1 HW 6

1. In Cox PH model, the observed information matrix of the partial likelihood is

$$\hat{I}_n(\beta) = -\frac{\partial^2 \log PL(\beta)}{\partial\beta\partial\beta^T}.$$

Show that its khth element is of form

$$\hat{I}_{n,kh}(\beta) = \sum_{i=1}^{D}\left\{\sum_{j\in R(t_i)}\left(Z_{jk} - \overline{Z}_{ik}\right)\left(Z_{jh} - \overline{Z}_{ih}\right)w_{ij}\right\},$$

where $w_{ij} = \exp(Z_j^T\beta)/\sum_{l\in R(t_i)}\exp(Z_l^T\beta)$, and $\overline{Z}_{ik} = \sum_{l\in R(t_i)} Z_{lk}w_{il}$. Here Z_{lk} is the kth element of $\mathbf{Z}_l$.

2. The file HIV-clinical-trial-data.csv on Canvas in the Files > HW folder contains data on the occurrence of AIDs or death in HIV positive patients in a clinical trial comparing HIV treatments. The columns in this data sets contains (from left to right):

   - `id`: Patient identification
   - `time`: Time to AIDs or death (in days)
   - `censor`: Censoring indicator (1 if AIDS or death, 0 if right-censored)
   - `time_d`: Time of death
   - `censor_d`: Censoring indicator (1 if death, o if AIDS or right-censored)
   - `tx`: Treatment (1 if IDV, 0 if control)
   - `txgrp`: Treatment group (can be ignored: there are various sub-treatments that were applied)
   - `strat2`: CD4 stratum at screening (0 if CD4≤ 50, 1 if CD4> 50
   - `sex`: Sex (1 if male, 2 if female)
   - `raceth`: Race/ethnicity (1 if white, 2 if black, 3 if Hispanic, 4 if Asian, 5 if native American, 6 others)
   - `ivdrug`: IV drug use history (1 if never, 2 if currently, 3 if previously)
   - `hemophil`: Hemophiliac (1 if yes, 0 if no)
   - `karnof`: Karnofsky performance scale (100 indicates no evidence of disease, 90 minor symptoms, 80 some symptoms, 70 active work impossible)
   - `cd4`: Baseline CD4 count in cells

- `priorzdv`: months of prior use of ZDV (another drug)
- `age`: age at enrollment

In the following analysis, focus on `time` as our survival time.

(a) For two groups (IDV verus control) separately, plot the Kaplan-Meier estimator of the survival functions.

(b) Consider a Cox proportional hazard model with only treatment as the covariate and plot the estimated survival functions for two groups (IDV verus control) separately. Use the `survival` package's routine to fit the model: `coxph(Surv(time, censor) tx, data = data)`

(c) Compare your plots in (a) and (b). Why are the two plots are different? What can you say about the proportional hazard assumption in the Cox model?

(d) Now consider a Cox proportional hazard model with one term: `strat2`. What is the MPLE of the hazard ratio comparing someone with `strat2` equal to 1 vs. someone with `strat2` equal to 0. Construct a 95% confidence interval for this ratio.

(e) Now consider a Cox PH model with covariates: treatment, CD4 count, sex, IV drug use, Hamophiliac, Karnofsky score, months of prior use of ZDV, age at enrollment. Does any of those covariates have a significant effect on the survival?

(f) Now consider a Cox PH model with only significant covariates. Conduct a likelihood ratio test to decide whether to keep the reduced model or the full model.

(g) Use your selected model, try to interpret the effects of significant covariates.

3. Consider the colon cancer oncology clinical trial data (download from R library `survival`, `data(colon)`). The response variable of interest is `time`, and `status` is censoring indicator (1=Event, 0=Censoring). We are interested in modeling the survival time of these patients using Cox proportional hazard model. The data contains two event types, cancer recurrence and death. We'll focus only on death, so subset your data to include only records with `etype = 2`.

   (a) Fit a Cox PH model with the covariates: `age`, `sex`, `rx`, `nodes`, `obstruct`, `perfor`, `adhere`, `differ`, `extent`, `surg`. You'll need to change `differ` and `extent` to factors because these are categorical variables coded as integers. Plot the martingale residuals against `age`, `nodes`. Are there any patterns that suggest the linear form of these covariates is not appropriate? Try using R's `loess` function to add

univariate nonparametric regression lines to your bivariate residual plots to see if this elucidates any of the relationships (it may not!). Comment on what you find. How would you change your model to incorporate any patterns you find?

(b) Now verify if the proportional hazard assumption is appropriate for these continuous predictors by plotting Schoenfeld residuals against time. You can get the Schoenfeld residuals from a fitted Cox model in `surival` using `cox.zph`. This function returns an object with the element `y` that contains the matrix of scaled Schoenfeld residuals. Please provide appropriate plots and interpretation of the plot.

(c) Plot $\log(-\log(\hat{S}^{\mathrm{KM}}(t)))$ vs. $\log(t)$ stratified by `rx`. You can do this using the command: `plot(survfit(Surv(time, status) rx, data = subset(colon, etype == 2)),fun="cloglog", col = c(1,2,3))` Does the proportional hazards assumption hold for these groups?

(d) Use the `dfbeta` function to generate the approximate change in the coefficients if one observation is dropped and scale it by the appropriate matrix (see the class notes on the influence function) to identify any outliers you found. Which, if any, data points are outliers?

(e) Based on your investigations in parts a), b), c), update the model to include more terms, a stratified model, or time-dependent covariates.

(f) Conduct a likelihood ratio test to decide whether to keep the initial model or the full model.

(g) Use your selected model, interpret the effects of significant covariates.