PHS 2000B Problem Set 2

Due Thursday, February 13th, 2025 by 11:59pm

This problem set is intended to reinforce many of the exciting results that Issa showed us these past two weeks and to help you develop better intuition about the estimation of causal effects for time-fixed and time-varying treatments.

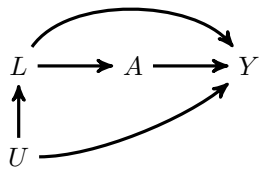
```
knitr::opts_chunk$set(
    echo = TRUE,
    class.output="shadebox",
    warning=FALSE,
    error=FALSE,
    message=FALSE)

# Install Relevant Packages & Datasets
if (!require("pacman")) install.packages("pacman")
pacman::p_load(MedDataSets, arsenal, tidyverse, ggplot2, ggpubr, ggsci, geepack)
```

1 Time-fixed treatment (14 points)

We are interested in the effect of smoking during pregnancy on the risk of infant low birth weight. Data is from the `birthwt_df` dataset in the `meddatasets` package. We will be using IPTW to estimate the risk of infant low birth weight had all mothers in the source population smoked during pregnancy compared to if all mothers in the source population had not smoked during pregnancy.

- `smoke` (A): mother smoked during pregnancy (binary)
- `age`(L): mother's age (continuous)
- `race` (L): mother's race (categorical)
- `first.tri.visit` (L): mother had physician visit in first trimester (binary)
- `low` (Y): infant's birth weight is low (binary, low = 1 if weight <2500 grams)

L is a vector of age, race, and physician visit in first trimester to simplify the DAG.




```
### Load the dataset and perform minimal data processing

df <- birthwt_df
#Generate id variable
df$id <- 1:nrow(df)
```

```
# Move id to first column
df <- df[,c(ncol(df),1:(ncol(df)-1))]
# Create indicator for whether mother had a physician visit during the first trimester
df$first.tri.visit <- ifelse(df$ftv > 0, 1, 0)
```

Question 1

Estimate the denominator of nonstabilized inverse probability of treatment weights, i.e., estimate the probability of receiving the treatment one actually received conditional on past covariates (hereinafter the "estimated treatment probability") for each observation in the observational dataset using a logistic regression model. Assume that exchangeability of observed treatment and counterfactual outcomes holds conditional on `age` (linear and quadratic terms), `race`, and `first.tri.visit`. Do not use product terms. Print the estimated treatment probabilities for the first five observations. (1 point)

Code:

```
# denom <- glm(smoke == 1 ~ age + I(age^2) + as.factor(race) + first.tri.visit,
#   family = binomial(), data = df)
```

Question 2

(a) Plot two histograms of the estimated treatment probabilities, one for the treated observations and one for the untreated observations. Either create a mirrored histogram or make sure the x-axes of the two histograms are the same so you can compare the distributions of the estimated treatment probabilities. (1 point)

Code:

```
# ggplot(subset(df, smoke == 0), aes(x = p.smoke, fill = factor(smoke))) +
#   geom_histogram(aes(y = -after_stat(density)), binwidth = 0.01) +
#   geom_histogram(data = subset(df, smoke == 1),
#       aes(x = p.smoke, y = after_stat(density), fill = factor(smoke)), binwidth = 0.01) +
#   ylab("Density") + xlab("Pr[A = 1 | L = l]") +
#   theme_pubr() + scale_fill_nejm(name = "Smoked during pregnancy") +
#   geom_hline(yintercept = 0) +
#   scale_x_continuous(n.breaks = 6) + scale_y_continuous(n.breaks = 10) +
#   coord_cartesian(xlim = c(0,1))
```

(b) Compare the two plots from part (2)(a). How well do the estimated treatment probability seem to predict whether the observation is treated or untreated? What does that tell you about the data? Evaluate the region of common support. (2 points)

Question 3

(a) Compute inverse probability of treatment weights for every observation in the observational dataset, using the estimated treatment probabilities from question (2). Print the first five rows of the dataset for treated observations and untreated observations. (2 point)

(b) Calculate summary statistics for your weights (mean, min, max, and quartiles). Comment on what you observe. Is the mean what you expect it to be? (1 points)

(c) Estimate the association between `smoke` and `first.tri.visit` in the original dataset and in the pseudo-population (i.e., weighted dataset). Is this what you expect? Why or why not? (2 point)

Code

```
# in the original dataset
# prop.table(xtabs( ~ df$smoke + df$first.tri.visit),2)
#
```

```
# # in the pseudo-population
# prop.table(xtabs(df$weight ~ df$smoke + df$first.tri.visit),2)
```

(d) Estimate the average treatment effect by fitting a weighted logistic regression model. Weight the model by the inverse probability of treatment weights. Report the ATE on the **risk difference** and **risk ratio** scale and interpret this effect estimate. Report the standard error on the log odds scale (i.e., from `summary(weighted.glm)`))

Technical point: Because the pseudo-population is larger than the original sample (in this study, about twice as large), the standard errors from the usual linear regression model will result in invalid 95% confidence intervals that are too narrow. To obtain valid, though conservative, 95% confidence intervals, we can use the robust variance estimator to estimate standard errors, as is used for generalized estimating equation (GEE) models with an independent working correlation.

(3 points)

Code

```
# weighted.glm <- geeglm(low == 1 ~ smoke , data = df,
#                weights = weight, family = binomial(), id = id, corstr = "ind")
```

Question 5

Consider the approach below to estimate the ATE using the g-formula.The following models are fitted.

$$E[Y \mid A = 0, \boldsymbol{L} = \boldsymbol{l}] = \beta_0 + \boldsymbol{\beta_1 L}$$
$$E[Y \mid A = 1, \boldsymbol{L} = \boldsymbol{l}] = \psi_0 + \boldsymbol{\psi_1 L}$$

where $\boldsymbol{L}$ is a vector of three covariates and $\boldsymbol{\beta_1}$ and $\boldsymbol{\psi_1}$ are vectors of their coefficients

$\widehat{ATE} = \widehat{E}[Y^{a=1}] - \widehat{E}[Y^{a=0}]$ is then estimated in the code below.

```
# # Fit outcome models for each level of treatment
# model_treated <- glm(low ~ age + I(age^2) + as.factor(race) + first.tri.visit,
#                      data = df, family = binomial(), subset = smoke == 1)
#
# model_untreated <- glm(low ~ age + I(age^2) + as.factor(race) + first.tri.visit,
#                      data = df, family = binomial(), subset = smoke == 0)
#
# # Compute E[Y | A = a, L = l]
# df$p.YA1 <- predict(model_treated, newdata = df, type = "response")
# df$p.YA0 <- predict(model_untreated, newdata = df, type = "response")
#
# # Compute E[E[Y | A = a, L = l]]
# mean(df$p.YA1)
# mean(df$p.YA0)
#
# # Compute the difference between E[E[Y | A = 1, L = l]] and
# #E[E[Y | A = 0, L = l]] as an estimate of the ATE
# mean(df$p.YA1)-mean(df$p.YA0)
```

Knowing the equivalence of the g-formula and inverse probability weighting, why might the results of the g-formula and the IPW be different in this real data application? (2 points)

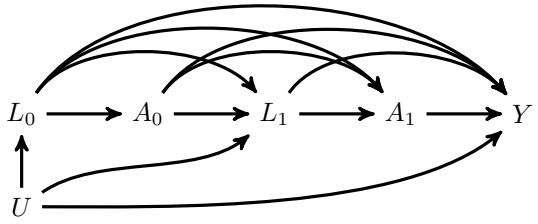
2 Time-Varying Treatment (23 points)

We are interested in the effect of a two-dose antibiotic regimen on the 5-year risk of colorectal cancer among people age 18-30 years with strep throat. Guidelines recommend the antibiotic be taken the day the infection begins and the day after. However, individuals with mild symptoms may decide to not take the antibiotic on one or both days. Antibiotics can cause a relief in strep throat symptoms.

Data for the following variables has been collected in an observational study. Assume there is no loss to follow-up, no competing events, no measurement error, and the DAG below is correct. In general, direct focus to the methodology and not to the substantive area.

Let

- A_0 indicate whether a person took the antibiotic on day 0,
- A_1 indicate whether a person took the antibiotic on day 1,
- L_0 indicate whether a person had mild symptoms on day 0,
- L_1 indicate whether a person had mild symptoms on day 1,
- U represent an unmeasured variable
- Y indicate diagnosis with colorectal cancer



Question 1

Consider the following research questions. For each, complete the following tasks: 1) write the causal effect of interest (i.e., using counterfactuals), 2) state if conventional outcome regression models would be sufficient to answer the question and fully explain your reasoning, 3) if a conventional outcome regression model would suffice to answer the question, write the model that you would fit; you do not need to show any other steps (e.g., transforming odds to probability, standardizing to obtain marginal estimates); assume no product terms are needed in all models.

(a) What is the average treatment effect of taking antibiotic on day 1 (A_1) on the 5-year risk of colorectal cancer (Y)? (3 points)

(b) What is effect of taking antibiotics on day 0 (A_0) and day 1 (A_1) versus not taking antibiotics on either day on the 5-year risk of colorectal cancer (Y)? (3 points)

(c) What is the effect of taking antibiotics on day 0 (A_0) on the 5-year risk of colorectal cancer (Y)? (2 points)

(d) What is effect of taking antibiotics on day 0 (A_0) but not on day 1 (A_1) versus not taking antibiotics on either day on the 5-year risk of colorectal cancer (Y)? (3 points)

Question 2

(a) Write out, in terms of counterfactual outcome Y^{a_0,a_1} and the variables in the DAG, the exchangeability (independence) assumptions that are required to identify the mean counterfactual outcomes under treatments A_0 and A_1. (2 points)

(b) Interpret these assumptions in words (you may interpret each one separately or explain what they mean as a whole). (2 points)

Question 3

In (a)-(d) write all the models needed to estimate the ATE ($E[Y^{a_0=1,a_1=1} - E[Y^{a_0=0,a_1=0}]$) using IPTW and specify the marginal structural model that the IPTW model is estimating.

(a) Model to estimate the probability of taking antibiotic on day 0 (2 points)

(b) Model to estimate the probability of taking antibiotic on day 1 (2 points)

(c) Model for the outcome weighted by observations' IPTW. We will assume no product terms are needed. (2 points)

(d) Write out the marginal structural model (i.e., use counterfactuals). We will assume no product terms are needed. (2 points)