Statistical Inferences from Nonprobability or Low Response-rate Probability Surveys: A Discussion

Phillip S. Kott
pkott@rti.org

$$\sum_{S_1} d_k \alpha(\mathbf{x}_k{}^T \mathbf{g})\mathbf{z}_k = \sum_U \mathbf{z}_k,$$

where

$$p_k = [\alpha(\mathbf{x}_k{}^T\mathbf{g})]^{-1} = \frac{1 + \exp(\mathbf{x}_k{}^T\mathbf{g})\,/\mathrm{U}}{L + \exp(\mathbf{x}_k{}^T\mathbf{g})}$$

Introduction

I received the slides for all four talks in time,

but they (and the papers supporting them) are *hard*;

so I will mostly talk about my own work.

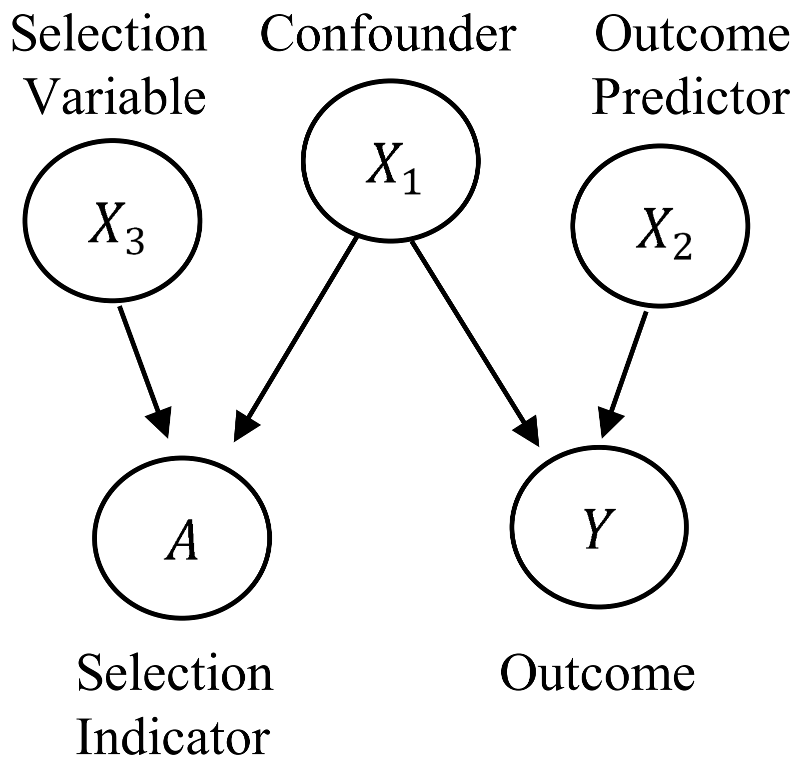
I will mention the slides/papers in passing.

Each has something interesting to say.

$\hat{t}$

Riff on a DAG graph

I hate DAG graphs, but this one is helpful for my discussion:



Riff on IPW

IPW stands for *inverse probability weights*.

This is what biostatisticians and economists call design weights.

For nonresponse and/or nonprobability samples, the selection probabilities need to be modeled:

Perhaps using kernel regression

Perhaps using splines

Perhaps using black-box ML methods

But I like fitting a bounded logistic regression with a calibration equation

What that means

Find a **g** so that the following equation holds

$$\sum\nolimits_{S_1} d_k \alpha(\mathbf{x}_k{}^T \mathbf{g})\mathbf{z}_k = \hat{\mathbf{T}}_{\mathbf{z}} \,,$$

may include probability estimates and population aggregates

where

$$p_k = [\alpha(\mathbf{x}_k{}^T \mathbf{g})]^{-1} = \frac{1 + \exp(\mathbf{x}_k{}^T \mathbf{g}) \,/\mathrm{U}}{L + \exp(\mathbf{x}_k{}^T \mathbf{g})}$$

How does West/Andridge fit in?

$$\sum_{S_1} d_k \alpha(\mathbf{x}_k{}^T \mathbf{g})\mathbf{z}_k = \sum_U \mathbf{z}_k$$

For them:

$$\mathbf{z}_k = (1 \;\; \hat{y}_k)^T$$

$$\mathbf{x}_k = (1 \;\; [(1-\phi)\hat{y}_k + \phi yk]\,)^T$$

$\alpha(t) = 1 + t \quad \leftarrow \quad$ the GREG weight adjustment

More on West/Andridge

Notice that (in the absence of design weights)

$$\text{SMUB}(0) = \frac{Cov_1(\hat{y},y)}{Var_1(\hat{y})}(\bar{\bar{\hat{y}}} - \hat{y}^{(1)})$$ regresses y on $\hat{y}$ using OLS

$$\text{SMUB}(1) = \frac{Var_1(y)}{Cov_1(y,\hat{y})}(\bar{\bar{\hat{y}}} - \hat{y}^{(1)})$$ regresses y on $\hat{y}$ using IVLS

with y as the instrument

Final Comments

When modeling selection:

The ingredients (the right variables) are usually more important than the recipe (functional form).

We need to worry at least as much about
Type 2 error (excluding a covariate that it needed) as
Type 1 error (including a covariate that is not).