Time-dependent covariates

Patrick Breheny

December 5

Introduction

- Our previous lecture introduced the idea of time-varying coefficients, $\beta(t)$
- A related idea is that the covariates themselves can change across time: $X(t)$
- For example, we saw in the previous lecture that the effect of Karnofsky status in the VA lung data waned over time
- Presumably, this is because we used Karnofsky status at baseline, which becomes more and more unrelated to Karnofsky status at time t as t increases

Introduction (cont'd)

- Suppose instead that we decided to re-measure Karnofsky status every three months
- In this scenario, perhaps the proportional hazards assumption would remain reasonably satisfied, with Karnofsky status retaining its predictive relevance over time
- Note that, unlike time-dependent coefficients, this is something that would have to be planned from the outset of the study as it requires additional data collection and not simply changing the modeling assumptions

The "proportional hazards" model

- As an aside, note that both of the ideas we're covering this week, time-varying coefficients $\beta(t)$ and time-dependent covariates $X(t)$, extend the Cox model in a way that relaxes the proportional hazards assumption
- In other words, while the basic Cox model assumes that for two subjects i and j, $\lambda_i(t)/\lambda_j(t)$ remains constant over time, this is not necessarily true if time-dependent covariates or coefficients are present in the model
- For this reason, as you may have noticed, Kalbfleisch and Prentice refuse to call Cox regression a "proportional hazards" model (this is not a sentiment shared by most other authors; Cox models are widely referred to as proportional hazards models in the literature)

Predicting the future

- The golden rule to keep in mind as we begin to deal with time-dependent covariates (i.e., with including into our analysis data that was not available at the outset) is that, while the model can change in any way based on past data, *it can never use data that comes from the future*
- For example, in our hypothetical variation of the VA lung study in which Karnofsky status was measured every 3 months, it's fine to use this information to model the three-month Karnofsky score to model a subject's risk of death at 4 months, but definitely not OK to use that information to model risk at 2 months

The IMPROVE-IT study

- Using future information can introduce bias into a model in a variety of ways
- As an example, in a 2015 randomized controlled trial of the drug Vytorin (used to treat high cholesterol), the researchers wished to determine whether the achievement of a cholesterol reduction target (LDL < 70 mg/dL) was associated with improved survival
- Whether a patient achieves the target is not known at the outset, so we need to be careful here that we don't unintentionally use future data in modeling the outcome (here, the time until death, heart attack, or stroke)

Naïve analysis

- A naïve analysis would be to simply compare the two groups (those who achieved the target vs. those who did not) using, say, a log-rank test
- This analysis, however, is seriously biased by the fact that the longer a person lives, the greater chance they have to achieve the target
- In particular, individuals who die early will be preferentially assigned to the "failed to achieve target" group, making that group appear to be at higher risk

Immortal time bias

- To put it another way, individuals in the "achieved target" group are essentially immortal until they achieve that target (this is sometimes referred to as "immortal time bias")
- As a consequence, even if achieving the target and death are completely independent, achieving the target will appear to be beneficial in this analysis
- Again, the bias arises from trying to use future data (target achievement) to predict survival (at times prior to the achievement of the target)

Landmarking

- One way around this is to select a fixed time – say, 1 month after randomization – by which patients much achieve the target in order to be in the "achieved target" group
- Then, we compare survival between the two groups from that point on, ignoring all deaths that occurred prior to 1 month
- This approach, which was in fact the approach used in the Vytorin study, is known as a *landmark analysis*, and is a valid way of dealing with time-dependent covariates

Time-dependent covariates

- The landmark method has several inherent limitations:
  - Need to choose an arbitrary landmark time
  - Need to throw away all data from before landmark time
  - Lose the ability to distinguish between subjects who achieve target after 1 month and people who never achieve target
- Cox modeling with time-dependent covariates is a much more flexible alternative

Mathematical details

- Like time-dependent coefficients, the general mathematical details are straightforward, although the bookkeeping from a software perspective is somewhat complicated
- When a new value of a covariate X becomes available at time t^*, that value is used to calculate $\eta(t)$ for all $t \geq t^*$ (or until an even newer value becomes available
- Note that, as with time-varying coefficients, this means that the linear predictor changes as a function of time:

$$\eta_i(t) = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ij}(t)\beta_j$$

where $x_{ij}(t)$ denotes the most recent measured value of the coefficient

Equivalent formulation

- An equivalent way of expressing the Cox partial likelihood in the case of time-dependent covariates is that we observe subject i from time $t = 0$ until time $t = t^*$, at which point the subject is censored
- Then we observe a new subject, starting at time t^* (i.e., left truncated at time t^*), mostly the same as subject i, except that $X_j(t)$ is now replaced with the new value
- This may seem like we are artificially increasing the sample size, but it leads to the exact same partial likelihood, so there is no need for any adjustment to the analysis

Pertussis vaccine

- To see an example of time-dependent covariates in action, we will consider a study of the link between preterm birth and receiving the Tdap vaccine during pregnancy, with data coming from California health records
- In response to several outbreaks of pertussis among newborns (in whom it is very serious and occasionally fatal), the CDC now recommends that pregnant women receive the Tdap vaccine during pregnancy
- The purpose of this study was to investigate whether this new recommendation has any unintended side effects
- I was unable to get the actual data from this study, so we will be analyzing simulated data intended to resemble the findings of the original data

Data

- The data set contains three variables:
  - The time of delivery (in weeks since conception)
  - Whether the woman received the Tdap vaccine
  - The time that the woman received if vaccine (if applicable)
- Deliveries prior to 37 weeks are considered preterm; therefore, we will stop follow-up at that point, treating all full-term births as censored at 37 weeks.

Naïve analysis

- A naïve analysis is subject to the same sorts of biases that we encountered earlier, as it attempts to use future data (vaccination) to make predictions of preterm birth that occur prior to vaccination
- In this case, the immortal time bias causes vaccination to appear protective:

```
> coxph(Surv(Time, Status) ~ Vac)
          coef exp(coef) se(coef)     z     p
Vac     -0.445     0.641    0.114 -3.88 1e-04
n= 1000, number of events= 364
```

Setting up time-dependent covariates

- The `survival` package supports time-dependent covariates
- To use them, you have to set up the data according to the subject duplication with start/stop times scheme that we referred to earlier
- This is something of a pain, so the `survival` package supplies a function called `tmerge` to assist with the process

tmerge

- tmerge supports more complicated operations than what we need here, such as merging two data sets with an ID key to match on, so the code looks a bit strange:

```
tmerge(Data, Data, id=ID, Vac=tdc(tVac),
       PTB=event(Time, Status))
```

- The result is that rows like this in the original:

```
  Time Status tVac
1 36.1      1 35.3
```

become rows like this:

```
  tstart  tstop PTB Vac
1    0.0   35.3   0   0
2   35.3   36.1   1   1
```

Results

All the work for fitting models with time-dependent covariates goes into setting up the data; once this is done, the fitting and inference are just regular Cox regression, with a slight modification to the `Surv` object:

```
> coxph(Surv(tstart, tstop, PTB) ~ Vac, tdData)
      coef exp(coef) se(coef)    z    p
Vac 0.121     1.129    0.115 1.05 0.29
n= 1326, number of events= 364
```

Comments

- In the simulation, vaccination and preterm delivery were independent, so this is the correct result
- Also, it agrees with what the researchers found in the real study (which also used Cox regression with time-dependent covariates), where they estimate a hazard ratio of 1.03 (95% CI: 0.97-1.09) in a sample size of 123,494
- Remark: Note that n is artificially increased when using Cox regression with time-dependent covariates in the `survival` package; this can lead to incorrect calculations for some quantities, such as R^2

Predictable time-dependent covariates

- A common question concerning time-dependent covariates is: "What should I do with covariates like age or years since diagnosis? Aren't these time-varying?"
- Yes and no; they are varying with time, of course, but in a predictable manner, and perhaps more importantly, varying for all subjects in the same way
- Thus, if age is in the model as a linear term, it makes no difference whether it is treated as time-dependent or not:

$$\frac{\exp\{(x_j + t)\beta\}}{\sum_{i \in R(t)} \exp\{(x_i + t)\beta\}} = \frac{\exp\{x_j\beta\}}{\sum_{i \in R(t)} \exp\{x_i\beta\}}$$

- If the effect of age is nonlinear, this cancellation no longer occurs; you may wish to use `\tt()` in such cases

Interval vs. external covariates

- Finally, another consideration worth being aware of is the difference between internal and external covariates:
  - An example of an external covariate would be air pollution as a predictor of asthma
  - Karnofsky score, on the other hand, would be an example of an internal covariate

  additional details, including a formal definition, are given in the book
- The distinction matters because even if we had access to Karnofsky status, adjusting for it might distort the meaning of other coefficients (in general, one should be careful about adjusting for variables in the causal pathway)