BIOS 7323 - Exam 2 Review

Chapters 1-4, 7

- Relationships between pdf, survival function, hazard function, cumulative hazard function
- Recognize commonly used distributions (Exponential, Weibull, Gamma)
- Non-parametric estimates of basic quantities, esp. Kaplan-Meier survival estimates
- Hypothesis tests, esp. log-rank test

Chapter 8 - Semi-parametric Proportional Hazards (PH) models

- Know form of PH model: $h(t|Z) = h_0(t)C(\beta' Z)$, usually $C(t) = \exp(t)$
- Recognize and be able to calculate the partial likelihood for distinct event time data

  $\mathcal{PL} = \prod_{i=1}^{D} \frac{\exp(\beta' Z_{(i)})}{\sum_{j \in R(t_i)} \exp(\beta' Z_{(j)})}$, D are uniq. death times

  $R(t_i) = \{j : 1 \leq j \leq n, T_j \geq t_i\}$ ind. in study just prior to t_i

  Note: PH model uses ranks and censoring, not actual times

- Know the three methods for dealing with ties (Breslow, Efron, and Cox) and how they differ

  with ties $d_i \geq 1$ at each $t_i, i = 1, \ldots, D$, $D(t_i)$=set of ind. who die at t_i

  Breslow - Use naive PL, assume no ties. good w/ few ties

  $\mathcal{PL}_1(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' Z_{(i)})}{[\sum_{j \in R(t_i)} \exp(\beta' Z_{(j)})]^{d_i}}$

  Efron - Based on discrete hazard model. Closer to correct PL than Breslow. Breslow & Efron similar with small # ties.

  $\mathcal{PL}_2(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' S_i)}{\prod_{j=1}^{d_i} [\sum_{k \in R(t_i)} \exp(\beta' Z_k) - \frac{j-1}{d_i} \sum_{k \in D(t_i)} \exp(\beta' Z_k)]}$

  Cox - Exact, but complicated & computationally intensive. Q_i is set of all subsets of the d_i individuals who could be selected from risk set. $q = \{q_1, \ldots, q_{d_i}\}$ and $S_q^* = \sum_{j=1}^{d_i} Z_{qj}$

  $\mathcal{PL}_3(\beta) = \frac{\exp(\beta' S_i)}{\sum_{q \in Q_i} \exp(\beta' S_q^*)}$

- Three tests for PH regression model parameters (Wald, Partial LR or Score Test). Score test with one binary covariate is the same as log-rank.
- PH Regression model building

  Possible Criteria: Wald test, LR test, score test, AIC

  AIC$= -2 \log \mathcal{L} + kp$, k is penalty, p is number of parameters

- Estimation of Survivor function

  $W(t_i, \hat{\beta}) = \sum_{j \in R(t_i)} e^{\hat{\beta}' Z_j}$

  $\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{W(t_i, \hat{\beta})}$ (Breslow cum. hazard est.)

  $\hat{S}_0(t) = \exp(-\hat{H}_0(t))$ (Baseline survival function)

  $\hat{S}(t|Z = Z_0) = [\hat{S}_0(t)]^{\exp(\hat{\beta}' Z_0)}$

Chapter 9 - Refinements of Semi-parametric PH models

- Know form of PH model with time-dependent covariates: $h(x|Z(t), t \leq x) = h_0(x) \exp(\beta' Z(x))$ and how to interpret
- Recognize and interpret `R coxph` models using `tt()` or counting process (start, end] intervals
- Approaches to deal with non-proportional hazards 1. piecewise PH model w/ TD vars 2. stratified models
- Know form of stratified PH model: $h_j(x|Z(t), t \leq x) = h_{0j}(x) \exp(\beta' Z(x))$

  LR Test for assumption of common β across j strata:
  $\chi^2_{(s-1)p} = 2[\sum_{j=1}^{s} LL_j(\hat{\beta}_j) - \sum_{j=1}^{s} LL_j(\hat{\beta})]$
  1st term from ind. models for each strata, 2nd term from stratified model
- PH regression with left-truncation - condition hazard on $X > L$, modify risk set $R(t) = \{j : L_j < t \leq T_j\}$. In `R` use `Surv(entry,failtime,status)` syntax

Chapter 11 - Regression Diagnostics

- Overall Fit

  1. Cox-Snell residuals $r_j = \hat{H}_0(T_j) \exp(\hat{\beta}' Z_j)$
  2. Plot $\hat{H}_r(r_j)$ (cum. hazard based on $\{r_j, \delta_j\}$) vs. r_j. line through origin w/ slope 1 if good fit

  in `R`, `cs_res<-delta-resid(fit,type="martingale")`
- Functional form of covariates

  1. Get martingale residuals (diff. between obs and exp deaths in $(0,t_i)$) $\hat{M}_j = \delta_j - \hat{H}_0(T_j) \exp(\hat{\beta}' Z_j)$ (for RC and time ind. var) from model where form of Z_1 is not known
  2. Scatterplot of $\hat{M}_j$ vs. Z_1 for jth obs. & apply smoother
  3. smoothed curve suggests form for $f(Z_1)$

  in `R`, `mg_res<-resid(fit,type="martingale")`
- PH assumption

  Approach 1 - Use time dependent covariate. 1. Multiply fixed covar by function of time $g(t)$ to create TD covar 2. fit PH model with fixed and TD covar; significant TD indicates PH violation

  Approach 2 - Cumulative Hazard plots. Discretize Z_1 into K groups and fit models stratified on Z_1, $\log\{\hat{H}_{g0}(t)\}$ for $g = 1, \ldots, K$
  $\log\{\hat{H}_{g0}(t)\}$ vs. t should be parallel
  $\log\{\hat{H}_{g0}(t)\} - \log\{\hat{H}_{10}(t)\}$ vs. t for $g = 2, \ldots, K$ should be roughly constant
  $\hat{H}_{g0}(t)$ vs. $\hat{H}_{10}(t)$ for $g = 2, \ldots, K$ should be straight lines through origin (Andersen plot)

  Approach 3 - Arjas plot for categorical covar Z_1

  Approach 4 - Score residuals plot define process $U_k(t)$ for each covar. Plot of $U_k(t)$ vs. t should fluctuate around 0 if PH holds. (within $\pm$ 1.358 - prob from Brownian bridge)

  in `R`, `sch_res<-resid(fit,type="schoenfeld")`
  `stdsc_res<-cumsum(sch_res)*sqrt(fit$var)`
- Outliers

  Deviance residuals less skewed than martingale residuals. Plot risk score vs. deviance resid. Large vals of deviance resid are outliers

  in `R`, `dev_res<-resid(fit1,type="deviance")`

- Influential points
  $\hat{\beta} - \hat{\beta}_{(j)}$ vs. j where $\hat{\beta}_{(j)}$ is model w/o j. approximate using score residuals $I(\hat{\beta})^{-1}(S_{j1}, \ldots, S_{jp})'$
  in R, `diff_betas<-resid(fit1,type="dfbetas")`

Chapter 12 - Parametric Regression models

- Accelerated Failure time representation
  $S(x|Z) = S_0[\exp(\theta' Z)x]$, where $\exp(\theta' Z)$ is accel. factor
  $X_{0.5}^{(Z)} = \frac{X_{0.5}^{(0)}}{\exp(\theta' Z)}$
- Linear log time representation
  $Y = \log X = \mu + \gamma' Z + \sigma W$, where W is known dist.
  If $S_0(x)$ is survival function of $\exp(\mu + \sigma W)$ then linear log time model $\Leftrightarrow$ AFT model with $\theta = -\gamma$.
  Weibull: W is standard extreme value distribution. Has linear log time, AFT, and PH representations
  $h(x|Z) = \alpha\lambda x^{\alpha-1}\exp(\beta' Z)$
  Convert between linear log time and hazard parameters:
  $\alpha = 1/\sigma \quad \lambda = \exp(-\mu/\sigma) \quad \beta_j = -\gamma_j/\sigma,\ j = 1, \ldots, p$
  in R, have $\log(\hat{\sigma})$ convert from $Cov(\hat{\mu}, \log(\hat{\sigma}))$ to $Cov(\hat{\mu}, \hat{\sigma})$:
  $Cov(\hat{\mu}, \hat{\sigma}) = Cov(\hat{\mu}, \log(\hat{\sigma}))\hat{\sigma} \quad Var(\hat{\sigma}) = Var(\log\hat{\sigma})\hat{\sigma}^2$
  Log-logistic: W is standard logistic distribution. Has linear log time, AFT, and prop. odds representations
  $S(x|Z) = \frac{1}{1+\lambda e^{\beta' Z} x^{\alpha}}$
  $\frac{S(x|Z)}{1-S(x|Z)} = \exp(-\beta' Z)\frac{S(x|Z=0)}{1-S(x|Z=0)}$
  Same parameter conversion from linear log time as Weibull

Sample Size and Study Design

- Know steps to calculate sample size :
- Crude estimate based on survival at fixed point:
  $N_{arm} = \frac{\left(z_{1-\alpha/2}\sqrt{2\bar{P}(1-\bar{P})} + z_{1-\beta}\sqrt{P_e(1-P_e)+P_c(1-P_c)}\right)^2}{(P_c-P_e)^2}$
  P_c: prob of event in control arm by time t
  P_e: prob of event in "experimental" arm by time t
  $\bar{P} = (P_e + P_c)/2$
- Sample size based on log-rank test:
  HR: $\theta = e^{\beta} = \frac{\lambda_1(t)}{\lambda_0(t)}$
  Number of events, d, needed for power $1-\beta$ with two-sided α level test is $d = \frac{4(z_{1-\alpha/2}+z_{1-\beta})^2}{[\log(\theta)]^2}$
  Estimate θ from desired R-year survival in group 1, $S_1(R)$ and group 0, $S_0(R)$ (under exponential distribution)
  $\frac{\log(S_1(R))}{\log(S_0(R))} = \frac{-\lambda_1 R}{-\lambda_0 R} = \frac{\lambda_1}{\lambda_0} = \theta$

Estimate θ from desired improvement in median survival from M_0 months to M_1 months (under exponential distribution)
$\lambda_i = \frac{-\log(0.5)}{M_i}$, $i = 0, 1$

How many patients? for follow-up time F,
$d = (N/2)(1 - e^{-\lambda_0 F}) + (N/2)(1 - e^{-\lambda_1 F})$

- More realistic accrual (not all entries on same day) for accrual period, A.

  to get P_c and P_e solve $P_i = 1 - \frac{\exp(-\lambda_i F)(1-\exp(-\lambda_i A))}{\lambda_i A}$ or $P_i \approx 1 - \exp[-\lambda_i(A/2 + F)]$ where $i = c, e$

  then $N = \frac{2d}{P_c + P_e}$ $\quad N = \frac{8(z_{1-\alpha/2}+z_{1-\beta})^2}{[\log(\theta)]^2(P_c+P_e)}$

  Vary A and F to find study design that has large enough sample and is feasible given expected accrual

- Freedman approx. (conservative)
  $N = \frac{2(z_{1-\alpha/2}+z_{1-\beta})^2}{P_e+P_c}\left(\frac{\theta+1}{\theta-1}\right)^2$

□