Coxův model proporcionálního rizika a logistická regrese

Iveta Selingerová

Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

22.4.2013

Proměnné

- **Dichotomická proměnná** např. pohlaví (muž, žena),
- **Kvalitativní proměnné - faktory** n skupin kódujeme pomocí n proměnných (přeparametrizovaný model) nebo pomocí $n-1$ proměnných,
  např. barva vlasů (blond, černé, hnědé, zrzavé)
- **Spojité proměnné** např. věk
- **Interakce faktorů**
  např. pohlaví a barva pleti (černý muž, černá žena, bílý muž, bílá žena)

Kódování pomocí referenční kategorie

- **Dichotomická proměnná** např. pohlaví (muž, žena)
  - proměnná Z
  - $Z = 1$ pro muže
  - $Z = 0$ pro ženy (referenční kategorie)
- **Kvalitativní proměnné - faktory** např. barva vlasů (blond, černé, hnědé, zrzavé)
  - proměnné Z_1, Z_2, Z_3
  - $Z_1 = 1$ blond, $Z_1 = 0$ jinak
  - $Z_2 = 1$ černé, $Z_2 = 0$ jinak
  - $Z_3 = 1$ hnědé, $Z_3 = 0$ jinak
  - zrzavé referenční kategorie

Barva vlasů	blond	černé	hnědé	zrzavé
Z_1	1	0	0	0
Z_2	0	1	0	0
Z_3	0	0	1	0

Kódování sigma omezená (deviation, effect)

- **Dichotomická proměnná** např. pohlaví (muž, žena)
  - proměnná Z
  - $Z = 1$ pro muže
  - $Z = -1$ pro ženy (referenční kategorie)
- **Kvalitativní proměnné - faktory** např. barva vlasů (blond, černé, hnědé, zrzavé)
  - proměnné Z_1, Z_2, Z_3
  - $Z_1 = 1$ blond, $Z_1 = 0$ jinak
  - $Z_2 = 1$ černé, $Z_2 = 0$ jinak
  - $Z_3 = 1$ hnědé, $Z_3 = 0$ jinak
  - zrzavé referenční kategorie

Barva vlasů	blond	černé	hnědé	zrzavé
Z_1	1	0	0	-1
Z_2	0	1	0	-1
Z_3	0	0	1	-1

Úvod

- testujeme vliv faktoru na přežití $\Rightarrow$ porovnání křivek přežití přežití (log-rank či Gehan-Wilcoxonův test)
  - např. srovnáváme přežití mužů a žen
- chceme studovat více faktorů najednou nebo máme kvantitativní proměnné $\Rightarrow$ regresní model
  - např. přežití může záviset na pohlaví, věku, výsledcích vyšetření, typu léčby, ...

Regresní model

- Parametrický model - předpokládáme, že známe rozdělení přežití (Normální, exponenciální, lognormální, ...)
- Semiparametrický model - založen pouze na poměru rizik (Coxův model)
- Neparametrický model, např. jádrové vyhlazování

Úvod

- testujeme vliv faktoru na přežití $\Rightarrow$ porovnání křivek přežití přežití (log-rank či Gehan-Wilcoxonův test)
  - např. srovnáváme přežití mužů a žen
- chceme studovat více faktorů najednou nebo máme kvantitativní proměnné $\Rightarrow$ regresní model
  - např. přežití může záviset na pohlaví, věku, výsledcích vyšetření, typu léčby, ...

Regresní model

- Parametrický model - předpokládáme, že známe rozdělení přežití (Normální, exponenciální, lognormální, ...)
- Semiparametrický model - založen pouze na poměru rizik (Coxův model)
- Neparametrický model, např. jádrové vyhlazování

Definice Coxova modelu

$(T_i, \delta_i, \mathbf{Z}_i(t))$

T_i pozorovaný čas pro i-tého jedince

δ_i indikátor cezorování pro i-tého jedince

$\mathbf{Z}_i(t)$ vektor kovariátů nebo rizikových faktorů pro i-tého jedince, které mohou mít efekt na přežití

- časově závislý, např. výsledek stejného vyšetření při jednotlivých kontrolách
- konstantní - známý v čase 0, např. pohlaví $\mathbf{Z}_i(t) = \mathbf{Z}_i$

Máme p nezávisle proměnných $\mathbf{Z} = (Z_1, \ldots, Z_p)$.

Definice Coxova modelu

Coxův model má tvar

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp(\beta^{\mathbf{T}}\mathbf{Z}) = \lambda_0(t)\exp\left(\sum_{k=1}^{p}\beta_k Z_k\right)$$

$\lambda(t|\mathbf{Z})$ riziková funkce pro jedince v čase t v závislosti na proměnných $\mathbf{Z}$

$\lambda_0(t)$ základní riziková funkce

$\beta^T = (\beta_1, \ldots, \beta_k)$ vektor parametrů

Definice Coxova modelu

Poměr rizik (hazard ratio)

$$\frac{\lambda(t|\mathbf{Z})}{\lambda(t|\mathbf{Z}^*)} = \frac{\lambda_0(t)\exp\left(\sum_{k=1}^{p}\beta_k Z_k\right)}{\lambda_0(t)\exp\left(\sum_{k=1}^{p}\beta_k Z_k^*\right)} = \exp\left(\sum_{k=1}^{p}\beta_k(Z_k - Z_k^*)\right)$$

např. Z_i léčebný efekt
$Z_i = 1$ pacient je léčen, $Z_i = 0$ použito placebo

$$\frac{\lambda(t|\mathbf{Z})}{\lambda(t|\mathbf{Z}^*)} = \exp(\beta_i)$$

Odhad a testování parametrů

Metoda maximální věrohodnosti

$$\frac{\partial L}{\partial L \beta_k} = 0, k = 1, \ldots, p$$

Řešení se obvykle provádí numericky (Newton-Raphsonova či jiné iterační metody)

Testování hypotézy

$$H_0 : \beta_1 = \beta_{10}, \ldots, \beta_q = \beta_{q0}$$

- Waldův test
- Test věrohodnostním poměrem
- Skórový test

Za platnosti nulové hypotézy mají statistiky těchto testů rozdělení χ^2 s q stupni volnosti

Výstavba modelu

Věrohodnostní poměr $LR = -2 \log L$, L je hodnota věrohodnostní funkce pro odhadnuté parametry
Akeikeho informační kritérium $AIC = -2 \log L + kp$, k je počet regresních koeficientů, p je nějaká konstanta (většinou 2)
Schwarzovo informační kritérium $SBC = -2 \log L + k \log n$, n je počet pozorování

- krokový výstavbový princip dopředu
- krokový výstavbový princip dozadu
- krokový výstavbový princip kombinovaný

Odhad funkce přežití

Coxova regrese poskytuje odhad rizikové funkce

$$\hat{\lambda}(t|\mathbf{Z}) = \widehat{\lambda}_0(t) \exp\left(\sum_{k=1}^{p} \hat{\beta}_k Z_k\right)$$

Vztah mezi funkcí přežití a rizikovou funkcí
$S(t) = \exp(-\int_0^t \lambda(s)\mathrm{d}s)$

Odhad funkce přežití

$$\hat{S}(t|\mathbf{Z}) = \widehat{S}_0(t)^{\exp\left(\sum_{k=1}^{p} \hat{\beta}_k Z_k\right)}$$

Logistická regrese

- Závislá proměnná Y, $Y = 1$ nastal sledovaný jev, $Y = 0$ nenastal sledovaný jev
- Pravděpodobnost $p = P(Y = 1|\mathbf{Z})$
- Šance $odds(p) = \frac{p}{1-p}$
- logitová transformace $\ln \frac{p}{1-p}$

$$\ln \frac{p}{1-p} = \beta^{\mathbf{T}}\mathbf{Z} = \sum_{k=1}^{p} \beta_k Z_k$$

$$P(Y = 1|\mathbf{Z}) = \frac{\exp(\beta^{\mathbf{T}}\mathbf{Z})}{1 + \exp(\beta^{\mathbf{T}}\mathbf{Z})}$$

Logistická regrese

- cutpoint=hranice pravděpodobnosti pro zařazení předpovědi
  např. cutpoint=0,5
  předpovíme 1, pokud $P(Y = 1|\mathbf{Z}) \geq 0,5$
  předpovíme 0, pokud $P(Y = 1|\mathbf{Z}) < 0,5$
- poměr šancí (odds ratio)

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{\exp(\beta^{\mathbf{T}}\mathbf{Z})}{\exp(\beta^{\mathbf{T}}\mathbf{Z}^*)} = \exp\left(\sum_{k=1}^{p} \beta_k(Z_k - Z_k^*)\right)$$