PH C240B: Assignment 3

Rachael Phillips

11/9/2017

Problem 1

Show the CAR condition, $x \to Pr(O = o|X = x)$ for $x \in \mathcal{C}(o)$ is constant implies $Pr(X = x|O = o) = Pr(X = x|x \in \mathcal{C}(o))$. You may assume all random variables here are discrete for simplicity.

Intuitively, the CAR assumption says that the observation of $O = o$ is not influenced by the specific value of X in $\mathcal{C}(O)$ which was taken, only by the fact that X did take a value in $\mathcal{C}(O)$. Thus, the CAR condition is equivalent to

$$Pr(O = o|X = x) = Pr(O = o|X \in C(O)) \text{ for all } o, x \in \mathcal{C}(o)$$

which can be rewritten as,

$$Pr(O = o|X = x \text{ and } X \in C(O)) = Pr(O = o|X \in C(O)) \text{ for all } x \in \mathcal{C}(o)\ .$$

clearly identifying the conditional independence assumption: given $X \in \mathcal{C}(O)$, the events $X = x$ and $O = o$ are conditionally independent. By symmetry of conditional independence we have,

$$Pr(X = x|O = o \text{ and } X \in \mathcal{C}(O)) = Pr(X = x|X \in \mathcal{C}(O))\ .$$

But, since the former is equal to $Pr(X = x|O = o)$, we have that CAR is equivalent to

$$Pr(X = x|O = o) = Pr(X = x|X \in \mathcal{C}(O)) \text{ for all } x \in \mathcal{C}(o)\ .$$

So we see that this conditional probability of the full data, X, given O is only dependent the distribution of P_X Thus the observation of $O = o$ tells us no more, in the sense of what is now the conditional distribution of X, than the obvious $X \in \mathcal{C}(O)$

Here, we prove the factorization of the likelihood as it will come in handy in Problem 2 and in Problem 3.

CAR involves the conditional probability of O given X so, since O is a function of C and X, CAR places an assumption on the conditional probability of C given X, the censoring mechanism, where $C|X \sim G(.|X)$. X is given so the randomness is only through C which is why the conditional probability only depends on G. Obviously, $Pr(O = o|X = x) = 0$ if $x \notin \mathcal{C}(O)$ and if $\mathcal{C}(O)$ is a coarsening of X. Thus,

$$p(o) = P(O = o) = \int_{x \in \mathcal{C}(o)} P(O = o|X = x) dP_X(x)$$

Since $P(O = o|X = x)$ is constant for $x \in \mathcal{C}(o)$, $P(O = o|X = x)$ is only a function of O and it depends on the censoring mechanism, G, as explained above. It follows that

$$p(o) = P(O = o) = h_G(o) \int_{x \in \mathcal{C}(o)} dP_X(x)$$

for $h_G(o) = P_G(O = o|X)$. We also see that $\int_{x \in \mathcal{C}(o)} dP_X(x)$ is just the probability that X falls in the coarsening, $\mathcal{C}(O)$. So,

$$p(o) = P(O = o) = h_G(o)Q_X(o)$$

for $Q_X(o) = P_X(\mathcal{C}(o))$, the full-data probability measure for the set $\mathcal{C}(O)$, the P_X-factor.

We have shown that we have a factorization of the likelihood under CAR which is that the density of o is the parameter of the full-data (defining P_X for the coarsenings) times the parameter of the censoring mechanism. That is, $p = Q_X h_G$. This factorization will come in handy when solving the log-likelihood we obtain $log(Q_X) + log(h_G)$ so we can maximize the likelihood seperately depending on our parameter of interest.

Problem 2

Let $P_{X,\epsilon}$ be a path through P_X, the distribution of the full data, X, and having score $S_1(X)$. This then defines a path $P_{P_{X,\epsilon},G}$ through the observed data distribution, $P_{P_X G}$. Show that the scores generated by these paths are $E[S_1(X)|O = o]$.

$$\frac{d}{d\epsilon} log dP_{P_{X,\epsilon}G}/dP_{P_X G}|_{\epsilon=0} = \frac{d}{d\epsilon} log dP_{P_{X,\epsilon}G}|_{\epsilon=0}$$

$$= \int \frac{dP(O = o|X = x)S_1(x)dP(x)d\nu(x)}{dP_{P_X G}(o)}$$

$$= \int S_1(x)dP(x|O = o)d\nu(x)$$

$$= E\big[S_1(X)|O = o\big] \ .$$

Problem 3

Let G_ϵ be a path through G, the distribution of the censoring time, C, given X, having score $S_2(C, X)$. This then defines a path $P_{P_X G_\epsilon}$ through the observed data distribution, $P_{P_X G}$. Show that the scores generated by these paths are $E[S_2(C, X)|O = o]$.

$$\frac{d}{d\epsilon} log dP_{P_X G_\epsilon}/dP_{P_X G}|_{\epsilon=0} = \frac{d}{d\epsilon} log dP_{P_X G_\epsilon}|_{\epsilon=0}$$

$$= \int \frac{S_2(c, x)dP(O = o|X = x)dP(x)d\nu(x)}{dP_{P_X G_\epsilon}(o)}$$

$$= \int S_2(c, x)dP(x|O = o)d\nu(x)$$

$$= E\big[S_2(C, X)|O = o\big] \ .$$

Problem 4

This problem involves simulating data under a general Cox model. Let's make the assumption we have a conditional hazard of death at time, t, given by $\lambda(t|X) = \lambda_0(t)exp(f_\beta(X))$ where X is a set of covariates and f_β is a function indexed by β, say finite dimensional. Assume the baseline hazard is $\lambda_0(t) = exp(rt)$ for positive r. Given X, what is the distribution of death times? Prove your answer.

Problem 5

Complete the first problem from LabCox in the lab section of the files on bCourses.

Bonus

Assume a CAR model for full data consisting of survival time, censoring time, the continuous baseline covariates and randomly assigned treatment indicator. We have observed data $min(T, C), \Delta$ along with the covariates and treatment indicator. Someome receives a data set of 1000 independent subjects drawn from this model from an RCT and runs a Cox Proportional hazards regression with treatment as the only covariate, showing a significantly negative coefficient. Can you convince this person he may be wrong via simulation? Explain how you set up your simulation and turn in your code to show the results.