STAT347: Generalized Linear Models
Lecture 15

Today's topics: Survival analysis

- Parametric model for survival function estimation
- Log-rank test: compare between two survival curves
- Proportional hazards regression model

1 Parametric model for the survival functions

We can also assume that T follows some parametric distribution. The most common ones are

- Exponential distribution: $f(t) = \lambda e^{-\lambda t} (\lambda > 0)$
  Then the survival function is $S(t) = e^{-\lambda t}$ and the hazard rate is $h(t) = \lambda$
- Weibull distribution: $f(t) = \kappa \lambda t^{\kappa-1} e^{-\lambda t^{\kappa}} (\lambda, \kappa > 0)$
  Then the survival function is $f(t) = e^{-\lambda t^{\kappa}}$ and the hazard rate is $h(t) = \kappa \lambda t^{\kappa-1}$

The Weibull distribution is relatively simple and allow the hazard rate to either increase/decrease with t.

How to estimate the unknown parameters? A very natural way is to maximize the likelihood, but there is censoring and we only observe $y_s = \min(T_s, C_s)$ for samples $s = 1, 2, \cdots, n$.

1.1 Constructing the likelihood with censoring

For each sample s, assume we observe (y_s, δ_s). We build a likelihood for each sample conditional on C_s (treat C_s as fixed):

- If $\delta_s = 1$, then we observe $T_s = y_s$, the likelihood is $L_s = f(y_s) = S(y_s)h(y_s)$
- If $\delta_s = 0$, then we only observe $T_s \geq y_s$, the likelihood is $L_s = S(y_s)$

Thus the total likelihood is

$$L = \prod_s L(s) = \prod_{s=1}^{n} S(y_s) h(y_s)^{d_s}$$

Specifically

- When T_s follows a common Exponential distribution with unknown parameter λ, $L = \prod_s e^{-\lambda y_s} \lambda^{d_s}$. The log-likelihood is
  $$\log L = -\lambda \sum_s y_s + \log \lambda \sum_s \delta_s$$
  The MLE is
  $$\widehat{\lambda} = \frac{\sum_s d_s}{\sum_s y_s}$$
  When there is no censoring, it is 1 over the average death time.

- When T_s follows a common Weibull distribution with unknown parameters λ and κ, $L = \prod_s e^{-\lambda y_s^\kappa}[\kappa\lambda y_s^{\kappa-1}]^{\delta_s}$. The log-likelihood is

$$\log L = -\lambda \sum_s y_s^\kappa + (\log\kappa + \log\lambda)\sum_s \delta_s + (\kappa - 1)\sum_s \delta_s \log y_s$$

In principle, we can still solve the score equations to find out MLE.

How can we check from the data whether we should use Exponential distribution or Weibull distribution?

- Likelihood ratio test
- Visualize the Kaplan-Meier curve
  - Exponential distribution: $\log S(t) = -\lambda t$, we can check the linearity between t and $\log(\widehat{S}(t))$
  - Weibull distribution: $\log[-\log S(t)] = \log\lambda + \kappa\log t$, we can check the linearity between $\log t$ and $\log[-\log\widehat{S}(t)]$

2 Log-rank test

How to compared between two distributions? In the NCOG data, how can we compare the survive curves of Arm A vs Arm B? We may want to know if the whole survival curve of Arm B is significantly larger than the whole curve of Arm A.

Here, we only consider the simplest null hypothesis: for two groups 1 and 2, we test if the two curves are exactly the same:

$$H_0 : S_1(t) \equiv S_2(t)$$

Let's first discuss the discrete survival time, or we can discretize the survival time into bins. For each bin i or discrete survival time i, assume we observe r_{i1} and r_{i2} samples that are still alive at the beginning of this time bin for each group respectively, and d_{i1} and d_{i2} death during this time bin for two groups respectively. Assume that drop-outs happen at the end of each time bin.

For each bin i, it is basically a 2×2 table

	death	alive	total at risk
Group 1	d_{i1}	$r_{i1} - d_{i1}$	r_{i1}
Group 2	d_{i2}	$r_{i2} - d_{i2}$	r_{i2}
Total	d_i	$r_i - d_i$	r_i

The Cochran-Mantel-Haenszel log-rank test is to test whether the group has no effect on death rate in each table. If the margins of this table are considered fixed, then under H_0, d_{i1} follows a Hypergeometric distribution, with (check the Wikipedia page)

$$E(d_{i1}) = \frac{d_i}{r_i} r_{i1}, \quad \mathrm{Var}(d_{i1}) = \frac{r_{i1} r_{i2} d_i (r_i - d_i)}{r_i^2 (r_i - 1)}$$

The log-rank test statistics is

$$X^2_{CMH} = \frac{\{\sum_i (d_{i1} - r_{i1} d_i / r_i)\}^2}{\sum_i r_{i1} r_{i2} d_i (r_i - d_i)/[r_i^2 (r_i - 1)]}$$

Because across i the data is "almost independent", asymptotically we have under H_0, X^2_{CMH} follows χ^2_1, and we can reject the hypotheses when X^2_{CMH} is too large.

For continuous survival time, we can make the bin finer and finer, and in the limit, the Cochran-Mantel-Haenszel log-rank test statistics is

$$X^2_{CMH} = \frac{\left\{\sum_{j=1}^K (d_{j1} - r_{j1} d_j / r_j)\right\}^2}{\sum_{j=1}^K r_{j1} r_{j2} d_j (r_j - d_j)/[r_j^2 (r_j - 1)]}$$

where $\{\tau_1, \tau_2, \cdots \tau_K\}$ is the set of K distinct uncensored failure times observed in the sample including both two groups, d_{j1} and d_{j2} are the number of death at τ_j for each group respectively, and r_{j1} and r_{j2} are the total number of people who are at risk right before τ_j for each group respectively. $r_j = r_{j1} + r_{j2}$ and $d_j = d_{j1} + d_{j2}$.
Some remarks:

- The asympotitics work when the total number of samples n goes to ∞, so we can have either a fixed K or a growing number of K
- For each 2×2 table, there can be many different tests for the group effect or death, for example testing for the odds ratio being 1 with a logistic regression, the challenge is to combine K different tables and have valid inference when each y_j is very small (exactly 1 when there is no tie).
- The CMH log-rank test is powerful when the survive curves does not across each other. It is most powerful when $h_2(t) = \alpha h_1(t)$
- the Log-rank test is non-parametric, and only depends on the ranks
- A class of weighted Log-rank tests:

$$X_W^2 = \frac{\left\{\sum_{j=1}^K w_j(d_{j1} - r_{j1}d_j/r_j)\right\}^2}{\sum_{j=1}^K w_j^2 r_{j1}r_{j2}d_j(r_j - d_j)/[r_j^2(r_j - 1)]}$$

3 Proportional hazards regression model

Finally, we deal with the covariates. For each sample s, we observe (y_s, X_s, d_s) where X_s is the covariate (for example, group indicators). The proportional hazards (PH) model assumes that

$$h_s(t) = e^{X_s^T \beta} h_0(t)$$

- The model is proposed by David Cox (1972, 1975)
- This is a semi-parametric model as we have no assumption on the baseline hazard function $h_0(t)$
- X does not include the intercept for identifiablity
- proportional hazard:

$$\log\left\{\frac{h_s(t)}{h_0(t)}\right\} = X_s^T \beta$$

The benefit of building a model on the hazard rate instead of survival function is that the survival function need to be less than 1, while the hazard rate does not have that constraint. The benefit of having a proportional model is that there is no constraint on the range of β to have the hazard rate positive.

3.1 the Partial likelihood

For simplicity, we assume no ties. When there are ties, people use the same idea with some adjustments (omitted here).
We look at each failure time, denote the risk set as $\mathcal{R}(t) = \{s : y_s \geq t\}$, which is the set of people that is at risk at time t. Then, for each y_s with $d_s = 1$ where the even is observed, there are $\mathcal{R}(y_s)$ individuals that are at risk, conditional on the fact that there is exactly one person die, the probability that individual s is chosen is then

$$L_s = \frac{h_s(y_s)}{\sum_{l \in \mathcal{R}(y_s)} h_l(y_s)} = \frac{e^{X_s^T \beta}}{\sum_{l \in \mathcal{R}(y_s)} e^{X_l^T \beta}}$$

The partial likelihood for the samples is

$$L = \prod_s L_s^{\delta_s}$$

It is "partial" because it ignores all the non-events, times when nothing happened or there were losses to follow-up.

Remember that the full likelihood is

$$L = \prod_s L(s) = \prod_{s=1}^{n} S_s(y_s) h_s(y_s)^{\delta_s} = \prod_{s=1}^{n} \left(\frac{h_s(y_s)}{\sum_{l \in \mathcal{R}(y_s)} h_l(y_s)} \right)^{\delta_s} \left(\sum_{l \in \mathcal{R}(y_s)} h_l(y_s) \right)^{\delta_s} S_s(y_s)$$

If we ignore the last two terms in the product, we get the partial likelihood. Cox (1972) argued that the first term in this product contained almost all of the information about β, while the last two terms contained the information about $h_0(t)$, the baseline hazard. Efron further justifies this with reasonable assumptions (Efron 1977, JASA 557–565).

3.2 Estimation and inference with the partial likelihood

The log-likelihood:

$$l(\beta) = \log L = \sum_{s=1}^{n} \delta_s \left[X_s^T \beta - \log \left\{ \sum_{t \in \mathcal{R}(y_s)} e^{X_s^T \beta} \right\} \right]$$

$\widehat{\beta}$ can be solved $\dot{l}(\beta) = 0$ and people has taken a lot of effort to show that it has asymptotic distribution (not a trivial result)

$$\widehat{\beta} \dot{\sim} N(\beta, \ddot{l}(\widehat{\beta})^{-1})$$