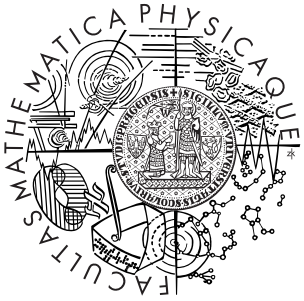
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Petr Novák

Regresní modely pro intenzity poruch v analýze spolehlivosti

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Doc. Petr Volf, CSc.
ÚTIA AV ČR

Studijní program: Matematika
Studijní obor:
Pravděpodobnost, matematická statistika a ekonometrie

2009

Rád bych na tomto místě poděkoval vedoucímu práce Doc. Petru Volfovi, CSc. za cenné podněty, rady a připomínky, trpělivost a poskytnutí literatury a dat.

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

V Praze dne 15. dubna 2009 Petr Novák

Obsah

Název práce: Regresní modely pro intenzity poruch v analýze spolehlivosti
Autor: Petr Novák
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Doc. Petr Volf, CSc., ÚTIA AV ČR
e-mail vedoucího: volf@utia.cas.cz

Abstrakt: V předložené práci studujeme regresní modely pro analýzu spolehlivosti. Srovnáváme Coxův model proporcionálního rizika, Aalenův aditivní model, model se zrychleným časem a jejich kombinace. U každého z modelů uvádíme postupy pro odhady parametrických i neparametrických částí rizikových funkcí a metody testování vhodnosti modelů, vycházející z postupů klasické regrese i z teorie čítacích procesů. Tyto metody pak demonstrujeme na simulovaných i reálných datech, zaměřujeme se na postup pro nalezení modelu, který daná data nejlépe popisuje.

Klíčová slova: analýza spolehlivosti, regresní modely, testy dobré shody

Title: Regression Models for Failure Intensities in Reliability Analysis
Author: Petr Novák
Department: Department of Probability and Mathematical Statistics
Supervisor: Doc. Petr Volf, CSc., ÚTIA AV ČR
Supervisor's e-mail address: volf@utia.cas.cz

Abstract: In the present work we study regression models in reliability analysis. We compare the Cox proportional hazards model, Aalen additive model, accelerated failure time model and their combinations. For each model we present procedures for estimating parametric and non-parametric risk function parts and goodness-of-fit tests based on classic regression routines and counting process theory. We demonstrate those tests on both real and simulated data and we focus on procedures how to find the model with the best fit.

Keywords: Reliability analysis, regression models, goodness-of-fit tests

Úvod

Analýza spolehlivosti respektive analýza přežití je důležitým nástrojem matematické statistiky, který umožňuje vyhodnocovat data reprezentující čas přežití nebo výdrže jedinců ve sledovaném výběru, ať pacientů v lékařských studiích nebo součástek v průmyslových testech. V této práci se zaměříme na studování a porovnávání regresních modelů pro spolehlivost, tedy metod popisujících vliv vysvětlujících proměnných na čas do sledované události. V medicinských studiích mohou mít na délku dožití vliv např. věk, pohlaví nebo výška, při průmyslovém testování součástek může mít na výdrž vliv např. zátěžový tlak, použitý materiál nebo teplota.

V první kapitole shrneme teoretické základy analýzy spolehlivosti, kde hraje důležitou roli teorie čítacích procesů a martingalů.

Vliv vysvětlujících proměnných je možné interpretovat mnoha způsoby. Ve druhé kapitole popíšeme tři základní modely, kterými jsou Coxův model proporcionálního rizika, model se zrychleným časem a Aalenův aditivní model. U každého modelu uvádíme metody, jak odhadovat jeho parametrické a neparametrické části a způsoby, jak testovat vhodnost modelu, tj. jak provést test dobré shody modelu s daty.

Ve třetí kapitole zkoumáme kombinace základních modelů. Věnujeme se podrobněji rozdílům mezi Coxovým modelem a modelem se zrychleným časem a uvádíme také Cox-Aalenův model kombinující proporcionální a aditivní vlivy na riziko.

Ve čtvrté kapitole předvedeme implementaci představených modelů. Na simulovaných datech demonstrujeme funkčnost testových procedur. Pro reálná data z praxe vyzkoušíme hledání a interpretaci nejvhodnějšího modelu.

Kapitola 1

Základy analýzy přežití

Zabýváme se studiem nezáporných náhodných veličin, které reprezentují čas od počátku sledování do nějaké předem definované události. Základní úlohou bude odhadnout rozdělení těchto veličin na základě pozorovaných dat. Často se takto analyzují data z lékařských výzkumů, např. čas od projevení vážné nemoci do smrti pacienta, nebo data z technických studií týkající se např. životnosti součástek.

Mějme tedy hodnoty $T_1, ..., T_n$, o nichž předpokládáme, že jsou nezávislé a stejně rozdělené (iid) s distribuční funkcí F(t). V analýze spolehlivosti se obvykle pracuje s *funkcí přežití* - pravděpodobností, že se subjekt dožije určitého času:

$$S(t) := P(T \geq t) = 1 - F(t),$$

rizikovou funkcí:

$$\alpha(t) := \lim_{h \to 0^+} \frac{P(t \leq T < t + h | T \geq t)}{h}$$

a *kumulativní rizikovou funkcí*:

$$A(s) := \int_0^t \alpha(s) ds$$

Pokud je rozdělení dat spojité s hustotou f, zjistíme, že

$$\alpha(t) = \lim_{h \to 0^+} \frac{P(t \leq T < t + h \wedge T \geq t)/h}{P(T \geq t)} = \frac{f(t)}{S(t)},$$

dále platí

$$S(t) = exp(- \int_0^t \alpha(s) ds).$$

Většinu modelů navíc sestavujeme tak, aby zahrnovaly i cenzorovaná data, tj. data, kdy ne všechny jedince pozorujeme kompletně od začátku až do události. Zde budeme pracovat se zprava cenzorovanými daty, kdy u některých jedinců bylo sledování ukončeno, aniž by se projevila událost. Když testujeme například životnost součástek, není často možné čekat s uzavřením studie až do doby, než se porouchá poslední sledovaná. Proto se studie ukončí po určitém čase s tím, že u zbylých součástek víme jen, že se do tohoto času neporouchaly.

Formálně pak uvažujeme skutečné časy událostí $T_1^*, ..., T_n^*$, časy cenzorování $C_1, ..., C_n$, časy ukončení pozorování $T_i = min(T_i^*, C_i)$ a indikátory událostí $\Delta_i = I(T_i^* < C_i)$. V případě nezávislého cenzorování, tedy že T_i^* a C_i jsou navzájem nezávislé, pozorujeme nezávislé, stejně rozdělené dvojice $(T_i, \Delta_i)_{i=1}^n$.

Zde se budeme zabývat regresními modely, tedy chováním dat v závislosti na vysvětlujících proměnných - kovariátách. Data proto budeme mít ve tvaru $(T_i, \Delta_i, \boldsymbol{X}_i)_{i=1}^n$, kde $\boldsymbol{X}_i$ představuje vektor hodnot kovariát. Časy událostí pak uvažujeme nezávislé podmíněně na $\boldsymbol{X}_i$. V některých modelech budeme pracovat i s kovariátami měnícími se v čase $\boldsymbol{X}_i(t)$.

Čítací procesy pro události

Užitečným nástrojem v analýze přežití je teorie čítacích procesů a martingalů. Pro pozorovaná data můžeme zavést procesy

$$N_i(t) = I(T_i \leq t, \Delta_i = 1),$$

tj. proces, který bude nejprve nulový a v okamžiku necenzorované události i-tého jedince skočí na jedničku, dále

$$N_\bullet(t) = \sum_{i=1}^{n} N_i(t),$$

udávající počet necenzorovaných událostí do času t v celém souboru, a $N(t)$ n-rozměrný proces se složkami $N_i(t)$. Všechny procesy nechť jsou definovány na $t \in [0, \tau]$, často se uvažuje $\tau = \infty$.

Historii (filtraci) událostí do času t označíme $\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), 0 \leq s \leq t\}$, kde $Y_i(t) = I(t \leq T_i)$, tedy indikátor, zda je i-tý jedinec v čase t ještě v ri-

ziku. Budeme také pracovat s $\mathcal{F}_{t^-} = \sigma\{N_i(s), Y_i(s), 0 \leq s < t\}$.

Dále budeme pracovat s přírůstkovými procesy. Pro čítací proces $N_i(t)$ bude $dN_i(t)$ procesem, který v případě změny $N_i(t)$ nabývá hodnoty příslušného skoku, jinak je nulový.

K čítacímu procesu lze ve standardních případech nalézt kompenzátor. Budeme pracovat se spojitým rozdělením dat, kdy existence kompenzátoru plyne z faktu, že $N_i(t)$ je zprava spojitý $\mathcal{F}_t$-submartingal. Označíme-li $\Lambda_i(t)$ kompenzátor procesu $N_i(t)$, dostaneme martingaly

$$M_i(t) = N_i(t) - \Lambda_i(t).$$

Existuje-li $\lambda_i(t)$ tak, že $\Lambda_i(t)$ se dají zapsat ve tvaru

$$\Lambda_i(t) = \int_0^t \lambda_i(s)ds,$$

nazveme jej procesem intenzity procesu $N_i(t)$. Ukáže se (Flemming & Harrington, 1992), že

$$\lambda_i(t) = Y_i(t)\alpha_i(t),$$

kde α_i představuje rizikovou funkci pro i-tého jedince.

Martingaly $M_i(t) = N_i(t) - \Lambda_i(t)$, kde $\Lambda_i(t) = \int_0^t Y_i(s)\alpha_i(s)ds$ si můžeme představit jako reziduální proces rozdílu mezi pozorovanými a očekávanými daty, rovnosti

$$N_i(t) = \Lambda_i(t) + M_i(t), \quad i = 1, ..., n$$

chápeme jako

$$data = model + chyba.$$

Vliv kovariát zahrnujeme do $\alpha_i(t)$. Často se testování vhodnosti modelu zakládá na martingalových reziduálech $M_i(t)$, přesněji na porovnání jejich rozdělení za platnosti modelu a hodnot jejich odhadů $\hat{M}_i(t)$.

Věrohodnost v analýze spolehlivosti

Většinou ale neznáme předem $\Lambda_i(t)$ resp. $\alpha_i(t)$, proto je potřeba je nejdřív odhadnout z dat. Budeme uvažovat případy parametrických, neparametrických i semiparametrických odhadů. Základem je zde teorie maximální věrohodnosti. Když máme n pozorování (T_i, Δ_i, X_i), věrohodnostní funkci můžeme psát jako

$$L = \prod_{i=1}^{n} f_i(T_i)^{\Delta_i} S_i(T_i)^{1-\Delta_i},$$

kde první část odpovídá necenzorovaným časům a druhá cenzorovaným. $f_i(t)$ a $S_i(t)$ značí hustotu resp. funkci přežití při daných hodnotách kovariát. Logaritmická věrohodnost má tvar

$$l = \sum_{i=1}^{n} \Delta_i \log(f_i(T_i)) + \sum_{i=1}^{n} (1 - \Delta_i) \log S_i(T_i).$$

Vzhledem k tomu, že $f_i(t) = \alpha_i(t) S_i(t)\ \forall i = 1, ..., n$, máme

$$l = \sum_{i=1}^{n} \Delta_i \log(\alpha_i(T_i)) + \sum_{i=1}^{n} \log S_i(T_i),$$

$$l = \sum_{i=1}^{n} \Delta_i \log(\alpha_i(T_i)) - \sum_{i=1}^{n} \int_0^{T_i} \alpha_i(t) dt.$$

Věrohodnost v řeči čítacích procesů

Věrohodnost je dobré přepsat pomocí čítacích procesů. Pro každého jedince máme $N_i(t) = \int_0^t dN_i(s)$, kde $dN_i(t) = Y_i(t) I(T_i = t)$ je proces, který je jedna v čase, kdy i-tý jedinec měl necenzorovanou událost, jinak je vždy nula. Zjistíme, že logaritmická věrohodnost má tvar

$$l = \sum_{i=1}^{n} \int_0^{\tau} \log \alpha_i(t) dN_i(t) - \sum_{i=1}^{n} \int_0^{\tau} Y_i(t) \alpha_i(t) dt.$$

Pro první část totiž platí:

$$\int_0^{\tau} \log \alpha_i(t) dN_i(t) = \begin{cases} \log \alpha_i(T_i) & \Delta_i = 1 \\ 0 & \Delta_i = 0 \end{cases}$$

a pro druhou část platí

$$\int_0^\tau Y_i(t)\alpha_i(t)dt = \int_0^{T_i} \alpha_i(t)dt = -\log S(T_i).$$

Vyjádření log-věrohodnosti můžeme odvodit alternativně. Vyjdeme z toho, že

$$P(dN(t)|\mathcal{F}_t) = \prod_{i=1}^{n} \alpha_i(t)^{dN_i(t)}(1-\alpha_i(t))^{Y_i(t)(1-dN_i(t))}.$$

Pomocí součinového integrování

$$L = \prod_{i=1}^{n}\prod_{t=0}^{\tau} \alpha_i(t)^{dN_i(t)}(1-\alpha_i(t))^{Y_i(t)(1-dN_i(t))},$$

z vlastností součinového integrálu pak

$$L = \prod_{i=1}^{n}\left(\prod_{t=0}^{\tau} \alpha_i(t)^{dN_i(t)}\right)\exp\left(-\int_0^\tau Y_i(t)\alpha_i(t)dt\right).$$

Logaritmováním se tak dostaneme ke stejnému výrazu jako předtím. Pro parametrické odhady dostaneme skórovou funkci derivováním podle parametrů:

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}}\sum_{i=1}^{n}\left(\int_0^\tau \log\alpha_i(t)dN_i(t) - Y_i(t)\alpha_i(t)dt\right).$$

Odhad kumulované rizikové funkce:

Uvažujme případ bez regrese, tedy pouze *iid* data (T_i, Δ_i), $i = 1, ..., n$. Označme

$$N_\bullet(t) = \sum_{i=1}^{n} N_i(t), \quad Y_\bullet(t) = \sum_{i=1}^{n} Y_i(t), \quad \Lambda_\bullet(t) = \sum_{i=1}^{n}\Lambda_i(t).$$

Čítací proces $N_\bullet(t)$ bude mít kompenzátor

$$\Lambda_\bullet(t) = \int_0^t Y_\bullet(s)\alpha(s)ds,$$

a tedy

$$M_\bullet(t) = N_\bullet(t) - \Lambda_\bullet(t)$$

je martingal, přičemž $M_\bullet(t) = \sum_{i=1}^{n} M_i(t)$.

Přírůstkový proces $dM_\bullet(t)$ má nulovou střední hodnotu, můžeme proto psát

$$E(dN_\bullet(t)|\mathcal{F}_{t^-}) = Y_\bullet(t)dA(t) = dN_\bullet(t).$$

Protože $A(t) = \int_0^t \alpha(s)ds$, můžeme kumulovanou rizikovou funkci odhadnout jako:

$$\hat{A}(t) = \int_0^t \frac{J(s)}{Y_\bullet(s)} dN_\bullet(s),$$

kde $J(s) = I(Y_\bullet(s) > 0)$. Tento odhad se nazývá *Nelson-Aalenův* (Aalen, 1975). Ve skutečnosti se jedná o součet

$$\hat{A}(t) = \sum_{T_i<t} \frac{\Delta_i}{Y_\bullet(T_i)}.$$

Kapitola 2

Základní regresní modely v analýze spolehlivosti

Naším hlavním úkolem bude modelovat intenzitu událostí v závislosti na různých regresorech. U medicinských dat můžeme zkoumat, zda má vliv váha, věk, pohlaví nebo jiné veličiny. Při testování technických součástek můžeme objekty vystavovat různému tlaku, napětí apod. Prozkoumáme nejdřív několik nejzákladnějších modelů, včetně toho, jak testovat, zda dobře popisují data.

Nechť máme data ve tvaru $(T_i, \Delta_i, \boldsymbol{X}_i(t))$, kde T_i jsou nezávislé podmíněně na $\boldsymbol{X}_i(t)$. Uvažujme nezávislé cenzorování zprava. $\boldsymbol{X}_i(t)$ je vektor kovariát i-tého jedince, přitom připouštíme, že regresory mohou nabývat různých hodnot v průběhu pozorování.

2.1 Coxův model proporcionálního rizika

Model proporcionálního rizika (Cox, 1972) vyjadřuje, že jednotlivé kovariáty působí multiplikativně přímo na rizikovou funkci, resp. intenzitu. Konkrétně tedy uvažujeme rizikovou funkci ve tvaru:

$$\alpha_i(t) = \alpha_0(t) \exp(\boldsymbol{X}_i^T(t)\boldsymbol{\beta}), \quad t \in [0, \tau],$$

kde $\alpha_0(t)$ je základní riziková funkce. Nejčastěji se uvažují $\boldsymbol{X}_i$ konstantní v čase. Můžeme model přepsat také pomocí intenzity:

$$\lambda_i(t) = Y_i(t)\alpha_0(t) \exp(\boldsymbol{X}_i^T(t)\boldsymbol{\beta}).$$

Míru vlivu jednotlivých složek popisují hodnoty koeficientů $\boldsymbol{\beta}$, při pevných ostatních kovariátách můžeme psát:

$$exp(\beta_1) = \frac{\alpha(t, X_1 + 1)}{\alpha(t, X_1)}.$$

Tento model je možné zobecnit tím, že závislost místo lineární formou $\boldsymbol{X}^T(t)\boldsymbol{\beta}$ modelujeme jinou funkcí:

$$\alpha_i(t) = \alpha_0(t) r(\boldsymbol{X}_i(t)).$$

Základním Coxovým modelem můžeme zjistit, zda je vliv dané kovariáty významný, popřípadě jaký je trend vlivu. Závislost ale může být složitější.

Odhad parametrů

Abychom mohli odhadovat parametry, je potřeba vyjádřit věrohodnost a skórovou funkci. Zavedeme nejdříve

$$S^{(0)}(s, \boldsymbol{\beta}) = \sum Y_i(s) e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(s)},$$

$$S^{(1)}(s, \boldsymbol{\beta}) = \sum \boldsymbol{X}_i(s) Y_i(s) e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(s)},$$

$$S^{(2)}(s, \boldsymbol{\beta}) = \sum \boldsymbol{X}_i(s) (\boldsymbol{X}_i(s))^T Y_i(s) e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(s)},$$

$$E(s, \boldsymbol{\beta}) = \frac{S^{(1)}(s, \boldsymbol{\beta})}{S^{(0)}(s, \boldsymbol{\beta})}.$$

Použijeme takzvanou *parciální věrohodnost*. Věrohodnost má tvar

$$L = \prod_{i=1}^{n} \alpha_i(T_i, \boldsymbol{\beta})^{\Delta_i} S_i(T_i, \boldsymbol{\beta}) =$$

$$= \prod_{i=1}^{n} \left(\int_0^\tau \alpha_i(s, \boldsymbol{\beta}) dN_i(s) \right)^{\Delta_i} \exp\left(- \int_0^\tau Y_i(s) \alpha_i(s, \boldsymbol{\beta}) \right) =$$

$$= \prod_{i=1}^{n} \left(\int_0^\tau \frac{e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(s)} dN_i(s)}{S^{(0)}(s, \boldsymbol{\beta})} \right)^{\Delta_i} \prod_{i=1}^{n} \int_0^\tau \left(\alpha_0(s) S^{(0)}(s, \boldsymbol{\beta}) dN_i(s) \right)^{\Delta_i} \times$$

$$\times \exp\left(- \int_0^\tau \alpha_0(s) S^{(0)}(s, \boldsymbol{\beta}) ds \right).$$

Budeme maximalizovat první součin, protože závisí na $\boldsymbol{\beta}$, ale nezávisí na základní rizikové funkci $\alpha_0(t)$. Budeme jej nazývat *parciální věrohodnostní funkcí*, $\widetilde{L}(\boldsymbol{\beta})$. Můžeme ji přepsat jako

$$\widetilde{L}(\boldsymbol{\beta}) = \prod_{\Delta_i=1} \frac{e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(T_i)}}{S^{(0)}(T_i, \boldsymbol{\beta})}.$$

Logaritmováním pak

$$\widetilde{l}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \left(\boldsymbol{\beta}^T \boldsymbol{X}_i(T_i) - \log S^{(0)}(T_i, \boldsymbol{\beta}) \right),$$

skórová funkce má tvar:

$$\widetilde{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \left(\boldsymbol{X}_i(T_i) - \frac{S^{(1)}(T_i, \boldsymbol{\beta})}{S^{(0)}(T_i, \boldsymbol{\beta})} \right) =$$

$$= \sum_{i=1}^{n} \Delta_i \left(\boldsymbol{x}_i(T_i) - E(T_i, \boldsymbol{\beta}) \right) = \sum_{i=1}^{n} \int_0^\tau \left(\boldsymbol{x}_i(s) - E(s, \boldsymbol{\beta}) \right) dN_i(s).$$

Koeficienty $\boldsymbol{\beta}$ můžeme odhadnout řešením soustavy skórových rovnic

$$\widetilde{U}(\boldsymbol{\beta}) = \mathbf{0}.$$

Derivujeme-li opačnou hodnotu skóre ještě jednou podle všech parametrů, získáme Fisherovu informační matici:

$$\widetilde{I}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \int_0^\tau \frac{\delta}{\delta \boldsymbol{\beta}} \left(\boldsymbol{x}_i(t) - \frac{S^{(1)}(s, \boldsymbol{\beta})}{S^{(0)}(s, \boldsymbol{\beta})} \right) dN_i(s) =$$

$$= \sum_{i=1}^{n} \int_0^\tau \left(\frac{S^{(2)}(s, \boldsymbol{\beta})}{S^{(0)}(s, \boldsymbol{\beta})} - \left(\frac{S^{(1)}(s, \boldsymbol{\beta})}{S^{(0)}(s, \boldsymbol{\beta})} \right)^{\otimes 2} \right) dN_i(s) = \int_0^\tau V(s, \boldsymbol{\beta}) dN_\bullet(s).$$

Za určitých podmínek můžeme použít výsledky jako u klasické teorie maximální věrohodnosti:

Věta 1. *Budiž $\boldsymbol{\beta}_0$ skutečná hodnota parametru $\boldsymbol{\beta}$. Nechť existuje $\mathcal{B}$ okolí bodu $\boldsymbol{\beta}_0$ tak, že:*
(a) $E\left(sup_{t\in[0,\tau],\boldsymbol{\beta}\in\mathcal{B}} Y_i(t) \left| X_{ij}(t) X_{ik}(t) \right| \exp(\boldsymbol{X}_i^T(t)\boldsymbol{\beta}) \right) < \infty \; \forall j, k = 1, ..., p$
(b) $P(Y_i(t) = 1 \forall t \in [0, \tau]) > 0$
(c) Existuje pozitivně definitní matice Σ, že

$$n^{-1} \int_0^\tau V(t, \boldsymbol{\beta}_0) S^{(0)}(t, \boldsymbol{\beta}_0) d\Lambda_0(t) \xrightarrow{P} \Sigma.$$

Pak pro $n \to \infty$ platí:

$$n^{-1/2} U(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, \Sigma),$$

$$n^{-1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, \Sigma^{-1}),$$

navíc $n^{-1} I(\hat{\beta})$ je konzistentním odhadem matice Σ.

Důkaz: Viz Andersen & Gill (1982), Martinussen & Scheike (2006), kap.6, str.184.

Máme tedy zaručenou konzistenci a asymptotickou normalitu odhadů. Můžeme proto aplikovat asymptotické metody pro testování hypotéz o hodnotách parametrů ($H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$):
Skórová statistika:

$$\widetilde{U}(\boldsymbol{\beta}_0)^T I(\boldsymbol{\beta}_0)^{-1} \widetilde{U}(\boldsymbol{\beta}_0),$$

Waldova statistika:

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T I(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

a věrohodnostní poměr:

$$-2 \log \left(\frac{\widetilde{L}(\boldsymbol{\beta}_0)}{\widetilde{L}(\hat{\boldsymbol{\beta}})} \right)$$

mají totiž za platnosti hypotézy asymptoticky rozdělení χ^2_p.

Odhad základní rizikové funkce

Vyjdeme z martingalové dekompozice za platnosti modelu:

$$N_i(t) = M_i(t) + \Lambda_i(t).$$

Z martingalové vlastnosti máme

$$E(N_i(t)) = E(\Lambda_i(t)) = E(\int_0^t Y_i(s) \alpha_i(s, \beta) ds) = E(\int_0^t Y_i(s) e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(s)} dA_0(s)),$$

pro všechna data potom

$$E(N_\bullet(t)) = E(\int_0^t S^{(0)}(s, \boldsymbol{\beta}) dA_0(s)).$$

Kdybychom od odhadu $\hat{A}_0(t, \boldsymbol{\beta})$ chtěli, aby

$$N_\bullet(t) = \int_0^t S^{(0)}(s, \boldsymbol{\beta}) d\hat{A}_0(s, \beta),$$

dostaneme tzv. *Breslowův* odhad ve tvaru

$$\hat{A}_0(t,\beta) = \int_0^t \frac{1}{S_0(s,\boldsymbol{\beta})} dN_\bullet(s).$$

Za stejných předpokladů jako ve Větě 1 pro $n \to \infty$ platí (Andersen & Gill, 1982):

$$n^{1/2}(\hat{A}_0(t,\hat{\boldsymbol{\beta}}) - A_0(t)) \xrightarrow{\mathcal{D}} U(t),$$

kde $U(t)$ je Gaussovský proces s nulovou střední hodnotou a varianční funkcí odhadnutelnou jako

$$\hat{\phi}(t) = n \int_0^t S_0(s,\hat{\boldsymbol{\beta}})^{-2} dN_\bullet(s) +$$

$$+n \int_0^t E(s,\hat{\boldsymbol{\beta}})^T d\hat{A}_0(s,\hat{\boldsymbol{\beta}}) (n^{-1} I(\hat{\boldsymbol{\beta}}))^{-1} \int_0^t E(s,\hat{\boldsymbol{\beta}}) d\hat{A}_0(s,\hat{\boldsymbol{\beta}}).$$

Testy dobré shody modelu s daty

Může nastat několik možností, při nichž by Coxův model nevystihoval chování dat dobře. Chyba může být například v předpokladu exponenciální závislosti. Je také možné, že se s časem mění míra vlivu jednotlivých kovariát. Tuto možnost bychom museli ošetřit zavedením koeficientů závislých na čase. Zde se zaměříme na testování proporcionality rizika při různých úrovních jednotlivých kovariát, což je podstatou modelu.

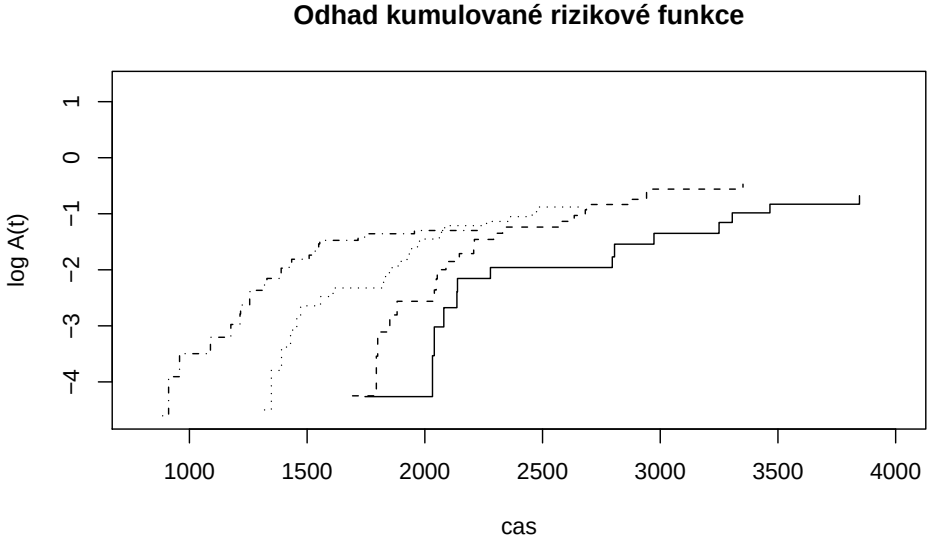
Grafické testy - stratifikovaná data

Uvažujme kovariáty konstantní v čase. Stratifikujme data do K skupin podle příbuzných hodnot kovariát, tedy že v rámci každé ze skupin bude $\boldsymbol{X}^T\boldsymbol{\beta}$ relativně podobné. Protože $\alpha_i(t) = e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\alpha_0(t)$, je $A_i(t) = e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}A_0(t)$, tedy

$$\log A_i(t) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \log A_0(t).$$

V každé z našich K skupin spočítáme Nelson-Aalenův odhad základní kumulované rizikové funkce $A_k(t)$ a vyneseme do jednoho grafu hodnoty

$$(t, \log A_k(t)), \quad k = 1, ..., K.$$



Obrázek 2.1: Odhady kumulované rizikové funkce pro data rozdělená podle kvartilů kovariáty X

Pokud Coxův model odpovídá chování dat, měly by být jednotlivé křivky zhruba rovnoběžné. Získáme tak přibližný náhled, ale ne přesné zhodnocení kvality modelu.

Na obr.2.1 jsou odhadnuté kumulované rizikové funkce pro data generovaná z Coxova modelu se základním rozdělením $\Gamma(20, 1/100)$, parametrem $\beta = 1$ a X_i generovanými z $N(0, 1)$. Data jsme rozdělili do čtyř skupin podle kvartilů X.

Martingalové reziduály

Za platnosti Coxova modelu máme martingaly

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(\boldsymbol{X}_i^T(s)\boldsymbol{\beta})\alpha_0(s)ds,$$

dosazením $\hat{\boldsymbol{\beta}}$ a Breslowova odhadu základní rizikové funkce dostaneme jejich

odhad

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(\boldsymbol{X}_i^T(s)\hat{\boldsymbol{\beta}}) d\hat{A}_0(s) ds =$$

$$= N_i(t) - \int_0^t Y_i(s) \exp(\boldsymbol{X}_i^T(s)\hat{\boldsymbol{\beta}}) \frac{1}{S_0(s,\hat{\beta})} dN_\bullet(s).$$

Pro přírůstky martingalových reziduálů tedy platí

$$d\hat{M}_i(s) = dN_i(s) - \frac{Y_i(s) \exp(\boldsymbol{X}_i^T(s)\hat{\boldsymbol{\beta}})}{S_0(s,\hat{\beta})} dN_\bullet(s).$$

Vynásobíme rovnosti kovariátami $\boldsymbol{X}_i(t)$, vyintegrujeme do času t a sečteme přes všechny jedince:

$$\sum_{i=1}^n \int_0^t \boldsymbol{X}_i(s) d\hat{M}_i(s) =$$

$$= \sum_{i=1}^n \int_0^t \boldsymbol{X}_i(s) dN_i(s) + \int_0^t \frac{\sum_{i=1}^n Y_i(s) \exp(\boldsymbol{X}_i^T(s)\hat{\boldsymbol{\beta}}) \boldsymbol{X}_i^T(s)}{S_0(s,\hat{\beta})} dN_\bullet(t) =$$

$$= \sum_{i=1}^n \int_0^t \left(\boldsymbol{X}_i(s) - E(s,\hat{\boldsymbol{\beta}}) \right) dN_i(s) = \tilde{U}(\hat{\boldsymbol{\beta}}, t).$$

Uvedený součet přesně odpovídá skórovému procesu do času t v bodě $\hat{\boldsymbol{\beta}}$.

Asymptotické rozdělení skórového procesu za platnosti modelu je možné odhadnout dle následujícího tvrzení:

Věta 2. *Pokud existuje $e(t, \boldsymbol{\beta}_0)$ limita v pravděpodobnosti $E(t, \boldsymbol{\beta}_0)$, je proces $n^{-1/2}\tilde{U}(\hat{\boldsymbol{\beta}}, t)$ v Coxově modelu asymptoticky ekvivalentní procesu*

$$n^{-1/2}(M_1(t) - I(t,\hat{\boldsymbol{\beta}}) I^{-1}(\tau,\hat{\boldsymbol{\beta}}) M_1(\tau)),$$

kde jsme označili

$$M_1(t) = \sum M_{1i}(t) = \sum_{i=1}^n \int_0^t (\boldsymbol{X}_i(u) - e(u,\hat{\boldsymbol{\beta}}_0)) dM_i(u).$$

Proces $n^{-1/2}M_1(t)$, $t \in [0,\tau]$ je asymptoticky ekvivalentní (tj. má stejnou limitu v distribuci) s:

$$n^{-1/2} \sum_{i=1}^n \int_0^t (\boldsymbol{X}_i(u) - E(u,\hat{\boldsymbol{\beta}})) dN_i(u) G_i,$$

respektive s

$$n^{-1/2}\sum_{i=1}^{n}\int_0^t (\boldsymbol{X}_i(u) - E(u,\hat{\boldsymbol{\beta}}))d\hat{M}_i(u)G_i,$$

kde G_i jsou iid $N(0,1)$.

Důkaz: Viz Bagdonavičius & Nikulin (2002), kap.12, str.239.

Nyní můžeme generovat hodnoty $M_1(t)$ pro odhad asymptotického rozdělení $n^{-1/2}\widetilde{U}(\hat{\beta},t)$ za předpokladu Coxova modelu. Jako testovou statistiku pro ověření modelu můžeme vzít např.

$$sup_{t\in[0,\tau]}|\widetilde{U}_j(\hat{\boldsymbol{\beta}},t)| \qquad nebo$$

$$sup_{t\in[\delta,\tau-\delta]}|\frac{\widetilde{U}_j(\hat{\boldsymbol{\beta}},t)}{\widehat{var}\widetilde{U}_j(\hat{\boldsymbol{\beta}},t)}|, \qquad j=1,...,p,$$

kde δ bereme malé kladné, abychom předešli problémům na krajích intervalu $[0,\tau]$ a $\widehat{var}\widetilde{U}_j(\hat{\boldsymbol{\beta}},t)$ je nějaký konzistentní odhad rozptylu skórového procesu.

V případě konstantních kovariát máme vlastně součet reziduálů

$$\widetilde{U}(\hat{\boldsymbol{\beta}},t) = \sum_{i=1}^{n}\boldsymbol{X}_i\hat{M}_i(t).$$

Coxův model s koeficienty proměnlivými v čase

Jako vsuvku uveďme rozšíření Coxova modelu. Je totiž možné, že vliv regresorů se během času mění. Coxův model potom lze rozšířit tak, že místo konstantních koeficientů budeme brát koeficienty $\boldsymbol{\beta}(t)$ jako funkce času:

$$\alpha_i(t) = \alpha_0(t)exp(\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^*(t)).$$

Za předpokladu $\alpha_0(t) > 0$ lze rizikovou funkci přepsat jako

$$\alpha_i(t) = \exp(\boldsymbol{X}_i^T(t)\boldsymbol{\beta}(t)),$$

protože základní riziková funkce je obsažena v koeficientech. Pomocí iteračních algoritmů je možné odhadnout kumulované koeficienty

$$\boldsymbol{B}(t) = \int_0^t \boldsymbol{\beta}(s)ds$$

a dají se i určit asymptotické vlastnosti jejich odhadů.
Pak lze testovat hypotézy jako

$$H_1 : \boldsymbol{\beta}_j(t) \equiv \boldsymbol{\beta}_j,$$

$$H_2 : \boldsymbol{\beta}_j(t) \equiv 0.$$

Odhady jsou ale relativně složité, uveďme proto jen speciální případ:

Jednoduchou závislost koeficientů na čase můžeme do modelu vnést tak, že zavedeme

$$\boldsymbol{\beta}_j(t) = \boldsymbol{\beta}_{1j} + \boldsymbol{\beta}_{2j}t$$

(Therneau & Grambsch, 2000). Uvědomme si, že

$$\alpha_i(t) = \exp(\boldsymbol{X}_i^T(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 t))\alpha_0(t) = \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}_1 + \boldsymbol{X}_i^T t\boldsymbol{\beta}_2)\alpha_0(t),$$

což je vlastně stále Coxův model, jen u $\boldsymbol{\beta}_2$ je kovariáta závislá na čase $\boldsymbol{X}_i t$. Můžeme tedy odhadnout jednotlivé parametry a testovat nezávislost na čase jako test $H_0 : \boldsymbol{\beta}_{2j} = 0$, $j = 1, ..., p$.

Když rozdělíme skórovou funkci na část pro $\boldsymbol{\beta}_1$ a pro $\boldsymbol{\beta}_2$, tj. $\widetilde{U} = (\widetilde{U}_1, \widetilde{U}_2)$, bude statistika skórového testu nulovosti vektoru β_2

$$\widetilde{U}_2^T(\hat{\boldsymbol{\beta}}_1, 0)I_{22}^{-1}(\hat{\boldsymbol{\beta}}_1, 0)\widetilde{U}_2(\hat{\boldsymbol{\beta}}_1, 0),$$

kde I_{22}^{-1} je příslušný blok inverze výběrové informační matice. Statistika bude mít za nulové hypotézy rozdělení χ_p^2. Můžeme také testovat nulovost jednotlivých parametrů zvlášť, stejně jako ve standardním Coxově modelu.

2.2 Model se zrychleným časem

Model se zrychleným časem (Accelerated Failure Time - AFT, Miller (1976), Buckley & James (1979)) vychází z představy, že hodnota kovariát určuje, jak rychle pro daného jedince subjektivně běží čas. Například když testujeme výdrž součástek, můžeme si představit, že pro součástku zatíženou vyšším napětím, tlakem, teplotou apod. poběží čas jakoby rychleji než pro součástku zatíženou méně, a proto více zatížená součástka vydrží v průměru objektivně kratší dobu. V běžném případě máme formálně vpodstatě lineární regresní model pro $log(T^*)$ s kovariátami $\boldsymbol{Z}_i = (Z_{i1}, ..., Z_{ip})$, parametry $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ a neznámým rozdělením ϵ:

$$\log(T_i^*) = -\boldsymbol{Z}_i^T\boldsymbol{\beta} + \epsilon_i,$$

což odpovídá transformaci času $t \to t\exp(\boldsymbol{Z}^T\boldsymbol{\beta})$. T_i^* představují skutečné časy událostí. Pokud je výběr cenzorovaný, máme ale k dispozici pouze hodnoty $T_i = min(T_i^*, C_i)$, není proto možné použít standardní regresní metody.

Když označíme F_0 distribuční funkci $\exp(\epsilon)$, máme

$$\begin{aligned} F(t) = P(T^* < t) = P(\exp(-\boldsymbol{Z}^T\boldsymbol{\beta})\exp(\epsilon) < t) = \\ = P(\exp(\epsilon) < t\exp(\boldsymbol{Z}^T\boldsymbol{\beta})) = F_0(t\exp(\boldsymbol{Z}^T\boldsymbol{\beta})), \\ f(t) = F'(t) = f_0(t\exp(\boldsymbol{Z}^T\boldsymbol{\beta}))\exp(\boldsymbol{Z}^T\boldsymbol{\beta}), \\ \alpha_i(t) = \frac{f(t)}{1-F(t)} = \alpha_0(t\exp(\boldsymbol{Z}_i^T\boldsymbol{\beta}))\exp(\boldsymbol{Z}_i^T\boldsymbol{\beta}). \end{aligned}$$

Rizikovou funkci uvažujme pro $t \in [0, \tau]$. Při pevných ostatních kovariátách můžeme míru vlivu jedné složky vyjádřit jako:

$$\exp(\beta_1) = \frac{E(T_{Z_1}^*)}{E(T_{Z_1+1}^*)}.$$

Model můžeme zobecnit tím, že místo logaritmické transformace budeme uvažovat monotónní transformaci h:

$$h(T^*) = -\boldsymbol{Z}^T\boldsymbol{\beta} + \epsilon.$$

Obecná transformace by mohla vystihovat data lépe, odhad parametrů by ale byl složitější.

AFT model se dá rozšířit i pro práci s kovariátami, které se mění v čase, ale pak se složitěji interpretují regresní koeficienty i motivace modelu.

Odhad parametrů

Vedle čítacích procesů událostí $N_i(t)$ a indikátorů rizika $Y_i(t)$ definujme jejich transformované varianty:

$$N_i^*(t) = N_i(t\exp(-\boldsymbol{Z}_i^T\boldsymbol{\beta})),\ \ i = 1, ..., n,$$

$$Y_i^*(t, \boldsymbol{\beta}) = Y_i(t\exp(-\boldsymbol{Z}_i^T\boldsymbol{\beta})),\ \ i = 1, ..., n.$$

Kumulovaná intenzita procesu $N_i^*(t)$ pak je

$$\Lambda_i^*(t) = \Lambda_i(t\exp(-\boldsymbol{Z}_i^T\boldsymbol{\beta})),$$

proto intenzita bude vypadat:

$$\lambda_i^*(t) = \frac{\partial}{\partial t}\Lambda_i^*(t) = \exp(-\boldsymbol{Z}_i^T\boldsymbol{\beta})\lambda_i(t\exp(-\boldsymbol{Z}_i^T\boldsymbol{\beta})) =$$

$$= Y_i(t\exp(-\boldsymbol{Z}_i^T\boldsymbol{\beta}))\exp(-\boldsymbol{Z}_i^T\boldsymbol{\beta})\alpha_0(t\exp(\boldsymbol{Z}_i^T\boldsymbol{\beta})\exp(-\boldsymbol{Z}_i^T\boldsymbol{\beta}))\exp(\boldsymbol{Z}_i^T\boldsymbol{\beta}) =$$

$$= Y_i^*(t, \boldsymbol{\beta})\alpha_0(t).$$

Označme ještě $M_i^*(t)$ transformované varianty martingalů M_i. Protože

$$N_\bullet^*(t) - \Lambda_\bullet^*(t) = M_\bullet^*(t)$$

a $M_i^*(t)$ mají nulovou střední hodnotu, bude $Y_\bullet^*(t, \boldsymbol{\beta})dA_0(t) \approx dN_\bullet^*(t)$. Základní kumulovanou rizikovou funkci můžeme proto odhadnout jako:

$$\hat{A}_0(t) = \int_0^t \frac{J(s)}{Y_\bullet^*(s, \boldsymbol{\beta})}dN_\bullet^*(s),$$

kde $J(s) = I(Y_\bullet^*(s, \boldsymbol{\beta}) > 0)$. Označme

$$S_0^*(t, \boldsymbol{\beta}) = \sum_{i=1}^n Y_i^*(t, \boldsymbol{\beta}), \qquad S_1^*(t, \boldsymbol{\beta}) = \sum_{i=1}^n Y_i^*(t, \boldsymbol{\beta})\boldsymbol{Z}_i,$$

$$E^*(t, \boldsymbol{\beta}) = \frac{S_1^*(t, \boldsymbol{\beta})}{S_0^*(t, \boldsymbol{\beta})}, \qquad W(t) = \left(\frac{\alpha_0'(t)t}{\alpha_0(t)} + 1\right).$$

Pro odhad parametrů $\boldsymbol{\beta}$ vyjdeme ze standardní skórové funkce (viz kapitola 1):

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}}\sum_{i=1}^n \int_0^\tau (\log(\alpha_i(t))dN_i(t) - Y_i(t)\alpha_i(t)dt) =$$

$$= \sum_{i=1}^{n} \int_0^\tau \frac{\partial}{\partial \boldsymbol{\beta}} (\alpha_i(t)) \frac{1}{\alpha_i(t)} (dN_i(t) - Y_i(t)\alpha_i(t)dt) =$$

$$= \sum_{i=1}^{n} \int_0^\tau \frac{\partial}{\partial \boldsymbol{\beta}} \left(\alpha_0(t \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta})) \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}) \right) \times$$

$$\times \frac{dN_i(t) - Y_i(t)\alpha_0(t \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta})) \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}) dt}{\alpha_0(t \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta})) \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta})} =$$

$$= \sum_{i=1}^{n} \int_0^\tau \left(\frac{\alpha_0'(t \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta})) t \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}) \boldsymbol{Z}_i + \alpha_0(t \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta})) \boldsymbol{Z}_i}{\alpha_0(t \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}))} \right) \times$$

$$\times (dN_i(t) - Y_i(t)\alpha_0(t \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta})) \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}) dt)$$

Pro $s = t \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta})$ dostaneme

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\tau \left(\frac{\alpha_0'(s)s}{\alpha_0(s)} + 1 \right) Z_i (dN_i^*(s) - Y_i^*(s, \boldsymbol{\beta}) dA_0(s)).$$

Když dosadíme za $dA_0(s)$ odhad $d\hat{A}_0(s, \beta)$, máme

$$U_W(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\tau W(s) \boldsymbol{Z}_i (dN_i^*(s) - \frac{Y_i^*(s, \boldsymbol{\beta})}{S_0^*(s, \boldsymbol{\beta})} dN_\bullet^*(s)) =$$

$$= \sum_{i=1}^{n} \int_0^\tau W(s)(\boldsymbol{Z}_i - E^*(s, \boldsymbol{\beta})) dN_i^*(s).$$

Abychom mohli použít skórovou funkci k odhadům, musíme buďto odhadnout α_0 a α_0' a dosadit do $W(s)$, případně použít jinou váhovou funkci, např. $W(s) \equiv 1$ nebo $W(s) = n^{-1} S_0^*(s, \boldsymbol{\beta})$. Zde budeme pracovat s $W(s) \equiv 1$. Uvědomme si, že

$$\widetilde{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\tau (\boldsymbol{Z}_i - E^*(t \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}), \boldsymbol{\beta})) dN_i(t) =$$

$$= \sum_{i=1}^{n} \Delta_i \left(Z_i - \frac{\sum_{j=1}^n I(T_i \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}) \leq T_j \exp(\boldsymbol{Z}_j^T \boldsymbol{\beta})) \boldsymbol{Z}_j}{\sum_{j=1}^n I(T_i \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}) \leq T_j \exp(\boldsymbol{Z}_j^T \boldsymbol{\beta}))} \right).$$

Skórová funkce tedy není spojitá v $\boldsymbol{\beta}$, což znamená, že nemusí existovat řešení rovnic $\widetilde{U}(\boldsymbol{\beta}) \equiv 0$. Jako odhad je možné vzít například takové $\hat{\boldsymbol{\beta}}$, které minimalizuje vzdálenost skóre od nuly, tedy $\left\| \widetilde{U}(\beta) \right\|$.

Pro inferenci o parametrech je dobré odhadnout asymptotické vlastnosti odhadů:

Věta 3. *Za podmínek regularity (Lin a kol., 1998) platí, že pokud je* $\boldsymbol{\beta}_0$ *skutečná hodnota parametru* $\boldsymbol{\beta}$ *a existuje* $\varphi(s, \boldsymbol{\beta}_0)$*, pro kterou je splněno*

$$n^{-1} \sum_{i=1}^{n} (\boldsymbol{Z}_i - E^*(s, \boldsymbol{\beta}_0))^{\otimes 2} Y_i^*(s) \xrightarrow{P} \varphi(s, \boldsymbol{\beta}_0),$$

a pozitivně definitní matice Σ *taková, že*

$$\Sigma = \int_0^\tau \varphi(s, \boldsymbol{\beta}_0) dA_0(s),$$

pak pro $n \to \infty$*:*

$$n^{-1/2} \tilde{U}(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, \Sigma),$$

$$n^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, C^{-1} \Sigma C^{-1}),$$

kde $C = \frac{\partial}{\partial \boldsymbol{\beta}}(U(\boldsymbol{\beta}_0))$ *a* $\hat{\boldsymbol{\beta}}$ *jsou odhady parametrů* $\boldsymbol{\beta}$*.*

Důkaz: Viz Lin a kol. (1998), Bagdonavičius & Nikulin (2003), kap.6 str.140.

Matici Σ můžeme odhadnout dosazením odhadu kumulované základní rizikové funkce, tj.

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} \int_0^\tau (\boldsymbol{Z}_i - E^*(s, \hat{\boldsymbol{\beta}}))^{\otimes 2} Y_i^*(s, \hat{\boldsymbol{\beta}}) d\hat{A}_0(s).$$

Derivujeme-li skóre ještě jednou podle $\boldsymbol{\beta}$, zjistíme po několika úpravách, že

$$C = \int_0^\tau s\varphi(s, \boldsymbol{\beta}_0) \alpha_0'(s) ds + \Sigma.$$

C můžeme odhadnout dosazením podobně jako B až na člen $\alpha_0'(s)$. Ten je možné aproximovat pomocí jádrového odhadu z odhadu $A_0(t)$, ale to nemusí být moc šikovné.

Rozdělení odhadu $\boldsymbol{\beta}$ lze taky odhadnout resamplingem. Platí totiž, že když vezmeme $\hat{\boldsymbol{\beta}}^*$ jako řešení rovnice

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\tau (\boldsymbol{Z}_i - E^*(t, \hat{\boldsymbol{\beta}})) d\hat{M}_i^*(t, \hat{\boldsymbol{\beta}}) G_i,$$

kde G_i jsou *iid* $N(0, 1)$, bude asymptotické rozdělení $n^{1/2}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*)$ shodné s rozdělením $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ (Lin a kol., 1998). Když tedy budeme opakovaně

simulovat G_i a řešit uvedenou soustavu, dostaneme odhad rozdělení $\hat{\boldsymbol{\beta}}^*$ a tedy i $\hat{\boldsymbol{\beta}}$.

Test dobré shody modelu s daty

AFT model může být nedostačující z mnoha důvodů, především proto, že by závislost času na kovariátách byla jiná než log-lineární.

Grafické testy

Stratifikujme data jako při grafických testech Coxova modelu do K skupin, ve kterých budou mít jedinci podobné hodnoty kovariát. Na základě hodnot $\log T_i$ v každé skupině spočtěme Nelson-Aalenův odhad kumulativní rizikové funkce $\hat{A}_k^{(\log)}(\log t)$. Protože

$$F_i^{(\log)}(\log t) = P(\log T_i^* < \log t) = P(-\boldsymbol{Z}_i^T\boldsymbol{\beta} + \epsilon_i < \log t) =$$

$$= P(\epsilon_i < \log t + \boldsymbol{Z}_i^T\boldsymbol{\beta}) = F_0^{(\log)}(\log t + \boldsymbol{Z}_i^T\boldsymbol{\beta}),$$

platí

$$A_i^{(\log)}(\log t) = A_0^{(\log)}(\log t + \boldsymbol{Z}_i^T\boldsymbol{\beta}).$$

V každé ze skupin by člen $\boldsymbol{Z}_i^T\boldsymbol{\beta}$ měl mít podobné hodnoty. Když vyneseme do jednoho grafu hodnoty

$$\left(\log t, \hat{A}_k^{(\log)}(\log t)\right), \quad k = 1, ..., K$$

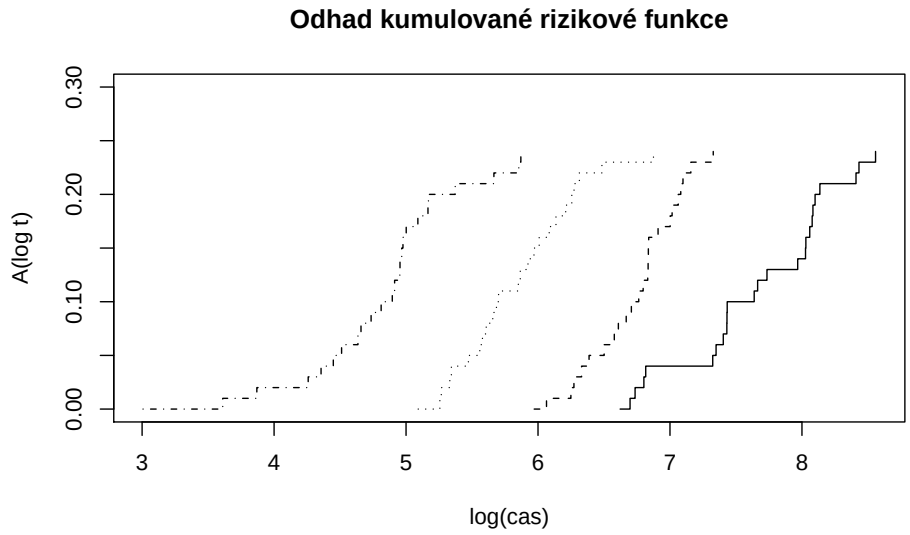
pro všechny skupiny, měly by být za platnosti modelu jednotlivé křivky přibližně rovnoběžné, navzájem posunuté ve vodorovné ose. Získáme tak obecnou představu, zda je model v pořádku.

Na obr.2.2 jsou odhadnuté kumulované rizikové funkce pro data generovaná z AFT modelu se základním rozdělením $\Gamma(5, 1/100)$, parametrem $\beta = 1$ a X_i generovanými z $N(0, 1)$. Data jsme rozdělili do čtyř skupin podle kvartilů X.

Regresní testy

AFT model lze do jisté míry otestovat podobnými metodami jako se testují lineární regresní modely. Uvažujme model bez cenzorování

$$\log T_i = -\boldsymbol{Z}_i^T\boldsymbol{\beta} + \epsilon_i.$$



Obrázek 2.2: Odhady kumulované rizikové funkce pro data rozdělená podle kvartilů kovariáty X

Když dosadíme maximálně věrohodný odhad $\hat{\beta}$, dostaneme rezidua

$$\log T_i + \boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}} = r_i.$$

Pokud budeme předpokládat konečné druhé momenty ϵ_i, měly by mít rezidua za platnosti modelu shodnou střední hodnotu a rozptyl. Můžeme tedy rozdělit data do skupin, např. podle hodnot jedné z kovariát, a následně testovat shodu rozptylu a střední hodnoty reziduí r_i.

Kdyby r_i odpovídaly normálnímu rozdělení, můžeme použít v případě dvou skupin t-test na shodu středních hodnot a F-test na shodu rozptylů, případně analýzu rozptylu pro více skupin. Normalitu nemůžeme obecně předpokládat, musíme ji otestovat. Kdyby rezidua testem normality neprošla, museli bychom použít neparametrické metody. V případě reziduí rozdělených do dvou skupin nejlépe Wilcoxonův test proti alternativě posunutí a Kolmogorov-Smirnovův test proti alternativě různého rozptylu nebo tvaru, pro více skupin pak Kruskal-Wallisův test.

Pokud ale máme cenzorovaná data, nelze použít rezidua přímo. Buďto je možné testovat pouze na necenzorovaných pozorováních, nebo můžeme odhadnout skutečné časy událostí T_i^* (Buckley & James, 1979):

$$\log \hat{T}_i^* = \Delta_i \log T_i + (1 - \Delta_i) E(\log T_i^* | \boldsymbol{Z}_i, T_i^* > T_i).$$

Rezidua pak odhadneme tak, že dosadíme za $\log T_i^*$ a odečteme $-\boldsymbol{Z}_i^T \boldsymbol{\beta}$:

$$\hat{\epsilon}_i = \Delta_i \epsilon_i + (1 - \Delta_i) E(\epsilon_i | \epsilon_i > \epsilon_i^C),$$

$$\hat{r}_i = \Delta_i r_i + (1 - \Delta_i) E(\epsilon | \epsilon > r_i^C),$$

kde $\epsilon_i^C = \log T_i + \boldsymbol{Z}_i^T \boldsymbol{\beta}$ a $r_i^C = \log T_i + \boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}$ pro cenzorované časy. Odhad rozdělení ϵ získáme buď z odhadu $A_0(t)$, nebo $E(\epsilon | \epsilon > r_i^C)$ odhadneme jako průměr všech reziduí vyšších než r_i^C (pouze z necenzorovaných pozorování).

2.3 Aalenův aditivní model

Model aditivního rizika (Aalen, 1980) vychází z myšlenky, že každá kovariáta ovlivňuje část rizikové funkce. Rizikovou funkci v tomto případě uvažujeme jako součet

$$\alpha_i(t) = \boldsymbol{X}_i^T(t)\boldsymbol{\beta}(t), \quad t \in [0, \tau],$$

kde $\boldsymbol{X}_i(t) = (X_{i1}(t), ..., X_{ip}(t))^T$ jsou hodnoty kovariát pro i-tého jedince. Nevýhodou tohoto modelu je, že narozdíl od předchozích případů musíme jednotlivé části rizikové funkce a $\boldsymbol{\beta}(t) = (\beta_1(t), ..., \beta_p(t))^T$ odhadovat neparametricky.

Aditivní model dobře vystihuje situaci, kdy uvažujeme sériový systém nezávislých komponent, např. zapojení technologického obvodu, ve kterém selhání jedné části znamená poruchu celého systému. Totiž pokud i-tá součástka bude mít funkci přežití $S_i(t)$ a rizikovou funkci $\beta_i(t)$, pro celý systém platí:

$$S(t) = \prod_{i=1}^{p} S_i(t),$$

$$\alpha(t) = -\frac{\partial}{\partial t} \log S(t) = \sum_{i=1}^{p} -\frac{\partial}{\partial t} \log S_i(t) = \sum_{i=1}^{p} \beta_i(t).$$

Součtový tvar rizika si taky můžeme představit jako Taylorův rozvoj prvního řádu okolo nuly vzhledem k hodnotám kovariát, tj.

$$\alpha(t, \boldsymbol{X}(t)) = \alpha(t, 0) + \boldsymbol{X}(t)^T \boldsymbol{\alpha}'(t, \boldsymbol{X}(t)^*).$$

Odhad rizikové funkce

Předpokládejme, že $\int_0^\tau |\beta_j(t)|dt < \infty$, $j = 1, ..., p$. Metodou vážených nejmenších čtverců budeme odhadovat kumulované části rizikové funkce

$$B_j(t) = \int_0^t \beta_j(s)ds.$$

Za platnosti modelu chceme $N_i(t) = \Lambda_i(t) + M_i(t)$. Za této podmínky platí

$$dN_i(t) = \lambda_i(t)dt + dM_i(t) = Y_i(t)\boldsymbol{X}_i(t)\boldsymbol{\beta}(t)dt + dM_i(t).$$

Když označíme $X_Y(t) = (Y_1(t)\boldsymbol{X}_1(t), ..., Y_n(t)\boldsymbol{X}_n(t))^T$, můžeme zapsat rovnost vektorově jako

$$dN(t) = X_Y(t)\boldsymbol{\beta}(t)dt + dM(t) = X_Y(t)dB(t) + dM(t).$$

Označme $X_Y^-(t) = (X_Y(t)^T(t)W(t)X_Y(t))^{-1}X_Y(t)^T(t)W(t)$ váženou pseudoinverzní matici k $X_Y(t)$. ($W(t)$ budiž pozitivně definitní váhová matice, k tomu jak ji volit se vrátíme později.) V bodech, kde pseudoinverze neexistuje, dodefinujeme nulovou maticí. Budiž tedy $J(t)$ indikátor, že pseudoinverze v daném bodě existuje. Vynásobíme obě strany rovnosti zleva maticí $X_Y^-(t)$ a dostaneme

$$X_Y^-(t)dN(t) = X_Y^-(t)X_Y(t)dB(t) + X_Y^-(t)dM(t).$$

Protože $X_Y^-(t)X_Y(t) = J(t)I_p$ a $M(t)$ je martingal s nulovou střední hodnotou, můžeme použít odhad

$$d\hat{B}(t) = X_Y^-(t)dN(t),$$

tedy

$$\hat{B}(t) = \int_0^t X_Y^-(s)dN(s).$$

Když bude mít matice $X_Y(t)$ plnou hodnost, bude $\hat{B}(t)$ nestranným odhadem $B(t)$, protože $M(t)$ má nulovou střední hodnotu a

$$\hat{B}(t) = \int_0^t J(s)dB(s) + \int_0^t X_Y^-(s)dM(s).$$

Za určitých podmínek odhadnuté hodnoty asymptoticky konvergují ke Gaussovskému procesu:

Věta 4. *Nechť platí*
(a) $sup_{t\in[0,\tau]}E(Y_i(t)W_i^2(t)X_{ij}(t)X_{ik}(t)X_{il}(t)) < \infty\ \forall j,k,l = 1,...,p$
(b) $r_2(t) = E(Y_i(t)W_i(t)\boldsymbol{X}_i^{\otimes 2}(t))$ *je regulární* $\forall t \in [0,\tau]$
potom pro $n \to \infty$ *platí:*

$$n^{1/2}(\hat{B}(t) - B(t)) \xrightarrow{\mathcal{D}} U,$$

kde U je Gaussovský martingal s kovarianční funkcí

$$\Phi(t) = \int_0^t \phi(s)ds,$$

kde

$$\Phi(t) = r_2^{-1}(t)E(Y_i(t)W_i^2(t)\boldsymbol{X}_i^{\otimes 2}(t)\boldsymbol{X}_i^T(t)\boldsymbol{\beta}(t))r_2^{-1}(t).$$

Dále platí, že

$$\hat{\Phi}(t) = n\int_0^t X_Y^-(s)diag(dN(s))(X_Y^-(s))^T$$

je konzistentním odhadem $\Phi(t)$*.*

Důkaz: Viz Martinussen & Scheike (2006), kap.5, str.110.

Konfidenční pás pro jednotlivé $B_j(t)$ o spolehlivosti $1-\alpha$ tedy můžeme sestavit jako

$$\hat{B}_j(t) \pm n^{-1/2}u_{1-\alpha/2}(\hat{\Phi}_{jj}^{1/2}(t)),$$

kde $u_{1-\alpha/2}$ značí kvantil normálního rozdělení $N(0,1)$.

Kdybychom vyšli z log-věrohodnostní funkce, derivovali podle $\boldsymbol{\beta}(t)$ a předpokládali v první části věrohodnosti $\lambda_i(t)$ známé, dostaneme stejný odhad jako metodou vážených nejmenších čtverců s tím, že

$$W(t) = diag(Y_i(t)/\lambda_i(t)).$$

Protože tento výraz ale závisí na neznámých $\lambda_i(t)$, doporučuje se nejprve odhadnout $\boldsymbol{\beta}(t)$ užitím $W(t) = I$, a pak znovu spočíst odhady s dosazenou $\hat{W}(t) = diag(Y_i(t)/\hat{\lambda}_i(t))$.

Test dobré shody modelu s daty

Aalenův model je velmi flexibilní, protože je neparametrický. Může se však stát, že by nevystihoval situaci dobře, například proto, že by vliv kovariát na jednotlivé části rizika nebyl lineární, nebo v situaci, kdy by bylo potřeba přidat do modelu interakce mezi kovariátami.

Martingalové reziduály

Za platnosti modelu máme následující martingaly:

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\boldsymbol{X}_i(s)^T\boldsymbol{\beta}(s)ds = N_i(t) - \int_0^t Y_i(s)\boldsymbol{X}_i(s)^T dB(s).$$

Dosadíme odhad komponent rizikové funkce a dostaneme (vektorově):

$$\hat{M}(t) = N(t) - \int_0^t X_Y(s) d\hat{B}(s) =$$

$$= N(t) - \int_0^t X_Y(s) X_Y^-(s) dN(s) = \int_0^t G(s) dN(s)$$

kde $G(t) = I - X_Y(t) X_Y(t)^-$. Protože ale $N(t) = X_Y(t)\boldsymbol{\beta}(t) + M(t)$ a $G(t)X_Y(t) = \mathbf{0}$, dostaneme

$$\hat{M}(t) = \int_0^t G(s) dM(s).$$

Derivováním zjistíme, že $d\hat{M}(t) = G(t)dM(t)$. Vynásobíme-li rovnost zleva $X_Y^T(t)$, dostaneme:

$$X_Y^T(t) d\hat{M}(t) = X_Y^T(t) G(t) dM(t) = \mathbf{0}.$$

Pokud budou kovatiáty nezávislé na čase, tj. $X_Y(t) = diag(Y_i(t))X$, bude

$$X^T diag(Y_i(t)) d\hat{M}(t) = X_Y^T(0) G(t) dM(t) = \mathbf{0}.$$

Proto musí platit

$$\sum_{i=1}^n Y_i(t) d\hat{M}_i(t) = 0,$$

tedy že součet přírůstků reziduálů prvků, které jsou v čase t v riziku, bude nulový. Odsud vyjdeme při testování, zda model dobře vystihuje data. Odhadneme rozdělení různých kumulativních součtů martingalových reziduálů.

Zvolme $n \times m$ rozměrný proces (maticově) $K(t) = (K_{i,j}(t)), i = 1, ..., n, j = 1, ..., m$ a definujme K-kumulativní residuální proces

$$M_K(t) = \int_0^t K^T(s) d\hat{M}(s) = \int_0^t K^T(s) G(s) dM(s).$$

Rozdělení $M_K(t)$ můžeme odhadnout pomocí simulací. Za platnosti modelu a za předpokladů věty 4 je totiž asymptotické rozdělení $M_K(t)$ shodné s asymptotickým rozdělením

$$\sum_{i=1}^n G_i \int_0^t \left(K_i(s)^T - K^T(s) X_Y(s) (X_Y^T(s) X_Y(s))^{-1} \boldsymbol{X}_i(s) \right) d\hat{M}_i(s),$$

kde G_i, $i = 1...n$ jsou $iid\ N(0,1)$. (viz Martinussen & Scheike (2006), kap.5, str.153)

Vhodnost modelu pak lze ověřit např. pomocí statistik

$$\sup_{t\in[0,\tau]} |M_{K_j}(t)|, \quad resp. \quad \int_0^\tau (M_{K_j}(t))^2 dt, \quad j = 1, ..., m.$$

Chceme, aby vysoké hodnoty testových statistik vypovídaly o porušení nulové hypotézy ve prospěch konkrétní alternativy, např. že závislost dat na k-té kovariátě není modelem dobře vysvětelná. Pro takový test můžeme vzít K konstantní v čase jako matici obsahující indikátory hodnot příslušné kovariáty zdiskretizované do skupin, např. po kvartilech.

$$K_i^k = (I(X_{ik} < 1.q(X_k)), I(1.q(X_k) < X_{ik} < 2.q(X_k)),$$
$$I(2.q(X_k) < X_{ik} < 3.q(X_k)), I(3.q(X_k) < X_{ik})), i = 1, ..., n.$$

Potom

$$M_{K_j}^k(t) = \int_0^t \sum_{i=1}^n K_i^k d\hat{M}(s) = \sum_{i:X_{ik}\in j.q(X_k)} X_{ik}\hat{M}_i(t), j = 1, ..., 4,$$

což představuje součty reziduálů pro jedince s hodnotami zkoumané kovariáty spadajícími do daného kvartilu. Pokud statistiky $\sup_{t\in[0,\tau]} |M_K^k(t)|$ respektive $\int_0^\tau (M_K^k(t))^2 dt$ spočítané z dat přesáhnou $1 - \alpha - kvantil$ simulovaných statistik, není závislost na k-té kovariátě na hladině α modelem dostatečně vysvětlena.

Testování submodelu

Protože odhady jednotlivých komponent $\boldsymbol{\beta}(t)$ mohou být relativně složité, můžeme otestovat, zda by se nedal model zjednodušit, např.

$$H_{01} : \beta_j(t) \equiv 0 \quad nebo \quad H_{02} : \beta_j(t) \equiv \gamma$$

pro nějakou konstantu γ. Pro kumulované koeficienty lze hypotézu přepsat jako

$$H_{01} : B_j(t) \equiv 0, \quad resp. \quad H_{02} : B_j(t) \equiv \gamma t.$$

Pro test H_{01} můžeme vzít např. statistiku

$$sup_{t\in[0,\tau]}|\hat{B}_j(t)|, \quad resp. \quad sup_{t\in[0,\tau]}|\frac{\hat{B}_j(t)}{\hat{v}ar(\hat{B}_j(t))}|,$$

pro test H_{02}

$$sup_{t\in[0,\tau]}|\hat{B}_j(t) + \hat{B}_j(\tau)\frac{t}{\tau}|.$$

Konstantnost příslušné části rizikové funkce testujeme pro odhad $\hat{\gamma} = \hat{B}_j(\tau)/\tau$.

Rozdělení těchto statistik můžeme odhadnout resamplingem podle následující věty:

Věta 5. *Pokud platí předpoklady Věty 4 a navíc $Y_i(t)\boldsymbol{X}_i(t)$ jsou stejně omezené s omezeným rozptylem a G_i, $i = 1, ..., n$ je iid výběr z $N(0, 1)$, má pak statistika*

$$n^{1/2}(\hat{\boldsymbol{B}}(\boldsymbol{t}) - \boldsymbol{B}(t))$$

stejné asymptotické rozdělení jako

$$\Delta_1(t) = n^{-1/2}\sum_{i=1}^{n}\hat{\epsilon}_i(t)G_i,$$

kde

$$\hat{\epsilon}_i(t) = \int_0^t (n^{-1}X_Y^T(s)X_Y(s))^{-1}\boldsymbol{X}_i(s)d\hat{M}_i(s).$$

Důkaz: Viz Martinussen & Scheike (2006), kap.5, str.117.

Nulovost nebo konstantnost j-té části rizikové funkce tak můžeme testovat porovnáním příslušné statistiky počítané z odhadu se simulovaným rozdělením $\Delta_{1j}(t)$ nebo jeho transformací.

Kapitola 3

Kombinace regresních modelů

V této kapitole se budeme zabývat možnostmi, jak spojit základní modely dohromady a jak v takových případech rozlišit, podle kterého modelu se data chovají. Uvedeme také některé situace ve kterých se modely shodují.

3.1 Proporcionální riziko vs. zrychlený čas

Coxův model vychází z toho, že hodnota kovariát působí na rizikovou funkci tak, že pro jedince s vyššími hodnotami kovariát momentální míra rizika proporcionálně roste (pro $\beta > 0$) respektive klesá. Model se zrychleným časem se zakládá na myšlence, že jedinci s vyššími hodnotami kovariát budou mít poměrně nižší (pro $\beta > 0$) resp. vyšší střední dobu dožití.

Shoda Coxova a AFT modelu

Ačkoliv jsou oba přístupy odlišné co se týče motivace, dá se ukázat, že v určitých případech modely splývají. Budeme pracovat s kovariátami konstantními v čase. Připomeňme rizikové funkce. Pro Coxův model

$$\alpha_i(t) = \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_C)\alpha_0(t),$$

pro AFT

$$\alpha_i(t) = \alpha_0(t \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_A)) \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_A).$$

Když bude základní rozdělení Weibullovo, tj. s hustotou

$$f_0(t) = \gamma\delta t^{\delta-1} \exp(-\gamma t^\delta),$$

budou oba modely vpodstatě totožné. Totiž

$$\alpha_0(t) = \frac{f_0(t)}{1 - F_0(t)} = \frac{\gamma\delta t^{\delta-1}\exp(-\gamma t^\delta)}{\exp(-\gamma t^\delta)} = \gamma\delta t^{\delta-1}.$$

Dosadíme do rizikových funkcí. Pro Coxův model:

$$\alpha_i(t) = \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}_C)\delta\gamma t^{\delta-1},$$

a pro AFT

$$\alpha_i(t) = \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}_A)\delta\gamma t^{\delta-1}(\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}_A))^{\delta-1} = \delta\gamma t^{\delta-1}(\exp(\boldsymbol{X}_i^T\delta\boldsymbol{\beta}_A)).$$

Rizikové funkce se shodují pro $\boldsymbol{\beta}_C = \delta\boldsymbol{\beta}_A$.

Abychom otestovali, zda je základní rozdělení skutečně Weibullovo, odhadneme nejprve základní rizikovou funkci v jednom nebo druhém modelu. Z ní spočteme odhad základní distribuční funkce a otestujeme její shodu s distribuční funkcí Weibullova rozdělení pro vhodné γ a δ, např. pomocí Kolmogorovova-Smirnonova testu. Alternativně můžeme základní distribuční funkci odhadnout v AFT modelu jako empirickou distribuční funkci $exp(\hat{r}_i)$ z necenzorovaných pozorování.

Teprve když test shodu zamítne, je třeba rozhodnout, který z modelů použít.

Grafické testy

Podobně jako v první kapitole můžeme rozdělit pozorování do K skupin s podobnými hodnotami kovariát. Ať bychom uvažovali Coxův nebo AFT model, v rámci každé skupiny by měly být tedy hodnoty $\boldsymbol{X}_i^T\boldsymbol{\beta}$ zhruba podobné. Skloubíme grafické postupy pro oba modely. Pro kumulovanou rizikovou funkci pro T a $\log T$ platí:

$$A(t) = -\log S(t) = -\log S^{(\log)}(\log t) = A^{(\log)}(\log t).$$

V Coxově modelu je

$$\log A_i(t) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \log A_0(t) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \log A_0^{(\log)}(\log t),$$

v AFT modelu

$$\log A_i(t) = \log A_0^{(\log)}(\log t + \boldsymbol{X}_i^T\boldsymbol{\beta})$$

V každé skupině tedy spočtěme Nelson-Aalenův odhad kumulované rizikové funkce založený na $\log T_i$ a do jednoho grafu vynesme hodnoty

$$(\log t, \log \hat{A}_k^{(log)}(\log t)), \quad k = 1, ..., K.$$

Pokud jeden z modelů popisuje data dobře, měly by být křivky zhruba rovnoběžné. Jak jsme ukázali, pro Coxův model by měly být navzájem posunuty vertikálně, pro AFT model horizontálně. Pokud se oba modely shodují, tj. pokud je základní rozdělení Weibullovo, budou křivky zhruba lineární.

Model kombinující zrychlený čas a proporcionální riziko

Jako vsuvku uveďme model, který kombinuje Coxův a AFT model tak, že intenzita je vyjádřena jako

$$\lambda_i(t) = Y_i(t)\alpha_0(t \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_1)) \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_2)$$

(Chen & Jewell, 2001). Model je v jistém smyslu přeurčený, protože odhadujeme dva parametry u každé kovariáty, hlavním cílem je zde ale testování

$$H_{01} : \beta_{1j} = 0 \qquad j = 1, ..., p,$$

$$H_{02} : \beta_{2j} = \beta_{1j} \qquad j = 1, ..., p.$$

Kdyby platila H_{01} pro všechny kovariáty, máme Coxův model, naopak kdyby platila H_{02}, máme model se zrychleným časem.

Odhad parametrů

Použijeme transformované čítací procesy podobně jako pro model se zrychleným časem:

$$N_i^*(t) = N_i(t \exp(-\boldsymbol{X}_i^T \boldsymbol{\beta}_1)), \quad Y_i^*(t) = Y_i(t \exp(-\boldsymbol{X}_i^T \boldsymbol{\beta}_1)),$$

dále

$$\Lambda_i^*(t) = \Lambda_i(t \exp(-\boldsymbol{X}_i^T \boldsymbol{\beta}_1)), \quad M_i^*(t) = M_i(t \exp(-\boldsymbol{X}_i^T \boldsymbol{\beta}_1)).$$

Zjistíme, že

$$\lambda_i^*(t) = \frac{\partial}{\partial \beta} \Lambda_i^*(t) = e^{-\boldsymbol{X}_i^T \boldsymbol{\beta}_1} \lambda_i(t e^{-\boldsymbol{X}_i^T \boldsymbol{\beta}_1}) = Y_i^*(t) e^{\boldsymbol{X}_i^T (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)} \alpha_0(t).$$

Protože $M_i^*(t) = N_i^*(t) - \Lambda_i^*(t)$ je martingal s nulovou střední hodnotou, chceme, aby

$$dN_i^*(t) \approx d\Lambda_i^*(t) = Y_i^*(t)e^{\boldsymbol{X}_i^T(\boldsymbol{\beta}_2-\boldsymbol{\beta}_1)}dA_0(t).$$

Základní riziko proto můžeme odhadnout jako

$$\hat{A}_0(t) = \int_0^t \frac{J(t)}{S_0^*(t,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)} dN_\bullet^*(t),$$

kde $S_0^*(t,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) = \sum_{i=1}^n Y_i^*(t)e^{\boldsymbol{X}_i^T(\boldsymbol{\beta}_2-\boldsymbol{\beta}_1)}$ a $J(t) = I(S_0(t,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)) > 0$. Označme ještě

$$S_1^*(t,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) = \sum_{i=1}^n Y_i^*(t)e^{X_i^T(\boldsymbol{\beta}_2-\boldsymbol{\beta}_1)}X_i \qquad E^*(t,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) = \frac{S_1^*(t,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)}{S_0^*(t,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)}.$$

Dosadíme odhad $\hat{A}(t)$ do věrohodnostní funkce (viz kapitola 1, max. věrohodnost). Zjistíme, že derivováním podle $\boldsymbol{\beta}_1$ a $\boldsymbol{\beta}_2$ dostaneme:

$$U_1(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\infty W(t)(\boldsymbol{X}_i - E^*(t,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2))dN_i^*(t),$$

$$U_2(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\infty (\boldsymbol{X}_i - E^*(t,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2))dN_i^*(t),$$

kde $W(t) \equiv \frac{t\alpha_0'(t)}{\alpha_0(t)}$. Řešením rovnic $U(\boldsymbol{\beta}) \equiv 0$ získáme požadované odhady. Problém je, že α_0 a α_0' pro $W(t)$ neznáme, odhady ale budou dávat smysl i pro jiné volby, např. $W(t) = t$, resp. W(t)=t/(1+t).

Podobně jako u AFT modelu, $U(\beta)$ není spojitá v $\boldsymbol{\beta}_1$, proto často není možné najít přesné řešení skórových rovnic. Jako odhad můžeme vzít např. takové $\hat{\boldsymbol{\beta}}$, aby $\|U(\boldsymbol{\beta})\|$ bylo co nejmenší.

Za podmínek regularity pro $n \to \infty$ platí:

$$n^{-1/2}U(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, \Sigma^{-1}(\boldsymbol{\beta}_0)),$$

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, D^{-1}\Sigma(\hat{\boldsymbol{\beta}})D^{-1}),$$

kde $\Sigma(\boldsymbol{\beta})$ je Fisherova matice, kterou můžeme odhadnout pomocí

$$\hat{\Sigma}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \int_0^{\infty} ((X_i^T, \hat{W}(t) X_i^T)^T +$$

$$-(E^*(t, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T, \hat{W}(t) E^*(t, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T)^T)^{\otimes 2} dN_i^*(t).$$

Přesnou formulaci podmínek regularity a důkaz asymptotických vlastností je možné nalézt v Chen & Jewell (2001), stejně jako tvar matice D. Ten je ale relativně složitý a navíc obsahuje člen $\alpha_0(t)$ a $\alpha_0'(t)$, které je potřeba aproximovat pomocí jádrových odhadů, což může způsobit nepřesnosti pokud počet pozorování není velký.

3.2 Cox-Aalenův model

Relativně obecnou kombinací modelů je spojení Coxova modelu se základní rizikovou funkcí $\alpha_0(t)$, kovariátami $\boldsymbol{Z}_i(t), i = 1, ..., n$ a vektorovým parametrem $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$

$$\lambda_i(t) = Y_i(t) \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_i(t))\alpha_0(t)$$

a Aalenova aditivního modelu s kovariátami $\boldsymbol{X}_i(t), i = 1, ..., n$ a komponentami rizika $\boldsymbol{\alpha}(t) = (\alpha_1(t), ..., \alpha_q(t))$

$$\lambda_i(t) = Y_i(t)\boldsymbol{X}_i^T(t)\boldsymbol{\alpha}(t).$$

Jednou z možností jak spojit proporcionální a aditivní vliv je Cox-Aalenův model (Scheike & Zhang, 2002) s intenzitou

$$\lambda_i(t) = Y_i(t) \exp(\boldsymbol{Z}_i^T(t)\boldsymbol{\beta})\boldsymbol{X}_i^T(t)\boldsymbol{\alpha}(t), \quad t \in [0, \tau],$$

tedy vpodstatě Coxův model, kde základní intenzita závisí na dalších kovariátách v aditivním tvaru. Tento model je prvním členem Taylorova rozvoje modelu

$$\lambda(t) = Y(t) \exp(\boldsymbol{\beta}^T \boldsymbol{Z}(t))\lambda(t, \boldsymbol{X}(t)),$$

podle hodnot kovariáty $\boldsymbol{X}$, který byl navržen pro zkoumání proporcionálního vlivu kovariát $\boldsymbol{Z}_i(t)$ oproštěného od vlivu hodnot kovariát $\boldsymbol{X}_i(t)$.

Odhad parametrů

Budeme postupovat podobně jako u Aalenova modelu. Definujme matici

$$Y(\boldsymbol{\beta}, t) = (Y_1(t) \exp(\boldsymbol{Z}_1(t)^T\boldsymbol{\beta})\boldsymbol{X}_1(t), ..., Y_n(t) \exp(\boldsymbol{Z}_n(t)^T\boldsymbol{\beta})\boldsymbol{X}_n(t)).$$

Log-věrohodnost bude mít tvar

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\tau \log(Y_i(t) \exp(\boldsymbol{Z}_i(t)^T\boldsymbol{\beta})\boldsymbol{X}_i(t)^T dA(t))dN_i(t)$$

$$- \sum_{i=1}^{n} \int_0^\tau Y_i(t) \exp(\boldsymbol{Z}_i(t)^T\boldsymbol{\beta})\boldsymbol{X}_i(t)^T dA(t).$$

Derivováním se dostaneme ke skórovým rovnicím pro $\boldsymbol{\beta}$ a $dA(t)$:

$$\int_0^\tau \boldsymbol{Z}(t)^T dN(t) - Y(\boldsymbol{\beta}, t)dA(t) = 0,$$

$$Y(\boldsymbol{\beta},t)^T W(t)(dN(t) - Y(\boldsymbol{\beta},t)dA(t)) = 0,$$

kde $W(t)$ je diagonální matice s prvky $w_i(t) = \frac{Y_i(t)}{\lambda_i(t)} = \frac{Y_i(t)\exp(-\boldsymbol{Z}_i^T\boldsymbol{\beta})}{\boldsymbol{X}_i^T\boldsymbol{\alpha}(t)}$. Když označíme

$$Y^-(\boldsymbol{\beta},t) = (Y(\boldsymbol{\beta},t)^T W(t) Y(\boldsymbol{\beta},t))^{-1} Y(\boldsymbol{\beta},t)^T W(t)$$

váženou pseudoinverzi matice $Y(\boldsymbol{\beta},t)$, můžeme ze druhé skórové rovnice odhadnout složky vektoru kumulované základní intenzity jako

$$\hat{A}(\boldsymbol{\beta},t) = \int_0^t Y^-(\boldsymbol{\beta},s)dN(s).$$

Dosazením do první rovnice získáme

$$U(\boldsymbol{\beta}) = \int_0^\tau \boldsymbol{Z}^T(t)(I - Y(\boldsymbol{\beta},t)Y^-(\boldsymbol{\beta},t))dN(t) = \int_0^\tau \boldsymbol{Z}^T(t)G(\boldsymbol{\beta},t)dN(t),$$

kde jsme označili $G(\boldsymbol{\beta},t) = I - Y(\boldsymbol{\beta},t)Y^-(\boldsymbol{\beta},t)$. Problémem je, že Y^- závisí jak na parametrech $\boldsymbol{\beta}$, tak na neznámých $\boldsymbol{\alpha}(t)$ skrz váhy $w_i(t)$. Váhy můžeme přepsat jako

$$w_i(t) = Y_i(t)h_i(t)exp(-\boldsymbol{Z}_i(t)^T\boldsymbol{\beta}).$$

Když místo původní části $h_i(t) = (\boldsymbol{X}_i(t)^T\boldsymbol{\alpha}(t))^{-1}$ použijeme pro odhad $h_i(t) \equiv 1$, nebude skórová funkce pro $\boldsymbol{\beta}$ záviset na $\boldsymbol{\alpha}(t)$.

Doporučuje se spočíst $\hat{\boldsymbol{\beta}}$ na základě skórových rovnic se zjednodušenými vahami a tyto odhady pak použít k odhad $\hat{A}(\hat{\boldsymbol{\beta}},t)$. Dále je pak možné odhady dosadit do původních vah a vyjádřit odhady založené na původních vahách.

Asymptotické vlastnosti odhadů $\hat{\boldsymbol{\beta}}$ a $\hat{A}(t)$ shrnují následující tvrzení, přesné znění předpokladů lze najít v Scheike & Zhang (2002):

Věta 6. *Za podmínek regularity v Cox-Aalenově modelu platí, že*

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0,\Sigma),$$

kde Σ lze odhadnout jako

$$\hat{\Sigma} = nI^{-1}(\hat{\boldsymbol{\beta}},\tau)\left\{\sum_{i=1}^n \hat{\epsilon}_{1i}^{\otimes 2}(\tau)\right\} I^{-1}(\hat{\boldsymbol{\beta}},\tau),$$

přičemž I je odhad informační matice o $\boldsymbol{\beta}$ a $\hat{\epsilon}_{1i}$ je odhadem

$$\epsilon_{1i} = \int_0^t (\boldsymbol{Z}_i(s) - Z^T(s)Y(\boldsymbol{\beta}_0, s)(Y^T(\boldsymbol{\beta}_0, s)W(s)Y(\boldsymbol{\beta}_0, s))^{-1}\boldsymbol{X}_i(s)^T)dM_i(s).$$

Důkaz: Viz Scheike & Zhang (2002).

Když tedy odhadneme varianční matici odhadů, můžeme testovat například, zda $\beta_j = 0$, tedy zda je vliv jednotlivých kovariát v proporcionální části modelu významný.

Věta 7. *Za podmínek regularity v Cox-Aalenově modelu platí, že*

$$n^{1/2}(\hat{A}(\hat{\boldsymbol{\beta}}, t) - A(t))$$

konverguje ke Gaussovskému procesu s varianční funkcí $\Psi(t)$, kterou lze odhadnout jako

$$\hat{\Psi} = n\sum_{i=1}^{n} \hat{\epsilon}_{2i}^{\otimes 2}(t),$$

přičemž

$$\epsilon_{2i}(t) = \epsilon_{3i}(t) + H^T(\boldsymbol{\beta}_0, t)I(\boldsymbol{\beta}_0, t)^{-1}\epsilon_{1i}(t),$$

$$\epsilon_{3i}(t) = \int_0^t (Y^T(\boldsymbol{\beta}_0, s)W(s)Y(\boldsymbol{\beta}_0, s))^{-1}\boldsymbol{X}_i(s)dM_i(s),$$

$$H(\beta, t) = \int_0^t Y^-(\boldsymbol{\beta}, s)diag\left(Y(\boldsymbol{\beta}, s)Y^-(\boldsymbol{\beta}, s)dN(s)\right)\boldsymbol{Z}(s)$$

a $\hat{\epsilon}_{ij}$ jsou příslušné odhady, tj. při dosazení $\hat{\boldsymbol{\beta}}$, $\hat{A}_j$ resp. $\hat{\alpha}_j$ a $\hat{M}_i$.

Důkaz: Viz Scheike & Zhang (2002).

Dále platí, že asymptotické rozdělení $n^{1/2}(\hat{A}(\hat{\boldsymbol{\beta}}, t) - A(t))$ bude stejné jako asymptotické rozdělení

$$n^{1/2}\sum_{i=1}^{n} \hat{\epsilon}_{2i}(t)G_i,$$

kde $G_i, i = 1, ..., n$ jsou *iid* $N(0, 1)$. Pro inferenci o komponentách $A(t)$ je tedy možné jako u samotného Aalenova modelu simulacemi testovat, zda $\boldsymbol{\alpha}_j(t) \equiv 0$ pomocí statistiky

$$sup_{t\in[0,\tau]}|\hat{A}_j(t)| \quad resp. \quad sup_{t\in[0,\tau]}|\frac{\hat{A}_j(t)}{\widehat{var}\hat{A}_j(t)}|,$$

nebo zda $\boldsymbol{\alpha}_j(t) \equiv \gamma$, tj. $A_j(t) = \gamma t$ pomocí

$$sup_{t\in[0,\tau]}|\hat{A}_j(t) - \frac{\hat{A}_j(t)}{\tau}t|,$$

kde γ odhadneme jako $\hat{\gamma} = \frac{\hat{A}_j(t)}{\tau}$.

Test dobré shody modelu s daty

Testováním hypotéz o nulovosti parametrů a komponent můžeme model významně zjednodušit, nezbavili jsme se ale nutnosti na začátku určit, které kovariáty mají mít aditivní a které proporcionální vliv.

Můžeme např. kovariáty postupně přesouvat a porovnat, při jakém sestavení kovariát model vystihuje data nejlépe. Počet možností jak kovariáty rozdělit mezi aditivní a proporcionální část roste ale exponenciálně s počtem kovariát, takže při větším počtu by už vyzkoušení všech variant modelu trvalo relativně dlouho.

Protože proporcionální vliv kovariát se interpretuje snáze, budeme upřednostňovat následující postup: Nejprve sestavíme Coxův model pro všechny kovariáty, otestujeme u kterých můžeme použít porporcionalitu, ty pak použijeme v proporcionální části Cox-Aalenova modelu a ostatní dosadíme do aditivní části.

Odvození následujících testových statistik je relativně složité, uvádíme zde proto jen základní myšlenky. Přesné odvození a důkazy asymptotických vlastností jsou k nalezení v Scheike & Zhang (2002).

Test pro aditivní část

Chceme otestovat, jestli model vystihuje dobře závislost dat na kovariátách obsažených v aditivní části modelu. Z martingalové dekompozice máme

$$M_i(t) = N_i(t) - \int_0^t Y_i(t)\exp(\boldsymbol{Z}_i(s)^T\boldsymbol{\beta}_0)\boldsymbol{X}_i(s)^T dA(s).$$

Když dosadíme odhady $\hat{\boldsymbol{\beta}}$ a $\hat{A}(t)$, dostaneme odhad

$$\hat{M}(t) = N(t) - \int_0^t Y(\hat{\boldsymbol{\beta}}, s)Y^-(\hat{\boldsymbol{\beta}}, s)dN(s) = \int_0^t G(\hat{\boldsymbol{\beta}}, s)dN(s).$$

Pomocí Taylorova rozvoje dostaneme po několika úpravách aproximaci

$$\tilde{M}(t) = \int_0^t G(\hat{\boldsymbol{\beta}}, t) dM(s) + B(\hat{\boldsymbol{\beta}}, t) U(\hat{\boldsymbol{\beta}}),$$

kde

$$B(\boldsymbol{\beta}, t) = -\int_0^t G(\boldsymbol{\beta}, s) diag\left\{Y(\boldsymbol{\beta}, s) Y^{-}(\boldsymbol{\beta}, s) dN(s)\right\} \boldsymbol{Z}(s) I^{-1}(\boldsymbol{\beta}, \tau).$$

Podobně jako u Aalenova modelu můžeme zkusit sečíst martingalové reziduály přes různé úrovně jednotlivých kovariát, tj. zavést $n \times m$ rozměrný proces $K(t)$ (nebo jen $n \times m$ matici K) a kumulativní reziduální proces

$$M_K(t) = \int_0^t K^T(s) d\hat{M}(s).$$

Když bude K matice konstant, dostaneme lineární kombinace martingalových reziduálů, protože pak

$$M_K(t) = K^T \int_0^t d\hat{M}(s) = K^T \hat{M}(t).$$

Dá se ukázat, že rozdělení $M_K(t)$ je asymptoticky ekvivalentní s rozdělením

$$n^{-1/2} \sum_{i=1}^{n} \int_0^t G_i \left\{K_i(s) - K^T(s) \left(Y(\hat{\boldsymbol{\beta}}, s) W(s) Y(\hat{\boldsymbol{\beta}}, s)\right)^{-1} \boldsymbol{X}_i(s)\right\} d\hat{M}_i(s) +$$

$$-n^{-1/2} B_K(\hat{\boldsymbol{\beta}}, t) \sum_{i=1}^{n} \hat{\epsilon}_{1i},$$

kde

$$B_K(\boldsymbol{\beta}, t) = \int_0^t K^T(s) G(\boldsymbol{\beta}, s) diag\left\{Y(\boldsymbol{\beta}, s) Y^{-}(\boldsymbol{\beta}, s) dN(s)\right\} Z(s) I^{-1}(\boldsymbol{\beta})$$

a G_i, $i = 1, ..., n$ jsou iid z $N(0, 1)$.

Jako testovou statistiku použijeme

$$\sup_{t \in [0,\tau]} |M_{K_j}(t)|, \quad resp. \quad \int_0^\tau (M_{K_j}(t))^2 dt, \quad j = 1, ..., m.$$

Můžeme tedy pro každou kovariátu v aditivní části zvlášť vzít $M^k_{K_j}(t)$ jako součet martingalových reziduálů přes jednotlivé kvartily jejích hodnot. Když

testová statistika přesáhne konfidenčních meze, nevysvětluje model situaci dostatečně přesně. Buďto tedy má j-tá kovariáta spíše proporcionální vliv nebo je vliv složitější než můžeme modelem postihnout.

Test pro proporcionální část

Podobně jako u Coxova modelu založíme test na skórovém procesu. Dá se ukázat, že rozdělení normovaného skórového procesu $n^{-1/2}U(\hat{\boldsymbol{\beta}}, t)$ za platnosti modelu je asymptoticky shodné s rozdělením procesu

$$n^{-1/2} \sum_{i=1}^{n} \left(\hat{\epsilon}_{1i}(t) + I(\hat{\boldsymbol{\beta}}, t) I^{-1}(\hat{\boldsymbol{\beta}}, \tau) \hat{\epsilon}_{1i}(\tau) \right) G_i,$$

kde G_i jsou *iid* $N(0, 1)$ a $\hat{\epsilon}_{1i}$ jsou odhady z Věty 6. Pro j-tou kovariátu pak spočítáme testovou statistiku

$$\sup_{t \in [0,\tau]} |U_j(\hat{\boldsymbol{\beta}}, t)|,$$

jejíž rozdělení můžeme asymptoticky odhadnout pomocí opakovaného simulování veličin G_i. Když tedy statistika nebude v konfidenčních mezích, je vliv j-té kovariáty složitější. Můžeme ji například zkusit dát do aditivní části nebo vhodně transformovat.

Kapitola 4

Příklady

V této části vyzkoušíme uvedené regresní modely na simulovaných i reálných datech. Výpočty byly prováděny v softwaru R. Coxův, Aalenův a Cox-Aalenův model jsou v R k dispozici v knihovně Timereg (autor Thomas Scheike), AFT model byl implementován přímo, pomocí metod popsaných v části 2.2. V příloze je vysvětleno, jak se příslušné funkce u všech modelů používají. Zdrojové kódy funkce pro AFT model, simulací a zpracování dat jsou přiloženy na CD.

4.1 Použití modelů na simulovaná data

Základní modely

Vyzkoušíme, zda testy vhodnosti modelů fungují tak, jak mají. Nagenerujeme data podle konkrétního základního modelu a otestujeme, kterým modelům data odpovídají. Mějme jednu spojitou kovariátu X s hodnotami simulovanými z rozdělení $N(0,1)$. Pro jednoduchost neuvažujme cenzorování.

Coxův model testujeme pomocí skórového procesu, porovnáním simulovaných hodnot $sup_{t\in[0,\tau]}\|U(\hat{\beta},t)\|$ se skutečnou. AFT model ověřujeme neparametrickými testy, pro Kruskal-Wallisův test shody rozdělení reziduí rozdělíme data podle kvartilů X, pro Wilcoxonův a Kolmogorov-Smirnovův porovnáváme rezidua pod a nad mediánem X. Aalenův model testujeme pomocí supremového a integrálního testu založeného na simulovaných hodnotách kumulativního reziduálního procesu $M_K(t)$. Matici K volíme jako indikátory hodnot kvartilů X, reziduální proces je tedy součtem martingalových re-

ziduálů přes jednotlivé kvartily (viz kapitola 1). Pro testy Coxova a Aalenova modelu vezmeme vždy 1000 replikací zkoumaných procesů. Pro Coxův a AFT model jsou p-hodnoty příslušných testů v tabulce 4.1 a pro Aalenův model v tabulce 4.2.

Data podle Coxova modelu

Uvažujme nejprve 1000 pozorování z Coxova modelu s $\beta_1 = 1$ s Weibullovým základním rozdělením $Wb(\gamma = 10^6, \delta = 5)$ ($S_0(t) = \exp(-\gamma t^\delta)$). Dosazením do Coxova modelu jsme dostali odhad $\hat{\beta}_C = 1.00$, 95%−konfidenční interval $(0.922, 1.087)$, p-hodnota testu proporcionality vyšla nevýznamná, tj. že model nezamítáme. Použitím AFT modelu jsme dostali odhad $\hat{\beta}_A = 0.204$, p-hodnoty pro testy vhodnosti opět vyšly nevýznamné, což nás nepřekvapí, vzhledem k tomu, že základní rozdělení je Weibullovo. Odhady parametrů přibližně korespondují s $\beta_C = \delta\beta_A$, kde δ je odpovídající parametr Weibullova rozdělení. Když jsme zkusili Aalenův model, vyšly ve všech kvartilech p-hodnoty pro supremový i integrální test významné, model proto zamítneme.

Když jsme použili stejné hodnoty X a simulovali 1000 hodnot z Coxova modelu s $\beta_1 = 1$ se základním rozdělením gamma $\gamma(a = 1/100, p = 5)$ (tj. $f_0(t) \propto t^{p-1}e^{-at}$), dostali jsme pro Coxův model odhad $\hat{\beta}_C = 1.04$, 95%−konfidenční interval $(0.950, 1.121)$. Simulovaná p-hodnota testu proporcionality vyšla opět nevýznamná, model nezamítáme. P-hodnoty AFT modelu i Aalenova modelu byly < 0.05, na této hladině je tedy zamítneme.

Když jsme simulovali data z Coxova modelu s $\beta_1 = 1$ se základním rozdělením lognormálním $LN(\mu = 2, \sigma^2 = 1)$, dostali jsme v Coxově modelu odhad 0.967, 95%−konfidenční interval rovný $(0.889, 1.045)$. Coxův model jako jediný nebyl na hladině 0.05 zamítnut.

Data podle AFT modelu

Nejprve uvažujme stejné hodnoty kovariáty X a 1000 pozorování z AFT modelu s $\beta_1 = 1$ s Weibullovým základním rozdělením $Wb(\gamma = 10^6, \delta = 5)$. Dosazením dat do AFT modelu jsme dostali odhad $\hat{\beta}_A = 1.00$, p-hodnoty všech neparametrických testů byly nevýznamné. Použitím Coxova modelu vyšel odhad $\hat{\beta}_C = 4.93$, 95%−konfidenční interval rovný $(4.677, 5.178)$ a

simulovaná p-hodnota testu vhodnosti na hladině 0.05 také nevýznamná. Coxův model proto také nezamítáme, což je správně, protože máme Weibullovo rozdělení. P-hodnoty pro test Aalenova modelu byly významné ve většině kvartilů, proto jsme jej zamítli.

Dále jsme pro stejné X simulovali 1000 pozorování z AFT modelu s $\beta_1 = 1$ se základním rozdělením gamma $Ga(a = 1/100, p = 5)$ a lognormálním $LN(\mu = 2, \sigma^2 = 1)$. V obou připadech byly p-hodnoty všech testů vhodnosti při použití AFT modelu nevýznamné. Naopak Coxův a Aalenův model byly na hladině 0.05 zamítnuty.

Data podle Aalenova modelu

Abychom se vyhnuli problémům s generováním při záporných hodnotách kovariát, uvažujeme tentokrát hodnoty $exp(X)$. Vyrobíme 1000 simulovaných pozorování podle Aalenova modelu $\alpha_i(t) = \beta_0(t) + X_i\beta_1(t)$. Opět neuvažujeme cenzorování.

$\beta_0(t)$ i $\beta_1(t)$ jsme volili tak, aby odpovídaly jednak exponenciálnímu rozdělení $Exp(\lambda = 1/100)$, kdy riziková funkce měla tvar $\alpha_i(t) = \lambda_0 + \lambda_1 X_i$, jednak gamma rozdělení $Ga(a = 5, p = 1/100)$ a nakonec lognormálnímu rozdělení $LN(\mu = 2, \sigma^2 = 1)$.

Ve všech případech jedině Aalenův model popisoval data dobře, Coxův i Aalenův model jsme pokaždé na hladině $\alpha = 0.05$ zamítli.

Shrnutí

Testy fungují v uvedených případech tak, jak mají. Ani jednou nebyl zamítnut původní model. Pro Weibullovo základní rozdělení byl v pořádku Coxův i AFT model, pokud jsme generovali z jednoho z nich. Když jsme zkusili dosadit data s gamma a lognormálním základním rozdělením do jiného modelu, než ze kterého byla generována, testy model vždy na hladině 0.05 zamítly.

Generovaná data	P-hodnoty testů vhodnosti			
	Coxův model	AFT model		
	skórový proces	Wilcoxon	Kolmogorov--Smirnov	Kruskal--Wallis
Coxův model				
Základní rozd.				
Weibullovo	0.275	0.72	0.87	0.466
Gamma	0.327	0.028	0.001	0.171
Lognormální	0.553	0.002	< 0.001	< 0.001
AFT model				
Základní rozd.				
Weibullovo	0.056	0.587	0.538	0.223
Gamma	< 0.001	0.456	0.748	0.117
Lognormální	< 0.001	0.159	0.127	0.975
Aalenův model				
Základní rozd.				
Exponenciální	0.01	< 0.001	< 0.001	0.04
Gamma	0.006	< 0.001	< 0.001	< 0.001
Lognormální	0.015	< 0.001	< 0.001	0.006

Tabulka 4.1: P-hodnoty testů dobré shody Coxova a AFT modelu s daty generovanými ze základních modelů s různými základními rizikovými funkcemi

Generovaná data	P-hodnoty testů vhodnosti pro Aalenův model			
	supremový (horní řádek) a integrální (spodní řádek) test			
	I.kvartil	II.kvartil	III.kvartil	IV.kvartil
Coxův model				
Základní rozd.				
Weibullovo	0.005	< 0.001	< 0.001	< 0.001
	0.015	< 0.001	< 0.001	< 0.001
Gamma	< 0.001	< 0.001	< 0.001	< 0.001
	< 0.001	< 0.001	< 0.001	< 0.001
Lognormální	0.017	< 0.001	< 0.001	< 0.001
	0.409	< 0.001	< 0.001	< 0.001
AFT model				
Základní rozd.				
Weibullovo	< 0.001	< 0.001	< 0.001	< 0.001
	< 0.001	< 0.001	< 0.001	< 0.001
Gamma	< 0.001	< 0.001	< 0.001	< 0.001
	< 0.001	< 0.001	< 0.001	< 0.001
Lognormální	< 0.001	< 0.001	< 0.001	0.128
	0.042	< 0.001	< 0.001	0.190
Aalenův model				
Základní rozd.				
Exponenciální	0.449	0.371	0.539	0.288
	0.307	0.395	0.818	0.640
Gamma	0.053	0.303	0.185	0.478
	0.074	0.483	0.152	0.539
Lognormální	0.445	0.674	0.883	0.692
	0.317	0.486	0.842	0.860

Tabulka 4.2: P-hodnoty testů dobré shody Aalenova modelu s daty generovanými ze základních modelů s různými základními rizikovými funkcemi

Data podle Cox-Aalenova modelu

Budeme generovat data podle Cox-Aalenova modelu a zkusíme, jestli testy zařadí kovariáty do správné části. Mějme 1000 pozorování odpovídající modelu

$$\alpha_i(t) = e^{Z_i\beta_Z}(\beta_0(t) + X_i\beta_1(t)),$$

kde $\beta_Z = 1$, Z_i byly generovány $zN(0,1)$ a X_i jako $exp(N(0,1))$. Vyzkoušíme základní rizikové funkce odpovídající exponenciálnímu a log-normálnímu rozdělení.

Používáme testy dobré shody popsané v části 3.2, proporcionální část testujeme pomocí skórového procesu, aditivní část supremovou statistikou kumulativního reziduálního procesu sčítaného podle kvartilů dané proměnné. Používáme vždy 1000 simulací zkoumaného procesu.

Cox-Aalenův model - exponenciální rozdělení

Vezměme $\beta_0(t)$ a $\beta_1(t)$ odpovídající exponenciálnímu rozdělení $exp(1/100)$, tedy konstanty. Když jsme obě kovariáty dosadili do proporcionální části modelu, tedy když jsme data popsali Coxovým modelem, dostali jsme odhad $\hat{\beta}_Z = 0.975$, p-hodnoty testu proporcionality 0.300 pro Z a 0.047 pro X. Proto proporcionalitu pro Z správně nezamítneme a pro X na hladině 0.05 zamítneme.

Když byly obě kovariáty použity v aditivní části modelu, simulovali jsme rozdělení $M_K(t)$, kde K je opět indikátor kvartilů, dostali jsme pro X p-hodnoty supremového testu 0.163, 0.362, 0.111 a 0.464, pro Z pak 0.007, 0.001, 0.003 a 0.001. Proto aditivitu pro Z zamítneme a pro X ne.

Pro X v proporcionální a Z v aditivní části, tj. přesně opačně, než jsme generovali, vyšla p-hodnota testu proporcionality X rovná 0.051 a pro supremový test aditivity pro každý kvartil menší než 0.001. Proto aditivitu Z zamítáme a proporcionalitu X na hladině 0.05 těsně nezamítáme.

Pro správný model, tj. Z v proporcionální části a X v aditivní jsme dostali odhad $\hat{\beta} = 1.02$, p-hodnotu testu proporcionality 0.695 a testu aditivity 0.988, 0.952, 0.390 a 0.433, původní model tedy správně nezamítneme.

Cox-Aalenův modelu - lognormální rozdělení

Vezměme $\beta_0(t)$ a $\beta_1(t)$ odpovídající lognormálnímu rozdělení $LN(\mu = 0, \sigma^2 = 1)$. Když jsme použili Coxův model na obě kovariáty, dostali jsme odhad $\hat{\beta}_Z = 0.986$, p-hodnoty testu proporcionality 0.583 pro Z a 0.235 pro X. Proto proporcionalitu pro X ani pro Z nezamítneme.

Při použití Aalenova modelu vyšly p-hodnoty supremového testu aditivity pro jednotlivé kvartily pro X 0.320, 0.498, 0.947 a 0.805, pro Z pak pro všechny kvartily < 0.001. Pro Z aditivitu tedy zamítáme a pro X nezamítáme.

Když jsme dali kovariáty do opačných částí, tj. X do proporcionální a Z do aditivní, dostali jsme p-hodnotu testu proporcionality X rovnou 0.655 a pro supremový test aditivity pro všechny kvartily < 0.001. Proto aditivní část jasně zamítáme ale proporcionální ne.

Při dosazení do modelu, ze kterého jsme generovali, tj. Z v proporcionální části a X v aditivní, dostali jsme odhad $\hat{\beta} = 1.04$, p-hodnotu testu proporcionality 0.923 a testu aditivity 0.222, 0.673, 0.357 a 0.609, správný model tedy nezamítneme.

Cox-Aalenův model s chybějícími částmi

Mějme opět kovariáty X a Z. Budeme simulovat hodnoty z Coxova modelu s kovariátou Z a z Aalenova modelu z kovariátou X. Použitou kovariátu dosadíme do příslušné částí Cox-Aalenova modelu. Budeme chtít zjistit, zda model správně vyhodnotí nulovost příslušného regresního koeficientu nebo neparametrické části u nevýznamné proměnné, když ji dosadíme postupně do obou částí.

Hodnoty Z jsme zvolili jako výběr z $N(0, 1)$, hodnoty X jako $exp(N(0, 1))$. Generujme 1000 hodnot bez cenzorování, nejprve z Coxova modelu se základní rizikovou funkcí odpovídající rozdělení $exp(1/100)$, tedy $\alpha_i(t) = e^{Z_i}/100$.

Když jsme obě kovariáty popsali pouze Coxovým modelem, dostali jsme odhady $\hat{\beta}_Z = 0.994$, $\hat{\beta}_X = 0.015$. Konfidenční interval pro β_X byl $(-0.011, 0.041)$, p-hodnota Waldova testu nulovosti pro β_X byla rovná 0.257, což značí, že vliv X není významný. P-hodnota testu proporcionality vyšla pro Z 0.907 a

pro X 0.686, model proto nezamítneme.

Když jsme dosadili Z do Coxovy části a X do Aalenovy části, získali jsme odhad $\hat{\beta}_Z = 0.969$. Když jsme testovali, zda v Aalenově části $B_X(t) \equiv 0$ pomocí resamplingu statistiky $sup_{t\in[0,\tau]} \left| \frac{\hat{B}_X(t)}{\sqrt{\widehat{var}\hat{B}_X(t)}} \right|$, dostali jsme p-hodnotu 0.936, tedy že vliv X není významný.

Generujme nyní 1000 hodnot z Aalenova modelu s kovariátou X a oběma částmi základní intenzity odpovídající $exp(1/100)$, tedy $\alpha_i(t) = \frac{1}{100} + \frac{1}{100}X_i$.

Zkusme obě kovariáty dosadit do Aalenova modelu. P-hodnota testu nulovosti $B_X(t)$ založeném na $sup_{t\in[0,\tau]} \left| \frac{\hat{B}_X(t)}{\sqrt{\widehat{var}\hat{B}_X(t)}} \right|$ vyšla < 0.001, pro Z ale 0.273, tedy správně vyhodnotíme vliv Z jako nevýznamný.

Když jsme nechali X v Aalenově části a Z dosadili do Coxovy části, dostali jsme odhad $\hat{\beta}_Z = 0.00964$, konfidenční interval $(-0.054, 0.073)$ a p-hodnotu Waldova testu nulovosti 0.771, tedy taky že vliv Z není významný.

Shrnutí

Cox-Aalenův model byl ve většině případů schopen rozeznat, do které části která kovariáta patří. Nikdy nebyl zamítnut původní model, v několika případech test nezamítl proporcionalitu u kovariáty, která byla původně v aditivní části.

V obou případech, kdy byl model generován jen podle jedné kovariáty, byla nevýznamná část správně detekována.

4.2 Reálné problémy

V reálných problémech z praxe budeme hledat model, který data o přežití nebo výdrži vystihuje nejlépe. V uvedených postupech budeme aplikovat základní tvary regresních funkcí, jak jsou popsány v prvních dvou kapitolách. Uvažujeme tedy jen základní tvar Coxova modelu s proporcionálním vlivem na riziko $e^{X^T(t)\beta}$ a AFT modelu se zrychlením času $e^{Z^T\beta}$. Je možné, že jiný tvar regresní funkce by data popisoval lépe, například pokud bychom provedli vhodné transformace kovariát nebo času.

V případě jedné kovariáty vyzkoušíme postupně všechny tři základní modely. Ověříme, které z nich popisují data dobře, interpretujeme výsledky.

V případě více kovariát zkusíme nejprve samotný Coxův a AFT model, nevýznamné kovariáty přitom vynecháme. Pokud jeden z modelů vystihuje data dobře pro všechny kovariáty, rozhodneme se pro něj. Pokud ne, zkusíme Cox-Aalenův model a do Coxovy části dáme kovariáty u kterých proporcionalita nebyla zamítnuta, ostatní dáme do Aalenovy části.

Všechna použitá data jsou přiložena na CD.

Výdrž součástek z automobilů Tatra

Máme údaje o životnosti součástek z automobilů Tatra při tlakové zkoušce (počet cyklů) při dané zátěži (kp/cm^2) (Kovanic & Volf, 1992). Data jsou cenzorována ($\Delta_i = 1$ necenzorované, $\Delta_i = 0$ cenzorované pozorování), viz tabulka 4.3.

Nejprve zkusíme Coxův model. Získáme odhad $\hat{\beta} = 0.315$, 95%−konfidenční interval je $(0.141, 0.488)$. $\exp(\hat{\beta}) = 1.37$, to znamená, že při vyšší zátěži jsou součástky ve větším riziku. Při 1000 replikacích vyšla simulovaná p-hodnota statistiky $sup_{t\in[0,\tau]}\|U(\hat{\beta}, t)\|$ rovná 0.215. Coxův model proto na hladině 0.05 nezamítáme.

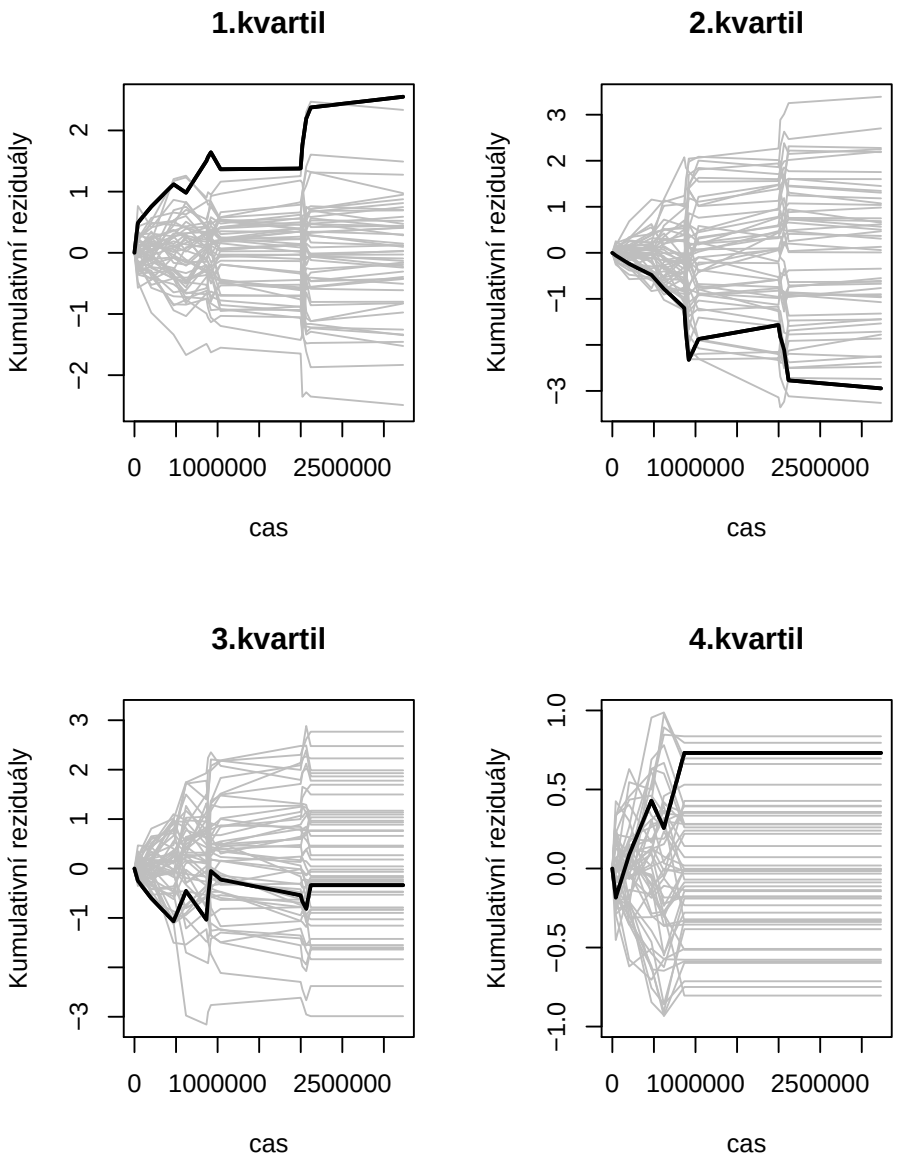
Když zkusíme model se zrychleným časem, získáme odhad $\hat{\beta} = 0.22$. Když ale testujeme shodu rozdělení reziduí necenzorovaných pozorování pod a nad mediánem hodnot zátěže, zjistíme že p-hodnota Wilcoxonova testu je 0.045. Na hladině 0.05 tedy AFT model zamítneme.

Výdrž	Δ_i	Zátěž
41200	1	45.40
470100	1	35.60
865800	1	35.60
202400	1	35.30
620000	1	27.70
884900	1	27.75
919300	1	27.75
2119900	1	27.75
1998900	1	27.50
1036200	1	25.00
11390000	0	25.00
14443900	0	25.00
2020000	1	22.70
2065900	1	22.70
3231400	1	20.20
10064800	0	20.20
9219000	0	17.66

Tabulka 4.3: Výdrž (počet cyklů tlakové zkoušky), zátěž (kp/cm^2) a indikátor necenzorované události pro data o součástkách z automobilů Tatra

Aalenův aditivní model můžeme otestovat opět simulováním kumulativního reziduálního procesu $M_K(t)$. Sčítali jsme reziduály přes kvartily hodnot zátěže, p-hodnota pro $sup_{t\in[0,\tau]}\|M_{Kj}(t)\|$ pro jednotlivé kvartily vyšla rovná 0.003, 0.060, 0.547 a 0.122. Aalenův model tedy nevystihuje závislost dobře (obr.4.1), především pro nízké hodnoty zátěže.

Rozhodneme se tedy pro Coxův model.



Obrázek 4.1: Kumulovaný reziduální proces $M_K(t)$ pro jednotlivé kvartily hodnot zátěže součástek z automobilů Tatra (černě) a jeho hodnoty simulované za hypotézy Aalenova modelu (šedě)

Stávka slévačů

Máme data o výdrži 572 tavících van používaných na zpracování hliníku (počet dní do doby, než musí být vana vyřazena jako příliš opotřebená) (Kalbfleisch & Struthers, 1982). Při generální stávce slévačů v roce 1967 došlo k náhlému odstavení velké části van. Po obnovení výroby začaly být vany méně spolehlivé, patrně vlivem rychlého ochlazení nebo kvůli zanesení ztuhlým materiálem. Máme záznamy o tom, jak dlouho fungovala daná vana před stávkou a po ní a chceme zjistit, zda odstavení mělo významný vliv na životnost. Máme také údaje o životnosti 104 van, které stávkou neprošly.

Zavedeme kovariátu proměnnou v čase, která bude indikovat, zda je daná pec v čase t před nebo po stávce. Nemůžeme použít model se zrychleným časem, protože neumožňuje kovariáty proměnné v čase.

Zkusíme aplikovat nejprve Coxův model. Odhad parametru vyjde $\hat{\beta} = 0.27$, $95\%-$konfidenční interval $(0.119, 0.420)$, tj. že vliv je významný. P-hodnota testu vhodnosti modelu založeném na skórovém procesu (1000 simulací) vyjde 0.256. Na hladině 0.05 proto proporcionalitu nezamítáme. Protože $\exp(\hat{\beta}) = 1.31$, je po stávce riziko zhruba 1.31 krát větší než před stávkou. Na obr.4.2 vidíme odhad základní kumulované rizikové funkce a změnu jejího průběhu, pokud v čase $t = 700$ prošla pec náhlým odstavením.

Mohli bychom zkusit data popsat také Aalenovým aditivním modelem. Uvědomme si ale, že Coxův model je za této situace jeho speciálním případem. V Coxově modelu totiž uvažujeme rizikovou funkci jako (s_i značí čas stávky i-tého jedince):

$$\alpha_i(t) = \begin{cases} \alpha_0(t) & t \leq s_i \\ e^{\beta}\alpha_0(t) & t > s_i, \end{cases}$$

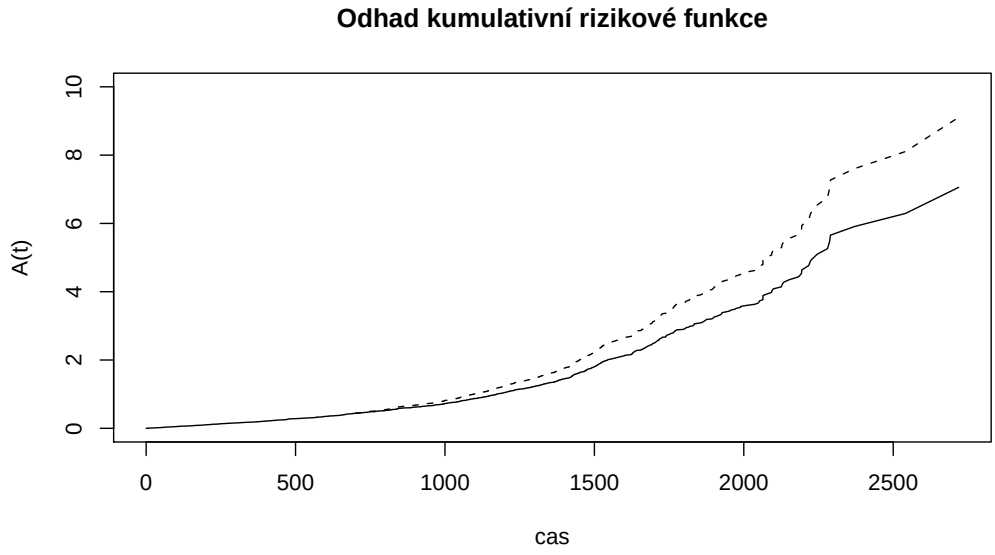
zatímco v Aalenově modelu ve tvaru:

$$\alpha_i(t) = \begin{cases} \alpha_0(t) & t \leq s_i \\ \alpha_0(t) + \alpha_1(t) & t > s_i. \end{cases}$$

α_0 je riziková funkce před stávkou, v obou modelech by měla být tedy stejná. Pro

$$\alpha_1(t) = (e^{\beta} - 1)\alpha_0(t)$$

jsou modely shodné. Vzhledem k tomu, že Coxův model popisoval data dobře a jeho interpretace je přímočařejší, zůstaneme u něj.



Obrázek 4.2: Odhad kumulované základní rizikové funkce výdrže tavících van (plná čára) a její změny v případě, že v čase $t = 700$ došlo k náhlému odstavení (čárkovaně).

Operace tumoru

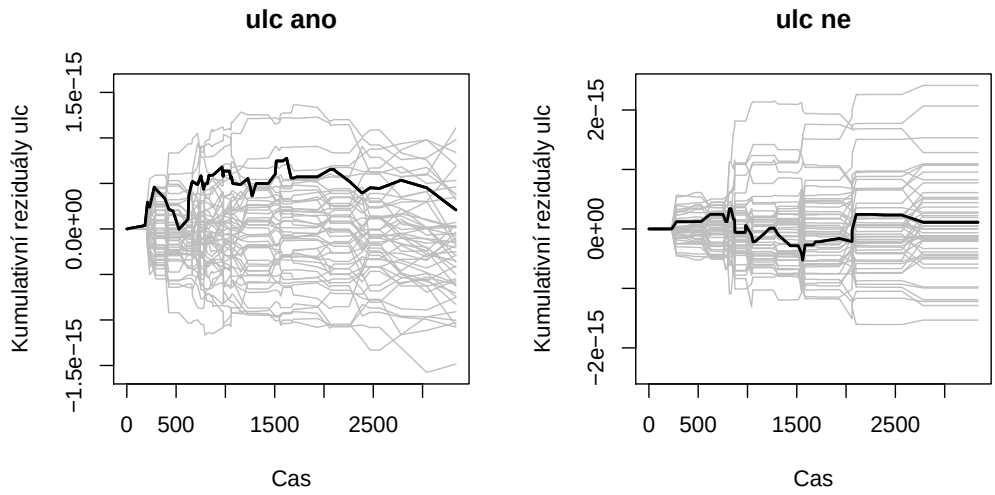
Máme data o počtu dní dožití 205 pacientů po operaci zhoubného nádoru (Drzewiecki (1990), obsaženo v R v knihovně Timereg). K dispozici máme údaj o tloušťce nádoru (`thick`, v setinách milimetru), pohlaví pacienta (`sex`) a indikátor, zda po operaci došlo k hnisavým komplikacím (`ulc`).

Když aplikujeme na data Coxův model, dostaneme odhady $\hat{\beta}^C_{ulc} = 1.1668$, $95\%-$konfidenční interval $(0.5564, 1.7773)$, $\hat{\beta}^C_{thick} = 0.0011$, $95\%-$konfidenční interval $(0.0004, 0.0019)$, $\hat{\beta}^C_{sex} = 0.4595$, $95\%-$konfidenční interval $(-0.0633, 0.9823)$. P-hodnoty testů proporcionality vyšly pro `ulc` 0.046, pro `thick` 0.117 a pro `sex` 0.257. Vliv hnisavých komplikací tedy bude potřeba popsat jiným modelem. Vliv pohlaví není na hladině 0.05 významný.

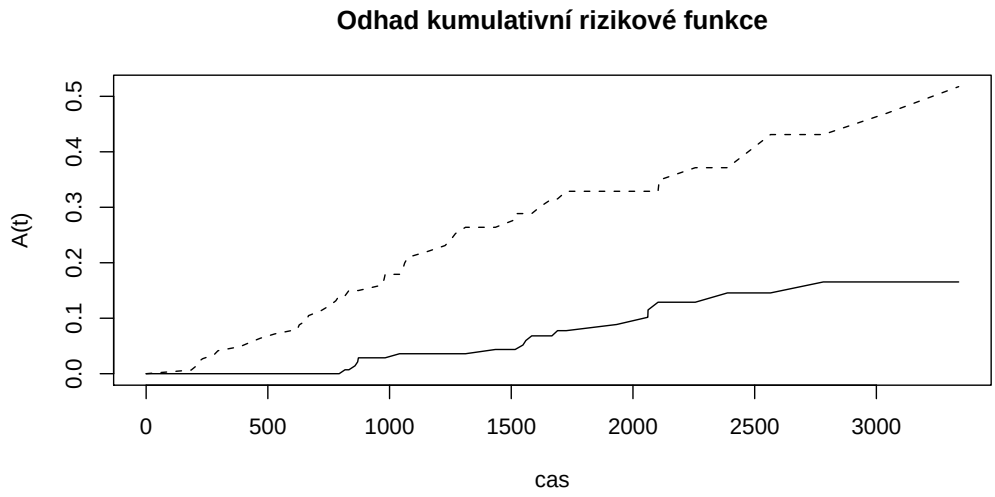
Dosaďme tedy ostatní data do AFT modelu. Dostaneme odhad $\hat{\beta}^A_{ulc} = 0.045$ a $\hat{\beta}^A_{thick} = 0.002$, p-hodnoty Kolmogorov-Smirnovova testu 0.02 pro `ulc` a < 0.001 pro `thick`, Wilcoxonova testu 0.790 pro `ulc` a 0.036 pro `thick`. Testovali jsme proti sobě rezidua pro hodnoty pod a nad mediánem hodnot tloušťky tumoru a rezidua u osob s komplikacemi a bez nich. Kruskal-Wallisův test se pro `ulc` shoduje s Wilcoxonovým, protože máme jen dvě úrovně, pro `thick` dá p-hodnotu 0.058. Závislost na přítomnosti hnisavých komplikacích ani na tloušťce tumoru není AFT modelem popsána dobře.

Zkusme ještě Cox-Aalenův model s kovariátou `ulc` v aditivní a `thick` v proporcionální části. Získáme odhad $\hat{\beta}^C_{thick} = 0.00108$ s konfidenčním intervalem $(0.0004, 0.0018)$, p-hodnota testu proporcionality je 0.356. Aditivní část otestujeme sečtením martingalových reziduálů zvlášť přes jedince, kde se hnisání projevuje, a zvlášť přes ty, kde ne. Dostali jsme simulované p-hodnoty supremového testu 0.840 a 0.416 a integrálního testu 0.945 a 0.406, simulované hodnoty vidíme na obr.4.3.

U tohoto modelu zůstaneme. $exp(0.00108) = 1.0011$, tj. s každou setinou milimetru tloušťky vyoperovaného nádoru stoupá riziko o zhruba 0.11%. Pacienti, u kterých byly zjištěny po operaci hnisavé komplikace, mají také vyšší riziko. Odhad základní rizikové funkce pro pacienty s komplikacemi i bez nich vidíme na obr.4.4.



Obrázek 4.3: Kumulovaný reziduální proces $M_K(t)$ pro jedince s hnisavými komplikacemi a bez nich po operaci zhoubného tumoru (černě) a jeho replikace za platnosti modelu (šedě)



Obrázek 4.4: Odhad základní kumulované rizikové funkce pro pacienty s hnisavými komplikacemi (čárkovaně) a bez nich (plná čára) po operaci zhoubného nádoru

Rakovina plic

Zkoumáme délku života pacientů od projevení pokročilého stádia rakoviny plic (Loprinzi, 1994). Budeme zkoumat závislost přežití na pohlaví pacienta a na jeho věku v době projevení příznaků. Data jsou k dispozici v knihovně Survival v R. Autoři studie doporučují modelovat věk v logaritmické transformaci.

Nejprve zkusíme Coxův model. Odhady vyšly $\hat{\beta}^C_{logage} = 1.0033$, 95%−konfidenční interval $(-0.0897, 2.0963)$, $\hat{\beta}^C_{sex} = 0.5129$, 95%−konfidenční interval $(0.1847, 0.8411)$, přičemž jako základní úroveň bereme ženy, koeficient značí úpravu pro muže. P-hodnota testu proporcionality nebyla významná ani pro věk ani pro pohlaví, Coxův model proto nezamítáme. P-hodnoty testů jsou uvedeny v tabulce 4.4.

Když jsme zkusili AFT model, obdrželi jsme $\hat{\beta}^A_{logage} = 0.760$ a $\hat{\beta}^A_{sex} = 0.387$, p-hodnoty skórového testu nulovosti 0.212 pro věk a 0.007 pro pohlaví. P-hodnota žádného neparametrického testu dobré shody také nebyla významná.

Ani jeden z modelů nebyl zamítnut, základní rozdělení bude patrně relativně blízko Weibullovu. Oba modely označily vliv věku na hladině 0.05 za nevýznamný. Máme $e^{\hat{\beta}^C_{sex}} = 1.670$ a $e^{-\hat{\beta}^A_{sex}} = 0.679$. Závislost na pohlaví můžeme buďto interpretovat tak, že riziková funkce pro muže je 1.67 krát větší než pro ženy, případně že střední doba dožití mužů je pouze 0.679 krát střední doba dožití žen.

Data	P-hodnoty testů vhodnosti			
	Coxův model	AFT model		
	skórový proces	Wilcoxon	Kolmogorov--Smirnov	Kruskal--Wallis
log(věk)	0.176	0.716	0.299	0.509
pohlaví	0.274	0.843	0.287	0.843

Tabulka 4.4: P-hodnoty testů dobré shody Coxova a AFT modelu s daty popisujícími závislost přežití po projevení rakoviny plic na věku a pohlaví

Závěr

Regresní modely v analýze spolehlivosti slouží ke zkoumání vlivu vysvětlujících veličin, které máme k dispozici, na dobu přežití nebo výdrže. Uplatňují se mimo jiné v medicinských a technických studiích.

Volba modelu je klíčová ke kvalitnímu popisu chování dat a následné predikci času přežití dalších jedinců nebo času výdrže dalších součástek. Shrnuli jsme zde základní i některé pokročilé regresní modely, které se výrazně liší původní motivací a ve většině případů i následnou interpretací výsledků. Zatímco u Coxova a Aalenova modelu se zkoumají primárně vlivy kovariát na intenzitu poruch, v modelu se zrychleným časem je vysvětlována střední doba přežití.

Hlavní přínos této práce spočívá jednak ve shrnutí a rozebrání metod jak testovat, zda modely data popisují dostatečně dobře, a především pak v předvedení volby a interpretace regresních modelů v praxi. Postupy pro testování modelů jsou založeny jednak na metodách klasické regrese, a jednak na moderní teorii čítacích procesů a simulačním přístupu. V reálných případech je potřeba nasadit dostatečnou výpočetní sílu, zvlášť pokud máme mnoho dat. Prozkoumali a doplnili jsme proto možnosti, jak modely implementovat v běžném statistickém softwaru.

Pro data, která jsme studovali, bylo vždy možné najít model, který zkoumané závislosti popisoval dostatečně dobře. Pokud se nepodaří najít vhodný model podle metod, které jsme zde používali, přicházely by na řadu další rozšíření. V takových případech můžeme vyzkoušet transformace kovariát nebo času, u Coxova modelu uvažovat koeficienty měnící se v čase nebo změnit tvar závislosti. To by bylo předmětem dalšího zkoumání.

Literatura

[1] Aalen O.O.: *Statistical inference for a family of counting processes*, Univ. of California, Berkeley, 1975.

[2] Aalen O.O.: *A model for non-parametric regression analysis of counting processes*, Mathematical Statistics and Probability Theory, 1–25. Springer Verlag, New York, 1980.

[3] Andersen P. K., Gill R.D.: *Cox's regression model for counting processes: A large sample study*, Ann. Statist. 10, 1100–1120, 1982.

[4] Buckley J., James I.R.: *Linear regression with censored data*, Biometrika 66, 429–436, 1979.

[5] Cox D.R.: *Regression models and life tables*, J. Roy. Statist. Soc. Ser. B 34, 187–220, 1972.

[6] Drzewiecki a kol.: *Malignant melanoma. Changing trends in factors influencing metastasis-free survival from 1964 to 1982*, Cancer 65, 362-366, 1990.

[7] Chen Y., Jewell N.: *On a general class of semiparametric hazards regression models*, Biometrika 88, 687–702, 2001.

[8] Fleming T. R., Harrington D. P.: *Counting Processes and Survival Analysis*, Wiley, New York, 1991.

[9] Kalbfleisch J.D., Struthers C.A.: *An Analysis of the Reynold's Metals. Company Data.*, Canadian Journal of Statistics, 10, 237-259, 1982.

[10] Kovanic P., Volf P.: *Robustní identifikace životnostního modelu*, Robust, 1992.

[11] Lin D.Y., Wei L.J., Ying Z.: *Accelerated failure time models for counting processes*, Biometrika 85, 605–618, 1998.

[12] Loprinzi C.L a kol.: *Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group.*, Journal of Clinical Oncology 12(3), 601-607, 1994.

[13] Martinussen T., Scheike T.H.: *Dynamic Regression Models for Survival Data*, Springer, New York, 2006.

[14] Miller R.G: *Least squares regression with censored data*, Biometrika 63, 449–464, 1976.

[15] Nikulin M., Bagdonavičius V.: *Accelerated Life Models*, Chapman&Hall, 2002.

[16] Scheike T.H., Zhang,M.-J.: *An additive-multiplicative Cox-Aalen model*, Scand. J. Statist. 28, 75–88, 2002.

[17] Therneau T., Grambsch P.: *Modeling Survival Data: Extending the Cox Model*, Springer Verlag, New York, 2000.

Přílohy

Knihovna Timereg

Coxův, Aalenův a Cox-Aalenův model jsou implementovány v R v knihovně TIMEREG
autor: Thomas Scheike,
www: `http://staff.pubhealth.ku.dk/~ts/timereg.html`).
Ukažme, jak potřebné funkce použít.

Coxův model se zavádí jako podmodel Cox-Aalenova modelu, kde v aditivní části je jen základní riziková funkce. Ukažme zavedení a výstup pro příklad se součástkami z automobilů Tatra (výdrž T, zátěž X, indikátory cenzorování D). Chceme 1000 simulovaných hodnot pro odhad rozdělení skórového procesu:

```
> fit<-cox.aalen(Surv(T,D)~prop(X),n.sim=1000)
> summary(fit)
Cox-Aalen Model
...
Test for non-significant effects
            sup|  hat B(t)/SD(t) | p-value H_0: B(t)=0
(Intercept)                 0.717               0.717

Test for time invariant effects
            sup| B(t) - (t/tau)B(tau)| p-value H_0: B(t)=b t
(Intercept)                  0.000158                 0.684

Proportional Cox terms :
        Coef.     SE Robust SE D2log(L)^-1    z   P-val
prop(X) 0.315 0.0692    0.0542      0.0886 4.54 5.5e-06
Test for Proportionality
        sup|  hat U(t) | p-value H_0
prop(X)               7       0.215


  Call:
cox.aalen(Surv(T, D) ~ prop(X), n.sim = 1000)
```

Obdržíme po řadě test nulovosti a konstantnosti základní rizikové funkce, odhady parametrů a jejich rozptylů, testovou statistiku a p-hodnotu Waldova testu nulovosti a především test vhodnosti modelu pro každou kovariátu, založený na skórovém procesu.

Aalenův model se zavádí jako

```
fit<-aalen(Surv(T,D)~X,n.sim=1000,residuals=1))
```

přičemž se automaticky přidává i absolutní člen, tj.

$$\alpha_i(t) = \beta_0(t) + X_i\beta_1(t).$$

parametr `residuals=1` zaručí, že budou k dispozici martingalové reziduály pro testování modelu. Výstup inference získáme jako

```
> summary(fit)
Additive Aalen Model

Test for non-significant effects
            sup|  hat B(t)/SD(t) | p-value H_0: B(t)=0
(Intercept)                  4.59                   0
X                            4.81                   0

Test for time invariant effects
            sup| B(t) - (t/tau)B(tau)| p-value H_0: B(t)=b t
(Intercept)                     2.850                 0.030
X                               0.131                 0.023
            int  (B(t)-(t/tau)B(tau))^2dt p-value H_0: B(t)=b t
(Intercept)                       5970000                 0.068
X                                    8880                 0.106
```

Dostaneme testové statistiky a p-hodnoty testů nulovosti a konstantnosti jednotlivých aditivních částí rizikové funkce. Vykreslíme je přes

```
>plot(fit)
```

Testy vhodnosti modelu pomocí kumulativního reziduálního procesu $M_K(t)$ pro požadovanou matici K (zde matice indikátorů kvartilů X_1):

```
> K<-model.matrix(~-1+cut(X,quantile(X),include.lowest=T))
> colnames(K)<-c("1.kv","2.kv","3.kv","4.kv")
> resids<-cum.residuals(fit,modelmatrix=K,n.sim=1000)
> summary(resids)
Test for cumulative MG-residuals

Grouped Residuals consistent with model
```

```
     sup|  hat B(t) | p-value H_0: B(t)=0
1.kv            2.551               0.003
2.kv            2.945               0.060
3.kv            1.070               0.547
4.kv            0.731               0.122

     int ( B(t) )^2 dt p-value H_0: B(t)=0
1.kv          11143924               0.007
2.kv          13862725               0.063
3.kv           1128603               0.616
4.kv           1458250               0.086
```

Dostaneme testové statistiky a p-hodnoty supremálního i integrálního testu dobré shody pro jednotlivé sloupce K (zde pro jednotlivé kvartily X_1). Skutečné i simulované hodnoty $M_K(t)$ zobrazíme pomocí

```
plot(resids,score=1)
```

Cox-Aalenův model pro X v aditivní a Z v proporcionální části implementujeme následovně (výsledky pro simulovaná data):

```
> summary(fit<-cox.aalen(Surv(T,D)~prop(Z)+X,n.sim=1000,residuals=1))
Cox-Aalen Model

Test for non-significant effects
            sup|  hat B(t)/SD(t) | p-value H_0: B(t)=0
(Intercept)                  9.83                   0
X                            7.66                   0

Test for time invariant effects
            sup| B(t) - (t/tau)B(tau)| p-value H_0: B(t)=b t
(Intercept)                      16.1                 0.067
X                                11.1                 0.183

Proportional Cox terms :
        Coef.     SE Robust SE D2log(L)^-1    z P-val
prop(Z)  1.02 0.0401    0.0425      0.0415 25.1     0
Test for Proportionality
        sup|  hat U(t) | p-value H_0
prop(Z)             16.2        0.695


  Call:
cox.aalen(Surv(tx, d) ~ prop(z) + x, n.sim = 1000, residuals = 1)
```

Pro jednotlivé části dostaneme výstupy jako v předchozích případech. Testy vhodnosti založené na kumulovaném reziduálním procesu $M_K(t)$ získáme úplně stejně jako u samotného Aalenova modelu.

```
> K<-model.matrix(~-1+cut(X,quantile(X),include.lowest=T))
> colnames(K)=c("1.kv","2.kv","3.kv","4.kv")
> resids<-cum.residuals(fit,modelmatrix=K,n.sim=1000)
> summary(resids)
Test for cumulative MG-residuals

Grouped Residuals consistent with model

     sup|  hat B(t) | p-value H_0: B(t)=0
1.kv           18.718               0.988
2.kv           26.798               0.952
3.kv           17.039               0.390
4.kv            7.920               0.433

     int ( B(t) )^2 dt p-value H_0: B(t)=0
1.kv        436268.149               0.852
2.kv        721631.131               0.943
3.kv         50833.714               0.425
4.kv           923.127               0.453
```

Implementace AFT modelu

Pomocí metod popsaných v části 2.2 jsme implementovali v R odhady parametrů a testy vhodnosti modelu se zrychleným časem

$$log(T_i^*) = -\boldsymbol{Z}_i^T\boldsymbol{\beta} + \epsilon_i.$$

Použití:

`aft(tx,d,x,c.test=FALSE)`

Argumenty:

`tx`	vektor pozorovaných časů T_i
`d`	vektor indikátorů cenzorování Δ_i (1=událost, 0=cenzorování)
`x`	matice nebo data frame kovariát Z. Kovariáty můžou být spojité, nebo faktory o dvou úrovních, označených 0 a 1.
`c.test`	logická proměnná, indikující zda se do testů vhodnosti mají (`TRUE`) nebo nemají (`FALSE`) zavést cenzorovaná pozorování.

Výstupy:

data frame, každý řádek odpovídá jedné kovariátě, ve sloupcích je obsaženo:

`covariate`	název kovariáty
`coefficient`	odhad regresních koeficientů $\boldsymbol{\beta}$, pro jednu kovariátu použita funkce `uniroot`, pro více kovariát Nelder-Meadova iterační metoda minimalizace euklidovské normy skóre pomocí funkce `optim`
`pval_wx`	p-hodnota Wilcoxonova testu srovnání reziduí, u spojité kovariáty se porovnávají rezidua s hodnotami příslušné kovariáty pod a nad jejím mediánem, pro faktorovou proměnnou rezidua obou faktorových tříd

`pval_ks`	p-hodnota Kolmogorov-Smirnovova testu srovnání reziduí, ve skupinách stejně jako pro Wilcoxonův test
`pval_kw`	p-hodnota Kruskal-Wallisova testu, pro dvouúrovňovou faktorovou kovariátu shodná s p-hodnotou Wilcoxonova testu, pro spojitou porovnává rezidua rozdělená podle kvartilů hodnot příslušné kovariáty
`pval_0`	p-hodnota skórového testu nulovosti regresních koeficientů

Poznámky:

Jedinci s chybějícími hodnotami některé z kovariát jsou z odhadů a testování vyjmuti.

Testy vhodnosti jsou standardně počítány z pouze necenzorovaných pozorování. Pokud zvolíme `c.test=TRUE`, jsou rezidua u cenzorovaných jedinců nahrazena průměrem všech vyšších reziduí necenzorovaných dat (viz kapitola 2).

Příklad:

```
## delka preziti ve  dnech od projeveni pokrocileho stadia
## rakoviny plic v zavislosti na veku (resp. logaritmu veku)
## a pohlavi (1=muz,2=zena, upraveno na 1 a 0)
## indikator statusu puvodne 1 cenzorovano, 2 udalost,
## upraveno na 0 a 1

> library(survival)
> data(lung)
> attach(lung)
> aft(time,status-1,data.frame(log(age),sex==1),c.test=T)
  covariate coefficient pval_wx pval_ks pval_kw      pval_0
1  log.age.       0.760   0.716   0.299   0.509 0.211663082
2  sex....1       0.387   0.843   0.287   0.843 0.007252759
```