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 $X \in \mathcal{C}(O)$

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

$$\begin{aligned}
\frac{d}{d\epsilon} logdP_{P_X G_\epsilon}/dP_{P_X G}|_{\epsilon=0} &= \frac{d}{d\epsilon} logdP_{P_X G_\epsilon}|_{\epsilon=0} \\
&= \int \frac{S_2(c,x)dP(O=o|X=x)dP(x)d\nu(x)}{dP_{P_X G_\epsilon}(o)} \\
&= \int S_2(c,x)dP(x|O=o)d\nu(x) \\
&= E\big[S_2(C,X)|O=o\big] \ .
\end{aligned}$$

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.

$$\lambda(t|X) = \lambda_0(t)exp(f_\beta(X)) = lim_{dt\to 0}\frac{Pr(t \leq T < t + dt|T \geq t, X)}{dt}$$

The conditional probability in the numerator may be written as the ratio of the joint probability that T is in the interval $[t, t + dt)$, $T \geq t$, and X (which is, of course, the same as the probability that $t|X$ is in the interval), to the probability of the condition $T \geq t|X$. The former may be written as $f(t|X)dt$ for small dt, while the latter is $S(t|X)$ by definition ($S(t|X) = Pr(T \geq t|X) = 1 - F(t|X) = \int_t^\infty f(s)ds$). Dividing by dt and passing to the limit gives the result

$$\lambda(t|X) = \frac{f(t|X)}{S(t|X)}$$

In words, given X, the rate of occurrence of the event at duration t equals the density of events at t given X, divided by the probability of surviving to that duration without experiencing the event given X. Note from the definition of the survival function that $-f(t|X)$ is the derivative of $S(t|X)$. Thus, we can rewrite the above equation as

$$\lambda(t|X) = -\frac{d}{dt}logS(t|X)$$

If we now integrate from 0 to t and introduce the boundary condition $S(0) = 1$ (since the event is sure not to have occurred by duration 0), we can solve the above expression to obtain a formula for the probability of surviving to duration t given X as a function of the hazard at all durations up to t given X:

$$S(t|X) \sim U[0,1] = exp(-\int_0^t \lambda(s|X)ds)$$

Notice that $\int_0^t \lambda(s|X)ds = \Lambda(t|X)$, the cumulative hazard.

Now we need to plug in our given conditional hazard of death at time, t given the covariates X and solve for t in order to find the distribution of death times.

$$S(t|X) \sim U[0,1] = exp(-\int_0^t \lambda_0(s) exp(f_\beta(X))ds)$$

$$S(t|X) \sim U[0,1] = exp(-\int_0^t exp(rs) exp(f_\beta(X))ds)$$

$$log\big(S(t|X) \sim U[0,1]\big) = (-\int_0^t exp(rs) exp(f_\beta(X))ds)$$

$$log\big(S(t|X) \sim U[0,1]\big) = -exp(f_\beta(X)) \int_0^t exp(rs)ds$$

$$log\big(S(t|X) \sim U[0,1]\big) = -exp(f_\beta(X))\Big(\frac{1}{r} exp(rs)|_0^t\Big)$$

$$log\big(S(t|X) \sim U[0,1]\big) = -exp(f_\beta(X))\Big(\frac{1}{r} exp(rt) - \frac{1}{r}\Big)$$

$$-log\big(S(t|X) \sim U[0,1]\big) = exp(f_\beta(X))\frac{1}{r}\Big(exp(rt) - 1\Big)$$

$$\frac{r\Big(-log\big(S(t|X) \sim U[0,1]\big)\Big)}{exp(f_\beta(X))} + 1 = exp(rt)$$

$$\frac{log\left(\frac{r\left(-log\left(S(t|X) \sim U[0,1]\right)\right)}{exp(f_\beta(X))} + 1\right)}{r} = t$$

We will can write $S(t|X) \sim U[0,1]$ as $U[0,1]$ because we know that in order to simulate the survival curve at time t given covariates X, we just need to draw from a $U[0,1]$ distribution. Thus,

$$\frac{log\left(\frac{r\left(-log\left(U[0,1]\right)\right)}{exp(f_\beta(X))} + 1\right)}{r} = t$$

is the distribution of death times given covariates X.

Problem 5

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

Specifically, we are asked to use the model generated in the lab to simulate 1000 draws of $n = 1000$ and check coverage on the coefficient, which represents the log of the hazard proportion between treated and non-treated, in this case. Also, according to the first problem from LabCox, “We also know that the true proportional hazard of treated to untreated is exp(-1), the coefficient of treatment being -1”.

```r
# we repeat the experiment 1000 times
N = 1000

# allocating memory to store the experiment
exp <- data.frame(Truth = numeric(N), Estimate = numeric(N), CI_lower = numeric(N),
                  CI_upper = numeric(N), CI_indicator = numeric(N))

for(i in 1:N){

  # generating the original model
  n = 1000

  # draw the W from standard normals
  W1 = rnorm(n)
  W2 = rnorm(n)
  g0 = function(W1,W2) plogis(.1 + W1*W2)
  A = rbinom(n, 1, g0(W1,W2))

  # draw T and C --both generated with random uniforms
  # as perviously described but C could
  # be generated anyway independent of T given W for
  # identifiability purposes
  C = -log(runif(n))/(.01*exp(.3*W1))
  T =  -log(runif(n))/(.02*exp(2*W1^2 - A))

  Ttilde = pmin(T, C)
  Delta = C >= T & T<= 100

  # Create the survival object
  S = Surv(time = Ttilde, event = Delta, type = "right")
  data = data.frame(A = A, W1 = W1, W2 = W2)
  coxfit = coxph(S ~ ., data = data)

  truth <- exp(-1)
  conf_interval <- as.data.frame(summary(coxfit)[8])
  lower <- conf_interval$conf.int.lower..95[1]
  upper <- conf_interval$conf.int.upper..95[1]
  indicator <- as.numeric(truth >= lower && truth <= upper)

  # populating the data frame where we store our experiment
  exp$Estimate[i] <- conf_interval$conf.int.exp.coef.[1]
  exp$Truth[i] <- exp(-1)
  exp$CI_upper[i] <- upper
  exp$CI_lower[i] <- lower
  exp$CI_indicator[i] <- indicator
}

# calculating the coverage
print(paste('Our coverage is', mean(exp$CI_indicator)))
```

```
## [1] "Our coverage is 0"
```

Our coverage on the coefficient, which represents the log of the hazard proportion between treated and non-treated, is 0, meaning none of our confidence intervals covered the true proportional hazard of treated to untreated. This is because we simulated data according to a model that defies the Cox assumption. That is, we assume the hazard of an individual with covariates A, W_1, W_2 at time, t, is $.02exp(2W_1^2 - A)$, a main terms *non-linear* function of the hazard and the Cox assumption requires the main terms to be a *linear* functional form of the hazard.

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.

```r
# we repeat the experiment 10,000 times
N <- 1000

# allocating memory to store the experiment
exp <- data.frame(Truth = numeric(N), Estimate = numeric(N), CI_lower = numeric(N), CI_upper = numeric(

for(i in 1:N){

  n <- 1000

  # simulate a pseudo-truth
  W1 <- rnorm(n)
  W2 <- rnorm(n)
  W3 <- rnorm(n)
  W4 <- rnorm(n)
  W5 <- rnorm(n)
  A <- rbinom(n, 1, .5)

  # we make C and T dependent on
  # covariates and we also
  # make C dependent on A,
  # such that drop out is effected
  # by treatment
  C <- abs(W1+W2+W3+W4+W5+(20*A))
  T <- abs(W2+W3+W4+W5+W1)
  Ttilde <- pmin(T, C)
  Delta <-  C >= T & T <= 1
  S <- Surv(time = Ttilde, event = Delta, type = "right")

  # because T is not dependent on A,
  # the true coefficient for A is
  # exp(0) = 1
  truth <- 1

  # they run a Cox Proportional Hazards regression
  # with treatment as the only covariate
  incomplete_data <- data.frame(A)
  coxfit <- coxph(S ~ A, data = incomplete_data)
  # did they cover the truth in their misspecified model?
  conf_interval <- as.data.frame(summary(coxfit)[8])
  upper <- conf_interval$conf.int.upper..95[1]
  lower <- conf_interval$conf.int.lower..95[1]
  indicator <- as.numeric(truth >= lower && truth <= upper)

  # populating the data frame where we store our experiment
  exp$Estimate[i] <- conf_interval$conf.int.exp.coef.[1]
  exp$Truth[i] <- 1
  exp$CI_upper[i] <- upper
  exp$CI_lower[i] <- lower
  exp$CI_indicator[i] <- indicator
}

# calculating the coverage
mean(exp$CI_indicator)
```

```
## [1] 0.108
```

We see that when there is dependence on the censoring mechanism from the treatment and there are covariates, the Cox Proportional Hazards regression with treatment as the only covariate poorly covers (only about 10% of the time in our case) the corresponding true coefficient of the treatment variable.

Collaborators & Resources

Tommy Carpenito, Yue You, Jonathan Levy

Survival Model online notes: http://data.princeton.edu/wws509/notes/c7.pdf

Gill, van der Lann, Robins, Coarsening at Random: Characterizations, Conjectures, Counter-Examples (1997)