Week5: Cox Regression I

November 23, 2020

Contents

A. The Cox model as a linear model

Consider independent subjects that follow a Cox model

$$\lambda_i(t, X_i) = \lambda_0(t) \exp(X_i^T \beta)$$

such that the observed right censored survival data (T_i, δ_i, X_i) are i.i.d. Where as always $T_i = \min(\tilde{T}_i, C_i)$ and $\delta_i = I(\tilde{T}_i \leq C_i)$. $\tilde{T}_i \sim \lambda_i(t, X_i)$ and $C_i \sim \lambda_c(t, X_i)$ and with independence given X_i. We further denote the survival distribution of C_i given X_i as $G_c(t, X_i)$ and assume that $G_c(t, x) > \tilde{\epsilon} > 0$ for $t \leq \tau$ and all x. We also assume that all $C_i \leq \tau$ some fixed limited follow-up.

Let the related counting processes be denoted as $N_i(t) = (T_i \leq t, \delta_i = 1)$ and the at-risk processes be $Y_i(t) = I(t \leq T_i)$. Define $N_\bullet(t) = \sum_i N_i(t)$, $Y_\bullet(t) = \sum_i Y_i(t)$. And let $S_j(t, \beta) = \sum_i Y_i(t) \exp(X_i^T \beta) X_i^2$ for $j = 0, 1, 2$, where $X_i^0 = 1$, $X_i^j = X_i$, $X_i^2 = X_i X_i^T$ (for X_i a $p \times 1$ vector). Let $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$.

1. What is the survival function of $\Lambda_0(\tilde{T}_i) \exp(+X_i^T \beta)$ given X_i.

2. Show that $Y_i = \log(\Lambda_0(\tilde{T}_i))$ can be written as as linear model

   $$Y_i = \alpha - X_i^T \beta + W_i$$

   where W_i is extreme value distributed with survival distribution $P(W_i > w) = exp(-exp(w))$. Hint: show that survival distributions are the same.

3. Let $\Delta_i = I(\tilde{T}_i \leq C_i)$, show that

$$E(\frac{\Delta_i}{G_c(T_i, X_i)}) = 1$$

Hint: repeated conditioning.

4. Use the previous result to show that

$$E(\frac{\Delta_i}{G_c(T_i, X_i)}(Y_i - (\alpha - X_i^T\beta + W_i))) = 0$$

How can this be used in practice. Construct an estimating equation for β based on this.

5. We now consider a linear regression model of log-transforms such that, given X, $log(V) = -X^T\gamma + \epsilon$ where ϵ has hazard $\nu(t)$. Derive the hazard of V.

B. This exercise is about understanding the partial likelihood.

Consider independent subjects that follow a Cox model

$$\lambda_i(t, X_i) = \lambda_0(t)\exp(X_i^T\beta)$$

such that the observed right censored survival data (T_i, δ_i, X_i) are i.i.d. With independent censoring given X. Let the related counting processes be denoted as $N_i(t) = (T_i \leq t, \delta_i = 1)$ and the at-risk processes be $Y_i(t) = I(t \leq T_i)$. Define $N_\bullet(t) = \sum_i N_i(t)$, $Y_\bullet(t) = \sum_i Y_i(t)$. And let $S_j(t, \beta) = \sum_i Y_i(t)\exp(X_i^T\beta)X_i^j$ for $j = 0, 1$.

Let $\tau_1, \tau_2, ..., \tau_d$ denote the ordered death times of the sample, that is the ordered jump times of $N_\bullet(t)$, let $\mathcal{R}(\tau_j)$ denote the indeces of those subjects under risk at τ_j, and let D_j denote the index of the subject that died at time τ_j for $j = 1, .., d$.

1. What is the intensity for $N_\bullet(t)$. Using the analogy from the Nelson-Aalen estimator suggest an estimator for $\Lambda_0(t)$ based on a moment equation for $N_\bullet(t)$. This is for known β. Indicate with martingale arguments why this is a good estimator.

2. What is the likelihood for the data using the hazard functions.

3. Compute $\pi_j(i) = P(D_j = i|\mathcal{R}(\tau_j), \tau_j = t)$, that is the probability that subject "i" dies given that we have an event at time "t" and given who are under risk and their covariates. Hint: write out the probability and see that we get back to the intensities conditioning on those that are under risk. Note knowing $\mathcal{R}(\tau_j)$ tells us who satisfies $T_i \geq \tau_j$.

4. What is the "partial likelihood" the probability of the seeing the observed D_j for $j = 1, .., d$, and write up a likelihood the observed data forward in time, using the D_j's and the $N_\bullet$.

5. See that 2 and 4 are the same, by re-arranging 2.

6. What is the expected covariate (the mean of X's) for the subject dying given we have a death at time τ_j and $\mathcal{R}(\tau_j)$, as well as their covariates. Use the probability distribution from 3 (everything else if fixed when we condition on covariates and risk set).

7. What is the variance of the X's under risk at the j th death time, that is the X's from $\mathcal{R}(\tau_j)$.

C. The Cox model

Do exercise 6.3 of MS. Hint: for b) write up the related estimating equation $U(\theta)$ and derive the asymptotic distribution of $U(\theta_0)$ using Martingale theory, adding and subtracting the compensator. The variance is then the limit of second derivative squared times the variance of the martingale. In addition

1. Write up the score test, that is the test based on evaluating the score function for $H_0 : \theta = 1$.

2. In the case where $N_{1i}(t) \sim \alpha_i(t)$ and $N_{2i}(t) \sim \theta\alpha_i(t)$. Estimate the θ by calculating and multiplying partial likelihoods for each i, so the probability of seeing who died at the the first jump time in all pairs. Show that this is also a mean zero estimating equation, by rewriting the score using martingales. Hint: use Martingale magic.

D. The Cox model in action

Considering the TRACE data with time to death as the outcome. We wish to understand how vf and chf (two covariates) are important for survival.

1. Choose a time-scale, here there are at-least two possibilities, age and follow-up time.

2. Do a Kaplan-Meier for the 4 groups based on vf and chf, and a log-rank test. What do you conclude ? Estimate also the cumulatives hazards using the Nelson-Aalen estimates and compare the survival estimates based on these with the Kaplan-Meier's.

3. Do a Cox regression for vf and chf and make conclusions under the model. How do we interpret the regression coefficients.

   - Estimate the survival for the 4 groups based on vf and chf and compare with the Kaplan-Meier estimates.
   - Should we have included interactions between vf and chf in the model. Do a formal test.
   - Estimate the effect of chf in a stratified Cox model (stratified after vf), look at the baselines and try to conclude wether they are proportional.

4. We shal now consider a Cox model with time-dependent covariates. It is expected that vf has a predictive strong effect only within the first 2 months (say), and that chf has an effect that is strong the first 6 monthts, and then different. Set up the data with)stop,start) such that you can fit a Cox model with time-dependent covariates of this type. Specifically we wish to consider the regression model

   $$\lambda_i(t, X_i) = \lambda_0(t) \exp(VF_i\beta_1 + VF_i I(t > 2)\beta_2 + CHF_i\beta_3 + CHF_i I(t > 6)\beta_4)$$

   What tests could be of interest here, and how do we interpret the coefficients.

   - Hint: in R you can use survSplit to get the data on the needed form. Part of this question is to explain how putting the data on this form and fitting a Cox model would be achieving what we are after.

- Fit the model and compare the survival predictions with the Kaplan-Meier estimates.