Survival Models

Lecture V. Model diagnostics. Akaike's information criterion. Residuals for Cox proportional hazards model

Akaike's information criterion (AIC), 1

Akaike's information criterion (AIC) is a statistic defined by

$$AIC = -2\ln(\text{maximum likelihood}) + k \cdot p,$$

where

- p is the number of parameters in a model under consideration
- k is predetermined constant

This statistic can be used to compare the possible candidate models, it can be considered as a tradeoff between

- goodness of fit (measured by the log-likelihood)
- model complexity (measured by p)

Akaike's information criterion (AIC), 2

Remark

The choice of k seems to be flexible. Collett (1994) states that the choice of $k = 3$ in the AIC is roughly equivalent to using a 5% significance level in judging the difference between the values of $-2\ln$(maximum likelihood) for two nested models which differ by one to three parameters. He recommends $k = 3$ for general use.

Kim&Tableman (2004) use $k = 2$ in examples. We will follow this choice.

Akaike's information criterion (AIC), 3

Taking $k = 2$ and addressing the parametric regression models discussed earlier, the AIC can be rewritten as

$$AIC = -2\ln(\text{maximum likelihood}) + 2 \cdot (a + b),$$

where

- a is the number of parameters in the specific model
- b is the number of (one-dimensional) covariates

For example

- for the exponential model, $a = 1$
- for the Weibull, log-logistic and log-normal models, $a = 2$

AIC. Motorette data example (1)

Data: motorettes' temperature stress test (*Nelson and Hahn, 1972*)

Goal of the trial: to estimate the failure time distribution of motorettes at design temperature of 130°C.

Setup.

- To accelerate the process, the Mottorettes are tested at 4 temperatures: 150°C, 170°C, 190°C, 220°C
- The temperature is constant for any particular motorette over time
- one single regressor variable is used: $x = 1000/(273.2 + {}^\circ\mathrm{C})$

AIC. Motorette data example (2)

Table 4.1: *Hours to failure of Motorettes*

Temperature	Times
150^0C	All 10 motorettes without failure at 8064 hours
170^0C	1764, 2772, 3444, 3542, 3780, 4860, 5196
	3 motorettes without failure at 5448 hours
190^0C	408, 408, 1344, 1344, 1440
	5 motorettes without failure at 1680 hours
220^0C	408, 408, 504, 504, 504
	5 motorettes without failure at 528 hours
$n = 40$,	n_u = no. of uncensored times = 17

Table 4.2: *Results of fitting parametric models to the Motorette data*

Model	log-likelihood		AIC	
exponential	intercept only	-155.875	311.750 + 2(1)	= 313.750
	both	-151.803	303.606 + 2(1 + 1)	= 307.606
Weibull	intercept only	-155.681	311.363 + 2(2)	= 315.363
	both	-144.345	288.690 + 2(2 + 1)	= 294.690
log-logistic	intercept only	-155.732	311.464 + 2(2)	= 315.464
	both	-144.838	289.676 + 2(2 + 1)	= 295.676
log-normal	intercept only	-155.018	310.036 + 2(2)	= 314.036
	both	-145.867	291.735 + 2(2 + 1)	= 297.735

What is a motorette?

Figure: The 1947 Motorette (Model 20)

Picture taken from `http://www.3wheelers.com/motorette.html`

Cox PH model revisited. Partial likelihood (1)

Let us recall that for the Cox PH model, the hazard function is given by

$$h(t|\underline{x}) = h_0(t) \cdot \exp(\underline{x}'\underline{\beta}),$$

As the baseline hazard is not specified in the Cox model, the likelihood function cannot be fully specified.
This follows directly from $f(\cdot) = h(\cdot)S(\cdot)$: if $h(\cdot)$ depends on baseline hazard, so does the pdf $f(\cdot)$

Therefore Cox (1975) defines a likelihood based on conditional probabilities and free of the baseline hazard. The estimates in Cox's model are found maximizing this partial likelihood.

Cox argues that most of the relevant information about the coefficients $\underline{\beta}$ for regression with censored data is contained in this partial likelihood. Also, to analyze the effect of covariates, there is no need to estimate the baseline hazard function $h_0(t)$

Cox PH model revisited. Partial likelihood (2)

We will derive the partial likelihood heuristically. Let us denote

- t^* – a time at which a death has occured
- $\mathcal{R}(t^*)$ – risk set at time t^* (indices of individuals who are alive and not censored just before t^*)

Now, estimate the probability of one death at t^* (given the risk set $\mathcal{R}(t^*)$) as follows:

$$\begin{aligned}\mathbb{P}\{\text{one death at } t^* | \mathcal{R}(t^*)\} &= \sum_{l \in \mathcal{R}(t^*)} \mathbb{P}\{T_l = t^* | T_l \geq t^*\} \\ &\approx \sum_{l \in \mathcal{R}(t^*)} h(t^* | \underline{x}_l) = \sum_{l \in \mathcal{R}(t^*)} h_0(t^*) \exp(\underline{x}_l' \underline{\beta})\end{aligned}$$

Cox PH model revisited. Partial likelihood (3)

Let us denote

- $t_{(j)}$, $j = 1, \ldots, r \leq n$ – the distinct ordered (uncensored) death times
- $\underline{x}_{(j)}$ – the vector of covariates corresponding to individual who dies at $t_{(j)}$

Then, for each j, we have

$$L_j(\underline{\beta}) = \mathbb{P}\{\text{individual with } \underline{x}_{(j)} \text{ dies at } t_{(j)} | \text{one death in } \mathcal{R}(t_{(j)}) \text{ at } t_{(j)}\}$$

$$= \frac{\mathbb{P}\{\text{individual with } \underline{x}_{(j)} \text{ dies at } t_{(j)} | \text{individual in } \mathcal{R}(t_{(j)})\}}{\mathbb{P}\{\text{one death at } t_{(j)} | \mathcal{R}(t_{(j)})\}}$$

$$= \frac{h_0(t_{(j)}) \exp(\underline{x}'_{(j)}\underline{\beta})}{\sum_{l \in \mathcal{R}(t_{(j)})} h_0(t_{(j)}) \exp(\underline{x}'_l \underline{\beta})} = \frac{\exp(\underline{x}'_{(j)}\underline{\beta})}{\sum_{l \in \mathcal{R}(t_{(j)})} \exp(\underline{x}'_l \underline{\beta})}$$

Cox PH model revisited. Partial likelihood (3)

Now, the partial likelihood function is the product of obtained expressions for $L_j(\underline{\beta})$ over all uncensored death times:

$$L_c(\underline{\beta}) = \prod_{j=1}^{r} L_j(\underline{\beta}) = \prod_{j=1}^{r} \frac{\exp(\underline{x}'_{(j)}\underline{\beta})}{\sum_{l \in \mathcal{R}(t_{(j)})} \exp(\underline{x}'_l \underline{\beta})}$$

Using the failure indicator δ_i ($\delta_i = 1$ if y_i is observed and $\delta_i = 0$ if y_i is censored), we can write

$$L_c(\underline{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\underline{x}'_i \underline{\beta})}{\sum_{l \in \mathcal{R}(y_i)} \exp(\underline{x}'_l \underline{\beta})} \right)^{\delta_i}$$

Cox-Snell residuals (1)

Let us first recall that for any survival time T, the random variable $F(T)$ is distributed uniformly on unit interval $(0, 1)$.

This implies that the random variable $H(T)$ is distributed exponentially with $\lambda = 1$. Now, for individual i with covariate $\underline{x}_i$, we have

$$H(T_i|\underline{x}_i) \sim Exp(1)$$

Hence, if the Cox PH model is correct, then given $\underline{x}_i$, it holds that

$$H(T_i|\underline{x}_i) = H_0(T_i)\exp(\underline{x}_i'\underline{\beta}) \sim Exp(1)$$

Cox-Snell residuals (2)

These observations motivate the use of the following residuals:

Definition (Cox-Snell residuals)

The Cox-Snell residuals (Cox&Snell, 1968) are defined as

$$r_{Ci} = \hat{H}_0(y_i) \exp(\underline{x}_i' \underline{\hat{\beta}}), \quad i = 1, \ldots, n,$$

where $y_i = \min(t_i, c_i)$ and the coefficients $\underline{\hat{\beta}}$ are obtained maximizing Cox's partial likelihood.

Notice that the Cox-Snell residuals are not residuals in the sense of linear models because they are not the difference between the observed and fitted values (nor is their mean equal to zero)

Cox-Snell residuals (3)

If the PH model is correct and the estimated coefficients vector $\hat{\underline{\beta}}$ is close to true vector $\underline{\beta}$, the residuals r_{Ci} should resemble a censored sample from a unit exponential distribution.

Thus, we can construct the following visual test.

- Let us have
  - $H_E(t)$ – the cumulative hazard rate for unit exponential ($Exp(1)$), i.e. the theoretical candidate for cumulative hazard
  - $\hat{H}_{r_C}(t)$ – a consistent estimator of cumulative hazard rate for r_{Ci}-s, i.e. the cumulative hazard estimated from sample
- Then $\hat{H}_{r_C}(t)$ should be close to $H_E(t) = 1 \cdot t$ and for each uncensored r_{Ci}, we should have $\hat{H}_{r_C}(r_{Ci}) \approx r_{Ci}$
- Now, one can simply plot $\hat{H}_{r_C}(r_{Ci})$ against r_{Ci} and observe if the plot is close to a 45° line through origin.

Cox-Snell residuals. General remarks

- The Cox-Snell residuals are most useful for examining the overall fit of a model.
- The closeness of the distribution of the r_{Ci}-s to the unit exponential depends heavily on the assumption that, when $\underline{\beta}$ and H_0 are replaced by their estimates, the probability integral transform $F(T)$ still yields uniform $(0, 1)$ distributed variates. Thus, one must be cautios in case of a small sample.
- Furthermore, departures from the unit exponential distribution may be partly due to the uncertainty in estimating the parameters $\underline{\beta}$ and H_0. This uncertainty is largest in the right-hand tail of the distribution and for small samples.

Martingale residuals. Definition

The martingale residuals are a slight modification of the Cox-Snell residuals.

Definition (Martingale residuals)

When the data is subject to right censoring and the covariates are time-independent (fixed at the start of the study), the martingale residuals are defined by

$$\hat{M}_i = \delta_i - r_{Ci} = \delta_i - \hat{H}_0(y_i)\exp(\underline{x}_i'\underline{\hat{\beta}}), \quad i = 1, \ldots, n,$$

where δ_i is the failure indicator and r_{Ci} is the Cox-Snell residual

In practice, martingale residuals are used

- to identify the best functional form of a covariate (also discretization of a covariate)
- to examine overall goodness-of-fit of a Cox model

Martingale residuals. General remarks

- One can interpret the martingale residual for a subject as the difference between the observed and the expected number of deaths for the individual:
  - δ_i – the number of observed deaths (0 or 1)
  - r_{Ci} – cumulative hazard in $(0, y_i)$ can be interpreted as "how many times we expect the subject to die in this interval"

- The martingale residuals sum to zero; that is, $\sum_{i=1}^{n} \hat{M}_i = 0$.
  For large n, the $\hat{M}_i$-s are an uncorrelated sample from a population with mean zero. However, they are not symmetric around zero because they take values in $(-\infty, 1)$

- For a censored time y_i, the value of martingale residual is always negative

Deviance residuals

Definition (Deviance residuals)

The deviance residuals (Therneau, Grambsch & Fleming, 1990) are defined by

$$D_i = \text{sgn}(\hat{M}_i)\sqrt{-2(\hat{L}_i - \hat{L}_{si})}, \quad i = 1, \ldots, n,$$

where $\hat{M}_i$ is the martingale residual, $\hat{L}_i$ is the likelihood of i-th individual by current model and $\hat{L}_i$ is the likelihood of i-th individual by saturated model.

- Deviance residuals are much more symmetrically shaped than martingale residuals
- Deviance residuals do not necessarily sum to zero
- Therneau, Grambsch & Fleming (1990) note that "*When censoring is minimal, less than 25% or so, these residuals are symmetric at zero. For censoring greater than 40%, a large bolus of points with residuals near zero distorts the normal approximation but the transform is still helpful in symmetrizing the set of residuals.*"
- Deviance residuals are used to detect possible outliers

Schoenfeld residuals (1)

Definition (Schoenfeld residuals)

The kth Schoenfeld residual (Schoenfeld, 1982) defined for the kth subject on the jth explanatory variable $x^{(j)}$ is given by

$$r_{s_{jk}} = \delta_k \{x_k^{(j)} - a_k^{(j)}\},$$

where

- δ_k is the failure indicator of subject k
- $x_k^{(j)}$ is the value of the jth explanatory variable on the kth individual in the study
- $a_k^{(j)} = \frac{\sum_{m \in \mathcal{R}(y_k)} \exp(\underline{x}_m' \underline{\hat{\beta}}) x_m^{(j)}}{\sum_{m \in \mathcal{R}(y_k)} \exp(\underline{x}_m' \underline{\hat{\beta}})}$
- $\mathcal{R}(y_k)$ is the risk set at time y_k

The MLE $\underline{\hat{\beta}}$ is obtained from maximizing the Cox's partial likelihood function $L_c(\underline{\beta})$

Schoenfeld residuals (2)

Note that a Schoenfeld residual $r_{s_{jk}}$ can be considered as the difference between $x^{(j)}$ and a weighted average of the values of explanatory variables over individuals at risk at time y_k.

The weight used for the mth individual in the risk set at y_k is

$$\frac{\exp(\underline{x}'_m\underline{\hat{\beta}})}{\sum_{m\in\mathcal{R}(y_k)}\exp(\underline{x}'_m\underline{\hat{\beta}})},$$

which is the contribution from this individual to the maximized Cox's partial likelihood.

Schoenfeld residuals. Remarks

- Schoenfeld calls these residuals the partial residuals as these residuals are obtained from maximizing the partial likelihood function. Collett (1994, page 155), among others, calls these residuals the score residuals as the first derivative of the partial likelihood can be considered as the efficient score.
- If the assumption of proportional hazards holds, a plot of these residuals against ordered death times should look like a tied down random walk. Otherwise, the plot will show too large residuals at some times.
- Nonzero residuals only arise from uncensored observations.
- Schoenfield residuals are used to examine the fit and detect outlying covariate

Grambsch and Therneau's test for PH assumption

General idea and usage:

- model time-varying coefficients: $\underline{\beta}(t) = \underline{\beta} + \underline{\theta} g(t)$
- main tool: describe the functional form of $\beta(t)$ using scaled Schoenfeld residuals
- under H_0 the function is expected to be constant, i.e. the PH assumption holds
- in R: use `cox.zph` and `plot` to study the behaviour of constructed function