Supplementary Material to:

A two-stage strategy to accommodate general patterns of confounding in the design of observational studies

SEBASTIEN HANEUSE

Department of Biostatistics, Harvard School of Public Health, Boston, MA

shaneuse@hsph.harvard.edu

JONATHAN SCHILDCROUT

Department of Biostatistics, Vanderbilt University, Nashville, TN

jonathan.schildcrout@vanderbilt.edu

DANIEL GILLEN

Department of Statistics, University of California - Irvine, Irvine, CA

dgillen@uci.edu

A Introduction

In this document we provide substantive and technical detail for the manuscript *A two-stage strategy to accommodate general patterns of confounding in the design of observational studies*. Specifically, the remainder of the document is organized as follows

- Section B provides summary information on the infant mortality data described in Section 2 of the manuscript.
- Section C outlines the inputs required by the simulation-based algorithm, as well as a simple strategy for specifying the numerical value of β_0.
- Section D presents the algorithm.
- Section E provides a simple framework for constructing multivariate distributions that are composed of mixtures of continuous and categorical covariates
- Section F provides a graphical representation of the sampling distribution of the estimated power using the proposed two-stage strategy under the set-up explained in Section 4.2 of the main manuscript.

We note that, to distinguish this document from the main manuscript, we have used numeric labels for sections and equations in the main paper and alpha-numeric labels in this document. Finally, we note that the approach for specifying β_0, in Section C and the algorithm of Section D are both implemented in the tpsDesign package for the statistical programming language R (http://cran.r-project.org/).

B North Carolina infant mortality data

Table SM-1: Population characteristics of the North Carolina infant mortality data and results from a multivariate logistic regression analysis of infant mortality. The former provides marginal information on the joint distribution of the exposure and six confounders, denoted $\Pr(X, Z_1, \ldots, Z_6)$ in the main manuscript; the latter provides information on the assumed confounder effects of model (C.1) (i.e. $\{\beta_{z_1}, \ldots, \beta_{z_p}\}$).

	Population characteristics N (%) or median (IQR)		Multivariate logistic regression
	Births	Deaths	OR (95% CI)
Total	225,152	1,752	
Race (X)			
caucasian	171,714 (76.3%)	986 (56.3%)	REF
African-American	53,438 (23.7%)	766 (43.7%)	1.14 (1.02, 1.28)
Gender (Z_1)			
Female	109,895 (48.8%)	761 (43.4%)	REF
Male	115,257 (51.2%)	991 (56.6%)	1.25 (1.12, 1.39)
Mothers age, years (Z_2)	27 (22 - 31)	25 (21 - 31)	0.89 (0.85, 0.93)
Smoking during pregnancy (Z_3)			
No	196,506 (87.3%)	1,382 (78.9%)	REF
Yes	28,646 (16.7%)	370 (21.1%)	1.65 (1.44, 1.89)
Weight gained by mother, lbs (Z_4)	30 (20 - 40)	18 (10 - 30)	0.86 (0.82, 0.90)
Low birth weight (Z_5)			
No	205,154 (91.1%)	488 (27.9%)	REF
Yes	19,998 (8.9%)	1,264 (72.1%)	2.24 (1.90, 2.64)
Gestation, weeks (Z_6)	39 (38 - 40)	28 (23 - 37)	0.27 (0.26, 0.29)

Note, the interpretation of the odds ratios for the three continuous confounders are for the following contrasts:

Z_2 odds ratio is for a 5 year difference in age

Z_4 odds ratio is for a 10lb difference in weight

Z_6 odds ratio is for a 4 week difference in gestational duration

C Algorithm inputs

Suppose a case-control study is conducted to estimate the association between a binary outcome Y and an exposure X, adjusting for a set of p potential confounders $Z = \{Z_1, \ldots, Z_p\}$. Given data collected via the case-control design, the primary analysis would consist of fitting a logistic regression model of the form:

$$\text{logit}\,\Pr(Y = 1|\ X, Z) \ = \ \beta_0 \ + \ \beta_x X \ + \ \sum_{j=1}^{p} \beta_{z_j} Z_j. \tag{C.1}$$

The simulation-based algorithm of Section D, and used throughout the main manuscript, requires the following inputs:

(i) β_x, the log-odds ratio coefficient for the primary exposure in model (C.1).

(ii) $\{\beta_{z_1}, \ldots, \beta_{z_p}\}$, the p log-odds ratio confounder coefficients in model (C.1).

(iii) β_0, the intercept in model (C.1), or $\tilde{\pi}_y = \Pr(Y = 1)$, the overall outcome prevalence.

(iv) $\Pr(X, Z_1, \ldots, Z_p)$, the joint exposure/confounder distribution in the population of interest.

(v) n_0 and n_1, the control and case sample sizes, respectively.

In practice, it will often be easier to elicit the overall outcome prevalence, $\tilde{\pi}_y$, than the value of the intercept, β_0, which corresponds to the outcome prevalence for the subset of the population with $X = Z_1 = \ldots = Z_p = 0$. This is particularly problematic when one changes the structure of the model that underlies the power calculations. For example, in many settings researchers will vary β_x to explore power at various effect

sizes. Doing so without modifying β_0 (holding everything else constant) will lead to differing induced $\tilde{\pi}_y$ in the underlying population across the various scenarios. The also applies when one varies the confounder coefficients. Hence, for a fixed outcome prevalence, β_0 will need to be re-calculated for each scenario that modifies β_x or any of the β_{z_j}. Given $\tilde{\pi}_y$, together with β_x, $\{\beta_{z_1}, \ldots, \beta_{z_p}\}$ and $\Pr(X, Z_1, \ldots, Z_6)$, one can determine the induced value of β_0 via

$$\beta_0 \;=\; \text{argmin}_{\beta_0} \left| \tilde{\pi}_y \;-\; \int_x \int_z \Pr(Y = 1|\; X = x, Z = z) \Pr(X = x, Z = z) \partial x \partial z \right| \quad \text{(C.2)}$$

where $P(Y = 1|\; X, Z)$ is given by model (C.1). Practically, the integration can be replaced by simulating a large dataset from $\Pr(X, Z_1, \ldots, Z_6)$ and averaging over the $\Pr(Y = 1|\; X = x, Z = z)$ (see steps (a) and (b) of the algorithm in Section D). Specifically, generate a large dataset consisting of $\{X_i, Z_{1,i}, \ldots, Z_{p,i}\}$ for N individuals, and calculate the corresponding π_i=$\Pr(Y_i = 1|\; X_i, Z_i)$. Noting the latter are a function β_0, via model (C.1), find the value of β_0 that minimizes:

$$\left| \tilde{\pi}_y \;-\; \frac{1}{N} \sum_{i=1}^{N} \pi_i \right|. \quad \text{(C.3)}$$

D Simulation-based power algorithm

One fundamental conceptual challenge in the development of sample size/power methods for case-control studies is that the design is *retrospective* whereas the model of interest is typically *prospective*; that is, exposures/confounders are random, rather than the outcome. Given the inputs outlined in Section C, the following provides a general-purpose algorithm for power calculations under the case-control design that overcomes this challenge:

(a) Construct a large 'population' of N individuals with joint exposure/confounder distribution $\Pr(X, Z_1, \ldots, Z_6)$. That is, construct a dataset consisting of $\{X_i, Z_{i,1}, \ldots, Z_{i,p}\}$ for $i = 1, \ldots, N$.

(b) Given $\{\beta_0, \beta_x, \beta_{z_1}, \ldots, \beta_{z_p}\}$, calculate $\pi_i = \Pr(Y_i = 1)$ for all N individuals using model (C.1).

(c) For each individual, generate a random draw from a Bernoulli(π_i) distribution.

(d) Stratify population according outcome status, to give N_1 cases with Y=1 and N_0 non-cases with Y=0 (note, $N_0 + N_1 = N$).

(e) Sample n_0 individuals from the N_0 non-cases and n_1 from the N_1 cases, and 'record' their exposure/confounder values.

(f) Fit model (C.1) and record whether or not the null hypothesis (i.e. H_0: β_x=0) is rejected.

(g) Repeat steps (c)-(f) R times.

Power is estimated as the percent of R instances where the null hypothesis is rejected. Since R is (necessarily) finite, the simulation-based estimate is subject to uncertainty, often referred to as Monte Carlo Error (MCE). Koehler et al. (2009) describe various techniques for characterizing MCE and strategies for determining the value of R.

References

Koehler, E., E. Brown, and S. Haneuse (2009). On the assessment of Monte Carlo Error in simulation-based statistical analyses. The American Statistician 63(2), 155–162.

E Constructing multivariate distributions for $\Pr(X, Z_1, \ldots, Z_6)$

Stage I of the proposed framework in Section 4 of the manuscript seeks to establish initial estimates of power under a range of different scenarios for confounding. Scenarios for confounding are defined, primarily, in terms of the associations between the confounder(s) and both the exposure of interest and outcome. The former requires the specification and construction of $\Pr(X, Z_1, \ldots, Z_6)$; the latter requires $\{\beta_{z_1}, \ldots, \beta_{z_p}\}$. While elicitation and specification of $\Pr(X, Z_1, \ldots, Z_6)$ and $\{\beta_{z_1}, \ldots, \beta_{z_p}\}$ (as well as β_x) pose important scientific challenges, constructing $\Pr(X, Z_1, \ldots, Z_6)$ such that the relationships that define the underlying confounding can be pre-specified and preserved also poses a difficult technical challenge. This is particularly the case as $(X, Z_1, \ldots, Z_p)$ is multivariate and may consist of a mixture of continuous and categorical covariates (as in Table SM-1 above).

As a simple way forward, we exploit the multivariate Normal distribution and generate data from the following joint exposure/confounder distribution:

$$\begin{pmatrix} X \\ Z_1 \\ \vdots \\ Z_p \end{pmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \Sigma_{xz_1} & \ldots & \Sigma_{xz_p} \\ \Sigma_{xz_1} & \sigma_{z_1}^2 & \ldots & \Sigma_{z_1 z_p} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{xz_p} & \Sigma_{z_1 z_p} & \ldots & \sigma_{z_p}^2 \end{bmatrix} \right). \tag{E.1}$$

Suppose the primary exposure of interest is binary (as with the race indicator for the infant mortality example), while the p confounders are all taken to be continuous. Given $\boldsymbol{\Sigma}$, generating random deviates from (E.1) is straightforward in most statistical packages. One can generate binary exposure with a pre-specified marginal prevalence

by dichotomizing the continuous random deviate at the appropriate quantile of the Normal(0, σ_x^2) distribution. Further, one can specify the off-diagonals of $\mathbf{\Sigma}$ to ensure some desired odds ratio association between the binary exposure and each of the confounders. To do this we note that the induced odds ratio is given by:

$$\phi_{xz} \;=\; \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}$$

where, for example,

$$\pi_{11} \;=\; \int_{q_{1-\tau_x}}^{\infty} \int_{q_{1-\tau_z}}^{\infty} f_{X,Z}(x,z)\; \partial x \partial z,$$

τ_x and τ_z are the marginal prevalences of X and Z, respectively, and $f_{X,Z}(\cdot,\cdot)$ is the MVN density.

F Operating characteristics for the two-stage design

Figure SM-1: Boxplots of the distribution of estimated power to detect θ_x=1.3, based on a case-control design with n_0=n_1=2,500 and using the fully adjusted model for the infant mortality example, using information on $\Pr(X, Z_1, \ldots, Z_6)$ and $\{\beta_{z_1}, \ldots, \beta_{z_6}\}$ obtained via internal pilot data with m=250, 500 and 1,000.