# Diagnostické metody pro model zrychleného času

Petr Novák

Katedra aplikované matematiky, Fakulta informačních technologií ČVUT

14. září 2016

# Analýza přežití

- Pozorujeme $i = 1, ..., n$ jedinců.

- $\mathcal{T}_i$: časy selhání
  $C_i$: časy cenzorování - nezávislé na $\mathcal{T}_i$
  $T_i = \min(\mathcal{T}_i, C_i)$: konec pozorování
  $\Delta_i = I(T_i \leq C_i)$: indikátor necenzorované události
  $\boldsymbol{Z}_i$: regresory.

- $\mathcal{T}_i$ nezávislé, spojité, s hustotami $f_i(t)$, distribučními funkcemi $F_i(t)$, funkcemi přežití $S_i(t) = 1 - F_i(t)$, rizikovými funkcemi

$$\alpha_i(t) = \lim_{h \to 0+} \frac{P(t \leq \mathcal{T}_i < t + h | \mathcal{T}_i \geq t)}{h}.$$

  a kumulovanými rizikovými funkcemi $A_i(t) = \int_0^t \alpha_i(s) ds$.

- Chceme rozumný popis závislosti doby do selhání na regresorech.

# AFT - model zrychleného času

- Accelerated Failure Time - AFT, Buckley & James (1979):

$$\log \mathcal{T}_i = -\boldsymbol{Z}_i^T \boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i \sim \texttt{i.i.d.}$$

- Vnitřní čas každého jedince plyne rychleji nebo pomaleji než pozorovaný, v závislosti na hodnotách regresorů.

- Můžeme si představit jako transformaci času mezi pozorovaným a virtuálním časem

$$t \to t e^{\boldsymbol{Z}_i^T \boldsymbol{\beta}}.$$

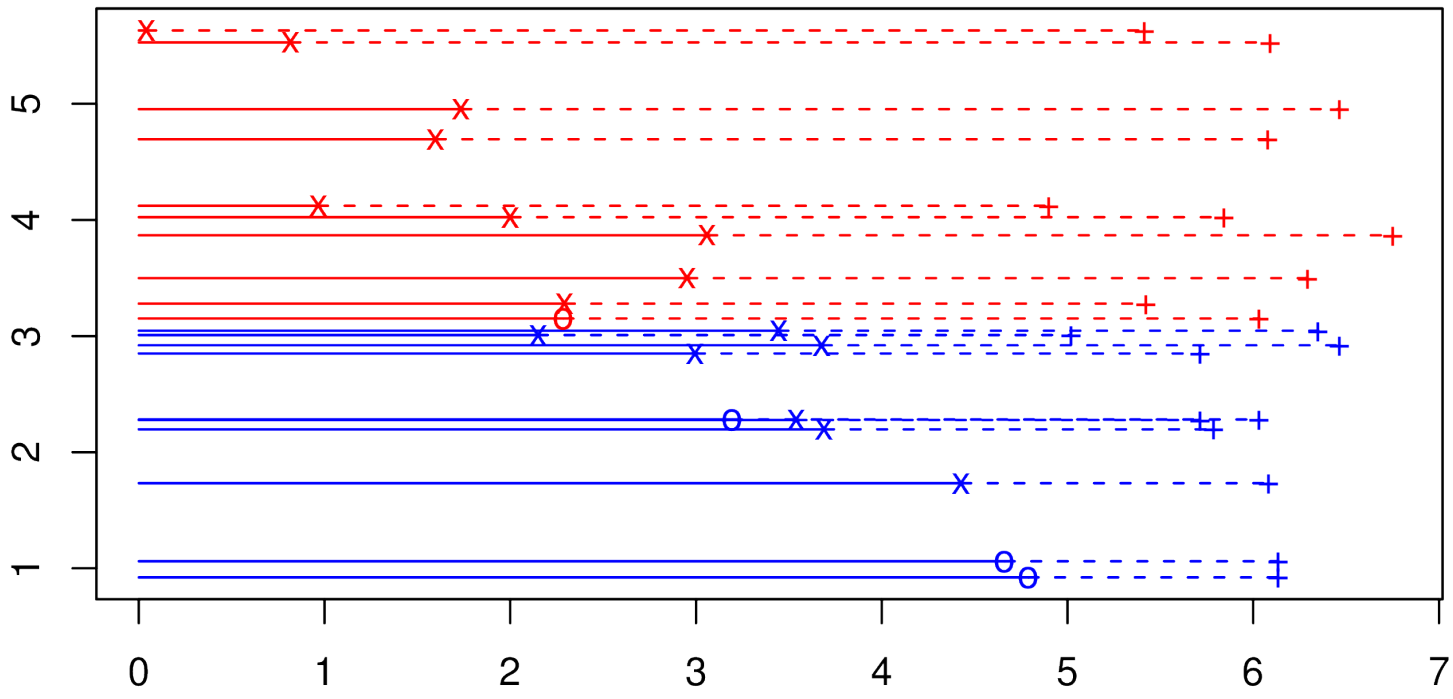
- Riziková funkce

$$\alpha_i(t) = e^{\boldsymbol{Z}_i^T \boldsymbol{\beta}} \alpha_0(t e^{\boldsymbol{Z}_i^T \boldsymbol{\beta}}),$$

kde $\alpha_0(t)$ je základní riziková funkce odpovídající $exp(\epsilon_i)$.

- Pro Weibullovo základní rozdělení je AFT shodný s Coxovým modelem (Cox, 1972).

# Data dle AFT modelu - virtuální a skutečný čas přežití




data from the AFT model (+) and log(T) (x – failure, o – censoring)

# AFT - možnosti inference

Jak odhadnout parametry $\beta$?

(a vyhnout se parametrizování základního rizika)

# AFT - možnosti inference

Jak odhadnout parametry $\beta$?

(a vyhnout se parametrizování základního rizika)

- Metoda maximální věrohodnosti + čítací procesy (Tsiatis, 1990).
- Metoda nejmenších čtverců + korekce pro cenzorovaná pozorování (Buckley & James, 1979).

# AFT - možnosti inference

Jak odhadnout parametry $\beta$?

(a vyhnout se parametrizování základního rizika)

- Metoda maximální věrohodnosti + čítací procesy (Tsiatis, 1990).
- Metoda nejmenších čtverců + korekce pro cenzorovaná pozorování (Buckley & James, 1979).

Existuje mezi těmito metodami spojitost?

# Metoda I - Maximální věrohodnost a čítací procesy

Data lze přepsat pomocí čítacích procesů na transformované škále:

- $N_i^*(t, \beta) = \Delta_i I(e^{\mathbf{Z}_i^T \beta} T_i \leq t)$ indikátor necenzorované události do času $t$,
- $Y_i^*(t, \beta) = I(e^{\mathbf{Z}_i^T \beta} T_i \geq t)$ indikátor, zda je v čase t jedinec ještě v riziku.

Pro kumulativní základní rizikovou funkci pak lze použít Nelson-Aalenův odhad:

$$\hat{A}_0(t) = \int_0^t \frac{dN_\bullet^*(s, \beta)}{\sum_i Y_i^*(s, \beta)}.$$

# Metoda I - Maximální věrohodnost a čítací procesy

Věrohodnost má části pro necenzorované a cenzorované události:

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

Logaritmujeme, dosadíme rizikovou funkci z AFT modelu a přepíšeme pomocí čítacích procesů. Derivujeme, získáme skóre. Místo neznámé základní rizikové funkce dosadíme její odhad. Dostaneme

$$\widetilde{U}(\beta) = \sum_{i=1}^{n} \int_0^{\tau} w_0(s) \left( \boldsymbol{Z}_i - \frac{\sum_{k=1}^{n} Y_k^*(t,\beta) \boldsymbol{Z}_k}{\sum_{k=1}^{n} Y_k^*(t,\beta)} \right) dN_i^*(s,\beta),$$

kde $w_0(s) = (1 + \frac{t\alpha'(t)}{\alpha(t)})$, což se těžko odhaduje.

- Místo $w_0$ lze použít pro odhad libovolnou funkci s rozumnými vlastnostmi, např. $w_1(t) \equiv 1$ (Tsiatis, 1990).
- Odhady získáme minimalizací $\|\widetilde{U}(\beta)\|$.
- Platí $\sqrt{n}(\hat{\beta} - \beta_0) \to \mathcal{N}(\mathbf{0}, \Sigma_w)$, rozptyl závisí na použité váhové funkci $w(t)$.

# Metoda II - Lineární regrese s updatovanou odezvou

- AFT model je modelem cenzorované lineární regrese.
- Pro jednoduchost zaveďme $\mathcal{Y}_i = \log \mathcal{T}_i$ pro skutečná data a $Y_i = \log T_i$ pro cenzorovaná.
- Získáme běžný model

$$\mathcal{Y}_i = -\boldsymbol{Z}_i^T \boldsymbol{\beta} + \epsilon_i,$$

  ve kterém navíc uvažujeme absolutní člen a nulovou střední hodnotu odchylek $\epsilon_i$.
- Kdybychom neuvažovali cenzorování, odhadli bychom $\hat{\boldsymbol{\beta}} = -(Z^T Z)^{-1} Z^T \boldsymbol{\mathcal{Y}}$.
- Místo cenzorovaných pozorování použijeme jejich odhad (Buckley & James, 1976).

# Metoda II - Lineární regrese s updatovanou odezvou

Uvažujme regresní rezidua

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

Pro cenzorovaná pozorování neposkytují aproximaci odchylek $\epsilon_i$ protože jsou systematicky nižší. Použijeme korekci

$$r_i^* := \Delta_i r_i + (1 - \Delta_i)\hat{E}(\epsilon|\epsilon > r_i),$$

kde $E(\epsilon|\epsilon > r_i)$ odhadneme jako

$$\hat{E}(\epsilon|\epsilon > r_i) = \frac{\sum_{j:r_j > r_i} \Delta_j r_j d\hat{F}_0(r_j)}{1 - \hat{F}_0(r_i)}.$$

Zde $\hat{F}_0$ je Kaplan-Meierův odhad distribuční funkce základního rozdělení

$$\hat{F}_0(t) = 1 - \prod_{j:r_j \leq t} \left( \frac{n - rank(r_j)}{n - rank(r_j) + 1} \right)^{\Delta_j}.$$

Tím lze odhadnout skutečné doby událostí jako

$$\mathcal{Y}_i^* = \begin{cases} Y_i & \text{non-cens.} \\ -\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}} + r_i^* & \text{cens.} \end{cases}$$

# Metoda II - Lineární regrese s updatovanou odezvou

Hledáme $\beta$ pomocí metody nejmenších čtverců s updatovanými odezvami, tedy řešíme rovnici

$$\boldsymbol{Z}^T(\mathcal{Y}^* + \boldsymbol{Z}^T\beta) = \mathbf{0}.$$

Updatovaná odezva se dá zapsat vektorově jako (Aziz, 2015):

$$\mathcal{Y}^* = -\boldsymbol{Z}^T\beta + Q(\boldsymbol{Y} + \boldsymbol{Z}^T\beta),$$

kde Q je matice s

$$q_{ij} = \begin{cases} \Delta_i & i = j \\ (1-\Delta_i)\frac{\Delta_j r_j d\hat{F}_0(r_j)}{1-\hat{F}_0(r_i)} & r_i < r_j \\ 0 & \text{jinak} \end{cases}$$

Dosadíme do řešené rovnice a získáme (z $QQ = Q$ a $Q\mathcal{Y} = Q\boldsymbol{Y}$):

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

Zádrhel: $Q$ závisí na $\beta$.

Nutno postupovat iterativně (dosadit $\hat{\beta}^{(m)}$ do $Q$, spočítat $\hat{\beta}^{(m+1)}$, ...).

# Souvislost mezi metodami

Odhady metodou maximální věrohodnosti získáme minimalizací normy skóre

$$\widetilde{U}(\beta, w) = \sum_{i=1}^{n} \int_0^{\tau} w(s) \left( \boldsymbol{Z}_i - \frac{\sum_{k=1}^{n} Y_k^*(t, \beta) \boldsymbol{Z}_k}{\sum_{k=1}^{n} Y_k^*(t, \beta)} \right) dN_i^*(s, \beta).$$

Odhady metodou vážených nejmenších čtverců získáme řešením

$$\Psi(\beta, s) = \sum_{i=1}^{n} \boldsymbol{Z}_i^T s(\mathcal{Y}_i^* + \boldsymbol{Z}_i^T \beta) = \mathbf{0},$$

kde s(t) je váhová funkce, výše $s(t) = t$.

# Souvislost mezi metodami

Odhady metodou maximální věrohodnosti získáme minimalizací normy skóre

$$\widetilde{U}(\beta, w) = \sum_{i=1}^{n} \int_0^{\tau} w(s) \left( \boldsymbol{Z}_i - \frac{\sum_{k=1}^{n} Y_k^*(t, \beta) \boldsymbol{Z}_k}{\sum_{k=1}^{n} Y_k^*(t, \beta)} \right) dN_i^*(s, \beta).$$

Odhady metodou vážených nejmenších čtverců získáme řešením

$$\Psi(\beta, s) = \sum_{i=1}^{n} \boldsymbol{Z}_i^T s(\boldsymbol{\mathcal{Y}}_i^* + \boldsymbol{Z}_i^T \beta) = \mathbf{0},$$

kde s(t) je váhová funkce, výše $s(t) = t$. Dá se ukázat (Ritov, 1990), že za platnosti modelu je

$$\frac{1}{\sqrt{n}} \Psi(\beta, s) = \frac{1}{\sqrt{n}} \widetilde{U}(\beta, w) + o_P(1)$$

pokud

$$w(t) = s(t) - \frac{\int_t^{\infty} s(u) dF_0(u)}{1 - F_0(t)}.$$

Odhady $\hat{\beta}_{MLE}$ a $\hat{\beta}_{LS}$ mají stejné asymptotické vlastnosti.

# Souvislost mezi metodami - využití

Tedy ať získáme odhady tak nebo tak, $\sqrt{n}(\hat{\beta} - \beta_0)$ má asymptoticky nulovou střední hodnotu a konečný rozptyl, který lze odhadnout např. pomocí resamplingu skóre z MLE (Lin et al, 1998).

Váhovou funkci ve skóre odpovídající $s(t) = t$ můžeme bez změny asymptotických vlastností odhadů nahradit

$$\hat{w}(t) = t - \frac{\int_t^\infty u d\hat{F}_0(u)}{1 - \hat{F}_0(t)}.$$

# Regresní diagnostika

Rezidua modelu lze zapsat jako

$$\begin{aligned} \boldsymbol{r}^* &= \boldsymbol{\mathcal{Y}}^* + \boldsymbol{Z}^T\hat{\beta} \\ &= \boldsymbol{\mathcal{Y}}^* - \boldsymbol{Z}^T(\boldsymbol{Z}^T Q\boldsymbol{Z})^{-1}\boldsymbol{Z}^T Q\boldsymbol{Y} \\ &= (I - \boldsymbol{Z}^T(\boldsymbol{Z}^T Q\boldsymbol{Z})^{-1}\boldsymbol{Z}^T Q)\boldsymbol{\mathcal{Y}}^*. \end{aligned}$$

Tedy získáme hat-matici $H^* = \boldsymbol{Z}^T(\boldsymbol{Z}^T Q\boldsymbol{Z})^{-1}\boldsymbol{Z}^T Q$ a $M^*$ matici $M^* = I - H^*$, které lze využít k regresní diagnostice. Např. podle diagonálních prvků (leverage)

$$h^*_{ii} = \boldsymbol{Z}_i^T(\boldsymbol{Z}^T Q\boldsymbol{Z})^{-1}\boldsymbol{Z}^T q_{\bullet i}$$

lze identifikovat vlivná pozorování dle $h^*_{ii} > 2(p+1)/n$ (Smith, 2004).

# Regresní diagnostika

Rezidua modelu lze zapsat jako

$$\begin{aligned} \boldsymbol{r}^* &= \mathcal{Y}^* + \boldsymbol{Z}^T\hat{\beta} \\ &= \mathcal{Y}^* - \boldsymbol{Z}^T(\boldsymbol{Z}^T Q\boldsymbol{Z})^{-1}\boldsymbol{Z}^T Q\boldsymbol{Y} \\ &= (I - \boldsymbol{Z}^T(\boldsymbol{Z}^T Q\boldsymbol{Z})^{-1}\boldsymbol{Z}^T Q)\mathcal{Y}^*. \end{aligned}$$

Tedy získáme hat-matici $H^* = \boldsymbol{Z}^T(\boldsymbol{Z}^T Q\boldsymbol{Z})^{-1}\boldsymbol{Z}^T Q$ a $M^*$ matici $M^* = I - H^*$, které lze využít k regresní diagnostice. Např. podle diagonálních prvků (leverage)

$$h^*_{ii} = \boldsymbol{Z}_i^T(\boldsymbol{Z}^T Q\boldsymbol{Z})^{-1}\boldsymbol{Z}^T q_{\bullet i}$$

lze identifikovat vlivná pozorování dle $h^*_{ii} > 2(p+1)/n$ (Smith, 2004).

- Pro cenzorovaná pozorování je $h^*_{ii} = 0$, takže je za vlivná neoznačíme.
- $H^*$ závisí na odhadnutých parametrech přes $Q$, tedy výsledky budou jen přibližné.

# Testy dobré shody - vzdálenost odhadů

- Za platnosti modelu by obě metody by měly dávat blízké odhady.
- Myšlenka: Odhadneme parametry oběma metodami a testujeme, zda nejsou významně vzdálené pomocí statistiky

$$(\hat{\beta}_{MLE} - \hat{\beta}_{LS})^T \hat{\Sigma}^{-1}_{\hat{\beta}_{MLE} - \hat{\beta}_{LS}} (\hat{\beta}_{MLE} - \hat{\beta}_{LS}),$$

  která bude mít za platnosti modelu $\chi^2_{p+1}$ rozdělení.
- Pro odhad rozptylu využijeme replikace odhadů pomocí bootstrapové replikace skóre dle Lin et al (1998) pro MLE a dle Ritov (1990) pro LS.
- Lze generovat $\hat{\beta}^G_{MLE}$ a $\hat{\beta}^G_{LS}$, pro které má $\sqrt{n}(\hat{\beta}_{MLE} - \hat{\beta}_{LS})$ asymptoticky stejné rozdělení jako $\sqrt{n}(\hat{\beta}^G_{MLE} - \hat{\beta}^G_{LS})$.

# Testy dobré shody - rezidua

- Zkoumáme, zda je závislost dat na určité proměnné popsaná modelem dobře.
- Myšlenka: Rozdělíme data do dvou skupin podle mediánu zkoumané proměnné.
- Spočteme updatovaná rezidua $r_i^*$.
- Za platnosti modelu by rezidua v obou skupinách měla mít shodnou střední hodnotu.
- Kdybychom znali $\beta_0$ a neměli cenzorování, mohli bychom použít vhodný dvouvýběrový test shody středních hodnot (např. t-test nebo Wilcoxonův test).

# Testy dobré shody - rezidua

- Zohledníme vliv odhadnutých parametrů a korekce cenzorování.
- Pro test použijeme statistiku

$$T = \sqrt{n}(\bar{r}^*_{s_1} - \bar{r}^*_{s_2}) = \sqrt{n}(\frac{1}{n_1}\sum_{s_1}(\mathcal{Y}^*_i + \boldsymbol{Z}_i^T\hat{\boldsymbol{\beta}}) - \frac{1}{n_2}\sum_{s_2}(\mathcal{Y}^*_i + \boldsymbol{Z}_i^T\hat{\boldsymbol{\beta}}))$$

$$= \sqrt{n}(\frac{1}{n_1}\sum_{s_1}(\mathcal{Y}_i + \boldsymbol{Z}_i^T\boldsymbol{\beta}_0) - \frac{1}{n_2}\sum_{s_2}(\mathcal{Y}_i + \boldsymbol{Z}_i^T\boldsymbol{\beta}_0))$$

$$+\sqrt{n}(\frac{1}{n_1}\sum_{s_1}(\mathcal{Y}^*_i - \mathcal{Y}_i) - \frac{1}{n_2}\sum_{s_2}(\mathcal{Y}^*_i - \mathcal{Y}_i) + \sqrt{n}(\bar{\boldsymbol{Z}}_{s_1} - \bar{\boldsymbol{Z}}_{s_2})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

- První část odpovídá původním i.i.d. datům, druhá rozdílům mezi updatovanými a skutečnými pozorováními (jen pro cenzorovanou část) a vlastnosti třetí známe z MLE.
- $T$ má asymptoticky normální rozdělení s nulovou střední hodnotou a konečným rozptylem, který lze odhadnout kombinací metod shora.
- Alternativně, statistiku lze přepsat jako $\sqrt{n}\boldsymbol{v}^T\boldsymbol{r}^* = \sqrt{n}\boldsymbol{v}^T M^*\mathcal{Y}^*$, s rozptylem aproximovatelným z $\boldsymbol{v}^T M^*\boldsymbol{v}\sigma^2$, kde $\sigma^2 = \operatorname{var}\epsilon_i$. Matice $M^*$ závisí na $Q$ a tedy i na $\beta$, tedy jejím odhadem ztrácíme přesnost testu.

# Conclusions & Outlook

Závěrem

- Existuje přímá souvislost mezi věrohodnostním přístupem a metodou nejmenších čtverců pro AFT model.
- Asymptotické vlastnosti odhadů jsou shodné.
- Lze využít pro diagnostiku modelu.

Outlook:

- Prozkoumat empirické vlastnosti testů - simulace.
- Vyzkoumat další využitelnost reziduí (log-rank,...).
- Vyzkoušet pro regresory proměnné v čase.

# Reference

**[1]** Aziz, N.: *Analysis and Diagnostics for Censored Regression and Multivariate data*. Ph.D. thesis, Victoria University of Wellington, 2010.

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

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

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

**[5]** Ritov, Y.: *Estimation in a linear regression model with censored data*. Annals of Statistics 18: 303–328, 1990.

**[6]** Smith P.J.: *Using linear regression techniques with censored data*. International Journal of Realibility, Quality and Safety Engineering 11(2): 163–173, 2004.

**[7]** Tsiatis A.A.: *Estimating regression parameters using linear rank tests for censored data*. Annals of Statistics 18: 354–372, 1990.