BSTT536: Survival Data Analysis

Instructor: Hua Yun Chen, PhD

Division of Epidemiology and Biostatistics
School of Public Health
University of Illinois at Chicago

Table of Content

Cox-Snell residuals

1. Idea: If T has survival distribution $S(t)$, then $-\log S(T)$ is exponentially distributed with unit intensity.

$$r_{ci} = \hat{H}(x_i)\exp(\hat{\beta} z_i)$$

   is approximately exponentially distributed with unit intensity if the model is correct.

2. To account for censoring, the Cox-Snell residual is modified as

$$r'_{ci} = (1 - \delta_i)u + r_{ci},$$

   where $u = 1$ (mean of the exponentially distributed survival time) or as suggested by Crowley and Hu (1977), $u = 0.693$ (median of the exponential distribution).

Cox-Snell residuals

1. The Cox-Snell residuals are always positive with with range $(0, +\infty)$. They do not have the properties of conventional residuals such as symmetic around zero and have mean zero.
2. A test of deviation from exponential distribution with unit intensity indicating potential problems.
3. Graphically, it is better to transformed it into normal by

$$r_i = \Phi^{-1}\{\exp(-r_{ci})\}.$$

4. Probability plot may be useful: plot ordered r_i against expected $(i-1)/(n+1)$.
5. Different variability of the residuals can greatly distorted the picture.

Martingale residuals

1. (Cumulative) Martingale residuals are defined as

$$r_{Mi} = \delta_i - r_{ci}.$$

*It is based on the fact that

$$dM_i(t) = dN_i(t) - Y_i(t)\exp(\beta Z_i)dH_0(t)$$

is a martingale with respect to the natural filtration $\mathcal{F}_t$, *i.e.*, $E(dM_i|\mathcal{F}_t) = 0$, where $N_i(t) = 1_{\{X_i \geq t, \delta_i = 1\}}$ and $Y_i(t) = 1_{\{X_i \geq t\}}$. Therefore,

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)dH_0(u)\exp(\beta Z_i).$$

2. Martingale residuals have mean zero. But the range is between $(-\infty, 1)$. Martingale residuals are also skewed.

Plot of martingale residuals*

1. r_{mi} is not much different from r_{ci}. Plot r_{mi} versus X_i or a transformation of it does not yield much more information.
2. It is more informative to plot

$$M_i(t) = N_i(t) - \sum_{i=1}^{n} \exp(\beta Z_i) H_0\{\min(X_i, t)\}$$

   against time t, where β and H are replaced by estimates.
3. Individual residual may be rescaled the martingale residual by its variance,

$$V_i(t) = \int_0^t d < M_i > (t) = \int_0^t Y_i(u) dH_0(u),$$

   as $M_i(t)/\sqrt{V_i(t)}$ before the plot. The calculation may be done for each failure time.

Deviance residuals and diagnostics

1. Idea: Analog to the generalized linear model, compute the minus twice of the difference of the fitted model versus the full model.
2. Deviance residuals are defined as

$$r_{Di} = \text{sgn}(r_{Mi})[-2\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}]^{1/2}.$$

3. See a simulation in R for comparison.

Schoenfeld residuals

1. Idea: Use the partial likelihood score as the residuals. As a result, the residuals are defined covariate-wise and is nonzero only for failed subjects.
2. Define the Schoenfeld residuals (or partial residuals) for subject i with respect to covariate Z_{ij} as

$$r_{Pij} = \delta_i \left\{ z_{ij} - \frac{\sum_{j=1}^{n} Y_j(X_i) z_{ij} \exp(\hat{\beta} z_i)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\hat{\beta} z_i)} \right\}.$$

3. Rescale Schoenfeld residuals is defined as

$$r_{Pi}^{*} = dVar(\hat{\beta}) r_{Pi},$$

where d is the total number of deaths, $r_{Pi} = (r_{Pi1}, \cdots, r_{Piq})$ and q is the dimension of the covariates.

Schoenfeld residuals explained

1. Schoenfeld residuals are used for diagnosis of proportional hazards voilation. Let the true effect be time-varying as $\beta(t)$.
2. For $\delta_i = 1$, the residuals

$$r_{Pij} = \left\{ z_{ij} - \frac{\sum_{j=1}^{n} Y_j(X_i) z_{ij} \exp(\beta(X_i) z_i)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\beta(X_i) z_i)} \right\} + \frac{\sum_{j=1}^{n} Y_j(X_i) z_{ij} \exp(\beta(X_i) z_i)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\beta(X_i) z_i)} - \frac{\sum_{j=1}^{n} Y_j(X_i) z_{ij} \exp(\hat{\beta} z_i)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\hat{\beta} z_i)}$$

3. Expanding $\beta(t)$ around $\hat{\beta}$ leads the approximation

$$E(r_{Pi} \mid \mathcal{F}_{X_i}) \approx U(\hat{\beta}, X_i)(\beta(X_i) - \hat{\beta}),$$

where

$$U(\hat{\beta}, X_i) = \frac{\sum_{j=1}^{n} Y_j(X_i) z_i^{\otimes 2} \exp(\hat{\beta} z_i)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\hat{\beta} z_i)} - \left\{ \frac{\sum_{j=1}^{n} Y_j(X_i) z_{ij} \exp(\hat{\beta} z_i)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\hat{\beta} z_i)} \right\}^{\otimes 2}$$

Schoenfeld residuals explained (continuing)

1. The weighted schoenfeld residuals

$$wR_{Pi} = U^{-1}(\hat{\beta}, X_i)R_{Pi}.$$

2. An approximation is

$$U(\hat{\beta}, X_i) \approx \sum_{i=1}^{n} \delta_i U(\hat{\beta}, X_i) / \sum_{i=1}^{n} \delta_i.$$

3. The implies

$$U^{-1}(\hat{\beta}, X_i) \approx d\hat{var}(\hat{\beta}),$$

where d is the total number of death and $\hat{var}(\hat{\beta})$ is the conventional estimate of the variance.

Score residuals (modified Schoenfeld residuals)

1. Score residuals

$$r_{Sji} = r_{Pji} + \exp(\hat{\beta} Z_i) \sum_{\{r|X_r \le X_i\}} \frac{(\hat{a}_{jr} - Z_{ji})\delta_r}{\sum_{\{l|X_l \ge X_r\}} \exp(\hat{\beta} Z_l)},$$

where

$$\hat{a}_{jr} = \frac{\sum_{i=1}^{n} Y_i(X_r) Z_{ji} \exp(\hat{\beta} Z_i)}{\sum_{i=1}^{n} Y_i(X_r) \exp(\hat{\beta} Z_i)}.$$

2. *Derivation

$$\begin{aligned}
&\sum_{i=1}^{n} \exp(\hat{\beta} Z_i) \sum_{\{r|X_r \le X_i\}} \frac{(\hat{a}_{jr} - Z_{ji})\delta_r}{\sum_{\{l|X_l \ge X_r\}} \exp(\hat{\beta} Z_l)} \\
&= \sum_{i=1}^{n} \exp(\hat{\beta} Z_i) \sum_{r=1}^{n} Y_i(X_r) \frac{(\hat{a}_{jr} - Z_{ji})\delta_r}{\sum_{l=1}^{n} Y_l(X_r) \exp(\hat{\beta} Z_l)} \\
&= \sum_{r=1}^{n} \delta_r \frac{\hat{a}_{jr} \sum_{i=1}^{n} Y_i(X_r) e^{\hat{\beta} Z_i} - \sum_{i=1}^{n} Y_i(X_r) Z_{ji} e^{\hat{\beta} Z_i}}{\sum_{l=1}^{n} Y_l(X_r) \exp(\hat{\beta} Z_l)} \\
&= 0.
\end{aligned}$$

Example on residual diagnosis

1. Compute the residuals in fitting the Cox regression model to the myeloma data.
2. Use SAS to perform the computation.
3. Plot the results and interpretation.
4. Remedy to any problem identified.

Diagonstics based on residual plots

1. Cox-snell residals plot agaist standard exponential distribution to see any deviations.
2. In martingale residual plot against covariates or the linear predictor $Z\hat{\beta}$, fit a moving average means. If the mean estimates changes substantially, it indicates a problem with the model.
3. In deviance residual plot, large deviance residuals indicate lack of fit.
4. In schoenfeld residual plots, fit a moving average mean function against time, deviation indicates possible missing time-dependent covariates.

Identify influential observations

1. Idea: To identify observations that have big influence on the parameter estimates.
2. Implement: Leave one observation out and fit the same model. Compare parameters estimates with and without the particular observation.

$$D\beta_{ki} = \hat{\beta}_k - \hat{\beta}_{k(-i)},$$

where $\hat{\beta}_{k(-i)}$ is the parameter estimate for β_k with the ith observation excluded from the model fit.

3. An approximate formula to obtain $D\beta_{ki}$ with a single run of the model fit.

$$D\beta_{ki} = \text{ the kth component of } r_{Si} var(\hat{\beta}).$$

Standardized influence is $D\beta_{ki}/se(\hat{\beta}_k)$.

4. The influence of the ith observation on the likelihood is

$$LD_i = 2\{\log L(\hat{\beta} - \log L(\hat{\beta}_{(-i)})\} \approx r'_{Si} Var(\hat{\beta}) r_{Si}.$$

Example on identifying influential observations

1. Compute $D\beta_{(-i)}$ for the model fitted to the myeloma data set.
2. Index plots for $D\beta$.
3. Check for influential observation.
4. Single out the influential observations for further check.

Check proportional hazard assumption

1. Plot $\log(-\log S(t|Z))$ to see if they are parallel for different Z values.
2. Use time dependent-coefficient to check. This can be done by plotting $r^*_{Pij} + \hat{\beta}_j$ against failure times, where r^*_{Pij} is the scaled shoenfeld residual. Dependence on time suggest the violation.
3. Add time-dependent covariate to check.

Residuals for parametric regression models

1. Parametric model

$$\log T_i = \mu + \beta_1 Z_{i1} + \cdots + \beta_p Z_{ip} + \sigma \epsilon_i,$$

   where ϵ_i has a known distribution.

2. Standardized residual

$$r_{Si} = (\log X_i - \mu - \beta_1 Z_{i1} - \cdots - \beta_p Z_{ip})/\hat{\sigma}.$$

3. Cox-Snell residuals

$$r_{Ci} = -\log \hat{S}_i(X_i),$$

   where

$$\hat{S}_i(t_i) = S_\epsilon(r_{Si}).$$

4. Martingale residuals:

$$r_{Mi} = \delta_i - r_{Ci}.$$

5. Deviance residuals:

$$r_{Di} = \mathrm{Sgn}(r_{Mi})[-2\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}]^{1/2}.$$

Residuals and distributions for ϵ

Score residuals: Individual score for the parameters from the parametric log-likelihood.

$$\begin{aligned}
\frac{\partial \log L}{\partial \mu} &= \frac{1}{\sigma}\sum_{i=1}^{n} g(X_i), \\
\frac{\partial \log L}{\partial \sigma} &= \frac{1}{\sigma}\sum_{i=1}^{n} \{Z_i g(X_i) - \delta_i\}, \\
\frac{\partial \log L}{\partial \mu} &= \frac{1}{\sigma}\sum_{i=1}^{n} Z_{ji} g(X_i),
\end{aligned}$$

where

$$g(t) = \frac{(1-\delta) f_\epsilon(t)}{S_\epsilon(t)} - \frac{\delta f_\epsilon'(t)}{f_\epsilon(t)}.$$

Residuals and distributions for ϵ (continuing)

1. Weibull distribution

$$S_\epsilon(t) = \exp(-\exp(t)).$$

2. Log-logistic distribution

$$S_\epsilon(t) = \{1 + \exp(t)\}^{-1}.$$

3. Log-normal distribution

$$S_\epsilon(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right).$$

Identify influential observations

1. $D\theta$ for measuring influence of ith observation on parameter estimator (the difference of with and without the ith observation in the parametric model),

$$D\theta = V(\hat{\theta})s_i,$$

   where $\theta = (\mu, \sigma, \beta)$ and s_i is the likelihood score for the ith observation.

2. Hall's F for measuring the influence of ith observation,

$$F_i = \frac{s_i' R^{-1} s_i}{(p+2)\{1 - s_i' R^{-1} s_i\}}$$

   where $R = \sum_{i=1}^{n} s_i s_i'$.

Identify influential observations (continuing)

1. Hall's C for measuring the influence of ith observation,

$$C_i = \frac{s_i' V(\hat{\theta}) s_i}{\{1 - s_i' V(\hat{\theta}) s_i\}^2}$$

   where $R = \sum_{i=1}^{n} s_i s_i'$.
2. If large F_i or C_i indicates influential of the ith observation. The further analysis can concentrate on the influence of the ith observation on the estimates of the model parameters.

Examples

1. Fit exponential or Weibull model to the myeloma data.
2. Compute the diagnostic statistics.
3. Display and interpret the results.

*Derivation of Deviance residuals

The full model assumes different β for different observations.

$$\log L_{full}(h_i, \forall i) = \sum_{i=1}^{n} \delta_i \left\{ h_i Z_i + \log \lambda_0(X_i) - \int_0^{X_i} e^{h_i Z_i} d\Lambda_0(t) \right\},$$

which is maximized at $\delta_i = \delta_i \Lambda_0(X_i) \exp(\hat{h}_i Z_i)$. The maximum is

$$\log L_{full}(\hat{h}_i, \forall i) = \sum_{i=1}^{n} \left\{ -\delta_i + \delta_i \log \frac{\lambda_0(X_i)}{\Lambda_0(X_i)} \right\},$$

*Derivation of Deviance residuals (continuing)

1. The deviance

$$\begin{aligned} D &= 2\sum_{i=1}^{n}\delta_i\left\{(\hat{h}_i-\hat{\beta})Z_i-\int_0^{X_i}(e^{\hat{h}_i Z_i}-e^{\hat{\beta}Z_i})d\Lambda_0(t)\right\} \\ &= -2\sum_{i=1}^{n}\left\{\delta_i-\Lambda_0(X_i)e^{\hat{\beta}Z_i}+\delta_i\log\Lambda_0(X_i)e^{\hat{\beta}Z_i}\right\} \\ &= -2\sum_{i=1}^{n}\left\{r_{Mi}+\delta_i\log(\delta_i-r_{Mi})\right\}. \end{aligned}$$

2. Deviance residuals

$$r_{Di}=\mathrm{sgn}(r_{Mi})\left[-2\left\{r_{Mi}+\delta_i\log(\delta_i-r_{Mi})\right\}\right]^{1/2}, i=1,\cdots,n$$

satisfy $D=\sum_i r_{Di}^2$

References for diagnostics

1. Barlow, W. E. and prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika*, **75**, 65-74.
2. Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.
3. Schoenfeld, D. Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241.
4. Therneau, T. M., Grambsch, P. M., Fleming, T. R. (1990). Martingal-based residuals for survival models. *Biometrika*, **77**, 147-160.
5. Winnett, A. and SASieni, P. (2001). A note on scaled Schoenfeld residuals for the proportional hazards model. *Biometrika*, **88**, 565-571.