Maximum Likelihood Estimation of Structure Nested Logistic Models with an Instrumental Variable

Roland A. Matsouaka
Dept. of Epidemiology, Harvard School of Public Health
rmatsoua@hsph.harvard.edu

Joint work with **Eric J. Tchetgen Tchetgen**
Depts. of Biostatistics and Epidemiology, Harvard School of Public Health

UAI Approaches to Causal Structural Learning Workshop
Bellevue, Washington
July 15, 2013

Outline

Confounding and Instrumental Variable (IV)

- In observational study or in clinical trial with non-compliance, treatment assignment is not completely under the control of the investigators;
- Confounding or non-ignorable selection of treatment may lead to spurious estimate of the treatment effect, providing biased and inconsistent results;
- IV study designs have been used abundantly to estimate treatment effect when confounding is present or suspected.
- In terms of potential outcomes, an IV for the effect of X on Y is a variable Z s. t. (1) $X \not\perp Z$; (2) $Y_{xz} = Y_x$ a. s. $\forall\, x, z$; and (3) $Y_{xz} \perp Z | L$, $\forall\, x, z$.

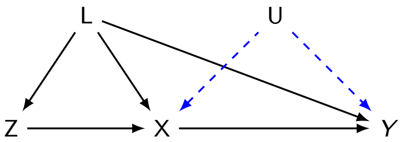

Figure 1: A graph showing an instrumental variable Z, a measured covariate L, an unmeasured confounding U, an exposure X, and an outcome Y.

G-estimation and SNMMs

- G-estimation and SNMMs: introduced by Robins (1989, 1994) to infer causality in studies where confounding might be an issue;
- G-estimation is a semiparametric method to estimate parameters in SNMMs.
- Let $b =$ link function; we define SNMMs as

$$b(E(Y_x|x, z, l)) - b(E(Y_0|x, z, l) = \gamma(x, z, l).$$

  $b =$id, log, and logit for, resp., additive, multiplicative, and logistic SNMMs.
- The contrast γ is a conditional causal effect: compares the average potential outcomes under active and inactive trt values on a scale given by b for the subset of the population with x, z, l.

Identification

- IV assumptions (1) to (3) not sufficient to identify γ nonparametrically when $b =$id, log, or logit (Robins (1994), Robins and Rotnitzky (2004)).
- Additional assumption is required: (4) $\gamma(x, z, l)$ is restricted i.e. $\gamma(x, z, l) = \gamma(x, z, l; \psi)$ s.t. it is identified under assumptions (1) to (3).

  For logistic SNMMs i.e. when $b =$logit,
  - Vansteelandt and Goetghebeur (2005) study impact of violations of (4) ;
  - Alternative identification conditions other than (4) are of great interest and are an ongoing research topic (Tchetgen Tcheten and Vansteelandt (2013));
  - Richardson and Robins (2010), Richardson et al. (2011) elucidate the issue of identification and conduct a careful analysis for a binary IV model.
- Throughout this presentation, we will assume that the parametric model $\gamma(x, z, l; \psi)$ is correctly specified, with unkown ψ to be estimated.

Back to G-estimation

- Under assumptions (1) to (4) trt effects for additive and multiplicative SNMMs can be estimated via G-estimation (Robins (1984, 1994) and RR);
- Logistic SNMMs cannot be estimated with G-estimation (VG, RR);
- VG propose an "association" model $\text{logit}E(Y|x, z, l) = m(x, z, l; \eta)$, which is used together with a model for the IV to estimate γ.
- Caveat: the association model must be saturated. A non-saturated association model implies further identifying assumptions that may be incompatible with the logistic SNMM model (a. k. a. uncongeniality.)

Robins and Rotnitzky's Parametrization

Robins and Rotnitzky (2004) propose a parametrization based on the contrast

$$\text{logit}E(Y_0|x,z,l) - \text{logit}E(Y_0|x=0,z,l) = q(x,z,l;\eta)$$

which is always guaranteed to be congenial. This parametrization is such that

$$\text{logit}P(Y=1|x,z,l) = \gamma(x,z,l) + q(x,z,l) + v(z,l),$$

with $v(z,l)$= $\text{logit}(E(Y_0|x=0,z,l))$ the unique solution to the integral equation

$$\text{logit}\int \text{expit}\{q(x^*,z,l) + v(z,l)\}dF(x^*|z,l) = \text{logit}\{E(Y_0|l)\} = t(l).$$

- ☞ Integral equation must be solved for v, for each observation. Unfortunately, it can't be solved in closed form for most choices of q, F_x and t, except for few cases (e.g. X is binary)

RR's Parametrization: Numerical Optimization

Integral equation can't be solved in closed form for most choices of q, F_x and t.

Numerical optimization of the likelihood for the joint density of the observables using the parametrization proposed by RR involves solving numerically an integral equation for each observed (Z, L), within each iteration of the algorithm.

- If the trt X takes more than 2 values, is continuous or multivariate, this approach becomes computationally challenging; particularly when:
  - Z is a continuous IV;
  - there are a large number of covariates L
- These numerical drawbacks have impeded the widespread use of this approach despite its mathematical and theoretical underpinning.

So What Do We Propose?

- A new parametrization and relate it to VG and RR. This parametrization
  - is congenial and circumvents computational complexity of RR's parametrization;
  - provides MLE of the joint density of observables, using standard softwares L
- A goodness-of-fit test statistic evaluating certain parametric assumptions of the fitted likelihood model
- A simulation study to illustrate our method and provide some insights about our approach

New Parametrization

As in RR, we consider the congenial parametrization

$$\text{logit}P(Y=1|x,z,l)=\text{logit}f_y(1|x,z,l)=\gamma(x,w;\psi)+q(x,z,l)+v(z,l)$$

$$\begin{aligned}\text{where}\quad v(z,l) &= \text{logit}P(Y_0=1|x=0,z,l))\\ &= -\left[\text{logit}P(Y_0=1|z,l)-\text{logit}P(Y_0=1|x=0,z,l)\right]+\text{logit}P(Y_0=1|z,l)\\ &= \log\left[\frac{P(Y_0=1|z,l)}{1-P(Y_0=1|z,l)}\right]-\log\left[\frac{1-P(Y_0=1|x=0,z,l)}{P(Y_0=1|x=0,z,l)}\right]+t(l).\end{aligned}$$

Since

$$\text{ODDS}(z,l)=\int \text{ODDS}(x,z,l)dF_{x,0}(x|w,Y_0=0)=\int\frac{P(Y_0=1|x,z,l)}{P(Y_0=0|x,z,l)}dF_{x,0}(x|w,Y_0=0)$$

with $F_{x,0}(x|w,Y_0=0)$ the CDF of X given Z,L and $Y_0=0$, it follows that

$$\begin{aligned}v(w) &= -\log\int \text{ODDS}(x,z,l)dF_{x,0}(x|w,Y_0=0)+\log\left[\text{ODDS}(x=0,z,l)\right]+t(l)\\ &= -\log\int\frac{\text{ODDS}(x,w)}{\text{ODDS}(x=0,w)}dF_{x,0}(x|w,Y_0=0)+t(l)=-\log\int\exp[q(x,w;\eta)]dF_{x,0}+t(l)\\ &= -\overline{q}(z,l)+t(l).\end{aligned}$$

Thus, $\text{logit}f_y(1|x,z,l)=\gamma(x,w;\psi)+q(x,z,l)-\overline{q}(z,l)+t(l)$

New Parametrization (Continued)

The observed data likelihood factorizes as $f_y(Y_X|X,W;\psi)f_x(X|W)f_z(Z|L)f_l(L)$

The result

$$\text{logit} f_y(1|x,z,l) = \gamma(x,w;\psi) + q(x,z,l) - \overline{q}(z,l) + t(l)$$

means that we are free to choose (parametric) models for $q(x,w;\eta)$, $t(l;\omega)$, and $f_{x,0}(x|w, Y_0 = 0;\alpha)$. However, the distribution of the exposure X is given by

$$\begin{aligned} f_x(x|z,l) &= f_{x,0}(x|z,l,Y_0=0)P(Y_0=0|z,l) + f_{x,0}(x|z,l,Y_0=1)P(Y_0=1|z,l) \\ &= f_{x,0}(x|z,l,Y_0=0)\left[(1-\Phi(t(l)) + \frac{\exp(q(x,z,l))}{\exp(\overline{q}(z,l))}\Phi(t(l))\right], \end{aligned}$$

where $\Phi(u) = b^{-1}(u) = 1/(1+exp(-u))$.

Maximum Likelihood Estimation

The resulting MLE is given by

$$\prod_i f_y(Y_i|X_i, Z_i, L_i; \psi, \eta, \alpha, \omega) f_x(X_i|Z_i, L_i; \alpha, \omega, \eta) f_z(Z|L; \kappa)$$

- The above likelihood can be maximized using standard softwares such as PROC NLMIXED in SAS or the *optim* function in R.
- Even if the choice of the models does not permit a closed form expression for the integral

$$\overline{q}(z, l) = \int \exp[q(x, z, l; \eta)] dF_{x,0}(x|z, l, Y_0 = 0),$$

it can still be estimated using Gauss-Hermite quadrature integral approximation (see Liu and Pierce (1985))

Goodness-of-fit Test Statistic

RR characterize the set of influence functions for γ for the semiparametric model defined by the assumptions (1) to (4). We use a scalar function from this set to construct a semiparametric GOF test statistic for the likelihood model. Let

$$\widehat{M}_1 = \gamma(X, Z, L; \widehat{\psi}) + q(X, Z, L; \widehat{\eta})\overline{q}(Z, L; \widehat{\eta}, \widehat{\alpha}) + t(L, \widehat{\omega}),$$
$$\widehat{M}_2 = q(X, Z, L; \widehat{\eta}) - \overline{q}(Z, L; \widehat{\eta}, \widehat{\alpha}) + t(L, \widehat{\omega})$$

where $(\widehat{\psi}, \widehat{\eta}, \widehat{\alpha}, \widehat{\omega})$ is the MLE obtained in the previous section.

The GOF test statistic is given by

$$T = \frac{\sum_i \widehat{U}_i}{\sqrt{\sum_i \widehat{U}_i^{\,2}}}$$

where $\widehat{U} = (Z - \mathbb{E}(Z|L; \widehat{\kappa}))\left[\frac{\widehat{\Phi}(M_2)(1-\widehat{\Phi}(M_2))}{\widehat{\Phi}(M_1)(1-\widehat{\Phi}(M_1))}(Y-\Phi(\widehat{M}_1))+\Phi(\widehat{M}_2)\Phi(t(L,\widehat{\omega}))\right].$

Simulated Data

- $L = (L_1, L_2)$ s.t. $L_1 \sim N(3,1)$, $L_2 \sim N(2,1)$
- $Z \sim Bernoulli(p_z)$, $\text{logit}(p_z) = \kappa_0 + \kappa_1 L_1 + \kappa_2 L_2 = -0.1 + 0.5L_1 + 0.2L_2$.
- $p_t = b^{-1}(t(L,\omega))$, $\text{logit}(p_t) = \omega_0 + \omega_1 L_1 + \omega_2 L_2 = -1 + 0.5L_1 + 0.3L_2$
- $q(X, Z, L; \eta) = \eta X = -0.4X$.
- $X|Z, L, Y_0 = 0 \sim Bernoulli(p_{x0})$, $\text{logit}\ (p_{x0}) = -0.4 - 0.3L_1 + 0.3L_2 + Z$.
  - $X \sim Bernoulli(p_x)$, $p_x = (1 - p_t)p_{x0} + p_t p_{x1}$ and $\text{logit}(p_{x1}) = \text{logit}(p_{x0}) - 0.4$.
- $X|W,\ Y_0 = 0 \sim N(\mu_{x0}, \sigma^2)$ with $\mu_{x0} = 1 - 2L_1 + L_2 + 3Z$.
  - $X \sim (1 - p_t)N(\mu_{x0}, \sigma^2) + p_t N(\mu_{x0} - 0.4\sigma^2, \sigma^2)$.
- $Y \sim Bernoulli(p_y)$, $\text{logit}(p_y) = (1 - 0.4)X - \bar{q}(Z, L) + t(L)$,

Parameter Estimates

Table 1: Results: Binary Exposure

MLE	Bias	MSE	Coverage	S.E.
ψ	0.002	0.102	0.96	0.319
η	-0.004	0.106	0.95	0.326
α_0	0.005	0.007	0.95	0.083
α_1	-0.002	0.001	0.95	0.038
α_2	-0.001	0.001	0.94	0.033
α_3	0.001	0.003	0.95	0.054
ω_0	0.006	0.031	0.96	0.177
ω_1	0.000	0.002	0.95	0.041
ω_2	-0.001	0.002	0.95	0.040
κ_0	-0.001	0.004	0.95	0.065
κ_1	0.000	0.001	0.94	0.032
κ_2	0.001	0.000	0.95	0.030

GOF test: Type I error $= 0.01$

Table 2: Results: Continuous Exposure

MLE	Bias	MSE	Coverage	S.E.
ψ	0.012	0.001	0.95	0.042
η	0.000	0.003	0.95	0.052
α_0	0.000	0.004	0.94	0.062
α_1	0.000	0.000	0.95	0.016
α_2	0.000	0.000	0.95	0.014
α_3	-0.001	0.002	0.94	0.042
ω_0	0.003	0.029	0.94	0.172
ω_1	0.000	0.005	0.95	0.073
ω_2	0.002	0.003	0.95	0.054
κ_0	-0.004	0.019	0.95	0.138
κ_1	0.001	0.001	0.95	0.041
κ_2	0.002	0.001	0.95	0.040
σ	-0.001	0.000	0.95	0.010

GOF test: Type I error $= 0.039$.

Power of Goodness-of-fit Test Statistic

Table 3: Goodness-of-fit Test: Power

Misspecified Model	Missing covariates [a]	Parameter Values [a]	Power
	(1) *Binary Exposure*		
$q(X,Z,L)$	X^2	1.5	0.15
	$Z,\ X\times Z$	-0.6, 1.5	0.41
$t(L)$	L_2^2	1.5	0.40
	$L_1\times L_2$	0.7	0.03
	L_2	1.5	0.89
	(2) *Continuous Exposure*		
$q(X,Z,L)$	X^2	-0.4	0.95
	$Z,\ X\times Z$	0.6, -1.5	0.62
$t(L)$	L_2^2	0.6	0.43
	$L_1\times L_2$	0.6	0.06
	L_2	0.6	0.14

[a] Covariates used in the generated model, but omitted in the fitted model.

Summary

- We have proposed a new parametrization for logistic SNMMs, presented a corresponding MLE approach, and a GOF test statistic
- Our approach builds upon the theoretical frameworks of VG and RR. Unlike VG, but similar to RR, our approach is guaranteed to always be congenial
- Unlike RR we obviate the need to solve numerically an integral equation, which can be computationally cumbersome and is not easily scalable with the dimension of the exposure.
- Our approach is readily implemented using standard statistical softwares.

Thank You!

Roland Matsouaka
rmatsoua@hsph.harvard.edu

References

1. Liu, Q., Pierce, D.A.(1994), "A note on Gauss-Hermite quadrature", *Biometrika*, **81**: 624-629.

2. Richardson, T. S., Evans, R. J., and Robins, J. M. (2011), "Transparent parametrizations of potential outcome models (with discussion)," in Bayesian Statistics 9, J.M. Bernardo, M.J. Bayarri, J. O. Berger, A. P. Dawid, D. Hekerman, A. F. M. Smith and M. West (Eds.), pp. 569-610.

3. Richardson, T. S. and Robins, J. M. (2010), "Analysis of the binary instrumental variable model," Heuristics, Probability and Causality: a tribute to Judea Pearl. R. Dechter, H. Geffner, and J. Halpern, Editors, College Publications, UK, pp. 415-444.

4. Robins, J. M. (1986), "A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect," *Math. Mod.*, **7**: 1393-.

5. Robins, J. M. (1989), "The Analysis of Randomised and Non-Randomised AIDS Treatment Trials Using a New Approach To Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, eds. L. Sechrest, H. Freeman, and A. Mulley, Washington, DC: US Public Health Service, National Center for Health Services Research, pp. 113-159.

References (Cont.)

6. Robins, J. M. (1994), "The Analysis of Randomised and Non-Randomised AIDS Treatment Trials Using a New Approach To Causal Inference in Longitudinal Studies," *Comm. Statist. - Theory and Methods*, **23**: 2379-2412.

7. Robins, J. M., and Rotnitzky, A. (2004), "Estimation of Treatment Effects in Randomised Trials With Non-Compliance and a Dichotomous Outcome Using Structural Mean Models," *Biometrika*, **91**: 763 - 783.

8. Tchetgen Tchetgen E., Robins J.M., Rotnitsky A. (2010). "On doubly robust estimation in a semiparametric odds ratio model" *Biometrika*, **97**(1):171-180.

9. Tchetgen Tchetgen and Vansteelandt (2013), "Alternative identification and inference of the effect of treatment on the treated with an instrumental varable." Submitted.

10. Vansteelandt, S. and Goetghebeur, E. (2003), "Causal Inference with Generalized Structural Mean Models", *J. R. Stat. Soc.*, Series B, **65**, 817-835.

11. Vansteelandt, S., Bowden, J, Babanezhad, M., and Goetghebeur, E. (2011), "On instrumental variable estimation of causal odds ratios ", *Statist. Sci.*, **26**, 403-422.