Homework: Probabilistic Bias Analysis

Due 2 May 2024 at 11:59pm on Gradescope

Read all questions carefully before answering. You may work in small groups of no more than 3 individuals and turn in a single assignment (and everyone in the group will receive the same grade). Work through the entire assignment individually first, then come together to discuss and collaborate. Please maintain numbering on sub-questions, type your responses, and **please keep answers brief.**

Load required packages and read data:

```
library(survival)
load("senssamp.rdata")
colnames(senssamp) <- tolower(colnames(senssamp))
```

We will replicate and modify part of the bias analysis described in the 2003 *Epidemiology* paper by Lash and Fink.[1] The data were obtained from the companion website to the text.

The variables in the dataset are:

```
#    Variable    Type    Len    Pos    Format     Label
-------------------------------------------------------------------------------
4    adjchem     Num       8     24    T2X.       Adjuvant chemotherapy
13   agecat1     Num       8     96    T2X.       65 to 74
14   agecat2     Num       8    104    T2X.       75 to 90
10   allcomor    Num       8     72               Sum of comorbidities
7    axil        Num       8     48    T2X.       Axillary Dissection
9    bccause     Num       8     64    T2X.       Death from breast cancer
2    bcssurg     Num       8      8    T2X.       Breast conserving surgery
6    defnther    Num       8     40    DEFINF.    Definitive therapy
12   excat1      Num       8     88    T2X.       Regional
5    grphos      Num       8    112    T2X.       Hospital w/o tumor registry
3    mastsurg    Num       8     16    T2X.       Mastectomy
11   mortime2    Num       8     80               Time to death (years)
1    reid        Num       8      0    T2X.       Reidentified
5    stopped     Num       8     32    T2X.       Chemo terminated, non-death
8    ttorad      Num       8     56               Time to radiation therapy
```

(**NOTE** that the `defnther` variable for the exposure denotes receiving less than definitive therapy (=1) vs the reference of definitive therapy (=0).)

[1] Lash, Timothy L., and Aliza K. Fink. "Semi-automated sensitivity analysis to assess systematic errors in observational data." Epidemiology (2003): 451-458.

Standard analysis

Estimate the relationship between receiving less than definitive therapy (`defnther=1`) on breast cancer-related mortality in this cohort. We will adjust models for age (categorical, `agecat1` and `agecat2`), and stage (regional vs. local, `excat1`).

Fit a Cox proportional hazards model for the association of interest:

```
obs.model <- coxph(Surv(mortime2, bccause) ~ defnther + excat1 + agecat1 + agecat2,
                   ties="efron", data=senssamp)
exp(coef(obs.model))[1]
exp(confint.default(obs.model))[1,]
```

Bias analysis for unmeasured confounding

There is a concern that this analysis is subject to *confounding by indication*, a source of systematic error where the treatment (or exposure) is assigned based on a clinical indication that is related to the outcome. In this case, Lash and Fink state that women who received less-than-definitive therapy may have had clinical reasons for receiving that therapy, which may have been related to greater risk of death.

We will perform a probabilistic bias analysis to determine the potential influence of an unmeasured confounder that may be indicative of greater propensity to be treated with less than definitive therapy, and greater risk of death.

1. Prepare to sample over the distribution of bias parameters 5000 times:

```
N.obs <- nrow(senssamp) # Determine number of observations in data
N.samp <- 5000 # Specify the number of samples
```

2. To define the relevant biasing relationships, we need to specify the prevalence of the unmeasured confounder U based on exposure (X, definitive therapy), and outcome (death, Y). The parameters used by Lash and Fink are:

- p_{00}: Prevalence of U given definitive therapy (`defnther=0`) and no death (`bccause=0`).
- p_{10}: Prevalence of U given less than definitive therapy (`defnther=1`) and no death (`bccause=0`).
- p_{01}: Prevalence of U given definitive therapy (`defnther=0`) and death (`bccause=1`).
- p_{11}: Prevalence of U given less than definitive therapy (`defnther=1`) and death (`bccause=1`).

The authors specify these with **Uniform** distributions, with p_{00} ranging from 30%–40%, p_{10} ranging from 45%–55%, p_{01} ranging from 45%–55%, and p_{11} ranging from 60%–70%.

On your own: Complete the following code to create variables for the above parameters (named `p.U00`, `p.U10`, `p.U01`, and `p.U11`, respectively) that contain `N.samp` random samples from Uniform distributions with the ranges described above. Before the first random sample, set the random number seed to the value 123.

```
set.seed(123) # Set the random seed.
# Prior distributions for unmeasured confounder given therapy and death:
p.U00 <- # Definitive therapy, did not die
p.U10 <- # Less than definitive therapy, did not die
```

```
p.U01 <- # Definitive therapy, died
p.U11 <- # Less than definitive therapy, died
```

3. For each sampled value of the bias parameters you constructed above, a) sample values of the unmeasured confounder U for each individual in the dataset, given their value of treatment and outcome, b) fit a Cox model that adjusts for this variable, and c) accumulate the parameter estimate for the effect of interest, along with the coefficient on U. Note that we calculate the effect estimates corrected for systematic error only, and resample from the sampling distribution to yield an estimate that incorporates sampling error.

On your own: complete the following loop, by finishing the pieces indicated to:

- Fit a logistic regression model of the samples of U on `defnther`, and accumulate the model coefficients.
- Fit a Cox model, adjusting for the simulated values of the unmeasured covariate U (**you do not need to use an offset term for this case**).

```
# Initialize storage vectors for the parameter estimates
HR.systematic <- # for HR corrected for systematic error only
  HR.total <-    # for HR corrected for systematic + accounting for sampling error
  HR.U <- matrix(NA, ncol=1, nrow=N.samp) # For the Y-X (defnther) and Y-U relationships
a.X <- matrix(NA, ncol=2, nrow=N.samp)   # For the U-X (defnther) relationship.

attach(senssamp)
for (i in 1:N.samp){

  # Calculate probability of U for current value of bias parameters
  # given therapy (defnther) and mortality (bccause):
  p.U <-  p.U00[i]*(1-defnther)*(1-bccause) +   # Definitive therapy, no death
          p.U10[i]*defnther*(1-bccause) +       # Less than definitive therapy, no death
          p.U01[i]*(1-defnther)*bccause +       # Definitive therapy, death
          p.U11[i]*defnther*bccause             # Less thanefinitive therapy, death

  U <- rbinom(N.obs, 1, p.U) # Sample the unmeasured confounder

  # ***** COMPLETE THIS *****
  U.model <- ##### Fit a logistic regression model here,
             ##### for the outcome U and a single precitor defnther

  bias.model <- ##### Fit a Cox proportional hazards model here, parameterized as above
                ##### but including the unmeasured confounder just sampled.
  # ************************

  # Accumulate coefficients
  a.X[i,] <- coef(U.model) # Save coefficients from U model

  b.systematic <- coef(bias.model)[1] # beta coefficient on therapy variable
  se.systematic <- sqrt(vcov(bias.model)[1,1]) # standard error of therapy estimate
```

```
  HR.U[i] <- exp(coef(bias.model)[5]) # HR for U

  HR.systematic[i] <- exp(b.systematic) # HR accounting for systematic error only

  b.total <- rnorm(1,b.systematic, se.systematic) # Adding in random variability
  HR.total[i] <- exp(b.total) # HR accounting for systematic + random error
}
detach(senssamp)
```

4. Summarize the coefficients for the corrected Cox model, and the model for U as a function of `defnther` with quantiles (50^{th}, 2.5^{th}, and 97.5^{th} percentiles).

```
round(quantile(HR.systematic, c(.5, .025, .975)), 2)
round(quantile(HR.total, c(.5, .025, .975)), 2)

round(quantile(HR.U, c(.5, .025, .975)), 2) ## Changed to HR.U (from b.U) 4/18/21

colnames(a.X) <- c("Intercept","defnther")
t(apply(a.X, 2, quantile, c(.5, .025, .975)))
```

5. A good practice is to examine the influence of the prior distributions of the bias parameters on the results of the sensitivity analysis. One approach could be to change the range of parameters, but here we will explore the change in the *shape* of the distribution of the bias parameters. Instead of assuming the bias parameters have a Uniform distribution, we will assume a Triangular distribution.

On your own: redo the above exercise using a triangle distribution with the same ranges on each of the bias parameters (you should only need to change the `runif` command to `rtriangle` after loading the `triangle` package).

Questions

1. Complete the following table with the hazard ratio and corresponding 95% quantile-based interval estimates in each cell (place the results for the standard analysis in the right most column of row 1) (**10 points**):

Table 1: Hazard ratios and 95% intervals for standard analysis and bias analysis of unmeasured confounding for relationship between less-than-definitive therapy and breast cancer mortality.

Analysis	Systematic Error Only	Systematic + Random Error
Standard analysis	N/A	
Bias analysis-Uniform		
Bias analysis-Triangle		

2. **From the model that assumed uniform distributions on the bias parameters:**
   a. Report the median and 95% quantile interval of: 1) the parameter estimates (intercept and coefficient on `defnther`) from the logistic model for U and 2) the HR for the mortality-U relationship. (**5 points**)
   b. Based on these, does the specification of the distribution of bias parameters seem to be consistent with a *confounding by indication* scenario? Why or why not? (**10 points**)
3. Between the analyses with the Uniform distribution and the Triangular distribution on the bias parameters, which one has the more narrow bias-corrected interval estimates for the HR between `defnther` and mortality (it won't be much, but should be evident)? Why do you think that is the case? (*Hint: consider the shape of the priors on the bias parameter distributions.*) (**10 points**)
4. **In no more than 4 sentences** provide an interpretation of the results of your bias analysis that you might include in a manuscript. Make references to 1) the original (uncorrected) estimate, 2) the treatment of the bias parameters, 3) how the bias corrected estimates compare to the original estimates, including the direction of the potential bias (not necessarily corresponding to individual sentences). You may focus on the results from the analysis assuming Uniform distributions on the bias parameters. (*Hint: see slide set for introduction to bias analysis.*) **(15 points)**