokojivé.

4.2. **Nula-jedničkový prediktor.** Nechť Z je veličina nabývající hodnot 0 a 1 s $\mathrm{P}[Z = 1] = p_Z$ a W nechť je nula-jedničkový náhradní prediktor pro Z. Označme sensitivitu W pro Z jako $\eta \equiv \mathrm{P}[W = 1 \mid Z = 1]$ a specificitu jako $\nu \equiv \mathrm{P}[W = 0 \mid Z = 0]$. Podmíněná střední hodnota W za podmínky Z je $\mathrm{E}[W \mid Z] = 1-\nu+(\eta+\nu-1)Z$. $\mathrm{E}[W \mid Z]$ má tudíž požadovaný lineární tvar s parametry $\gamma_0 = 1-\nu$ a $\gamma_1 = \eta+\nu-1$. Předpokládáme, že $\gamma_1 \neq 0$. Podmíněný rozptyl W za podmínky Z je $V(Z) = \mathrm{var}[W \mid Z] = \nu(1-\nu) + [\eta(1-\eta) - \nu(1-\nu)]Z$, což je též lineární funkce Z. Protože všechny chybové parametry jsou funkcí sensitivity a specificity, budeme problém parametrisovat pomocí η a ν. Bude ale užitečné nadále psát γ_1 pro $\eta + \nu - 1$.

TABULKA 1. Simulační studie spojitého prediktoru s náhodnou chybou v modelu $\lambda(t \mid Z) = 3.4 + \beta_0 Z$, kde $\beta_0 = 0.3$.

Metoda[a]	α	σ	Prům. Odh.	Výběr. SE	Prům. $\widehat{\text{SE}}$	95% pokr.	Síla
F	*	*	0.300	0.157	0.162	0.960	0.468
CC	0.2	*	0.300	0.361	0.370	0.956	0.124
	0.5	*	0.296	0.232	0.230	0.943	0.256
N	0.2	0.5	0.245	0.142	0.148	0.939	0.378
		1	0.167	0.115	0.120	0.809	0.269
		2	0.071	0.075	0.078	0.166	0.143
	0.5	0.5	0.262	0.152	0.153	0.953	0.416
		1	0.190	0.131	0.132	0.864	0.293
		2	0.104	0.094	0.093	0.442	0.194
CS	0.2	0.5	0.299	0.173	0.178	0.961	0.398
		1	0.312	0.212	0.217	0.955	0.300
		2	0.315	0.291	0.314	0.961	0.148
	0.5	0.5	0.298	0.171	0.172	0.953	0.424
		1	0.290	0.194	0.189	0.946	0.348
		2	0.312	0.213	0.214	0.945	0.296

[a] F = úplná data, CC = valid. skup., N = naivní, CS = CS odhad.
* Odhad nezávisí na tomto parametru.

Upravený náhradní prediktor pro W je dán jako $\widehat{Z} = \gamma_1^{-1}(W - 1 + \nu)$. Nestranný odhad $V(Z)$ je $V(\widehat{Z}) = \nu(1-\nu) + [\eta(1-\eta) - \nu(1-\nu)]\widehat{Z}$. Nevalidační příspěvek do pseudoskórové funkce má tvar

$$\psi_i^{(NV)}(\beta) = \int_0^\tau [\widehat{Z}_i - \bar{R}(t)] \{dN_i(t) - \widehat{Z}_i \beta Y_i(t)\, dt\} + \beta \gamma_1^{-2} X_i V(\widehat{Z}_i).$$

Chybové parametry $\boldsymbol{\theta} = (\eta, \nu)^{\mathrm{T}}$ odhadneme jako $\widehat{\eta} = n_{11}/(n_{10} + n_{11})$ a $\widehat{\nu} = n_{00}/(n_{00}+n_{01})$, kde $n_{ij} = \sum_k \xi_k \mathbb{1}(Z_k = i, W_k = j)$. To znamená, že $\widehat{\boldsymbol{\theta}}$ je řešením $\sum \xi_i \phi_i(\boldsymbol{\theta}) = 0$ s $\phi_i(\boldsymbol{\theta}) = (Z_i W_i - \eta Z_i,\ (1 - Z_i)(1 - W_i) - \nu(1 - Z_i))^{\mathrm{T}}$.

Jednoduché výpočty ukazují, že

$$\begin{aligned}\Gamma^{\mathrm{T}} \phi_i(\boldsymbol{\theta}) \;=\; & \frac{\beta_0}{\eta + \nu - 1} \left\{ p_Z^{-1} \left[\int_0^\tau e(t)\bar{e}(t)\pi(t)\, dt - \mathrm{E}\, Z_i X_i \right] Z_i(\eta - W_i) \right. \\ & \left. + (1 - p_Z)^{-1} \left[\int_0^\tau e(t)\bar{e}(t)\pi(t)\, dt - \mathrm{E}\,(1 - Z_i) X_i \right] (1 - Z_i)[\nu - (1 - W_i)] \right\},\end{aligned}$$

TABULKA 2. Simulační studie nula-jedničkového binárního prediktoru v modelu $\lambda(t \mid Z) = 2.6 + \beta_0 Z$, kde $\beta_0 = 0.9$.

Metoda[a]	α	η	ν	Prům. Odh.	Výběr. SE	Prům. $\widehat{\text{SE}}$	95% pokrytí	Síla
F	*	*	*	0.897	0.314	0.299	0.937	0.842
CC	0.2	*	*	0.911	0.689	0.679	0.950	0.272
	0.5	*	*	0.911	0.444	0.425	0.944	0.576
N	0.2	0.7	0.7	0.178	0.131	0.130	0.000	0.288
		0.9	0.7	0.390	0.195	0.190	0.241	0.539
		0.9	0.9	0.616	0.253	0.247	0.780	0.707
	0.5	0.7	0.7	0.245	0.161	0.156	0.015	0.366
		0.9	0.7	0.481	0.231	0.216	0.512	0.598
		0.9	0.9	0.708	0.268	0.263	0.868	0.764
CS	0.2	0.7	0.7	0.940	0.554	0.537	0.944	0.411
		0.9	0.7	0.932	0.442	0.435	0.957	0.583
		0.9	0.9	0.900	0.371	0.360	0.946	0.711
	0.5	0.7	0.7	0.912	0.408	0.395	0.938	0.646
		0.9	0.7	0.902	0.387	0.362	0.934	0.693
		0.9	0.9	0.908	0.343	0.331	0.946	0.775

[a] F = úplná data, CC = valid. skup., N = naivní, CS = CS odhad.
* Odhad nezávisí na tomto parametru.

kde $\bar{e}(t) = 1 - e(t)$ a $\pi(t) = \mathrm{E}\, Y_i(t)$. Všechny tyto výrazy lze snadno odhadnout.

Chování CS odhadu při konečných výběrech jsme i v tomto případě zkoumali pomocí simulační studie. Generovali jsme 1000 datových souborů o 1000 pozorováních. Základní intensita byla i tentokrát konstantní a censorování rovnoměrné na takovém intervalu, že bylo zcensorováno přibližně 42% pozorování. Pravděpodobnost p_Z byla 0.5. Validační skupina zahrnovala buď 20% nebo 50% všech pozorování. Počítali jsme tytéž čtyři odhady jako v předchozí simulaci na chyby ve spojitém prediktoru. Výsledky jsou shrnuty v Tabulce 2.

Tabulka 2 uvádí výsledky pro různě velké chyby měření: uvažované dvojice sensitivity a specificity (0.7,0.7), (0.9,0,7) a (0.9,0.9) dávají chybné W po řadě u 30%, 20% a 10% pozorování. Tabulka naznačuje, že CS odhad úplně odstraňuje výchylku naivního odhadu. Směrodatná odchylka CS odhadu je vždycky menší než směrodatná odchylka odhadu založeného na validační skupině, a to i tehdy, je-li chyba měření velká a validační množina malá.

Pravděpodobnosti pokrytí skutečného parametru 95% intervalem spolehlivosti pro CS odhad se pohybují od 0.937 do 0.957, což naznačuje, že jak CS odhad tak odhad jeho rozptylu i v tomto případě pracují spolehlivě.

5. Příklad: Wilms Tumor Study Data

Jako příklad praktické aplikace CS odhadu uvádíme analysu doby do znovuobjevení nádoru u pacientů sledovaných v rámci dvou studií prováděných výzkumným centrem National Wilms Tumor Study Group v USA. Tyto studie nesou označení NWTSG-3 a NWTSG-4. NWTSG se zabývá zkoumáním Wilmsova nádoru, vzácného druhu rakoviny ledvin vyskytujícího se pouze u dětí. Wilmsův nádor má nyní naštěstí velmi dobrou prognózu, pokud je léčen vhodnou kombinační chemoterapií. Výsledky studie NWTSG-3 byly již publikovány v pracích D'Angio a kol. (1989) a Breslow a kol. (1991); článek o NWTSG-4 se objeví v blízké budoucnosti (Green a kol., 1997).

Jeden z nejdůležitějších prognostických faktorů pro dobu do rekurence u pacientů s Wilmsovým nádorem je histologický typ nádoru. Ten lze zhruba klasifikovat jako příznivý anebo nepříznivý. V NWTSG studiích byl histologický typ nejprve zjištěn místním patologem v té nemocnici, která daného pacienta léčila. Patologické centrum NWTSG potom histologický typ překlasifikovalo podle zaslaných vzorků tkáně. Tato dvě měření, místní a centrální, se shodovala u většiny pacientů, ale zdaleka ne u všech. Centrální měření lze přitom považovat za přesnější a spolehlivější než měření místní. Proto při analyse vlivu histologického typu na dobu do rekurence budeme považovat centrální měření za skutečný prediktor a místní měření za prediktor náhradní. Ačkoli v NWTSG studiích byla obě měření k disposici, je zajímavé se podívat, jak by vypadala statistická analysa, kdyby Patologické centrum vyhodnocovalo jen náhodný vzorek ze všech pacientů. Pak by totiž cenu i komplexnost těchto studií bylo možné podstatně snížit.

Naše data obsahují pozorování o 4119 pacientech (1911 v NWTSG-3 a 2208 v NWTSG-4) se známým rekurenčním stavem, dobou do ukončení sledování a oběma histologickými měřeními. Z nich mělo 11.5% nepříznivou centrální histologii (11.6% v NWTSG-3 a 11.5% v NWTSG-4) a 14.5% mělo rekurenci (15.6% v NWTSG-3 a 13.5% v NWTSG-4). Sensitivita a specificita místní histologie pro centrální byla po řadě 0.72 a 0.98. Předběžná analysa potvrdila, že při daném výsledku centrální histologie nemělo místní měření žádný vliv na pravděpodobnost rekurence.

Nejprve jsme spočetli standardní AH odhad z celých dat s použitím jednak centrální a jednak místní histologie jako jediného prediktoru. Poté jsme náhodně vybírali validační skupiny a počítali jsme CS odhad a AH odhad předpokládajíce, že centrální histologie je známa pouze u validačních pozorování. Výsledky jsou uvedeny v Tabulce 3. CS odhad, AH odhad z validační skupiny a jejich směrodatné odchylky jsou dány jako průměry přes 500 simulovaných validačních skupin. AH odhad pro centrální histologii z celých dat

TABULKA 3. Analysa NWTSG dat: odhady rozdílu risika rekurence při nepříznivé centrální histologii.

Metoda[a]	α	Odhad	SE
F	1	0.0744	0.00683
F	0^b	0.0557	0.00609
CC	0.2	0.0755[c]	0.01555[c]
CC	0.5	0.0746[c]	0.00969[c]
CS	0.2	0.0768[c]	0.01058[c]
CS	0.5	0.0753[c]	0.00829[c]

[a] F = úplná data, CC = valid. skup., CS = CS odhad.

[b] Místní histologie jako jediný prediktor.

[c] Průměr přes 500 náhodných validačních skupin.

je 0.0744. Protože časová škála je v rocích, tento parametr znamená, že u 100 pacientů s nepříznivou histologií očekáváme každý rok o 7.44 více rekurencí než u stejného počtu pacientů s příznivou histologií. Pokud použijeme jako prediktor místní histologii, dostaneme 0.0557, což podceňuje skutečný parametr o více než 25%. Při validační skupině o velikosti 20% a 50% vycházel průměr CS odhadů po řadě 0.0768 a 0.0753. Jejich průměrné směrodatné odchylky však byly po řadě o 55% a 21% větší než u AH odhadu z celých dat. Ovšem toto zvětšení rozptylu je pouze daní za to, že Patologické centrum má nyní o 3200 či o 2000 vzorků méně práce.

6. Poznámky

Navržený odhad metodou upravené pseudoskórové funkce je konsistentní, snadno se spočítá a je eficientní v určité, i když dosti úzké, třídě možných odhadů. Popsali jsme jeho asymptotické vlastnosti a navrhli jsme odhady pro asymptotický rozptyl založené na pozorovaných datech. Důkazy všech zde uvedených tvrzení jsou součástí disertační práce autora (Kulich, 1997).

V tomto článku jsme definovali CS odhad pro jediný prediktor, nicméně jeho zobecnění na více prediktorů není obtížné. Pokud například mezi několika prediktory existuje jeden obtížně změřitelný a pro ten máme k disposici náhradní prediktor, provedeme příslušnou korekci pseudoskórové funkce pouze v její odpovídající složce. Musíme pak ovšem zobecnit váhovou konstantu w na matici. Podobně lze řešit případ několika prediktorů měřených s chybou. Klíčová je vždy platnost podmínek analogických k podmínkám 1 a 2. Podrobnosti viz Kulich (1997).

CS odhad klade jen relativně slabé podmínky na skutečný a náhradní prediktor. Pro jeho konsistenci je však důležitá podmínka nezávislosti chyb měření na době do události a do censorování (Podmínka 4). Její porušení má takřka vždy za následek ztrátu konsistence. V případě, že chyby měření závisí na ostatních prediktorech nebo jiných pozorovaných proměnných, stačí rozšířit lineární model, kterým popisujeme rozdělení chyb (viz Podmínky 1 a 2), o další proměnné.

CS odhad by šlo též zobecnit zavedením obecnějšího výběrového schématu pro selekci validační skupiny. Zahrnutím více necensorovaných pozorování do validační skupiny anebo zvětšením pravděpodobnosti výběru u subjektů s určitými charakteristikami by bylo možné dosáhnout menšího asymptotického rozptylu. Tento přístup je zatím předmětem dalšího výzkumu.

Literatura

[1] Breslow N. E. a Day N. E. (1987). *Statistical Models in Cancer Research.* **2**. *The Design and Analysis of Cohort Studies.* Lyon: IARC.

[2] Breslow N., Sharples K., Beckwith J. B., Takashima J., Kelalis P. P., Green D. M. a D'Angio G. J. (1991). Prognostic factors in nonmetastatic, favorable histology Wilms tumor. *Cancer*, **68**, 2345–2353.

[3] Carroll R. J., Ruppert D. a Stefanski L. A. (1995). *Measurement Error in Nonlinear Models.* London: Chapman & Hall.

[4] Cox D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

[5] Cox D. R. a Oakes D. (1984). *Analysis of Survival Data.* London: Chapman & Hall.

[6] D'Angio G. J., Breslow N., Beckwith J. B. a kol. (1989). Treatment of Wilms tumor: Results of the third National Wilms Tumor Study. *Cancer*, **64**, 349–360.

[7] Green D. M., Breslow N. E., Beckwith J. B. a kol. (1997). A comparison between single dose and divided dose administration of dactinomycin and doxorubicin. A report from National Wilms Tumor Study Group. *J. Clin. Oncology*, submitted.

[8] Kulich M. (1997). *Additive Hazards Regression With Incomplete Covariate Data.* PhD dissertation. Seattle: University of Washington.

[9] Lin D. Y. a Ying Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61–71.

[10] Nakamura T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, **48**, 829–838.

[11] Prentice R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, **69**, 331–342.

[12] Wang C. Y., Hsu L., Feng Z. D. a Prentice R. L. (1997). Regression calibration in failure time regression. *Biometrics*, **53**, 131–145.

[13] Zhou H. a Pepe M. S. (1995). Auxiliary covariate data in failure time regression. *Biometrika*, **82**, 139–149.