Chapter 8 Semi-parametric Proportional Hazards (PH) Regression with Fixed Covariates

PubH 7450

©Wei Pan

Email: weip@biostat.umn.edu

Http: www.biostat.umn.edu/∼weip

§8.1 Introduction

- So far, assume event times $X_i \overset{iid}{\sim} S$ (or H or h).
  Observe n iid $(T_i, \delta_i)'s \Longrightarrow \hat{S}, \hat{H}, \hat{h}, ...$
- In practice, we observe some covariates (or predictors or independent or explanatory variables) Z_i, and the distribution of X_i possibly depends on Z_i.
  E.g., X_i: survival time; Z_i: age, gender, smoking status,...
- Goal: model $X_i \sim Z_i$, a problem of ...
  Then can explain possible effects of Z_i on survival...
- Some commonly used models; see also §2.6.
  1. Accelerated Failure Time (AFT) model (Chapter 12 for parametric approaches):
     - An analog of linear models
     - Because X_i is non-negative $\Longrightarrow Y_i = \log(X_i)$.

- AFT model:

$$Y_i = \beta_0 + Z_i'\beta + \epsilon_i,$$

  $\epsilon_i \overset{iid}{\sim} F_0$ with mean 0.

- Model interpretation: as before...
- Model implication: suppose $S_0(x)$ is the survival function of $\exp(\beta_0 + \epsilon_i)$; i.e., $S_0(x) = Pr(\exp(\beta_0 + \epsilon_i) > x)$.

$$\begin{aligned} S(x|Z_i) &= Pr(X_i > x|Z_i) = Pr(Y_i > \log x|Z_i) \\ &= Pr(Y_i - Z_i'\beta > \log x - Z_i'\beta|Z_i) \\ &= Pr(\exp(\beta_0 + \epsilon_i) > \exp(\log x - Z_i'\beta)|Z_i) \\ &= S_0[x\exp(-Z_i'\beta)] \end{aligned}$$

- The effect of Z_i is to change the time scale by a factor of $\exp(-Z_i'\beta)$: if $Z_i'\beta < 0$, then the failure time is accelerated!

- By definition,

$$h(x|Z) = \frac{f(x|Z)}{S(x|Z)} = h_0[x\exp(-Z'\beta)]\exp(-Z'\beta).$$

- With censoring, AFT model is easy to fit with a parametric assumption on S_0–contents of Chapter 12; o/w, harder, no standard software?

2. Proportion hazards model (PHM): by Cox (1972); by far the most popular.

   - PHM:

     $$h(x|Z) = h_0(x)\exp(Z'\beta),$$

     though exp(.) can be any non-negative function. No error term?

   - Model implication:

     $$\frac{h(x|Z_1)}{h(x|Z_2)} = \frac{h_0(x)\exp(Z_1'\beta)}{h_0(x)\exp(Z_2'\beta)} = const$$

—-PH!
$S(x|Z) = \exp(-\int_0^x h(t|Z)dt) =$
$\exp(-\int_0^x h_0(t)\exp(Z'\beta)dt) = [S_0(x)]^{\exp(Z'\beta)} \Longrightarrow$
$\log[-\log S(x|Z)] = Z'\beta + \log[-\log S_0(x)]$, can be used to check the PH assumption.

3. Additive hazard rate model: an alternative to PHM (Chapter 10)

$$h(x|Z) = h_0(x) + Z'\beta.$$

Some constraints have to be put on $Z'\beta$ so that $h(x|Z)$ is non-negative. no standard software?

4. Proportional Odds Model (POM):

$$Logit S(x|Z) = \log Odds(x|Z) = \log\frac{S(x|Z)}{1-S(x|Z)} = G(x) + Z'\beta.$$

Note: different from the POM for ordinal outcomes.

$\log \frac{Odds(x|Z_1)}{Odds(x|Z_2)} = (Z_1 - Z_2)'\beta = const!$

5. Linear Transformation Models:

$$g(X) = Z'\beta + \epsilon.$$

A general class of models!

Case I: $g = \log \Longrightarrow$ AFT;

Case II: g unknown, $\epsilon \sim$ extreme value distribution $\Longrightarrow$ PHM;

Case III: g unknown, $\epsilon \sim$ (standard) logistic distribution $\Longrightarrow$ POM.

§8.2 Model interpretation

- I). A binary covariate, e.g. gender.
  Create a dummy variable; how?–why does it even matter?

- i) use a reference group: $Z = 1$ for M; $= 0$ for F.
  PHM: $h(x|Z) = h_0(x)\exp(Z\beta)$.
  $Z = 0 \Longrightarrow h(x|F) = h_0(x)$;
  $Z = 1 \Longrightarrow h(x|M) = h_0(x)\exp(\beta) \Longrightarrow$

  $$\text{Relative risk (RR, or HR)} = \frac{h(x|M)}{h(x|F)} = \exp(\beta),$$

  or, $\beta = \log RR$ (M vs F).
  Q: is fitting this PHM equivalent to fitting two K-M curves for the two groups?
  Note:
  $\beta = 0 <==> h(x|F) = h(x|M) <==> S(x|F) = S(x|M)$;
  hence, can be used for two-sample comparison!

Q: how is it related to previous K-sample comparisons?

- ii) "sum-to-0" coding: $Z = 1$ for M; $= -1$ for F.
  $RR = \frac{h(x|M)}{h(x|F)} = \frac{h_0(x)\exp(\beta)}{h_0(x)\exp(-\beta)} = \exp(2\beta) \Longrightarrow \beta = \frac{1}{2}RR$ (M vs F).
- II. A categorical variable with K categories: C_1,..., C_K.
- i) Choose C_K (or any other one) as the ref group:
  $Z_j = 1$ for C_j; $Z_j = 0$ o/w; $j = 1, ..., K-1$.
  PHM: $h(x|Z) = h_0(x)\exp(\sum_{k=1}^{K-1} Z_j\beta_j)$.
  $h(x|C_j) = h_0(x)\exp(\beta_j)$ for any $j = 1, ..., K-1$.
  $h(x|C_K) = h_0(x) \Longrightarrow$
  $\beta_j = \log RR$ (C_j vs C_K).
  Q: $\log RR$ (C_j vs C_m)= ... for $1 \le j \ne m \le K-1$?
- ii) Sum-to-0 coding,...
- III. Continuous variables, e.g. age
  PHM: $h(x|Z) = h_0(x)\exp(Age \cdot \beta_1 + \sum_{j=1}^{K} Z_j\beta_j)$.

$\frac{h(x|Age=a+1,Z_2,...,Z_K)}{h(x|Age=a,Z_2,...,Z_K)} = \exp(\beta_1)$;
$\beta_1 = \log RR$ with one-unit increase of Age *after adjusting for other variables* Z_2,...,Z_K!
Assumption: Z_2,...,Z_K can be fixed while Age is changed; always possible?
Q: What is β_1 in PHM $h(x|Age) = h_0(x) \exp(Age \cdot \beta_1)$?
Q: How is this related to confounding?

- IV. Interactions:
  Consider a simple case with two binary variables: 1) gender, $Z_1 = 1$ for F, $= 0$ for M; 2) race, $Z_2 = 1$ for B, $= 0$ for Others. $Z_3 = Z_1 * Z_2$.
  PHM: $h(x|Z) = h_0(x) \exp(\sum_{j=1}^{3} Z_j \beta_j)$.
  $\frac{h(x|F,Others)}{h(x|M,Others)} = \frac{h_0(x)\exp(\beta_1)}{h_0(x)} = \exp(\beta_1)$;
  $\frac{h(x|F,B)}{h(x|M,B)} = \frac{h_0(x)\exp(\beta_1+\beta_2+\beta_3)}{h_0(x)\exp(0+\beta_2+0)} = \exp(\beta_1+\beta_3)$;
  Q: how to relate this to effect-modification? Consider Race as a stratifier...

Note: the above model is equivalent to coding for the 4 groups: FB, MB, FO, MO.

- Finally, note that there is no intercept term in any PHM; should we add one? why or why not?

§8.3 Partial likelihood

- Given: 1) (T_i, δ_i, Z_i), $i = 1, ..., n$;
  2) PHM: $h(x|Z) = h_0(x) \exp(Z'\beta)$.
- Goal: to infer β semi-parametrically. How?
- Approach 1: write down the nonparametric likelihood for (h_0, β), then get NPMLE $\hat{h}_0$, $\hat{\beta}$,...
  Downside: complicated because NPL depends on infinite-dimensional h_0, while most often, interest is in β, not h_0.
- Approach 2: use partial likelihood (PL) proposed by Cox (1972).
  h_0 is treated as a nuisance parameter and eliminated from the PL.
  PL is not a standard likelihood, but it has (almost) all nice properties of a standard likelihood, e.g., asymptotics.

- What is PL?
- Notation:

  1) Suppose that distinct event times are $t_1 < t_2 < ... < t_D$ and no tied event times.

  2) $R(t_i)$: risk set at t_i; subjects who are still in the study at t_i^-.

  3) $Z_{(i)}$: covariate (vector) of the subject who has event at t_i.
- PL

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(Z'_{(i)}\beta)}{\sum_{j \in R(t_i)} \exp(Z'_j \beta)}.$$

- The maximum partial likelihood estimator (MPLE)

$$\hat{\beta} = argmax_\beta \log L(\beta),$$

  which solves the score equation $U(\beta) = (U_1, ..., U_p)' = 0$ with

$$U_k = \frac{\partial \log L(\beta)}{\partial \beta_k} = \sum_{i=1}^{D} \left(Z_{(i)k} - \frac{\sum_{j \in R(t_i)} Z_{jk} \exp(Z'_j \beta)}{\sum_{j \in R(t_i)} \exp(Z'_j \beta)} \right).$$

Information matrix is $I = (I_{gh})_{p\times p}$ with

$$I_{gh} = -\frac{\partial^2 \log L(\beta)}{\partial\beta_g \partial\beta_h} = ...$$

see p.254.

- Then we use the usual Wald test, score test and LRT to test H_0: $\beta = \beta_0$.
  The usual likelihood theory applies!

- Example 8.1.
  Data: (T_i, δ_i, Z_i), $i = 1, ..., n$; no tied event times;
  Z_i is binary: $Z_i = 1$ for group 1; $Z_i = 0$ for group 0.
  PHM: $h(x|Z) = h_0(x)\exp(Z\beta)$.

$$\log L(\beta) \quad = \quad \sum_{i=1}^{D} \left\{ Z_{(i)}\beta - \log[\sum_{j\in R(t_i)} \exp(Z_j\beta)] \right\}$$

$$= \quad d_1\beta - \sum_{i=1}^{D} \log(y_{0i} + y_{1i}e^{\beta})$$

where $d_1 = \sum_{i=1}^{D} Z_{(i)}$ =# events in group 1;
y_{0i} =# subjects from group 0 at risk at t_i^-;
y_{1i} =# subjects from group 1 at risk at t_i^-.
$U(\beta) = d_1 - \sum_{i=1}^{D} \frac{y_{1i}e^{\beta}}{y_{0i}+y_{1i}e^{\beta}}$.
$I(\beta) = \sum_{i=1}^{D} \left(\frac{y_{1i}e^{\beta}}{y_{0i}+y_{1i}e^{\beta}} - \frac{(y_{1i}e^{\beta})^2}{(y_{0i}+y_{1i}e^{\beta})^2} \right)$.
To test H_0: $\beta = 0$, apply the score test:
$U(0) = d_1 - \sum_{i=1}^{D} \frac{y_{1i}}{y_{0i}+y_{1i}} = d_1 - \sum_{i=1}^{D} \frac{y_{1i}}{y_i} * d_i^*$ since $d_i^* = ...$
$I(0) = \sum_{i=1}^{D} \frac{y_{0i}y_{1i}}{y_i^2}$,
$\chi_S^2 = \frac{U(0)^2}{I(0)}$=...... statistic!

- SAS example.

- Some justifications for PL: $L = \prod_{i=1}^{D} L_i$

- 1) Interpretation of L_i:

$$\begin{aligned} L_i &= \frac{\exp(Z'_{(i)}\beta)}{\sum_{j\in R(t_i)} \exp(Z'_j\beta)} \\ &= Pr(\text{subject } (i) \text{ has event at } t_i | \text{one event at } t_i). \end{aligned}$$

Why?

$$\begin{aligned} & Pr((i) \text{ has event in } (t_i, t_i+\Delta t) | \text{an event in } (t_i, t_i+\Delta t)) \\ = \quad & \frac{h(t_i|Z_{(i)})\Delta t \prod_{m\in R(t_i), m\neq (i)} [1 - h(t_i|Z_m)\Delta t]}{\sum_{j\in R(t_i)} h(t_i|Z_j)\Delta t \prod_{m\in R(t_i), m\neq j} [1 - h(t_i|Z_m)\Delta t]} \\ \xrightarrow{\Delta t \to 0} \quad & \frac{h(t_i|Z_{(i)})}{\sum_{j\in R(t_i)} h(t_i|Z_j)} \\ \overset{PHM}{=} \quad & L_i \end{aligned}$$

why called “partial” likelihood? compared to a full likelihood.

- 2) PL is a profile likelihood from the full likelihood $L_f(\beta, h_0)$: $L(\beta) = L_f(\beta, \hat{h}_0(\beta))$.
for details, see p.258.

$$\begin{aligned}
L_f(\beta, h_0) &= \prod_{j=1}^{n} f(T_j|Z_j)^{\delta_j} S(T_j|Z_j)^{1-\delta_j} \\
&= \prod_{j=1}^{n} h(T_j|Z_j)^{\delta_j} S(T_j|Z_j) \\
&\overset{PHM}{=} \prod_{j=1}^{n} [h_0(T_j)\exp(Z_j'\beta)]^{\delta_j} \exp[-H_0(T_j)\exp(Z_j'\beta)].
\end{aligned}$$

- 3) PL is a marginal likelihood for observed event ranks; see p.127 of Miller (1981).

§8.4 PL with tied event times

- Problem: previous derivation of PL is based on the assumption that there are no tied event times. In practice, because of rounding, it can happen. There are several ways to handle tied event times, giving different PLs.

- Notation:
  D_i: subjects having event at t_i; $d_i = |D_i|$.
  R_i: risk set at t_i;
  Q_i: set of all size-d_i subsets of R_i ;

- 1) Brelsow's approximation:
  $L(\beta) = \prod_{i=1}^{D} \prod_{k=1}^{d_i} \frac{\exp(Z_k'\beta)}{\sum_{j \in R_i} \exp(Z_j'\beta)}$.

- 2) Efron's approximation:
  $L(\beta) = \prod_{i=1}^{D} \prod_{k=1}^{d_i} \frac{\exp(Z_k'\beta)}{\sum_{j \in R_i} \exp(Z_j'\beta) - \frac{k-1}{d_i} \sum_{m \in D_i} \exp(Z_m'\beta)}$.

- 3) Cox's approximation based on a discrete model:

$L(\beta) = \prod_{i=1}^{D} \frac{\prod_{k=1}^{d_i} \exp(Z_k'\beta)}{\sum_{S \in Q_i} \prod_{m \in S} \exp(Z_m'\beta)}$.

4) Exact: L_i is the average of the partial likelihoods corrsponding to all d_i! possible permutations of D_i; in SAS

- An example: $D_i = \{1, 2\}$;
  $L_{i,1}$ is PL for that subject 1 a had an event first, then subject 2;
  $L_{i,1} = \frac{exp(Z_1'\beta)}{\sum_{j \in R_i} \exp(Z_j'\beta)} \frac{exp(Z_2'\beta)}{\sum_{j \in (R_i - \{1\})} \exp(Z_j'\beta)}$.
  $L_{i,2}$ is PL for that subject 2 a had an event first, then subject 1;
  $L_{i,2} = \frac{exp(Z_2'\beta)}{\sum_{j \in R_i} \exp(Z_j'\beta)} \frac{exp(Z_1'\beta)}{\sum_{j \in (R_i - \{2\})} \exp(Z_j'\beta)}$.
  $L_i = (L_{i,1} + L_{i,2})/2$.
- Example 8.4
- PHM may not hold. In Example 8.4, how to check the PH assumption?
  $\log H(x|Z = 1) - \log H(x|Z = 0) = ... = \beta =$const.

Then apply ...

Fig 8.1.

§8.5 Local tests

- PHM: $h(t|Z) = h_0(t)\exp(Z'\beta)$.
- Partition $\beta = (\beta_1', \beta_2')'$
- Global test: H_0: $\beta = \beta_0$
- Local test: H_0: $\beta_1 = \beta_{10}$

  More generally, H_0: $C\beta_1 = b_0$ with a specified matrix C and vector b_0.
- Example: H_0: $\beta_1 = \beta_2 = \beta_3$. $\Longrightarrow \beta_1 = \beta_2$ and $\beta_2 = \beta_3$

  $\Longrightarrow C\beta = (0, 0)'$ with $C = ...$
- Score test: least demanding in terms of computation; but least popular in computer packages.

  Read Example 8.2 and formula (8.5.3) on p.264.
- LRT: easiest to apply with a computer package.

  i) fit a full model $\Longrightarrow -2\log L_F$;

ii) fit a reduced model under $H_0 \Longrightarrow -2\log L_R$;
iii) $2\log L_F - 2\log L_R \overset{a.}{\sim} \chi^2_{p_1}$ under H_0, where p_1= difference of # parameters in H_0 and H_1, usually $p_1 = dim(\beta_1)$, but NOT always.

- Wald test: most commonly used.
  To test H_0: $\beta_1 = \beta_{10}$,
  i) fit a full model $\Longrightarrow \hat{\beta}_1, Cov(\hat{\beta}_1)$;
  ii)$\chi^2_W = (\hat{\beta}_1 - \beta_{10})' Cov(\hat{\beta}_1)^{-1}(\hat{\beta}_1 - \beta_{10}) \overset{a.}{\sim} \chi^2_{p_1}$ under H_0.
  To test H_0: $C\beta_1 = b_0$,
  $\chi^2_W = (C\hat{\beta}_1 - b_0)'[CCov(\hat{\beta}_1)C']^{-1}(C\hat{\beta}_1 - b_0) \overset{a.}{\sim} \chi^2_{p_1}$ under H_0, with $p_1 = rank(C)$.

- Example 8.2: SAS

- Example 8.2: R

§8.8 Estimation of the survival function

- Main idea: PL $\Longrightarrow \hat{\beta} \Longrightarrow \hat{H}_0(t)$, $\hat{S}_0(t) \Longrightarrow \hat{S}(t|Z)$ for any Z.
- Use the same notation with t_i, d_i and $R(t_i)$ as before; Brelsow estimator

$$\hat{H}_0(t) = \sum_{t_i \le t} \frac{d_i}{\sum_{j \in R(t_i)} \exp(Z_j' \hat{\beta})}.$$

  Notes: 1) reduces to the N-A estimator if $\hat{\beta} = 0$; 2) derived based on a profile likelihood argument.

- Baseline survival

$$\hat{S}_0(t) = \exp(-\hat{H}_0(t)).$$

- For any Z,

$$\hat{S}(t|Z) = \hat{S}_0(t)^{\exp(Z'\hat{\beta})}.$$

- Example 8.2: SAS

- Q: for a 2-sample problem, does applying the N-A estimator to each group give the same result as using a PHM?

§8.7 Model building

- Practice: model selection$\Longrightarrow$model checking $\Longrightarrow$...
  Here (and often), model selection is mainly on variable selection.
- Why do model selection?
  try to find an approximately correct model.
- What happens if not?
- A too small model:
  Consequence:
  True: $E(Y|X_1, X_2) = b_0 + b_1 X_1 + b_2 X_2$
  Working: $E(Y|X_1) = \beta_0 + \beta_1 X_1$
  Consequence: $E(\hat{\beta}_1) \neq b_1$ unless X_1 and X_2 are ... (i.e. when X_2 is NOT a ...)
  For a non-linear model (e.g. logistic regression or PHM), $E(\hat{\beta}_1) \neq b_1$ always! one reason: interpretation of β_1 and b_1 are

...

- A too large model:
  True: $E(Y|X_1, X_2) = b_0 + b_1 X_1$
  Working: $E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  Note: the working model is not really wrong: it covers the true model as a special case with $\beta_2 = 0$. Does it mean that a larger model is always better?
  Consequences: 1) $E(\hat{\beta}_1) = b_1$, unbiased!
  2) $Var(\hat{\beta}_1) \geq Var(\hat{b}_1)$
  3) $MSE(\hat{\beta}_1) \geq MSE(\hat{b}_1)$
  $MSE(\hat{\beta}_1) = E[(\hat{\beta}_1 - \beta_1)^2] = E[(\hat{\beta}_1 - E(\hat{\beta}_1))^2] + [E(\hat{\beta}_1) - \beta_1]^2 =$ bias^2 + Variance. —-celebrated bias-variance trade-off!

- For a small/finite sample size, is it possible to have $MSE(\hat{\beta}_1) < MSE(\hat{b}_1)$?

- In biomedical studies, because one (perhaps incorrectly) cares more about bias than variance, use a slightly smaller or larger model?

- Variable selection:
- 1) Sequential: forward, backward, stepwise
  Advantage: simple
  Disadvantage: because of its greedy nature, may miss the best model.
- 2) Best subset: computationally intensive.
  Give top-ranked models for each model size.
  How to compare two nested models? – H.T.
  How to compare two non-nested models?
  $AIC = -2\log L(\hat{\beta}) + 2p$
  $BIC = -2\log L(\hat{\beta}) + \log(n)p$
  $p = dim(\beta)$, $\hat{\beta}$ is MLE.
  Note, each criterion = "goodness-of-fit" + "penalty on model complexity". —strike a balance!

Why not choose a model with the highest GOF?

- Comparison b/w AIC and BIC:
  BIC tends to select a smaller model, because ...
  BIC is consistent if the true model is fixed.
  If more concerned with bias (e.g. caused by confounders), go with AIC.
  AIC/BIC: a difference less than 2 is not significant.
- Example: SAS

§Some questions

- Is there *the* correct model? Why or why not?
  true: $E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  $\Longrightarrow$

- More practically,
  George Box's famous quote: "all models are wrong, but some are useful."
  Box and Draper, *Empirical Model-Building and Response Surfaces* (1987), p. 74: Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.
  A key: keep in mind what is your question!
  Example 1: shoe size $\sim$ test score;
  Example 2: trt $\rightarrow$ blood pressure $\rightarrow$ heart disease;

- Why do model selection criteria (e.g. AIC vs BIC vs p-values) sometimes select different models?
  A short answer:*different* criteria are **different**!
  A key: nothing can replace knowledge!
- Can a wrong model beat a correct model (if exists)?
  Prediction: bias-variance trade-off!
  Inference: a surprise!